

**New approaches and software for
sequence function prediction and gene design
integrating specific knowledge on protein
structure and translation**

Dissertation

zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades

»**Doctor rerum naturalium**«

der Georg-August-Universität Göttingen

im Promotionsprogramm Computer Science (PCS)
der Georg-August University School of Science (GAUSS)

vorgelegt von

Dominic Alexander Simm

aus Dresden

Göttingen, 2022

Betreuungsausschuss

Prof. Dr. Stephan Waack
Institut für Informatik, Abtl. Theoretische Informatik
Georg-August-Universität Göttingen

Prof. Dr. Gerhard H. Braus
Institut für Mikrobiologie und Genetik, Abtl. Molekulare Mikrobiologie und Genetik
Georg-August-Universität Göttingen

PD Dr. Martin Kollmar
Forschungsgruppe Systembiologie der Motorproteine
Max-Planck-Institut für biophysikalische Chemie, Göttingen

Mitglieder der Prüfungskommission

Referent/in: Prof. Dr. Stephan Waack
Institut für Informatik, Abtl. Theoretische Informatik
Georg-August-Universität Göttingen

Korreferent/in: PD Dr. Martin Kollmar
Forschungsgruppe Systembiologie der Motorproteine
Max-Planck-Institut für biophysikalische Chemie, Göttingen

Weitere Mitglieder der
Prüfungskommission: Prof. Dr. Gerhard H. Braus
Institut für Mikrobiologie und Genetik, Abtl. Mol. Mikrobiologie und Genetik
Georg-August-Universität Göttingen

Prof. Dr. Carsten Damm
Institut für Informatik, Abtl. Theoretische Informatik
Georg-August-Universität Göttingen

Prof. Dr. Burkhard Morgenstern
Institut für Mikrobiologie und Genetik, Abtl. Bioinformatik
Georg-August-Universität Göttingen

Prof. Dr. Dieter Hogrefe
Institut für Informatik, Abtl. Telematik
Georg-August-Universität Göttingen

Tag der mündlichen Prüfung: 16. Februar 2022

I hereby declare that this thesis was written independently and with no other sources and aids than quoted.

Göttingen, January 3rd, 2022

Dominic Simm

Abstract

In recent years, advances in sequencing techniques resulted in an explosive increase in sequencing data. By now, there are billions of generated protein sequences available in public databases to a large extent without experimentally proven structure and function. Computational approaches and bioinformatic analysis, like the ones presented here, complement and assist experimentally intensive research efforts by processing, analyzing and interpreting the produced amounts of biological data and by giving first forecasts to uncover hidden knowledge from the data and help answering fundamental biological questions.

Coiled-coil prediction software has an important role in one of the first steps of the structural annotation of newly generated protein sequences with unknown molecule structure. Unfortunately established software have been shown to have a rather limited applicability especially in terms of the prediction quality with regard to large-scale coiled-coil analyses. The web-application »Waggawagga« was developed for the comparative visualization of coiled-coil predictions generated by different software packages (Simm et al., 2015). As a basis for decision-making over a coiled-coil domain in question, the strength of the majority consensus of multiple and freely combinable prediction tools builds the central aspect of this comparative approach to overcome the limitations of single applications. Supportive hints are provided by the specially developed SAH prediction algorithm, that enables a discrimination between putative coiled coils and actual single α -helix domains (SAH) and helps in the identification of real coiled-coil domains. The developed SAH prediction algorithm is both part of the web-application and available as a stand-alone version for the command line, termed »Waggawagga-CLI«. Its function has been tested and evaluated in detail in two studies that investigated the distribution and evolution of predicted SAH domains in the myosin motor protein family (Simm et al., 2017) and in two dozen eukaryotic organisms across the tree of life (Simm and Kollmar, 2018). The results revealed that SAH-domains occur in 0.5 to 3.5% of the protein-coding content per investigated species and are particularly present in longer proteins supporting their function as structural building block in multi-domain proteins. In addition, a large-scale in-depth prediction analysis was performed by testing the most relevant softwares of the field against the most comprehensive reference data set available, the entire Protein Data Bank, and tracked down the results to each amino acid and its secondary structure (Simm et al., 2021). Comparing the binary classifications metrics with naïve coin-flip models suggests that the tested tools' performance is close to random. This implicates that the tools' predictions have only limited informative value, should be treated very cautiously and need to be supported and validated by experimental evidence.

Heterologous protein expression is often applied in the investigation of cellular functions, in genetic circuit engineering, in overexpressing proteins for biopharmaceutical applications and structural biology research. One of the key factors for heterologous expression represents the degeneracy of the genetic code, which enables a single protein to be encoded by a multitude of synonymous gene sequences and simultaneously allows adjusting gene sequences without changing the protein sequences, substantial uncertainty exists concerning the details of this phenomenon. With the de-

velopment of the software »Odysseus«, we realized a new probabilistic approach that exploits this particular genetic property to design typical genes for heterologous expression applications in common model organisms (Simm et. al, submitted). The Markov model based approach is highly configurable and can be operated with pre-trained genome profiles to control protein expression levels by the codon usage adaptation of genes. We evaluated the influence of the profiled codon usage adaptation approach on protein expression levels in the eukaryotic model organism *Saccharomyces cerevisiae*. Therefore, we selected green fluorescent protein (GFP) and human α -synuclein (α Syn) as representatives for stable and intrinsically disordered proteins as representing a benchmark and a challenging test case. GFP was expressed at high levels, and the toxic α Syn could be adapted to endogenous, low-level expression. The new software is publicly available as a web-application for performing host-specific protein adaptations to a set of the most commonly used model organisms.

Preface

This dissertation is the result of my PhD project in the »*Doctoral Program Computer Science (PCS)*« at the Georg-August-Universität Göttingen. It was developed under the supervision of Prof. Dr. Stephan Waack and PD Dr. Martin Kollmar as a collaborative project in the research groups »*Theoretical Computer Science and Algorithmic Methods*« of the Institute of Computer Science at the University of Göttingen and »*Systems Biology of Motor Proteins*«, which was located in the Department of NMR-based Structural Biology at the Max Planck Institute for Biophysical Chemistry in Göttingen.

The research group »*Theoretical Computer Science and Algorithmic Methods*« focuses on research topics in which basic theoretical algorithms and concepts, such as probabilistic models and information-theoretic methods, can be applied and investigated. One of their core research topics is statistical sequence analysis, which finds its application in bioinformatics.

The research group »*Systems Biology of Motor Proteins*« dealt with the elucidation of motor protein-mediated processes in eukaryotic cells, which play an important role in muscle functions, neuronal transport and cell division. To this end, in recent years the group has developed experimental methods to gain insight into the function of the dynein/dynactin motor protein complex at the atomic level and computational methods to determine motor protein content in eukaryotes.

Contents

| | |
|---|------------|
| Abstract | IV |
| Preface | VI |
| Contents | VII |
| 1. Introduction | 1 |
| 1.1. Motivation | 2 |
| 1.1.1. Heterologous gene expression | 2 |
| 1.1.2. Protein function and structure prediction | 3 |
| 1.2. Scope of the thesis | 4 |
| 1.2.1. Comparative domain prediction of coiled coils and SAHs | 4 |
| 1.2.2. Design of typical genes for heterologous gene expression | 5 |
| 1.3. Impact | 5 |
| 1.4. Structure of the thesis | 7 |
| 2. Foundations | 8 |
| 2.1. Background | 8 |
| 2.1.1. Protein structure and domain prediction | 10 |
| 2.1.2. Heterologous gene expression | 11 |
| 2.2. Biological foundations | 13 |
| 2.2.1. The genetic code | 13 |
| 2.2.2. Proteins | 14 |
| Green fluorescent protein | 15 |
| Alpha-synuclein | 16 |
| 2.2.3. Coiled coils | 17 |
| 2.2.4. Stable single-alpha helices | 20 |
| 2.3. Bioinformatical foundations | 22 |
| 2.3.1. Probabilistic models | 22 |
| Markov chains | 22 |
| 2.3.2. Protein structure und function prediction | 24 |
| 2.3.3. Coiled-coil prediction | 25 |
| Position-specific scoring matrix (PSSM) | 27 |
| Pairwise residue correlations (extended PSSM-variant) | 28 |
| Hidden Markov model (HMM) | 28 |
| 2.3.4. Analysis software for secondary structure prediction | 30 |
| Database of secondary structure assignments (DSSP) | 30 |
| SOCKET | 31 |

| | |
|--|-----------|
| 2.3.5. Biological databases and exchange formats | 32 |
| Protein Data Bank | 32 |
| CC+ Database | 34 |
| PaxDB | 35 |
| REBASE | 36 |
| 2.4. Computational approaches | 37 |
| 3. Publications and Manuscripts | 38 |
| 3.1. Comparative protein structure prediction of coiled-coil and SAH domains | 39 |
| 3.1.1. Waggawagga: comparative visualization of coiled-coil predictions and detection of stable single α -helices (SAH domains) | 39 |
| 3.1.1.1. Abstract | 40 |
| 3.1.1.2. Introduction | 40 |
| 3.1.1.3. Features | 41 |
| 3.1.1.4. Implementation | 42 |
| 3.1.1.5. Acknowledgement | 42 |
| 3.1.1.6. References | 42 |
| 3.1.2. Distribution and evolution of stable single α -helices (SAH domains) in myosin motor proteins | 43 |
| 3.1.2.1. Abstract | 44 |
| 3.1.2.2. Introduction | 44 |
| 3.1.2.3. Results and discussion | 45 |
| 3.1.2.4. Methods | 54 |
| 3.1.2.5. Acknowledgements | 57 |
| 3.1.2.6. References | 57 |
| 3.1.3. Waggawagga-CLI: A command-line tool for predicting stable single α -helices (SAH-domains), and the SAH-domain distribution across eukaryotes | 60 |
| 3.1.3.1. Abstract | 61 |
| 3.1.3.2. Introduction | 62 |
| 3.1.3.3. Results and discussion | 63 |
| 3.1.3.4. Conclusions | 74 |
| 3.1.3.5. Materials and methods | 75 |
| 3.1.3.6. Acknowledgements | 78 |
| 3.1.3.7. References | 78 |
| 3.1.4. Critical assessment of coiled-coil predictions based on protein structure data | 80 |
| 3.1.4.1. Abstract | 80 |
| 3.1.4.2. Authour summary | 80 |
| 3.1.4.3. Introduction | 80 |
| 3.1.4.4. Results | 80 |
| 3.1.4.5. Discussion | 80 |
| 3.1.4.6. Methods | 80 |
| 3.1.4.7. References | 80 |
| 3.2. Heterologous gene expression | 99 |
| 3.2.1. Design of typical genes for heterologous gene expression | 99 |
| Abstract | 100 |
| Introduction | 100 |

| | |
|---|------------|
| Materials and Methods | 101 |
| Results and Discussion | 108 |
| Conclusions | 116 |
| Data Availability | 117 |
| References | 117 |
| 3.3. Genome annotation and genomic databases | 121 |
| 3.3.1. diArk – the database for eukaryotic genome and transcriptome assemblies in 2014 | 121 |
| 3.3.1.1. Abstract | 122 |
| 3.3.1.2. Introduction | 122 |
| 3.3.1.3. Conclusions | 125 |
| 3.3.1.4. Acknowledgement | 126 |
| 3.3.1.5. References | 126 |
| 3.3.2. The landscape of human mutually exclusive splicing | 128 |
| 3.3.2.1. Abstract | 129 |
| 3.3.2.2. Introduction | 129 |
| 3.3.2.3. Results | 130 |
| 3.3.2.4. Discussion | 140 |
| 3.3.2.5. Materials and Methods | 142 |
| 3.3.2.6. Acknowledgements | 147 |
| 3.3.2.7. References | 148 |
| 3.3.3. Identifying Sequenced Eukaryotic Genomes and Transcriptomes with diArk | 151 |
| 3.3.3.1. Abstract | 152 |
| 3.3.3.2. Introduction | 152 |
| 3.3.3.3. Methods | 154 |
| 3.3.3.4. References | 170 |
| 3.3.4. The Protein-Coding Human Genome: Annotating High-Hanging Fruits | 171 |
| 3.3.4.1. Introduction | 172 |
| 3.3.4.2. Do Current Approaches Annotate Functional Regions, Junk, or Both? | 172 |
| 3.3.4.3. The Number of Human Protein-Coding Genes Went Up and Down in the Pre-Genome Era | 174 |
| 3.3.4.4. Evidence for Protein-Coding Genes Comes from Multiple Sources | 176 |
| 3.3.4.5. The Genome Is Annotated at Multiple Overlapping Layers? | 176 |
| 3.3.4.6. Is There Consensus on the Low-Hanging Fruits? | 177 |
| 3.3.4.7. Finding All Protein-Coding Segments Remains Difficult | 178 |
| 3.3.4.8. To Each Individual Its Own Annotation | 181 |
| 3.3.4.9. Conclusions and Outlook | 181 |
| 3.3.4.10. References | 182 |
| 4. Conclusions | 186 |
| 4.1. Contributions to prediction of coiled-coil and SAH domains | 186 |
| 4.2. Contributions to heterologous expression of genes | 189 |
| References | 191 |
| Acknowledgements | 197 |

| | |
|---|------------|
| List of Figures | 198 |
| A. Appendix | 200 |
| A.1. Supplementary information | 201 |
| A.2. Supplementary figures | 203 |
| A.3. Abbreviations and acronyms | 219 |
| A.4. Curriculum vitae | 221 |

1

Introduction

Over the last decades, the emergence of modern molecular genetics has continuously changed the face of biology with far-reaching impact on numerous of its subfields. Especially, the rapid development in the field of whole-genome sequencing contributed a lot to this change. It enabled for the first time to get deep insights and an understanding of the organization and genetic composition of entire genomes, function-encoding regions and the capability to compare sequenced genomes of different organisms with each other. Major reasons for that advance are the considerable decrease in costs, technological improvements and a significant increase in speed enabled by modern high-throughput sequencing centers. As a consequence of this development, large-scale DNA-sequencing efforts from genome and metagenome projects such as the Human Genome Project (Venter et al., 2001) or the Molecular digitization of a botanical garden (Liu et al., 2019), to name only two exemplary projects, are nowadays able to generate immense amounts of data. In case of the Human Genome an amount of nearly 15-billion bp in DNA information were produced. The latter mentioned venture published in the beginning of 2019 a broad dataset with 54 terabytes of high-depth whole-genome sequencing data of 689 plant species and provides as a side-effect insight into the requirements of the next level, the »planetary-scale« projects such as the 10.000 Plant Genomes Project (Cheng et al., 2018), the Earth BioGenome Project (Lewin et al., 2018) or the 100.000 Genomes Project (Turnbull et al., 2018). The tendency is obvious, the arising sequence information is exponentially increasing and has reached an order of magnitude, where the demands for sophisticated analyses of biological sequence data are getting higher and the research fields bioinformatics and computational molecular biology are becoming more and more important.

But not only the field of sequencing technology has strongly evolved over time concerning capability and affordability, also the counterpart the synthesis of physical sequences from virtual templates. Research fields, that profit a lot from this advance are the synthetic biology and artificial gene synthesis. The technological progress helped to overcome the consisting difficulty to read and write long perfect DNA sequences, which limited previous applications to a rather small scale. In 2012 the milestone of utilizing synthetic DNA as reliable storage of digital file information could be reached enabled by next-generation synthesis. It became feasible to encode and store the information of small computer files with capacities of around 700 kilobytes (Church et al., 2012; Goldman et al., 2013) in DNA sequences including also the recovery step of the stored information by sequencing and decoding the artificial produced DNA. Designed as a feasibility study in summer of 2019, already the whole Wikipedia text information with its nearly 16 gigabytes could be written into DNA strands (Shankland, 2019) demonstrating the future potential of the technology and the fact, that the efficient and reliable synthesis of virtual genes is no obstacle anymore.

The need for computational methods supporting these kinds of applications range from information theoretical principles of modelling, over efficient structures and forms of data storage to biological sequence analysis methods. Approaches to handle and process these types of biological data are

continuously being developed. Nowadays, state of the art techniques use powerful sequence analysis methods based on principles of probabilistic modelling. An important area of application is for instance the annotation of genome assemblies. Tasks that need to be covered for this purpose reach from the identification of distant members of sequence families, the determination of the significance of sequence alignments, the inference of phylogenetic trees as well to the assignment of biological functions by protein structure prediction, to name a few examples.

1.1. Motivation

The advances in the fields next-generation sequencing and synthetic biology formed the basis for in-depth investigations to explain cellular processes, molecular structures and genetic relationships. Concerning the process of protein expression, these methods contributed a lot to the understanding of the relative roles of initiation, elongation, degradation and misfolding of synthesized molecules. The mere availability of the sequenced genome data in digital form facilitates further research fields to derive substantial knowledge and get deeper insights into the genetic information of an organism. Computational methods and statistical analyses applied to this sort of data provide different routes to uncover underlying characteristics and striking patterns. For example the special composition of each genome, even each gene comprises, regarded with a focus on the smallest building blocks, significant deeper information.

1.1.1. Heterologous gene expression

Heterologous expression of protein-coding genes is an important method in structural biology research and pharmaceuticals. Its diverse applications range from proving the homology of protein functions up to biopharmaceutical studies on overexpression of proteins, improvement of design principles for vaccine development or gene therapy. Profiting as well from the advances in next-generation sequencing and synthetic biology, the approach is generally based on the close interaction between bioinformatic concepts, computational methods and experimental techniques of molecular biology such as molecular cloning. In addition to the experimental efforts, consisting among others in the integration of the foreign gene into the DNA of an heterologous organism and the determination of the synthesised protein, the computational part for adapting the gene to the characteristics of the expression system plays an essential role in the regulation of protein expression. Therefore, the understanding of genetic key requirements and the development of appropriate computation-based models is an important research topic in bioinformatics.

One central aspect is the codon usage distribution of a sequenced genome. It is common knowledge, resulting from the redundant encoding of the 20 amino acids by 61 triplet nucleotide codons, that the genetic code is degenerated and builds the basis of the codon usage. As a consequence, the degeneracy enables a single protein to be encoded by a multitude of synonymous gene sequences and has an important role in regulating protein expression (Boël et al., 2016). The codon usage of a genome is unique for each species and behaves like a fingerprint reflecting the effects of mutations, genetic drift and evolutionary selection commonly known under the term »genome hypothesis« (Grantham et al., 1980; Sharp et al., 1988). Due to these effects, the distribution of alternative synonymous codons for each amino acid differs between distinct organisms, which means that the identity of more and less frequent synonymous codons is unequal. This inequality among species is called the codon usage bias (Hershberg and Petrov, 2008). But not only inter-species differences exist, also the frequencies of

alternative codons across genes within a genome and even between sites in a gene are not generally consistent.

Although originally presumed to be identical in function and neutral for protein folding, there exist two hypotheses explaining why certain synonymous codons are more favoured and increasingly recognized to be influential for protein biogenesis (Hershberg and Petrov, 2009; Plotkin and Kudla, 2011). One focuses on accuracy and assumes that synonymous codons differ in their rate of mistranslation, whereby the potential to synthesize misfolded molecules is lowered by incorporating more accurately translated codons. The other hypothesis describes the selection with the translation efficiency, a more systematic variation of the codon usage, meaning that synonymous codons are translated with different speeds. In general, it can be stated that the use of efficient codons leads to an increase of the translation elongation rate as well as the accuracy. But there are exceptions, rare codons on the other hand result in a reduction of translational errors by slowing down the general protein translation process, but this strongly depends on the respective gene or the part of the gene (Brase and Ridge, 2019). For the design and expression of adapted genes this influence on the translation elongation rate in expression systems plays an important role, because the optimization of the synonymous codon choice through minimization of rare codons leads often to significantly increased protein yields, especially for single-domain proteins (Burgess-Brown et al., 2008; Gorochowski et al., 2015). In the protein biosynthesis process, the simultaneous presence of two tRNAs in the A and P positions of the ribosome is necessary for the formation of peptide bonds (Nierhaus et al., 1998; Stark et al., 1997). Due to steric reasons not all combinations of codons and tRNAs are equally compatible to the ribosome surface, which means that certain codon pairs are processed more efficiently than others. The phenomenon of the non-random utilization of codon pairs is called the »codon context« and is assumed in addition to correlate with the translation elongation rate in a way that rare codon pairs decrease the rate (Coleman et al., 2008).

Former research from the literature has shown, that there are global and local factors that influence the expression level of genes in heterologous hosts. The most important global factors are the differences in codon context, the codon usage and the GC contents in coding genes between species. On this account the adaptation of the codon usage of the source gene to the global codon usage of the foreign host organism (expression system) or to a subset of highly expressed genes is assumed to play an important role for the heterologous gene expression. The most important local variables decreasing or abandoning the protein expression that need to be handled, are the mRNA secondary structures close to translation initiation sites, internal repeats and patterns resembling ribosomal binding sites. Some of these factors may lead to instable mRNAs or need to be concerned for practical reasons like the cloning process (e.g. occurrence of restriction sites) and have to be removed in advance. Currently available software for gene design use the most commonly used synonymous codons, so called »major codons«, for the target gene and allow to exclude only a few of the local restraints.

1.1.2. Protein function and structure prediction

An important topic in the field of bioinformatics and computational molecular biology is the determination of protein structures and functions to contribute to the functional annotation of genomes. Generated genome sequence assemblies consist on the level of their smallest components of »contiguous blocks of millions of sequential nucleotides encompassing every chromosome of the sequenced organisms« (Watson et al., 2014). But without any further processing these assemblies have no deeper

meaning than randomly concatenated stretches of the genomic building blocks. One step in the annotation of genomes resolves this by the systematic identification and assignment of protein-coding regions referred to as genes, its associated regulatory sequences and the locations of non-coding regions. Being aware of the genes and their locations, makes it similarly possible to assign known or rather potential functions by comparison with related model organisms or by prediction approaches based on probabilistic models.

One of the most occurring and best understood structural motifs in proteins is the coiled coil. It is involved in a broad palette of biological processes, that controls protein-protein-interactions for example during the transcription or the remodelling of the membrane and reaches up to the mediation of structure and stability in cells and tissues. Presumably it plays an important role in several transcription factors and proteins, which participate in the transport of vesicles. Furthermore it is assumed, that coiled coils function as well as structural spacers in protein complexes functioning as stabilizing components and are therefore an essential structure to be identified.

1.2. Scope of the thesis

The present thesis contributes to the research field of bioinformatics and its applications in biological sequence analysis. The general aim of this cumulative work lies in the development of algorithms and probabilistic concepts with main focus on the principles of modelling and their practical suitability. This is to be implemented and investigated on the basis of two interdisciplinary projects from the fields of a classic domain of bioinformatics, the area of protein structure and function prediction and the more experimental-oriented area, the heterologous gene expression. At first glance both fields differ relatively strong in the nature of their application, one aims to predict functional or structural regions in inspected biological sequences and the other attempts to figure out characteristic genetic patterns and solutions to control the expression of specific proteins in foreign model organisms. But from an informatics point of view, both research areas provide well suited circumstances for the application and investigation of probabilistic models.

1.2.1. Comparative domain prediction of coiled coils and SAHs

The studies of the first project on protein structure and function prediction deal with the »comparative domain prediction of coiled coils and SAHs«. High quality and reliable prediction of coiled coils in protein sequences with the relevant available software packages remains a challenge. Difficulties lie not only in the prediction of the correct sequence regions, but also in whether regions with coiled-coils are recognized at all and, if so, are not mistaken for SAH domains. The result depends significantly on the software being used, because these often give different, and in part contradictory, prediction results. Here, we try to make a contribution for improvement by the development of a comparative visualization software for coiled-coil predictions with an algorithm for SAH domain detection, that is tested and investigated in two studies and by conducting an in-depth investigation to clarify the »status quo« of coiled-coil prediction on a comprehensive structural reference dataset. In summary, the research questions addressed are as follows:

- **RQ 1.1:** As a »proof of concept«, does the developed SAH prediction algorithm allow a reliable determination of SAH domains?

- **RQ 1.2:** Is the new SAH prediction algorithm in terms of quality and performance suitable for genome-wide analyses?
- **RQ 1.3:** What is the current »status quo« of the established coiled-coil prediction software and what relevance has this for the genome-wide annotation of this type of domain?

1.2.2. Design of typical genes for heterologous gene expression

The second project treats and evaluates the suitability of an information theory based approach for the design of typical gene sequences aiming to improve the requirements in heterologous protein expression. In order to meet that objective, basically a generative probabilistic model, in form of highly configurable Markov chain, has been developed that is capable to capture and simulate the genetic and stochastic characteristics of a sequenced genome and provides an interface to re-design protein-coding genes into host-typical DNA sequences. Originally being entitled »Information theory based design of typical genes for heterologous gene expression and evaluation of experimental studies«, the project covers, besides the conception and development of a software, the experimental evaluation under laboratory conditions. This evaluation has been realized in close collaboration with the Department for Molecular Microbiology. By being a computation-based approach with a stronger focus on the biological application and usability, this process has been an essential part for the success of the project. It enabled to test assumptions and predictions for the iterative improvement and refinement of the software model during the development process as well as the general suitability of the model for the chosen application scenario. The addressed research questions are as follows:

- **RQ 2.1:** Is the developed gene design model suitable for the generation of typical gene sequences and has influence on the experimental expression regulation?
- **RQ 2.2:** As a »proof of concept«, does the host-specific adaptation of genes allow to regulate the expression level of a specific protein in the expression system in an intended way for instance by increasing or decreasing the amounts of produced protein yields?

1.3. Impact

During the course of this cumulative work, intermediate results have been submitted and published in the following peer reviewed journal articles:

- **Dominic Simm**, Klas Hatje and Martin Kollmar. (2015) "Waggawagga: comparative visualization of coiled-coil predictions and detection of stable single α -helices (SAH domains)". *Bioinformatics*, Volume 31 (5), 767-769. [Online]. Available: <http://bioinformatics.oxfordjournals.org/content/31/5/767.long>
- **Dominic Simm**, Klas Hatje and Martin Kollmar. (2017) "Distribution and evolution of stable single α -helices (SAH domains) in myosin motor proteins". *PLoS ONE*, Volume 12 (4), e0174639, 1-16. [Online]. Available: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0174639>
- **Dominic Simm** and Martin Kollmar. (2018) "Waggawagga-CLI: A command-line tool for predicting stable single α -helices (SAH-domains), and the SAH-domain distribution across eukaryotes". *PLoS ONE*, Volume 13 (2), e0191924, 1-19. [Online]. Available: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0191924>

- **Dominic Simm**, Klas Hatje, Stephan Waack and Martin Kollmar. (2021) "Critical assessment of coiled-coil predictions based on protein structure data". *Scientific Reports*, Volume 11 (12439), 1-18. [Online]. Available: <https://www.nature.com/articles/s41598-021-91886-w>
- **Dominic Simm**, Blagovesta Popova, Gerhard H. Braus, Stephan Waack and Martin Kollmar. (2021) "Design of typical genes for heterologous gene expression". Submitted.

Furthermore, the author of this thesis has also contributed to the following papers:

- Martin Kollmar, Lotte Kollmar, Björn Hammesfahr and Dominic Simm. (2015) "diArk – the database for eukaryotic genome and transcriptome assemblies in 2014". *Nucleic Acids Research*, Volume 43 D1, D1107–D1112. [Online]. Available: <https://academic.oup.com/nar/article/43/D1/D1107/2439940>
- Klas Hatje, Raza-Ur Rahman, Ramon O Vidal, Dominic Simm, Björn Hammesfahr, Vikas Bansal, Ashish Rajput, Michel Edwar Mickael, Ting Sun, Stefan Bonn and Martin Kollmar. (2017) "The landscape of human mutually exclusive splicing". *Molecular Systems Biology*, Volume 13 (12), 959. [Online]. Available: <https://www.embopress.org/doi/full/10.15252/msb.20177728>
- Martin Kollmar and Dominic Simm. (2018) "Identifying Sequenced Eukaryotic Genomes and Transcriptomes with diArk". *Eukaryotic Genomic Databases, Methods in Molecular Biology*, Volume 1757, Chapter 1. [Online]. Available: https://link.springer.com/protocol/10.1007/978-1-4939-7737-6_1
- Klas Hatje, Stefanie Mühlhausen, Dominic Simm and Martin Kollmar. (2019). "The Protein-Coding Human Genome: Annotating High-Hanging Fruits". *BioEssays*, Volume 10 (19), 959. [Online]. Available: <https://doi.org/10.1002/bies.201900066>

Finally, the following student projects and teaching courses were supervised and held during the work on this thesis.

Student projects

- Fabian Meyer, "An evaluation of tRNA predictions over the eukaryotic genomes database diArk". Student project, Institute of Computer Science, University of Goettingen. 2019.
- Ivana Gavrilova, "Literature research on new genome sequencing projects". Student project, Institute of Computer Science, University of Goettingen. 2020.
- Gregor Sommer, "Development of Django import modules for satellite resources and web frontend features for the Genometation project database" Student project, Institute of Computer Science, University of Goettingen. 2020/21.
- Lotte Kollmar, "Database content development and curation of genomic data in the Genometation project database". Student project, Institute of Computer Science, University of Goettingen. 2021.
- Paul Gräve, "Development of web frontend features for the representation of tRNA gene predictions in the Genometation project database". Student project, Institute of Computer Science, University of Goettingen. 2021.

Teaching classes

- Supervised as PhD student at the Max Planck Institute for Biophysical Chemistry:
 - Summer term 2014:
Methods course »Protein family analysis as basis for experiments and experimental data interpretation« (PD Dr. Kollmar).
 - Winter term 2014/15:
Methods courses »Protein family analysis as basis for experiments and experimental data interpretation« and »Molecular Biology« (PD Dr. Kollmar).
- Supervised as PhD student at the University of Goettingen:
 - Summer term 2015:
Exercises for lecture »Grundlagen der Informationstheorie« (Prof. Waack).
 - Summer term 2016:
Exam for lecture »Datenbanken« (Prof. May).
 - Summer term 2017:
Exercises for lecture »Theoretische Informatik« (Prof. Damm) and exam for lecture »Datenbanken« (Prof. May).
 - Winter term 2017/18:
Exercises for lecture »Betriebssysteme« (PD Dr. Brosenne) .
 - Summer term 2018:
Exercises for lecture »Theoretische Informatik« (Prof. Damm), exam for lecture »Datenbanken« (Prof. Damm) and practical course »Schwachstellenanalyse und Risikomanagement« (PD Dr. Wiese).
 - Winter term 2018/19:
Exercises for lecture »Betriebssysteme« and practical course »Datenbankanwendungsentwicklung« (PD Dr. Wiese).
 - Summer term 2019:
Exercises for lecture »Theoretische Informatik« (Prof. Damm) and exam for lecture »Datenbanken« (Prof. May).

1.4. Structure of the thesis

This dissertation is structured as follows. At first, the foundations and the theoretical background are addressed. For this purpose, an overview of the related work of the two research fields treated in this PhD project is given. In addition, the terminology and theoretical models that play an essential role in this thesis are introduced. This includes basic terms from molecular, structural and synthetic biology, probabilistic models from theoretical computer science, and computational methods from bioinformatics. This is followed by Chapter 3, which presents the results of this cumulative work. It consists of the four studies published on the »Comparative protein structure prediction of coiled-coil and SAH domains« during this PhD project and the manuscript on »Heterologous Gene Expression« recently submitted for publication. Furthermore, it includes four additional publications to which the author of this work has contributed. Finally, the work is summarized in a brief conclusion.

2 Foundations

This chapter gives an introduction into the scientific background, related work and the theoretical foundations, that are assumed to be known in order to understand the studies of the main projects treated in this PhD thesis. The central topics revolve around the two research fields of »structure prediction« and »heterologous expression« of proteins. Both serve to gain biological insight and to provide important contributions from different angles to elucidate the cellular functions and roles of proteins. Therefore the foundations are divided into three sections, starting with the biological part that recapitulates briefly essential information about these molecules. It ranges from the basics of their genetic encoding and translation, over the composition, folding and structure, including the specific structural domains of »Coiled Coils« and »stable single-alpha helices (SAHs)« up to sequence modification and heterologous expression in technical model organisms. The following section on bioinformatic methods treats probabilistic models with focus on Markov chains and approaches for sequence analysis and structure prediction supporting the scientific progress in the mentioned research fields. Finally an overview is given about the biological databases and technological methods, that come into use in the developed applications of the performed PhD projects.

2.1. Background

Proteins are essential to life. Representing the molecular machines in cells, these complex molecules enable practically all of the living functions of the known prokaryotic and eukaryotic organisms. Due to the numerous different functions and properties, their areas of application are highly diverse. Proteins can be both cause and solution in the treatment of pathological diseases such as Parkinson's or Alzheimer's (Meade et al., 2019), act as genetic markers in expression studies (Zimmer, 2002) or be used as DNA-editing enzymes in the modification of genomic information as for instance in the revolutionary CRISPR/Cas method (Jinek et al., 2012). At the same time, they have as well a high importance for other types of life science applications as catalytic enzymes in the production and processing of food or in the degradation of rubbish and waste water. One of the major challenges in biology is therefore to gain insight into the molecular structure, its mechanics and supported cellular processes to achieve a deeper understanding of their functions and roles. The function of a protein depends largely on the unique three-dimensional structure and the adopted shape is in turn closely linked with its coding sequence. This multi-faceted relationship between »sequence, structure and function« makes these molecules the subject of a wide variety of research fields, all of which try to contribute to a better structural and functional understanding from different perspectives.

The sequencing of whole genomes, including the capture of protein-coding sequences, is no longer a major challenge in the post-genomic era. Due to the rapid development of sequencing technologies in the last decades, there exist nowadays broad-equipped sequencing centres all over the world, that generate masses of sequenced genomic data at high-throughput and massively parallelized. In

the meantime, more than 20,000 officially published and sequenced organisms exist in the relevant public genome databases. This enormous development is also reflected in the significantly increasing amount of available protein sequence data, such as available in the UniProt/TrEMBL database (UniProt Consortium, 2018) with currently over 200 million entries, see Figure 2.1 for more detailed information. However, the field of experimental structure determination of sequenced proteins is unable to keep pace with progress on this scale. As a result, the wide field of computational biology has emerged and developed into an equally important field of research with its numerous subfields to address this discrepancy for instance through computational structure prediction.

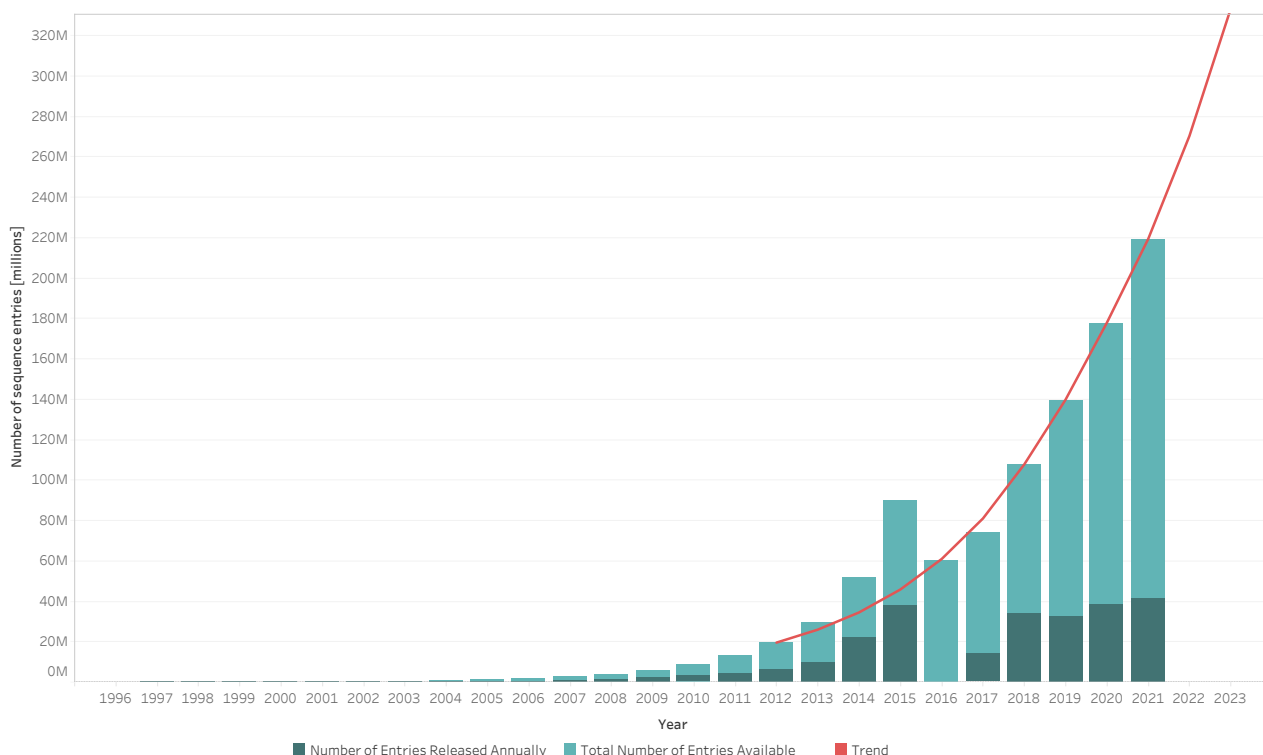


Figure 2.1. | Growth of the protein sequence knowledge base. The diagram illustrates the development of known protein sequences stored in the UniProtKB/TrEMBL protein database of the European Molecular Biology Laboratory (EMBL) from 1997 to 2021, one of the defacto standard resources for protein sequence and functional information. Shown here are the new released and the total number of sequence entries available on an annually basis. The numbers were obtained from the UniProt release statistics information pages (<https://www.ebi.ac.uk/uniprot/TrEMBLstats>).

In the context of this PhD thesis, the focus was placed on computational applications in the areas of »structure prediction« and »heterologous expression« of proteins, with both coming from different areas of the emerged field of computational biology. These two very distinct fields of research deal, on the one side, with the predictability of structure and function solely on the basis of the underlying protein sequence information, whereby the quality and reliability of prediction is of great interest. Especially since these applications are not only used for individual proteins, but also on a much larger scale, for instance for the structural annotation of entire genomes. On the other side, the feasibility of expressing computationally modified genes in different model organisms is being investigated. For this purpose, the focus lies on analyzing the influence of synonymously varied protein-coding

sequences on the regulation of heterologous protein expression. In both projects, besides the informatics methods and models used, the influence of additional domain-specific biological knowledge for structure prediction as well as the natural respectively *typical* modification of gene sequences are at the centre of the investigations.

2.1.1. Protein structure and domain prediction

The research field of »protein structure prediction« gained lately a high amount of public attention. Since a major milestone in this field has been reached recently by the publication of the breakthrough structure prediction algorithm Alphafold (Jumper et al., 2021) and its associated protein structure database (Tunyasuvunakool et al., 2021), that contains in its first release newly generated structural predictions for the human proteome and a selection of key model organisms. The neural-network based approach nearly solved the for more than 50 years open research problem to make accurate computational predictions for the three-dimensional folding of single protein structures with almost experimental quality solely based on its sequential information. A research problem, that was long believed to have no predictable solution due to the sheer endless number of folding possibilities (*Levinthal's paradox*, Levinthal, 1969). One of the crucial key factors to the success of this approach lies in the combination of large amounts of biological data information with the novel techniques of deep learning to extract essential information, interpret correlations and apply the knowledge gained for the prediction. For instance with regard to evolutionary relatedness on the basis of sequence-structure families between known and new molecular structures. In the process, the approach has clearly benefited from the numerous experimentally determined protein structures deposited in the Protein Data Bank (wwPDB consortium, 2019) and the wide range of available sequenced genome data sets generated by the rapidly advanced sequencing methods.

The relatively new and promising approach of deep learning has already been applied to one of the subfields of structure prediction, the »coiled-coil domain prediction« by the authors of DeepCoil (Ludwiczak et al., 2019). It is one of the latest developments in this field, but in contrast to the Alphafold approach it has been extensively trained mainly with structural information from the Protein Data Bank (PDB). However, since this type of structure prediction is not only about domains within a single protein, but to a large extent also about coiled-coil domains forming in complexes of several interacting proteins, such as dimers, trimers and tetramers, the chosen approach relying solely on the training of single structures unfortunately did not bring a comparable resounding success.

The prediction of coiled-coil domains using computational approaches has a long history that began in the early 1990s with the release of the COILS software (Lupas et al., 1991). Since then numerous coiled-coil prediction softwares were developed and published, a progress that continues to the present day. Each of the newly developed software has attempted to improve the quality of the coiled coil predictions either by using more advanced probabilistic models and algorithms or by incorporating and exploiting increasing amounts of characteristic biological information. The derived knowledge and rules build then the basis for the enhanced prediction approach. Unfortunately, each of these developed softwares has been shown to have a rather limited applicability especially in terms of the quality of predictions with regard to large-scale coiled-coil analyses, such as for genome-level annotations. This is either due to technical limitations in the chosen underlying model or deficiencies in the choice of reference data for training and fitting the model parameters. For this reason, the field of coiled-coil domain detection providing accurate and reliable predictions remains a challenge.

One of the both PhD projects treated in this thesis revolves around the topic of »coiled-coil domain prediction« in proteins and addresses the mentioned challenge. The four studies carried out, deal with the analysis and evaluation of the status quo of existing software packages as well as the development of approaches to improve the reliable coiled-coil domain detection.

2.1.2. Heterologous gene expression

Equally essential for a deeper understanding of the cellular functions of proteins are the underlying mechanisms from the encoding and transcription of genetic information through biosynthetic production by the ribosome to the correctly folded and functional protein.

An important method for analyzing these mechanisms and cell functions is »heterologous protein expression«. This experimental approach attempts to produce, respectively express, a specific protein in a host organism that misses this gene in its own genome. In the process, the gene of interest is first permanently inserted into the genome of the foreign host using recombinant DNA technologies. The host organism, also termed as expression system, thus serves to produce the encoded protein in a directed and controlled form of protein biosynthesis. Many of the common model organisms can be used as expression systems, frequently used are *E. coli* and *S. cerevisiae*. But genetic differences between prokaryotes and eukaryotes, such as *E. coli*, lead to difficulties in the expression of cloned genes e.g. originating from the human, which results in a need for strategies to cope with them (Goodman, 2007). The systematic modification of the gene sequences to be integrated has been shown to have a strong influence on the regulation of protein expression and is an important step for homologous and heterologous protein expression studies (Brule and Grayhack, 2017; Gustafsson et al., 2012; Hanson and Collier, 2018; Hershberg and Petrov, 2009; Nieuwkoop et al., 2020). Further studies, that this method enables are to investigate functions of homologous proteins (Hia et al., 2019), to synthetically construct genetic circuits (Hansen et al., 2014; Kato, 2019; Michalodimitrakis and Isalan, 2009), or to overexpress proteins for biopharmaceutical applications (Mauro, 2018) and structural biology research (Hedfalk, 2012).

At the DNA level, a protein-coding gene not only determines the specific amino acid sequence of the resulting protein and thus its three-dimensional shape and function, but also decisively influences the speed of protein biosynthesis and the produced amount of complete and functional proteins, solely through the selection of different types of synonymous codons. An effect that is caused by the »redundancy and degeneracy of the genetic code« enabling a single protein to be encoded by a multitude of equivalent gene sequences. The specific distribution of amino acid-coding codons over the entire genome is called the »codon usage«, it is a genomic key characteristic of organisms that plays an important role in the systematic generation and modification of gene sequences.

Each used codon usually has a specific tRNA counterpart that is responsible for the incorporation of the encoded amino acid. However, these tRNAs are present in the cell in different concentrations, some occur very frequently, others in very low concentrations, which means that some codons can be translated more efficiently than others. The different influx of tRNAs available for production slows down or speeds up the process of biosynthesis of a protein on the ribosome. A phenomenon known as ribosomal pausing, that has direct impact on the protein folding process and can therefore affect the final shape. The tRNA concentrations vary between different organisms as well as between cell types and developmental stages of cells and depend on several factors. One major factor is the copy number of the tRNA-coding genes located in the genome, these influence the amount of available

tRNAs on the production side. A higher number of gene duplicates usually automatically results in a higher number of available tRNAs. These two factors of codon-usage and tRNA abundance are in a way complementary to each other. If the organism-specific codon usage is adapted, for example through systematic modification of the gene sequence to express, other tRNAs are needed to produce the protein. If these are in turn present in different concentrations, this may lead to an increase or decrease in the expression rate. A property that can be used for the controlled regulation of protein biosynthesis. Thus, the genetic code used plays an essential role in heterologous protein expression.

The second main project on the "*Design of typical genes for heterologous gene expression*" (see Section 3.2) of this work examines the phenomenon of codon usage modification in heterologous expression applications as described above. In this context we analyzed the influence of a profiled codon usage adaptation approach, realized through a highly configurable Markov model, on the protein expression levels in the eukaryotic model organism *Saccharomyces cerevisiae*.

For experimental verification of the developed gene adaptation algorithm, a set of test proteins was required to demonstrate that intended protein expression rates can be reached under real conditions in an expression system. Requirements for this set contained the selection of proteins with significant different structural characteristics. Thus representatives of the two main structure classes of stable and intrinsically disordered proteins were chosen, due to the fact that the naturally unfolded human α -synuclein (α Syn) was already predetermined by the research interest of our collaborating group "Department for Molecular Microbiology". For the second class with contrary structural properties the compact and stable folding green fluorescent protein (GFP) was selected, which is proven to be reliable expressible in common model organisms. In the course of the project a public available web-application has been developed for performing host-specific protein adaptations to a set of the most commonly used model organisms.

2.2. Biological foundations

This section recapitulates briefly the main biological basics building the fundamentals of the two main projects on »coiled-coil domain prediction« and »heterologous expression« of proteins ranging from their genetic encoding and translation over the composition, folding and structure, including the specific structural domains of »Coiled Coils« and »SAHs«.

2.2.1. The genetic code

The genetic code is an essential part of the encoding of genetic information in DNA and represents the translation table for nucleotide sequences into amino acid sequences. It is also referred to as the central dogma of molecular biology. The code translates three nucleotides, so-called triplets or codons, into one amino acid. With the four different nucleotides, there are 64 possible triplet combinations, but only 20 naturally occurring amino acids. Due to the much higher number of triplets than amino acids, some amino acids are encoded by several synonymous codons. A maximum of up to six triplets code for the same amino acid. The genetic code is thus highly redundant and in this context we speak of the degeneracy of the genetic code (Figure 2.2). The details were already worked out 50 years ago and therefore reference is made to textbooks for further reading.

| | | second position | | | | |
|-------------------------|---|---|------------------------------|--|--|------------------|
| | | U | C | A | G | |
| first position (5' end) | U | UUU Phe UUC UUA Leu UUG | UCU Ser UCC UCA UCG | UAU Tyr UAC UAA* stop UAG* stop | UGU Cys UGC UGA* stop UGG Trp | U C A G |
| | C | CUU Leu CUC CUA CUG | CCU Pro CCC CAA CCG | CAU His CAC CAA Gln CAG | CGU Arg CGC CGA CGG | U C A G |
| | A | AUU Ile AUC AUA AUG ⁺ Met | ACU Thr ACC ACA ACG | AAU Asn AAC AAA Lys AAG | AGU Ser AGC AGA Arg AGG | U C A G |
| | G | GUU Val GUC GUA GUG | GCU Ala GCC GCA GCG | GAU Asp GAC GAA Glu GAG | GGU Gly GGC GGA GGG | U C A G |

Figure 2.2. | Genetic code. ⁺ Start codon, specifies the translation initiator formyl-Met-tRNA^{fMet}; * Chain-terminating or »nonsense« codons

2.2.2. Proteins

Proteins are essential to life. The complex macromolecules are responsible for the implementation of practical all biological functions in living organisms. Like DNA, a protein is a long-chain biomolecule but composed of 20 different building blocks. These monomers are amino acids that are linked and strung together by peptide bonds. They are therefore also referred to as amino acid or polypeptide chains. These polypeptide chains fold into a spatial structure that is determined by the sequence of amino acids. At the same time, the folding determines the biological function of the protein. Factors that influence folding are intramolecular interactions, for example in the form of hydrogen bonds between peptide backbone elements or attractions between the side chains of amino acids.

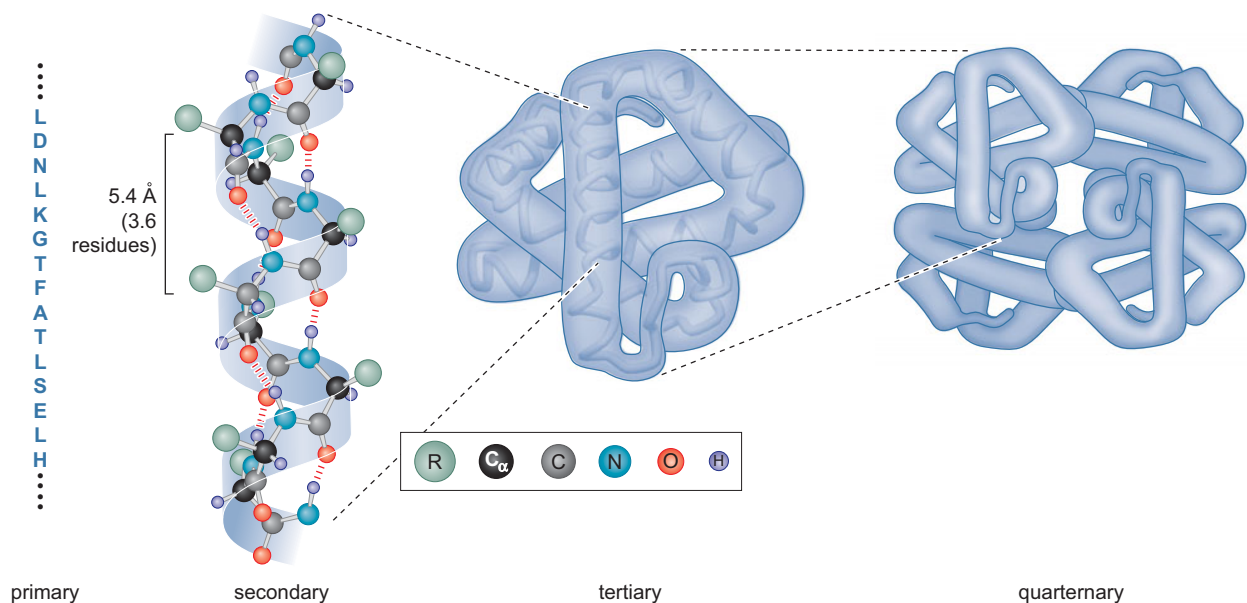


Figure 2.3. | The four structural levels of protein folding. Depicted are from left to right exemplary views of the primary, secondary, tertiary and quaternary structure as single structural states. (cf. Watson et al., 2014, p. 127)

The spatial structure of proteins can be divided into four observation levels: The primary structure refers to the pure sequential order of individual amino acids in the polypeptide chain. The secondary structure, the first form of the actual spatial arrangement, describes frequently occurring motifs, such as α -helices, β -sheets and other forms such as bends, loops and turns. They derive directly from the primary structure and are held together by interactions between amino acid residues such as disulfide bridges, covalent and non-covalent bonds, or hydrophobic, ionic, and van der Waals forces. The α -helices are right-handed twisted spirals with an average of 3.6 residues per helical turn. This corresponds to a pitch of 54 nm (5.4 Å) for one turn. This spacing ensures that amino acids that are three to four places apart in the primary structure are adjacent in the helix. A property of high relevance for the formation of coiled coils, as described in the following section 2.2.3. The tertiary structure describes the next higher structure stabilizing arrangement level, in which the secondary structure motifs arrange spatially with respect to each other by their position in the sequence and further interactions (Fig. 2.3). The last level, the quaternary structure of proteins, refers more to an assembly of already prefolded tertiary subunits of (independent) polypeptides, that driven by interaction forces form into a functional protein complex.

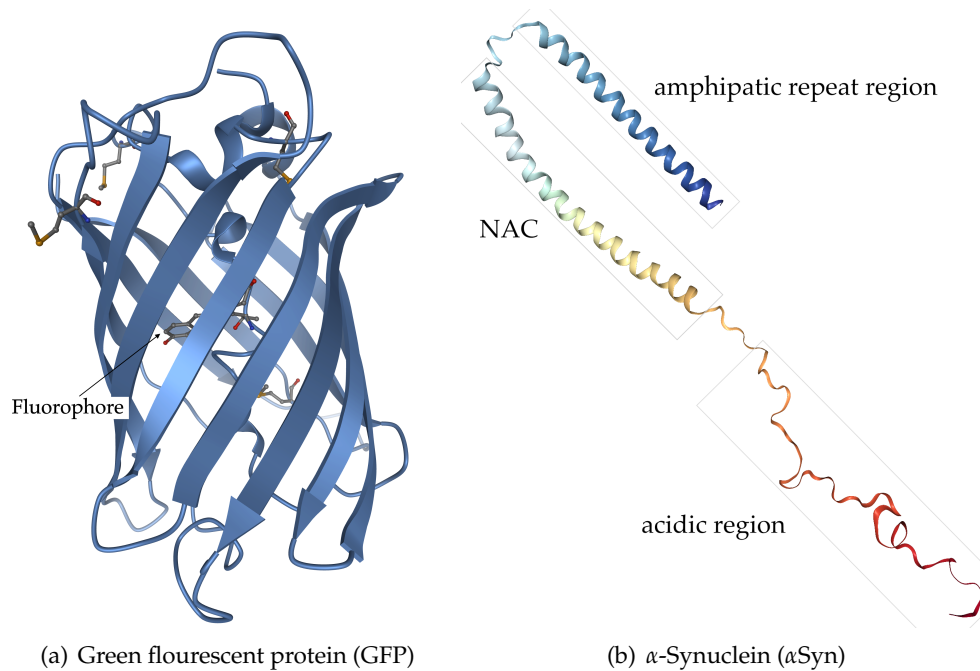


Figure 2.4. | **Representatives of stable and intrinsically disordered proteins.** Depicted are views of the 3-dimensional structure of two proteins from both classes, namely the stable GFP (a) and the IDP α -Synuclein (b). Images were created using Mol* (Sehnal et al., 2021) and NGL (Rose et al., 2018) and the structures of 1EMA and 1XQ8 retrieved from RCSB PDB (wwPDB consortium, 2019).

But the individual characteristics of the structural levels provide only folding tendencies and do not necessarily determine a fixed three-dimensional structure of the protein. Differences exist in the form and stability of the folding proteins, which can be summarized into two classes. On the one hand, there is the class of proteins with a native and predefined folding process, that have a stable structure with a certain shape at the end of folding, such as the GFP. On the other hand, there is the class of intrinsically disordered proteins (IDPs), that are missing a fixed and ordered three-dimensional structure, either only in parts or concerning the whole protein. The reasons for this lie mostly in the absence of interacting molecules, such as other proteins or RNA, that give external stability and influence the adopted conformation of the structure. These fully or partially unstructured molecules contain often randomly evolving domain structures such as coils as well as flexible loops and linkers.

Green fluorescent protein

A representative of the reliable and stable folding proteins is the green fluorescent protein (GFP), that originally stems from the Pacific Northwest jellyfish *Aequorea victoria*. It belongs with its luminous property to the class of fluorescent proteins and the ability that its cDNA is suitable for heterologous expression made it to one of the standard tools in cell and molecular biology applications with a widespread use as reporter for gene expression and localization of gene products in living cells (Chalfie et al., 1994; Tsien, 1998). The cellular location of GFP is revealed by the green fluorescence emission, when the protein is exposed to light from the blue or ultraviolet spectrum. Given the protein is properly expressed in the host organism and is folded in its intact three-dimensional shape. GFP can be expressed in a wide range of prokaryotic and eukaryotic cells and organisms,

among others into bacteria (*Escherichia coli*), yeasts (*Saccharomyces cerevisiae*) and plants (*Arabidopsis thaliana*). Therefore, its genetic information has to be introduced into the genome of the foreign organism through the use of transgenic techniques, where it remains and serves as blueprint for the protein expression.

The natural wild-type molecule has a length of 238 amino acids, a molecular mass of approximate 30 kDa and consists of a proteolysis-resistant single chain. The three-dimensional structure is characterized by a compact and stable 11-stranded β -barrel with a coaxial helix in its interior. In the middle of the molecule, it encloses the actual fluorescence emitting chromophore (or fluorophore), that is attached to the through-barrel winding α -helix. An advantage of the luminous unit is, it functions autonomously without additional external cofactors (Ormö et al., 1996; Zimmer, 2002), see for the GFP structure Figure 2.4(a). In its typical function as gene expression marker for genetic tagging, the GFP gene needs be fused to the gene of interest to monitor its expression under *in vivo* conditions. Having free outstanding ends at the amino and carboxyl terminals, the GFP can be used to build gene fusion complexes with the flexibility that GFP can be located either *in front* or *behind* the protein of interest as required. Therefore the coding DNA sequence of GFP needs to be cloned into an expression vector construct, a vehicle that facilitates the incorporation into the target genome. As a consequence of the gene fusion, the size and mass of the resulting protein complex increases and the natural behavior and function of the actual target protein may be disturbed (Goodman, 2007).

Alpha-synuclein

The class of "intrinsically disordered proteins" is represented by the neuronal protein α -synuclein (α Syn). The wild-type molecule has a length of 140 amino acids with a molecular mass of approximate 14 kDa and consists of seven 11-residue repeats and a hydrophilic tail dividing the single chain in three subregions. Being a natively unstructured protein in solution, in interaction with phospholipid membranes it shows a loose α -helical conformation in the head part of the protein. This conformation, also predicted by sequence analysis, is composed of two interrupted α -helical regions with a short loop in between (Weinreb et al., 1996), see Figure 2.4(b) for a structural view of α Syn. The origin of α -synuclein lies in the genome of the *Homo sapiens* and is encoded by the *SNCA* gene. It is also referred to as the precursor protein *NACP*, including the *NAC* fragment (non- $A\beta$ component) that is known to be related with the Alzheimer disease (AD) amyloid protein. But due to the unknown native structure, its function is also not yet completely understood. It is known that α Syn has a high abundance in the human brain and occurs to a quite lower extend in other tissues like the heart or in muscles. More precisely, α Syn was identified in detailed expression studies to be abundant in the synaptic terminals of neurons. This and the relationship with the AD amyloid protein led to the conclusion that it contributes to the pathogenesis of Parkinson's and other neurodegenerative diseases (Chandra et al., 2003; Meade et al., 2019; Xia et al., 2001).

2.2.3. Coiled coils

α -Helical coiled coils (CCs) are a special form of protein structural domains, that belong to the tertiary respectively quaternary level of protein structure, characterized by a strong regularity and stability. The flexible composition of this domain type allows the formation of a variety of different assembled structures. Due to their high structural diversity, coiled coils support a wide range of essential biological functions and interactions, such as forming mechanically rigid structures, the transduction of conformational changes or the transport of molecules (Lupas et al., 2017). These properties make them a research object of great interest, because the detailed understanding and reliable annotation of this fold type is crucial for studies of protein structure and function (Newman et al., 2000), the design and modelling of new structures (Wood and Woolfson, 2018) and systems for drug delivery (McFarlane et al., 2009).

Canonical coiled coils are α -helical oligomers consisting of two or more helices with parallel or antiparallel orientation, that stem from the same (*homo*) or from different polypeptide chains (*hetero*). The constituent α -helices twist around each other and form left-handed supercoiled bundles, that are stabilized through specific packing interactions, also known as "knobs-into-holes (KIH)" packing, in detail first defined by Crick (Crick, 1953). Each of the participating helices is commonly referred to as coiled-coil domain (CCD). The stability of these bundles results from the regular meshing of the side chains, in which a residue from one helix (*knob*) packs into a space formed by the side-chains of the facing helix (*hole*). The necessary regularity for this packing requires a recurrent structural motif of seven residues along the helix interface. This motif is called a »heptad repeat« and its positions are typically labelled with letters from *a* to *g* resulting in a sequence pattern of the form $(abcdefg)_n$. The core-oriented positions (*a* and *d*) of this motif are characterized by hydrophobic (*H*) residues, like valine, leucine and isoleucine, that lie at the interface between the supercoiling α -helices. The remaining positions (*b*, *c*, *e*, *f* and *g*) are occupied by charged and polar (*P*) amino acids at the outside of the molecule (see Figure 2.5D), leading to the occupancy order *HPPHPPP* of the heptad repeat (Lupas et al., 2017).

As a consequence of the above described heptad motif, the positions are arranged over two turns of a helix, that have on average a distance of 3.5 residues per turn. A single α -helix has a periodicity of 3.63, so supercoiling helices need to bend around a central axis to reduce the periodicity difference accordingly to be able to build the favored super-helical coiled-coil form. This bending of the participating helices results in a left-handed supercoil with a hydrophobic seam of unpolar residues (*H*), that winds along the surface of the actually right-handed single α -helices (Woolfson, 2017).

By their very nature, single α -helices are not really stable, but gain additional stability by winding around each other and building left-handed coiled coils, in which the hydrophobic seam is enclosed in the center of the molecule, an association additionally driven by the exclusion of water from the hydrophobic core (Delorenzi and Speed, 2002). This is clearly visible in Figure 2.5A and 2.5B by regarding the red and green highlighted areas for position *a* and *d* in the course of the depicted example of a classical, dimeric coiled-coil motif.

As mentioned at the beginning, coiled-coil assemblies can be composed of two or more α -helices of parallel or anti-parallel orientation, that may be formed from the same (*homo*) or different (*hetero*) helical sequences. These structural types are referred to as the oligomerization state (OS) of the coiled coils, which can take several forms, as shown in Figure 2.6. The most common and structural sim-

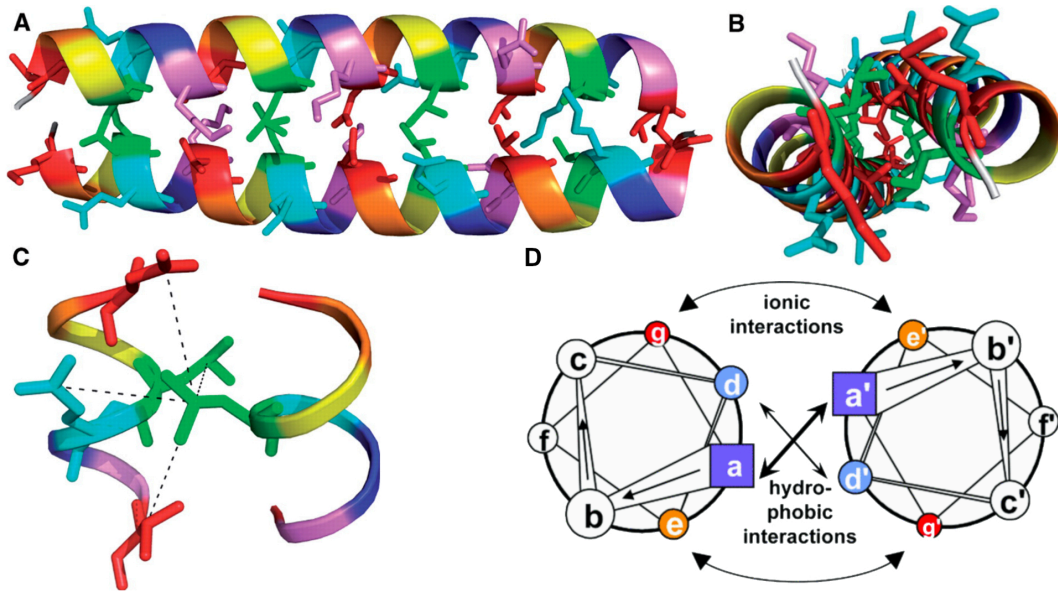


Figure 2.5. | Representations of a parallel dimeric coiled coil. (A) The coiled coil orthogonally seen from the side and (B) down the axis. Helical backbones are represented by ribbons, side chains that make up KIH interactions as sticks and the single heptad positions "a–g" coloured from red to violet, that highlights the hydrophobic core in "red" and "green". (C) A single KIH interaction between a "hole" formed by four residues from one helix (left) and a "knob" from the other (cf. Testa et al., 2009, p. D316 under CC BY 2.0). (D) Schematic representation in form of a "helical wheel" diagram. The graph is looking down the helical axis from the N- to the C-terminus. Residues at the "a" and "d" positions create a hydrophobic core while "e" and "g" residues favour dimerization through formation of stabilizing ionic interactions e.g. salt bridges (cf. Mahrenholz et al., 2010, p. 1 under CC BY 3.0).

plest states, that are known to occur in nature, are the classical dimer, which can occur in antiparallel (A) and parallel (B) conformations, as well as the trimer (C) and the tetramer (D). By parallel and antiparallel orientation is meant the orientation of the single α -helices involved. In the antiparallel case A, the participating helices run in the opposite direction. These four oligomeric states cover about 90% of the known coiled-coil assemblies (Vincent et al., 2013).

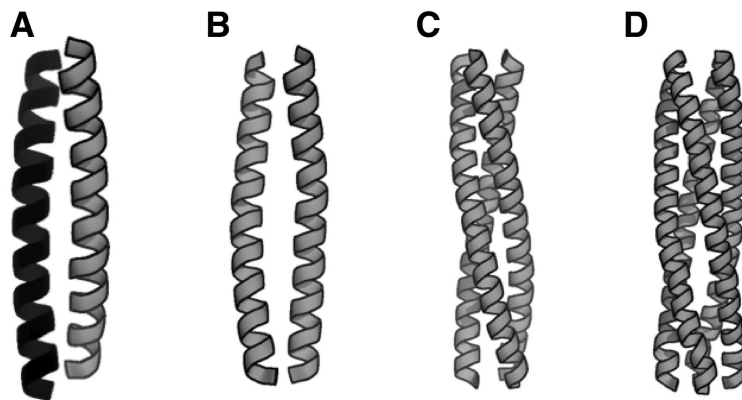


Figure 2.6. | Oligomeric states of coiled coils: antiparallel dimer, parallel dimer, trimer and tetramer. The most common states known to occur in nature are the antiparallel dimer (A), parallel dimer (B), trimer (C) and tetramer (D). Adopted with permission from Vincent et al. (2013, p. 1), Copyright 2021 Oxford University Press.

In addition to the four classical oligomerization states mentioned above, further forms with higher structural complexity exist. These complexes are documented and can be accessed in the CC+ database, see Section 2.3.5, and are depicted in detail in Figure 2.7.

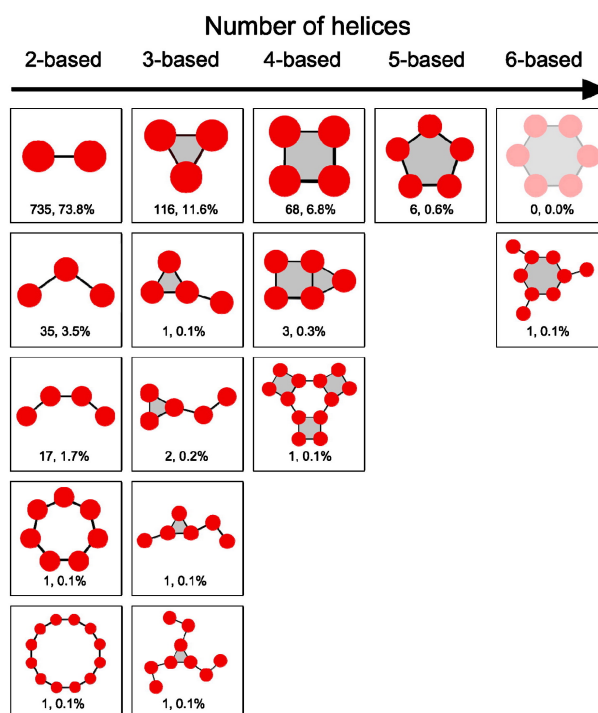


Figure 2.7. | The periodic table of coiled coils. Schematic overview of detected oligomerization states with higher order assemblies. Depicted are along the x -axis the number of helices participating in a coiled-coil and down the y -axis the increase in the order and complexity of the occurring assembly. Adopted and resized with permission from Moutevelis and Woolfson (2009, p. 729), Copyright 2008 Elsevier Ltd.

One of the first protein structures identified to contain coiled coils are the motor proteins of eukaryotic cells, such as in the myosin, kinesin or dynein families. These are classes of proteins that play an essential role in processes of muscle function, neuronal transport and cell division. They are all based on a similar structure, consisting of a head-forming motor domain and a long tail domain. In myosin II families, coiled coils often form in the tail domain between the two long C-terminal chains, where they wrap around each other. Later on, they were chosen to be part of the first reference sets characterizing the key features of coiled coils in the early developed prediction algorithms.

Until today many studies have been undertaken that tried to determine the overall amount of coiled coils at a genome-scale level, both for individual organisms and across related organism families. The intention behind this is a deeper understanding of the functional roles of the coiled-coil motif and an improvement of the global annotation state. For this the term "*coilome*" was chosen, it describes the complete subset of an organisms proteome predicted to contain one or more coiled-coil regions (Barbara et al., 2007). The resulting estimates for the *coilomes* on a single genome basis or whole genome families vary relatively strongly. They range from 10% in eukaryotes to 4-5% coiled coils in prokaryotes and Archaea based on predictions made with the program COILS (Liu and Rost, 2001). A newer study that evaluated taxonomic superfamilies (SUPERFAM, including 1227 completely sequenced

genomes) found an average coiled-coil proportion of 2.9% across all genomes (2.5% eukaryotes, 3.1% bacteria, 1.9% Archaea and 4.3% human), it used SpiriCoil for the predictions (Rackham et al., 2010). Besides the choice of the analysed single organism or superfamily, the used prediction method has a strong impact on the obtained estimates. The studies and researches that investigate this topic in detail are subject of this PhD work.

2.2.4. Stable single-alpha helices

Stable single-alpha helices (SAHs) are versatile structural elements acting in many prokaryotic and eukaryotic proteins as semi-flexible linkers and constant force springs. A stable single-alpha helix, as can be seen from Figure 2.3, initially describes one of the fundamental motifs of the secondary structures of proteins. Added to this are the properties that these helices consist of a significantly high proportion of charged amino acids and remain stable in shape in aqueous solution. Protein sequences forming this structure have a characteristic alteration of positive and negative charges, which leads to short- and long-range electrostatic interactions that have a stabilizing effect. The SAHs exhibit low structural complexity and are rich in these repetitive charged sequences. Amino acids that are very common are aspartic acid (*Asp, D*), glutamic acid (*Glu, E*), arginine (*Arg, R*) and lysine (*Lys, K*). Thus, SAHs structurally fall to some extent into the pattern of coiled coils, so that confusion in the prediction of this type of domains is quite likely (Gáspári et al., 2012; Simm et al., 2017).

Helices, which are not buried within certain globular or coiled-coil structures, usually need networks of charge interactions for stabilization in water, so stand alone SAHs are usually rather atypical to occur in proteins. However, it was discovered that predicted coiled-coil domains of some myosins are in fact stable single-alpha helices (Peckham and Knight, 2009). Consequently, these proteins also do not form coiled coils, as was previously assumed. Thus the assumption seems obvious, that SAHs could be often misidentified as coiled-coil domains by the common prediction softwares.

Several indications for the case, when a putative coiled coil is in fact a stable SAH domain, exist. Firstly, by comparing the occupied heptad positions with the position probabilities derived by an analysis of a large coiled coil dataset, such as retrievable from the CC+ database (Testa et al., 2009, see section 2.3.5). Furthermore, additional conclusions about the domain type can be drawn based on the amino acid pattern with focus on the positional interactions in a heptad network. Such helix-stabilizing forces are charged interactions (*salt bridges*) between residues at $(i, i+3)$ and $(i, i+4)$ spacing and hydrogen-bonding interactions between polar/charged residues at $(i, i+3)$ and $(i, i+4)$. Additional stability is obtained through networks of oppositely charged residues in $(i, i+3, i+6)$, $(i, i+3, i+7)$, $(i, i+4, i+7)$, or $(i, i+4, i+8)$ distances (Simm et al., 2017).

Such heptad networks are shown in Figure 2.8. Initially, the amino acid sequence under consideration is mapped on the repeating heptad motif $(abcdefg)_n$, which is aligned in the form of a mesh. The actual network reflects an α -helix sliced along the central axis with its amino acids located on the helix surface. The outer heptad position *f* is shown another time on the right side (*f*) of the mesh (»cutting edge«) to represent left and downward oriented interactions. For better illustration, an axis with degrees from 0° - 360° is plotted at the bottom of the graph. The sequence in the heptad mesh runs in its sequential order from top to bottom and in rows from right to left. If there are a large number of repetitive positively and negatively charged amino acids in the sequence, as shown in Figure 2.8B, it is highly probable that the respective region does not participate in a coiled coil, but in fact folds into a charged and stable SAH. In the opposite case (left, Figure 2.8A), charged amino acids can

also occur and interact with each other, but there are significantly fewer of them and the positions *a* and *d* are mainly occupied by non-polar amino acids. Such an occupancy, in which a hydrophobic seam is formed (shown in gray), is assumed to be a coiled coil.

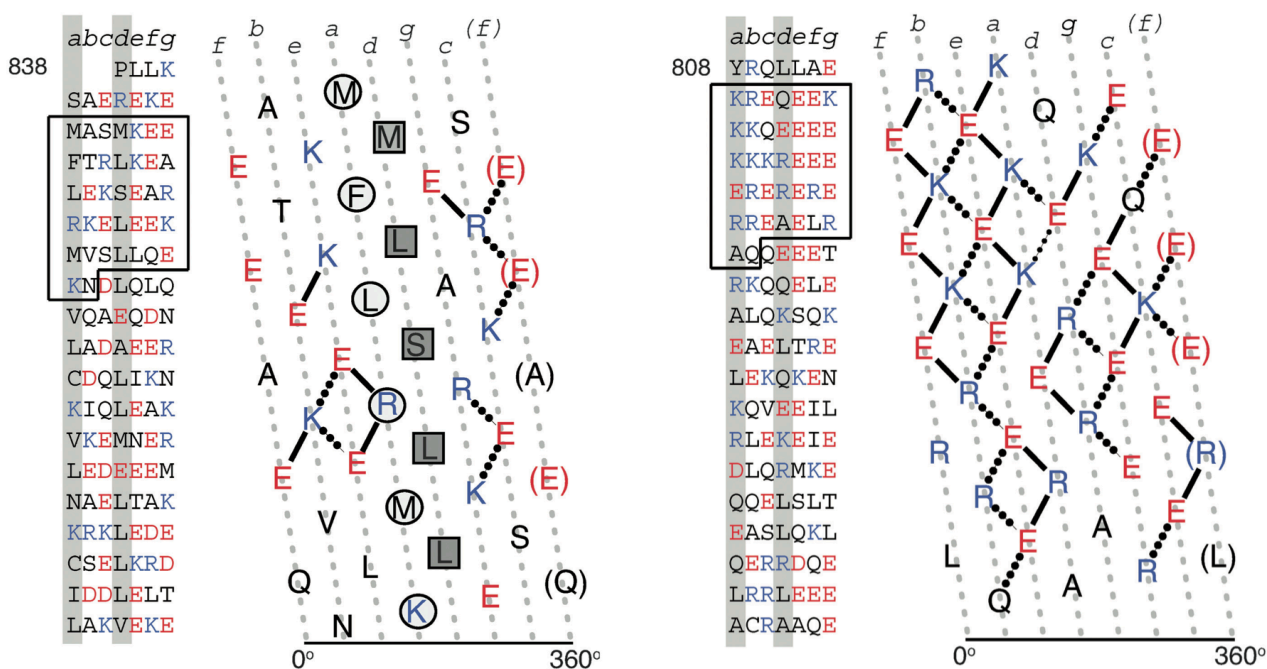


Figure 2.8. | Heptad net representation of a CC and a stable SAH. The figure shows two sequence regions in the heptad net representation (A: Coiled Coil; B: SAH). Charged amino acids are coloured in red/blue. (A) Heptad positions *a* and *d* are highlighted by circles and squares to illustrate the hydrophobic core. The solid connections between residues represent strong intrahelical and the dotted lines weak interactions. Adopted with permission from Peckham and Knight (2009, p. 2495), Copyright 2005 Royal Society of Chemistry.

2.3. Bioinformatical foundations

2.3.1. Probabilistic models

In this section the basic probabilistic models and approaches will be discussed, that have been applied in each of the both PhD projects. Here we focus mainly on the Markov chain model, which builds the core of the developed protein adaption approach to design host-specific and typical genes for the heterologous expression in commonly used model organisms.

Markov chains

The classical model of Markov chains represents a consistent development of conditional probabilities and plays a significant role in the statistical analysis of biological sequences. They are successfully used as models to characterize the influence of restriction enzymes or DNA modification systems on the composition of genomic DNA or to predict the location of genes (Merkl and Waack, 2003).

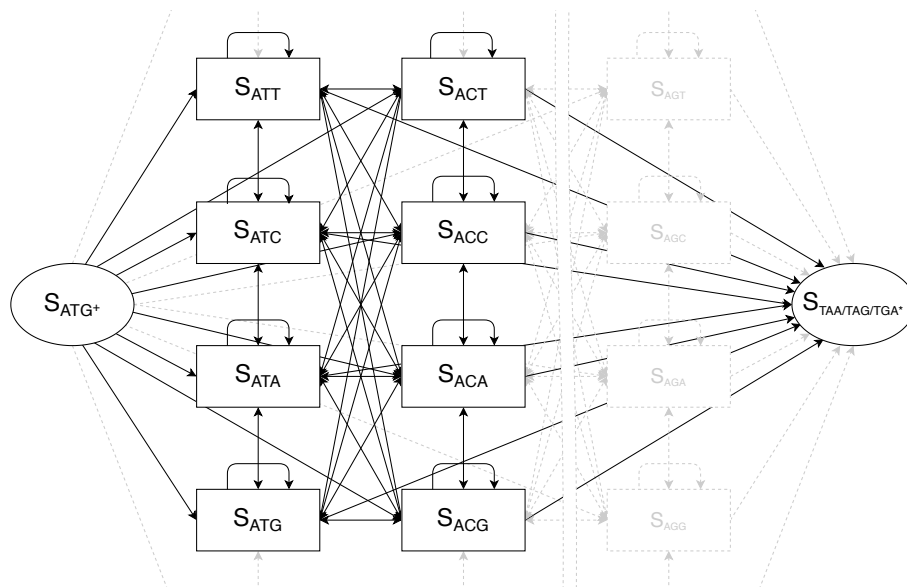


Figure 2.9. | Illustration of the state graph of the Markovian adapter. The figure shows an excerpt of the state graph of the Markovian adapter modelled for the design of typical genes. In the excerpt, only 8 of the actual 64 states are shown for ease of illustration. All of them are fully interconnected by bi-directed transition edges. Besides the 61 codon representing states, part of the graph are the additional states S_{ATG^+} and the three *stop-codon* states S_{TAA^*} , S_{TAG^*} and S_{TGA^*} (depicted as combined end state), that model the start and end states of the gene-modifying Markov chain.

A *first-order Markov chain* is a stochastic process with values from a finite set of *states*, that can be represented as a state graph. Dependent on the modeled sequence type, each state can correspond for instance to a particular nucleotide, amino acid residue or DNA codon, that are interconnected with each other. These connections are directed edges between the state nodes in the graph.

Given the implemented Markov chain of the adaption approach for the design of typical genes, in the following referred to as Markovian adapter (MA), the states of the adapter are the 64 possible and freely combinable DNA codons, which means that each state is connected to all other states. A

so called fully interconnected or complete digraph, illustrated as reduced excerpt in Figure 2.9. Each node in the graph is assigned a *stationary probability* based on the underlying codon distribution. A further probability parameter is associated with each connection in the graph, which determines the probability of a certain codon following another codon. These probability parameters are called transition probabilities and are obtained like the stationary probabilities by trainings from entire genomes of heterologous hosts. Let the *transition probabilities* be denoted by a_{st} :

$$a_{st} = P(x_i = t | x_{i-1} = s).$$

At any time i the Markov chain is in a specific state x_i and may change to a state x_{i+1} depending on the transition probabilities. The probability of a sequence for any probabilistic model of sequences can be written as:

$$P(x) = P(x_L, x_{L-1}, \dots, x_1) \quad (2.1)$$

$$= P(x_L | x_{L-1}, \dots, x_1) P(x_{L-1} | x_{L-2}, \dots, x_1) \cdots P(x_1). \quad (2.2)$$

The key property of a Markov chain (of order 1) is its "memorylessness". It can only remember 1 state transition in its history. This means the probability of each symbol x_i depends only on the value of the preceding symbol x_{i-1} , not on the entire previous sequence, e.g. $P(x_i | x_{i-1}, \dots, x_1) = P(x_i | x_{i-1}) = a_{x_{i-1}x_i}$. Using the preceding equations a simplified form comes out, resulting in the general equation for the probability of a specific sequence from any Markov chain (Durbin et al., 1998):

$$P(x) = P(x_L | x_{L-1}) P(x_{L-1} | x_{L-2}) \cdots P(x_2 | x_1) P(x_1) \quad (2.3)$$

$$= P(x_1) \prod_{i=2}^L a_{x_{i-1}x_i}. \quad (2.4)$$

For a better imagination, the stochastic process can be visualized by the extended state graph of the Markov chain, see Figure 2.10. It is closely related to the simple state graph with the difference, that the extended graph has for each time point t an own layer with all 64 codon states available. Between these layers exist forward directed connections from each state in layer $t - 1$ to all states in layer t making transitions possible. The probability of a specific generated codon sequence is then the product of the start probability $P(x_1)$ and the transition probabilities $a_{x_{i-1}x_i}$ of each hop between the layers of the chosen codon states, as described by equation 2.4.

The training of the parameters provides an elegant way to adapt the model to the characteristics of various expression systems simply by exchanging the model parameters with pre-computed values. As an input parameter it takes the amino-acid sequence of the protein of interest and generates host-adapted DNA proposals conditioned by the original sequence.

For the choice of the transition probabilities, several options are available. Concerning the biological ribosome model with its neighbored "A" and "P" positions from the protein biosynthesis, the di-codon-usage represents a good choice for the transitions between the codon-states. Because it reflects two characteristics at a time, firstly the frequencies of the di-codons-pairs play with their natural abundance an important role for the chosen "2-slot model" also with regard to ribosomal pausing and secondly, the di-codon-usage provides natively through its higher level of detail still the characteristic host-specific mono-codon-usage information.

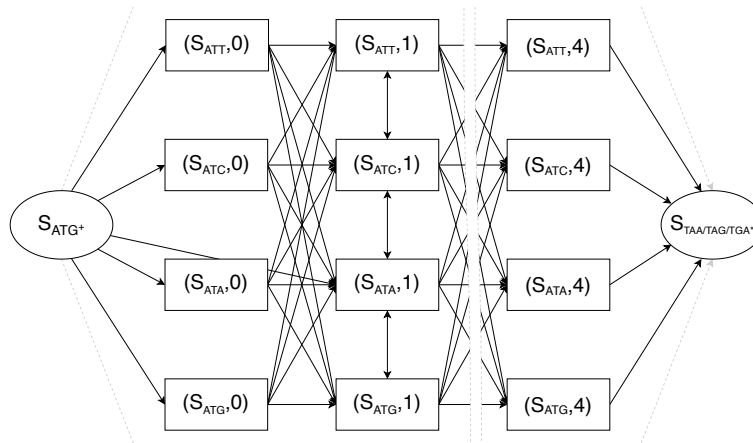


Figure 2.10. | Illustration of the extended state graph of the Markovian adapter. The figure represents an excerpt of the extended state graph of the Markovian adapter with $Z = \{S_{AAA}, S_{AAC}, \dots, S_{TTG}, S_{TTT}\}$ and $t = 5$. Shown are for each time point t 4 of the 61 possible codon states.

2.3.2. Protein structure und function prediction

Proteins are essential to life and were among the first macromolecules to be sequenced and studied in detail (Mount, 2004). A comprehensive knowledge about their structure is closely linked to the understanding of how they work and the uncovering of cellular functions. Different approaches have been developed, that help to clarify the questions about the relationship between their structure and function. The most accurate, but at the same time the one requiring the greatest efforts, are the experimental methods using elaborately prepared probes to resolve the physical structure of proteins into 3-dimensional models. This way to uncover the structure at an atomic level is based on very time- and resource-consuming methods. Among these fall techniques like crystal structure analysis, NMR spectroscopy or electron microscopy to name only the most used, that have been applied for structures provided in the Protein Data Bank (wwPDB consortium, 2019). Thereby should be noted that each of these experimental methods has specific additional limitations, which affect and complicate the structural resolution process. Dependent on the applied method, size-constraints exist for example, so that molecules being larger than 33 kDa (about 300 amino acids) cannot be resolved using the classic NMR-based techniques. All these factors combined ensure, of course depending on the complexity of investigated proteins, that it can take months or even years to resolve the wanted 3-dimensional structures.

At the same time, more and more genomic data is accumulated each year due to the advances in sequencing techniques. These developments led to an explosion in the amount of sequenced genome data. By now, there are billions of known protein sequences available in public databases (UniProt, cf. Figure 2.1), but the proportion for which an actual structure has been resolved is with a magnitude of a around 180,000 molecules comparatively small (cf. size of the PDB from 2021/01 in Figure 2.15). Thus the slow process of the experimental structure coverage represents a crucial bottleneck for the overall progress in this field.

An efficient way to accelerate the structural discovery of proteins is the application of computational approaches to generate structure predictions. The general idea behind these approaches is basically the same. Assuming the clear link between sequence information and structure, the starting point of

these methods is the amino acid sequence as sole 1-dimensional input. With these, the developed algorithms and models attempt to make predictions about the structure as precise as possible. The used methodology can be condensed on two complementary ways that focus either on modelling physical interactions within the protein structures, which are responsible for the final folding or on the evolutionary history of proteins, that allows to draw conclusions from homologous structures, that have been already solved (Jumper et al., 2021). Whereby for both applies that the amount of additional information, processed beforehand and integrated into the underlying models, has a significant influence on the accuracy and quality of the predictions. The major advantages of computation-based approaches lie in the significantly reduced time required to produce these structure predictions and the ability to easily scale to even large amounts of data as is necessary for large genome-wide studies. Limiting factors are mainly the access to sufficient amounts of computing and storage resources, which are nowadays in the era of the prevalent cloud infrastructure solutions and HPC clusters easily accessible to the research community.

The term "protein structure prediction" subsumes several types of applications. Included algorithms and models differ in the scope of the predicted structure, some are limited only to the prediction of specific protein structure domains, such as coiled-coil regions or basic secondary structure types, others try to predict the complete 3-dimensional protein structure at once, which is one of the hardest computational problems due to the manifold possibilities of folding (Levinthal, 1969). Regardless of the type of application, the resulting structure predictions in turn can then be used as a starting point to guide experimental methods, whereby with this prior knowledge the complex experiments can be significantly shortened and the entire process of the structural coverage accelerated. A procedure that has been successfully applied many times. Such as in genome scale studies for example on the evaluation of coiled-coils in the yeast genome (Newman et al., 2000), where the binding of proteins, that have been predicted to be interacting over coiled-coil domains, was experimentally verified.

2.3.3. Coiled-coil prediction

As one of the direct subfields of protein structure prediction, the "coiled-coil prediction" focuses, as its name implies, on the precise and reliable search for protein sequence regions playing a role in the formation of coiled-coil domains. Coiled coils are one of the earliest and best understood protein folding types with the unique characteristics of a high regular nature and unlike other structural types the ability to be computable by parametric equations. The periodic and repetitive pattern of these structures allows their prediction to be so accurate that individual residues can be assigned to the positions of the heptad repeat (see section 2.2.3). For this reason, coiled coils represent an ideal model for the investigation of sequence-structure relationships, which opens the possibility to develop computational methods for the prediction of their structural features solely based on sequence information. The group of coiled-coil prediction methods can be classified in two main categories, namely the "coiled-coil domain detection" and "coiled-coil oligomeric state prediction".

Being originally described in the literature in the beginning of the 1950s, the first computation-based contributions to the research field "prediction of coiled coils" have been published starting in the year 1991 with the release of the software COILS (Lupas et al., 1991). Since then various follow-up studies and improved prediction softwares were performed and published which lasts to the present day. To this point, the most popular programs for the "coiled-coil domain detection" besides COILS are PAIRCOIL (Berger et al., 1995), it improved COILS scoring method with pairwise correlation infor-

mation, MultiCoil (Wolf et al., 1997), which extended PAIRCOIL by adding the first oligomeric state classification and PCOILS (Gruber et al., 2005), a new version of COILS using profiles derived from multiple sequence alignments for comparison. Further developments leading to updated versions have been Paircoil2 (McDonnell et al., 2006) and Multicoil2 (Trigg et al., 2011). All of them make use of similar concepts based on the PSSM approach, but added more informational input to improve the prediction. In addition to mention is MARCOIL (Delorenzi and Speed, 2002), that introduced the first window-less HMM-based approach for short coiled-coil domain detection, as well as DeepCoil (Ludwiczak et al., 2019) and DeepCoil2, both neural network-based approaches trained with structural data from the PDB.

Representatives of the second category for the "oligomeric state prediction" are the programs SCORER and its successor SCORER 2.0 (Armstrong et al., 2011; Woolfson and Alber, 1995), as already mentioned above the Multicoil programs but to a very limited extent, PrOCOIL (Mahrenholz et al., 2010) and LOGICOIL (Vincent et al., 2013).

| Name / Reference | | Principle of coiled-coil prediction | Prediction type | optional settings | window |
|------------------|----------------------------|---|--|---|------------|
| COILS | Lupas et al. (1991) | PSSM (MTK, MTIDK) | CC | un-, weighted (a+d = b,c,e,d,g), window-size, PSSM matrices | 14, 21, 28 |
| PAIRCOIL | Berger et al. (1995) | PSSM (pairwise side chain correlations) | CC | pair-distances (default: 1, 2, 4) | 21, 28 |
| Paircoil2 | McDonnell et al. (2006) | | CC | pair-distances (default: 1, 2, 4) | 21, 28 |
| MultiCoil | Wolf et al. (1997) | | CC, OS (dimer, trimer) | pair-distances (dimer: 3, 4, 5, trimer: 2, 3, 4) | 21, 28 |
| Multicoil2 | Trigg et al. (2011) | HMM (pairwise side chain correlations) | CC, OS (dimer, trimer) | - | - |
| MARCOIL | Delorenzi and Speed (2002) | HMM and PSSM | CC | HMM parameters, PSSM matrices (MTK, MTIDK) | - |
| PCOILS | Gruber et al. (2005) | PSSM (profile-profile) | CC | PSSM matrices (MTK, MTIDK) | - |
| SpiriCoil | Rackham et al. (2010) | HMM | CC | - | - |
| DeepCoil | Ludwiczak et al. (2019) | Neuronal network (CNN) | CC | - | - |
| DeepCoil2 | | CNN | CC | - | - |
| Scorer 2.0 | Armstrong et al. (2011) | log-likelihood for di- and trimer PSSM-profiles * | OS (dimer, trimer) | - | - |
| PrOCOIL | Mahrenholz et al. (2010) | SVM * | OS (dimer, trimer) | - | - |
| LOGICOIL | Vincent et al. (2013) | Bayesian variable selection and multinomial probit regression * | OS (parallel & antip. dimer, tri-, tetramer) | - | - |

Figure 2.11. | Overview of the available coiled-coil prediction programs. CC = coiled coil, OS = oligomeric state; with * marked tools depend as input on the CC predictions of further softwares.

Over the years, a variety of probabilistic models and algorithms have been implemented to improve successively the prediction of coiled coils. The applied approaches range from position-specific scoring matrix (PSSM) over a PSSM-variant extended by pairwise residue correlations to machine-learning models like hidden Markov models (HMMs), support vector machines (SVMs) and recently convolutional neural networks (CNNs). These methods are explained briefly in the following sections using respectively one application as illustrating example.

Position-specific scoring matrix (PSSM)

The original method for coiled-coil detection in proteins was the "position-specific scoring matrix"-based approach. It was first implemented by Andrei Lupas in 1991 in the program COILS (Lupas et al., 1991) and is based on the prediction algorithm for coiled coils developed by David Parry (1982). The approach described the analysis of amino acid types in the heptad motif of known coiled coils and the usage of derived residue probabilities for the single heptad positions to predict coiled-coil domains in other unknown proteins. In COILS, based on this procedure, new protein sequences are compared with a database of known 2-stranded parallel coiled-coils by calculating a window-based similarity score. This determined score is compared with a distribution of scores from globular and coiled-coil proteins. Based on this, a probability is calculated that indicates whether the sequence forms a coiled coil. The COILS program can be run with two different scoring matrices for the detection of coiled-coil domains, the MTK and the MTIDK matrices. In the following, the MTK matrix is presented as an example (see Table 2.12). The included position scores were trained with myosin, tropomyosin and keratin sequences (intermediate filaments type I & II). The matrix has a score for each of the 20 amino acids for each position of the heptad motif (*a-g*), reflecting their probability to occur in the analyzed MTK sequences.

| | <i>a</i> | <i>b</i> | <i>c</i> | <i>d</i> | <i>e</i> | <i>f</i> | <i>g</i> |
|---|----------|----------|----------|----------|----------|----------|----------|
| L | 3.167 | 0.297 | 0.398 | 3.902 | 0.585 | 0.501 | 0.483 |
| I | 2.597 | 0.098 | 0.345 | 0.894 | 0.514 | 0.471 | 0.431 |
| V | 1.665 | 0.403 | 0.386 | 0.949 | 0.211 | 0.342 | 0.360 |
| M | 2.240 | 0.37 | 0.480 | 1.409 | 0.541 | 0.772 | 0.663 |
| F | 0.531 | 0.076 | 0.403 | 0.662 | 0.189 | 0.106 | 0.013 |
| Y | 1.417 | 0.090 | 0.122 | 1.659 | 0.19 | 0.13 | 0.1550 |
| G | 0.045 | 0.275 | 0.578 | 0.216 | 0.211 | 0.426 | 0.156 |
| A | 1.297 | 1.551 | 1.084 | 2.612 | 0.377 | 1.248 | 0.877 |
| K | 1.375 | 2.639 | 1.763 | 0.191 | 1.815 | 1.961 | 2.795 |
| R | 0.659 | 1.163 | 1.210 | 0.031 | 1.358 | 1.937 | 1.798 |
| H | 0.347 | 0.275 | 0.679 | 0.395 | 0.294 | 0.579 | 0.213 |
| E | 0.262 | 3.496 | 3.108 | 0.998 | 5.685 | 2.494 | 3.048 |
| D | 0.03 | 2.352 | 2.268 | 0.237 | 0.663 | 1.62 | 1.448 |
| Q | 0.179 | 2.114 | 1.778 | 0.631 | 2.55 | 1.578 | 2.526 |
| N | 0.835 | 1.475 | 1.534 | 0.039 | 1.722 | 2.456 | 2.280 |
| S | 0.382 | 0.583 | 1.052 | 0.419 | 0.525 | 0.916 | 0.628 |
| T | 0.169 | 0.702 | 0.955 | 0.654 | 0.791 | 0.843 | 0.647 |
| C | 0.824 | 0.022 | 0.308 | 0.152 | 0.180 | 0.156 | 0.044 |
| W | 0.24 | 0.0 | 0.0 | 0.456 | 0.019 | 0.0 | 0.0 |
| P | 0.00 | 0.008 | 0.0 | 0.013 | 0.0 | 0.0 | 0.0 |

(a) MTK matrix

| | <i>a</i> | <i>b</i> | <i>c</i> | <i>d</i> | <i>e</i> | <i>f</i> | <i>g</i> |
|---|----------|----------|----------|----------|----------|----------|----------|
| L | 2.998 | 0.269 | 0.367 | 3.852 | 0.510 | 0.514 | 0.562 |
| I | 2.408 | 0.261 | 0.345 | 0.931 | 0.402 | 0.440 | 0.289 |
| V | 1.525 | 0.479 | 0.350 | 0.887 | 0.286 | 0.350 | 0.362 |
| M | 2.161 | 0.605 | 0.442 | 1.441 | 0.607 | 0.457 | 0.570 |
| F | 0.490 | 0.075 | 0.391 | 0.639 | 0.125 | 0.081 | 0.038 |
| Y | 1.319 | 0.064 | 0.081 | 1.526 | 0.204 | 0.118 | 0.096 |
| G | 0.084 | 0.215 | 0.432 | 0.111 | 0.153 | 0.367 | 0.125 |
| A | 1.283 | 1.364 | 1.077 | 2.219 | 0.490 | 1.265 | 0.903 |
| K | 1.233 | 2.194 | 1.817 | 0.611 | 2.095 | 1.686 | 2.027 |
| R | 1.014 | 1.476 | 1.771 | 0.114 | 1.667 | 2.006 | 1.844 |
| H | 0.590 | 0.646 | 0.584 | 0.842 | 0.307 | 0.611 | 0.396 |
| E | 0.281 | 3.351 | 2.998 | 0.789 | 4.868 | 2.735 | 3.812 |
| D | 0.068 | 2.103 | 1.646 | 0.182 | 0.664 | 1.581 | 1.401 |
| Q | 0.311 | 2.290 | 2.330 | 0.811 | 2.596 | 2.155 | 2.585 |
| N | 1.231 | 1.683 | 2.157 | 0.197 | 1.653 | 2.430 | 2.065 |
| S | 0.332 | 0.753 | 0.930 | 0.424 | 0.734 | 0.801 | 0.518 |
| T | 0.197 | 0.543 | 0.647 | 0.680 | 0.905 | 0.643 | 0.808 |
| C | 0.918 | 0.002 | 0.385 | 0.440 | 0.138 | 0.432 | 0.079 |
| W | 0.066 | 0.064 | 0.065 | 0.747 | 0.006 | 0.115 | 0.014 |
| P | 0.004 | 0.108 | 0.018 | 0.006 | 0.010 | 0.004 | 0.007 |

(b) MTIDK matrix

Figure 2.12. | MTK and MTIDK matrices of the coiled coil prediction program COILS. Shown are the probabilities (scores) of the 20 amino acids to occur at each specific position of the heptad motif.

The application supports different window sizes in the calculation of the score. Possible values are 14, 21 and 28, which are whole multiples of a α -helix with two turns (length: 7 amino acids). The choice of the window size has a significant impact on the detection of coiled-coil domains. The greater window of length 28 usually gives clearer results, and is the default setting, being a compromise between sensitivity and specificity.

Pairwise residue correlations (extended PSSM-variant)

The second type of coiled-coil prediction programs is based on the method of pairwise residue correlations of amino acids in the heptad motif. The application in which this method was first implemented is PairCoil (Berger et al., 1995). The data basis of this method is again a database of known coiled-coil sequences composed of myosins, tropomyosins, and intermediate filaments. From these protein sequences, the propensity of occurrence (»frequency«) was calculated for each amino acid pair for each positional combination in the heptad motif. These pairwise frequency values form the coiled-coil database and are used to estimate the probability of a given amino acid pair for a specific positional combination in the heptad repeat. The probabilities are used to calculate the score S_k for an amino acid at position k , which is an indicator of whether that amino acid is in a coiled-coil.

To determine the score S_k , the maximum window score of all windows of length 30 in which the k th amino acid occurs is selected. When calculating the window score, the maximum of the sum of amino acid propensities over the seven possible heptad repeat positions in the window of 30 is selected (see Figure 2.13). The propensity (\approx slope, tendency; in terms of probability) of an amino acid to a given heptad position involves the correlations between the amino acid under consideration and the amino acids at the following structurally relevant positions with distances $i = 1, i = 2$, and $i = 4$. The propensity for the k -th amino acid is given by the following equation (simplified):

$$\frac{P(k, k + 1)P(k, k + 2)P(k, k + 4)}{P(k + 1)P(k + 2)P(k + 4)}.$$

The correlation positions that enter into the amino acid calculation are derived from the structure of the α -helices. These have a length of 3.5 amino acids per turn, so that positions 1 and 4 are superimposed along the helix axis (slightly offset) and form a line. With respect to the heptad motif, positions a and d are structurally related in this way. This arrangement, as already described in section 2.2.3, is a prerequisite for the formation of coiled-coils.

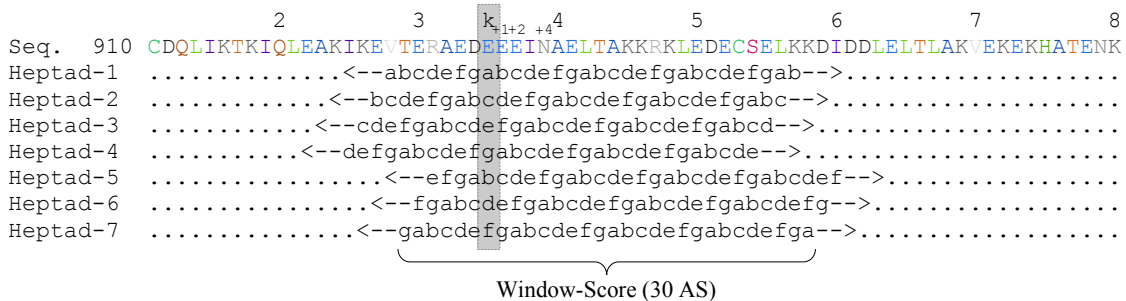


Figure 2.13. | Pairwise residue correlations: illustration of the maximum window score calculation. Displayed are the scoring windows for the different heptad repeat positions used to calculate the maximum window score based on the summed amino acid propensities. For clarification, a position k is marked with its structurally relevant positions $d = 1, 2, 4$.

Hidden Markov model (HMM)

One of the first applications whose coiled-coil prediction algorithm in protein sequences is based on a HMM is the program Marcoil (cf. Delorenzi and Speed, 2002). It was developed in 2001 by the authors Mauro Delorenzi and Terry Speed. The strength of this method has been the prediction

specifically of short CCDs and several related protein families, which could not be detected with previous developed programs using PSSM approaches. One reason for this is that the underlying HMM is more flexible, due to its window-less and state-based design. In contrast to the assessment of amino acid sequences with a fixed-width scoring window, the advantage of HMMs is the ability to evaluate protein sequences with dynamic lengths and more diverse patterns. The authors found that previous methods such as "residue-based correlations" are too specific for the detection of general as well as new coiled-coil classes because the window-based methods have due to their fixed-width design two weaknesses. First, if the window is longer than the domain, neighboring amino acids that do not belong to a coiled-coil domain will be included, resulting in a washed-out score, and second, if it is shorter, not enough evidence can be reached for the considered region.

The software Marcoil makes use of an hidden Markov model consisting of 64 states, that are divided into a background state "0" and 63 further states to model the behavior of being in a coiled-coil domain. The coiled-coil states build nine groups numbered from "1-9" consisting of seven heptad positions each, labeled with "a-g". The state "0" realizes the background or reference state, which includes all sequence regions outside putative coiled-coil domains. The groups "1-4" model the first four amino acids in a CCD and groups "6-9", the last four. Group "5" describes all CC amino acids included in between. Analyzable sequences must have a minimum length of nine amino acids, since each group must occur once, as well as begin and end again at state "0", which is not counted.

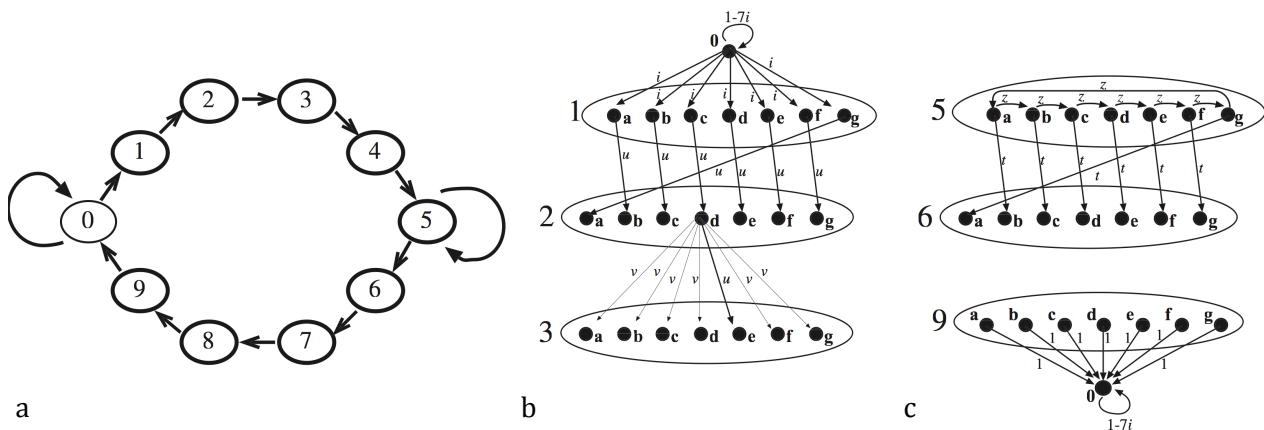


Figure 2.14. | Marcoil: representation of the HMM-based approach. Part (a) shows a simplified representation of the HMM depicting the logical order at the group level. Parts (b) and (c) illustrate the possible transitions of the heptad states to each other within and between the single heptad groups. Adopted with permission from Delorenzi and Speed (2002, p. 619 ff.), Copyright 2002 Oxford University Press.

Each of the shown heptad states of the groups "1-9" emits amino acids according to a state-specific probability distribution. Individual "transitions" between these heptad states can be traced in the right two Figures 2.14b and 2.14c. According to the principle shown, each state has specific transition probabilities with which it can move to its subsequent states either within the current group or into the next group. Transitions following the predefined order of the heptad motif (*transition u* or *z*: $a \rightarrow b, b \rightarrow c, c \rightarrow d, \dots$), which not necessarily need to start at heptad position *a*, are favored. Other non-heptad transitions, represented by *v* or *t*, that would cause a jump in the motif have lower probabilities. For a given amino acid sequence produced by the hidden Markov chain, to illustrate the transitions over two heptad repeats of a putative coiled-coil domain, the associated state chain could

have the following form: The sequence were to start at heptad position c , $0 - 1c - 2d - 3e - 4f - 5g - 5a - 5b - 5c - 5d - 5e - 6f - 7g - 8a - 9b - 0$. The joint probability for such a protein sequence α with a given state chain π , representing the heptad occupancy, for this HMM is calculated as follows:

$$1. P[\alpha, \pi] = \tau(0, \pi(1)) \prod_{t=1}^n [\epsilon(\pi(t), \alpha(t)) \cdot \tau(\pi(t), \pi(t+1))]$$

$\alpha(t)$ = Sequence at position t of length n ($1 \leq t \leq n$)

$\pi(t)$ = State chain at position t of length $n + 2$ ($0 \leq t \leq n + 1$)

$\tau(r, s)$ = Transition from state $r \rightarrow s$

$\epsilon(s, a)$ = Emission probability for amino acid a in state s

$$2. P[\pi(t) = s | \alpha] = \frac{P[\alpha, \pi(t) = s]}{P[\alpha]} \text{ (conditional/posterior probability)}$$

The actual parsing of a given protein sequence α to make a coiled coil prediction is based on posterior probabilities with the forward-backward algorithm (Rabiner, 1989). If the calculated probability for a single amino acid lies above a certain threshold, the software assigns it to putative coiled-coil domain similar to the principle used in the PSSM approach.

2.3.4. Analysis software for secondary structure prediction

A kind of software to detect secondary structure domains and regions (e.g. α -helices, β -sheets, coiled coils) in structural data of resolved protein molecules and that have been extensively used in the course of this work include the following programs.

Database of secondary structure assignments (DSSP)

One application that can be used to extract protein secondary structures from PDB files is the tool DSSP (Kabsch and Sander, 1983). Similar to the PDB database, the term DSSP is a database for which there is an application of the same name. The program DSSP computes for a 3D structure of a given PDB file the most likely applicable secondary structures. For this an algorithm is used, which recognizes hydrogen bond patterns. For this purpose, DSSP calculates the possible hydrogen bonds from all atoms of a protein and selects the best two from these for each atom. This structural information is used to assign the best-fitting secondary structure class to each amino acid. The algorithm underlying the tool follows clear calculations and makes no predictions.

DSSP can distinguish between the following seven secondary structure classes:

- B = Amino acid in isolated β -bridge
- E = Extended β -strand, that is part of a β -sheet
- G = 3/10-helix
- H = α -helix
- I = 5 π -helix
- S = Bend
- T = Hydrogen-bonded turn
- _ = Loop or irregularity

SOCKET

The program "SOCKET" developed by Walshaw and Woolfson (2001) is an application designed to find coiled-coil domains in structurally resolved proteins as collected and provided by the Protein Data Bank. In the program, an algorithm is applied that searches for "knobs-into-holes" packing interactions (KIH) in the 3D structural model of helical proteins, as described by Crick (1953). This refers to a model in which a simplified form of the helix side chains is assumed, designated as *knobs*. They are assigned the same shape and size for this purpose. The regions between the *knobs* are called *holes*. If two helices form a simple coiled-coil, they wrap around each other so that the *knobs* fit exactly into the *holes* and build a compact packing shape, for a more detailed description see section 2.2.3.

The algorithm is able to detect where and in what form the coiled-coil structures occur in the proteins under investigation. Also part of the analysis is the determination of the oligomerization state. It determines the number of helices with their orientation (parallel, anti-parallel) and the composition of the coiled-coils with respect to the helical sequences involved (homomer or heteromer).

A prerequisite for the calculation of coiled-coils using SOCKET is the prepared secondary structure analysis for a PDB file using the program DSSP, which was introduced in the previous section 2.3.4.

2.3.5. Biological databases and exchange formats

The explosion in the amount of biological data, independent of the type, goes hand in hand with the need to be accessible for all kinds of scientific research. For this reason there is a strong demand for reliable solutions to collect, provide and exchange the generated masses of data to ensure the idea of "open-access" and the FAIR principles for good scientific practice. The effortless usability helps in further studies to evaluate the gained biological insights and to clarify with these fundamental biological questions. In the following, relevant database resources and formats are addressed that made the developments and analyses of this work possible in the first place.

Protein Data Bank

The PDB is the single global archive of experimentally determined 3-dimensional structure data of biological macromolecules managed by the international wwPDB consortium. It exists in its infancy since the early 1970s, provided formerly by the Brookhaven National Laboratory and the Research Collaboratory for Structural Bioinformatics (RCSB), and was the first open-access resource for validated and biocurated structure data. Nowadays the PDB is one of the largest resources for structural data of biological macromolecules, particularly structurally resolved proteins, and is essential for biological and medical studies working on the deciphering of the functional roles of proteins. Structural data are usually obtained using experimental methods such as macromolecular crystallography (MX), nuclear magnetic resonance spectroscopy (NMR), 3D electron microscopy (3DEM) and electron tomography (ET) (Berman et al., 2000; wwPDB consortium, 2019). The database resource is available to the general public both as a website and in the form of further web services and protocols (e.g. SOAP, REST, FTP) over which the structural data can be obtained. The protein structures are available in individual files that can be accessed via a unique (four-digit, alpha-numeric) identifier. The database contains currently about 180,000 solved protein structures with increasing tendency (State: 2021/01, see Figure 2.15 for detailed information).

As described in the publication of Simm et al. (2021), see section 3.1.4 on the "Critical assessment of coiled-coil predictions based on protein structure data", the custom protein sequence and structure database of this work used as one part of the data basis a comprehensive set of all available records deposited in the PDB, more precisely the parsed and selective processed information of the individual PDB structures of this database resource.

PDB and PDBx/mmCIF

A file format with the same name exists for the PDB, in which the structural data of the proteins can be exchanged. The original PDB file format (*.pdb*) is a text-based exchange format for structurally solved macromolecules. In contrast to the FASTA or GenBank file format, there is a much more complex specification for the structure of PDB files (cf. *Atomic Coordinate Entry Format Description*). Each of these files describes a molecule complex constructed from one or multiple protein sequences (named *chains*) originating of the same or different interacting proteins. PDB files contain in the header section annotating meta-information about the molecule, such as the name and type of protein, a description of the analysis environment and molecule composition, the solution method (*NMR*, *X-Ray*, etc.), the number of chains contained, the journal of publication and some more. Furthermore, it contains the amino acid sequences for each of the included chains, as well as the "*accession numbers*" of the databases where the sequences can be found with the related location information (UniProt, SwissProt).

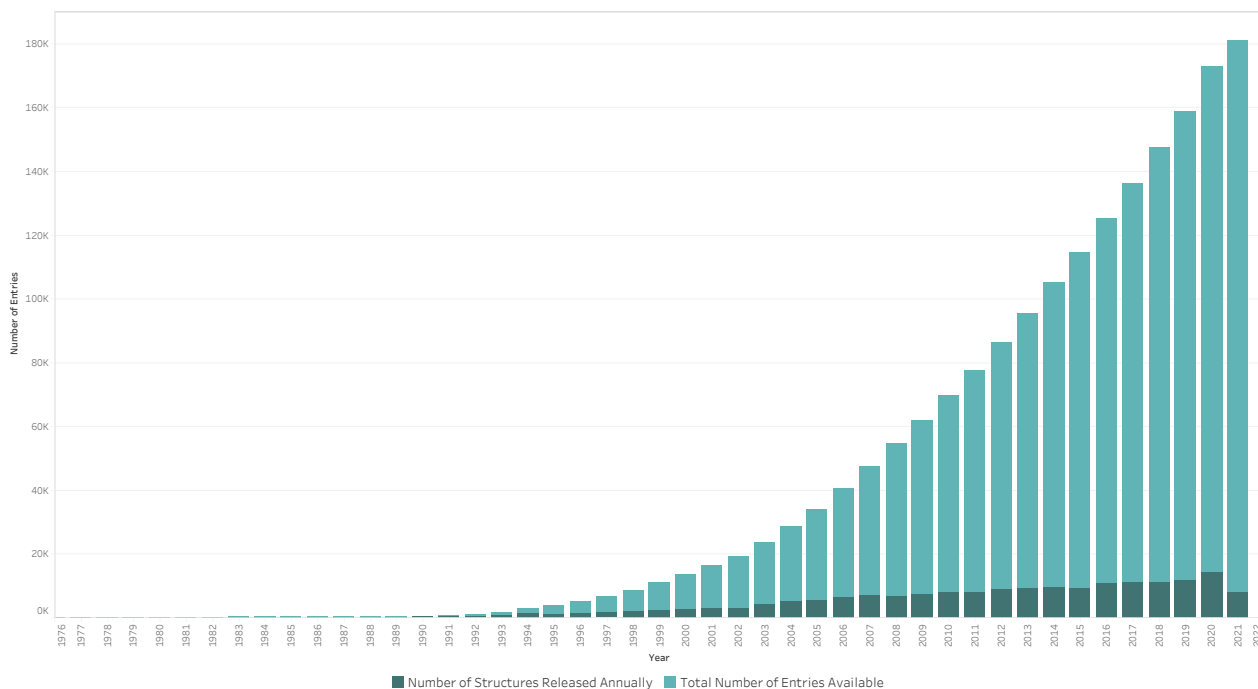


Figure 2.15. | Growth of the RCSB Protein Data Bank. The diagram illustrates the amount of released structures stored in the PDB of the Research Collaboratory for Structural Bioinformatics (RCSB) from 1976 to 2021. The shown information are the number of structures released and the total number of entries available on annually basis. The numbers were obtained from the PDB statistics information page (<https://www.rcsb.org/stats/growth/growth-released-structures>).

The important main component of the format is formed by position information of the atoms of the protein molecule in the 3-dimensional space. In Table 2.16 some important format elements are shown in extracts. In the upper part, the positions of the individual molecular atoms can be seen, with the atoms of single amino acids listed from the amino to the carboxyl terminus. There is a separate section for each chain, identifiable by the column "CH". The column entries from left to right have the following meaning: The first position shows the type of entry, position 2 represents a numbering for each "atom", column 3 gives the name of the atom, column 4 the amino acid residue to which it is assigned, column 5 the numbering of the amino acids in the overall sequence, columns 6-8 give the coordinates in the three dimensions (x , y , z) measured in Å, the penultimate column represents the temperature factor, and the last specifies the element type of the atom.

The section shown below indicates secondary structure elements identified in the molecule (α -helix, β -sheet). Basically, both have a similar structure. Column 1 specifies the type of entry, column 2 contains a sequential position number, and in the case of the *helix*-section, it is now followed by two blocks consisting of an amino acid, the chain, and a position in the sequence. The first block indicates the start position and the second the end position of the helix. The last column represents the length of the helix domain. In the *sheets* section, the strands involved in the β -fold are listed, again with start and end blocks, a "sense" column indicating the orientation of the current strand in the fold, and in the last two blocks, the positions over which the strands are related.

The constant technological improvement of the structure determination methods led over the years

| | | | | | | | | | | | | | | | | | | | |
|----|-------|--------|------|-----|-----|-------|--------|--------|---------|---------|-------|-----|-----|----|------|-----|-----|----|------|
| 1 | TYP | AT-Nr | NAME | AA | CH | AA-Nr | X | Y | Z | TempFac | EL | | | | | | | | |
| 2 | ATOM | 1 | N | VAL | A | 3 | 23.607 | 44.559 | 142.863 | 1.00 | 65.12 | N | | | | | | | |
| 3 | ATOM | 2 | CA | VAL | A | 3 | 22.448 | 44.382 | 143.759 | 1.00 | 63.91 | C | | | | | | | |
| 4 | ATOM | 3 | C | VAL | A | 3 | 22.675 | 43.119 | 144.620 | 1.00 | 61.91 | C | | | | | | | |
| 5 | ATOM | 4 | O | VAL | A | 3 | 22.190 | 43.110 | 145.783 | 1.00 | 63.90 | O | | | | | | | |
| 6 | ATOM | 5 | CB | VAL | A | 3 | 21.099 | 44.302 | 143.002 | 1.00 | 64.88 | C | | | | | | | |
| 7 | ATOM | 6 | CG1 | VAL | A | 3 | 20.910 | 45.344 | 141.894 | 1.00 | 63.70 | C | | | | | | | |
| 8 | ATOM | 7 | CG2 | VAL | A | 3 | 20.792 | 42.904 | 142.442 | 1.00 | 64.23 | C | | | | | | | |
| 9 | ATOM | 8 | N | GLN | A | 4 | 23.379 | 42.168 | 144.005 | 1.00 | 57.23 | N | | | | | | | |
| 10 | ATOM | 9 | CA | GLN | A | 4 | 23.639 | 40.903 | 144.713 | 1.00 | 52.30 | C | | | | | | | |
| 11 | ATOM | 10 | C | GLN | A | 4 | 25.083 | 40.778 | 145.190 | 1.00 | 48.05 | C | | | | | | | |
| 12 | ATOM | 11 | O | GLN | A | 4 | 26.049 | 41.216 | 144.554 | 1.00 | 46.24 | O | | | | | | | |
| 13 | ATOM | 12 | CB | GLN | A | 4 | 23.174 | 39.721 | 143.856 | 1.00 | 53.45 | C | | | | | | | |
| 14 | ATOM | 13 | N | ALA | A | 5 | 25.158 | 40.150 | 146.360 | 1.00 | 42.50 | N | | | | | | | |
| 15 | [...] | | | | | | | | | | | | | | | | | | |
| 16 | TYP | HLX-Nr | [AA | CH | ST] | [AA | CH | END] | | | LEN | | | | | | | | |
| 17 | HELIX | 1 | 1 | THR | A | 6 | ASP | A | 9 | 5 | 4 | | | | | | | | |
| 18 | HELIX | 2 | 2 | LEU | A | 15 | GLY | A | 19 | 1 | 5 | | | | | | | | |
| 19 | HELIX | 3 | 3 | ASP | A | 35 | GLY | A | 47 | 1 | 13 | | | | | | | | |
| 20 | HELIX | 4 | 4 | HIS | A | 54 | VAL | A | 59 | 1 | 6 | | | | | | | | |
| 21 | HELIX | 5 | 5 | ASP | A | 64 | GLY | A | 83 | 1 | 20 | | | | | | | | |
| 22 | HELIX | 6 | 6 | HIS | A | 96 | LYS | A | 100 | 5 | 5 | | | | | | | | |
| 23 | [...] | | | | | | | | | | | | | | | | | | |
| 24 | TYP | STR-Nr | CH | ID | [AA | CH | ST] | [AA | CH | END] | SEN | [AT | AA | CH | POS] | [AT | AA | CH | POS] |
| 25 | SHEET | 1 | A | 4 | GLY | A | 50 | THR | A | 52 | 0 | | | | | | | | |
| 26 | SHEET | 2 | A | 4 | PHE | A | 11 | GLY | A | 14 | 1 | 0 | PHE | A | 11 | N | GLY | A | 50 |
| 27 | SHEET | 3 | A | 4 | ARG | A | 289 | PHE | A | 291 | 1 | 0 | ARG | A | 289 | N | SER | A | 12 |
| 28 | SHEET | 4 | A | 4 | ASP | A | 245 | LEU | A | 246 | 1 | N | LEU | A | 246 | 0 | HIS | A | 290 |
| 29 | [...] | | | | | | | | | | | | | | | | | | |

Figure 2.16. | Extracts from a PDB flat file.: Listed are the detailed atomic positions of the single sequence residues and identified structures (e.g. α -helix, β -sheet) within the annotated macromolecule (PDB-ID: 9XIM).

to an increasing size of the solvable macromolecules with corresponding changes in the form of the structural data, which was accompanied by equally increasing requirements for the PDB file format. The basic type of the described format is called *PDB flat file* and is for this reason designed as a continuously extensible data format. The need to be ongoing extendable, leads at the same time to problems with the backward compatibility to former deposited structures. So that older versions of the format need to be reprocessed and corrected to be compatible to the new standards of the file format. To meet the new requirements, the same structural data can also be retrieved in alternative formats such as PDBML (XML version) and the extensible PDBx/mmCIF master format.

These circumstances bring a certain potential for read and parsing errors of the structural flat files with it, which indeed was a problem for the analyses in this work during the preparation of the relational structure database of the project holding a comprehensive copy of all deposited structures of the PDB, as well as SOCKET and coiled-coil prediction data.

CC+ Database

The CC+ Database (Testa et al., 2009) represents a specialized and comprehensive resource for coiled-coil structures and sequences. It has been developed with the aim to improve the understanding of the sequence-to-structure relationships for this type of domain. The contained structures are based on scans of the Protein Data Bank (wwPDB consortium, 2019) performed with the structure analysis program SOCKET (Walshaw and Woolfson, 2001), an algorithm that identifies coiled coils by the detection of KIH packing interactions of side chains between α -helices in the deposited protein

structural data. The analysis enabled a classification of numerous coiled-coil structures into a periodic table of coiled-coil architectures according to their oligomeric states and thus provides profound information about the assembling diversity of coiled coils (Moutevelis and Woolfson, 2009). Besides the PDB, the CC+ database is an important resource for the collection of representative reference sets for structural data-based studies, because it enabled due to its relational design efficient filtering mechanisms and a rapid compilation of subsets of coiled-coil structures.

PaxDB

The PaxDb (Wang et al., 2015) stands for "Protein Abundances Across Organisms" and is a public database resource for protein abundance information of the most common model organisms hosted by the Swiss Institute of Bioinformatics. It provides whole-proteome datasets being captured with different biophysical and mass spectrometry (MS) techniques, that originate from various unstructured online resources such as publication supplementals and custom data repositories. The necessity of the database lies in the circumstances that data processing and its reuse in proteomics is a challenging task due to the rapid technical advances in proteomics and the low requirements for standardized comparable data. For this reason, the PaxDB offers access to reprocessed, unified and quality-scored protein quantification data in one place. The normalized abundance data can be retrieved in the common text-file exchange formats CSV, TSV and JSON.

The provided datasets of the model organisms *Saccharomyces cerevisiae*, *Escherichia coli* and *Arabidopsis thaliana* build the data basis for the parameter training of the Markov model used in the project on heterologous protein expression. The quantified proteomic data allowed the generation of model profiles for different expression levels (e.g. low, mid and highly expressed proteins).

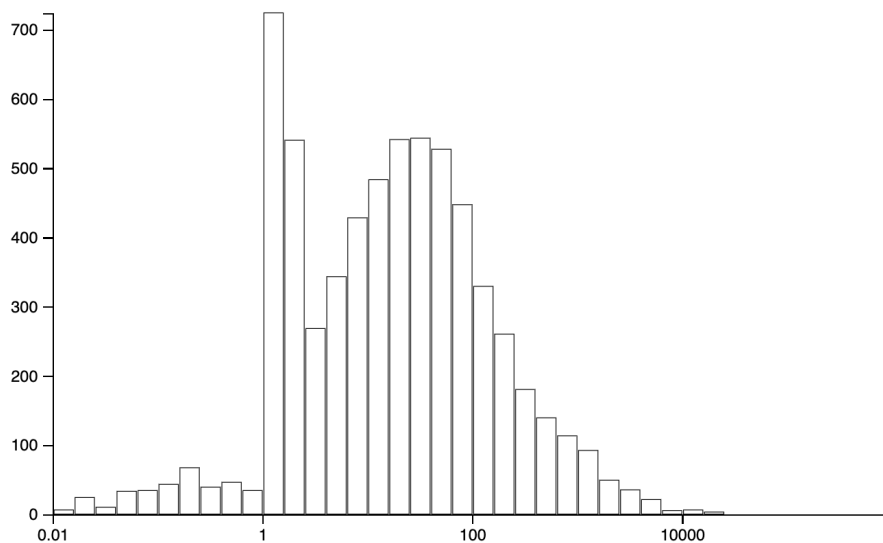


Figure 2.17. | Protein abundance histogram. The diagram illustrates the protein abundance distribution of the quantified proteome for the model organism *Saccharomyces cerevisiae*. The abundance of each measured protein (in ppm) is plotted against the number of proteins grouped in the respective abundance range. The numbers were obtained from the PaxDB entry for *S.cerevisiae* (<https://pax-db.org/dataset/4932/3>).

REBASE

The Restriction enzyme database (REBASE) (Roberts et al., 2010) is a comprehensive and manually curated resource, that provides information about restriction enzymes and related proteins, such as methylases, and the microorganisms from which they originate, dating back until the early 1950s and is updated on a daily basis. For the listed enzymes detailed information about recognition sequences, cleavage sites, methylation specificity and the commercial availability can be retrieved. The information can be accessed dynamically via a search interface or on file basis in the common text-file exchange formats.

2.4. Computational approaches

In this work, new approaches and analyses are presented to assess and improve the current state of the structural annotation of »coiled-coil domains« in proteins as well as computational prerequisites for experimental »protein expression« applications. The major goal behind these is to contribute to the deciphering of mechanistic and cellular functions of proteins to enable biological insights and a deeper understanding of living processes in the organisms of the tree of life. The performed approaches clearly profited from the ever increasing amount of sequential and structural protein data available in the relevant resources. A development that made biological investigations with a statistically profound and meaningful data basis possible in the first place, such as it was used in the study on the »Critical assessment of coiled-coil predictions based on protein structure data« by the usage of the comprehensive data set including all available protein structures deposited in the PDB, see Simm et al. (2021). But not only protein data builds a central aspect of the performed work, also sequenced genome data enabling applications like the profiled codon usage adaptation approach, that are based on capturing of whole genome characteristics are nowadays possible to be implemented. The approaches, tools and results developed during the course of this work are made available for the public as freely accessible and easy to use web-applications with plenty of interactive setting options.

The computational approaches developed were implemented on a broad basis of softwares, libraries and frameworks. The different requirements of the both projects and changing technological capabilities have led with Ruby, which was used for the whole development of »Waggawagga«, and Python (»Odysseus«) to the usage of different programming languages over time. For the implementation of the web-applications of both projects the web-frameworks Ruby on Rails and Flask came into use, both providing necessary packages to realize fast and modern web-services with builtin API interfaces for external data access. The standalone version »Waggawagga-CLI« is implemented with the portable Ruby environment (Traveling Ruby) and the mobile database SQLite to run the SAH prediction software easy installable on the machines of end-users. The extensive coiled-coil prediction dataset including the complete PDB protein structures of the Simm et al. (2021) study is stored in a high performant relational PostgreSQL database, that constitutes the data backend. Extensively used packages for PDB structure parsing and the visual inspection of the different coiled-coil predictions on single PDB structures, are BioRuby and the Web-GL based PDB-viewer JSMol and ChemDoodle Web Components. RNA secondary structures and minimum folding energies are predicted by RNAFold from the ViennaRNA package. Figures, plots and illustrations in Waggawagga and Odysseus are realized with the embeddable XML-based SVG file format, gnuplot and ImageMagick. The developed projects are realized as well as Docker images and containers for development and production operations, ensuring the simple installation and long-term, low-maintenance availability and operability. This enables operation on various operating systems such as Linux distributions, macOS and Windows systems.

Nowadays, sophisticated and reliable computational approaches to process, analyse and interpret the ever-increasing amounts of experimentally generated sequential and structural data build an essential part in the investigation of biological questions by complementing the elaborate experiments in the labs.

3

Publications and Manuscripts

This chapter includes the publications and manuscripts, that have been authored in the context of the PhD project, and presents them in thematic and chronological order. The conducted studies belong to the both major topics »comparative protein structure prediction of coiled-coil and SAH domains« and »heterologous gene expression« and are listed in the first two sections. These are followed by additional works located in the fields of »genome annotation« and »genomic databases«, in which minor contributions by the author were made.

The first section of this chapter comprises in multiple studies the development and evaluation of new approaches to improve the reliable »prediction of coiled-coil domains« for individual cases and on a large scale as for genome annotation purposes. In conclusion, it also presents an in-depth study for the critical assessment of the current state of »protein structure prediction of coiled-coil domains« performed and measured against one of the most comprehensive reference datasets for protein structural data based on the entire Protein Data Bank.

The second section includes the work on the design of typical genes for heterologous gene expression and its functional evaluation carried out and validated by experimental studies. For this purpose a configurable information theory based approach has been developed, that can be operated with pre-trained genome profiles to control protein expression levels by codon usage adaptation.

Contents

| | |
|--|-----|
| 3.1. Comparative protein structure prediction of coiled-coil and SAH domains | 39 |
| 3.1.1. Waggawagga: comparative visualization of coiled-coil predictions and detection of stable single α -helices (SAH domains) | 39 |
| 3.1.2. Distribution and evolution of stable single α -helices (SAH domains) in myosin motor proteins | 43 |
| 3.1.3. Waggawagga-CLI: A command-line tool for predicting stable single α -helices (SAH-domains), and the SAH-domain distribution across eukaryotes | 60 |
| 3.1.4. Critical assessment of coiled-coil predictions based on protein structure data | 80 |
| 3.2. Heterologous gene expression | 99 |
| 3.2.1. Design of typical genes for heterologous gene expression | 99 |
| 3.3. Genome annotation and genomic databases | 121 |
| 3.3.1. diArk – the database for eukaryotic genome and transcriptome assemblies in 2014 | 121 |
| 3.3.2. The landscape of human mutually exclusive splicing | 128 |
| 3.3.3. Identifying Sequenced Eukaryotic Genomes and Transcriptomes with diArk | 151 |
| 3.3.4. The Protein-Coding Human Genome: Annotating High-Hanging Fruits | 171 |

3.1. Comparative protein structure prediction of coiled-coil and SAH domains

3.1.1. Waggawagga: comparative visualization of coiled-coil predictions and detection of stable single α -helices (SAH domains)

Dominic Simm, Klas Hatje and Martin Kollmar *

* Corresponding author

Group Systems Biology of Motor Proteins, Department of NMR-based Structural Biology, Max-Planck-Institute for Biophysical Chemistry, Göttingen, Germany

Bioinformatics (2015) - Volume 31(5): 767-769

doi: 10.1093/bioinformatics/btu700

<http://bioinformatics.oxfordjournals.org/content/31/5/767.long>

Contribution

DS and MK designed the software with input from KH. DS developed the software, set up the web application, and performed data engineering and all computations. KH contributed to database design. DS and MK performed data analysis, designed the figures and drafted the manuscript. All authors discussed the results and commented on the manuscript.

Sequence analysis

Waggawagga: comparative visualization of coiled-coil predictions and detection of stable single α -helices (SAH domains)

Dominic Simm, Klas Hatje and Martin Kollmar*

Department of NMR-based Structural Biology, Max-Planck-Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on July 14, 2014; revised on October 17, 2014; accepted on October 20, 2014

Abstract

Summary: Waggawagga is a web-based tool for the comparative visualization of coiled-coil predictions and the detection of stable single α -helices (SAH domains). Overview schemes show the predicted coiled-coil regions found in the query sequence and provide sliders, which can be used to select segments for detailed helical wheel and helical net views. A window-based score has been developed to predict SAH domains. Export to several bitmap and vector graphics formats is supported.

Availability and implementation: <http://waggawagga.motorprotein.de>

Contact: mako@nmr.mpibpc.mpg.de

1 Introduction

Coiled coils are α -helical structural domains common to all domains of life and present in 2–12% of the proteins of a proteome (Liu and Rost, 2001). The classical coiled coils comprise dimerising long α -helices that wind around each other forming superhelices (Crick, 1952) as found in structural proteins such as α -keratins, muscle myosins and tropomyosin (KMTs). In the last years, this view has shifted to defining coiled-coil segments based on the presence of knobs-into-holes packing of side chains between α -helices resulting in many different architectures and topologies (Moutevelis and Woolfson, 2009). The basis for a coiled coil is an amino acid heptad ('abcdefg'), in which the 'a' and 'd' positions are occupied by hydrophobic residues.

Coiled coils were among the first structural domains to be predicted by algorithms (Lupas *et al.*, 1991). In this method a query sequence is compared with a database of known coiled-coil sequences, which were the KMTs available at that time. Subsequently, a similarity score is computed and a probability to form a coiled-coil calculated. In principle, this is an implementation of the basic ideas already presented nine years before, that residues show an asymmetric distribution within the heptad repeats and that this statistical

data can be used to predict coiled coils in other proteins (Parry, 1982). This position-specific scoring matrix (PSSM) approach has been improved both on the database site and on the feature site trying to disfavor the assignment of high coiled-coil probabilities to hydrophilic sequences.

Another approach to predict coiled coils is based on the pairwise residue probabilities as implemented in Paircoil (Berger *et al.*, 1995) and Paircoil2 (McDonnell *et al.*, 2006). Here, pairwise frequencies of heptad residues are calculated from known coiled-coil sequences and the probability of a pair of amino acids in a given sequence for a certain combination in the heptad is scored. With this approach better predictions for long coiled-coil regions could be obtained compared to PSSM predictions. Improvement of the prediction of short coiled coils has been reached by using a hidden Markov model as in Marcoil (Delorenzi and Speed, 2002), and, recently, Markov Random Fields as used in Multicoil2 (Trigg *et al.*, 2011). These general coiled-coil prediction approaches have been extended by software to predict different oligomerization states as Scorer (Armstrong *et al.*, 2011), PrOCoil (Mahrenholz *et al.*, 2011), and LOGICOIL (Vincent *et al.*, 2013). However, all these approaches provide different results and fail to distinguish between coiled coils and stable single α -helices (SAH domains). A special case of SAH

domains consisting of alternating repeats of four glutamic acid (E) residues and four positively charged residues of either lysine (K) or arginine (R) has been termed ER/K motif (Sivaramakrishnan *et al.*, 2008). SAH domains are in most cases mispredicted as coiled coils but have been shown to form stable monomeric structures in aqueous solution (Knight *et al.*, 2005; Süveges *et al.*, 2009). SAHs are highly enriched in E, Q, K and R residues, which stabilize the α -helices through intrahelical salt bridges (Peckham and Knight, 2009).

Waggawagga is designed to provide a direct schematic comparison of many coiled-coil prediction tools. Users can inspect the predictions in classical helical wheel (Schiffer and Edmundson, 1967) and helical net (Dunnill, 1968) representations. Visualization of the coiled-coil predictions in helical wheel schemes provides the possibility to fast and easily identify potential hydrophobic seams in 'a' and 'd' and oppositely charged residues in 'e' and 'g' positions, respectively. In contrast, SAH domains have patterns of highly charged residues only occasionally interrupted by hydrophobic or hydrophilic amino acids. Therefore, charged residues are highly enriched in the 'a' and 'd' positions. The hydrogen-bonded and charged interaction network in SAH domains is best seen in helical net representations showing the potential interactions along the helix. Waggawagga provides layouts that can easily be used in presentations and manuscripts.

2 Features

Waggawagga allows the comparative analysis of six coiled-coil prediction (Marcoil, Multicoil, Multicoil2, Ncoils, Paircoil, Paircoil2) and three oligomerization state prediction programs (Scorer, PrOCoil and LOGICOIL). In addition, Multicoil2 distinguishes dimers, trimers and non-coiled-coil oligomerization states. These tools can be run in any combination against single or multiple query sequences.

2.1 Domain view

The interactive domain view (Fig. 1A) is the main control element for setting the two analysis views, the helical wheel (Fig. 1B) and the helical net views (Fig. 1C). Separate schemes are generated for each of the results of the selected coiled-coil prediction programs visualizing the coiled-coil regions on top of the query sequence. For each region, the predicted oligomerization state is given depending on the selected tools. In the selected domain scheme, any specific region can be chosen by mouse clicks or by moving an interactive slider. The respective region is shown in detail in the helical wheel and the helical net views.

2.2 Helical wheel view

The exact sequence borders of the coiled-coil region and several different types of helix arrangements (parallel dimer, anti-parallel dimer, trimer) can be set in the configuration panel. In the helical wheel view, every α -helix is shown along the helix axis and the helices are arranged that the hydrophobic core is in the interface of the two or three helices. Each helical wheel represents 10 helix turns.

2.3 Helical net view

Intra-helical interactions are displayed in the helical net view, which is the representation of a helix split open along a line parallel to its axis and laid flat. Solid and dashed lines mark strong and weak interactions, respectively. Strong and weak interactions are mainly formed by oppositely charged residues from subsequent turns resulting in hydrogen-bonded networks in addition to charge interactions, with the strength of the interaction given by the distance between the residues and their relative orientation. Interaction networks lead to extra stabilization in addition to that of the component pairs, and hydrophobic seams favor helix association in contrast to single helices. The so-called SAH-score is calculated as the sum of the

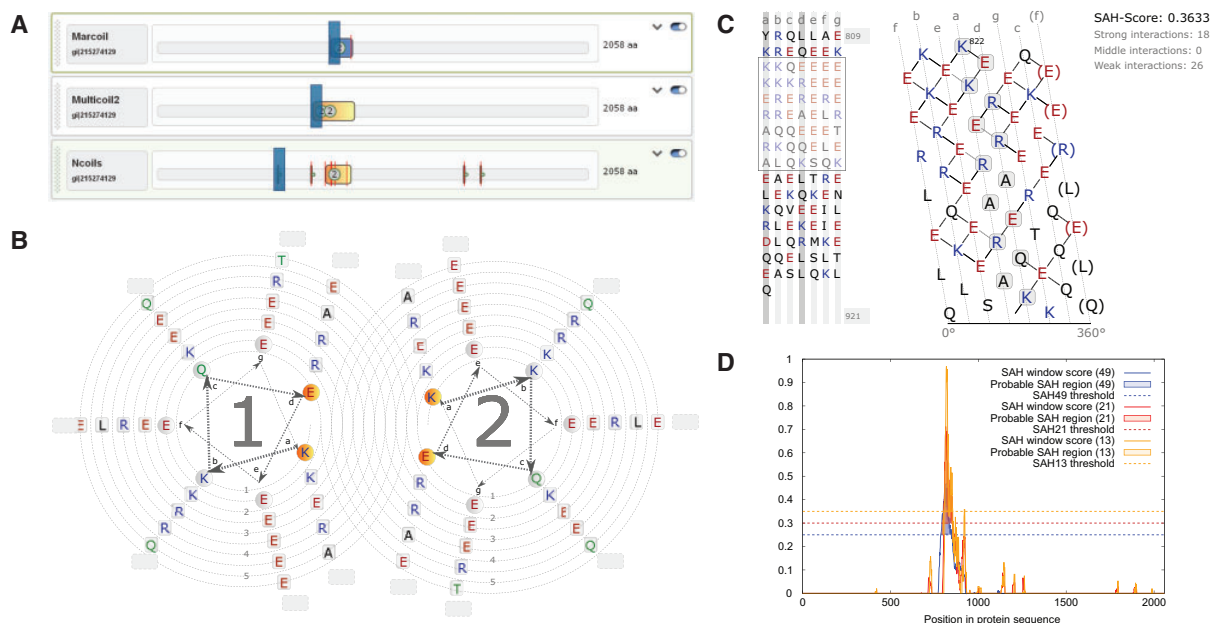


Fig. 1. (A) Interactive domain view of the coiled-coil prediction of human myosin-10, with the Marcoil prediction activated. (B) Helical wheel view. (C) Helical net view. 'a' to 'f' refer to the heptad positions. The 'f' residues are repeated on both sides of the net to better visualize intrahelical interactions with the residues in 'b' and 'c' positions. Solid and dotted lines between residues denote strong and weak intrahelical interactions, respectively, as have been introduced by others (Peckham and Knight, 2009). (D) SAH line plots for different window sizes

interactions divided by the window-size resulting in values from 0 to 1. SAHs typically have values greater than 0.25.

2.4 Line plots and tables

Here, the prediction scores are presented as line plots and tables (Fig. 1D). These are particularly useful to evaluate the relevance of SAH regions.

2.5 Performance and limitations

The SAH prediction scheme and score have been developed and tested against previously reported regions from human proteins (Peckham and Knight, 2009), SwissProt proteins (Süveges *et al.*, 2009), and thousands of cytoskeletal and motor proteins from all across the eukaryotes as available from CyMoBase (<http://www.cymobase.org>). There are other proteins that are mispredicted as coiled coils although forming monomeric α -helical structures, such as stathmin, which, however, are not enriched in E, Q, K, and R residues and do not form stable structures in solution (Honnappa *et al.*, 2006). These types of proteins do not contain SAH domains according to their definition (Peckham and Knight, 2009; Süveges *et al.*, 2009), and are consequently not detected and characterized as monomeric α -helical proteins in the current implementation of Waggawagga.

3 Implementation

The web application framework is Ruby on Rails. In order to present the user with a feature rich interface the site makes extensive use of Ajax (Asynchronous JavaScript and XML) using jQuery (<http://jquery.com>) and FancyBox (<http://fancybox.net>). Interactive schemes are drawn as SVG and graphs are generated with the graphical toolkit GnuPlot (<http://www.gnuplot.info>). Figure export is done through an intermediary SVG file, which is converted into various output formats using the Inkscape graphics package (<http://inkscape.org>). User-uploaded data is stored temporary on the server and deleted when leaving the application.

Acknowledgement

We would like to thank Christian Griesinger for his continuous generous support.

Conflict of Interest: none declared.

References

- Armstrong, C.T. *et al.* (2011) SCORER 2.0: an algorithm for distinguishing parallel dimeric and trimeric coiled-coil sequences. *Bioinformatics*, **27**, 1908–1914.
- Berger, B. *et al.* (1995) Predicting coiled coils by use of pairwise residue correlations. *Proc. Natl Acad. Sci. USA*, **92**, 8259–8263.
- Crick, F.H.C. (1952) Is alpha-keratin a coiled coil? *Nature*, **170**, 882–883.
- Delorenzi, M. and Speed, T. (2002) An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. *Bioinformatics*, **18**, 617–625.
- Dunnill, P. (1968) The use of helical net-diagrams to represent protein structures. *Biophys. J.*, **8**, 865–875.
- Honnappa, S. *et al.* (2006) Control of intrinsically disordered stathmin by multisite phosphorylation. *J. Biol. Chem.*, **281**, 16078–16083.
- Knight, P.J. *et al.* (2005) The predicted coiled-coil domain of myosin 10 forms a novel elongated domain that lengthens the head. *J. Biol. Chem.*, **280**, 34702–34708.
- Liu, J. and Rost, B. (2001) Comparing function and structure between entire proteomes. *Protein Sci.*, **10**, 1970–1979.
- Lupas, A. *et al.* (1991) Predicting coiled coils from protein sequences. *Science*, **252**, 1162–1164.
- Mahrenholz, C.C. *et al.* (2011) Complex networks govern coiled-coil oligomerization – predicting and profiling by means of a machine learning approach. *Mol. Cell. Proteomics*, **10**, M110.004994.
- McDonnell, A.V. *et al.* (2006) Paircoil2: improved prediction of coiled coils from sequence. *Bioinformatics*, **22**, 356–358.
- Moutevelis, E. and Woolfson, D.N. (2009) A periodic table of coiled-coil protein structures. *J. Mol. Biol.*, **385**, 726–732.
- Parry, D.A. (1982) Coiled-coils in alpha-helix-containing proteins: analysis of the residue types within the heptad repeat and the use of these data in the prediction of coiled-coils in other proteins. *Biosci. Rep.*, **2**, 1017–1024.
- Peckham, M. and Knight, P.J. (2009) When a predicted coiled coil is really a single α -helix, in myosins and other proteins. *Soft Matter*, **5**, 2493–2503.
- Schiffer, M. and Edmundson, A.B. (1967) Use of helical wheels to represent the structures of proteins and to identify segments with helical potential. *Biophys. J.*, **7**, 121–135.
- Sivaramakrishnan, S. *et al.* (2008) Dynamic charge interactions create surprising rigidity in the ERK alpha-helical protein motif. *Proc. Natl Acad. Sci. USA*, **105**, 13356–13361.
- Süveges, D. *et al.* (2009) Charged single alpha-helix: a versatile protein structural motif. *Proteins*, **74**, 905–916.
- Trigg, J. *et al.* (2011) Multicoil2: predicting coiled coils and their oligomerization states from sequence in the twilight zone. *PLoS One*, **6**, e23519.
- Vincent, T.L. *et al.* (2013) LOGICOIL—multi-state prediction of coiled-coil oligomeric state. *Bioinformatics*, **29**, 69–76.

3.1.2. Distribution and evolution of stable single α -helices (SAH domains) in myosin motor proteins

Dominic Simm^{1,2}, Klas Hatje^{1,#} and Martin Kollmar^{1,*}

* Corresponding author

¹ Group Systems Biology of Motor Proteins, Department of NMR-based Structural Biology, Max-Planck-Institute for Biophysical Chemistry, Göttingen, Germany

² Theoretical Computer Science and Algorithmic Methods, Institute of Computer Science, Georg-August University, Göttingen, Germany

Roche Pharmaceutical Research and Early Development, Pharmaceutical Sciences, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd., Basel, Switzerland

PLoS ONE (2017) - Volume 12(4): e0174639

doi: 10.1371/journal.pone.0174639

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0174639>

Contribution

Conceptualization: Dominic Simm, Klas Hatje, Martin Kollmar.

Data curation: Dominic Simm.

Formal analysis: Dominic Simm, Martin Kollmar.

Investigation: Dominic Simm, Klas Hatje, Martin Kollmar.

Methodology: Dominic Simm.

Software: Dominic Simm.

Visualization: Dominic Simm, Martin Kollmar.

Writing – original draft: Dominic Simm, Martin Kollmar.

Writing – review & editing: Dominic Simm, Klas Hatje, Martin Kollmar.

RESEARCH ARTICLE

Distribution and evolution of stable single α -helices (SAH domains) in myosin motor proteins

Dominic Simm^{1,2}, Klas Hatje^{1*}, Martin Kollmar^{1*}

1 Group Systems Biology of Motor Proteins, Department of NMR-based Structural Biology, Max-Planck-Institute for Biophysical Chemistry, Göttingen, Germany, **2** Theoretical Computer Science and Algorithmic Methods, Institute of Computer Science, Georg-August-University Göttingen, Göttingen, Germany

✉ Current address: Roche Pharmaceutical Research and Early Development, Pharmaceutical Sciences, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd., Basel, Switzerland

* mako@nmr.mpibpc.mpg.de



Abstract

Stable single-alpha helices (SAHs) are versatile structural elements in many prokaryotic and eukaryotic proteins acting as semi-flexible linkers and constant force springs. This way SAH-domains function as part of the lever of many different myosins. Canonical myosin levers consist of one or several IQ-motifs to which light chains such as calmodulin bind. SAH-domains provide flexibility in length and stiffness to the myosin levers, and may be particularly suited for myosins working in crowded cellular environments. Although the function of the SAH-domains in human class-6 and class-10 myosins has well been characterised, the distribution of the SAH-domain in all myosin subfamilies and across the eukaryotic tree of life remained elusive. Here, we analysed the largest available myosin sequence dataset consisting of 7919 manually annotated myosin sequences from 938 species representing all major eukaryotic branches using the SAH-prediction algorithm of Waggawagga, a recently developed tool for the identification of SAH-domains. With this approach we identified SAH-domains in more than one third of the supposed 79 myosin subfamilies. Depending on the myosin class, the presence of SAH-domains can range from a few to almost all class members indicating complex patterns of independent and taxon-specific SAH-domain gain and loss.

OPEN ACCESS

Citation: Simm D, Hatje K, Kollmar M (2017) Distribution and evolution of stable single α -helices (SAH domains) in myosin motor proteins. PLoS ONE 12(4): e0174639. <https://doi.org/10.1371/journal.pone.0174639>

Editor: Maria Gasset, Consejo Superior de Investigaciones Científicas, SPAIN

Received: June 30, 2016

Accepted: March 13, 2017

Published: April 3, 2017

Copyright: © 2017 Simm et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: KH performed all work related to this manuscript when he was affiliated at the Max-Planck-Institute for Biophysical Chemistry. He is now working at F. Hoffmann-La Roche Ltd. (current address). Hoffmann-La Roche did not have any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Introduction

Helices, which are not buried within globular structures or coiled-coil helical dimers, usually need networks of charge interactions for stabilization in water [1–5]. In the late 1980th and early 1990th many studies have been performed using poly-alanine peptide models aiming to resolve the conditions for helix formation and stabilization. Different amino acids (mainly aspartic acid, glutamic acid, lysine and arginine, but in some cases also glutamine) were introduced into these peptides alone and in all possible combinations at varying distances. The corresponding peptides were synthesised, their α -helicity experimentally determined by, for

Competing interests: KH's current address does not alter our adherence to PLOS ONE policies on sharing data and materials.

example, circular dichroism, and stabilization energies were obtained by fitting models to the data. Although poly-alanine peptides adopt α -helical conformations when sparsely interrupted by lysine [6], arginine [7] and glutamine residues [8], helices are especially stabilized by charged interactions (salt bridges) between residues at $(i, i+3)$ and $(i, i+4)$ spacing [1–3] and hydrogen-bonding interactions between polar/charged residues at $(i, i+3)$ and $(i, i+4)$ [4,5]. Additional stability is obtained through networks of oppositely charged residues in $(i, i+3, i+6)$, $(i, i+3, i+7)$, $(i, i+4, i+7)$, or $(i, i+4, i+8)$ distances [9,10]. In addition to these poly-alanine based peptides, studies have been performed on peptides with complex amino acid distributions [11,12]. However, each study used different combinations of residues and non-physiological experimental conditions were applied (e.g. salt concentrations).

Studies on natural proteins known to contain stable single α -helices (SAHs) have shown that the respective sequence regions mainly consist of repeated patterns of negatively and positively charged residues [13–18]. In congruence with the results of the poly-alanine based analyses, repeats of four negatively and four positively charged residues form the most stable α -helices, while peptides with repeats of two residues do not show helical content [15,19]. Genome-wide searches for SAH-domains revealed their presence in bacteria, archaea, and eukaryotes, with the largest number of potential SAH-domains identified in mammals [17,20]. Predictions of SAH-domains in human proteins range from 0.25% to 0.4% [20,21], of which many belong to the cytoskeletal and motor proteins.

Myosins are a diverse protein family characterised by a large motor domain, which in most subfamilies contains an ATP-hydrolysis and an actin-binding site, and extended so-called tail domains located both N- and C-terminal to the motor domain [22]. Many myosins contain regions with one or multiple IQ-motifs C-terminal to the motor domain for binding calmodulin-family proteins that together act as a lever to transmit the power of the working stroke into displacement of the tail alongside the actin-filaments. Mammalian myosins from the class-6 and class-10 subfamilies have experimentally been shown to contain SAH-domains subsequent to the IQ-motif regions functioning as extended levers and constant force springs [14,17,23,24]. However, myosin tail architectures are very divergent even within subfamilies and a comprehensive analysis of SAH-domains in myosins based on deep taxonomic sampling across the supposed 79 subfamilies is still missing. Here, we analysed the largest available myosin dataset and identified putative SAH-domains in more than one third of all subfamilies. Depending on the myosin class, the presence of SAH-domains can range from a few to almost all class members indicating complex patterns of independent and taxon-specific SAH-domain gain and loss.

Results and discussion

Composition of the SAH-score

Because of the different combinations of residues and experimental conditions used in studies determining peptide stabilization energies, it is difficult to obtain and tabulate experimentally validated stabilization energies for all possible combinations of amino acids in all positions. We tried to develop a score reflecting the stability of a single α -helix compared to random coils and helices stabilized by interactions with other structural elements (e.g. within a protein structure or by binding to another protein via a coiled-coil motif). To this end, published stabilization energies [1–5,10] were compiled and set into relation to define stabilization values for all types of salt bridges and hydrogen-bonding interactions. We distinguish three types of stabilizations, weak, medium and strong, and classified all possible amino acid interactions accordingly (S1 Fig). In addition to binary interactions, an additional stabilization value is added for networks of at least three residues of oppositely charged residues in $(i, i+3, i+6)$, $(i, i$

+3, $i+7$), ($i, i+4, i+7$), or ($i, i+4, i+8$) distances. In contrast, networks of hydrophobic residues at distances ($i, i+3, i+7$) would provide a hydrophobic seam around the helix and might cause dimerisation by coiled-coil formation. Such hydrophobic networks are thought to be destabilizing and are accounted for by negative stabilization values. In addition to these values describing interactions, we assign alanines positive stabilization and glycines and prolines negative stabilization values, according to their helix promoting and helix braking characteristics, respectively. All stabilization values within a given sequence window are summed up and the sum is subsequently normalized against an idealised SAH-domain containing only strong interactions to compute an SAH-score for the central amino acid of the sequence window [25].

For comparison, we computed SAH-scores for four different window sizes, specifically 14, 21, 28, and 49 (about half a pitch of a two-stranded coiled-coil [26,27]). Small window sizes allow detection of relatively short SAH-domains (down to roughly four helix turns) but in these cases single or few residues can have large effects on the SAH-score both in terms of detecting false positives (coiled-coil regions often contain stretches with high percentages of charged residues) and missing true positives (one or few hydrophobic residues are well accommodated in stable single α -helices although they do not contribute to the SAH-score as defined above). With the largest window (49 amino acids, around 14 helical turns) the effect of single or few residues on the SAH-score is diminished but short SAH-domains might not be detected. These SAH-scores considerably fluctuate from residue to residue. Independent of score fluctuations, SAH-domains as structural entities are uninterrupted helices. However, similar to coiled-coil regions SAH-domains do not have distinct borders such as globular protein domains and might have any length. Therefore, specifying SAH-domain start and end positions as well as the transition point from non-SAH-domain to SAH-domain are a matter of definition.

Determining SAH-domains in myosins

We obtained 7919 manually annotated myosin sequences from 938 species from CyMoBase (Kollmar and Mühlhausen, submitted), of which 7675 sequences contain tail regions with varying degree of completeness (6744 tails are complete, in 531 tails short sequence regions are missing, 490 tails are fragmented). The minimum length of a SAH-domain was set to 14 amino acids (around four helical turns). Every residue within a SAH-domain should have a SAH-score above a given threshold (here: 0.25). Known coiled-coil regions from muscle myosin heavy chain proteins, tropomyosins, and keratins clearly have lower SAH-scores, while known SAH-domains such as those from class-6 and class-10 myosins, GCP60, M4K4, INCENP, and Caldesmon-1 [13,14,16,18] have considerably higher SAH-scores. Therefore, a threshold of 0.25 for the SAH-score seemed reasonable. We allowed 20 percent of the scores of a putative SAH-domain to be below the threshold to avoid multiple successive regions interrupted by just one or two residues. Such residues within SAH-domains with SAH-scores below the threshold are, however, rare. Only 247 (1.2%) of 20449 ['high' SAH-domain-score cut-off] / 580 (1.7%) of 35002 ['low' SAH-domain-score cut-off] residues (numbers based on computing SAH-scores with a 14 amino acid window) within proposed SAH-domains have SAH-scores below the threshold.

To obtain comparable scores for entire SAH-domains but to avoid bias by peak values, we define the *SAH-domain-score* as the highest average of the SAH-scores of any 14 neighbouring amino acids within each SAH-domain. The SAH-domain-score is dependent on the size of the sequence window for computing each amino acid's SAH-score (Fig 1A, S1 Table). The distribution of the SAH-domain-scores does not indicate a clear distinction between SAH-domains and non-SAH-regions. This is the result of the low sequence and structural complexity of stable single α -helices, which allows all types of possible combinations of SAH-score-contributing

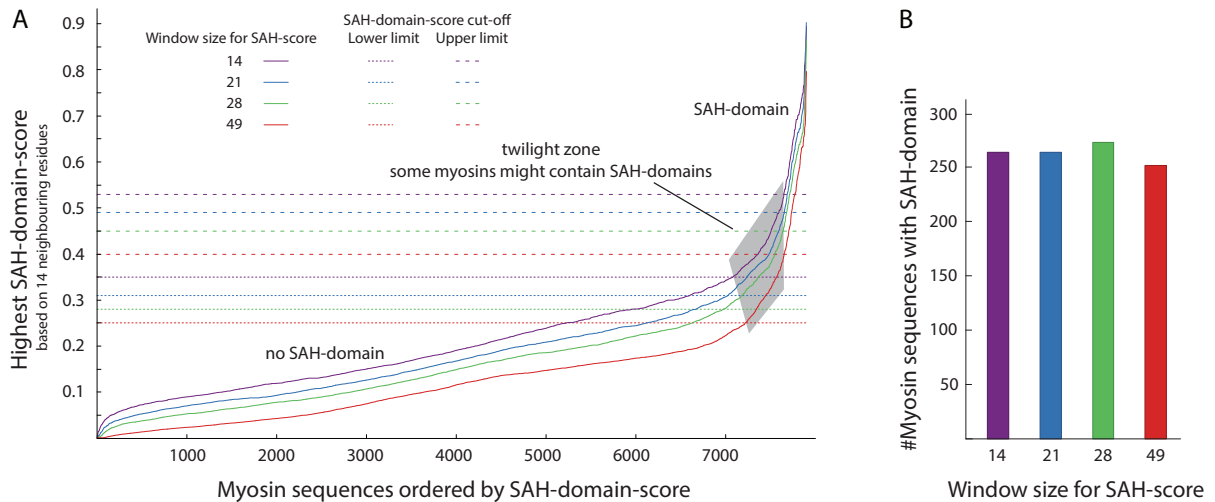


Fig 1. SAH-domains in myosins. A) For each myosin sequence, we determined the region with the highest SAH-domain-score, which is the average of the SAH-scores of 14 neighbouring amino acids. For comparison, we used SAH-scores computed for four different amino acid window sizes for SAH-domain-score determination. SAH-domain-score cut-offs were derived by sorting all sequences by SAH-domain-scores and determining the largest change in SAH-domain-score differences between two neighbouring sequences using kernel-regression (S2 Fig). B) Number of putative SAH-domains with minimum length of 14 amino acids in dependence of amino acid window sizes used to determine SAH-scores.

<https://doi.org/10.1371/journal.pone.0174639.g001>

amino acids at distances allowing charged and polar interactions. However, there is a clear region indicating unambiguous SAH-domains, and a large area indicating the absence of SAH-domains (Fig 1A). Sequences in the twilight zone might represent regions with high content of charged residues but nevertheless buried in tertiary and quaternary structures, or short single α -helices with atypically low numbers of charged and polar residues. It has been demonstrated that single amino acid mutations can cause small domains to switch between α -helix and β -sheet structures [28,29], and also cause coiled-coil segments to switch between different oligomeric states [30,31]. Accordingly, experimental evidence is needed to finally reveal the *in vivo* oligomeric state of the putative SAH-domains in the twilight zone. For further analyses, we use those SAH-domain-scores as cut-off, at which the changes in the increase of the score are highest (S2 Fig; 'high' SAH-domain-score cut-off). Applying these cut-offs, about 260 SAH-domains are detected for all window sizes of 14, 21, 28 and 49 amino acids (Fig 1B). The beginning of the twilight zone is defined by the first major change in the increase of the score (S2 Fig; 'low' SAH-domain-score cut-off).

Distribution of SAH-domains across myosin classes

SAH-domains were identified in 265 myosins (window size of 14 amino acids) from 25 myosin classes (32% of all classes). We also identified SAH-domains in 18 orphan myosins (Fig 2). The evaluation of the percentage of myosins with SAH-domain per class shows a tripartite distribution: A) A group of four classes, where most members contain SAH-domains, B) a group of about five classes with 15–40% of the myosins containing SAH-domains, and C) the remaining classes, of which only a small percentage or none of the class members contain SAH-domains (Fig 2). The classes with predominantly myosins with SAH-domains comprise the holozoan class-6, the opisthokont class-10, the stramenopiles class-38, and the amoebozoan class-45. The class-10 and the class-45 myosins are the only myosins where all members have SAH-domains (some class-10 SAH-domains have scores shortly below the high SAH-domain-score

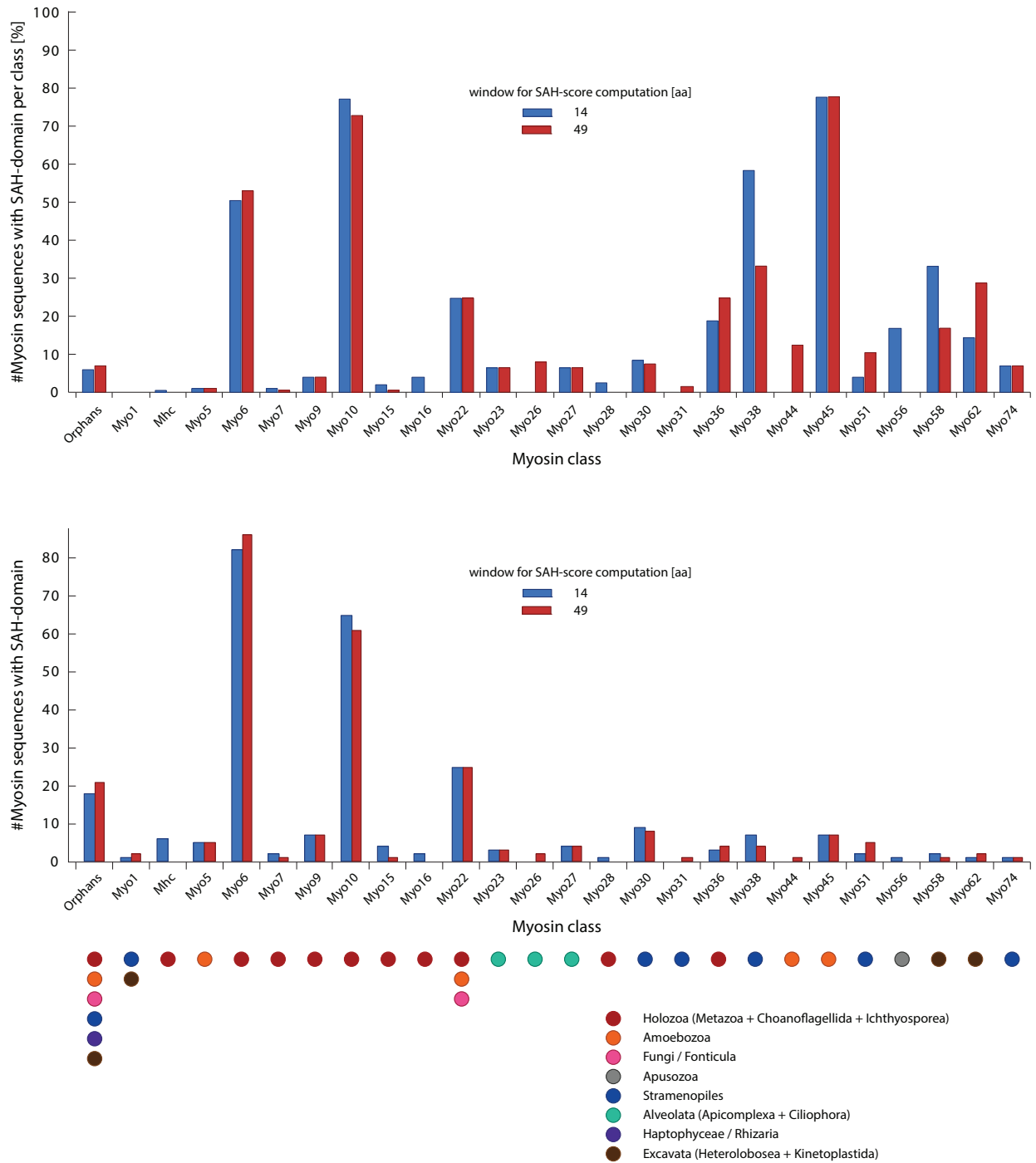


Fig 2. Distribution of SAH-domains with respect to myosin classes. The analysis of the percentage of myosins with SAH-domain per myosin class showed few classes, in which most or all myosins contain SAH-domains, while in most classes SAH-domains are present in only few, taxonomically restricted myosins. For comparison, the total number of myosins with SAH-domain is shown for each class. The taxonomic distribution of the myosins

with SAH-domains is indicated at the bottom for each class. Note, that the taxa represent the first occurrence of respective myosins with SAH-domain, which is not always identical to the first occurrence of the respective myosin class. Subsequently, the SAH-domains were independently lost in many subtaxa so that the respective myosin motor domain and SAH-domain combination is not present in every extant species. For comparison, the distribution of SAH-domains with respect to myosin classes using the low SAH-domain-score cut-off is shown in [S3 Fig](#).

<https://doi.org/10.1371/journal.pone.0174639.g002>

cut-off, and for some the respective regions comprising the SAH-domains are missing in the genome assemblies, [Fig 3](#)). Comparison of the SAH-domain predictions based on the 14 and 49 amino acid window sizes shows that most of the SAH-domains are long and detected with both window sizes (e.g. those of the class-6, class-10 and class-45 myosins), while in some classes such as the class-16, class-38, class-56, and class-58 the SAH-domains are particularly short ([Fig 2](#) and [S3 Fig](#)). In contrast to the results obtained with other algorithms [20], SAH-domains in homologous myosins of closely related species (e.g. all mammals) were consistently detected.

Gain and loss of SAH-domains in myosin classes

The current myosin data indicate that the last eukaryotic common ancestor only contained a class-1 myosin, of which members are present in almost all extant eukaryotes (the major exceptions are the plants and alveolates), and a second myosin of unknown class, which most probably had been the ancestor of all other classes [22]. The phylogenetic distribution of the classes with myosins containing SAH-domains implicates that the SAH-domains have been added to the tail domain architectures independently of each other. Within class-1 myosins, SAH-domains were detected in members of the Excavata and Stramenopiles taxa, also indicating independent gain. In the other, taxon-restricted classes, both independent gain in late-diverging branches and secondary loss were observed. Examples for independent gain are the SAH-domains in frog and fish Myo9B subtypes and the acorn worm *Saccoglossus kowalevskii* Myo9, examples for secondary SAH-domain loss events are the absence of SAH-domains in Platyhelminthes and many nematode class-6 myosins ([Fig 3](#) and [S4](#) and [S5 Figs](#)).

Characteristics of SAH-domains in myosins

The SAH-domains are on average 37–54 (low SAH-domain-score cut-off) / 62–81 (high SAH-domain-score cut-off) amino acids long, depending on the window size for SAH-score computation ([Fig 4A](#)). There is no preference for a certain SAH-domain length, and the number of myosins with SAH-domain lengths up to 80 amino acids is relatively constant. However, SAH-domains longer than 80 amino acids are rare. The longest SAH-domains of about 200 residues have been identified in cryptosporidian class-26 myosins. The lengths of the corresponding uninterrupted regular α -helices would be about 300 Å. Thus, the myosin SAH-domains have on average two-third of the mean-length of the average SAH-domains of all proteins, which are supposed to be about 75 residues long (according to protein sequences available in SwissProt and UniProt in 2012 [20]), although the distribution of number of sequences versus SAH-domain lengths is very similar [20].

The proposed SAH-domains contain only amino acids with high helical propensity [32], supporting that we identified only regions forming uninterrupted α -helices ([Fig 4B](#)). About 76% of the residues comprise charged amino acids with equal numbers of residues with positive and negative character, which is similar to the distribution of charged residues in a set of 47 proteins with supposed SAH-domains derived from SwissProt [16]. The percentage of charged residues is more than twice as high as observed in coiled-coil domains [21]. Compared to the amino acid distribution in α -helices in general [32] and coiled-coil regions in particular [16], hydrophobic residues are highly under-represented (only about 9.6% in the detected

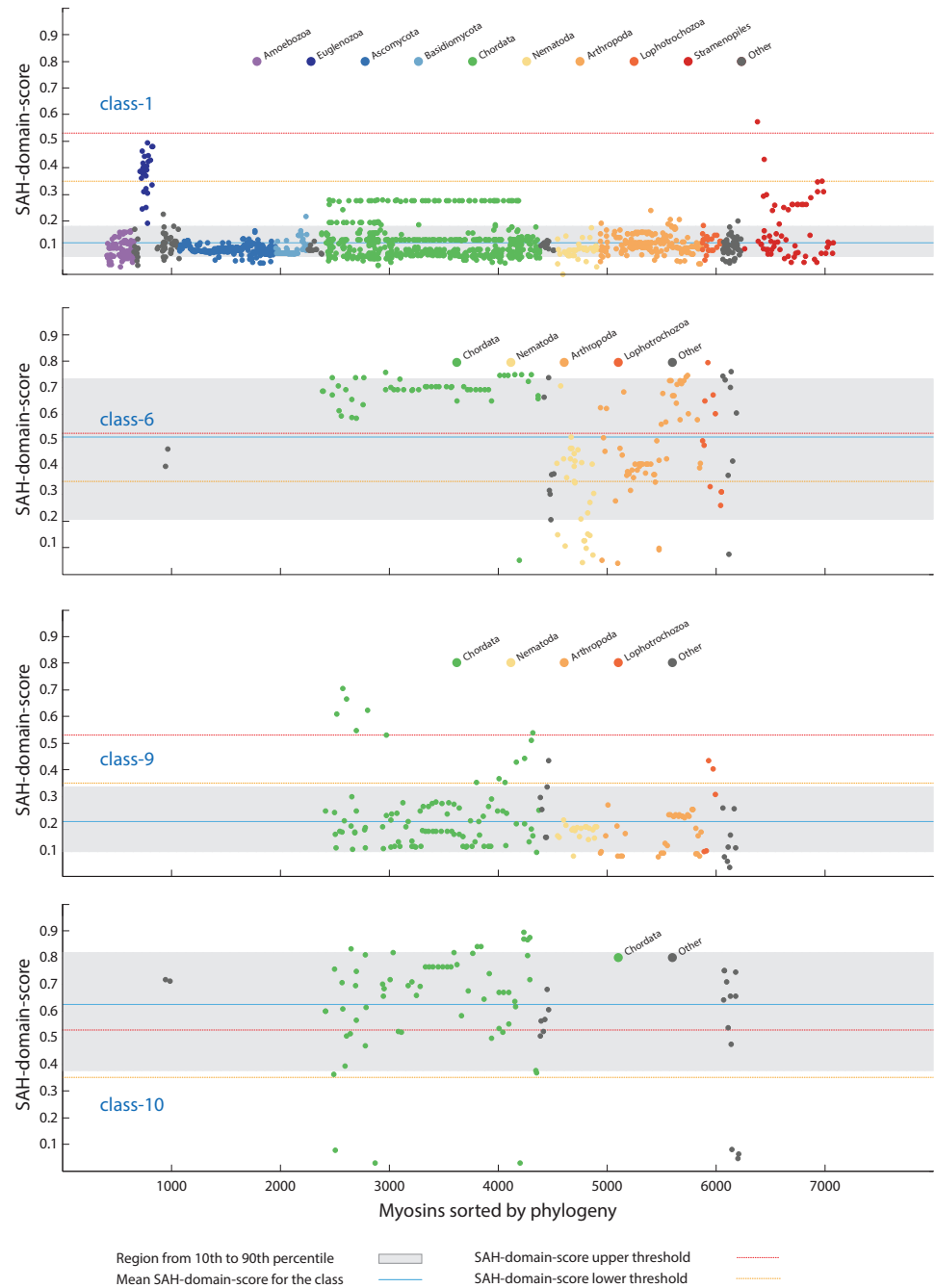


Fig 3. Phylogenetic distribution of the myosins. All myosins were sorted by taxonomy, and the highest SAH-domain-score for each myosin was plotted per class. Here, four examples based on the 14 amino acid window size for computing the SAH-score are shown representing classes with independent SAH-domain gain in late-branching taxa (class-1 and class-9), a class with multiple independent SAH-domain loss (class-6), and a class with all myosins containing SAH-domains (class-10; the sequences with SAH-domain-scores below 0.1 miss the putative SAH-domain regions because of genome and transcriptome assembly gaps). Major taxa are indicated by colour for better

orientation. The phylogenetic distribution of all SAH-domains sorted by classes is shown in [S4](#) and [S5](#) Figs (window sizes of 14 and 49 amino acids, respectively, for computing the SAH-score).

<https://doi.org/10.1371/journal.pone.0174639.g003>

SAH-domains). These findings strongly support that the predicted SAH-domains are not able to oligomerize but instead comprise real stable single α -helices. A few hydrophobic residues might well be accommodated within SAH-domains through hydrophobic interactions with the hydrophobic parts of the arginine and lysine side chains [33].

Functional implications

The importance of the SAH-domains as functional modules providing stiffness and modulating length of the lever in myosins has already been discussed in detail [16,20,21,23,34–36]. Our data show that the presence of SAH-domains in myosins is not myosin class-dependent but the result of a complex history of taxon- and species-specific gain and loss events. The only exceptions are the class-10 and class-45 myosins, of which all homologs identified to date contain SAH-domains. In all other cases, the myosins of interest have to be inspected manually. The data available in [S1 Table](#) and presented in [S4](#) and [S5](#) Figs could help in the evaluation. The SAH-domain-score cut-offs used throughout this study are rather conservative. Thus, although there might be few exceptions the SAH-domains with scores above these cut-offs can be regarded as veritable SAH-domains. This does not inevitably mean that the respective myosins are monomers. Dimerisation might happen through additional coiled-coil regions or other dimer-promoting domains. At least, coiled-coil predictions overlapping the SAH-domains as presented in the [S1 Table](#) can be regarded as mis-predictions caused by the intrinsic problems of all available coiled-coil prediction software to distinguish coiled-coils from SAH-domains [25].

Examples for SAH-domains and twilight cases

In the following, we present examples for veritable SAH-domains and twilight-zone cases ([Fig 5](#)). The *Acanthamoeba* class-22 myosins contain the most ideal SAH-domains with dense networks of potential charged interactions only sporadically interrupted by small (e.g. alanines) and hydrophobic (e.g. isoleucines and leucines) residues ([Fig 5A](#)). A few hydrophobic residues can well be accommodated in SAH-domains through hydrophobic interaction with the hydrophobic parts of lysine and arginine residues [33].

The Kinetoplastid class-1 myosins contain unique tail architectures consisting of, from N- to C-terminus, an IQ motif, a WW, a MyTH1, a FYVE, and a subsequent short SAH-domain. This short SAH-domain is around 28 residues long ([Fig 5B](#)) and its function is most probably the spatial separation of the FYVE domain from a short C-terminal domain or protein interaction motif.

A short region in the tails of the mammalian Mhc14 class-2 myosins (MYH14, non-muscle myosin 2C) represents a twilight case ([Fig 5C](#)). This region is not present in other vertebrate Mhc14 orthologs and has evolved in the ancient mammalian Mhc14 ([S6 Fig](#)). The SAH-domain-score of this region is slightly below our conservative cut-off. Such a putative short SAH-domain within an extended coiled-coil domain is supposed to have a different function than causing monomerization. Rather, its function could be to open up and interrupt the coiled-coil region allowing local unwinding or bending of the α -helices similar to the unstable coiled-coil region of the N-terminal S2 (subfragment-2) domains of the skeletal muscle myosins [37]. This hypothesis is supported by the presence of many residues with strong helix-breaking propensity such as glycines and serines directly C-terminal to the putative SAH-

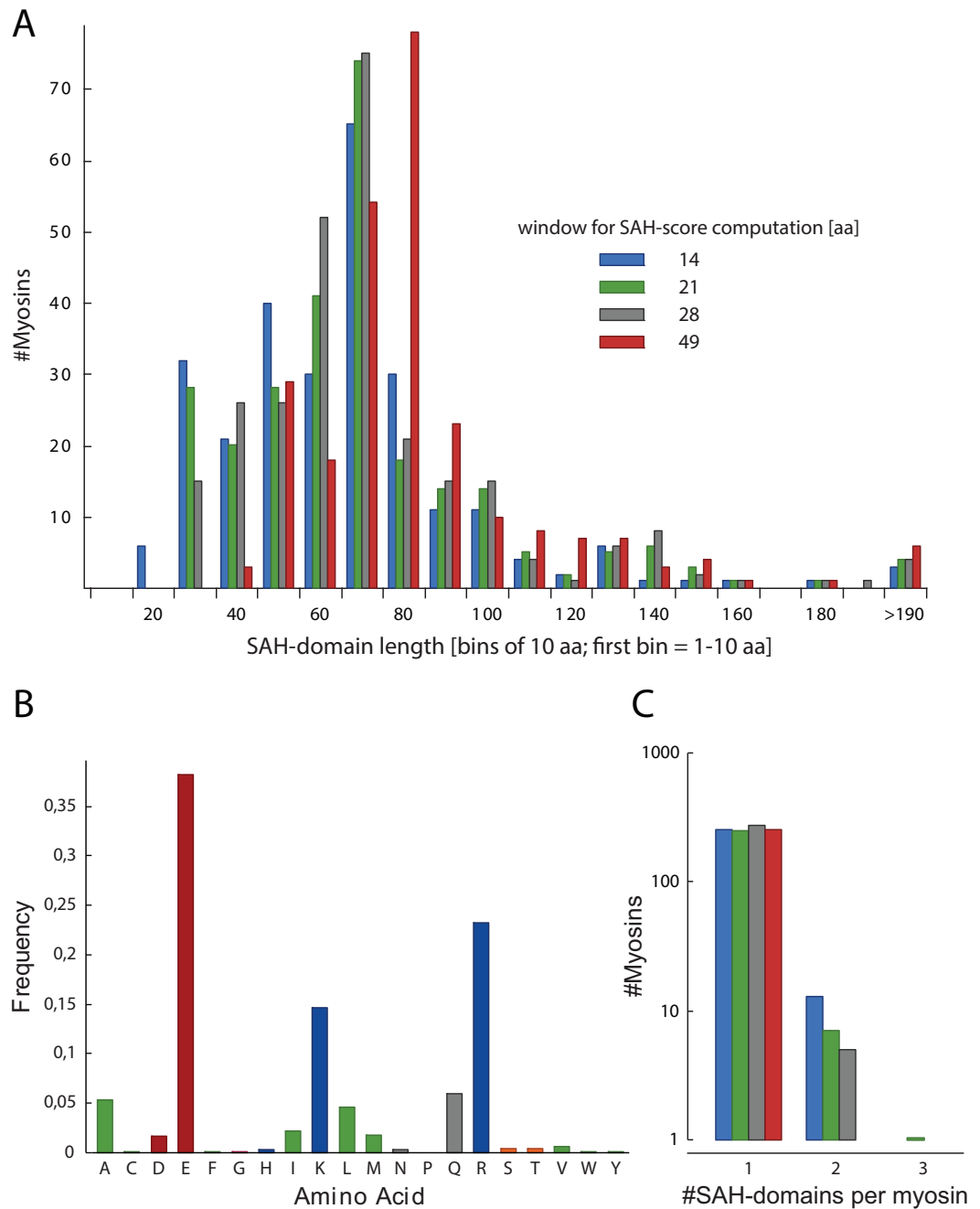


Fig 4. Characteristics of SAH-domains in myosins. A) Analysis of the length of SAH-domains. For better presentation, all SAH-domains with lengths within bins of ten amino acids were summed. B) Amino acid distribution counted over all 265 SAH-domains determined with the 14 amino acid window size SAH-score. C) Analysis of the number of SAH-domains per myosin.

<https://doi.org/10.1371/journal.pone.0174639.g004>

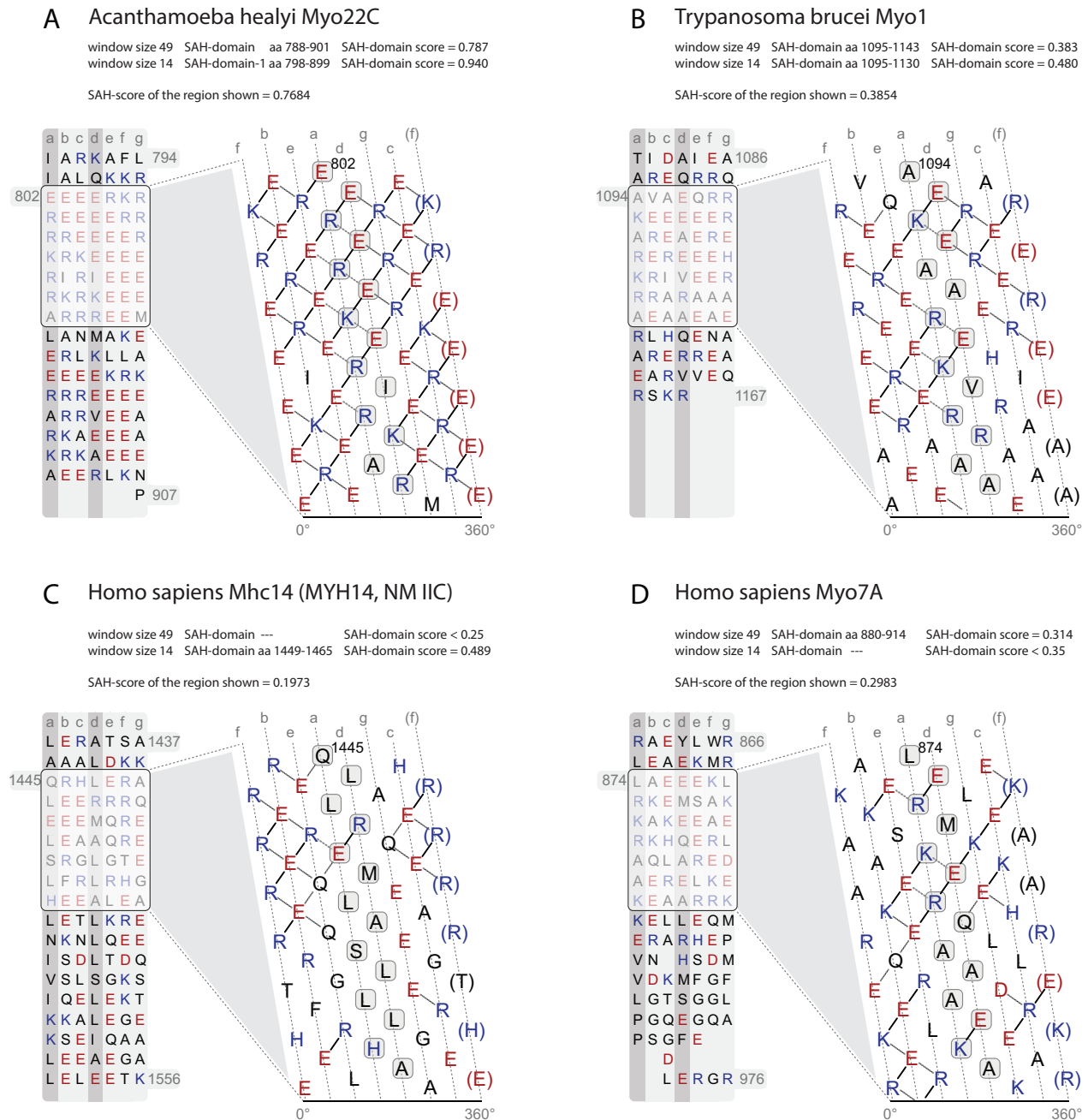


Fig 5. Examples of SAH-domains, and regions probably not containing SAH-domains. A) SAH-domain of *Acanthamoeba healyi* Myo22C, which is among the SAH-domains with the highest SAH-domain-scores. B) Example for a short SAH-domain, which is only detected with the 14, 21, and 28 amino acid window sizes for computing SAH-scores. This short SAH-domain is located C-terminal to the class-1 myosin-specific MyTH1 domain and therefore not part of the lever. Its putative function is to spatially separate the MyTH1 domain from another small domain or protein interaction motif of unknown function at the C-terminus. C) Example for a short SAH-domain with a SAH-domain-score in the twilight zone. This highly charged region is unique to mammalian Mhc14 class-2 myosins (MYH14, non-muscle myosin 2C; [S6 Fig](#)) and interrupts the long coiled-coil filament-forming region of the muscle myosins. The subsequent C-terminal sequence contains a large number of amino acids with high helix-breaking propensity such as glycines and serines. The putative

function of the short SAH-domain and the following glycine-rich region might be to open up the coiled-coil. D) Example of a region rich in charged amino acids but with an SAH-domain-score in the range of non-SAH-domains.

<https://doi.org/10.1371/journal.pone.0174639.g005>

domain (S6 Fig). The corresponding regions in the other mammalian muscle myosin homologs contain a conserved tryptophan at a “d” position (S6 Fig). This is one of the two conserved tryptophans in the muscle myosin rod region that were shown to be appreciably exposed to solvent and to be located in the least stable parts of the fibrous region [38–40]. Introducing a highly charged region in mammalian Mhc14 myosins could be an alternative solution to destabilize this part of the coiled-coil region.

The monomeric nature of the *Drosophila* Myo7A myosin *in vitro* [41,42] and the accumulation of charged residues in the region subsequent to the IQ motifs suggested this region also representing a SAH-domain [36]. However, we only observed veritable SAH-domains in cnidarian and placozoan class-7 myosins, and the short putative SAH-domains in *Drosophila* Myo7B myosins represent twilight cases (S1 Table, S4 and S5 Figs). The highest SAH-scores determined in all other class-7 myosins including *Drosophila* and human Myo7A are far below the upper SAH-domain-score cut-offs. The regions C-terminal to the IQ-motif region in class-7 myosins clearly do not contain hydrophobic residues ordered in a heptad pattern needed for coiled-coil formation (Fig 5D). However, these regions also do not have the extended and dense networks of charged amino acids characteristic of SAH-domains. These regions might therefore represent divergent types of SAH-domains not described and detected with current methods, or be part of a larger domain such as a helical bundle.

Myosins with unusual lever architectures

In class-38 myosins, we identified alternating IQ-motif regions and SAH-domains (Fig 6A). Multiple SAH-domains have also been determined in many other myosins (Fig 4C, S1 Table) supporting that IQ-motifs and SAH-domains are structural building blocks, which can be used in any order and number to adjust the lever length of myosins. Usually, myosins have at least a single IQ-motif C-terminal to the motor domains acting as canonical lever. Here, we detected a few myosins not containing a single IQ-motif but having SAH-domains directly following the motor domain, for example in the Stramenopiles *Aplanochytrium* Myo31A myosin (Fig 6B) and in an orphan myosin from the cryptophyte *Guillardia theta*.

Methods

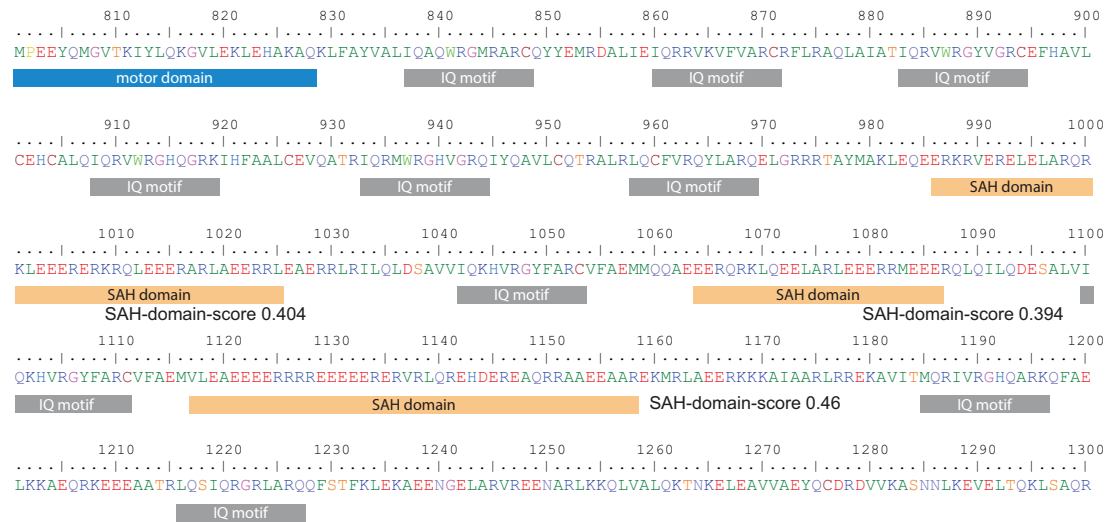
Determination of SAH-domains in myosins

7919 Myosin sequences were obtained from CyMoBase (www.cymobase.org, [43]), release 19th December 2016, and sorted by class and taxonomy. SAH-domain prediction was performed by querying the Waggawagga API [25]. The prediction provides an SAH-score for each amino acid within a certain window. Here, we used windows with lengths of 14, 21, 28, and 49 amino acids, and computed SAH-scores for each amino acid in each myosin sequence. Although there is no clear distinction between SAH-domains and multimeric α -helical structures (already a point mutation might lead to a switch between these two states), tests with known cases have shown that SAH-regions have scores above 0.25 [25]. We used this rather conservative value to distinguish amino acids within (SAH-score > 0.25) and outside (SAH-score < 0.25) SAH-regions.

Determination of SAH-domain-scores

SAH-domains are stretches of continuous amino acids with SAH-scores above 0.25. The criterion for a SAH-domain was to have at least 14 amino acids (around four helical turns). To not

A *Phytophthora ramorum* Myo38



B *Aplanochytrium kerguelense* Myo31A

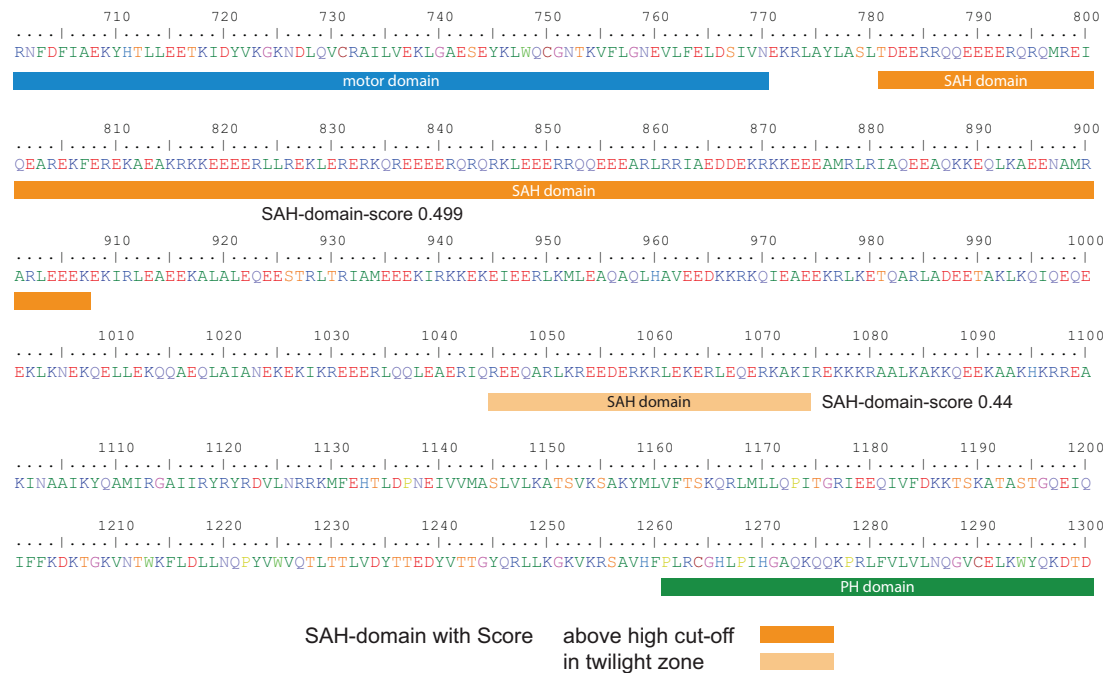


Fig 6. Examples of myosins with unusual lever architectures. A) Tail region of the *Phytophthora ramorum* Myo38 myosin containing alternating IQ-motifs and SAH-domains. For better orientation only the core motifs of the IQ-regions are indicated. For the SAH-domains, the predictions based on the 14 amino acid window were used (S1 Table). B) Tail region of the *Aplanochytrium kerguelense* Myo31A representing the neck region between the motor domain and the C-terminal PH domain. This myosin does not contain any IQ-motif.

<https://doi.org/10.1371/journal.pone.0174639.g006>

obtain regions with multiple predicted SAH-domains interrupted by only one or two residues with SAH-scores below the threshold, we allowed for 20 percent of the SAH-scores of a SAH-domain to be below the threshold. These lower scores were, however, not allowed to be at the start or at the end of the SAH-domain. To provide a score for each SAH-domain but not have peak values dominate the analysis, we defined the SAH-domain-score as the highest average value of the SAH-scores of 14 neighbouring amino acids. The SAH-domain-score depends on the window size used to compute the individual SAH-scores.

For SAH-domain-score based analyses we determined SAH-domain-score cut-offs by sorting the SAH-domain-scores in increasing order, estimating a kernel-regression function to smooth noisy scores, and then determining the cut-off by choosing the SAH-domain-score of the maximum slope change position (the maximum change in the difference between two subsequent SAH-domain-scores).

Supporting information

S1 Table. Lists of predicted SAH-domains. One subtable for each of the four window sizes for computing SAH-scores. In all tables, a SAH length cut-off of 14 amino acids has been applied. One additional subtable for window size of 14 with an SAH length cut-off of ten residues was added for comparison. All subtables contain the name of the myosin sequence, the SAH-domain-score, the SAH-domain sequence, the sequence start and end position of the SAH-domain, the SAH-scores for all residues within the SAH-domains, and the full taxonomy of the species according to NCBI.

(XLS)

S1 Fig. Composition of the SAH-score. Helical net view of a SAH-domain region. Some strong and weak interactions are indicated by arrows. The table lists the score of each interaction taken for computing the SAH-score.

(PDF)

S2 Fig. Kernel-regression plots to determine SAH-domain-score cut-offs. A separate plot is shown for each window size, with the corresponding SAH-domain-score cut-off.

(PDF)

S3 Fig. Distribution of SAH-domains with respect to myosin classes. In contrast to Fig 2, this plot is based on the low SAH-domain-score cut-off for accepting SAH-domains (see S2 Fig). The total number of myosins with SAH-domain is shown for each class. The taxonomic distribution of the myosins with SAH-domains is indicated at the bottom for each class. Note, that the taxons represent the first occurrence of respective myosins with putative SAH-domain, which is not always identical to the first occurrence of the respective myosin class. Subsequently, the SAH-domains were independently lost in many subtaxa so that the respective myosin motor domain and SAH-domain combination is not present in every extant species.

(PDF)

S4 Fig. Phylogenetic distribution of the myosins. This figure is similar to Fig 3 but contains the results of all myosin classes. Shortly, all myosins were sorted by taxonomy, and the highest SAH-domain-score for each myosin plotted class by class. The 14 amino acid window size was taken for computing the SAH-score. Major taxa are indicated by colour for better orientation.

(PDF)

S5 Fig. Phylogenetic distribution of the myosins. This figure is similar to S4 Fig except that the 49 amino acid window size was taken for computing the SAH-score. Major taxa are

indicated by colour for better orientation.
(PDF)

S6 Fig. Example for a putative short SAH-domain in class-2 myosins. Alignment of part of the tail region of all human class-2 myosins focusing on the region comprising a predicted short SAH-domain in HsMhc14. For comparison, further mammalian and some other vertebrate Mhc14 homologs are shown indicating that the respective region has evolved after separation of the mammals.
(PDF)

Acknowledgments

We would like to thank Prof. Christian Griesinger and Prof. Stephan Waack for their continuous generous support.

Author Contributions

Conceptualization: DS KH MK.

Data curation: DS.

Formal analysis: DS MK.

Investigation: DS KH MK.

Methodology: DS.

Software: DS.

Supervision: MK.

Validation: MK.

Visualization: DS MK.

Writing – original draft: DS MK.

Writing – review & editing: DS KH MK.

References

1. Marqusee S, Baldwin RL. Helix stabilization by Glu...Lys+ salt bridges in short peptides of de novo design. *Proc Natl Acad Sci U S A*. 1987; 84: 8898–8902. PMID: [3122208](#)
2. Huyghues-Despointes BM, Scholtz JM, Baldwin RL. Helical peptides with three pairs of Asp-Arg and Glu-Arg residues in different orientations and spacings. *Protein Sci Publ Protein Soc*. 1993; 2: 80–85.
3. Wang E, Wang CL. (i, i + 4) Ion pairs stabilize helical peptides derived from smooth muscle caldesmon. *Arch Biochem Biophys*. 1996; 329: 156–162. <https://doi.org/10.1006/abbi.1996.0204> PMID: [8638947](#)
4. Scholtz JM, Qian H, Robbins VH, Baldwin RL. The energetics of ion-pair and hydrogen-bonding interactions in a helical peptide. *Biochemistry (Mosc)*. 1993; 32: 9668–9676.
5. Smith JS, Scholtz JM. Energetics of polar side-chain interactions in helical peptides: salt effects on ion pairs and hydrogen bonds. *Biochemistry (Mosc)*. 1998; 37: 33–40.
6. Marqusee S, Robbins VH, Baldwin RL. Unusually stable helix formation in short alanine-based peptides. *Proc Natl Acad Sci U S A*. 1989; 86: 5286–5290. PMID: [2748584](#)
7. Lockhart DJ, Kim PS. Electrostatic screening of charge and dipole interactions with the helix backbone. *Science*. 1993; 260: 198–202. PMID: [8469972](#)
8. Shalongo W, Dugad L, Stellwagen E. Distribution of Helicity within the Model Peptide Acetyl(AAQAA)3amide. *J Am Chem Soc*. 1994; 116: 8288–8293.

9. Spek EJ, Bui AH, Lu M, Kallenbach NR. Surface salt bridges stabilize the GCN4 leucine zipper. *Protein Sci Publ Protein Soc.* 1998; 7: 2431–2437.
10. Olson CA, Spek EJ, Shi Z, Vologodskii A, Kallenbach NR. Cooperative helix stabilization by complex Arg-Glu salt bridges. *Proteins.* 2001; 44: 123–132. PMID: [11391775](#)
11. Lyu PC, Marky LA, Kallenbach NR. The role of ion pairs in α -helix stability: two new designed helical peptides. *J Am Chem Soc.* 1989; 111: 2733–2734.
12. Lyu PC, Liff MI, Marky LA, Kallenbach NR. Side chain contributions to the stability of α -helical structure in peptides. *Science.* 1990; 250: 669–673. PMID: [2237416](#)
13. Wang CL, Chalovich JM, Graceffa P, Lu RC, Mabuchi K, Stafford WF. A long helix from the central region of smooth muscle caldesmon. *J Biol Chem.* 1991; 266: 13958–13963. PMID: [1856225](#)
14. Knight PJ, Thirumurugan K, Xu Y, Wang F, Kalverda AP, Stafford WF, et al. The Predicted Coiled-coil Domain of Myosin 10 Forms a Novel Elongated Domain That Lengthens the Head. *J Biol Chem.* 2005; 280: 34702–34708. <https://doi.org/10.1074/jbc.M504887200> PMID: [16030012](#)
15. Sivaramakrishnan S, Spink BJ, Sim AYL, Doniach S, Spudich JA. Dynamic charge interactions create surprising rigidity in the ER/K α -helical protein motif. *Proc Natl Acad Sci U S A.* 2008; 105: 13356–13361. <https://doi.org/10.1073/pnas.0806256105> PMID: [18768817](#)
16. Süveges D, Gáspári Z, Tóth G, Nyitray L. Charged single α -helix: a versatile protein structural motif. *Proteins.* 2009; 74: 905–916. <https://doi.org/10.1002/prot.22183> PMID: [18712826](#)
17. Spink BJ, Sivaramakrishnan S, Lipfert J, Doniach S, Spudich JA. Long single α -helical tail domains bridge the gap between structure and function of myosin VI. *Nat Struct Mol Biol.* 2008; 15: 591–597. <https://doi.org/10.1038/nsmb.1429> PMID: [18511944](#)
18. Samejima K, Platani M, Wolny M, Ogawa H, Vargiu G, Knight PJ, et al. The Inner Centromere Protein (INCENP) Coil Is a Single α -Helix (SAH) Domain That Binds Directly to Microtubules and Is Important for Chromosome Passenger Complex (CPC) Localization and Function in Mitosis. *J Biol Chem.* 2015; 290: 21460–21472. <https://doi.org/10.1074/jbc.M115.645317> PMID: [26175154](#)
19. Lyu PC, Gans PJ, Kallenbach NR. Energetic contribution of solvent-exposed ion pairs to α -helix structure. *J Mol Biol.* 1992; 223: 343–350. PMID: [1731079](#)
20. Gáspári Z, Süveges D, Perczel A, Nyitray L, Tóth G. Charged single α -helices in proteomes revealed by a consensus prediction approach. *Biochim Biophys Acta.* 2012; 1824: 637–646. <https://doi.org/10.1016/j.bbapap.2012.01.012> PMID: [22310480](#)
21. Peckham M, Knight PJ. When a predicted coiled coil is really a single α -helix, in myosins and other proteins. *Soft Matter.* 2009; 5: 2493–2503.
22. Odrionitz F, Kollmar M. Drawing the tree of eukaryotic life based on the analysis of 2,269 manually annotated myosins from 328 species. *Genome Biol.* 2007; 8: R196. <https://doi.org/10.1186/gb-2007-8-9-r196> PMID: [17877792](#)
23. Baboolal TG, Sakamoto T, Forgacs E, White HD, Jackson SM, Takagi Y, et al. The SAH domain extends the functional length of the myosin lever. *Proc Natl Acad Sci.* 2009; 106: 22193–22198. <https://doi.org/10.1073/pnas.0909851106> PMID: [20018767](#)
24. Wolny M, Batchelor M, Knight PJ, Paci E, Dougan L, Peckham M. Stable single α -helices are constant force springs in proteins. *J Biol Chem.* 2014; 289: 27825–27835. <https://doi.org/10.1074/jbc.M114.585679> PMID: [25122759](#)
25. Simm D, Hatje K, Kollmar M. Waggawagga: comparative visualization of coiled-coil predictions and detection of stable single α -helices (SAH domains). *Bioinforma Oxf Engl.* 2015; 31: 767–769.
26. Phillips GN. What is the pitch of the α -helical coiled coil? *Proteins Struct Funct Bioinforma.* 1992; 14: 425–429.
27. Seo J, Cohen C. Pitch diversity in α -helical coiled coils. *Proteins.* 1993; 15: 223–234. <https://doi.org/10.1002/prot.340150302> PMID: [8456094](#)
28. Alexander PA, He Y, Chen Y, Orban J, Bryan PN. The design and characterization of two proteins with 88% sequence identity but different structure and function. *Proc Natl Acad Sci U S A.* 2007; 104: 11963–11968. <https://doi.org/10.1073/pnas.0700922104> PMID: [17609385](#)
29. He Y, Chen Y, Alexander PA, Bryan PN, Orban J. Mutational tipping points for switching protein folds and functions. *Struct Lond Engl* 1993. 2012; 20: 283–291.
30. Kammerer RA, Kostrewa D, Progiás P, Honnappa S, Avila D, Lustig A, et al. A conserved trimerization motif controls the topology of short coiled coils. *Proc Natl Acad Sci U S A.* 2005; 102: 13891–13896. <https://doi.org/10.1073/pnas.0502390102> PMID: [16172398](#)
31. Steinmetz MO, Jelesarov I, Matousek WM, Honnappa S, Jahnke W, Missimer JH, et al. Molecular basis of coiled-coil formation. *Proc Natl Acad Sci.* 2007; 104: 7062–7067. <https://doi.org/10.1073/pnas.0700321104> PMID: [17438295](#)

32. Nick Pace C, Martin Scholtz J. A Helix Propensity Scale Based on Experimental Studies of Peptides and Proteins. *Biophys J*. 1998; 75: 422–427. PMID: [9649402](#)
33. Andrew CD, Penel S, Jones GR, Doig AJ. Stabilizing nonpolar/polar side-chain interactions in the alpha-helix. *Proteins*. 2001; 45: 449–455. PMID: [11746692](#)
34. Peckham M. Coiled coils and SAH domains in cytoskeletal molecular motors. *Biochem Soc Trans*. 2011; 39: 1142–1148. <https://doi.org/10.1042/BST0391142> PMID: [21936779](#)
35. Swanson CJ, Sivaramakrishnan S. Harnessing the unique structural properties of isolated α -helices. *J Biol Chem*. 2014; 289: 25460–25467. <https://doi.org/10.1074/jbc.R114.583906> PMID: [25059657](#)
36. Batchelor M, Wolny M, Dougan L, Paci E, Knight PJ, Peckham M. Myosin tails and single α -helical domains. *Biochem Soc Trans*. 2015; 43: 58–63. <https://doi.org/10.1042/BST20140302> PMID: [25619246](#)
37. Li Y, Brown JH, Reshetnikova L, Blazsek A, Farkas L, Nyitray L, et al. Visualization of an unstable coiled coil from the scallop myosin rod. *Nature*. 2003; 424: 341–345. <https://doi.org/10.1038/nature01801> PMID: [12867988](#)
38. King L, Lehrer SS. Thermal unfolding of myosin rod and light meromyosin: circular dichroism and tryptophan fluorescence studies. *Biochemistry (Mosc)*. 1989; 28: 3498–3502.
39. Chang YC, Ludescher RD. Tryptophan photophysics in rabbit skeletal myosin rod. *Biophys Chem*. 1994; 49: 113–126. PMID: [8155813](#)
40. Zhou X, Maéda Y, Mabuchi K, Lehrer SS. Unfolding domains and tryptophan accessibility of a 59 kDa coiled-coil light meromyosin. *J Mol Biol*. 1998; 276: 829–838. <https://doi.org/10.1006/jmbi.1997.1571> PMID: [9500922](#)
41. Umeki N, Jung HS, Watanabe S, Sakai T, Li X, Ikebe R, et al. The tail binds to the head-neck domain, inhibiting ATPase activity of myosin VIIA. *Proc Natl Acad Sci U S A*. 2009; 106: 8483–8488. <https://doi.org/10.1073/pnas.0812930106> PMID: [19423668](#)
42. Yang Y, Baboolal TG, Siththanandan V, Chen M, Walker ML, Knight PJ, et al. A FERM domain autoregulates *Drosophila* myosin 7a activity. *Proc Natl Acad Sci U S A*. 2009; 106: 4189–4194. <https://doi.org/10.1073/pnas.0808682106> PMID: [19255446](#)
43. Odronitz F, Kollmar M. Pfarao: a web application for protein family analysis customized for cytoskeletal and motor proteins (CyMoBase). *BMC Genomics*. 2006; 7: 300. <https://doi.org/10.1186/1471-2164-7-300> PMID: [17134497](#)

3.1.3. Waggawagga-CLI: A command-line tool for predicting stable single α -helices (SAH-domains), and the SAH-domain distribution across eukaryotes

Dominic Simm ^{1,2} and Martin Kollmar ^{1,*}

* Corresponding author

¹ Group Systems Biology of Motor Proteins, Department of NMR-based Structural Biology, Max-Planck-Institute for Biophysical Chemistry, Göttingen, Germany

² Theoretical Computer Science and Algorithmic Methods, Institute of Computer Science, Georg-August University, Göttingen, Germany

PLoS ONE (2018) - Volume 13(2): e0191924

doi: 10.1371/journal.pone.0191924

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0191924>

Contribution

Data curation: Dominic Simm, Martin Kollmar.

Formal analysis: Dominic Simm.

Investigation: Dominic Simm, Martin Kollmar.

Methodology: Dominic Simm.

Software: Dominic Simm.

Validation: Dominic Simm, Martin Kollmar.

Visualization: Dominic Simm, Martin Kollmar.

Writing – original draft: Dominic Simm, Martin Kollmar.

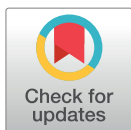
RESEARCH ARTICLE

Waggawagga-CLI: A command-line tool for predicting stable single α -helices (SAH-domains), and the SAH-domain distribution across eukaryotes

Dominic Simm^{1,2}, Martin Kollmar^{1*}

1 Group Systems Biology of Motor Proteins, Department of NMR-based Structural Biology, Max-Planck-Institute for Biophysical Chemistry, Göttingen, Germany, **2** Theoretical Computer Science and Algorithmic Methods, Institute of Computer Science, Georg-August-University Göttingen, Göttingen, Germany

* mako@nmr.mpibpc.mpg.de



Abstract

Stable single-alpha helices (SAH-domains) function as rigid connectors and constant force springs between structural domains, and can provide contact surfaces for protein-protein and protein-RNA interactions. SAH-domains mainly consist of charged amino acids and are monomeric and stable in polar solutions, characteristics which distinguish them from coiled-coil domains and intrinsically disordered regions. Although the number of reported SAH-domains is steadily increasing, genome-wide analyses of SAH-domains in eukaryotic genomes are still missing. Here, we present Waggawagga-CLI, a command-line tool for predicting and analysing SAH-domains in protein sequence datasets. Using Waggawagga-CLI we predicted SAH-domains in 24 datasets from eukaryotes across the tree of life. SAH-domains were predicted in 0.5 to 3.5% of the protein-coding content per species. SAH-domains are particularly present in longer proteins supporting their function as structural building block in multi-domain proteins. In human, SAH-domains are mainly used as alternative building blocks not being present in all transcripts of a gene. Gene ontology analysis showed that yeast proteins with SAH-domains are particularly enriched in macromolecular complex subunit organization, cellular component biogenesis and RNA metabolic processes, and that they have a strong nuclear and ribonucleoprotein complex localization and function in ribosome and nucleic acid binding. Human proteins with SAH-domains have roles in all types of RNA processing and cytoskeleton organization, and are predicted to function in RNA binding, protein binding involved in cell and cell-cell adhesion, and cytoskeletal protein binding. Waggawagga-CLI allows the user to adjust the stabilizing and destabilizing contribution of amino acid interactions in $i,i+3$ and $i,i+4$ spacings, and provides extensive flexibility for user-designed analyses.

OPEN ACCESS

Citation: Simm D, Kollmar M (2018) Waggawagga-CLI: A command-line tool for predicting stable single α -helices (SAH-domains), and the SAH-domain distribution across eukaryotes. PLoS ONE 13(2): e0191924. <https://doi.org/10.1371/journal.pone.0191924>

Editor: Alexandre G. de Brevern, UMR-S1134, INSERM, Université Paris Diderot, INTS, FRANCE

Received: September 27, 2017

Accepted: January 12, 2018

Published: February 14, 2018

Copyright: © 2018 Simm, Kollmar. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Waggawagga-CLI is available for download from <http://waggawagga.motorprotein.de>. The software and all results from data analysis as presented in this study (SAH-domain predictions and GO analyses) are also available from figshare (doi: [10.6084/m9.figshare.5435947](https://doi.org/10.6084/m9.figshare.5435947)).

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Stable single α -helices (SAHs) are extended helices that are not buried within globular structures or coiled-coil helical dimers [1–8]. Their most common function is to serve as rigid connectors or constant force springs between structural domains [1–5,7,9,10], but they also provide contact surfaces for protein-protein and protein-RNA interactions [7,8]. The latter function has been found for the inner centromere protein INCENP and for many regions of spliceosomal proteins in various complexes formed during the pre-mRNA splicing cycle. Independent of any tertiary interactions, SAH-domains are stable and monomeric in polar solvents. These features distinguish SAH-domains from other proteins that fold into α -helices only in the presence of binding partners such as stathmin which is an intrinsically disordered protein lacking any stable fold in the absence of binding partners, but forms an extended α -helix when binding to tubulin dimers [11,12].

SAH-domains are extremely rich in glutamate (E), lysine (K) and arginine (R) [4,6,13,14], which have been shown to stabilize poly-alanine peptides by charge interactions along the helix [15–20]. Although aspartate (D) can also form stabilizing interactions with K/R [21,22], aspartates occur less often than isoleucine, leucine, methionine, alanine and glutamine in predicted, highly likely SAH-domains [13,14]. Especially repeated patterns of four E followed by four K/R seem to stabilize α -helices, while peptides with repeats of two residues do not show helical content [3,17,23]. The specific $(E_4(R/K)_4)_n$ pattern has therefore been termed *ER/K motif* [6] but this is sometimes mixed up with the term *EK/R α -helix* (e.g. [8]), which had been introduced as alternative term to SAH [3]. Another term introduced in the field is *charged single α -helix* (CSAH) [4] but SAH-domains must not have an overall net charge. Protein regions with stable single α -helices must also not be uninterrupted helices but might include short breaks leading to multiple successive SAHs behaving as a worm-like chain [10]. To exclude misunderstandings because of term usage and to include all special cases, we will refer to these protein regions as SAH-domains from now on.

The experimental identification of SAH-domains first in caldesmon and then L9 ribosomal protein and class-10 myosin has fostered the idea that SAH-domains might be common structural motifs and be present in many other proteins. In first analyses with BLAST using the EK/R motif [3] and the SAH-domain of class-10 myosin [13], 123 distinct proteins in 137 archaea and eukaryotes and 36 human proteins, respectively, have been identified. In a more exhaustive search against UniProt using two newly developed software tools and requiring a minimum of 40 amino acids for an SAH-domain to be detected, SAH-domains were identified in all three kingdoms of life and it was estimated that their abundance is less than 0.2% of all proteins of a species [24]. This, however, was a very conservative approach and less stringent criteria might inevitably yield more SAH-domains. Interestingly, the most SAH-domains were found in the human proteome (165 proteins).

Waggawagga was developed as a web application to visually compare coiled-coil predictions from various tools using helical-wheel and helical-net representations [25]. These representations also allow distinguishing between predicted coiled-coils and SAH-domains. A score summarizing stabilizing and destabilizing interactions was introduced to discriminate SAH-domains from non-SAH-domains. Although most SAH-domains and non-SAH-domains can clearly be discriminated, a large-scale analysis of more than 7900 myosin sequences across all eukaryotes revealed a twilight-zone between the two extremes [14]. Sequences with SAH-domain-scores within this twilight-zone likely need experimental confirmation to demonstrate their SAH or non-SAH appearance. Here, we present a new version of Waggawagga, termed Waggawagga-CLI, intended for the command-line usage to investigate small- and large-scale protein sequence data.

Results and discussion

Waggawagga-CLI is the command-line version of the web application Waggawagga. It was developed to provide a tool for large-scale SAH-domain prediction and analysis and should, in principle, be able to manage protein sequence datasets of any size. Waggawagga-CLI has not been optimized for speed, but an SAH-domain prediction is likely to be performed only once for each dataset. The SAH-domain predictions are stored in a mobile database thus allowing repeated analyses in case the default parameters need to be adjusted. In contrast to the web application which predicts SAH-domains based on the heptad-repeat assignments of coiled-coil prediction tools, Waggawagga-CLI assumes each protein sequence to be a continuous α -helix and predicts SAH-domains in these.

The Waggawagga-CLI version contains all required secondary software libraries and a lightweight database, SQLite, and therefore does not require any further software installations by the user. The SQLite database is used for storing analysed sequence data during runtime, and can be queried by the advanced user afterwards in multiple ways. Waggawagga-CLI predicts SAH-domains for each of the sequences in the file and subsequently filters the hits by two cut-offs, the minimum SAH-score for each amino acid to be included in an SAH-domain (default: 0.25) and the minimum SAH-domain-score, which depends on the length of the SAH-domain window (default windows: 14, 21, 28, and 49 amino acids). Accordingly, the window size sets the minimum length of an SAH-domain to be detected. The default SAH-domain-scores are based on the results of a comprehensive analysis of more than 7900 myosin sequences across all eukaryotes [14] and range from 0.25 (window 49 aa) to 0.35 (window 14 aa). The results of a Waggawagga-CLI run are provided in text format (and optional gnuplot SVG images) for each sequence containing a predicted SAH-domain, and in summary tables, for each SAH-domain window separately. The summary tables comprise an SAH amino acid distribution analysis, a list of the predicted SAHs by length of SAH-domains, and a detailed list of all SAH-domains with amino acid sequences and SAH-scores for each amino acid.

To demonstrate the application of Waggawagga-CLI on large-scale datasets such as protein sequence datasets generated by whole-genome annotations, we selected protein annotation datasets from species across the eukaryotic tree of life (Table 1). The datasets were obtained from Ensembl Genomes release 87 [26]. The overall runtime per dataset ranged from a few hours to seven days depending on dataset size. The average runtime for single sequences ranged from 4.6 to 23.3 seconds. Across all datasets, the average runtime is 8.3 seconds per sequence which corresponds to 252 aa per second.

SAH-domain distribution in eukaryotes

SAH-domain validation depends on the length of the sequence window for supposed SAH-domains [14]. Short SAH-domains are not (or considerably less) detected using a large window (e.g. 49 amino acids), and long but less characteristic SAH-domains are often below the cut-off when using a short window (e.g. 14 amino acids). Therefore, by default Waggawagga-CLI determines SAH-domains with four windows, 14, 21, 28, and 49 amino acids. The numbers of detected SAH-domains in the analysed eukaryotic genomes are summarized in Table 2. The lowest total numbers of SAH-domains were found in *Cyanidioschyzon merolae* (28 SAH-domains, 21 aa window), and *Schizosaccharomyces pombe* (34 SAH-domains, 21 aa window), but these species also have the lowest gene numbers (Table 1). In contrast, *Plasmodium falciparum* with only a slightly higher number of genes contains six times more SAH-domains (Table 2). With respect to the percentage of SAH-domains found per dataset, and the percentage of sequences containing SAH-domains per dataset, these species represent the lower (0.5%) and upper (3.5%) limits of the range of SAH-domains per species (Fig 1). The total

Table 1. Number of sequences and runtime per protein dataset. All datasets were downloaded from Ensembl Genomes. According to the specifications at Ensembl, datasets specified by “all” represent the super-set of all translations resulting from Ensembl known or novel gene predictions, while datasets specified by “ab initio” include translations resulting from *ab initio* gene prediction software tools. Such *ab initio* predictions are based solely on the genomic sequence and not any other experimental evidence, and, therefore, not all predictions represent biologically real proteins. The sequences listed as *Seqs valid* represent the part of all sequences which meet the necessary conditions for processing (e.g. minimum length of sequence).

| Organism | Seqs total | Seqs valid | # AA | CPU hrs. | Time/Seq [sec] |
|--|------------|------------|-----------|----------|----------------|
| <i>Arabidopsis thaliana</i> [ab initio] | 20579 | 20517 | 42948060 | 46,5 | 8,16 |
| <i>Arabidopsis thaliana</i> [all] | 48321 | 47952 | 83278520 | 82,3 | 6,17 |
| <i>Caenorhabditis elegans</i> | 31574 | 31191 | 58037112 | 60,0 | 6,93 |
| <i>Chlamydomonas reinhardtii</i> | 14489 | 14473 | 26291352 | 25,9 | 6,43 |
| <i>Cyanidioschyzon merolae</i> | 4998 | 4963 | 10046308 | 9,7 | 7,00 |
| <i>Danio rerio</i> [ab initio] | 36087 | 35675 | 71517964 | 75,7 | 7,64 |
| <i>Danio rerio</i> [all] | 45336 | 44759 | 91282168 | 93,6 | 7,53 |
| <i>Dictyostelium discoideum</i> | 13267 | 13025 | 28052304 | 29,4 | 8,13 |
| <i>Drosophila melanogaster</i> [ab initio] | 36155 | 36056 | 66438088 | 75,2 | 7,50 |
| <i>Drosophila melanogaster</i> [all] | 30362 | 30194 | 80128800 | 110,1 | 13,13 |
| <i>Gallus gallus</i> [ab initio] | 50996 | 49017 | 85316640 | 88,5 | 6,50 |
| <i>Gallus gallus</i> [all] | 30252 | 30196 | 61283944 | 69,6 | 8,30 |
| <i>Giardia lamblia</i> | 7364 | 6730 | 12906580 | 43,5 | 23,28 |
| <i>Homo sapiens</i> [ab initio] | 50890 | 50200 | 84319768 | 92,8 | 6,66 |
| <i>Homo sapiens</i> [all] | 102915 | 97110 | 153405460 | 181,7 | 6,74 |
| <i>Leishmania major</i> | 8308 | 8307 | 20934752 | 21,2 | 9,18 |
| <i>Mus musculus</i> [ab initio] | 57111 | 56142 | 82405740 | 86,2 | 5,53 |
| <i>Mus musculus</i> [all] | 61440 | 59075 | 105206872 | 118,8 | 7,24 |
| <i>Oryza sativa</i> [ab initio] | 63510 | 63219 | 124444320 | 126,5 | 7,20 |
| <i>Oryza sativa</i> [all] | 42132 | 41596 | 55154072 | 52,8 | 4,57 |
| <i>Plasmodium falciparum</i> | 5352 | 5350 | 16345344 | 17,3 | 11,67 |
| <i>Saccharomyces cerevisiae</i> | 6692 | 6573 | 12023708 | 11,6 | 6,34 |
| <i>Schizosaccharomyces pombe</i> | 5146 | 5126 | 9548000 | 9,2 | 6,49 |
| <i>Tetrahymena thermophila</i> | 24725 | 24266 | 64078484 | 67,8 | 10,06 |

<https://doi.org/10.1371/journal.pone.0191924.t001>

number of SAH-domains and the number of sequences with SAH-domains are very similar for each dataset indicating that most sequences contain a single SAH-domain. While the datasets of the unicellular species contain one transcript per gene, the *ab initio* and *all* datasets also contain alternative transcripts. These transcripts contain identical, overlapping and independent SAH-domains (see section below). As a rough estimate, the total number of unique SAH-domains per genome is the total number of SAH-domains per datasets divided by the average number of transcripts per gene. In general, we identified less SAH-domains in the *ab initio* datasets than in the *all* datasets (Table 2), except for *Oryza sativa* and *Mus musculus*, where the total numbers of SAH-domains in the *all* and *ab initio* datasets are likely strongly under- and overestimated, respectively (Fig 1). Compared to the only other available genome-wide analysis of SAH-domains [24] we find two to six times more SAH-domains per genome. Given that about 1.5% of all genes of a species contain an SAH-domain, the SAH-domain is not a rare but a widely distributed and used building block for proteins.

The number of protein sequences with multiple SAH-domains linearly depends on the total number of sequences with SAH-domains (Fig 2). Thus, species-independent about 15% of the sequences with an SAH-domain contain at least one further SAH-domain. Only flowering plants, *Cyanidioschyzon merolae*, and *Saccharomyces cerevisiae* have considerably less genes with multiple SAH-domains.

Table 2. Number of predicted SAH-domains per protein dataset. SAH-domains were filtered by score for windows of 14, 21, 28, and 49 amino acids. Higher numbers for SAH-domains per dataset than sequences with SAH-domains per dataset indicate that some sequences contain multiple independent SAH-domains.

| Organism | 14 | | 21 | | 28 | | 49 | |
|--|------|------|------|------|------|------|------|------|
| | #SAH | #Seq | #SAH | #Seq | #SAH | #Seq | #SAH | #Seq |
| <i>Arabidopsis thaliana</i> [ab initio] | 151 | 138 | 185 | 173 | 132 | 123 | 86 | 81 |
| <i>Arabidopsis thaliana</i> [all] | 306 | 270 | 384 | 352 | 272 | 244 | 181 | 165 |
| <i>Caenorhabditis elegans</i> | 633 | 506 | 698 | 558 | 562 | 460 | 407 | 325 |
| <i>Chlamydomonas reinhardtii</i> | 119 | 100 | 133 | 113 | 95 | 83 | 63 | 55 |
| <i>Cyanidioschyzon merolae</i> | 22 | 21 | 28 | 27 | 9 | 9 | 5 | 5 |
| <i>Danio rerio</i> [ab initio] | 751 | 479 | 770 | 544 | 537 | 392 | 377 | 279 |
| <i>Danio rerio</i> [all] | 877 | 747 | 963 | 821 | 687 | 604 | 474 | 415 |
| <i>Dictyostelium discoideum</i> | 408 | 318 | 437 | 339 | 331 | 273 | 227 | 197 |
| <i>Drosophila melanogaster</i> [ab initio] | 383 | 330 | 432 | 360 | 289 | 249 | 168 | 148 |
| <i>Drosophila melanogaster</i> [all] | 620 | 546 | 694 | 592 | 456 | 398 | 302 | 266 |
| <i>Gallus gallus</i> [ab initio] | 546 | 462 | 619 | 515 | 456 | 391 | 305 | 280 |
| <i>Gallus gallus</i> [all] | 549 | 482 | 639 | 557 | 435 | 383 | 263 | 243 |
| <i>Giardia lamblia</i> | 36 | 30 | 47 | 39 | 27 | 23 | 17 | 16 |
| <i>Homo sapiens</i> [ab initio] | 887 | 694 | 962 | 767 | 663 | 562 | 441 | 379 |
| <i>Homo sapiens</i> [all] | 1262 | 1044 | 1412 | 1181 | 969 | 829 | 621 | 539 |
| <i>Leishmania major</i> | 70 | 57 | 93 | 71 | 56 | 41 | 42 | 29 |
| <i>Mus musculus</i> [ab initio] | 3403 | 2931 | 3591 | 3108 | 2780 | 2522 | 1847 | 1722 |
| <i>Mus musculus</i> [all] | 844 | 734 | 960 | 804 | 655 | 589 | 450 | 399 |
| <i>Oryza sativa</i> [ab initio] | 487 | 475 | 672 | 655 | 423 | 417 | 387 | 383 |
| <i>Oryza sativa</i> [all] | 161 | 156 | 184 | 180 | 126 | 123 | 77 | 75 |
| <i>Plasmodium falciparum</i> | 167 | 142 | 217 | 183 | 148 | 130 | 130 | 112 |
| <i>Saccharomyces cerevisiae</i> | 49 | 46 | 57 | 55 | 34 | 34 | 21 | 21 |
| <i>Schizosaccharomyces pombe</i> | 31 | 26 | 34 | 30 | 19 | 17 | 11 | 11 |
| <i>Tetrahymena thermophila</i> | 352 | 306 | 406 | 350 | 266 | 240 | 174 | 158 |

<https://doi.org/10.1371/journal.pone.0191924.t002>

SAH-domains in human alternative transcripts and evolution

SAH-domains are structural entities in proteins, and it seems likely that extended, combined, and altered SAH-domains might be obtained as result of alternative splicing. More simply, SAH-domains might either be present or absent in protein variants of a gene. A previous analysis of SAH-domains across human and mouse transcripts identified nine and seven cases, respectively, where the SAH-domains are either present or absent in alternative transcripts [24]. For a single human gene, *AFDN* (protein: afadin), transcripts resulting in SAH-domains of different length were found. Because the human genome annotation is likely the most complete comprising extensive alternative transcripts we used the human dataset “all” to analyse the role of SAH-domains as structural building block. We distinguish three cases of SAH-domains resulting from a single gene (presence/absence, including, overlapping; Fig 3), and all can happen at the same time if a gene is spliced in more than two alternative transcripts, or is spliced in two alternative transcripts and encodes multiple SAH-domains.

The human dataset “all” contains 97110 transcripts (Table 1) which are derived from 22622 genes. We identified 1262 SAH-domains (14 aa window, Table 2), of which 447 are unique with respect to genes. Of the 22622 genes, 5577 code for a single and 17045 code for multiple transcripts. 31 (0.55%) of the single and 265 (1.55%) of the multiple transcript genes encode 51 and 396 unique SAH-domains, respectively. 77 unique SAH-domains are present in all transcripts of the multiple transcript genes while 319 (80.6%) unique SAH-domains are absent in at least one of the transcripts (Fig 3). 116 unique SAH-domains are present in only a single

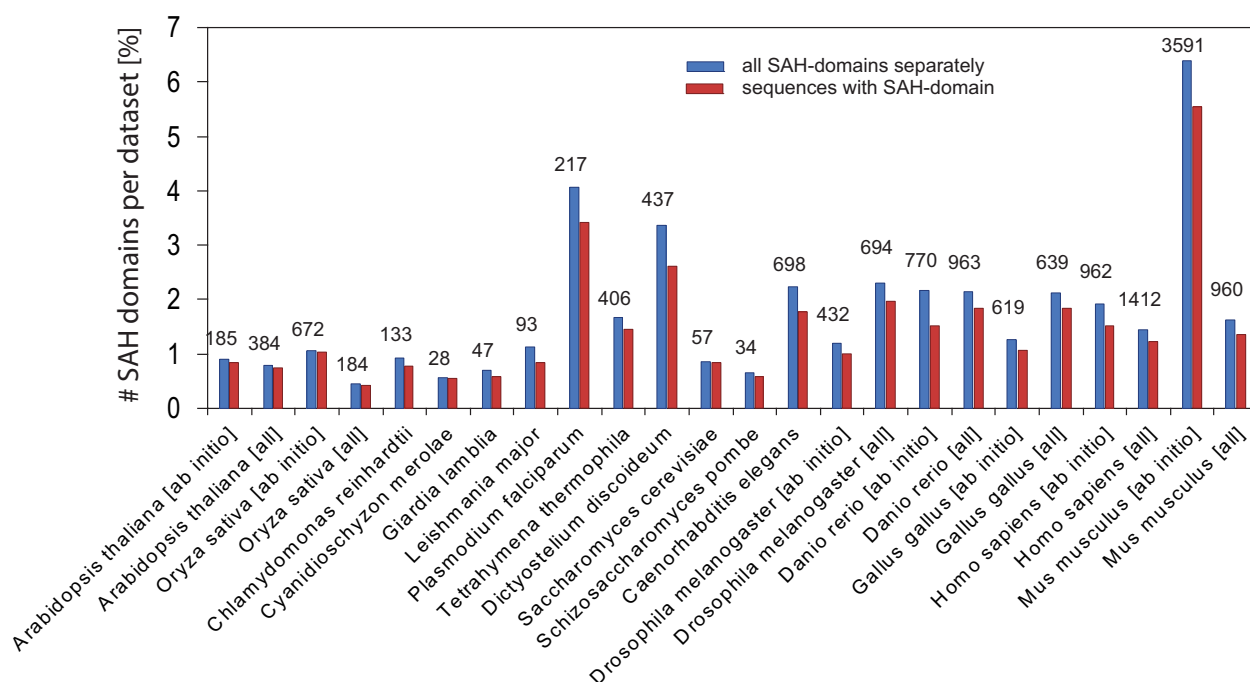


Fig 1. SAH-domains in eukaryotic genomes. The plot presents the percentage of SAH-domains with respect to total numbers of processed sequences per datasets. The blue bars represent all SAH-domains, while the red bars represent all sequences containing at least a single SAH-domain. The numbers on top of the blue bars denote the total numbers of predicted SAHs (see also Table 2). The window size for identifying SAHs was 21 amino acids. For an explanation of the difference between “all” and “ab initio” datasets see Table 1.

<https://doi.org/10.1371/journal.pone.0191924.g001>

from up to 24 transcripts. We identified only eight cases where unique SAH-domains are completely part of larger SAH-domains in other transcripts (case *including*). 28 unique SAH-domains overlap with unique SAH-domains of other transcripts of the same gene (case *overlapping*; minimum number of overlapping amino acids: 5).

The presence of an SAH-domain region in all transcripts of a gene indicates that the respective region is an essential structural entity in all resulting proteins. Our analysis shows that SAH-domains are indispensable in transcripts of only 19.4% of the genes. In the majority of the cases, SAH-domains are differentially included building blocks and add to the diversity of protein isoforms. Modulation of the SAH-domain lengths (cases *including* and *overlapping*) happens but is currently rare.

If SAH-domains represent structural building blocks similar to any other structural domain they should appear in transcripts of orthologous and paralogous proteins. If SAH-domains are part of exon shuffling processes they might appear in unrelated proteins. Because of the low amino acid and structural complexity in SAH-domains few mutations could turn these motifs into intrinsically disordered regions (which still might fold into α -helices upon interaction with binding partners) or α -helices that aggregate or even form coiled-coil structures. Thus, sequence homology based methods likely do not provide any specific relations. Instead, we searched for identical sequences of unique SAH-domains across all genes. Of the 447 human SAH-domains, 383 are unique with respect to the human genome. 30 SAH-domains are present in identical sequence in two to seven different genes. Most of these belong to gene paralogs such as multiple *golgin A6* family and *golgin A8* family genes, *tropomyosin 3* and *tropomyosin 4*,

and the calcium channel subunits *CACNA1H* and *CACNA1I*. In addition to these identical SAH-domains we identified 68 cases where a unique SAH-domain from one protein is part of a longer SAH-domain in another protein (*case including*).

Characteristics of SAH-domains

SAH-domains are regions with low sequence homology and low amino acid diversity, and their comparison is therefore restricted to some basic metrics. In all species analysed SAH-

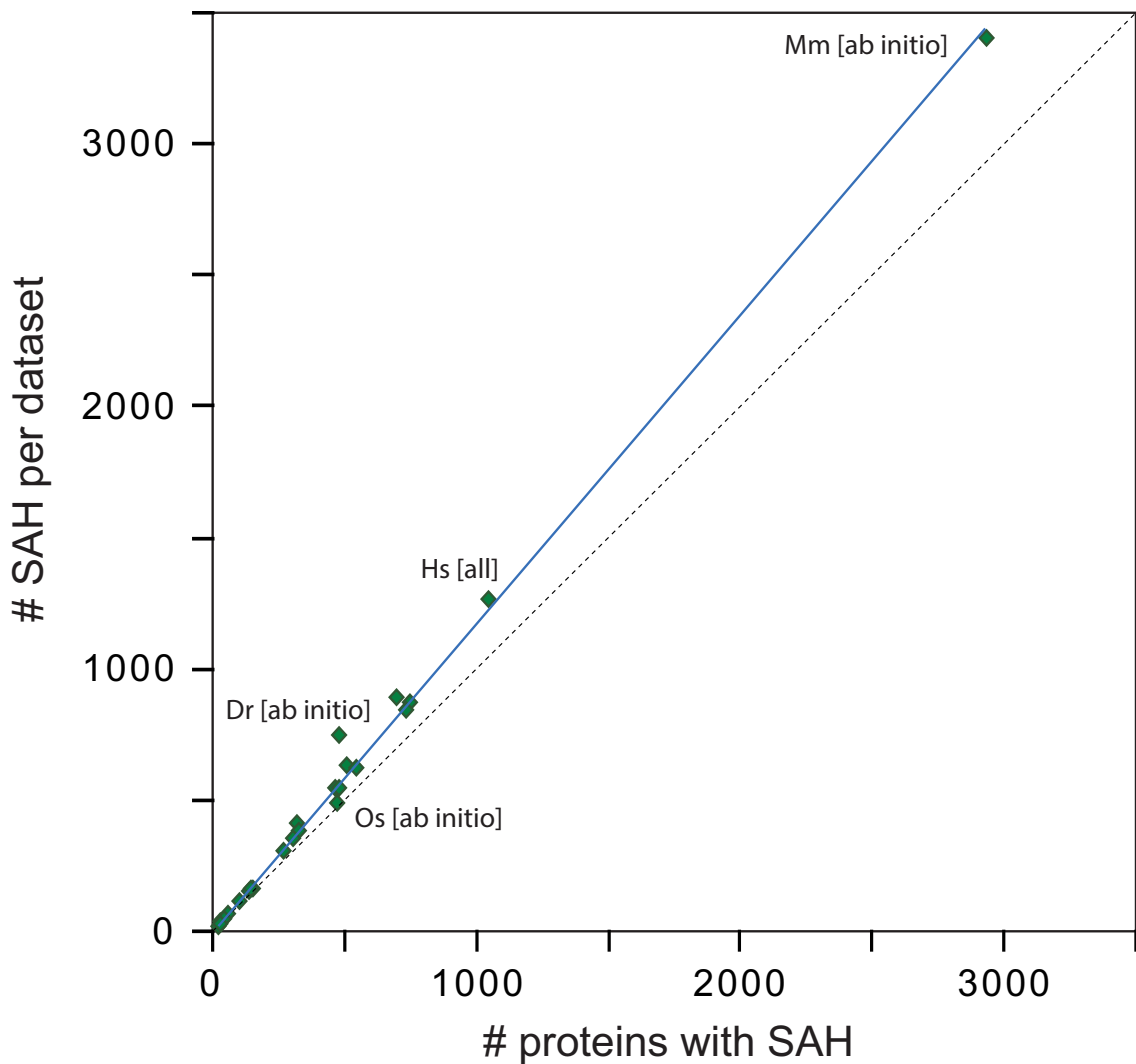


Fig 2. SAH-domains in eukaryotic genomes. The plot contrasts the total number of SAH-domains per dataset with the total number of sequences containing at least a single SAH-domain. Each diamond represents a dataset. The number of sequences with multiple SAH-domains is not a characteristic of certain species but depends on the total number of sequences with SAH-domains per species. The more sequences contain SAH-domains, the more sequences with multiple SAH-domains will be found. For orientation, labels were given for likely the best annotated dataset, human [all], the most extreme case mouse [ab initio], and the two datasets with the largest deviation from the line, danio [ab initio] and rice [ab initio]. The datasets with SAH-domains identified with a window size of 14 amino acids were taken. Abbreviations: Dr, *Danio rerio*; Hs, *Homo sapiens*; Mm, *Mus musculus*; Os, *Oryza sativa*.

<https://doi.org/10.1371/journal.pone.0191924.g002>

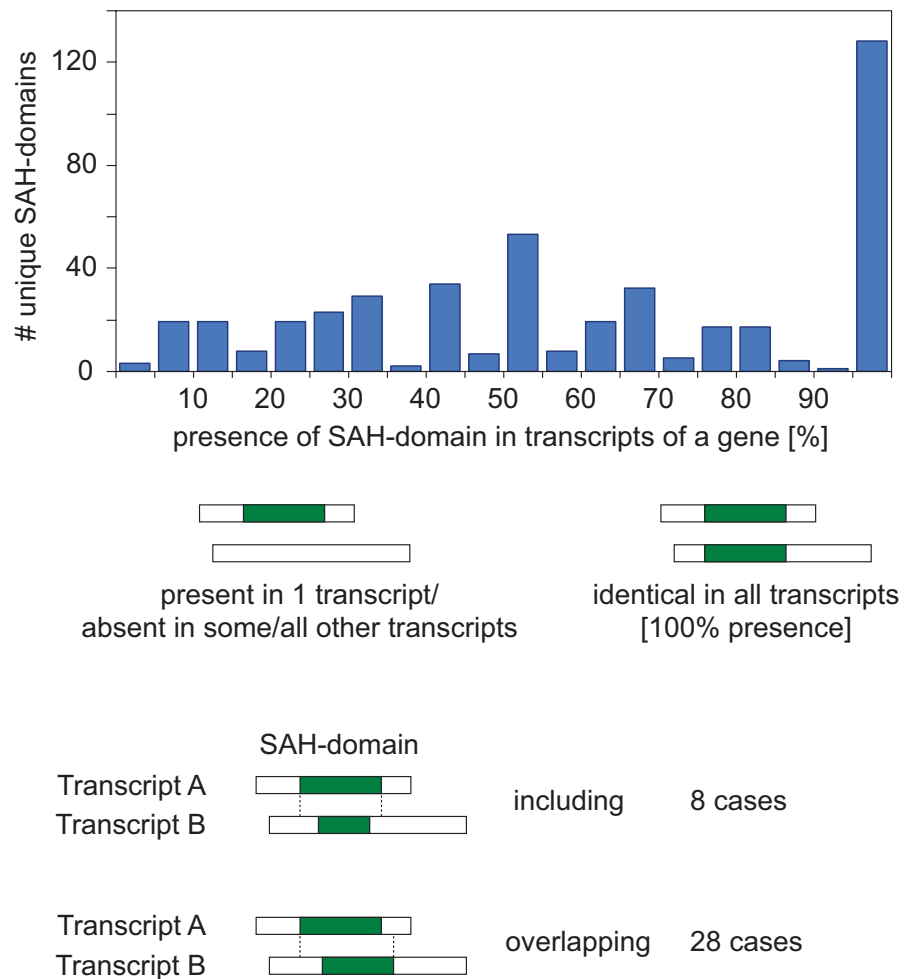


Fig 3. Presence of SAH-domains in alternative human transcripts. All predicted SAH-domains from the human “all” dataset were compared across transcripts from the same gene and combined to *unique* SAH-domains in case of identical sequences. For each unique SAH-domain its presence or absence in all annotated transcripts of the gene was determined. The histogram represents the fraction of unique SAH-domains present in alternative transcripts, with each bin combining values of a range of 5%. The dataset also contains a few cases where a unique SAH-domain is completely part of a larger unique SAH-domain of another transcript and where unique SAH-domains from different transcripts overlap.

<https://doi.org/10.1371/journal.pone.0191924.g003>

domains are enriched in longer proteins (Fig 4). This is consistent with their main function as connectors between other protein domains, although they might also function as direct binding site for other proteins. Because of their limited structural and amino acid diversity, direct protein-protein and protein-RNA binding via the SAH-domains is likely very rare. Mostly, SAH-domains are anchored to additional domains such as those in myosins [2,5], the Inner Centromere Protein (INCENP) [7], and in the spliceosomal proteins MFAP1 and Snu23 [8]. Because of this combination with other domains, which has also been found in an earlier analysis [24], it is evident that SAH-domains are rather found in longer protein sequences.

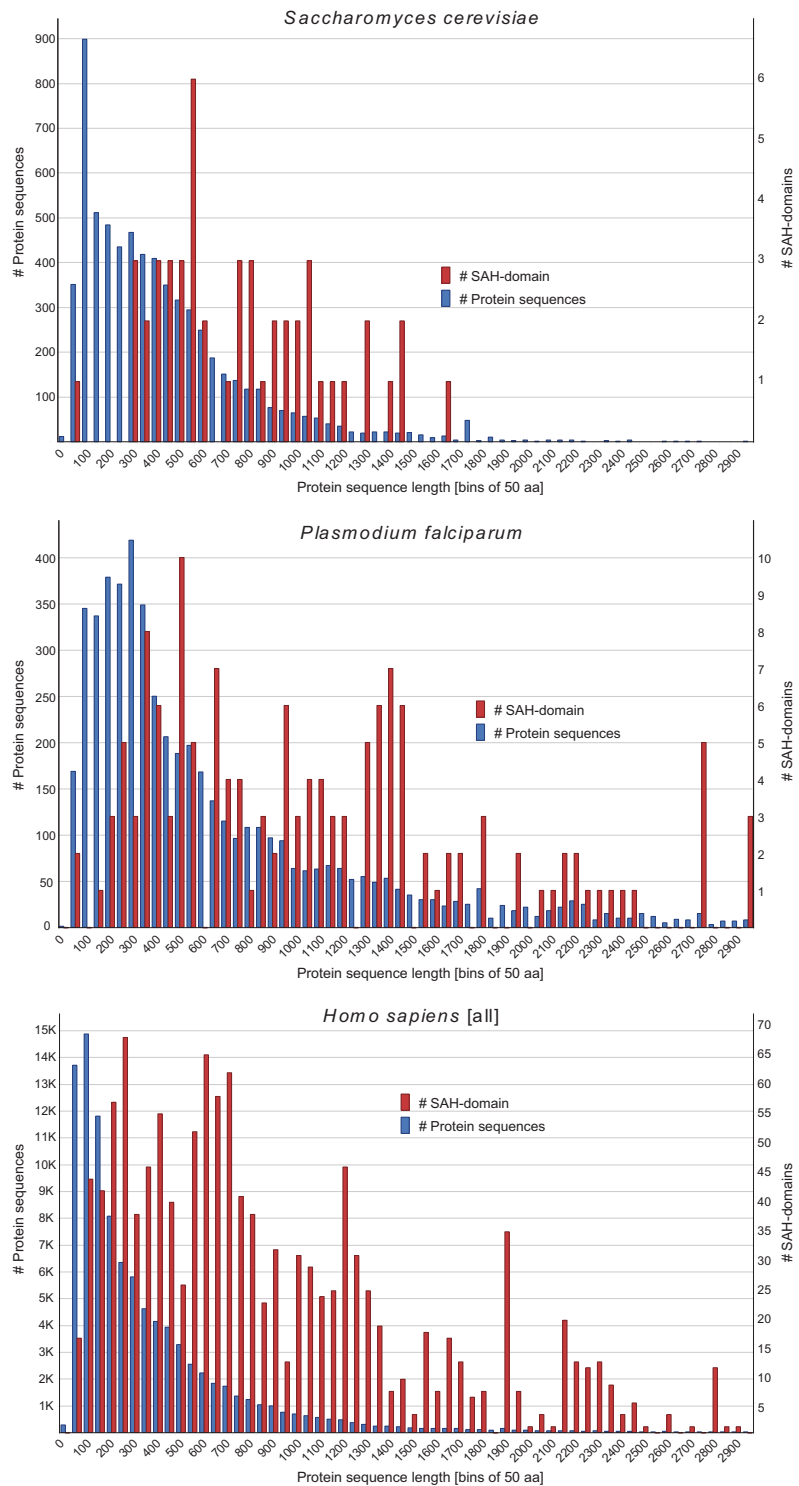


Fig 4. Distribution of SAH-domains with respect to protein sequence length. As examples for comparing the lengths of protein sequences in general with the presence of SAH-domains in proteins, the proteins from the *Plasmodium falciparum*, *Saccharomyces cerevisiae* and human [all] datasets were plotted. The latter proteins comprise the translations of all annotated gene transcripts. Proteins were combined in bins of 50 amino acids for better visualization, and proteins with length >3000 aa were omitted. All SAH-domains identified with the 14 aa window are counted and shown.

<https://doi.org/10.1371/journal.pone.0191924.g004>

Another characteristic of SAH-domains distinguishing them from any other domain is their amino acid distribution with up to 80% of the residues being E, K and R [4,13,24]. Of the various possibilities to build salt bridges, E \rightarrow R was shown in short peptides to be more α -helix stabilizing than E \rightarrow K [27], and both are more stabilizing than the reverse salt bridges R \rightarrow E and K \rightarrow E. E \rightarrow R salt bridges are also the most favourable for the speed of folding [22]. However, comparison of long repeats of AEEEEXXX with X being either K or R showed, that such peptides aggregated when two or three of the X were R [28]. A repeat including one arginine, however, was more helical and stable than a repeat with only lysines. Repeats with three or five alanines per heptad also aggregated [28]. Thus, it seems that a certain percentage of charged amino acids and a mixture of arginines and lysines are the most favourable to stabilize single α -helices.

The analysis of the amino acid distribution in the predicted SAH-domains across the eukaryotes shows that aspartate is rarely used in SAH-domains and that glutamate is the dominating negatively charged amino acid (Fig 5). This is consistent with earlier extrapolations from few data [4,13,24] and in agreement with studies on short peptides showing considerably less stabilizing effects of D \rightarrow K and D \rightarrow R salt bridges compared to their glutamate homologs [21,23]. Across the eukaryotes there are strong differences for the preference of the positively charged amino acid (Fig 5). In plants, green and red algae, diplomonads, kinetoplastids, and metazoans, arginines are preferred to lysines, while lysines are strongly preferred in alveolates, amoebae, and yeasts. The finding of twice as many arginines than lysines in red algae, diplomonads, and kinetoplastids seems to contract the findings of aggregating peptides with similar arginine to lysine proportions [28]. However, these peptides were based on repeats of three negatively and three positively charged amino acids, and natural peptides might form stable single α -helices if the same salt bridges were randomly distributed and not present in such ordered three-to-three repeats. In total, about 80% of the supposed heptad positions are occupied by charged residues or asparagine and glutamine, which corresponds to about 5,5 heptad positions (Fig 6). Species with, in total, less E/K/R have more aspartates, asparagines or glutamines. This also corresponds with the results obtained from the AEEEEXXX repeats that showed that peptides aggregated if three or more positions of the heptads are occupied by alanines [28]. Alanines show the widest distribution from all uncharged amino acids found to be present in SAH-domains (Fig 5), with the highest fractions found in green and red algae, diplomonads, and kinetoplastids. In contrast, alanines are rarely present in *Plasmodium falciparum* and *Dictyostelium discoideum* SAH-domains.

Next we were interested in identifying the most common patterns in natural SAH-domains (Fig 7). Compared to total numbers of SAH-domains and lengths of SAH-domains the most common heptad repeats are rather rare within each species. For example, the most common heptad pattern in human SAH-domains, RERERER, is present in only 3.6% of the sequences with SAH-domains and accounts for only 0.7% of the amino acids of all SAH-domains. This indicates that SAH-domains do not have certain, common patterns but instead are highly variable with respect to their sequences. The most common patterns across most species contain duplets of oppositely charged amino acids, [D/E][K/R]. To our knowledge, such repeats have not extensively been studied experimentally yet. Repeats of (EK)_n were shown to be at least partially helical [23], but it is not known whether these are monomeric or aggregate, or even form stable single α -helices at all. However, only a minor fraction of the SAH-domains

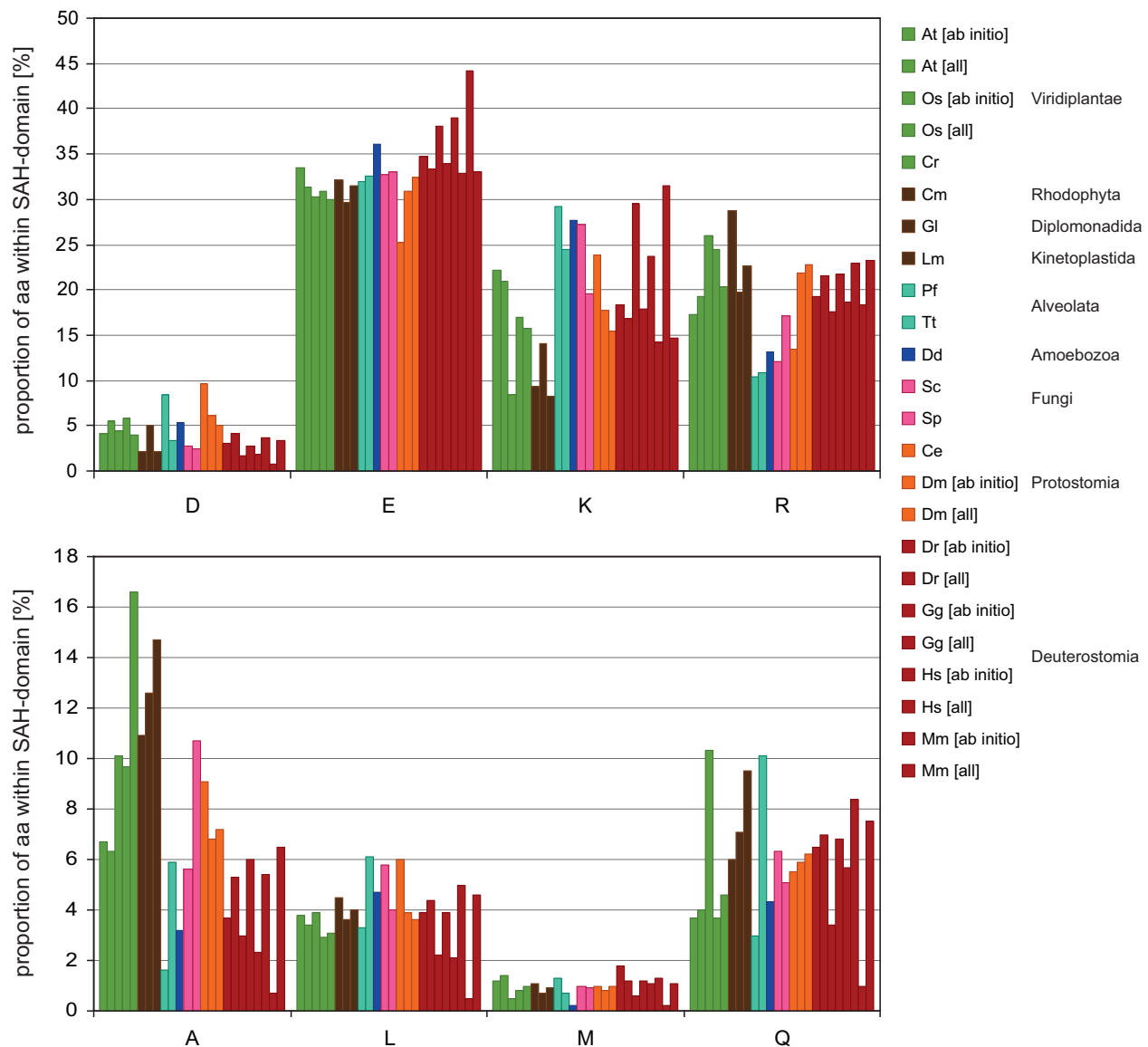


Fig 5. Proportion of the most abundant amino acids within SAH-domains with respect to dataset and eukaryotic domain. For computing the amino acid proportions, the SAH-domains predicted with the 14 aa window were used. The proportions within the four windows are almost identical for each species.

<https://doi.org/10.1371/journal.pone.0191924.g005>

exclusively consists of such $[D/E][K/R]$ repeats. In most cases, such repeats are part of predicted SAH-domains with variable sequence such as the SAH-domain in human Myo10, “AEKREQEEKKKQEEEEKKKREEEEREREREREAELRAQQEEETRKQEELEALQKSQKEAEL” (ENSP00000421280). We suspect that these $[D/E][K/R]$ repeats are enriched in the analysis because of their very simple pattern. If regions exclusively consisting of $[D/E][K/R]$ repeats do not form stable single α -helices, a small fraction of false-positive predictions might be present in a Waggawagga analysis.

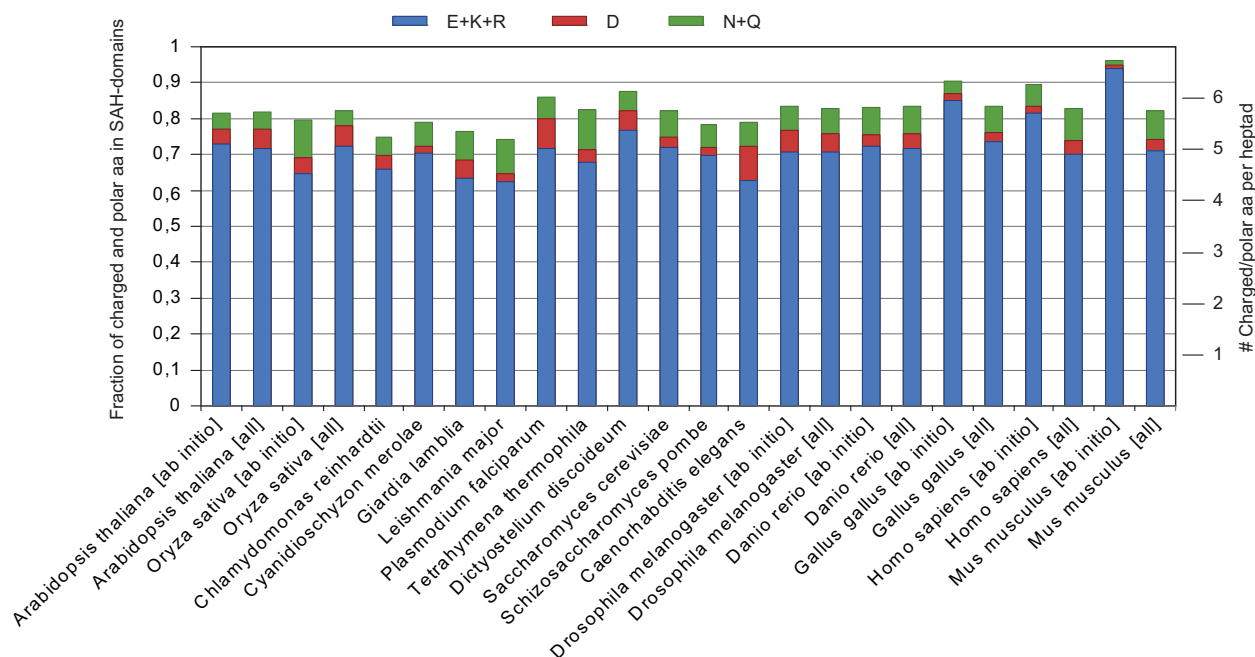


Fig 6. Fraction of charged and polar amino acids within SAH-domains across species. For better orientation, fractions are also given for distinct numbers of heptad positions fully occupied by charged/polar amino acids.

<https://doi.org/10.1371/journal.pone.0191924.g006>

Other common heptad patterns contain alanines and glutamines (Fig 7) and resemble the widely studied peptides based on poly-alanine backbones onto which oppositely charged amino acids were placed in all possible spacings. Heptad patterns based on oppositely charged amino acid triplets such as $(AE_3 [K/R]_3)_n$ [28] were not found. Octad patterns with quadruplets of glutamates and arginines/lysines $(E_4 [K/R]_4)$, which have systematically been studied experimentally [3,23], are extremely rare in natural SAH-domains of eukaryotes. Also, they are not present in repeated patterns indicating that these octad patterns appeared by chance but are in fact part of heptad patterns. If clusters of identical amino acids were found, then these appear in repeated heptad patterns from triplets and quadruplets (Fig 7, see patterns of *Arabidopsis thaliana* and *Saccharomyces cerevisiae*, for example).

Although heptad patterns are shown in Fig 7, most of these patterns are built from even-numbered smaller motifs. Heptad repeats tend to form supra-molecular structures because the repeat moves around the helix axis, as is evident from coiled-coil structures, where hydrophobic residues form a left-handed helical seam along the surface of the right-handed α -helix which is buried in the centre of the dimer. In contrast, patterns of even-numbered motifs are supposed to form straight α -helices and to rather aggregate in stacks, if at all, than in intertwined supra-molecular helices.

Functional analysis of SAH-domains in *Saccharomyces cerevisiae* and human

To determine whether proteins with SAH-domains are particularly involved in certain cellular processes, localizations and molecular functions, we performed gene ontology enrichment analyses of the *Saccharomyces cerevisiae* and human datasets. Yeast (*Saccharomyces cerevisiae*)

| <i>Arabidopsis thaliana</i> [all] | <i>Oryza sativa</i> [all] | <i>Chlamydomonas reinhardtii</i> | <i>Cyanidioschyzon merolae</i> | <i>Giardia lamblia</i> | <i>Leishmania major</i> |
|-----------------------------------|--------------------------------------|----------------------------------|---------------------------------|----------------------------------|---------------------------|
| RDRDRDR 15 | RERERER 23 | RERERER 11 | ADENRHR 1 | RAAEQAR 12 | AEEQARR 91 |
| ERERERE 14 | RDRDRDR 20 | EAKAKAE 10 | IADENRH 1 | ARRDEEA 12 | EAEEQAR 55 |
| DRDRDRD 12 | ERERERE 16 | ERERERE 9 | KIADENR 1 | QARRDEE 12 | EEQARRE 53 |
| RERERER 12 | DRDRDRD 14 | KAEAEAK 9 | RKIADEN 1 | EQARRDE 12 | REAEEOA 52 |
| EKKKEEE 7 | LRDRER 10 | EAEAKAK 8 | LRKIADE 1 | AEQARRD 12 | QARREAE 52 |
| KKKEEEE 7 | RDRDRER 8 | AEAEAKA 8 | ELRKIAD 1 | AAEQARR 12 | EQARREA 52 |
| RDRDRER 6 | RERERDR 8 | AKAKAEA 8 | EELRKIA 1 | ARAAEOA 12 | ARREAE 52 |
| EEAKRRE 5 | DRERERE 7 | KAKAEAE 7 | QRKREER 1 | EARAAEQ 11 | RREAEEO 51 |
| KREEEER 5 | EARERAA 7 | AEAKAKA 7 | RQRKREE 1 | EEARAAE 11 | EQARRVA 39 |
| EEEEARK 5 | REDRDR 7 | AKAEAEA 7 | ERQRKE 1 | DEEARAA 11 | ARRVAEE 39 |
| | | | | | |
| <i>Plasmodium falciparum</i> | <i>Tetrahymena thermophila</i> | <i>Dictyostelium discoideum</i> | <i>Saccharomyces cerevisiae</i> | <i>Schizosaccharomyces pombe</i> | |
| EKEKEKE 103 | KRLAEK 36 | EKEKEKE 638 | KKEKKEK 13 | AKREAE 16 | |
| KEKEKEK 98 | EEKRLAE 36 | KEKEKEK 619 | KEKKEKK 12 | KREAEK 12 | |
| RLKEEER 57 | AEEKRLA 36 | ERERERE 114 | EEEEKKK 11 | EKAKREA 11 | |
| EERLKEE 56 | EKRLAEE 35 | RERERER 103 | KKEEEEK 9 | KAKREAE 11 | |
| EEERLKE 55 | RERERER 30 | RDRDRDR 100 | EKKEKKE 8 | EKAKRE 11 | |
| KEEERLK 54 | ERERERE 29 | DRDRDRD 85 | KEEEKK 8 | AEEKAKR 11 | |
| LKEEERL 53 | LAKAEAE 28 | KERLEKE 72 | EEKKKKE 7 | EAEKAK 11 | |
| ERLKEEE 51 | KEAEKCR 28 | EKERLEK 64 | EKKKKEE 7 | REAEKA 11 | |
| RDRDRDR 51 | RLAEKA 28 | LEKERLE 57 | KKKKEEE 7 | EAEENAK 4 | |
| DRDRDRD 31 | EAEKRL 28 | KEKEKE 51 | KKKEEEE 7 | AENAKR 4 | |
| | | | | | |
| <i>Caenorhabditis elegans</i> | <i>Drosophila melanogaster</i> [all] | <i>Danio rerio</i> [all] | <i>Gallus gallus</i> [all] | <i>Homo sapiens</i> [all] | <i>Mus musculus</i> [all] |
| DDKLKQE 77 | RERERER 86 | ERERERE 100 | EKEKEKE 164 | RERERER 64 | ERERERE 99 |
| KLKQEA 75 | ERERERE 81 | RERERER 93 | KEKEKEK 152 | ERERERE 61 | RERERER 97 |
| DKLKQEA 73 | KDKDKDK 25 | RDRDRDR 39 | ERERERE 88 | EKIREQE 27 | RDRDRDR 27 |
| KQEADAK 72 | DKDKDKD 21 | ERLEKER 34 | RERERER 86 | EKIREQ 27 | DRDRDRD 22 |
| QEADAKL 71 | RERERDR 20 | EKEKEKE 31 | KRREEKR 76 | RDRDRDR 26 | EKERERE 17 |
| LKQEADA 71 | REDRDR 18 | LERERLE 29 | EKRREEK 74 | KIREQEE 26 | REKERER 13 |
| KDDKLKQ 66 | RDRDRDR 18 | ERERLEK 28 | EEKRREE 74 | QEEKIRE 23 | EREKERE 13 |
| ADAKLKK 58 | RDRERER 17 | KEKEKEK 28 | RREEKRR 73 | RDRDRER 17 | KDKDKDK 12 |
| EADAKLK 57 | RDRDRER 17 | RERLEKE 28 | REEKRRE 72 | IREQEEK 17 | EREREKE 12 |
| EKDDKLK 51 | ERRREE 16 | DRDRDRD 26 | KEKKRKE 47 | RDRERDR 16 | DKKDKDK 11 |

Fig 7. The ten most common heptad repeats found in SAH-domains per species. For each dataset of SAH-domains predicted using the 14 amino acid window all possible heptads were determined and ranked by frequency, given as numbers next to each heptad sequence.

<https://doi.org/10.1371/journal.pone.0191924.g007>

proteins with SAH-domains have not been analysed at a whole-genome level yet, because yeast was not among the species with highest numbers of SAH-domain containing proteins in an UniProt analysis [24]. Yeast proteins with SAH-domains are particularly enriched in the GO terms macromolecular complex subunit organization, cellular component biogenesis and RNA metabolic processes. Enriched GO terms also include nuclear and ribonucleoprotein complex localization, and function in ribosome and nucleic acid binding. Cytoskeleton associated processes and functions were not among the enriched terms for the yeast proteins. The

recent discovery of a few spliceosomal proteins with potential SAH-domains [8] is consistent with the observed enrichment of the yeast proteins.

The human proteins with SAH-domains have roles in all types of RNA processing (mRNA processing, mRNA metabolic processes, RNA splicing) and cytoskeleton organization. They are enriched in cytoskeleton and adherens junction localization terms, and are predicted to function in RNA binding, protein binding involved in cell and cell-cell adhesion, and cytoskeletal protein binding. In contrast to earlier analyses based on considerably smaller datasets of human proteins with SAH-domains that showed a wide distribution of cellular localizations and functions [4,13,24], the analysis of our extended whole-genome based dataset demonstrates a strong enrichment in RNA- and cytoskeleton-related processes and localizations.

Customized usage of Waggawagga-CLI

SAH-predictions are based on scoring statistics. In our previous analysis of almost 8000 myosins from species across all eukaryotes we showed that the scores from SAH-predictions were scoring-window-dependent and could be placed on a continuously increasing line [14]. The exponential shape of the curve allowed defining a lower and an upper cut-off, below which and above which the probability to not-have and to have, respectively, identified an SAH-domain is high. The scores between these cut-offs, in the so-called “twilight”-zone, do not allow a simple “is” or “is not” decision. To provide full flexibility to the prediction and analysis of SAH-domains, Waggawagga-CLI allows adjusting the scoring matrices and the cut-offs for result filtering. There are two amino acid matrix files, one each for amino acids in ($i, i+3$) and in ($i, i+4$) spacings, where stabilizing and destabilizing interaction scores can be defined for every amino acid combination. Modifying these files allows, for example, fine-tuning the search for specific sub-types of SAH-domains or adapting to species- or lineage-specific parameters such as general amino acid usage. Waggawagga-CLI also allows separating the steps of SAH-domain prediction and analyses so that users experienced in database interaction can easily design additional analyses.

Conclusions

Waggawagga-CLI is a complementary tool to predict functional domains in whole genome annotations. Our predictions included all previously proposed SAH-domains as far as they were particularly named such as the 36 human proteins identified by BLAST [13]. While BLAST is based on sequence homology, Waggawagga-CLI uses a dedicated scoring scheme for potential $i,i+3$ and $i,i+4$ interactions along the sequence and we were, therefore, able to predict more than ten times more SAH-domains. Waggawagga-CLI is thus similar to other SAH-domain prediction software such as SCAN4CSAH and FT_CHARGE [24] but, in contrast to these, allows users to modify the scoring schemes and to adjust cut-off parameters to individual analysis needs. Several proteins from spliceosomal complexes have recently been suggested to contain SAH-domains [8] but most of these rather represent isolated α -helical segments that folded upon binding to other proteins. In MFAP1, the regions with considerable SAH-score are shorter than the minimum window of 14 amino acids that we used in our analysis. There were few potential charged interactions along the α -helix of MFAP1 [8] and similar numbers of potential charged interactions are found for most predicted coiled-coil regions. We suppose that by altering Waggawagga’s scoring scheme to include MFAP1 and similar proteins many false positive coiled-coil proteins will also appear in the list. Adjusting the scoring scheme might, however, be useful if subsets of proteins are analysed that likely do not contain any coiled-coil proteins.

Materials and methods

Implementation

Waggawagga-CLI is available for the main operating systems Linux (arch. x86 and 64-Bit) and macOS (10.10 or higher) [and Windows]. It comes in precompiled system-specific packages with a portable Ruby environment (Traveling-Ruby) and a mobile SQLite-database, where the analysed sequence data are stored for direct or later use. Waggawagga-CLI runs out of the box, not requiring any further installations after downloading and extracting the tarball.

In general, to be fully functional the software requires Ruby version 1.9 or later, and the following gems: Active-Record ($> 4.0.x$), BioRuby (1.5.0) [29], and SQLite3 (1.3.9). Besides the textual result files, full-sequence prediction score graphs (depicted along the full sequences) are additionally available as SVG, if the graphical toolkit GnuPlot (<http://www.gnuplot.info>) is pre-installed on the user's system.

Waggawagga-CLI in contrast to webserver Waggawagga

The focus of Waggawagga-CLI is explicitly on the SAH-domain prediction in large protein sequence datasets (instead of just one at a time). The CLI-version analysis workflow separates prediction and analysis. Sequences are parsed, scored and imported in a single step and can be evaluated with different parameter preferences later-on. The scoring process itself can be adjusted as well, by modifying the pre-installed scoring matrices. The application should run with FASTA-files of any size, although it was tested with only the datasets presented here. The prediction result files are strictly organized in subfolders in the results directory named by the working title parameter ('id'). Only results with scores above the initial cut-off are kept for later inspection by the user.

SAH scoring algorithm

The SAH prediction is based on the classical *helical net* diagram which is the representation of a single α -helix opened along a line parallel to its axis and laid flat [30]. Each sequence is considered as a continuous right-handed α -helix and each position in the helix is assumed to be repeated every eighth amino acid. Thus, each sequence can be depicted as a repeat of heptads in the *helical net* diagram. The amino acid positions within each heptad are formalized as letters *a-g* and the first amino acid of each sequence is set to position *a* of the first heptad. This is an important difference to the Waggawagga-webinterface version, where the positions of the amino acids within the heptads are derived from the coiled-coil prediction tools and accordingly allow for gaps and any other pattern within the *helical net* diagram. According to the classical representation the α -helix in the *helical net* diagram is opened along the *f* column. In contrast to the classical representation, which is read from bottom to top, we adopted the view introduced for SAH-domains in which SAH sequences are depicted from top to bottom [2,13] and which seems to be used by most if not all in the SAH community.

For each position in the *helical net* amino acid interactions between interacting residues in $i,i+3$ and $i,i+4$ distance are drawn. These interactions are classified into strong, medium and weak stabilizing types, a helix-supporting type and types of destabilizing interactions. But this is only a linguistic differentiation; the interactions can be assigned every possible score. In Waggawagga-CLI, the values for possible amino acid interactions are taken from customizable scoring matrices. The standard scoring files are located in the *config*-directory, are named '*scoring_matrix_i_3.csv*' (for interactions in $i,i+3$ distance) and '*scoring_matrix_i_4.csv*' (for interactions in $i,i+4$ distance; S1 Fig), and can freely be edited. In addition to these binary amino acid interactions, we consider two types of interaction networks: i) Subsequent

hydrophobic amino acids (V, I, L, M, F, Y) in $i,i+3,i+6$, $i,i+3,i+7$, $i,i+4,i+7$ and $i,i+4,i+8$ distance are regarded as destabilizing (because of potentially stabilizing hydrophobic seams in coiled-coil dimers) and contribute a negative network-score. ii) Subsequent oppositely charged amino acids in $i,i+3,i+6$, $i,i+3,i+7$, $i,i+4,i+7$ and $i,i+4,i+8$ distance are known to stabilize α -helices more than the sum of the respective binary interactions, and thus contribute an additional positive network-score. These network-scores can be adjusted in the config-file. In contrast to the SCAN4CSH algorithm, we did not consider interactions of identically charged residues in $i,i+3$ and $i,i+4$ distance, and of oppositely charged amino acids in $i,i+1$ and $i,i+2$ distance, which were regarded as destabilizing [4]. Because such interactions are present in most patterns that are regarded as exemplary SAH-domains, e.g. EEEEEKKK, we suspect that inclusion of these interactions does not help in distinguishing between SAH-domains and non-SAH-domains.

Computing the SAH-score

For computing an SAH-score, all amino acid interaction scores in a *helical net* representation of a certain sequence window are summed up. By default, Waggawagga-CLI computes SAH-scores for windows of 14, 21, 28 or 49 amino acids. Because we consider the protein query sequence to be a continuous α -helix there are also interactions to and from amino acids at the first and the last helical turn within the scoring window. By definition we include all amino acid interactions from residues of the last helical turn of the window (positions *d-g*) to amino acids of the next heptad in the SAH-score. The sum of the interaction scores of each window is then normalized with respect to the highest possible score for the respective window, which is obtained by summing all interactions of an EEEEEKKK repeat and which is set to "1". If individual interaction scores are changed in the scoring matrices (see above), the highest possible scores for each window need to be adjusted accordingly. This is done in the config-file. The user needs to keep in mind that the heptad repeat pattern that will result in the highest possible score, might also change when the scoring matrices change. This interplay between individual interaction scores and accordingly unlimited possibilities for the pattern resulting in the highest possible score is the reason why we keep the window sizes for computing SAH-scores fixed. The SAH-score for the respective window is then assigned to the central amino acid (windows 21 and 49) or amino acids 8 and 15 (windows 14 and 28) of the window, respectively. Scores for the first and last 7, 10, 14 and 24 amino acids (depending on window size) of the protein sequences are calculated by filling the window with dummy amino acids at the N- or C-terminus. By definition, these dummy residues are strictly neutral and do not have interactions with other amino acids. The computed SAH-scores vary considerably from amino acid to amino acid along the sequence depending on how many interactions are lost and gained on each side of the amino acid window. To term a protein sequence region an SAH-domain, the scores of all amino acids within this region need to be above a user-defined SAH-score cut-off (default: 0.25).

Computing the SAH-domain-score

To be able to compare and rank SAH-domains, we developed the SAH-domain-score. By definition, SAH-domains have a minimum length of 14 amino acids with SAH-scores for each of the 14 amino acids above the SAH-score cut-off. 20% of the amino acids within the full SAH-domain are allowed to have SAH-scores below the cut-off to avoid splitting long SAH-domains into multiple short SAH-domains separated by just one or a few amino acids. However, SAH-domains need to start and to end with amino acids having SAH-scores above the cut-off (default: 0.25). Taking the average of all SAH-scores within an SAH-domain as SAH-domain-

score would introduce strong bias by peak values and therefore strongly influence a comparison of short with long SAH-domains. To allow length-independent comparison of SAH-domains we thus developed a score based on a certain amino acid window. Accordingly, the SAH-domain-score is the maximum of the scores computed as the average of SAH-scores within a window of neighbouring amino acids (default: 14 amino acids). For example, given an SAH-domain of 20 amino acids (scores of all 20 aa above the 0.25 score cut-off) six SAH-domain-scores (each possible 14 aa window) are calculated and the highest of these scores is taken as SAH-domain-score for the respective SAH-domain. The length of the window for determining SAH-domain-scores can be adjusted (Advanced Mode parameter).

Data sources and analyses

The following protein sequence files were downloaded from Ensembl release-87 (<ftp://ftp.ensembl.org/pub/release-87/>): *Arabidopsis thaliana*.TAIR10.pep.abinitio, *Arabidopsis thaliana*.TAIR10.pep.all, *Caenorhabditis elegans*.WBcel235.pep.all, *Chlamydomonas reinhardtii*.v3.1.pep.all, *Cyanidioschyzon merolae*.ASM9120v1.pep.all, *Danio rerio*.GRCz10.pep.abinitio, *Danio rerio*.GRCz10.pep.all, *Dictyostelium discoideum*.dicty_2.7.pep.all, *Drosophila melanogaster*.BDGP6.pep.abinitio, *Drosophila melanogaster*.BDGP6.pep.all, *Gallus gallus*.Gallus_gallus-5.0.pep.abinitio, *Gallus gallus*.Gallus_gallus-5.0.pep.all, *Giardia lamblia*.GCA_000002435.1.pep.all, *Homo sapiens*.GRCh38.pep.abinitio, *Homo sapiens*.GRCh38.pep.all, *Leishmania major*.ASM272v2.pep.all, *Mus musculus*.GRCm38.pep.abinitio, *Mus musculus*.GRCm38.pep.all, *Oryza sativa*.IRGSP-1.0.pep.abinitio, *Oryza sativa*.IRGSP-1.0.pep.all, *Plasmodium falciparum*.ASM276v1.pep.all, *Saccharomyces cerevisiae*.R64-1-1.pep.all, *Schizosaccharomyces pombe*.ASM294v2.pep.all, *Tetrahymena thermophila*.JCVI-TTA1-2.2.pep.all. The SAH-domain predictions presented and analysed in this study were produced with the latest Waggawagga-CLI version using default parameters.

Gene Ontology enrichment analysis

Gene Ontology enrichment analyses were done with WebGestalt [31]. The lists of unique genes in gene symbol format were uploaded to WebGestalt and the GO Enrichment Analysis selected. The entire *Saccharomyces cerevisiae* and human genome annotations, respectively, were set as background and 0.05 as threshold for the p-value for the significance test using the default statistical method "hypergeometric".

Software and data availability

Waggawagga-CLI is available for download from <http://waggawagga.motorprotein.de>. The software and all results from data analysis as presented in this study (SAH-domain predictions and GO analyses) are also available from figshare (doi: [10.6084/m9.figshare.5435947](https://doi.org/10.6084/m9.figshare.5435947)).

Supporting information

S1 Fig. Composition of the SAH-score. The figure shows publications and the respective peptides and amino acid interactions analysed. From the reported observed helicities and measured energies of side chain interactions we build the scoring scheme shown in the table at the bottom. The table lists the score of each interaction taken for computing the SAH-score.

(PDF)

Acknowledgments

We would like to thank Prof. Christian Griesinger and Prof. Stephan Waack for their continuous generous support.

Author Contributions

Conceptualization: Martin Kollmar.

Data curation: Dominic Simm, Martin Kollmar.

Formal analysis: Dominic Simm.

Investigation: Dominic Simm, Martin Kollmar.

Methodology: Dominic Simm.

Software: Dominic Simm.

Supervision: Martin Kollmar.

Validation: Dominic Simm, Martin Kollmar.

Visualization: Dominic Simm, Martin Kollmar.

Writing – original draft: Dominic Simm, Martin Kollmar.

References

1. Wang CL, Chalovich JM, Graceffa P, Lu RC, Mabuchi K, Stafford WF. A long helix from the central region of smooth muscle caldesmon. *J Biol Chem.* 1991; 266: 13958–13963. PMID: [1856225](#)
2. Knight PJ, Thirumurugan K, Xu Y, Wang F, Kalverda AP, Stafford WF, et al. The Predicted Coiled-coil Domain of Myosin 10 Forms a Novel Elongated Domain That Lengthens the Head. *J Biol Chem.* 2005; 280: 34702–34708. <https://doi.org/10.1074/jbc.M504887200> PMID: [16030012](#)
3. Sivaramakrishnan S, Spink BJ, Sim AYL, Doniach S, Spudich JA. Dynamic charge interactions create surprising rigidity in the ER/K alpha-helical protein motif. *Proc Natl Acad Sci U S A.* 2008; 105: 13356–13361. <https://doi.org/10.1073/pnas.0806256105> PMID: [18768817](#)
4. Süveges D, Gáspári Z, Tóth G, Nyitrai L. Charged single alpha-helix: a versatile protein structural motif. *Proteins.* 2009; 74: 905–916. <https://doi.org/10.1002/prot.22183> PMID: [18712826](#)
5. Spink BJ, Sivaramakrishnan S, Lipfert J, Doniach S, Spudich JA. Long single alpha-helical tail domains bridge the gap between structure and function of myosin VI. *Nat Struct Mol Biol.* 2008; 15: 591–597. <https://doi.org/10.1038/nsmb.1429> PMID: [18511944](#)
6. Swanson CJ, Sivaramakrishnan S. Harnessing the unique structural properties of isolated α -helices. *J Biol Chem.* 2014; 289: 25460–25467. <https://doi.org/10.1074/jbc.R114.583906> PMID: [25059657](#)
7. Samejima K, Platani M, Wolny M, Ogawa H, Vargiu G, Knight PJ, et al. The Inner Centromere Protein (INCENP) Coil Is a Single α -Helix (SAH) Domain That Binds Directly to Microtubules and Is Important for Chromosome Passenger Complex (CPC) Localization and Function in Mitosis. *J Biol Chem.* 2015; 290: 21460–21472. <https://doi.org/10.1074/jbc.M115.645317> PMID: [26175154](#)
8. Ulrich AKC, Seeger M, Schütze T, Bartlick N, Wahl MC. Scaffolding in the Spliceosome via Single α Helices. *Structure.* 2016; 24: 1972–1983. <https://doi.org/10.1016/j.str.2016.09.007> PMID: [27773687](#)
9. Kuhlman B, Yang HY, Boice JA, Fairman R, Raleigh DP. An exceptionally stable helix from the ribosomal protein L9: implications for protein folding and stability. *J Mol Biol.* 1997; 270: 640–647. <https://doi.org/10.1006/jmbi.1997.1146> PMID: [9245593](#)
10. Sivaramakrishnan S, Spudich JA. Systematic control of protein interaction using a modular ER/K α -helix linker. *Proc Natl Acad Sci U S A.* 2011; 108: 20467–20472. <https://doi.org/10.1073/pnas.1116066108> PMID: [22123984](#)
11. Gigant B, Curmi PA, Martin-Barbey C, Charbaut E, Lachkar S, Lebeau L, et al. The 4 A X-ray structure of a tubulin:stathmin-like domain complex. *Cell.* 2000; 102: 809–816. PMID: [11030624](#)
12. Chui AJ, López CJ, Brooks EK, Chua KC, Doupey TG, Foltz GN, et al. Multiple structural states exist throughout the helical nucleation sequence of the intrinsically disordered protein stathmin, as reported by electron paramagnetic resonance spectroscopy. *Biochemistry (Mosc).* 2015; 54: 1717–1728. <https://doi.org/10.1021/bi500894q> PMID: [25715079](#)

13. Peckham M, Knight PJ. When a predicted coiled coil is really a single α -helix, in myosins and other proteins. *Soft Matter*. 2009; 5: 2493–2503. <https://doi.org/10.1039/B822339D>
14. Simm D, Hatje K, Kollmar M. Distribution and evolution of stable single α -helices (SAH domains) in myosin motor proteins. *PLoS One*. 2017; 12: e0174639. <https://doi.org/10.1371/journal.pone.0174639> PMID: 28369123
15. Marqusee S, Baldwin RL. Helix stabilization by Glu-...Lys+ salt bridges in short peptides of de novo design. *Proc Natl Acad Sci U S A*. 1987; 84: 8898–8902. PMID: 3122208
16. Marqusee S, Robbins VH, Baldwin RL. Unusually stable helix formation in short alanine-based peptides. *Proc Natl Acad Sci U S A*. 1989; 86: 5286–5290. PMID: 2748584
17. Lyu PC, Marky LA, Kallenbach NR. The role of ion pairs in α -helix stability: two new designed helical peptides. *J Am Chem Soc*. 1989; 111: 2733–2734. <https://doi.org/10.1021/ja00189a067>
18. Scholtz JM, Qian H, Robbins VH, Baldwin RL. The energetics of ion-pair and hydrogen-bonding interactions in a helical peptide. *Biochemistry (Mosc)*. 1993; 32: 9668–9676.
19. Wang E, Wang CL. (i, i + 4) Ion pairs stabilize helical peptides derived from smooth muscle caldesmon. *Arch Biochem Biophys*. 1996; 329: 156–162. <https://doi.org/10.1006/abbi.1996.0204> PMID: 8638947
20. Olson CA, Spek EJ, Shi Z, Vologodskii A, Kallenbach NR. Cooperative helix stabilization by complex Arg-Glu salt bridges. *Proteins*. 2001; 44: 123–132. PMID: 11391775
21. Huyghues-Despointes BM, Scholtz JM, Baldwin RL. Helical peptides with three pairs of Asp-Arg and Glu-Arg residues in different orientations and spacings. *Protein Sci Publ Protein Soc*. 1993; 2: 80–85. <https://doi.org/10.1002/pro.5560020108> PMID: 8443591
22. Meuzelaar H, Vreede J, Woutersen S. Influence of Glu/Arg, Asp/Arg, and Glu/Lys Salt Bridges on α -Helical Stability and Folding Kinetics. *Biophys J*. 2016; 110: 2328–2341. <https://doi.org/10.1016/j.bpj.2016.04.015> PMID: 27276251
23. Baker EG, Bartlett GJ, Crump MP, Sessions RB, Linden N, Faul CFJ, et al. Local and macroscopic electrostatic interactions in single α -helices. *Nat Chem Biol*. 2015; 11: 221–228. <https://doi.org/10.1038/nchembio.1739> PMID: 25664692
24. Gáspári Z, Sűveges D, Perczel A, Nyitrai L, Tóth G. Charged single α -helices in proteomes revealed by a consensus prediction approach. *Biochim Biophys Acta*. 2012; 1824: 637–646. <https://doi.org/10.1016/j.bbapap.2012.01.012> PMID: 22310480
25. Simm D, Hatje K, Kollmar M. Waggawagga: comparative visualization of coiled-coil predictions and detection of stable single α -helices (SAH domains). *Bioinformatics*. 2015; 31: 767–769. <https://doi.org/10.1093/bioinformatics/btu700> PMID: 25338722
26. Kersey PJ, Allen JE, Armean I, Boddu S, Bolt BJ, Carvalho-Silva D, et al. Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Res*. 2016; 44: D574–580. <https://doi.org/10.1093/nar/gkv1209> PMID: 26578574
27. Sommese RF, Sivaramakrishnan S, Baldwin RL, Spudich JA. Helicity of short E-R/K peptides. *Protein Sci Publ Protein Soc*. 2010; 19: 2001–2005. <https://doi.org/10.1002/pro.469> PMID: 20669185
28. Wolny M, Batchelor M, Bartlett GJ, Baker EG, Kurzawa M, Knight PJ, et al. Characterization of long and stable de novo single α -helix domains provides novel insight into their stability. *Sci Rep*. 2017; 7: 44341. <https://doi.org/10.1038/srep44341> PMID: 28287151
29. Goto N, Prins P, Nakao M, Bonnal R, Aerts J, Katayama T. BioRuby: Bioinformatics Software for the Ruby Programming Language. *Bioinformatics*. 2010; 26: 2617–2619. <https://doi.org/10.1093/bioinformatics/btq475> PMID: 20739307
30. Dunnill P. The Use of Helical Net-Diagrams to Represent Protein Structures. *Biophys J*. 1968; 8: 865–875. [https://doi.org/10.1016/S0006-3495\(68\)86525-7](https://doi.org/10.1016/S0006-3495(68)86525-7) PMID: 5699810
31. Wang J, Duncan D, Shi Z, Zhang B. WEB-based GENE SeT Analysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res*. 2013; 41: W77–83. <https://doi.org/10.1093/nar/gkt439> PMID: 23703215

3.1.4. Critical assessment of coiled-coil predictions based on protein structure data

Dominic Simm^{1,2}, Klas Hatje^{1,#}, Stephan Waack² and Martin Kollmar^{1,*}

* Corresponding author

¹ Group Systems Biology of Motor Proteins, Department of NMR-based Structural Biology, Max-Planck-Institute for Biophysical Chemistry, Göttingen, Germany

² Theoretical Computer Science and Algorithmic Methods, Institute of Computer Science, Georg-August University, Göttingen, Germany

Roche Pharmaceutical Research and Early Development, Pharmaceutical Sciences, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd., Basel, Switzerland

Scientific Reports (2021) - Volume 11(12439)

doi: 10.1038/s41598-021-91886-w

<https://www.nature.com/articles/s41598-021-91886-w>

Contribution

DS and MK designed the study with input from KH. DS developed the software, set up the data viewer, and performed data engineering and all computations. KH contributed to database design. DS and MK performed data analysis, designed the figures and drafted the manuscript. SW was involved in statistical analyses. All authors discussed the results and commented on the manuscript.



OPEN

Critical assessment of coiled-coil predictions based on protein structure data

Dominic Simm^{1,2}, Klas Hatje^{1,3}, Stephan Waack² & Martin Kollmar^{1,2}✉

Coiled-coil regions were among the first protein motifs described structurally and theoretically. The simplicity of the motif promises that coiled-coil regions can be detected with reasonable accuracy and precision in any protein sequence. Here, we re-evaluated the most commonly used coiled-coil prediction tools with respect to the most comprehensive reference data set available, the entire Protein Data Bank, down to each amino acid and its secondary structure. Apart from the 30-fold difference in minimum and maximum number of coiled coils predicted the tools strongly vary in where they predict coiled-coil regions. Accordingly, there is a high number of false predictions and missed, true coiled-coil regions. The evaluation of the binary classification metrics in comparison with naïve coin-flip models and the calculation of the Matthews correlation coefficient, the most reliable performance metric for imbalanced data sets, suggests that the tested tools' performance is close to random. This implicates that the tools' predictions have only limited informative value. Coiled-coil predictions are often used to interpret biochemical data and are part of in-silico functional genome annotation. Our results indicate that these predictions should be treated very cautiously and need to be supported and validated by experimental evidence.

Coiled coils consist of two or more α -helices that twist around each other and give rise to a multitude of supercoiled quaternary structures^{1,2}. Coiled-coil regions are characterised by hydrophobic residues at the interface between the supercoiled α -helices and by charged and polar amino acids at the outside. This pattern is usually found in heptads (with the amino acids marked as *abcdefg*) where the hydrophobic residues are located in *a* and *d* positions, but slightly different patterns from hendecad and pentadecad repeats are also observed^{3–7}. The coordinates of the smallest building block, two closely packing α -helices, can be calculated from parametric equations⁸. This might explain why the coiled-coil dimer was likely the first structural element, for which a sequence—structure—function relationship could be established^{9,10}. Accordingly, one of the first tools for predicting protein structure was COILS, which allowed the identification of coiled-coil regions from protein sequences alone¹¹. Coiled-coil structures are claimed to be better understood than those of any other fold^{12,13} and are increasingly used as building blocks in the emerging fields of synthetic biology and de novo protein design^{14–18}. The most advanced design case so far is likely a coiled coil that can switch between pentameric and hexameric states upon pH-change¹⁹. Thus, it seems well possible now to design amino acid sequences forming coiled-coil structures with dedicated oligomeric states.

The complementary problem of detecting coiled-coil regions in amino acid sequences is considered to have been solved as well given the deep biochemical understanding of this structural motif. Even if the oligomeric state is not predicted correctly, it is expected that at least the presence and position of the coiled coil is properly recognized. Multiple prediction programs have been developed using different approaches. COILS¹¹ and its successor NCOILS (COILS2.2)²⁰ match sequences against a fixed length position-specific scoring matrix derived from frequencies at heptad positions. PairCoil²¹ and MultiCoil²² expand this concept by adding pairwise residue correlations to the matrix. PairCoil2 is similar to PairCoil but trained with more coiled-coil sequences²³. Using a different approach, Marcoil calculates posterior probabilities from a windowless hidden Markov model (HMM)²⁴. MultiCoil2 is an advancement to MultiCoil and combines the pairwise correlations with a HMM into a Markov Random Field²⁵. SOSUIcoil uses a unique concept by discriminating coiled-coil regions from other types of regions applying the canonical discriminant analysis²⁶. PCOILS²⁷ is an alternative to NCOILS substituting the

¹Group Systems Biology of Motor Proteins, Department of NMR-Based Structural Biology, Max-Planck-Institute for Biophysical Chemistry, Göttingen, Germany. ²Theoretical Computer Science and Algorithmic Methods, Institute of Computer Science, Georg-August-University Göttingen, Göttingen, Germany. ³Present address: Roche Pharmaceutical Research and Early Development, Pharmaceutical Sciences, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd., Basel, Switzerland. ✉email: mako@nmr.mpibpc.mpg.de

sequence-profile comparison with a profile-profile comparison. The tools CCHMM²⁸ and CCHMM-PROF²⁹ also use HMMs, trained with sequences and profiles, respectively. The latest development, DeepCoil, uses a neural network-based method³⁰. SpiriCoil does not really predict coiled-coil regions, but scores query sequences against protein profiles of the SUPERFAMILY database, which is thought to represent all proteins of known structure³¹, and passes SUPERFAMILY's coiled-coil assignments to the new sequence³². Some tools can also predict oligomeric states of coiled coils, such as LOGICOIL, PrOCoil, Scorer, and RfCoil, but predicting the oligomeric state is still error prone because of the low number of available protein structures with complex coiled-coil arrangements for training the algorithms^{33–36}. In contrast to these many coiled-coil prediction tools, there is a single software, termed SOCKET, that detects knobs-into-holes packing in protein structures³⁷.

Given the many available prediction tools, there are multiple studies indicating high sensitivity and specificity in comparative analyses^{32,38,39}. The benchmark data used, however, were limited by restriction to a selection of SCOP protein families containing coiled-coil regions (SCOP = Structural Classification of Proteins database)^{40,41}, or intersections of SCOP and SOCKET hits. These approaches assessed the sensitivity and specificity against highly restricted data sets and, therefore, strongly overestimated the proportion of true positive coiled-coil predictions. These results do not allow to even estimate the number of false negative cases (no prediction where a coiled coil is present) and false positive predictions (prediction of a coiled coil where there is none) in a representative proteome.

Coiled-coil predictions are part of the standard tool box for in-silico functional genome annotation. In the most extensive comparative study available today, SpiriCoil was used to predict coiled coils in the proteomes of more than 1200 sequenced genomes suggesting that 0.33 to 6.53% of a species' proteins contain at least one coiled-coil region³². A proteome-wide prediction with NCOILS suggested similar proportions⁴². Because most of the proteins predicted to contain coiled coils do not belong to the protein families with known extended coiled-coil regions such as muscle myosin heavy chain and intermediate filament proteins, we wondered how many of these proteins really contain true coiled-coil domains. The most promising approach to evaluate coiled-coil predictions is to compare the predictions with known protein structures. Therefore, we assessed the current status of coiled-coil prediction accuracy by running most of the available coiled-coil prediction tools against all sequences, for which protein structures are known: the entire PDB. Each software was used with default parameters as recommended by the developers and as commonly done in genome annotation pipelines.

Results

Prediction of coiled coils in protein structures and their sequences. To create a ground truth to rely on and compare against, we used SOCKET³⁷, the de facto standard to detect coiled-coil regions within PDB structures. Instead of using SOCKET data generated by the CC+ database⁴³, the Periodic Table of Coiled Coils⁴⁴ or the Atlas of Coiled Coils⁴⁵ we generated our own reference to include the latest PDB release possible and for easier integration with the other prediction data generated. The number of coiled coils that SOCKET might miss is expected to be relatively small compared to the size of the PDB. Such cases could be NMR structures with which SOCKET sometimes has trouble dealing with³². Within 144,270 PDB files (PDB status 12/2018), SOCKET detected 59,693 components (27,803 coiled coils) in 10,684 (7.4%) PDB files (Fig. 1A; Supplementary Table S1). This means that most PDB files with SOCKET hits contain multiple coiled-coil domains, in different copies of the same biological unit, in different regions of the same protein, or in different proteins if biological units consist of multiple different proteins. From all PDB files we extracted 187,021 unique sequences, meaning that exact copies of the same sequence were removed independent of whether they were present in the same or a different PDB file, while slightly different sequences (e.g. longer N- or C-terminus, insertions) remained. With respect to these unique 187,021 sequences, there are 14,117 (7.5%) distinct coiled-coil sequences (components) as found by SOCKET. This occurrence is similar to that of predicted coiled coils in reported analyses of genome annotations³².

To assess the accuracy of coiled-coil predictions from sequences alone, we compared the SOCKET reference set with the results from NCOILS (COILS2.2), PairCoil, PairCoil2, MultiCoil, MultiCoil2, and Marcoil. We did not include SOSUcoil because it is not accessible anymore. The tools CCHMM and CCHMM-PROF were also excluded. A short test for prediction performance of CCHMM-PROF using the globular and coiled-coil free myosin motor domain resulted in many predicted coiled-coil regions, which are obviously not correct (Supplementary Fig. S1A). CCHMM-PROF requires protein profiles as input. SpiriCoil is also only available via a web interface, caused server errors when used, and was therefore excluded. We also refrained from using the tool PCOILS, the successor of NCOILS, because it runs very slowly and is thus not applicable for the amount of data to be analysed. The latter problem is likely the reason why NCOILS is the tool commonly used in genome annotation projects. In addition, PCOILS was found to be less accurate than Marcoil³⁸ and showed the highest overlap with predictions of intrinsically disordered regions⁴⁶. We did not include DeepCoil because running of the software available at the time of performing this analysis (December 2019) was not possible without execution errors. In addition, DeepCoil is limited to sequence lengths of 500 amino acids, which is only slightly larger than the median eukaryotic protein length and much shorter than the length of classical coiled-coil containing proteins such as myosins and kinesins. A test for the prediction performance using a 500 aa region of the myosin-X motor protein⁴⁷ via the DeepCoil web server resulted in mis-prediction of a coiled-coil region at the first IQ motif, which is a calmodulin binding-site, and mis-prediction of a coiled-coil region, where these class-10 myosins contain extended SAH (single α -helix) domains⁴⁸ (Supplementary Fig. 1B). The 500 aa query limit is removed in DeepCoil2 (v. 2.0.1., 30 Nov 2020), but the available (and maybe not completely finished) version of DeepCoil2 does not predict any coiled-coil region in mouse MyoX (Supplementary Fig. 1B; in sharp contrast to DeepCoil "v.1") and no coiled-coil regions in many of the classical coiled-coil proteins such as the muscle and non-muscle myosin heavy chain proteins (Supplementary Fig. 2). While DeepCoil2 might perform very well

3. Publications and Manuscripts

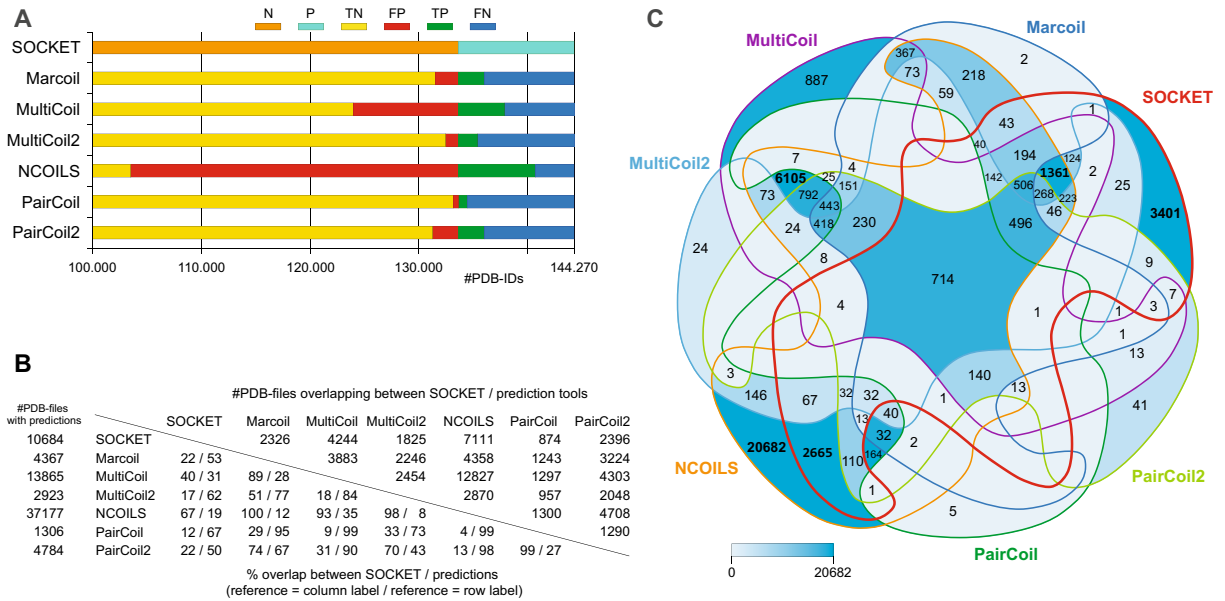


Figure 1. Coiled-coil regions identified by SOCKET and predicted with the respective tools at the level of PDB files. **(A)** Of the 144,270 PDB files 10,684 contain a coiled coil according to SOCKET (P = positives; the remaining 133,586 PDB files represent the negatives = N). Based on the overlap with these SOCKET hits, the coiled-coil predictions of the tools were categorized into four classes: true positives (TP = predictions in same PDB files as SOCKET hits), false positives (FP = predicted coiled coils in PDB files where SOCKET did not detect any), true negatives (TN = no prediction and no SOCKET hit in PDB files), and false negatives (FN = no prediction in PDB file in which SOCKET identified a coiled coil). **(B)** The tools predicted coiled coils in 1307 (PairCoil) to 37,177 PDB files (NCOILS; left column). The numbers in the matrix denote the overlap between SOCKET and every tool, and between any two tools, at the level of hits within the same PDB file (upper part: total numbers; lower triangle: percentage overlap with respect to SOCKET or tool). **(C)** The 7-way Venn diagram shows the subsets of PDB files with SOCKET hits and predicted coiled coils found by the respective combinations of tools colored by number in intersection. The intersection of PDB files with coiled coils predicted by all tools is 714 PDB files, and 1210 PDB files when ignoring PairCoil, the tool with the least predictions.

on PDB data (it was trained on 90% of all SOCKET hits from the July-2020 version of PDB), DeepCoil2 fails on the tested non-PDB sequences from myosins. In its current status, DeepCoil2 is not a fair competitor in an exclusively PDB-based benchmark study.

To best simulate a common functional protein annotation, we used default parameters for each tool as recommended by the developers, except for setting 21 amino acids as sliding window in all tools for comparability (21 is default in NCOILS).

Comparing coiled-coil predictions across PDB files. Before investigating the performance of the coiled-coil prediction tools we wanted to get a first glimpse on whether the tools predict coiled coils in the same structures or in different structures. This should result in all possible intersections of reference and prediction tools. Therefore, we analyzed whether coiled coils are found by SOCKET and the prediction tools in sequences of the same PDB file. For this question, only part of a coiled coil (e.g. only one of the sequences in a heterodimeric coiled coil) needs to be identified to classify a PDB file as “coiled coil present”. In addition, this approach ignores whether coiled-coil predictions overlap between tools and the SOCKET reference, and whether sequences contain one or more coiled-coil regions. Accordingly, by this very simplified approach the intersection between reference and predictions is highly overestimated (the tools look better than their results are). In this scenario, there are 10,684 PDB files containing at least one coiled coil found by SOCKET. Surprisingly, the various tools predict coiled coils in strikingly different total numbers of PDB files with PairCoil predicting coiled coils in fewest (1307) and NCOILS in most PDB files (37,177; Fig. 1B and Supplementary Table S1). 33.1% (PairCoil) to 80.9% (NCOILS) of the tools’ predictions were found in PDB files where SOCKET did not find any hit. SOCKET hits are exclusive in 3401 PDB files (31.8% of all SOCKET hits).

Ignoring PairCoil, with which by far the fewest coiled coils were predicted, the minimum overlap of PDB files with coiled-coil predictions from any two tools is 2048 (overlap of MultiCoil2 and PairCoil2 predictions; Fig. 1B). Although this number suggests considerable overlap of predictions in the same PDB files, the opposite is found (Fig. 1C, Supplementary Fig. S3). The predictions overlap by only 11.6% (PairCoil) to 66.6% (NCOILS) with the PDB files containing SOCKET hits (Fig. 1B) indicating that the tools did not predict any coiled-coil regions in the vast majority of the PDB files where SOCKET identified coiled coils. The intersection of PDB files

with SOCKET hits and coiled coils predicted by all tools is only 714 PDB files. This number increases to 1210 PDB files if PairCoil, the tool with the fewest predictions, is ignored. All tools predicted coiled coils in 230 PDB files (648 when ignoring PairCoil) where SOCKET did not identify any, potentially indicating structures outside SOCKET's default cut-off and structures difficult to resolve by SOCKET (e.g. some NMR structures). 9399 PDB files contained coiled coils predicted by at least two tools but did not contain SOCKET hits. This number is considerably higher than that of SOCKET hits overlapping with at least one of the tools (7283 PDB files). While almost all PairCoil and Marcoil predictions overlapped with SOCKET hits or predictions of at least one other tool, PairCoil2, MultiCoil, MultiCoil2, and NCOILS exclusively predicted coiled-coil regions in 41 (0.9% of PairCoil2 predictions), 887 (6.4%), 24 (0.8%), and 20,682 (55.6%) PDB files, respectively (Fig. 1C). Irrespective of the individual performance of each coiled-coil prediction tool, these intersection data demonstrate that the tools predict coiled-coils in very different sequences.

No effect of sequence redundancy on binary classification metrics. To evaluate the performance of the coiled-coil prediction tools, we analysed their overlap with SOCKET hits. Of the 187,021 unique sequences in the PDB files, 14,117 contain SOCKET hits, which are defined as positives here. For simplicity, we required a single amino acid overlap between SOCKET hit and coiled-coil prediction for the prediction to be classified as "true positive". Redundancy in the data is only a problem, if it does not apply to all binary classification categories similarly. Therefore, every filter based on a user-defined criterion, for example selecting proteins whose SCOP family contains coiled-coils, excluding all-beta-strand proteins, or preferentially selecting coiled-coil containing proteins from clusters of otherwise similar proteins, would introduce a bias on the data set taking effect on only the positives or the negatives. Previous comparative analyses used such filters and it has not been investigated whether these filters influenced the performance metrics (Supplementary Notes). To exclude that sequence redundancy in the PDB influences coiled-coil tool evaluation we reduced the redundancy of the 187,021 unique sequences with CD-HIT⁴⁹ applying 90%, 70% and 50% sequence identity cut-offs resulting in 49,311, 39,394 and 30,397 unique sequences, respectively. The drastic reduction by 74% in sequence space from no redundancy to 90% sequence identity nevertheless did not result in considerable changes in the performance of the binary classification (Fig. 2, Supplementary Table S2). The sensitivity of the prediction tools increased by two to five percent and the corresponding miss rates decreased by the same numbers. All other metrics are almost identical. Most notably, there is no change in any of the performance metrics if the sequence redundancy is further decreased from the 90% to 70% and 50% sequence identity. Although values over 90% for several metrics such as specificity and accuracy indicate strong performance of the coiled-coil prediction tools (except for NCOILS), a close look at other metrics demonstrates that the overall performance is instead close to random and might even strongly misguide interpretation of bioinformatics analyses and biological experiments.

Performance of the coiled-coil prediction tools. All described performance metrics measure the classification quality of either of true and false positives and negatives, and all rely on the characteristics of the data set, i.e. the proportion of positives and negatives. Because the benchmark data set is the entire PDB and all its unique sequences, the number of negatives is much higher than the number of positives (6.2% positives at 90% sequence identity, 7.6% at 100% identity). Accordingly, the specificity and accuracy of the tools are very high, while sensitivity and precision are rather low. It is obvious that the performance of the tools cannot be evaluated just based on these metrics. In case of imbalanced data, the Matthews Correlation Coefficient (MCC) represents the best overall measure to evaluate the performance of binary classifiers. The MCC score ranges from -1 (totally wrong classification, or perfect classification of the opposite) to 0 (random classification) to $+1$ (perfect classification)⁵⁰. Here, requiring overlap of only a single amino acid between SOCKET hit and coiled-coil prediction, the MCC indicates random prediction in case of NCOILS (MCC of 0.02) and close to random prediction for all other tools (MCC of 0.22 for MultiCoil2 being the highest value; Fig. 2). It is important to note that also the MCCs are independent of sequence redundancy reduction.

While random prediction, by wording, suggests a flipped coin chance to have a coiled coil in a sequence when tools predict one, there are two other metrics very important for the experimental biologist. The false discovery rates (FDR), which denote the percentage of false predictions compared to all predictions, show that the actual percentage of false predictions is considerably higher than random, with 83% for MultiCoil and 91% for NCOILS (Fig. 2). For the other coiled-coil prediction tools, the chance of having predicted a true coiled coil is slightly better than flipping a coin (precision of 36–54%, Fig. 2). For the experimental biologist this means that the chance is higher that a coiled coil predicted with one of the many available NCOILS web server is in fact not a coiled coil than the chance that the predicted coiled coil might really be present in the protein. The other important metric is the miss rate, which denotes the percentage of elements in the reference data that were not predicted. The analysis shows that the prediction tools missed from 59 to 63% (NCOILS; range across the redundancy reduced data sets) to 92–93% (PairCoil) of the SOCKET reference coiled coils. For the experimental biologist this means that chances are considerably higher than flipping a coin that a coiled coil is present in a sequence of interest where the prediction tools did not predict any.

Performance of the coiled-coil prediction tools compared to naïve models. Because values over 90% for some of the performance metrics such as specificity and accuracy indicate good performance we compared these with the results of three naïve classification models (Fig. 3). First, we calculated the same metrics assuming that all sequence is classified as coiled coil. Second, we used a coin flip model where half of the cases are coiled coils and the others are not. Third, we assumed that we know the proportion of true coiled coils in the data and randomly predict coiled coils with the same proportion. While the first model assumes an extreme case and the second would provide a reasonable baseline for a balanced data set, the third model provides a good

3. Publications and Manuscripts

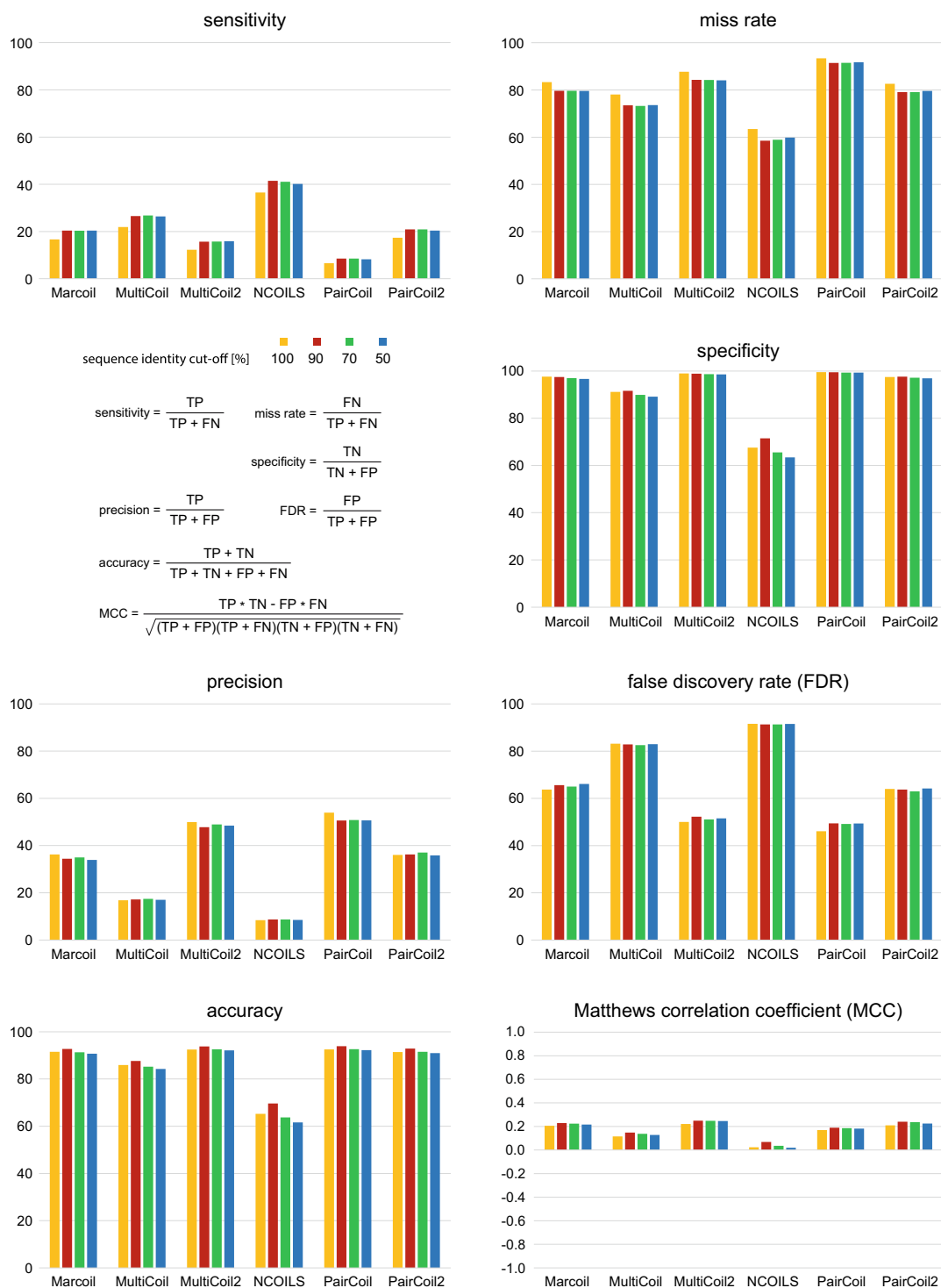


Figure 2. Performance of coiled-coil prediction tools in dependence of PDB sequence redundancy. The performance was analysed for four data sets with decreasing levels of sequence redundancy. With respect to these data sets, the sequences containing a coiled coil according to SOCKET are classified as positives while the remaining sequences represent the negatives. Based on the minimum requirement for overlap with these SOCKET hits (a single amino acid overlap), the coiled-coil predictions of the tools were categorized into four classes: true positives (TP = predictions overlap SOCKET hits), false positives (FP = predicted coiled coils do not overlap SOCKET hits), true negatives (TN = no prediction and no SOCKET hit in sequence), and false negatives (FN = no prediction in sequence region where SOCKET identified a coiled coil). The plots show the performance of the coiled-coil prediction tools based on commonly used statistical measures. A Matthews correlation coefficient (MCC) of +1 indicates a perfect prediction, predictions with MCCs around 0 are no better than random, and a MCC of -1 represents total disagreement between prediction and reference.

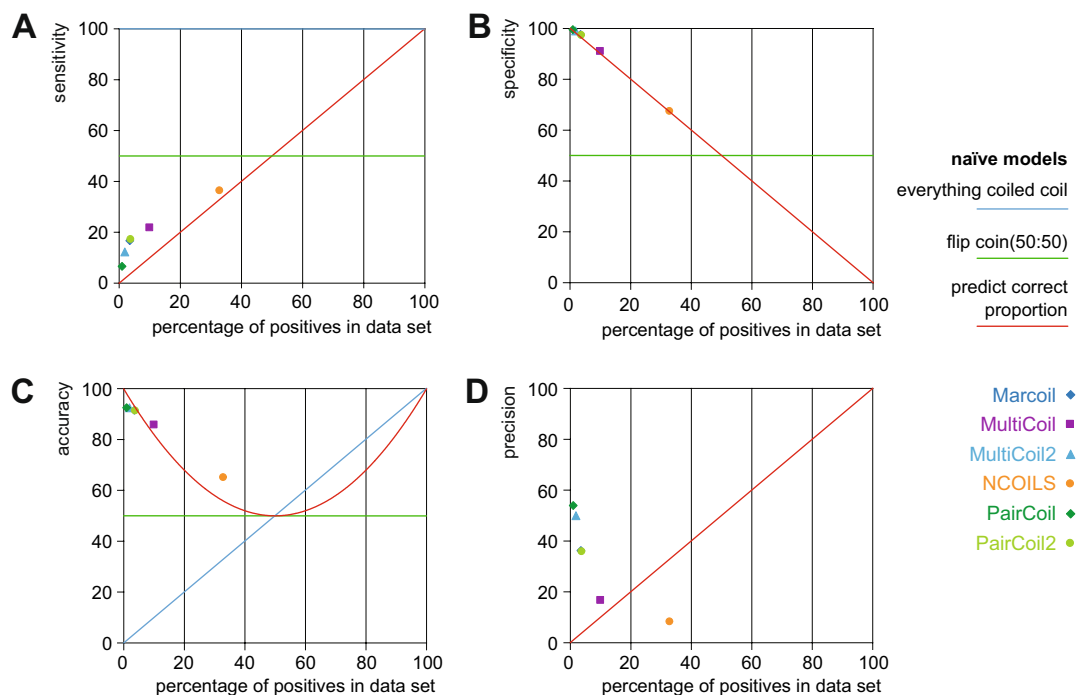


Figure 3. Performance of coiled-coil prediction tools compared to three naïve models. The plots present statistical measures based on three naïve models as described in the legend. The respective performance of the coiled-coil prediction tools based on the unreduced data is plotted for comparison.

baseline for the coiled-coil prediction tools. Therefore, the base level to estimate the performance of each prediction tool would be the performance obtained if the identical proportion were predicted randomly. Compared to this naïve model, the sensitivity of all prediction tools is slightly, but not considerably better (Fig. 3A). The specificity is almost exactly identical to the specificity obtained if the same number of coiled coils were predicted randomly (Fig. 3B). The accuracy of Marcoil, MultiCoil2, PairCoil and PairCoil2 is lower than in the naïve model, while that of MultiCoil and NCOILS is slightly higher (Fig. 3C). The precision of the tools is highest for the tools with the lowest number of predictions (Fig. 3D). The precision of all tools is considerably better than the naïve model, except for NCOILS, which is considerably worse. From all these metrics it is clear that the coiled-coil prediction tools do not significantly outperform the naïve model of predicting the same proportion randomly.

Effect of a length cut-off filter on coiled-coil prediction performance. Because SOCKET might identify considerably more short coiled coils than prediction tools do, two sequence length cut-offs were applied and reference SOCKET hits and coiled-coil predictions shorter than 14 amino acids and 21 amino acids were excluded from each data set (Supplementary Fig. S4, Supplementary Table S2). As discussed above, this is not a scientific decision based on data that show that short coiled coils do not exist but a subjective decision based on the observation that coiled-coil prediction tools perform bad in predicting these types of coiled coils. With respect to reducing redundancy of the PDB sequence space the two cut-offs don't have any effect. At each cut-off, the tools show the same performance for the 100, 90, 70 and 50% sequence identity data sets (Supplementary Fig. S4), indicating again that reducing redundancy is not necessary for evaluating benchmark studies based on PDB sequence data. However, the prediction tools show increasing performance when applying the 14 and then the 21 amino acid cut-offs compared to no length cut-off (Supplementary Fig. S4). The sensitivities increase considerably, and this is the main reason for the increased MCCs, although the values are below 0.4 for all tools and all cut-offs. The increase in sensitivity is a direct consequence of the cut-offs excluding the short coiled coils that the tools rarely predict. NCOILS highly overpredicts coiled coils, and therefore shows the highest specificity and lowest miss rate. The miss rates for the other tools are still above 50% in case of the 14 amino acid cut-off, and decrease to 34% (MultiCoil) to 71% (PairCoil) for the 21 amino acid cut-off (Supplementary Fig. S4). The increase in sensitivity comes to the cost of precision, which decreases by 5–17%. Accordingly, the false discovery rates increase by the same numbers. This analysis demonstrates that sequence redundancy in the PDB does not matter, but, of course, applying a cut-off or filter on the positives does. The results show the performance of the tools with respect to predicting long coiled-coils, and not with respect to predicting coiled coils in general.

Extent of structural overlap between coiled-coil predictions and SOCKET hits. In contrast to simple absence/presence counting, evaluation of overlap is getting more difficult the shorter the regions and peptide segments are and the more complex the overlapping patterns become. Overlap patterns such as length

and/or contiguity of the predictions depend on prediction tools and parameters. Therefore, considerable care must be taken that the scoring scheme for evaluation does not prefer one over the other pattern. For example, the muscle myosin heavy chain proteins are well known to assemble into homo-dimers based on their long, coiled coil forming tail domains. Coiled-coil predictions on a human adult skeletal muscle myosin heavy chain protein sequence result in different sets of coiled-coil regions with different start and end positions, different switches in the predicted heptad registers, and different predictions of oligomerisation state from dimer and trimer to tetramer (Supplementary Fig. S5). It is obvious that predictions of long, uninterrupted coiled-coil regions should be preferred, and that breaks in these very extended coiled-coil regions might be attributed to over- and under-winding of the twisted α -helices. This is very different in the opposite case, for example α -actinin, which forms intra-sequence coiled coils by folding into helix-loop-helix segments. There, the prediction of interrupted coiled coils would be highly preferred over single long, uninterrupted coiled coils. Coiled-coil prediction tools cannot distinguish between intra-sequence and inter-sequence coiled coils.

A method to evaluate very short patterns is the segment overlap score (SOV) that has been developed to compare known and predicted secondary structural elements^{51–53}. In contrast to a per-residue score, which judges the percentage of individual overlapping positions, the SOV positively weights contiguous longer overlapping segments and reduces the influence of considered less significant features such as slightly different segment lengths and/or positions. In the current version from 1999⁵², contiguous segments are rated higher than multiple shorter segments when compared to a long reference segment even if the total number of positions is considerably lower, compared to the initial version from 1994⁵¹. However, it is not clear why there is a strong difference in weighting between the case where the reference consists of multiple segments and the prediction is contiguous and the case where the reference is contiguous and the predictions are multiple segments, why additional splits in overlapping segments do not contribute linearly to down weighting (and why they contribute differently with respect to a contiguous reference or prediction), and why false positives of different features (e.g. α -helix versus β -strand) are less disfavoured than false positives of no feature. Because coiled coils are intermediate between single large features and (potentially) multiple short segments, we decided to apply a percentage overlap to each coiled coil. This approach slightly favors the coiled coil predictions, which are usually longer than the reference because of the window-based prediction algorithms, and does not average coiled-coil evaluations in case sequences contain multiple reference coiled coils and/or predictions.

Considering possible tool-dependent bias in determining start and end positions of the predictions we determined the number of predictions overlapping SOCKET hits in dependence of reference (either SOCKET or tool) and degree of overlap (Fig. 4, Supplementary Fig. S6, and Supplementary Table S3). At least, each predicted coiled-coil sequence should overlap with a single amino acid of a SOCKET region. At that minimum level only 19,036 (26.7%) of the 71,393 coiled-coil regions predicted by NCOILS and found by SOCKET in the same PDB files overlap. This indicates that the majority of the NCOILS predicted coiled coils do not overlap with SOCKET hits, although SOCKET hits and predicted coiled coils were found in the same PDB file. The percentage of overlapping regions (single amino acid criterion) is highest for MultiCoil2 and SOCKET hits (5590 of 8346 regions, 67.0%; Fig. 4 and Supplementary Fig. S6). Taking the prediction tools as reference and requiring overlap of at least 50% of their predicted sequence regions with SOCKET hit regions, only 13.4% (MultiCoil) to 34.0% (PairCoil) of the tools' predictions match this criterion. When requiring at least 80% overlap between predicted coiled-coil and SOCKET hit regions, the fraction of overlapping regions decreases further to 5.4% (NCOILS) to 17.5% (PairCoil). When taking SOCKET hits as reference, 23.6% (NCOILS) to 65.0% (MultiCoil2) of the predictions overlap with at least 50% of respective SOCKET hit regions. The percentages of overlapped SOCKET hit regions only slightly decrease with increasing size of overlap when using SOCKET hits as reference. This shows that predicted coiled-coil regions are in general considerably longer than SOCKET hit regions and therefore reach into protein structural regions without knobs-into-holes packing indicative of coiled-coil helix-helix interactions.

Requiring more overlap between reference and prediction decreases prediction performance. In the performance analyses shown so far, only a single amino acid overlap was required. By this minimum requirement, many predictions matching unrelated structural regions were also considered true positives although they are, by inspecting the structures, in fact false positives. Because SOCKET hits are usually shorter than coiled-coil predictions, using the SOCKET hits as reference for the comparison of the overlap will result in better performance metrics for the prediction tools. Accordingly, we increased the required overlap in steps of 5% for all data sets described before, the four data sets with decreased sequence redundancy and the data sets with additional coiled-coil length cut-off. With respect to reducing redundancy of the PDB sequence space changing the required overlap shows a marginally effect on the performance metrics (Supplementary Figs. S7 and S8; Supplementary Table S4). At each overlap proportion, the tools show almost the same performance for the 100, 90, 70 and 50% sequence identity data sets. However, the performance decreases when increasing the required overlap (Supplementary Fig. S8).

Patterns of coiled-coil predictions. The heptad pattern characteristic for almost all coiled-coil regions is found in all SOCKET regions (Fig. 5 and Supplementary Fig. S9). There is a clear preference for leucines in *d* positions compared to *a* positions, and increased propensity for isoleucines and valines in *a* positions compared to *d* positions. There is strong discrimination against glutamates in *a* positions and lysines and arginines in both *a* and *d* positions. However, there are significantly more leucines in *e* and *g* positions than in *b*, *c*, and *f* positions, and there is no discrimination against glutamates in *d* positions (Fig. 5). SOCKET might also detect knobs-into-holes packed α -helices buried within e.g. globular structures, which might not be detected by coiled-coil prediction tools. However, the amino acid distributions at heptad positions were almost identical between SOCKET hits, which overlap regions where tools also predict coiled coils, and SOCKET hits where tools did not identify

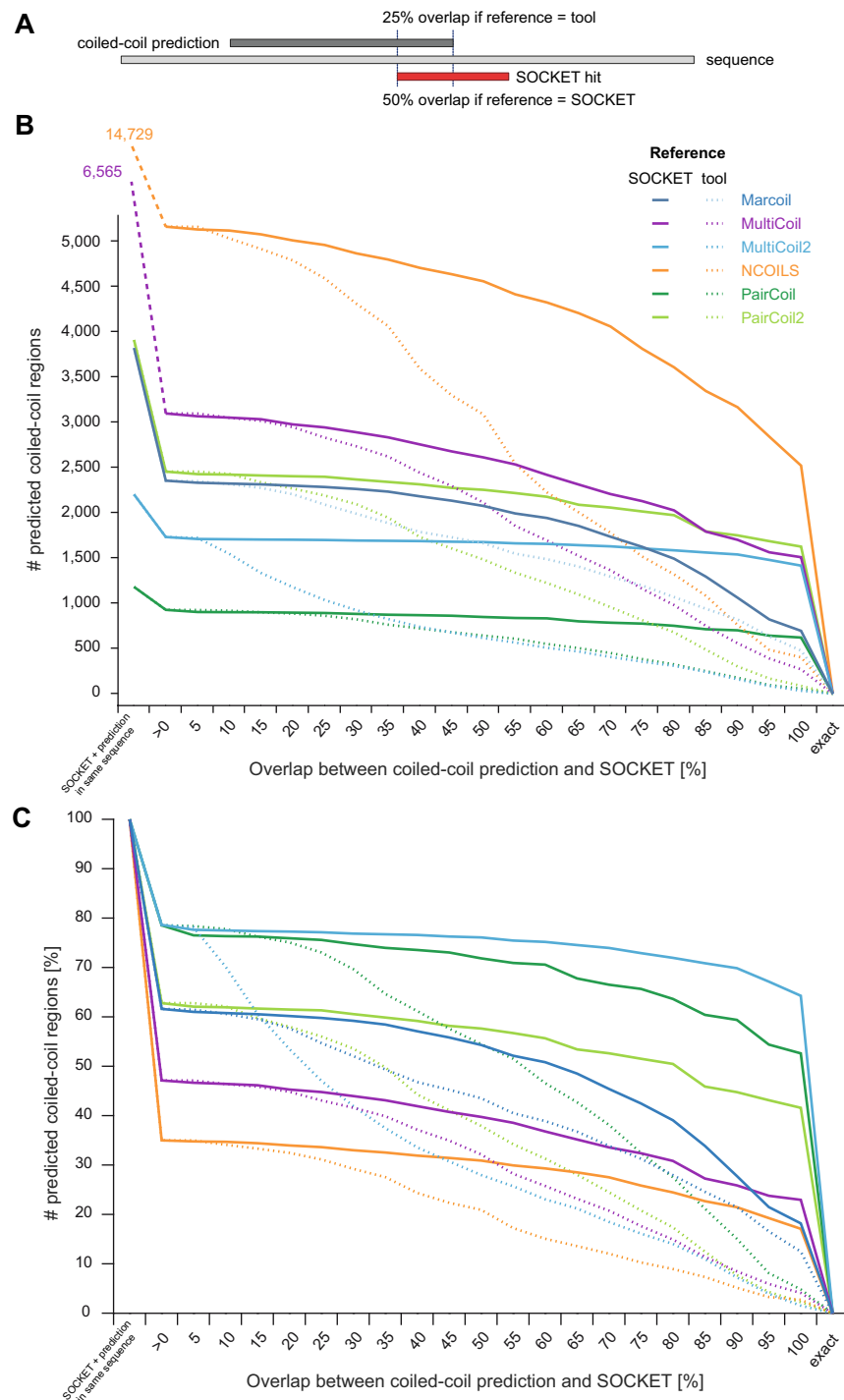


Figure 4. Overlap of coiled-coil predictions with SOCKET hit regions. **(A)** Schematic drawing of a coiled-coil prediction overlapping a SOCKET hit. The ratio of overlap between prediction and SOCKET hit is different depending on whether the prediction or the SOCKET hit is taken as reference. **(B)** The plot shows the total number of coiled-coil predictions in dependence of the degree of overlap with the SOCKET hits. The values at the left side indicate the number of PDB sequences containing both a SOCKET hit and a coiled-coil prediction. For each tool, the number of overlapping hits is counted once with taking the coiled-coil prediction as reference (solid lines) and once with the SOCKET hits as reference (dashed lines). **(C)** This plot is similar to **(B)** but shows the percentage of overlapping regions with respect to the overlap ratio. The number of sequences containing both a coiled-coil prediction and a SOCKET hit is set to 100% for each tool. A similar plot with the number of predictions overlapping a SOCKET hit with at least a single amino acid set to 100% is shown in Supplementary Fig. S4.

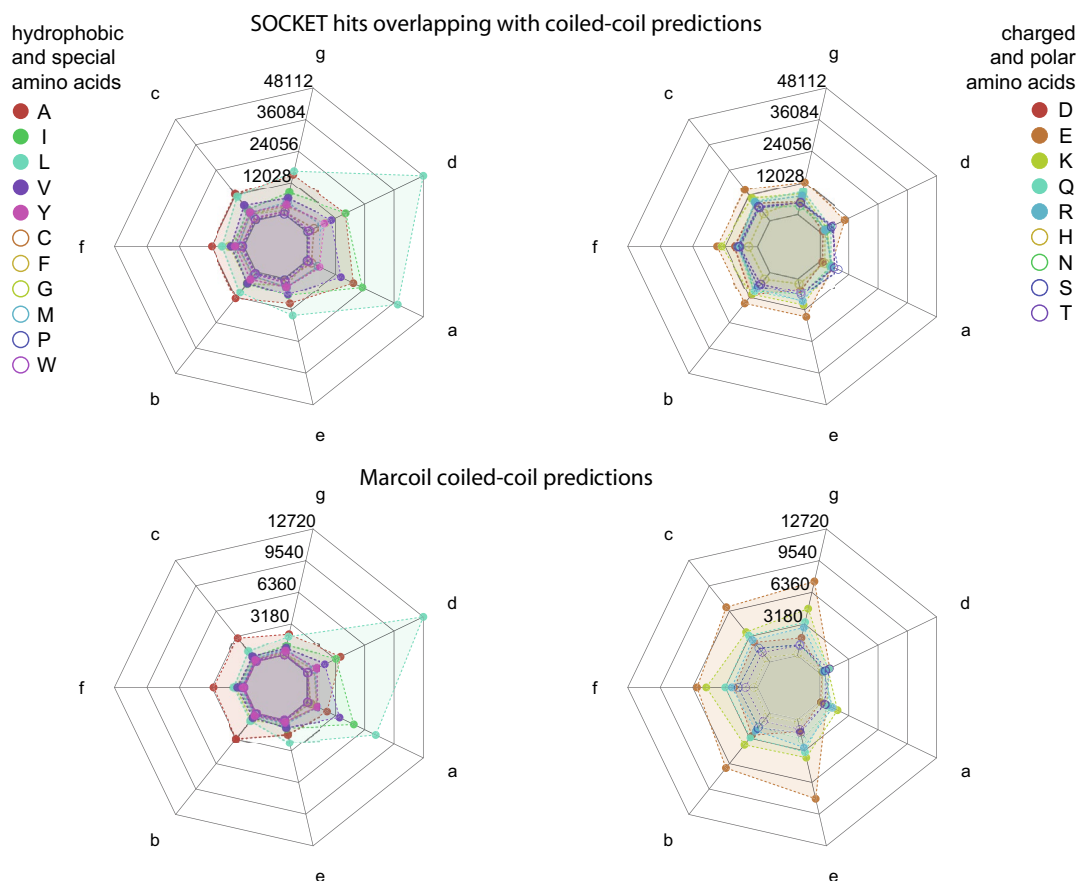


Figure 5. Amino acid preferences at heptad positions *abcdefg*. The spider plots show the distribution of amino acids at each heptad position. Hydrophobic and special amino acids are shown on the left and hydrophobic and polar amino acids on the right for better orientation. SOCKET not only detects “classical” coiled coils but interacting α -helices within globular protein structures. To reveal possible differences in the amino acid distributions among these two groups, SOCKET hits overlapping (shown here) and not overlapping (Supplementary Fig. S4) coiled-coil predictions were analysed separately. The distribution of hydrophobic and charged amino acids is slightly less biased in the latter structures. Marcoil predictions show strong bias for leucine and isoleucine at the interior positions *a* and *d*, and for glutamate at all other positions. The heptad patterns of the other predicted coiled coils show similar distributions. The letters at the axes denote the heptad register positions. Data values at grid lines refer to amino acid counts at each heptad position over all heptads. Amino acids with high proportions are shown as filled circles and amino acids with low proportions as unfilled circles.

any (Fig. 5 and Supplementary Fig. S9). The strongest difference between the two patterns is the slightly lower preference for leucines and less discrimination against charged residues. This comparison indicates that from the perspective of amino acid distribution at heptad positions the SOCKET-determined coiled coils in the 3401 PDB files, where the tools did not predict any coiled coil, are not very different from the SOCKET hits overlapping coiled-coil predictions. The patterns of the predicted coiled-coil regions are very similar with respect to each other and to the SOCKET pattern although more discriminating against glutamates and lysines in *a* and *d* positions (Fig. 5).

The different curve shapes for SOCKET hits overlapping predictions and predictions overlapping SOCKET hits (Fig. 4) already showed that coiled-coil predictions are longer regions than SOCKET hits. This is supported by the length distributions of the coiled-coil regions (Fig. 6). Most SOCKET hits are 10 to 19 amino acids long, while most Marcoil, MultiCoil and NCOILS regions are 20 to 29 amino acids long, and most PairCoil and PairCoil2 regions are 30 to 39 residues long. MultiCoil2 regions show a very different length distribution with no specific preference for a certain length. Instead, MultiCoil2 seems to combine consecutive helices, e.g. the repeat regions of helical bundle forming proteins such as spectrin and α -actinin, into single super-long coiled-coil regions.

Coiled-coil predictions map to all types of secondary structural elements. The considerably deviating matchings of SOCKET-determined coiled-coil regions and predicted coiled coils with respect to the PDB

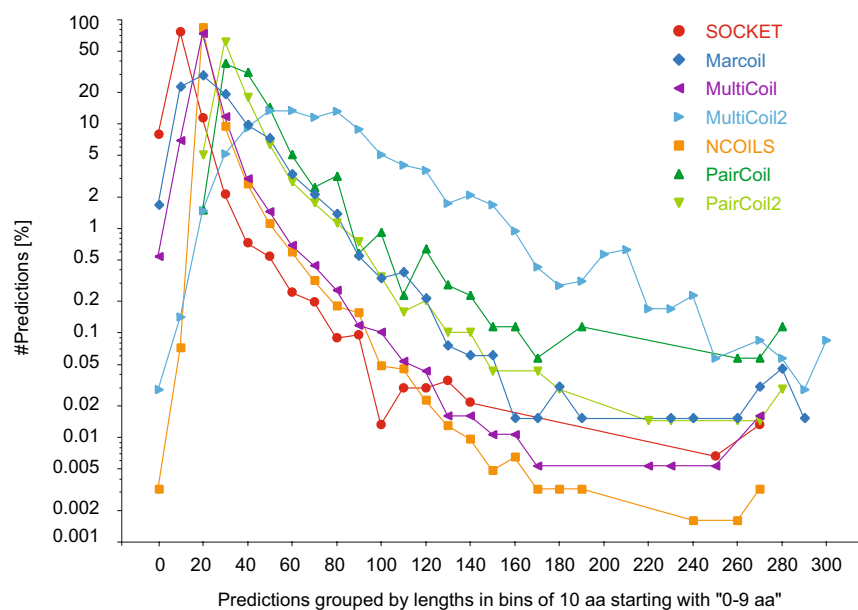


Figure 6. Length of SOCKET hits and coiled-coil predictions. SOCKET hits and coiled-coil predictions were grouped by length in bins of ten amino acids, and the number of hits/predictions in each bin plotted in percent with respect to the total number of hits/predictions of the respective tool.

structures prompted us to look at the secondary structural elements of matched regions. As ground truth for the secondary structure we relied on DSSP (Define Secondary Structure of Proteins)^{54,55}. DSSP assigns secondary structure elements to amino acids based on hydrogen-bonded and geometrical features extracted from X-ray coordinates and is the standard tool at the Protein Data Bank (RCSB PDB) for assigning secondary structures. As expected by SOCKET's algorithm, which selects α -helical regions assigned by DSSP and uses these to detect knobs-into-holes packing in the protein structures, 99.994% of the amino acids within SOCKET hits match to α -helical regions (H in DSSP notation) while the remaining 0.006% match to loops ("blank "; Fig. 7A and Supplementary Table S5). In contrast, 20.8% (Marcoil) to 31.5% (NCOILS) of the regions predicted to be coiled coils do not fall into α -helices, and considerable parts of MultiCoil (2.6%), MultiCoil2 (2.6%), and NCOILS (5.0%) predictions match to β -strands (Fig. 7A and Supplementary Table S5). To allow visual inspection of these rather surprising results and detection of potential mis-assignments or systematic deviations we implemented a web-interface to the analysis database providing a search interface, a structure viewer, and sequence-based representations of all predictions in comparison. The web-interface can freely be accessed at <https://waggawagga.motorprotein.de/pdbccviewer>.

Parallel and antiparallel coiled coils, and oligomeric states. Coiled coils can have parallel and antiparallel arrangements of the α -helices and take part in many different oligomeric assemblies⁴⁴. The "classical" coiled coil is a parallel homodimer, and accordingly the protein training data of the first prediction tools consisted of parallel homo- and heterodimers such as myosins, tropomyosins, kinesins, and intermediate filament proteins. In the PDB by far the most detected arrangement is the antiparallel 2-stranded coiled coil (72.1%) followed by the parallel 2-stranded coiled coil (14.1%; Fig. 7B). NCOILS' predictions have the largest overlap with SOCKET hits and thus the most similar distribution of arrangements with SOCKET. PairCoil shows the strongest bias of all prediction tools for detecting parallel 2-stranded coiled coils. Marcoil, MultiCoil, MultiCoil2, and PairCoil2 all have similar distributions with bias towards parallel 2-stranded and 3-stranded coiled coils (Fig. 7B). These data show that the prediction algorithms do not exclude certain arrangements, be it the direction or the number of involved α -helices.

The polyglutamine puzzle. It has been suggested that (Q/N)-rich prions and polyQ-expanded proteins form coiled-coil structures based on the polyQ and neighbouring regions⁵⁶. There are 23 structures now in the PDB containing stretches of at least six glutamines (Table 1). In only one of these structures, 3PJS, the polyQ region is the α -helical extension of a tetrameric coiled-coil structure, while in the others these regions do not take part in any oligomerisation. In 3PJS, the polyQ region is part of an α -helix, but this region does not interact with any neighbouring region. The polyQ region is therefore not part of a coiled coil but part of a single α -helix (SAH). In addition, the polyQ region is not present in any natural sequence of this protein class but the result of multiple mutations to facilitate structural and biochemical analyses. In five structures, the polyQ region is the C-terminal end and extends as SAH domain into the solvent. In the remaining structures, the first two to four glutamines succeeding an extended α -helix often extend this α -helix while the remaining glutamines are mostly

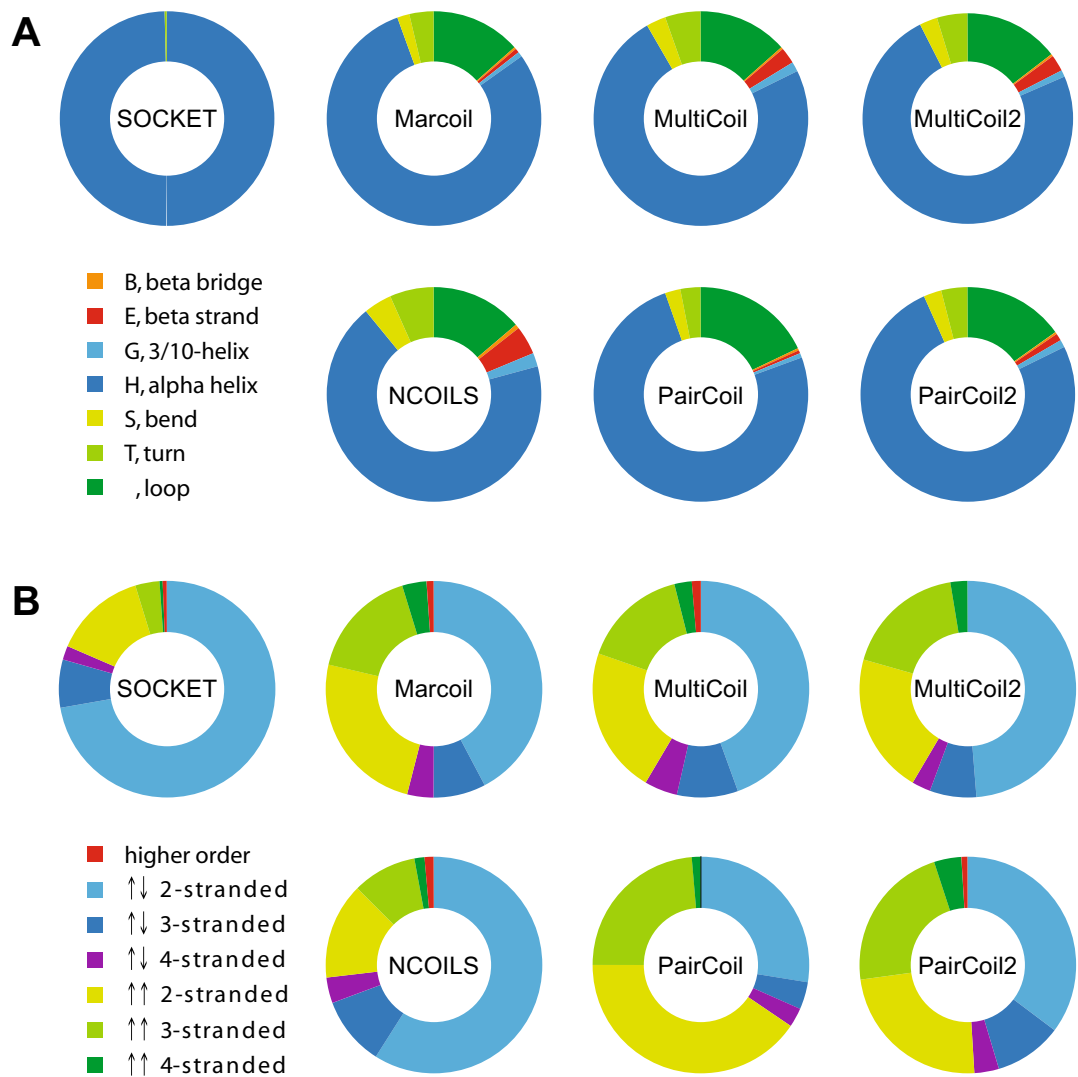


Figure 7. SOCKET hits and coiled-coil predictions matching protein structures. (A) The plots represent the matching of all SOCKET hits and all coiled-coil predictions with secondary structure elements as determined by DSSP. The secondary structure assignment for each amino acid was read from the DSSP output, the assignments summed up for each element, and the distribution of elements determined for each tool in percent. (B) Coiled-coil predictions on sequence alone are not biased for certain oligomeric states. As reference, the distribution of oligomeric states as determined by SOCKET is shown, with parallel and antiparallel arrangement of the α -helices separated. For each coiled-coil prediction tool, only those predictions were selected that overlap at least 50% of a SOCKET hit, and the oligomeric state assignments of the respective matched hits were collected.

flexible and not visible in the crystal structures. Thus, there is no structural support for polyQ-regions forming coiled coils yet.

In the initial report of (Q/N)-rich and polyQ regions in human proteins forming coiled-coil structures in vitro, the coiled-coil propensities of these regions were predicted with COILS and PairCoil⁵⁶. However, it has already been reported in 1995 and was described as major intention to develop a new algorithm, that COILS predicts coiled coils in homopolymers of charged amino acids²¹. This did not change with the second and current COILS version, NCOILS. NCOILS predicts coiled coils in homopolymeric polyK, polyE, polyN, and polyQ peptides, and in polyR and polyA peptides if additional amino acids are inserted somewhere in the homopolymer. A single leucine within polyR is enough to turn this homopolymer into a “coiled coil”, and for polyA to become a “coiled coil” two glutamates or a glutamate and a lysine are needed. The latter polyA[+2E/EK] is also predicted to be a “coiled coil” by Multicoil. Marcoil also predicts “coiled coils” for homopolymeric polyK, polyE, and polyQ peptides. These predictions can easily be reproduced by the reader using Waggawagga, a webserver for the comparative analysis and visualisation of coiled-coil predictions of the most common coiled-coil prediction software packages⁵⁷. Thus, the prediction of coiled coils for such homopolymers is rather an artefact of these software.

3. Publications and Manuscripts

| PDB id | Poly“X” stretch | Observed structure |
|--|---|---|
| 1U6F | QLQQLQ ₆ | Turn, bend |
| 2DMS | Q ₇ | Turn, bend |
| 2NB1, 4A9Z | Q ₆ HQ | Helical, SAH |
| 2OTU, 2OTW | Q ₁₀ G | Turn, bend |
| 1QB3 | Q ₁₆ HQ ₁₀ TQ | No structure |
| 3IO4, 3IO6, 3IOR, 3IOT, 3IOU, 3IOV, 3IOW | Q ₁₇ | Multiple conformations, mainly turn and bend, partially helical |
| 3PJS | QE ₆ Q ₆ | Helical, SAH |
| 5LTY, 6ES2, 6ES3 | Q ₆ | Helical, SAH |
| 4FE8, 4FEB, 4FEC, 4FED, 4WTH | Q ₇ HQH ₁₀ HQ ₂₇ | Helical, turn, β-hairpin |
| 1QBK | ED ₆ EID ₄ | Disordered, turn, helical, SAH |
| 2WGO | D ₆ | Disordered |
| 5GAP, 5GAN | DIDEVD ₆ | Turn, beta-bridge |
| 5GRQ | D ₆ ND | Bend, disordered |
| 1AP0, 1GUW | E ₆ | Turn, disordered |
| 1L0L, 1NTK | E ₈ | Helical, SAH |
| 1MHS | EDDEDDEDID, E ₆ | Bend, disordered |
| 2MKF, 2MKG, 2RR9 | QE ₆ | Helical, SAH |
| 2XZE | EDE ₆ | Helical, coiled coil with oppositely charged (multiple K) α-helix |
| 6EYC, 3JA8, 5BK4, 3JC5, 3JC7, 5XF8, 5H7I, 5U8S, 5U8T, 6F0L, 6HV9 | E ₆ | Helical, orthogonal to another helix |
| 5A6C | NE ₆ D | Turn |
| 5E26 | E ₆ | Helical, SAH |
| 5IY6, 5IY7, 5IY8, 5IY9, 5IVW | DKDE ₆ | Turn, disordered |
| 5XIS, 5XIT, 5YDK | E ₆ | Helical, SAH |
| 5XTE | E ₈ | Helical, SAH |

Table 1. Crystal structures of proteins containing stretches of at least six glutamines, aspartates or glutamates.

Inspection of stretches of at least six consecutive aspartates and glutamates in protein structures supports this conclusion (Table 1). Homopolymers of at least six asparagines are not present in the PDB.

While polyK, polyE, polyQ, and polyN regions might transiently fold into partially α-helical structures, it is extremely unlikely that these cause specific protein interactions or form coiled coils. Even if such regions formed an α-helix, the helix surface would be indistinguishable and would lead to uncontrolled aggregation in all directions. However, organisms with massive homopolymeric regions such as *Dictyostelium discoideum*⁵⁸ do not show more protein aggregation than any other species, which in turn suggests that these homopolymeric regions rather form single α-helices and not aggregates. Therefore, we suggest using the term “coiled coil” exclusively in the original sense coined by Francis Crick for two or more α-helices in a dedicated structural packing, and not for any stretch of amino acids that might partially fold into α-helices and aggregate. In this sense, polyK, polyE, polyQ, and polyN (and polyR and polyA) regions do not form coiled coils.

Discussion

Coiled coils consist of a minimum building block of just an α-helix and can be designed on a drawing board based on a poly-alanine backbone and subsequently substituting alanines by hydrophobic, charged, and polar amino acids to obtain structures with certain characteristics, mainly a certain length and topology^{14,15}. The simplicity in design should, in principle, allow a relatively accurate and precise prediction of these motifs in real-world sequences. For the evaluation of the performance of coiled-coil predictions in the context of a functional genome annotation the reference data set should be large and diverse, should contain only a few percent sequences with coiled-coil regions, and should allow structural verification. To our knowledge the protein structure databank (PDB) represents the most comprehensive reference data set given the broad sampling of species, protein families and protein folds. As ground truth and reference we used the coiled coils detected by SOCKET, which identifies knobs-into-hole packings of α-helices within protein structures. We evaluated the performance of coiled-coil prediction tools against all unique sequences within the PDB using all common binary classification metrics. The specificity and accuracy of all prediction tools is very high, which is a natural result from the large proportion of true negatives within the data set. In contrast, the sensitivity and precision are rather low. This is, in part, due to the lower total numbers of predictions compared to the number of SOCKET hits (however, NCOILS predicted about four times more coiled coils), but more importantly the result of the large proportion of false positive predictions (predicted coiled coils where SOCKET did not find any). Coiled coils were predicted in 31,040 PDB files where no SOCKET hits were found, and thousands of coiled coils were detected within PDB files that do not overlap with SOCKET hits. The structures for which SOCKET might fail explain some dozens of these predicted coiled coils and even a few hundreds, but not the thousands of false positive predictions. It is

highly unlikely that SOCKET missed many “classical coiled coils”, which are the supposed primary target of the prediction tools. Without having inspected all predictions manually, we suspect that it is more likely that most of these predictions are in fact false positive hits. Very obvious cases of false positive hits include the prediction of coiled coils in polyQ regions, which are not supported by structural data, and the prediction of coiled-coil regions in sequences that form β -strands, loops and other non- α -helical structures.

The low sensitivity of the prediction tools comes with a high number of false negatives (SOCKET-detected coiled coils that were not predicted). In 3401 PDB files coiled coils were exclusively found by SOCKET. At first instance, the most likely explanation for these cases is that the coiled-coil prediction tools are thought to be specific for solvent-exposed, left-handed coiled-coil dimers, and are not expected to detect types of coiled-coil α -helices buried within globular domains or as part of transmembrane structures. And because most coiled-coil prediction tools were developed before next-generation sequencing boosted sequence databases, the relatively low number of training data for tool development could have also been limiting in detecting more divergent coiled-coil types. However, the false negative rates (1—sensitivity; also called miss rate) of the individual tools at the sequence level are in the range of 63.5% (NCOILS) to 93.4% (PairCoil) indicating that the majority of even the classical coiled coils are not detected. Our analysis also shows that the amino acid patterns at the heptad positions of SOCKET hits overlapping and not overlapping with predictions are very similar. This implies that classical coiled coils and coiled coils within globular structures have similar amino acid distributions suggesting that most of the SOCKET hits in the 3401 PDB files could have been identified by coiled-coil prediction tools just as well as those that were detected.

The discussed metrics depend on the proportion of true and false positives and negatives in the benchmark data set. As discussed, if the fraction of true positives (coiled coils) is low compared to true negatives (no coiled coil), specificity and accuracy will automatically be high, if only 50% of the predictions are correct (Fig. 1). If the fraction of true positives is high compared to true negatives, the sensitivity will automatically be high. Fortunately, there is a metric termed the “Matthews correlation coefficient” (MCC) that is insensitive to the proportion of true positives (coiled coils) in the data set and that gives a balanced assessment of the performance⁵⁰. According to this metric, the performance of the coiled-coil prediction tools at the PDB file level was rather poor (MCCs between -0.05 and 0.19 when requiring only the overlap of a single amino acid between SOCKET hits and coiled-coil predictions) and did not significantly change at the sequence level (MCCs between 0.02 and 0.22). The MCCs do not considerably increase if the 3401 PDB files with exclusive SOCKET hits are regarded as true negatives (no coiled coils to be detected by prediction tools) and if the 230 PDB files with coiled coils predicted by all tools but not detected by SOCKET are regarded as true positives. Requiring only a single amino acid overlap is a rather weak criterion and balances possible biasing effects from software parameters such as the SOCKET packing-cutoff and window sizes or cutoffs from prediction tools. When requiring a more realistic overlap of at least 50% of predictions and SOCKET hits the quality of the prediction tools is no better than random, based on the MCCs (Supplementary Fig. S8). Independent of whether comparisons and analyses of coiled-coil prediction tools report high sensitivities, specificities, accuracies, and precisions, this analysis of the entire PDB using the MCC as a balanced measure demonstrates that it is random whether a predicted coiled coil in an unknown sequence is a coiled coil or not. In fact, the performance of the tools is very similar to a naïve model assuming that the prediction is random but knows and reproduces the proportion of the reference category in the data set.

The finding that coiled-coil prediction tools show low performance when benchmarked with sequences from heterogeneous protein structures is not completely new. A comparison of SpiriCoil, Marcoil, and PairCoil2 revealed a similar low absolute performance, with SpiriCoil, the supposed best performing coiled-coil prediction tool in this comparison, displaying a sensitivity of 41.7% and an FDR of 84.6% at the level of sequences³². In this comparison, 2.7% of the sequences in the test data contained coiled coils, which is slightly lower than the percentage of likely coiled-coil regions in our data set (7.4% of 144,270 PDB files contained SOCKET hits). However, SpiriCoil just passes the coiled-coil assignment from SUPERFAMILY protein profiles on query sequences based on global sequence comparisons without ever verifying the presence of a coiled coil. This approach therefore ignores domain gain, loss, and rearrangement processes, which are very common in eukaryotic genomes. In contrast, the latest comparison showing good sensitivity and specificity of the prediction tools was based on a highly biased sequence data set with 63.4% of the 1643 test sequences containing coiled coils, and each of these sequences containing 2.09 coiled-coil regions on average³⁹. Already an even and random assignment of coiled coils to sequences of this data set would result in sensitivity and specificity of 79% and 100%, respectively. Because of the biased benchmark dataset, the implied quality of the coiled-coil prediction tools based on the excellent values for the metrics is completely misleading. In addition to these general shortcomings in approach and data set, in both previous studies the precise location of the coiled coils with respect to reference SOCKET hits and secondary structural elements was not determined.

Given this analysis and the application of the evaluated prediction tools, especially NCOILS, in the functional annotation of genomes it is highly questionable that many of the proteins with predicted regions really contain coiled-coil domains. In addition it is highly likely that coiled-coil domains have been missed in many proteins. Given the broad application of coiled-coil prediction tools, as citation rates suggest, and the high interest in this structural motif, as publication numbers suggest, we see a high demand for accurate coiled-coil prediction. We suggest improving the tools’ performance against unbiased and not pre-selected data and to use approaches that combine sequence profiles and secondary structure assignments, or that discriminate against certain atypical features. Secondary structure information, for example, was included in WDSP, a pipeline to predict WD40 repeats and domains⁵⁹. WD40 repeats have very low sequence homology, are therefore notoriously difficult to detect, and are usually present in a chain of seven repeats folding into a domain^{60,61}. In WDSP, protein sequences are filtered by selecting fragments with β -sheets according to PSIPRED⁶². Subsequently, WD40 repeats are detected using a profile generated by aligning repeats by secondary structure elements and not global similarity, and finally WD40 domains are assigned when chains of at least six WD40 repeats are present. As another

example, Waggawagga uses the discriminative approach to detect stable single α -helices (SAH domains), which coiled-coil prediction tools mis-predict as coiled coils^{57,63,64}. Waggawagga searches for networks of oppositely charged residues and discriminates against helix-breaking residues, networks of residues with identical charge, and networks of hydrophobic residues as found in the hydrophobic seams of coiled coils. Approaches similar to those used by WDSP and Waggawagga could be implemented to improve coiled-coil predictions. For example, protein sequence regions could be pre-filtered and/or coiled-coil predictions could be post-filtered by secondary structure predictions. A selection filter for potentially coiled-coil domain containing regions could also be the detection by at least two tools. Training the prediction tools against unbiased data such as the entire PDB could also improve tools' performance. The evaluation of multiple tools to predict the pathogenicity of SNPs⁶⁵ and protein stability⁶⁶ also demonstrated low performance (low to medium MCCs), but the results stimulated substantial tool improvement with respect to the benchmark data sets.

In conclusion, at best, the evaluated tools predict coiled-coil regions in well described and well analysed coiled-coil forming proteins with reasonable accuracy. For predicting coiled coils in large data sets with balanced proportion of all protein folds, such as present in gene prediction datasets, the tested tools have only limited applicability. One possibility to reduce the number of false predictions in such functional genome annotations would be to only accept coiled-coils regions if predicted by multiple prediction tools and to only predict coiled coils in regions not already covered by other protein domain predictions.

Methods

Benchmark dataset. To benchmark the performance of coiled-coil prediction software as fairly and reliably as possible, we created a copy set of the current state (15/12/2018) of all available 147,073 PDB structures from the RCSB Protein Data Bank⁶⁷. The flatfiles were downloaded, stored locally and parsed with BioRuby v.1.5.1⁶⁸. Removing nucleotide-only structures and some PDB-files with handling issues reduced the number of usable structures to 144,270. Main reasons for handling issues were unavailability (moved/renamed/discontinued) at the RCSB servers (761 files), and BioRuby parsing issues with some of the structures in the PDBx format, some early structures from last century, and structures with non-natural amino-acids. We refrained from any attempt to manually remove PDB files, which could be part of the training data of some of the tools. All six tools benchmarked used mainly collections which are known to contain coiled-coil sequences, such as intermediate filament proteins, muscle myosins, kinesins, tropomyosins, dyneins (for an extended list see²³), and, if at all, only a few sequences derived from the PDB. Given the number of sequences used in this benchmark here, the effect of the few PDB sequences already present in the training data of some of the tools should be marginal. In any case, the performance of the tools might look slightly better in the benchmark than is in practice. The information from PDB files and all additional data generated were stored in a PostgreSQL database. To facilitate data handling and analysis of the PDB, DSSP, SOCKET, and coiled-coil prediction information a relational database scheme was designed, which stores the relevant data for the evaluation with low redundancy and depicts each data type into its assigned classes (Supplementary Fig. S10). Protein sequences were extracted from the ATOM records of the PDB files. Identical sequences (from start to end including identity of possible gaps) were removed independent of whether they were present in the same or different PDB files. This means that sequences differing by a single amino acid at, for example, the N- or C-terminus because of their presence in different structures or independent molecules within the asymmetric unit are treated as different sequences. The remaining sequences break down to 187,776 unique sequences. In order to create a broad, representative data set, these unique sequences remained unreduced in terms of similarity or other criteria, even very short sequences were left in the data set. Only the 755 sequences containing amino acids labeled "unknown", one-letter code "X", which are handled very differently by the coiled-coil prediction tools, were removed from the analysis. The remaining 187,021 sequences were used as reference for all analyses. Accordingly, the secondary structure for every single amino acid within the reference sequences is known. The overall size of the database for the current PDB structure data set amounts to 2.9 GB, the flatfiles in combination with the predictions sum up to around 159 GB in 2.8 million files.

Running coiled-coil software. The SOCKET algorithm detects knobs-into-holes packed α -helices based on secondary structure assignments from DSSP v.2.0.4 (Define Secondary Structure of Proteins)^{54,55}, which we generated according to the software documentation. DSSP only needs specification of an input and an output file. For the determination of coiled-coil regions SOCKET³⁷ was run with the recommended parameter settings, especially the packing-cutoff was left at the default 7 Å as described in the documentation. The coiled coils were, according to the database model, split into their superordinate structure and building/participating components, which contain registers, sequences and position information. To prevent mis-assignment of any amino acid due to sequence gaps or other unexpected shifts, the register assigned sequence of each SOCKET-determined coiled-coil component is searched in the respective PDB sequence and a potential offset is added to the component database entry.

Each coiled-coil prediction software was run with its recommended default settings and a search window of 21 amino acids, except for Marcoil and MultiCoil2, which are implemented to run without a respective window setting. Accordingly, quite conservative thresholds were set for coiled-coil selections. For Marcoil a lower limit of 90.0 (minimum) was chosen (HMM training file 9FAM, with default transition and emission parameters). For MultiCoil and MultiCoil2, the "CoiledCoil-Threshold" was set to 0.25 (minimum). For NCOILS, the "CoiledCoil-Threshold" was 0.5 (minimum), and the latest provided MTIDK-matrix was used. For PairCoil and PairCoil2, the "CoiledCoil-Threshold" was 0.84 (minimum) and 0.025 (maximum), respectively. Because the reference sequences do not contain the gap information as present in protein structures, coiled-coil prediction tools handle all sequences as continuous entities. This might affect the length of some predicted coiled coils.

Data reduction. For generating data sets with reduced sequence redundancy, the PostgreSQL database was copied three times and the unique sequences were subjected to CD-hit⁴⁹ applying 90%, 70% and 50% sequence identity cut-offs, respectively. Each of the now four databases were copied another two times and region length cut-offs of 14 and 21 amino acids were applied to the stored reference SOCKET regions and the coiled-coil predictions.

Determining overlap between assignments and predictions. Assigning SOCKET hits, DSSP assignments, and coiled-coil predictions globally to PDB files and chains is trivial. However, precisely determining overlapping regions within sequences is more challenging because amino acid numbering schemes change with data parsing and gaps in structures. Numbering of amino acids in DSSP features and SOCKET hits follows the numbering of the amino acids in the structures, which either start with the first amino acid of the protein construct, the first amino acid of the sequence of interest (excluding any terminal amino acids from protein expression plasmids), or follow the numbering of the analysed protein with respect to the numbering in the gene or transcript. The numbering of amino acids in the structures is usually aware of gaps. When extracting sequences from PDB files all position-wise numbering information including gaps is discarded, only start- and end-positions in PDB numbering are retained. Sequences themselves are stored plain without any numbering in the database, meaning a sequential numbering starting with “1” when referred to from other tools. Instead of fitting the results from the prediction tools to the complex numbering in the structure files, the numbering of the initial register sequences of the SOCKET hits was shifted to the position-independent numbering as described. Accordingly, the matching between SOCKET hits and coiled-coil predictions is independent of any peculiarities in structure numbering and independent of any gaps in the structures. Similarly, the sequences corresponding to every contiguous DSSP feature in a structure are located in the number-less sequences and the DSSP features are subsequently numbered according to the matching. A problem with this approach could be that very short DSSP features (1–3 amino acids), which are surrounded by sequence without feature (“loop”), might match to more than one position. However, these very short DSSP features are very likely not part of SOCKET hits or coiled-coil predictions so that the matching of DSSP and SOCKET/coiled-coil prediction is not affected.

Handling difficult cases. There are a few cases which cannot be resolved consistently without introducing multiple subcategories, which in turn would considerably detract from the main message without adding additional understanding. One of these problems is handling cases of overlapping SOCKET hits and coiled-coil prediction when one overlaps multiple of the other. In such cases we treated every overlap independently. For example, a long coiled-coil prediction could overlap with a SOCKET hit in its N-terminal half and another SOCKET hit in its C-terminal half. Such cases were treated as two independent overlap instances.

The actual data categories and assigned categories (true and false positives and negatives) for computing the Matthews correlation coefficient at the level of PDB files are clearly defined. It is, however, difficult to define the same categories (true and false positives and negatives) in case of evaluating the cases of overlapping SOCKET hits and coiled-coil predictions. The problem is that multiple hits were found in many PDB files, and those hits can be overlapping and non-overlapping. As a rough approximation for an upper bound we computed for each tool the percentage of overlapping hits within PDB files, and applied this percentage onto the number of overlapping PDB files. With this approach we rather overestimate the number of true overlapping hits, because for most tools there are more non-overlapping than overlapping hits in PDB files with both SOCKET hits and coiled-coil predictions, and the false positive PDB files contain, to some extent, multiple coiled-coil predictions whose contribution is ignored.

Viewer for coiled coils mapped to PDB structures. For visual inspection of SOCKET hits and coiled-coil predictions, a simplified search interface to the analysis database and a 3D molecule viewer were integrated into the coiled-coil project site Waggawagga. Structures can be searched by PDB ID and results are displayed for each structure on a single page. The result page presents the structure in JSmol, a JavaScript—only version of Jmol⁶⁹, for interactive viewing, some general information about the PDB file, and all SOCKET hits and coiled-coil predictions if found. The SOCKET hits and coiled-coil predictions are shown in the sequence-based and interactive Waggawagga format³⁷ providing access to all information down to the amino acid level. SOCKET hits and coiled-coil predictions can be loaded and combined into the JSmol viewer by simple selection. Intersecting regions between SOCKET hits and predictions are marked separately allowing the structure-based visual inspection of coiled-coil assignment by SOCKET versus prediction tools.

Handling of multimers not present in the PDB files. Having finished the entire benchmarking study we were made aware by a reviewer comment that we might have missed a substantial part of coiled coils because we did not include multimers, which are not present in the PDB files. PDB files contain the coordinates of the molecules present in the asymmetric units. In few cases, a biologically relevant multimer might be present in the crystal structure, of which only a monomer is present in the asymmetric unit. These multimers can be reconstructed using the transformation matrices present in the BIOMT part of the PDB files. To reconstruct all possible multimers we used the tool MakeMultimer.py (<http://watcut.uwaterloo.ca/tools/makemultimer/index>). 7180 PDB files were missing BIOMT information. 760 PDB files generated a “__main__.PdbError: invalid pdb code “ error, 675 generated a “list index out of range” error, and 10 generated key or value errors. Subsequently, DSSP and SOCKET were run on all multimers. DSSP failed on 2081 of the generated multimer PDB files. 875 PDB files were obtained that contain coiled coils that were not present in the initial data set. Of these, 9 are not based on multimers but the result of various bugs in the SOCKET output that we were not aware of when checking the SOCKET output files of the benchmark data. In contrast, 487 coiled coils are only present in the initial

dataset and not in the makeMultimer.py generated dataset due to the missing BIOMT data and the error messages described above. The total number of missed coiled coils is small compared to the number of coiled coils in the benchmark data, and the coiled coils likely distribute on true positives and false negatives similar to the benchmark data. A few false positives might turn to true positives, but most of the missed coiled coils (SOCKET) will add to the false negatives (no coiled coil predicted), thus numbers of the predictions will mainly shift from true negatives to false negatives. For an example see Supplementary Fig. S11. Given the low number of PDB files with coiled coils, which were predicted by all tools but are not present in the SOCKET reference dataset (see Venn diagram in Fig. 1C, only presence in same PDB file was tested but not overlap) the missed coiled coils will show a similar distribution for true positives and false negatives for each prediction tool as show the coiled coils in the analysed data set.

Data availability

The data are freely available at figshare <https://doi.org/10.6084/m9.figshare.9994706>.

Received: 3 February 2021; Accepted: 28 May 2021

Published online: 14 June 2021

References

1. Woolfson, D. N. The design of coiled-coil structures and assemblies. *Adv. Protein Chem.* **70**, 79–112 (2005).
2. Lupas, A. N. & Gruber, M. The structure of alpha-helical coiled coils. *Adv. Protein Chem.* **70**, 37–78 (2005).
3. Brown, J. H., Cohen, C. & Parry, D. A. Heptad breaks in alpha-helical coiled coils: stutters and stammers. *Proteins* **26**, 134–145 (1996).
4. Hicks, M. R., Holberton, D. V., Kowalczyk, C. & Woolfson, D. N. Coiled-coil assembly by peptides with non-heptad sequence motifs. *Fold. Des.* **2**, 149–158 (1997).
5. Gruber, M. & Lupas, A. N. Historical review: another 50th anniversary—new periodicities in coiled coils. *Trends Biochem. Sci.* **28**, 679–685 (2003).
6. Kühnel, K. *et al.* The VASP tetramerization domain is a right-handed coiled coil based on a 15-residue repeat. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 17027–17032 (2004).
7. Liguori, A. *et al.* NadA3 structures reveal undecad coiled coils and LOX1 binding regions competed by meningococcus B vaccine-elicited human antibodies. *MBio* **9**, e01914-18 (2018).
8. Crick, F. H. C. The Fourier transform of a coiled-coil. *Acta Crystallogr.* **6**, 685–689 (1953).
9. Parry, D. A. Coiled-coils in alpha-helix-containing proteins: analysis of the residue types within the heptad repeat and the use of these data in the prediction of coiled-coils in other proteins. *Biosci. Rep.* **2**, 1017–1024 (1982).
10. Parry, D. A. D., Fraser, R. D. B. & Squire, J. M. Fifty years of coiled-coils and α -helical bundles: a close relationship between sequence and structure. *J. Struct. Biol.* **163**, 258–269 (2008).
11. Lupas, A., Van Dyke, M. & Stock, J. Predicting coiled coils from protein sequences. *Science* **252**, 1162–1164 (1991).
12. Woolfson, D. N. Coiled-coil design: updated and upgraded. *Subcell. Biochem.* **82**, 35–61 (2017).
13. Lupas, A. N. & Bassler, J. Coiled coils—a model system for the 21st century. *Trends Biochem. Sci.* **42**, 130–140 (2017).
14. Marsden, H. R. & Kros, A. Self-assembly of coiled coils in synthetic biology: inspiration and progress. *Angew. Chem. Int. Ed. Engl.* **49**, 2988–3005 (2010).
15. Fletcher, J. M. *et al.* A basis set of de novo coiled-coil peptide oligomers for rational protein design and synthetic biology. *ACS Synth. Biol.* **1**, 240–250 (2012).
16. Thomson, A. R. *et al.* Computational design of water-soluble α -helical barrels. *Science* **346**, 485–488 (2014).
17. Boyken, S. E. *et al.* De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity. *Science* **352**, 680–687 (2016).
18. Mravic, M. *et al.* Packing of apolar side chains enables accurate design of highly stable membrane proteins. *Science* **363**, 1418–1423 (2019).
19. Lizatović, R. *et al.* A De Novo designed coiled-coil peptide with a reversible pH-induced oligomerization switch. *Structure* **24**, 946–955 (2016).
20. Lupas, A. Prediction and analysis of coiled-coil structures. *Methods Enzymol.* **266**, 513–525 (1996).
21. Berger, B. *et al.* Predicting coiled coils by use of pairwise residue correlations. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 8259–8263 (1995).
22. Wolf, E., Kim, P. S. & Berger, B. MultiCoil: a program for predicting two- and three-stranded coiled coils. *Protein Sci.* **6**, 1179–1189 (1997).
23. McDonnell, A. V., Jiang, T., Keating, A. E. & Berger, B. Paircoil2: improved prediction of coiled coils from sequence. *Bioinformatics* **22**, 356–358 (2006).
24. Delorenzi, M. & Speed, T. An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. *Bioinformatics* **18**, 617–625 (2002).
25. Trigg, J., Gutwin, K., Keating, A. E. & Berger, B. Multicoil2: predicting coiled coils and their oligomerization states from sequence in the twilight zone. *PLoS ONE* **6**, e23519 (2011).
26. Tanizawa, H., Ghimire, G. D. & Mitaku, S. A high performance prediction system of coiled coil domains containing heptad breaks: SOSULcoil. *Chem-Bio Inform. J.* **8**, 96–111 (2008).
27. Gruber, M., Söding, J. & Lupas, A. N. REPPER—repeats and their periodicities in fibrous proteins. *Nucl. Acids Res.* **33**, W239–W243 (2005).
28. Fariselli, P., Molinari, D., Casadio, R. & Krogh, A. Prediction of structurally-determined coiled-coil domains with hidden Markov models. In *Bioinformatics Research and Development* (eds Hochreiter, S. & Wagner, R.) 292–302 (Springer, 2007).
29. Bartoli, L., Fariselli, P., Krogh, A. & Casadio, R. CCHMM_PROF: a HMM-based coiled-coil predictor with evolutionary information. *Bioinformatics* **25**, 2757–2763 (2009).
30. Ludwiczak, J., Winski, A., Szczepaniak, K., Alva, V. & Dunin-Horkawicz, S. DeepCoil—a fast and accurate prediction of coiled-coil domains in protein sequences. *Bioinformatics* **35**, 2790–2795 (2019).
31. Gough, J. & Chothia, C. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucl. Acids Res.* **30**, 268–272 (2002).
32. Rackham, O. J. L. *et al.* The evolution and structure prediction of coiled coils across all genomes. *J. Mol. Biol.* **403**, 480–493 (2010).
33. Mahrenholz, C. C., Abfalder, I. G., Bodenhofer, U., Volkmer, R. & Hochreiter, S. Complex networks govern coiled-coil oligomerization—predicting and profiling by means of a machine learning approach. *Mol. Cell. Proteom.* **10**, M110-004994 (2011).
34. Armstrong, C. T., Vincent, T. L., Green, P. J. & Woolfson, D. N. SCORER 2.0: an algorithm for distinguishing parallel dimeric and trimeric coiled-coil sequences. *Bioinformatics* **27**, 1908–1914 (2011).

3. Publications and Manuscripts

35. Vincent, T. L., Green, P. J. & Woolfson, D. N. LOGICOIL—multi-state prediction of coiled-coil oligomeric state. *Bioinformatics* **29**, 69–76 (2013).
36. Li, C., Wang, X.-F., Chen, Z., Zhang, Z. & Song, J. Computational characterization of parallel dimeric and trimeric coiled-coils using effective amino acid indices. *Mol. Biosyst.* **11**, 354–360 (2015).
37. Walshaw, J. & Woolfson, D. N. Socket: a program for identifying and analysing coiled-coil motifs within protein structures. *J. Mol. Biol.* **307**, 1427–1450 (2001).
38. Gruber, M., Söding, J. & Lupas, A. N. Comparative analysis of coiled-coil prediction methods. *J. Struct. Biol.* **155**, 140–145 (2006).
39. Li, C. *et al.* Critical evaluation of in silico methods for prediction of coiled-coil domains in proteins. *Brief. Bioinform.* **17**, 270–282 (2016).
40. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).
41. Andreeva, A. *et al.* Data growth and its impact on the SCOP database: new developments. *Nucl. Acids Res.* **36**, D419–D425 (2008).
42. Liu, J. & Rost, B. Comparing function and structure between entire proteomes. *Protein Sci.* **10**, 1970–1979 (2001).
43. Testa, O. D., Moutevelis, E. & Woolfson, D. N. CC+: a relational database of coiled-coil structures. *Nucl. Acids Res.* **37**, D315–D322 (2009).
44. Moutevelis, E. & Woolfson, D. N. A periodic table of coiled-coil protein structures. *J. Mol. Biol.* **385**, 726–732 (2009).
45. Heal, J. W., Bartlett, G. J., Wood, C. W., Thomson, A. R. & Woolfson, D. N. Applying graph theory to protein structures: an Atlas of coiled coils. *Bioinformatics* **34**, 3316–3323 (2018).
46. Szappanos, B., Süveges, D., Nyitray, L., Perczel, A. & Gáspári, Z. Folded-unfolded cross-predictions and protein evolution: the case study of coiled-coils. *FEBS Lett.* **584**, 1623–1627 (2010).
47. Kollmar, M. & Mühlhausen, S. Myosin repertoire expansion coincides with eukaryotic diversification in the Mesoproterozoic era. *BMC Evol. Biol.* **17**, 211 (2017).
48. Simm, D., Hatje, K. & Kollmar, M. Distribution and evolution of stable single α -helices (SAH domains) in myosin motor proteins. *PLoS ONE* **12**, e0174639 (2017).
49. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
50. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **405**, 442–451 (1975).
51. Rost, B., Sander, C. & Schneider, R. Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.* **235**, 13–26 (1994).
52. Zemla, A., Venclovas, C., Fidelis, K. & Rost, B. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins* **34**, 220–223 (1999).
53. Liu, T. & Wang, Z. SOV_refine: a further refined definition of segment overlap score and its significance for protein structure similarity. *Sour. Code Biol. Med.* **13**, 1 (2018).
54. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
55. Touw, W. G. *et al.* A series of PDB-related databanks for everyday needs. *Nucl. Acids Res.* **43**, D364–D368 (2015).
56. Fiumara, F., Fioriti, L., Kandel, E. R. & Hendrickson, W. A. Essential role of coiled coils for aggregation and activity of Q/N-rich prions and PolyQ proteins. *Cell* **143**, 1121–1135 (2010).
57. Simm, D., Hatje, K. & Kollmar, M. Waggawagga: comparative visualization of coiled-coil predictions and detection of stable single α -helices (SAH domains). *Bioinformatics* **31**, 767–769 (2015).
58. Eichinger, L. *et al.* The genome of the social amoeba *Dictyostelium discoideum*. *Nature* **435**, 43–57 (2005).
59. Wang, Y., Jiang, F., Zhuo, Z., Wu, X.-H. & Wu, Y.-D. A method for WD40 repeat detection and secondary structure prediction. *PLoS ONE* **8**, e65705 (2013).
60. Neer, E. J., Schmidt, C. J., Nambudripad, R. & Smith, T. F. The ancient regulatory-protein family of WD-repeat proteins. *Nature* **371**, 297–300 (1994).
61. Smith, T. F. Diversity of WD-repeat proteins. *Subcell Biochem.* **48**, 20–30 (2008).
62. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202 (1999).
63. Peckham, M. & Knight, P. J. When a predicted coiled coil is really a single α -helix, in myosins and other proteins. *Soft Matter* **5**, 2493–2503 (2009).
64. Simm, D. & Kollmar, M. Waggawagga-CLI: a command-line tool for predicting stable single α -helices (SAH-domains), and the SAH-domain distribution across eukaryotes. *PLoS ONE* **13**, e0191924 (2018).
65. Thusberg, J., Olatubosun, A. & Vihinen, M. Performance of mutation pathogenicity prediction methods on missense variants. *Hum. Mutat.* **32**, 358–368 (2011).
66. Khan, S. & Vihinen, M. Performance of protein stability predictors. *Hum. Mutat.* **31**, 675–684 (2010).
67. Burley, S. K. *et al.* RCSB Protein Data Bank: sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education. *Protein Sci.* **27**, 316–330 (2018).
68. Goto, N. *et al.* BioRuby: bioinformatics software for the ruby programming language. *Bioinformatics* **26**, 2617–2619 (2010).
69. Hanson, R. M., Prilusky, J., Renjian, Z., Nakane, T. & Sussman, J. L. JSmol and the next-generation web-based representation of 3D molecular structure as applied to proteopedia. *Isr. J. Chem.* **53**, 207–216 (2013).

Acknowledgements

We acknowledge support by the Open Access Publication Funds of the Göttingen University.

Author contributions

D.S. and M.K. designed the study with input from K.H., D.S. developed the software, set up the data viewer, and performed data engineering and all computations. K.H. contributed to database design. D.S. and M.K. performed data analysis, designed the figures and drafted the manuscript. S.W. was involved in statistical analyses. All authors discussed the results and commented on the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-91886-w>.

Correspondence and requests for materials should be addressed to M.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021

3.2. Heterologous gene expression

3.2.1. Design of typical genes for heterologous gene expression

Dominic Simm^{1,2}, Blagovesta Popova³, Gerhard H. Braus³, Stephan Waack¹ and Martin Kollmar^{1,2,*}

* Corresponding author

¹ Theoretical Computer Science and Algorithmic Methods, Institute of Computer Science, Georg-August University, Göttingen, Germany

² Group Systems Biology of Motor Proteins, Department of NMR-based Structural Biology, Max-Planck-Institute for Biophysical Chemistry, Göttingen, Germany

³ Molecular Microbiology and Genetics, Institute for Microbiology and Genetics and Göttingen Center for Molecular Biosciences (GZMB), Georg-August-University Göttingen, Göttingen, Germany

Manuscript submitted

Contribution

DS, SW and MK designed the study. DS developed the software, set up the web-application, and performed data engineering and all computations. SW contributed to the algorithm design. BP designed and performed the experimental studies. DS and MK performed data analysis and drafted the manuscript. DS, MK and BP designed the figures. GHB was involved in experimental study design and data interpretation. All authors discussed the results and reviewed the manuscript.

Abstract

Heterologous protein expression is an important method for analysing cellular functions of proteins, in genetic circuit engineering and in overexpressing proteins for biopharmaceutical applications and structural biology research. The degeneracy of the genetic code, which enables a single protein to be encoded by a multitude of synonymous gene sequences, plays an important role in regulating protein expression, but substantial uncertainty exists concerning the details of this phenomenon. Here we analyse the influence of a profiled codon usage adaptation approach on protein expression levels in the eukaryotic model organism *Saccharomyces cerevisiae*. We selected green fluorescent protein (GFP) and human α -synuclein (α Syn) as representatives for stable and intrinsically disordered proteins and representing a benchmark and a challenging test case. A new approach was implemented to design typical genes resembling the codon usage of any subset of endogenous genes. Using this approach, synthetic genes for GFP and α Syn were generated, heterologously expressed and evaluated in yeast. We demonstrate that GFP is expressed at high levels, and that the toxic α Syn can be adapted to endogenous, low-level expression. The new software is publically available as a web-application for performing host-specific protein adaptations to a set of the most commonly used model organisms (<https://odysseus.motorprotein.de>).

Introduction

Modifying gene sequence is an important step when generating sequences for homologous and heterologous protein expression (Brule and Grayhack, 2017; Gustafsson et al., 2012; Hanson and Coller, 2018; Hershberg and Petrov, 2009; Nieuwkoop et al., 2020). This allows, for example, to investigate functions of homologous proteins (Hia et al., 2019), to synthetically construct genetic circuits (Hansen et al., 2014; Kato, 2019; Michalodimitrakis and Isalan, 2009), or to overexpress proteins for biopharmaceutical applications (Mauro, 2018) and structural biology research (Hedfalk, 2012). These types of experiments have on the one hand highly profited from the exponentially accumulating sequence information from genome and transcriptome sequencing projects. On the other hand, the considerable increase in speed and decrease in costs for synthetic gene synthesis provides a convenient way to obtain physical genes encoding the desired proteins.

The genetic code redundancy allows adjusting gene sequences without changing the protein sequences. This principle is used for a long time for practical aspects such as facilitating cloning by adding or removing restriction sites or by removing internal Shine-Dalgarno consensus sequences. Here, just one or a few codons are altered. Adjusting all codons of a gene to a certain codon usage frequency is often referred to as codon optimization (Gould et al., 2014; Welch et al., 2009). In synthetic biology the optimization goal is mostly increased protein expression, whereas in many other biological applications overexpressed proteins might generate unwanted effects and adjustment to the codon usage frequency of lowly expressed proteins might be preferred. Most gene design tools optimize the codon adaptation index (CAI) (Sharp and Li, 1987), which is the deviation of a protein coding sequence from a set of reference genes and ranges from 0 to 1. Very simply, the CAI of a gene is optimized to perfection if only the most used codons of the reference gene sets are used. For most applications the reference gene sets just consist of a few to a few dozen genes, of which most encode ribosomal proteins (Jansen et al., 2003; Sharp and Li, 1987).

Instead of this rather statistical approach that evaluates gene sequences by codon counting, gene design can be driven by biochemical observations and deeper understanding of the ribosomal trans-

lation. In the protein biosynthesis process, the simultaneous presence of two tRNAs in the A and P positions of the ribosome is necessary for the formation of a peptide bond (Nierhaus et al., 1998; Rodnina, 2018; Stark et al., 1997). Due to steric reasons not all combinations of codons and tRNAs are equally compatible to the ribosome surface, which means that certain codon pairs are processed more efficiently than others. If this were the case it would be expected that the observed frequency of occurrence of a codon pair would significantly deviate from its statistically predicted mean value. Analysing 237 protein coding genes from *E. coli* demonstrated that some codon pairs were overrepresented while others were underrepresented in comparison with the theoretical predicted means (Gutman and Hatfield, 1989). This study has later been extended to all protein coding genes of the *E. coli* genome (Boycheva et al., 2003) and also to several hundred organisms from all three domains of life (Tats et al., 2008). The phenomenon of the non-random utilization of codon pairs is called the "codon context" and is assumed to correlate with the translation elongation rate in a way that rare codon pairs decrease the rate (Coleman et al., 2008). Only few gene design software use codon context information (Gaspar et al., 2012; Lanza et al., 2014; Taneda and Asai, 2020).

Here, we developed a software to design "typical genes". Typical genes are not optimized against parameters such as CAI or codon usage but are intended to show a similar codon distribution as compared to a reference gene set, which can be a selection of highly or lowly expressed genes or a selection of genes having a similar cellular context such as transmembrane or cytoskeletal proteins. Because many studies showed that heterologous proteins can strongly be overexpressed although they have a counterintuitively wrong codon usage, we developed a formalism to invert a selected codon usage. The new design algorithm was tested by designing typical genes for green fluorescent protein (GFP, Zimmer, 2002) and human α -synuclein (α Syn, Meade et al., 2019) and evaluating their expression in the unicellular budding yeast *Saccharomyces cerevisiae*.

Materials and Methods

Model for generating typical genes

Given a protein sequence a set of typical genes is generated using a Markov chain model. The Markov chain is built by using the relative synonymous di-codon usage frequencies (*RSdCU*) for the transition/emission probabilities (Figure 3.1). The relative codon usage refers to the usage of a codon with respect to all 61 sense codons, the relative synonymous codon usage refers to the usage of a codon within the set of codons coding for the same amino acid. The relative di-codon usage refers to the frequency of each set of two neighbouring codons. The *RSdCU* is defined here as the relative synonymous codon usage of the second codon of a di-codon with respect to all codons coding for the same second amino acid. All frequencies are normalized within each codon box. The reference for all the codon usage metrics is the codon usage within a set of genes. Usually, for heterologous gene expression the set of genes is taken from the host organism, to which the gene sequence obtained from another species should be adapted. But in principle any set of genes can be chosen. By allowing the user to define a selection of genes, the gene of interest can be adapted to the codon usage of any subset of genes of a host organism, e.g. the most highly expressed genes, genes involved in metabolism, or genes coding for transmembrane proteins. To speed up the process of generating the *RSdCU* matrices, the *RSdCU* is pre-calculated from data for a number of pre-defined subsets of genes such as the selection of the highest expressed yeast genes.

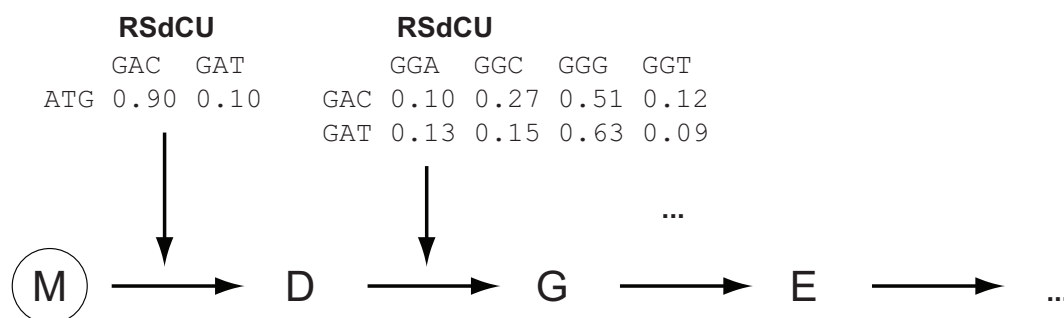


Figure 3.1. | Example of a Markov chain. For a protein sequence starting with M-D-G-E a typical gene sequence will be designed. The RSdCU frequencies are computed based on the set of selected sequences, which could be all genes of a species, the sub-section of the 10% most highly expressed genes of a species, the selection of all genes coding for trans-membrane proteins of a species, or any other user-specified set of genes. All frequencies are normalized within each codon box. The Markov chain is built by using the RSdCU for the transition/emission probabilities.

Collecting and processing codon usage and protein abundance data

Protein abundance datasets were obtained from the publicly available PaxDB database (Wang et al., 2015). PaxDB provides unified protein quantification data with proteome-wide coverage derived from biophysical and mass spectrometry studies for a broad range of organisms. The associated coding sequences (CDS) were collected via the Entrez-API (NCBI Resource Coordinators, 2018) from the National Center of Biotechnology Information (NCBI) in May 2018. The downloaded sequences were checked, and partial and invalid gene sequences as well as sequences with obvious problems (i.e. discontinued genes, internal reading-frame shifts, in-frame stop codons) were removed.

The protein abundance data allows sampling of the proteins by cellular protein abundance levels. The abundance is given in "ppm" (parts per million) and varies over several powers of ten. This means that by simply counting the corresponding codons in every subset of proteins the corresponding codons of the lowest expressed proteins would get the same weight as the corresponding codons of the highest expressed proteins, although their expression level varies considerably. To accomplish for the different abundance level of the proteins, each codon is therefore multiplied with the protein abundance resulting in the weighted codon-usage parameter-set *weighted-RSdCU*.

The annotation of the sequence data (e.g. cellular localization, biomolecular function) and the reference to the protein abundance data (from PaxDB) allows the generation of organism-specific and gene set-specific RSdCU computations for the Markov chain model. For example, based on the protein abundance information just the 50 highest expressed proteins of an organism could be selected, or the 2,000 least expressed proteins. In other use cases, for example, the ATPases, membrane proteins, or cytoskeletal proteins of an organism could be selected to generate the RSdCU matrix. This approach allows the flexible computation of the RSdCU for a diverse set of organisms as target hosts, for a set of proteins with similar expression level, and for a selected set of proteins with similar cellular function.

Inverting the codon usage

While we compared the codon usage of many non-yeast genes with the codon usage of yeast, we observed that the codon usage of the non-yeast genes is often different, different not by showing a random different codon usage but by kind of "inverting" the difference of the codon usage of the highly expressed yeast genes compared to the yeast codon usage. For example, let the yeast codon usage of CAA and CAG be 0.68 and 0.32, respectively. Selecting only the highest expressed genes the codon usage of CAA and CAG were 0.93 and 0.07, respectively. Thus, in highly expressed genes, the codon usage of CAA is increased by 0.25 while the codon usage of CAG decreases by 0.25. "Switching" the codon usage would result in a dramatic change of the codon usage to 0.07 for CAA and 0.93 for CAG, which is not observed. Rather, the codon usage of the non-yeast genes resembles a scheme, where the difference of 0.25 between highly expressed genes and yeast codon usage is inverted, resulting in codon usage of 0.43 and 0.57 for CAA and CAG, respectively. Therefore, we here coin the term "inverted codon usage" for codon usages, which are generated by reversing the codon usage frequencies within each set of synonymous codons with respect to a reference codon usage. As reference, we here used the genome-wide frequency of each codon as described in the example above. With this reference the inversion of the codon usage results in patterns of typical codon frequency distributions, and not in rather artificial codon usages as generated by "switching" the codon usage within synonymous codons. The inversion thus represents kind of a mirror operation on the values of the genome-wide frequencies used as reference. To make the computation of the inversion fail-safe, the inverted codon usage of the respective codon is set to 0.05 or 0.95, respectively, in case the inversion would result in a negative RCU or an RCU bigger than 1 (this can only happen when extremely rare codons in the reference are the most prevalent in the set of selected sequences and vice versa). The difference between the computed inverted codon usage and the 0.05 or 0.95 setting is then proportionally subtracted from or added to the frequencies of the other codons of the codon box. For generating the RSdCU matrix, the RCUs for each codon box are inverted in both dimensions of the matrix, e.g. first inverting the codon usage of the first codon of the di-codon and then inverting the codon usage of the second codon of the di-codon (Supplementary Figure A.16).

Post-processing and filtering the initial set of typical genes

The generated sequences might contain patterns unfavourable for subsequent experimental work (e.g. presence of enzyme restriction sites) and/or patterns unfavourable for translation initiation (e.g. strong base pairing at the 5'-end of the mRNAs). Such patterns are determined in several post-processing steps. Instead of modifying the respective sequences to remove the patterns, which would lead to local deviations from the RSdCU, sequences containing the patterns are removed and further typical sequences generated. To allow filtering for restriction sites wanted or eliminated for cloning and control the respective sequence pattern information has been collected from the Restriction Enzyme Database (REBASE, Roberts et al., 2015). To allow filtering for unfavourable base pairings, RNAfold from the ViennaRNA Package (Lorenz et al., 2011) was integrated for prediction of mRNA stability.

Software implementation

The gene reconstruction algorithm is written in Python 2.7 and available as software termed Odysseus (Figure 3.2). The software can be used via a web interface at <https://odysseus.motorprotein.de>,

and obtained from GitHub at <https://github.com/dsimm/Odysseus> for local installation and use. Odysseus requires input of a protein or cDNA sequence, the latter being translated subsequently. Next, the web interface allows selecting a host-organism and adjusting model-parameters such as selection of subsets of proteins. Typical genes are then generated using pre-computed or dynamically assembled codon-usage profiles. Pre-computed profiles are available for multiple organisms based on various expression level ranges, which were termed "Low", "Mid" and "High". If the user filters proteins for their cellular function or through a systematic selection using the annotated PaxDB expression information the profile-dataset is computed dynamically. This is the more time-consuming option and should only be considered, if there is need for a more specific adaptation of the model parameters of the Markov chain to characteristic protein groups of the targeted host organism.

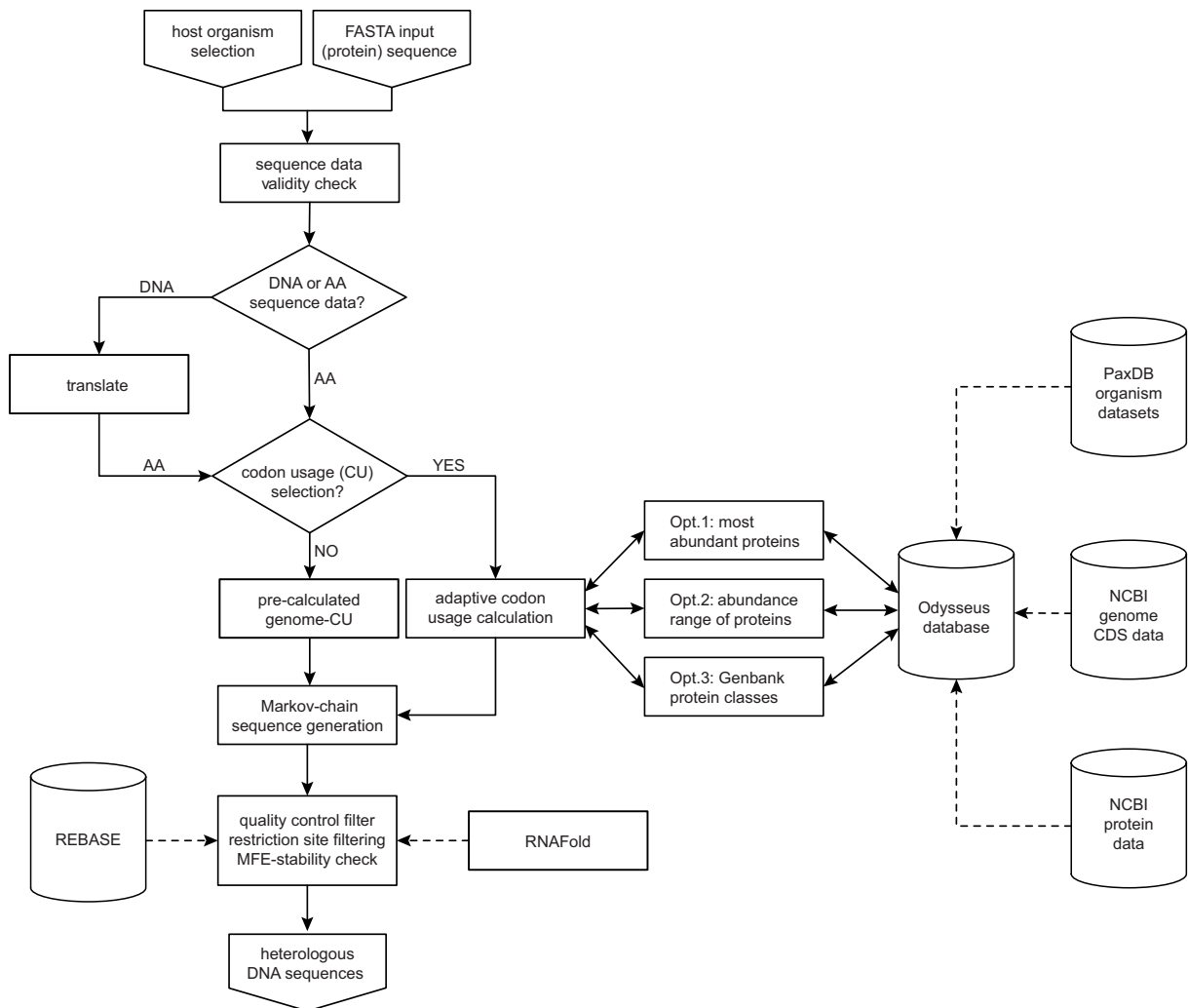


Figure 3.2. | Odysseus flowchart. The input for the process (top of the scheme) are a sequence (protein or DNA) in FASTA format and the selection of the host organism for which the gene will be designed. The resulting DNA sequence is the output of the process (bottom of the scheme). Computations during the process are represented by boxes, databases by cylinders, decisions by diamonds and the direction of data flow by arrows. Data input from external databases and computations with external software are represented by dotted lines.

Plasmid construction, Yeast Strains, Transformation and Growth Conditions

Plasmids and *Saccharomyces cerevisiae* strains are listed in Tables 1 and 2. DNA coding sequences were synthesized by Life Technologies, Darmstadt, Germany. The synthetic DNA fragments were cloned into the *SmaI* site of the integrative plasmid pRS306 using GENEART Seamless cloning and assembly kit (Life Technologies, Darmstadt, Germany). All constructs were verified by DNA sequencing. The *GAL1-SNCA* or *GAL1-GFP* sequences were integrated into the mutated *ura3-1* or *trp1-1* locus of *S. cerevisiae* W303-1A strain using an intact *URA3* or *TRP1* gene on the corresponding integrative plasmid for selection. The number of the integrated copies was determined by Southern hybridization as described previously (Petroi et al., 2012).

S. cerevisiae strain W303-1A was used for transformations performed by standard lithium acetate protocol (Gietz et al., 1992). All strains were grown in Synthetic complete (SC) medium (Guthrie and Fink, 1991) lacking the corresponding marker and supplemented with 2% raffinose or 2% glucose. α Syn or GFP expression was induced by shifting yeast cells cultivated overnight in raffinose to 2% galactose-containing medium ($OD_{600} = 0.1$).

Spotting Assay

For growth test on solid medium, yeast cells were pre-grown in SC-selection medium containing 2% raffinose to mid-log phase. Cells were normalized to equal densities, serially diluted 10-fold starting with an OD_{600} of 0.1, and spotted on SC-plates containing either 2% glucose or 2% galactose. After three days incubation the plates were photographed.

Immunoblotting

Yeast cells harboring α Syn or GFP-encoding genes were pre-grown at 30°C in SC-selection medium containing 2% raffinose. Cells were transferred to SC medium containing 2% galactose at $OD_{600} = 0.1$ to induce the *GAL1* promoter for 6 h. Total protein extracts were prepared as described (Knop et al., 1999) and the protein concentrations were determined with a Bradford assay. Equal amounts from each protein sample were subjected to 12% SDS-polyacrylamide gel electrophoresis and transferred to a nitrocellulose membrane. Membranes were probed with α Syn rabbit polyclonal antibody (Santa Cruz Biotechnology, USA) or GFP rat monoclonal antibody (Chromotek, Germany). GAPDH mouse monoclonal antibody (Thermo Fisher Scientific, USA) was used as loading controls. Pixel density values for Western quantification were obtained from TIFF files generated from digitized X-ray films (KODAK) and analyzed with the ImageJ software (NIH, Bethesda, USA). Before comparison, sample density values were normalized to the corresponding loading control.

Fluorescence microscopy and quantifications

Yeast cells harboring GFP were grown in SC-selective medium containing 2% raffinose overnight, and transferred to 2% galactose containing medium for induction of GFP expression for 6 h. Fluorescent images were obtained with Zeiss Observer. The Z1 microscope (Zeiss) was equipped with a CSU-X1 A1 confocal scanner unit (YOKOGAWA), QuantEM:512SC digital camera (Photometrics) and SlideBook 6.0 software package (Intelligent Imaging Innovations). At least 100 cells were mea-

sured per strain and per experiment for quantification of fluorescent intensities.

RNA isolation and quantitative real-time PCR

Total RNA was isolated using the "High Pure RNA Isolation Kit" (Roche Diagnostics GmbH, Mannheim, Germany) from yeast cells that were grown in SC-selective medium containing 2% galactose for induction of *GAL1* promoter for 6 hours. cDNA synthesis was performed in duplicates for each sample using 0.8 µg RNA and the QuantiTect Reverse Transcription Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. Amplification was performed with CFX Connect Real-Time System (Bio-Rad) with MESA GREEN qPCR MasterMix Plus for SYBR Assay (Eurogentec) and analyzed in three technical repeats. The expression of Histone h2A was used as reference.

Figure 3.3. | Plasmids used in this study.

| Plasmid | Description | Source |
|---------|---|------------|
| pRS306 | pRS306- <i>GAL1</i> -Promoter, <i>CYC1</i> -Terminator, <i>URA3</i> , integrative, <i>pUC</i> origin, <i>AmpR</i> | (36) |
| pME4859 | pRS306- <i>GFP</i> (low-expression-weighted; <i>gene1</i>) | This study |
| pME4860 | pRS306- <i>GFP</i> (high-expression-weighted; <i>gene2</i>) | This study |
| pME4861 | pRS306- <i>GFP</i> (high-expression-weighted-inverted; <i>gene3</i>) | This study |
| pME4853 | pRS306- <i>SCNA</i> (low expression; <i>gene4</i>) | This study |
| pME4854 | pRS306- <i>SCNA</i> (middle expression; <i>gene5</i>) | This study |
| pME4855 | pRS306- <i>SCNA</i> (high expression; <i>gene6</i>) | This study |
| pME4856 | pRS306- <i>SCNA</i> (low-expression-weighted; <i>gene7</i>) | This study |
| pME4857 | pRS306- <i>SCNA</i> (high-expression-weighted; <i>gene8</i>) | This study |
| pME4858 | pRS306- <i>SCNA</i> (high-expression-weighted-inverted; <i>gene9</i>) | This study |

Figure 3.4. | Yeast strains used in this study.

| Strain | Genotype | Source |
|---------|--|------------|
| W303-1A | <i>MATa</i> ; <i>ura3-1</i> ; <i>trp1-1</i> ; <i>leu2-3_112</i> ; <i>his3-11</i> ; <i>ade2-1</i> ; <i>can1-100</i> | EUROSCARF |
| RH3771 | W303 containing 1 genomic copy <i>GAL1::GFP</i> (low expression-weighted; <i>gene1</i>) in <i>ura3</i> locus | This study |
| RH3772 | W303 containing 2 genomic copy <i>GAL1::GFP</i> (low expression-weighted; <i>gene1</i>) in <i>ura3</i> locus | This study |
| RH3773 | W303 containing 3 genomic copy <i>GAL1::GFP</i> (low expression-weighted; <i>gene1</i>) in <i>ura3</i> locus | This study |
| RH3774 | W303 containing 1 genomic copy <i>GAL1::GFP</i> (high expression-weighted; <i>gene2</i>) in <i>ura3</i> locus | This study |
| RH3775 | W303 containing 2 genomic copy <i>GAL1::GFP</i> (high expression-weighted; <i>gene2</i>) in <i>ura3</i> locus | This study |
| RH3776 | W303 containing 3 genomic copy <i>GAL1::GFP</i> (high expression-weighted; <i>gene2</i>) in <i>ura3</i> locus | This study |

| | | |
|------------------|---|------------|
| RH3777 | W303 containing 1 genomic copy <i>GAL1::GFP</i> (<i>high expression-weighted-inverted; gene3</i>) in <i>ura3</i> locus | This study |
| RH3778 | W303 containing 2 genomic copy <i>GAL1::GFP</i> (<i>high expression-weighted-inverted; gene3</i>) in <i>ura3</i> locus | This study |
| RH3779 | W303 containing 3 genomic copy <i>GAL1::GFP</i> (<i>high expression-weighted-inverted; gene3</i>) in <i>ura3</i> locus | This study |
| RH3756 | W303 containing 1 genomic copy <i>GAL1::SNCA</i> (<i>low expression; gene4</i>) in <i>ura3</i> locus | This study |
| RH3757 | W303 containing 2 genomic copy <i>GAL1::SNCA</i> (<i>low expression; gene4</i>) in <i>ura3</i> locus | This study |
| RH3758 | W303 containing 1 genomic copy <i>GAL1::SNCA</i> (<i>middle expression; gene5</i>) in <i>ura3</i> locus | This study |
| RH3759 | W303 containing 2 genomic copy <i>GAL1::SNCA</i> (<i>middle expression; gene5</i>) in <i>ura3</i> locus | This study |
| RH3760 | W303 containing 1 genomic copy <i>GAL1::SNCA</i> (<i>high expression; gene6</i>) in <i>ura3</i> locus | This study |
| RH3761 | W303 containing 2 genomic copy <i>GAL1::SNCA</i> (<i>high expression; gene6</i>) in <i>ura3</i> locus | This study |
| RH3762 | W303 containing 1 genomic copy <i>GAL1::SNCA</i> (<i>low expression-weighted; gene7</i>) in <i>ura3</i> locus | This study |
| RH3763 | W303 containing 2 genomic copy <i>GAL1::SNCA</i> (<i>low expression-weighted; gene7</i>) in <i>ura3</i> locus | This study |
| RH3764 | W303 containing 3 genomic copy <i>GAL1::SNCA</i> (<i>low expression-weighted; gene7</i>) in <i>ura3</i> locus | This study |
| RH3765 | W303 containing 1 genomic copy <i>GAL1::SNCA</i> (<i>high expression-weighted; gene8</i>) in <i>ura3</i> locus | This study |
| RH3766 | W303 containing 2 genomic copy <i>GAL1::SNCA</i> (<i>high expression-weighted; gene8</i>) in <i>ura3</i> locus | This study |
| RH3767 | W303 containing 3 genomic copy <i>GAL1::SNCA</i> (<i>high expression-weighted; gene8</i>) in <i>ura3</i> locus | This study |
| RH3768 | W303 containing 1 genomic copy <i>GAL1::SNCA</i> (<i>high expression-weighted-inverted; gene9</i>) in <i>ura3</i> locus | This study |
| RH3769 | W303 containing 2 genomic copy <i>GAL1::SNCA</i> (<i>high expression-weighted-inverted; gene9</i>) in <i>ura3</i> locus | This study |
| RH3770 | W303 containing 3 genomic copy <i>GAL1::SNCA</i> (<i>high expression-weighted-inverted; gene9</i>) in <i>ura3</i> locus | This study |
| RH3780 | W303 containing 1 genomic copies <i>GAL1::SNCA</i> (<i>human</i>) in <i>trp1</i> locus | This study |
| RH3781 | W303 containing 2 genomic copies <i>GAL1::SNCA</i> (<i>human</i>) in <i>trp1</i> locus | This study |
| RH3465 | W303 containing 1 genomic copy <i>GAL1::GFP</i> in <i>ura3</i> locus | (32) |
| RH3466 RH3467 | W303 containing 1/2 genomic copy <i>GAL1::SNCA</i> (<i>human</i>)-GFP in <i>ura3</i> locus | (32) |

Results and Discussion

Most approaches to computationally reconstruct genes from heterologous protein sequences concentrate on optimizing the usage of so-called preferred codons (also called more frequent codons, optimal codons, or major codons) with preference usually being regarded as present in highly expressed proteins. Because there is no definition for highly expressed protein and the set of most strongly expressed proteins likely differs from species to species, multiple ways have been suggested to derive the subset of preferred codons. According to one of the earliest approaches developed in the 1980s the preferred codon is assigned to that codon of a codon box that is most frequently used across ribosomal genes. At that time this was likely the best approach given the limit in available sequences and expression data. However, while highly expressed and translated, ribosomal proteins are not representative for the cellular proteome. In another method, the overall codon bias of each gene is determined and those codons, whose frequencies within the gene most significantly positively correlate with the bias, are assigned as preferred codons (Hershberg and Petrov, 2009).

tRNA abundance does not correlate with codon usage in many codon boxes

In another approach, a codon is termed preferred codon if it is recognized by the tRNA that has either the highest copy number in the genome or the highest cellular expression level. This method of course does not work for synonymous codons that are recognized by tRNAs with equal copy numbers. In addition, decoding is highly redundant and wobble decoding is often as efficient as decoding of cognate codons (Johansson et al., 2008; Kollmar and Mühlhausen, 2017a; Rojas et al., 2018). In many codon boxes that codon, for which there are the most tRNA gene copies, is more used in highly expressed genes than any of the other synonymous codons of the box. In contrast, the usage of the wobble decoded GGU- and UGU-codons, for which cognate tRNAs are absent in all yeast genomes available to date, are more frequently used across all *S. cerevisiae* genes than the corresponding synonymous codons with the high-copy cognate tRNAs (Mühlhausen et al., 2018), is considerably increased in the highly expressed yeast genes. Also the frequency of, for example, the AUC-, ACC-, and GUC-codons, for which cognate tRNAs also do not exist, increases more strongly in those genes whose products are detectably present in the proteomes compared to the usage frequency of synonymous codons with cognate tRNAs. Similarly, the usage frequency of UUG triplicates in highly expressed genes while the frequency of UUA decreases although the number of cognate tRNAs is very similar (ten cognate tRNAs compared to seven, respectively). Thus, total tRNA abundance does not correlate in all codon boxes with codon usage in highly expressed genes. We therefore refrained from designing genes based on tRNA presence and abundance data.

Protein abundance as reference for codon usage bias

All these approaches are confined by extrapolation of limited data or assumptions on codon evolution models. To overcome these limitations, we suggest using the term preferred codon only for codons in genes whose protein products have the highest measured abundance in the cell. As a first and rough estimation available DNA microarray data can be used (Lanza et al., 2014). Even more, quantitative protein abundance data are available at PaxDB for many organisms and can be segmented by absolute or relative criteria. The difference between the codon usage derived from the genome versus that present in a selected proteome can best be visualized in GPome-plots (genome versus proteome plots; 3.5), which have been introduced recently (Mühlhausen et al., 2018). Proteins

with the highest abundance in *S. cerevisiae* show the largest codon usage bias, and genes with no detectable translation have a correspondingly inverted codon usage. The analysis also shows that several codons are almost not used at all, similar to the so-called [RIL]-codons in *E. coli* bacteria, but that there is not a single codon box, in which one of the synonymous codons is used exclusively (except for the trivial one-codon one-amino acid boxes). Accordingly, selecting only the preferred codons for heterologous gene design will cause highly biased and atypical codon usage patterns.

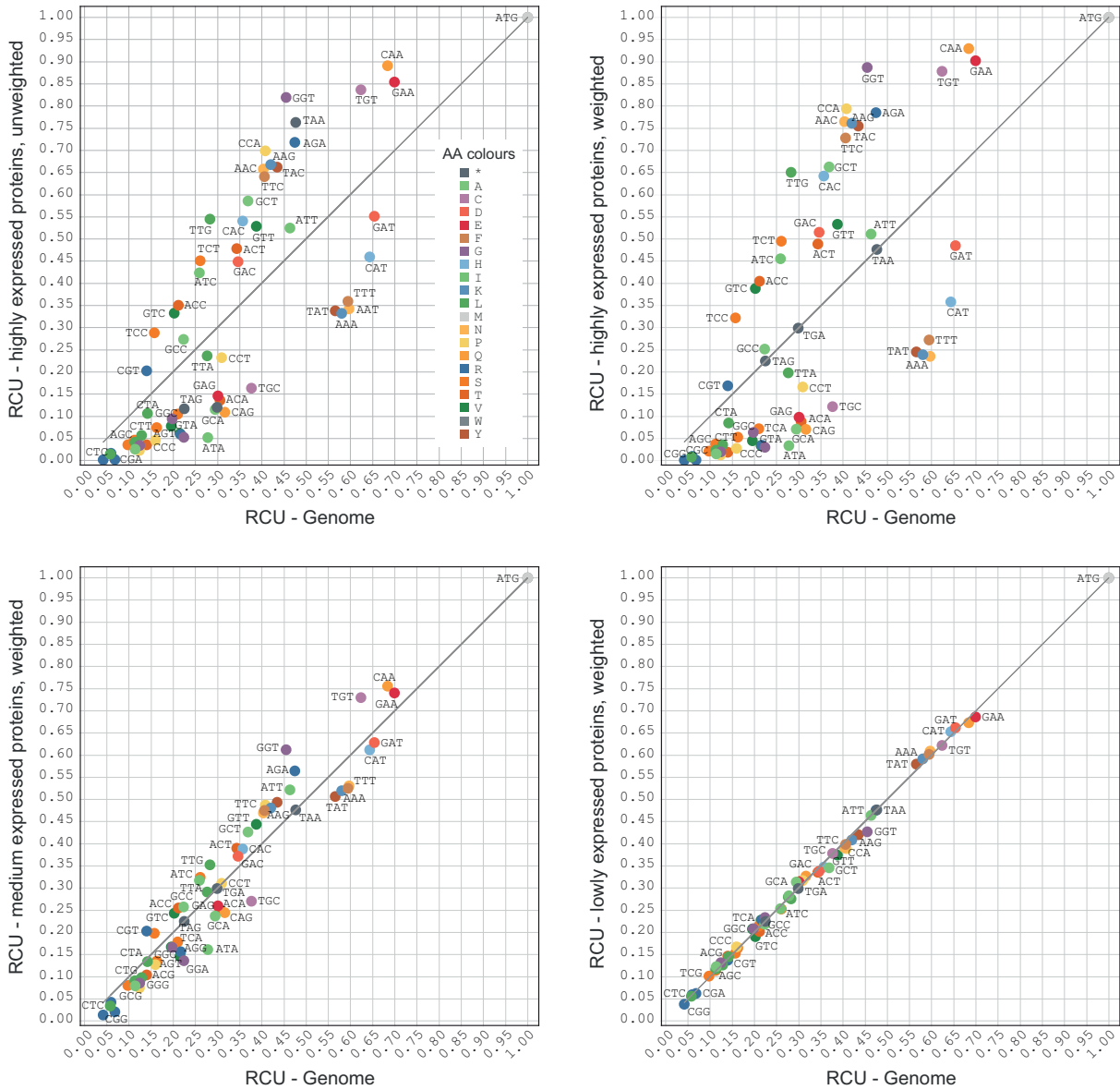


Figure 3.5. | GPome-plots (genome versus proteome plots). The plots show the relative codon usage (RCU) of the 50 most expressed proteins ("highly expressed"), the following 250 proteins with medium expression, and the 2500 least expressed proteins of *S. cerevisiae* plotted against the RCU of all predicted yeast genes (x-axis). For comparison, the RCUs of the highly expressed proteins are shown unweighted and weighted. Weighting means that each gene is multiplied by its absolute abundance as given by the PaxDB data.

The protein abundance does not decrease in steps but exponentially across all proteins (see PaxDB for data). Of course, the abundance is not a function of codon usage alone. In order to reflect the very different abundance across large sets of proteins (e.g. the 50 or 250 most highly expressed proteins), we introduced a weighting scheme. While in the unweighted RCU each codon of each gene in the selected set of proteins is counted once, in the weighted RCU each codon of each gene is multiplied with the abundance of the protein according to the PaxDB data (Figure 3.5).

Typical genes for every purpose

For synthetic biology applications there is not only need for protein production (e.g. highest protein expression) but also for functional studies at expression levels comparable to that of endogenous proteins, or at low levels to avoid toxic effects to name a few. Thus, it would be favourable to generate heterologous DNA sequences for every expression level or expression purpose. We suggest terming such heterologous sequences "typical genes" as they are intended to resemble other genes with a similar expression level. To generate such typical gene sequences we developed Odysseus, which is available online at <http://odysseus.motorprotein.de>. Its core feature is the adaptation of coding sequences to the characteristics of a pre-selected reference gene set of a host organism to increase or decrease the expression rates of the designed proteins. This is done by using a probabilistic model in form of a Markov chain with the RCU as stationary probabilities and the RSdCU as transition probabilities, both trained with host-specific codon-usage information. Odysseus does not generate a single, perfectly optimized gene but provides multiple genes that are all equally typical with respect to the selected reference gene set. The initial set of typical genes is subsequently filtered by excluding candidate genes containing user-defined features such as unwanted enzyme restriction sites. We think that excluding entire genes (and re-generating more typical genes in case all candidates contain unwanted enzyme restriction sites, for example) is the better solution compared to changing a typical gene sequence (to remove unwanted enzyme restriction sites, for example), because the latter might change the di-codon usage to non-typical patterns. In principle, any additional local feature could be implemented to be filtered at this stage as well, such as patterns resembling Shine-Dalgarno consensus sequences, premature poly(A) translation termination sites, CpG islands, cryptic splice sites, dyad repeat sequences or RNase E cleavage sites. The filtered sequences are subjected to RNA secondary structure prediction and the final set of typical genes is presented in a comparative view. Here, the user can inspect the characteristics of each sequence (e.g. RNA secondary structure, restriction sites) and select sequences for DNA synthesis. For validation of the new approach, we choose a highly structured protein, GFP, and an intrinsically disordered protein, human α Syn.

Expression of typical genes encoding GFP in *S. cerevisiae*

GFP is a compact, stable beta-barrel forming protein derived from the jellyfish *Aequorea victoria* that can be expressed in almost every organism (Tsien, 1998). Therefore, it is used as a *gold standard* for testing gene design algorithms and analysing protein expression characteristics such as translational efficiency and accuracy. To test the Odysseus algorithm, we generated genes based on the codon usage of the 5024 lowest expressed (abundance-threshold 88.8 ppm) and of the 308 highest expressed proteins (abundance-threshold 663.0 ppm) in *S. cerevisiae* as determined by PaxDB data (dataset 4932-WHOLE_ORGANISM-integrated.txt; weighted average of all *S. cerevisiae* WHOLE_ORGANISM datasets). In both cases we used the weighting scheme as described above.

For precise comparison of the protein levels, we avoided the typical plasmid-borne expression of the designed genes that might reflect variations in the plasmid copy number in different cell populations. Therefore, yeast strains were generated with genomically integrated either one, two or three copies of the designed GFP-encoding genes, driven by the inducible *GAL1* promoter (Figure 3.6, Supplementary Figures A.12 and A.13). As control, we analysed the original *A. victoria* GFP gene without any nucleotide changes. Expression of GFP was induced for 6 h and the protein levels were analysed by Western blot analysis (Figure 3.6A and 3.6B; Supplementary Figure A.14). The results of these expression test support our initial idea of generating typical genes for typical protein expression ranges. The expression of the gene based on the codon usage of the lowest expressed proteins (gene1) is considerably lower than the expression of the gene based on the codon usage of the highest expressed proteins (gene2). The expression of both proteins considerably increases when increasing genomic copy numbers. The expression level of the control is similar to the expression level of the gene based on the highest expressed proteins. Additionally, live-cell fluorescence microscopy was performed with cells, expressing GFP from different genes. Quantification of the GFP fluorescence intensity corroborated the results from the Western blot analysis and revealed similar differences in the GFP fluorescence depending on the codon context (Figure 3.6C and 3.6D).

If our experiments represented typical lowly and highly expressed genes, then genes N-terminally fused with GFP-tag would also represent highly or lowly expressed genes, depending on the GFP sequence. The first 30-50 codons at the 5' end are thought to determine the expression efficiency and protein level (also called ramp sequence) Tuller and Zur (2015) implying that the expression level of 3'-fused genes will be similar to that of GFP alone. These data might explain the observation that the expression of GFP-fused genes often depends on whether the genes are fused to the 5' or 3' end. Instead of supposed folding problems of the fused proteins, the difference in expression level might mainly depend on whether the fused gene resembles a typical lowly or highly expressed gene. Our experiments suggest that the designed GFP resembling lowly expressed genes could be used for studies of cellular protein expression if the expression level needs to resemble endogenous low levels.

Expression of human α -synuclein in *S. cerevisiae*

Intrinsically disordered proteins play important function in cellular signalling and regulation pathways (Wright and Dyson, 2015). As a test case for an intrinsically disordered protein, we choose human α Syn. The protein α Syn has a central role in the pathogenesis of Parkinson's disease (PD). Accumulation of this highly soluble protein leads to aggregation and proteotoxicity in several neurodegenerative diseases (Wong and Krainc, 2017). Expression of human α Syn in yeast faithfully reproduces the molecular mechanisms that results in aggregation and cellular toxicity (Popova et al., 2015; Tenreiro et al., 2017). Importantly, the toxicity is dose-dependent and directly correlates with α Syn expression level Outeiro (2003); Petroi et al. (2012). We used the advantages of this humanized yeast model, where the toxic effects depend on α Syn gene expression and assessed, whether the toxicity can be rescued by codon adaptation. First, we designed three genes based on subsets of the 5024 lowest expressed proteins in yeast (according to PaxDB; gene4), 1013 proteins with medium expression level (gene5), and the 308 highest expressed proteins (gene6; Figure 3.7). As reference we expressed α Syn from the human coding *SNCA* gene sequence. Yeast strains were generated with genomically integrated one or two copies of the designed α Syn-encoding genes, driven by *GAL1* pro-

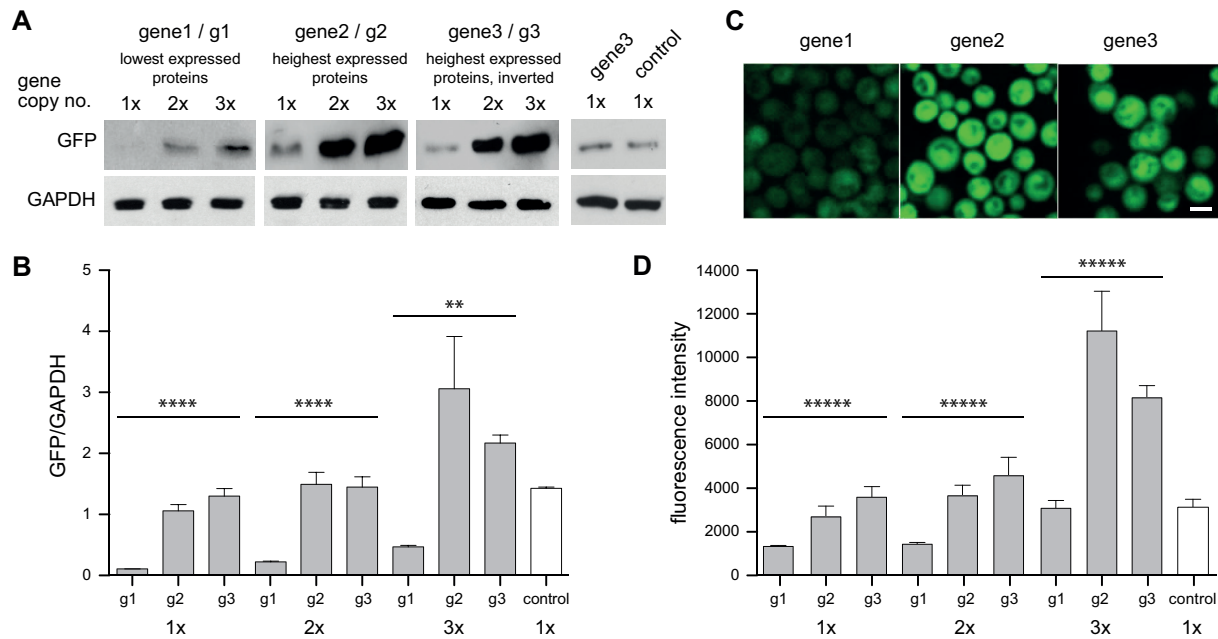


Figure 3.6. | Steady-state protein levels of GFP. Three types of gene design were tested in combination with one to three gene copies. All designed genes are based on the weighting scheme, by which each codon of a subset of genes is multiplied with its expression level as provided by PaxDB data. Gene 1 is based on the subset of the 2,500 least expressed genes, gene 2 is based on the 250 highest expressed genes, and gene 3 is based on the inversion of the codon usage of the highest expressed genes. **A.** Western blot analysis of crude protein extracts from yeast strains, expressing GAL1-driven GFP from one, two and three copies. Protein expression was induced for 6 h in galactose-containing medium, crude protein extracts were prepared and equal protein amounts from all samples were used for Western blotting. The membrane was probed with anti-GFP antibody. GAPDH antibody was used as a loading control. The full-sized blots are available in Supplementary Figure A.13. **B.** Quantification of the protein levels of GFP. Densitometric analysis of the immunodetection of GFP, relative to GAPDH loading control. The significance of the differences was calculated with a One-way Anova-test (**, $p = 0.002$; ****, $p < 0.0001$; $n = 3$). **C.** Life-cell fluorescence microscopy of yeast cells, expressing GFP from three copies. Scale bar: 5 μm . **D.** Quantification of the fluorescence intensity of GFP-expressing cells with different copy numbers and coding sequences. The mean fluorescence intensities were quantified using SlideBook6 software package ($n=100$ per strain, except $n=200$ for the control). The significance of the differences was calculated with a One-way Anova-test (****, $p = 0.0$).

motor (Supplementary Figure S1). Expression of αSyn was induced for 6 h and the protein levels were analysed by Western blot analysis (Figure 3.7A and 3.7B). Surprisingly, the designed genes showed considerably lower expression compared to the human reference, although their codon composition had been adapted to the yeast host organism. Even more surprisingly, the expression level decreased from the gene based on the lowest expressed proteins to the gene based on the highest expressed proteins. This indicates that adaptation of the codon usage of a gene of interest to the highest expressed proteins of a species does not always yield highest expression, which is consistent with observations of researchers trying to boost expression levels, e.g. for structural biology.

Expression of human α -synuclein in *S. cerevisiae* using weighted codon usages

To exclude that our observation depends on not having used the weighting of the protein abundance levels, we designed genes based on the lowest and highest expressed proteins, respectively, using the weighting scheme (gene7 and gene8, respectively; Figure 3.7C and 3.7D; Supplementary

Figure A.15). Again, for precise comparison of α Syn protein levels strains with one, two or three copies of the designed genes were generated. Western blot analysis revealed significantly lower protein levels than the human reference gene, similar to the results with genes 4 - 6. Next, we assessed whether the differences in α Syn expression level are mediated by an impact of the codon usage on transcription. Comparison of the mRNA levels in these strains showed similar gene expression of the designed genes and the human gene (Figure 3.7E). This suggest that the effect of different codon usage for α Syn is mainly due to its impacts on translation.

Expression of GFP and α -synuclein using an inverted codon usage pattern

To identify potential abnormalities within the human reference α Syn gene with respect to the designed yeast genes, we compared their codon usage. The human reference gene does not include many of the codons, which are preferentially used in the highest expressed genes, nor does it include many of the rare codons (Figure 3.8). Instead, it appears that the human α Syn gene resembles an inverted codon usage scheme of the highest expressed yeast genes. With inverted scheme it is not meant that the codon usage is switched (e.g. if the codon usage of CAA and CAG were 0.93 and 0.17, respectively, switching would mean 0.17 and 0.93 for CAA and CAG; Figure 3.8), but that the codon usage is inverted at the values of the genomic codon usage (e.g. if the genomic codon usage of CAA and CAG were 0.68 and 0.32 and the codon usage of these codons across the highest expressed genes were 0.93 and 0.07, respectively, inverting the codon usage would result in codon usage of 0.43 and 0.57 for CAA and CAG, respectively; Figure 3.8; Supplementary Figure A.16). To test whether genes designed with such an inverted codon usage scheme still resemble typical genes, we designed and tested a GFP gene with inverted codon usage (gene3; Figure 3.6). The protein expression level of this GFP gene rather resembled that of the gene based on the highest expressed yeast proteins than that of the gene based on the lowest expressed proteins. The designed α Syn gene based on the inverted codon usage did not result in higher expression, however, compared to the genes without inverting the codon usage (Figure 3.7C and 3.7D). These two test cases demonstrate that genes based on an inverted codon usage can well be expressed. This way rare codons are used more often, although not exclusively. The comparable expression levels of GFP based on the inverted and the highly expressed genes imply that tRNA abundance, which is commonly assumed to be lower for the rare codons, does not determine expression level. However, more tests with more types of genes are needed to fully assess, in which cases the inverted codon usage pattern might be preferable to other patterns.

Finally, we assessed whether the observed low protein levels affect α Syn induced toxicity, reflected as growth retardation. Growth assays were performed with α Syn, expressed from different gene copy numbers and codon context. Expression of α Syn from three copies of gene8 did not reveal toxic phenotype in contrast to its human counterpart that is toxic in yeast (Figure 3.7F) (Petroi et al., 2012). These results demonstrate that we could successfully implement the newly developed design algorithm for reducing the expression level of a toxic proteins in yeast, exemplified by the use of α Syn.

Comparison to other software

There are multiple services for adapting gene sequences to heterologous hosts or for designing gene sequences from scratch. For example, the TISIGNER software has been developed to adjust translation initiation sites by optimizing mRNA accessibility (reducing mRNA secondary structures) and can thus be used to optimize the 5'-end of a gene (Bhandari et al., 2021). TISIGNER requires at least

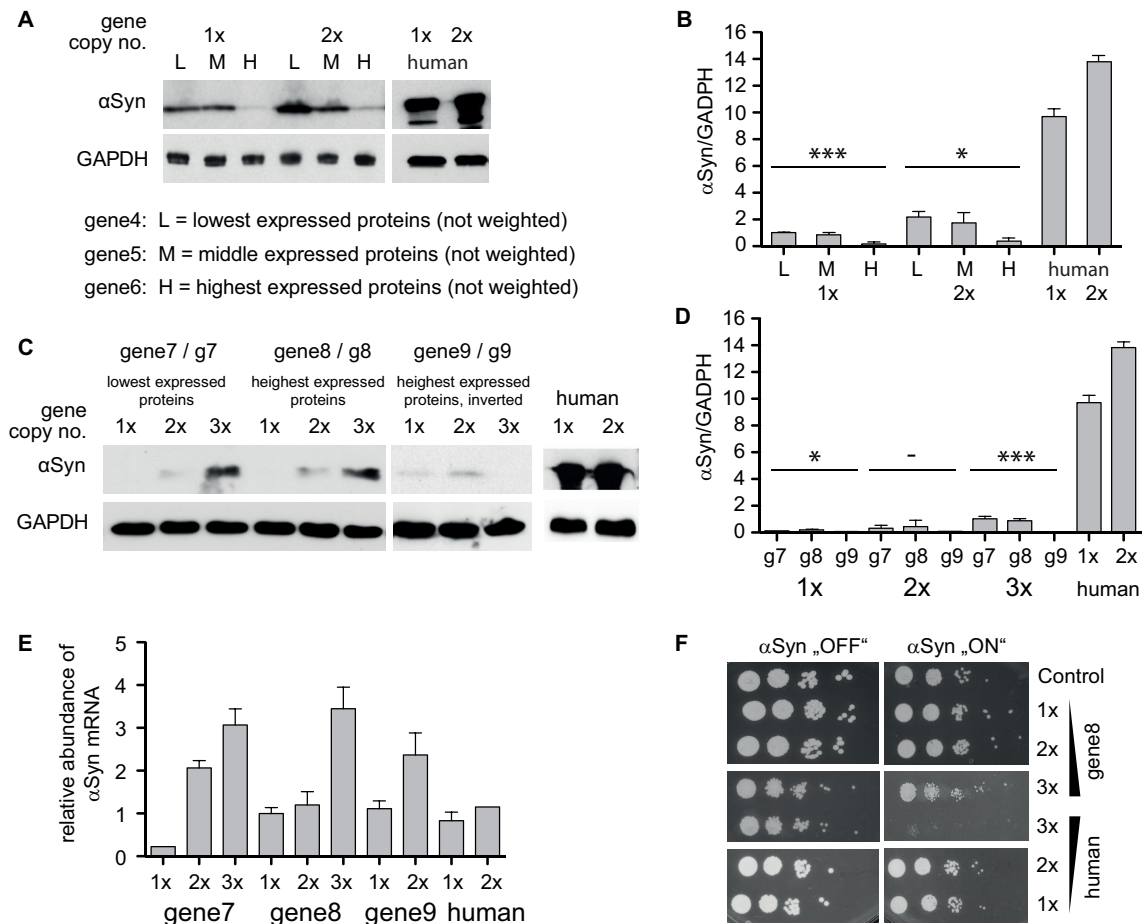


Figure 3.7. | Expression of designed and human α -synuclein. **A.** Western blot analysis for determination of the protein level of α Syn. Protein expression was induced for 6 h, crude protein extracts were prepared and the protein concentrations were determined with a Bradford assay. 160 μ g crude protein extract from samples gene 4 (L), gene 5 (M) and gene 6 (H), and 40 μ g from samples "human" were used for Western blotting. The membrane was probed with anti α Syn antibody. GAPDH antibody was used as a loading control. The full-sized blots are available in Supplementary Figure A.14. **B.** Quantification of the protein levels of α Syn. Densitometric analysis of the immunodetection of α Syn, relative to GAPDH loading control. The significance of the differences was calculated with a One-way Anova-test (*, $p = 0.0107$; ***, $p = 0.00014$; $n = 3$). **C.** Western blot analysis of crude protein extracts from yeast strains, expressing *GAL1*-driven α Syn from one, two and three copies. Protein expression was induced for 6 h, crude protein extracts were prepared and the protein concentrations were determined with a Bradford assay. 160 μ g crude protein extract from samples gene 7, gene 8 and gene 9, and 40 μ g from samples "human" were used for Western blotting. The membrane was probed with anti- α Syn antibody. GAPDH antibody was used as a loading control. The full-sized blots are available in Supplementary Figure A.15. **D.** Quantification of the steady-state protein level of α Syn. Densitometric analysis of the immunodetection of α Syn, relative to GAPDH loading control ($n=3$). The significance of the differences was calculated with a One-way Anova-test (*, $p = 0.038$; -, $p = 0.41$; ***, $p = 0.00034$; $n = 3$). **E.** Quantification of SNCA gene expression. RNA was prepared from yeast strains after 6 h induction of α Syn expression. Relative α Syn mRNA levels were determined by qRT-PCR and normalized against H2A. Expression values represent the mean of three replicates \pm standard error. **F.** Growth analysis of yeast cells expressing α Syn from one, two and three gene copies, driven by the inducible *GAL1*-promoter on non-inducing ("OFF": glucose) and inducing ("ON": galactose) SC-URA medium after 3 days. Yeast cells expressing GFP from the same promoter were used as a control.

part of the 5'-UTR as input in addition to the gene sequence. As discussed above, most software to design entire gene sequences optimize the CAI. However, the used codon usage tables most of-

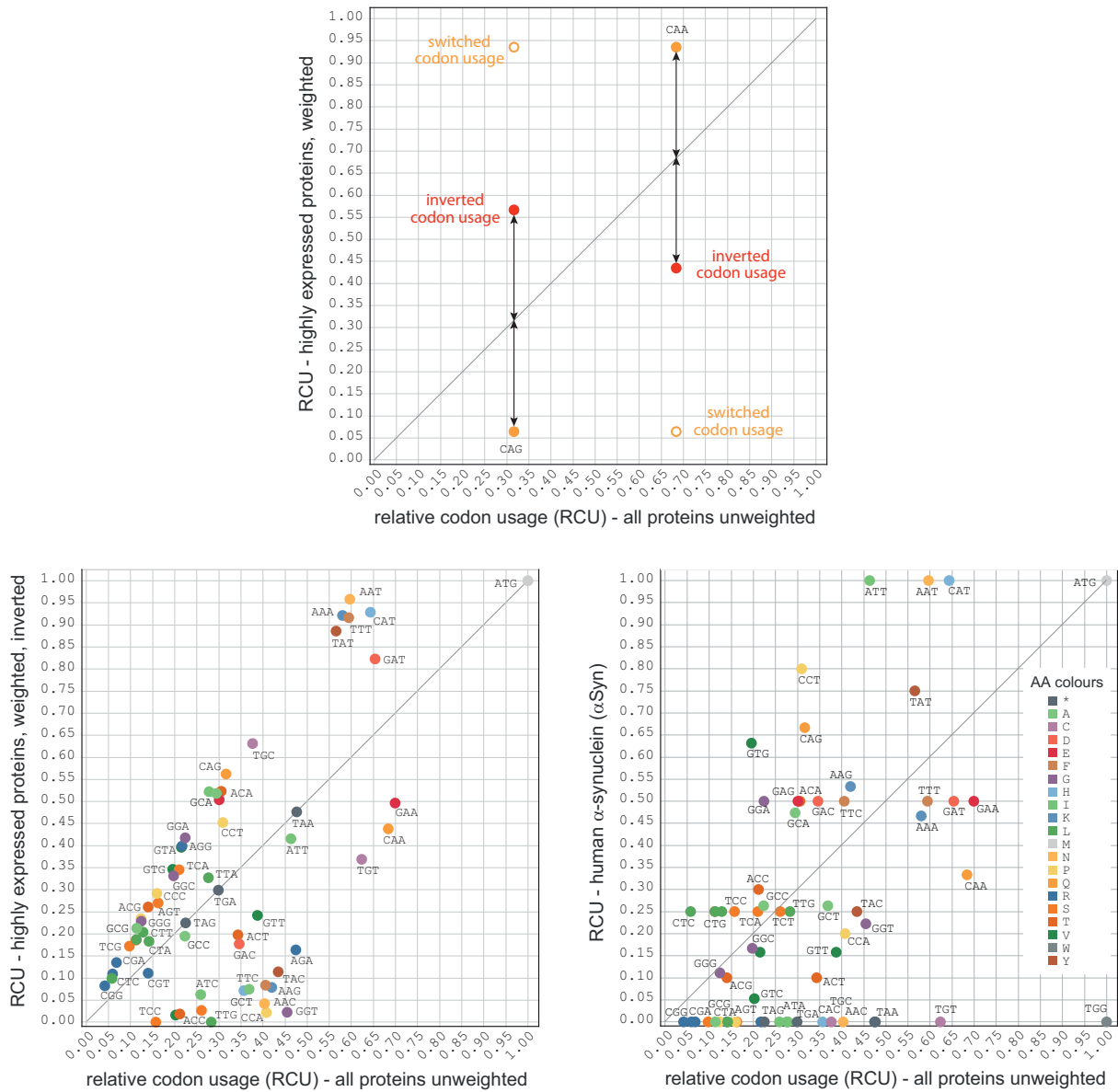


Figure 3.8. | The "inverted" codon usage. The schematic view at the top demonstrates the generation of an inverted codon usage compared to that of a switched codon usage. The plots at the bottom show the relative codon usage of the 308 highest expressed proteins, when weighted and inverted (left plot), and the relative codon usage of human α Syn (right plot).

ten do not refer to the codon usage of only the highly expressed genes but that of all genes, which rather resembles the usage of the lowly expressed genes. For example, the Gene Designer v1 software (the algorithm of v2 is not published and could be different) uses the frequency distribution for each codon box based on a codon usage table, which according to the online documentation does not correspond to the usage of the highly expressed genes in yeast (Villalobos et al., 2006). The tool OPTIMIZER allows the upload of a user-provided codon usage table and selection of always the most used codon, random selection by frequency distribution or manual selection of codons (Puigbò et al., 2007). In a very recent approach, ChimeraUGEM, a target gene is designed by comparing its protein

sequences by a longest substring approach to a set of reference sequences assuming that longer repetitive substrings became more optimized to the hosts translation machinery (Diament et al., 2019). The algorithm has been used for predicting gene expression levels, and has also been shown to be successful in increasing expression of a synthetic gene in green algae. The ChimeraUGEM approach seems to be most related to our typical gene approach, although ChimeraUGEM designs the genes substring by substring (not overlapping) and does not explicitly aim to optimize di-codons.

Limitations of the Odysseus implementation

It is well known that there is additional information in the coding sequence of a gene beyond the genetic code for translating nucleotide triplets (codons). For example, the UGA stop codon is translated to selenocysteine if a so-called SECIS pattern is present (Mariotti et al., 2013; Peng et al., 2021). There is also a process termed programmed ribosomal frameshifting by which the ribosome shifts the reading frame by one or two nucleotides in either the + or the – direction (Caliskan et al., 2015). The patterns of these two examples, selenocysteine decoding and ribosomal frameshifting, dictate the protein sequence. These patterns are not implemented yet. The Odysseus tool does also not allow to adjust the gene sequence to other genetic codes than the standard genetic code which could be a useful extension. It is recommended that users select one of those suggested typical genes, that do not contain the reassigned codon (e.g. does not contain a CTG codon if protein expression in *Candida albicans* is wanted). However, the designed genes might not be typical anymore in species that heavily use the reassigned codons such as several ciliates, that decode stop codons by glutamine (Kollmar and Mühlhausen, 2017b). In addition to codes leading to different protein sequences, there are also various regulatory signals on top of the coding region that affect the gene expression and protein translation levels (Bergman and Tuller, 2020). Odysseus is not aware of these signals and the designed genes might miss important signals (if wanted) or by chance introduce unwanted signals. As far as those signals or sequence patterns are known a user could manually detect this and select another of the set of typical genes that Odysseus generates.

Conclusions

Odysseus is a new software tool to design typical genes for heterologous protein expression. In contrast to most other tools, which intend to optimize the codon usage by selecting only codons from a few highly expressed proteins or by selecting only the codon with the highest relative codon usage from each codon box, Odysseus generates genes resembling the codon usage of a selected group of proteins. Such groups can be the highest or lowest expressed proteins of a species (with the cut-off free to choose), or even a subset of proteins with a certain function. We tested the new system by generating synthetic genes of the non-toxic, highly structured protein GFP and by evaluating their expression level. The expression level strongly increased from the gene based on the lowest expressed proteins to the gene based on the highest expressed proteins. This supports the general finding that protein expression is stronger when adapting a heterologous gene to the most used codons. Such a strong expression is, however, often not wanted and disfavoured when trying to express a toxic protein. To test our software for its use for expressing proteins at low endogenous protein expression levels, we designed synthetic genes for the toxic, non-structured protein α -synuclein and showed that human α Syn can be adapted to low expression levels. Although further tests with more proteins are needed our results suggest that Odysseus is a valuable tool for designing typical genes for heterologous protein expression.

Data Availability

The software can be used via a web interface at <http://odysseus.motorprotein.de>, and obtained from GitHub at <https://github.com/dsimm/Odysseus> for local installation and use.

Funding

GHB acknowledge the support of the Deutsche Forschungsgemeinschaft (DFG: BR1502/18-1).

Competing Interests Statement

The authors declare no competing interests.

References

- Bergman, S. and Tuller, T. (2020). Widespread non-modular overlapping codes in the coding regions. *Phys Biol*, 17(3):031002.
- Bhandari, B. K., Lim, C. S., and Gardner, P. P. (2021). TISIGNER.com: web services for improving recombinant protein production. *Nucleic Acids Res*, 49(W1):W654–W661.
- Boycheva, S., Chkodrov, G., and Ivanov, I. (2003). Codon pairs in the genome of Escherichia coli. *Bioinformatics*, 19(8):987–998.
- Brule, C. E. and Grayhack, E. J. (2017). Synonymous Codons: Choose Wisely for Expression. *Trends Genet*, 33(4):283–297.
- Caliskan, N., Peske, F., and Rodnina, M. V. (2015). Changed in translation: mRNA recoding by 1 programmed ribosomal frameshifting. *Trends Biochem Sci*, 40(5):265–274.
- Coleman, J. R., Papamichail, D., Skiena, S., Futcher, B., Wimmer, E., and Mueller, S. (2008). Virus attenuation by genome-scale changes in codon pair bias. *Science*, 320(5884):1784–1787.
- Diamant, A., Weiner, I., Shahar, N., Landman, S., Feldman, Y., Atar, S., Avitan, M., Schweitzer, S., Yacoby, I., and Tuller, T. (2019). ChimeraUGEM: unsupervised gene expression modeling in any given organism. *Bioinformatics*, 35(18):3365–3371.
- Gaspar, P., Oliveira, J. L., Frommlet, J., Santos, M. A. S., and Moura, G. (2012). EuGene: maximizing synthetic gene design for heterologous expression. *Bioinformatics*, 28(20):2683–2684.
- Gietz, D., St Jean, A., Woods, R. A., and Schiestl, R. H. (1992). Improved method for high efficiency transformation of intact yeast cells. *Nucleic Acids Res.*, 20(6):1425.
- Gould, N., Hendy, O., and Papamichail, D. (2014). Computational tools and algorithms for designing customized synthetic genes. *Front Bioeng Biotechnol*, 2:41.
- Gustafsson, C., Minshull, J., Govindarajan, S., Ness, J., Villalobos, A., and Welch, M. (2012). Engineering genes for predictable protein expression. *Protein Expr. Purif.*, 83(1):37–46.
- Guthrie, C. and Fink, G. (1991). Guide to yeast genetics and molecular biology. *Meth. Enzymol.*, 194:1–863.

- Gutman, G. A. and Hatfield, G. W. (1989). Nonrandom utilization of codon pairs in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.*, 86(10):3699–3703.
- Hansen, J., Mailand, E., Swaminathan, K. K., Schreiber, J., Angelici, B., and Benenson, Y. (2014). Transplantation of prokaryotic two-component signaling pathways into mammalian cells. *Proc Natl Acad Sci U S A*, 111(44):15705–15710.
- Hanson, G. and Collier, J. (2018). Codon optimality, bias and usage in translation and mRNA decay. *Nat Rev Mol Cell Biol*, 19(1):20–30.
- Hedfalk, K. (2012). Codon optimisation for heterologous gene expression in yeast. *Methods Mol Biol*, 866:47–55.
- Hershberg, R. and Petrov, D. A. (2009). General rules for optimal codon choice. *PLoS Genet.*, 5(7):e1000556.
- Hia, F., Yang, S. F., Shichino, Y., Yoshinaga, M., Murakawa, Y., Vandenbon, A., Fukao, A., Fujiwara, T., Landthaler, M., Natsume, T., Adachi, S., Iwasaki, S., and Takeuchi, O. (2019). Codon bias confers stability to human mRNAs. *EMBO Rep*, 20(11):e48220.
- Jansen, R., Bussemaker, H. J., and Gerstein, M. (2003). Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. *Nucleic Acids Res*, 31(8):2242–2251.
- Johansson, M. J. O., Esberg, A., Huang, B., Björk, G. R., and Byström, A. S. (2008). Eukaryotic wobble uridine modifications promote a functionally redundant decoding system. *Mol. Cell. Biol.*, 28(10):3301–3312.
- Kato, Y. (2019). Translational Control using an Expanded Genetic Code. *Int J Mol Sci*, 20(4).
- Knop, M., Siegers, K., Pereira, G., Zachariae, W., Winsor, B., Nasmyth, K., and Schiebel, E. (1999). Epitope tagging of yeast genes using a PCR-based strategy: more tags and improved practical routines. *Yeast*, 15(10B):963–972.
- Kollmar, M. and Mühlhausen, S. (2017a). How tRNAs dictate nuclear codon reassignments: Only a few can capture non-cognate codons. *RNA Biol*, 14(3):293–299.
- Kollmar, M. and Mühlhausen, S. (2017b). Nuclear codon reassignments in the genomics era and mechanisms behind their evolution. *Bioessays*, 39(5):1600221.
- Lanza, A. M., Curran, K. A., Rey, L. G., and Alper, H. S. (2014). A condition-specific codon optimization approach for improved heterologous gene expression in *Saccharomyces cerevisiae*. *BMC Syst Biol*, 8:33.
- Lorenz, R., Bernhart, S. H., Siederdissen, C. H. z., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, 6(1):26.
- Mariotti, M., Lobanov, A. V., Guigo, R., and Gladyshev, V. N. (2013). SECISearch3 and Seblastian: new tools for prediction of SECIS elements and selenoproteins. *Nucleic Acids Res*, 41(15):e149.
- Mauro, V. P. (2018). Codon Optimization in the Production of Recombinant Biotherapeutics: Potential Risks and Considerations. *BioDrugs*, 32(1):69–81.

- Meade, R. M., Fairlie, D. P., and Mason, J. M. (2019). Alpha-synuclein structure and Parkinson's disease – lessons and emerging principles. *Molecular Neurodegeneration*, 14(1).
- Michalodimitrakis, K. and Isalan, M. (2009). Engineering prokaryotic gene circuits. *FEMS Microbiol Rev*, 33(1):27–37.
- Mühlhausen, S., Schmitt, H. D., Pan, K.-T., Plessmann, U., Urlaub, H., Hurst, L. D., and Kollmar, M. (2018). Endogenous Stochastic Decoding of the CUG Codon by Competing Ser- and Leu-tRNAs in *Ascoidea asiatica*. *Curr. Biol.*, 28(13):2046–2057.e5.
- NCBI Resource Coordinators (2018). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, 46(D1):D8–D13.
- Nierhaus, K. H., Wadzack, J., Burkhardt, N., Jünemann, R., Meerwinck, W., Willumeit, R., and Stuhmann, H. B. (1998). Structure of the elongating ribosome: arrangement of the two tRNAs before and after translocation. *Proc. Natl. Acad. Sci. U.S.A.*, 95(3):945–950.
- Nieuwkoop, T., Finger-Bou, M., van der Oost, J., and Claassens, N. J. (2020). The Ongoing Quest to Crack the Genetic Code for Protein Production. *Mol Cell*, 80(2):193–209.
- Outeiro, T. F. (2003). Yeast Cells Provide Insight into Alpha-Synuclein Biology and Pathobiology. *Science*, 302(5651):1772–1775.
- Peng, J.-J., Yue, S.-Y., Fang, Y.-H., Liu, X.-L., and Wang, C.-H. (2021). Mechanisms Affecting the Biosynthesis and Incorporation Rate of Selenocysteine. *Molecules*, 26(23):7120.
- Petroi, D., Popova, B., Taheri-Talesh, N., Irniger, S., Shahpasandzadeh, H., Zweckstetter, M., Outeiro, T. F., and Braus, G. H. (2012). Aggregate clearance of α -synuclein in *Saccharomyces cerevisiae* depends more on autophagosome and vacuole function than on the proteasome. *J. Biol. Chem.*, 287(33):27567–27579.
- Popova, B., Kleinknecht, A., and Braus, G. (2015). Posttranslational Modifications and Clearing of α -Synuclein Aggregates in Yeast. *Biomolecules*, 5(2):617–634.
- Puigbò, P., Guzmán, E., Romeu, A., and Garcia-Vallvé, S. (2007). OPTIMIZER: a web server for optimizing the codon usage of DNA sequences. *Nucleic Acids Res.*, 35(Web Server issue):W126–131.
- Roberts, R. J., Vincze, T., Posfai, J., and Macelis, D. (2015). REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.*, 43(Database issue):D298–299.
- Rodnina, M. V. (2018). Translation in Prokaryotes. *Cold Spring Harb Perspect Biol*, 10(9).
- Rojas, J., Castillo, G., Leiva, L. E., Elgamal, S., Orellana, O., Ibba, M., and Katz, A. (2018). Codon usage revisited: Lack of correlation between codon usage and the number of tRNA genes in enterobacteria. *Biochem. Biophys. Res. Commun.*, 502(4):450–455.
- Sharp, P. M. and Li, W. H. (1987). The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, 15(3):1281–1295.
- Stark, H., Orlova, E. V., Rinke-Appel, J., Jünke, N., Mueller, F., Rodnina, M., Wintermeyer, W., Brimacombe, R., and van Heel, M. (1997). Arrangement of tRNAs in pre- and posttranslocational ribosomes revealed by electron cryomicroscopy. *Cell*, 88(1):19–28.

- Taneda, A. and Asai, K. (2020). COSMO: A dynamic programming algorithm for multicriteria codon optimization. *Comput Struct Biotechnol J*, 18:1811–1818.
- Tats, A., Tenson, T., and Remm, M. (2008). Preferred and avoided codon pairs in three domains of life. *BMC Genomics*, 9(1):463.
- Tenreiro, S., Franssens, V., Winderickx, J., and Outeiro, T. F. (2017). Yeast models of Parkinson's disease-associated molecular pathologies. *Current Opinion in Genetics & Development*, 44:74–83.
- Tsien, R. Y. (1998). The green fluorescent protein. *Annu Rev Biochem*, 67:509–544.
- Tuller, T. and Zur, H. (2015). Multiple roles of the coding sequence 5 end in gene expression regulation. *Nucleic Acids Res*, 43(1):13–28.
- Villalobos, A., Ness, J. E., Gustafsson, C., Minshull, J., and Govindarajan, S. (2006). Gene Designer: a synthetic biology tool for constructing artificial DNA segments. *BMC Bioinformatics*, 7:285.
- Wang, M., Herrmann, C. J., Simonovic, M., Szklarczyk, D., and von Mering, C. (2015). Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics*, 15(18):3163–3168.
- Welch, M., Villalobos, A., Gustafsson, C., and Minshull, J. (2009). You're one in a googol: optimizing genes for protein expression. *J. R. Soc. Interface*, 6(Suppl 4):S467–S476.
- Wong, Y. C. and Krainc, D. (2017). α -synuclein toxicity in neurodegeneration: mechanism and therapeutic strategies. *Nature Medicine*, 23(2):1–13.
- Wright, P. E. and Dyson, H. J. (2015). Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology*, 16(1):18–29.
- Zimmer, M. (2002). Green Fluorescent Protein (GFP): Applications, Structure, and Related Photo-physical Behavior. *Chemical Reviews*, 102(3):759–782.

3.3. Genome annotation and genomic databases

3.3.1. diArk – the database for eukaryotic genome and transcriptome assemblies in 2014

Martin Kollmar *, Lotte Kollmar, Björn Hammesfahr and Dominic Simm

* Corresponding author

Group Systems Biology of Motor Proteins, Department of NMR-based Structural Biology, Max-Planck-Institute for Biophysical Chemistry, Göttingen, Germany

Nucleic Acids Research (2014) - Volume 43(D1): D1107–D1112,

doi: 10.1093/nar/gku990

<https://academic.oup.com/nar/article/43/D1/D1107/2439940>

Own contribution

MK specified the requirements from a user's perspective, defined the rules for data handling, and collected all the data. DS updated the database scheme, set up the technical requirements and did the technical design and programming. DS prepared the figures. LK collected data. BH contributed to technical design and programming. MK wrote the manuscript. All authors read and approved the final manuscript.

diArk – the database for eukaryotic genome and transcriptome assemblies in 2014

Martin Kollmar*, Lotte Kollmar, Björn Hammesfahr and Dominic Simm

Group Systems Biology of Motor Proteins, Department of NMR-based Structural Biology, Max-Planck-Institute for Biophysical Chemistry, Göttingen, 37085, Germany

Received September 10, 2014; Accepted September 30, 2014

ABSTRACT

Eukaryotic genomes are the basis for understanding the complexity of life from populations to the molecular level. Recent technological innovations have revolutionized the speed of data generation enabling the sequencing of eukaryotic genomes and transcriptomes within days. The database diArk (<http://www.diark.org>) has been developed with the aim to provide access to all available assembled genomes and transcriptomes. In September 2014, diArk contains about 2600 eukaryotes with 6000 genome and transcriptome assemblies, of which 22% are not available via NCBI/ENA/DDBJ. Several indicators for the quality of the assemblies are provided to facilitate their comparison for selecting the most appropriate dataset for further studies. diArk has a user-friendly web interface with extensive options for filtering and browsing the sequenced eukaryotes. In this new version of the database we have also integrated species, for which transcriptome assemblies are available, and we provide more analyses of assemblies.

INTRODUCTION

Eukaryotic genome research enormously benefits from the increasing number of sequenced organisms. Whereas in the time of Sanger-sequencing single-species analyses and small-scale comparative projects dominated, the throughput of the Illumina technology allowed initiating and conducting the sequencing of thousands of species. Examples are the Genome 10K project (1), the i5k project (2) and the 959 Nematodes project (3) intending to provide the genome sequences of a broad range of species, and the 1001 Arabidopsis project (4), the 1000 bull project (5) and the 3000 rice project (6) aiming to reveal phenotypic and genetic differences of breeds and varieties of economically important animals and plants. Usually, genome assemblies are generated for new species, whereas in population studies the sequencing reads are mapped against reference genomes without producing independent genome assemblies.

NCBI/ENA/DDBJ are the central repositories for sequence read archives (SRAs), the ‘raw data’ for generating assemblies, but publishers and funding agencies often do not require assemblies to also be stored there. Thus, most large-scale sequencing centers like *The Broad Institute of MIT and Harvard* (Cambridge, MA, USA), the *DOE Joint Genome Institute* (Walnut Creek, CA, USA) and the *Wellcome Trust Sanger Institute* (Cambridge, UK) established own species- and taxa-dedicated databases such as Phytozome for plants (7) and the Fungal Genome Initiative project pages (8). Powered by research community efforts, there are also excellent databases dedicated to single species such as FlyBase (9), WormBase (10) and dictyBase (11), or repositories for species of certain taxonomic branches such as EuPathDB (12), VectorBase (13) and FungiDB (14). Although these databases only comprise model species and related organisms, they are well known far beyond their research communities. In contrast, dedicated databases have been set up for many of the newly sequenced species that are only known to small communities. In addition, for many species it takes years from the first release of a draft assembly to the publication of the genome analysis (e.g. the *Babesia bigemina* genome is available since 2003 but was published in 2014; the *Callithrix jacchus* genome has been made available in 2007 but published in 2014). Therefore, it is necessary to have a database to identify and access all the available data.

The two major manually curated genome project databases are GOLD (15) and diArk (16). Whereas GOLD is mainly focused on microbial genomes, we developed diArk as a central hub for all eukaryotes, for which large-scale transcriptome or genome assembly data have been produced and are available to the public. diArk provides measures and analyses of these assemblies, as well as links to the data generator repositories. Currently, diArk comprises 2577 eukaryotes and provides access to almost 6000 transcriptome and genome assemblies.

*To whom correspondence should be addressed. Tel: +49 551 201 2260; Fax: +49 551 201 2202; Email: mako@nmr.mpibpc.mpg.de

DIARK

Search options and data filtering

diArk provides three main options to search for sequenced eukaryotes. The most straight-forward way to search for a specific species is to use the autocompletion form at diArk's entry page, which is also available in the menu bar and as a browser search plugin. A *FastSearch* allows filtering for taxa, genome type (genome assemblies, EST data, RNA-seq data) and sequencing status (complete and incomplete assemblies), and is currently the most frequently used entry point to diArk. An extended *Search* form provides six search modules that can be combined in any way to search and filter diArk's content. The most widely used modules are the *Taxonomy* module with taxa/species selection integrated into an expandable phylogenetic tree and the *Genome Files* module that provides options to filter for sequencing and assembly methods, genome type, GC content, sequencing coverage and assembly release date. Search options and filters for the new data types generated in the last 3 years have been integrated into the existing modules.

Result views and statistics

The filtered species can be analysed and compared using seven *Result* tabs. The most frequently used are the pre-selected *Species* tab, the *Genome Files* and the *Sequencing Stats* tab. The *Species* tab provides an overview about species-related information such as alternative, common and anamorph (for many fungi) names, a full taxonomy, a list of all external sources providing access to respective sequence data, and all publications related to the respective species' sequencing. In the *Genome Files* view, extensive analyses of all assembly files are shown with clickable icons to inspect P50 and A50 plots, CGRs with resolution up to $k = 10$, and details of the used sequencing and assembly methods. All genome assemblies for a given species are listed below each other for fast comparison. While the other *Result* tabs list data species wise, the *Sequencing Stats* view offers many comparisons of the whole selected/filtered species (16).

CURRENT STATUS OF THE DATABASE

diArk's growth reflects the exponentially increasing availability of sequenced eukaryotes, now (September 5, 2014) comprising 2577 species (806 in 2011, 415 in 2007). For 1999 of these species (613 in 2011, 209 in 2007) whole genome assembly data are available, and for 429 species transcriptome shotgun assemblies (TSAs; Figure 1), of which the first became available end of 2012. Assembly data for 2017 (78.3%) of the eukaryotes are available at NCBI/ENA/DDBJ meaning that data for 560 (21.7%) species can currently only be accessed at other resources. The data for these 560 species have not yet been or might never be submitted to NCBI/ENA/DDBJ. These species include, for example, the recently published fish *Electrophorus electricus* (17) and the stick insect *Timema cristinae* (18), of which only the SRA data but not the genome assemblies have been deposited at NCBI/ENA/DDBJ, several species such as the snake *Boa constrictor constrictor*, whose genome assemblies are only

available at (Giga)ⁿDB database (19), and species whose genome assemblies are only available at the sequencing centers such as 31 nematode and 22 Platyhelminth genomes recently finished by *The Wellcome Trust Sanger Institute*. These examples underline the unique value of diArk for the eukaryote sequencing and research community in providing a central hub integrating data available from single species repositories to large-scale sequencing centers. In the last 3 years, RNA-Seq based transcriptome assemblies have essentially replaced EST and cDNA sequencing efforts. Due to the lower costs and faster accessibility, TSAs have almost passed EST/cDNA data in diArk (423 TSAs versus 654 EST/cDNA projects; Figure 1).

diArk links publications of sequence assemblies to species (Figure 1). At NCBI, there is a master record for each assembly to access the respective data, and to which respective publications are linked. Currently, these publication links (784 links) only comprise 79.6% of the publication links included in diArk (985 links). Examples for species, whose genomes have long been published but are still not linked to the NCBI genome entries, include *Ciona savignii* (submitted to NCBI in 2003, published 2007), *Filobasidiella neoformans B-3501A* (submitted 2004, published 2005), *Pristionchus pacificus* (submitted and published 2008), *Culex pipiens quinquefasciatus* (submitted 2007, published 2010) and *Uncinocarpus reesii* (submitted 2005, published 2009). Instead, the genome assemblies of these 200 species are marked as 'unpublished'. Published and unpublished genomes are important resources for the community but it is also important that data generators get credits for their efforts. On the other hand, embargo rules for unpublished data should indefinitely not prohibit specific analyses (20). At diArk, researchers find the most complete list of references to genome assemblies for proper citation or for selection of appropriate subsets of published species to avoid data usage issues.

In the last years, most of the available genomes have been sequenced with Illumina machines. However, the Sanger method is still used to assist in scaffold and chromosome assembling (Figure 1). Roche's 454 sequencing method is currently the most widely used method for transcriptome shotgun sequencing. Other methods such as SOLiD, PacBio or IonTorrent are still rarely used to generate *de novo* genome or transcriptome assemblies.

NEW DEVELOPMENTS

diArk hosts and analyses whole-genome and transcriptome assemblies. Currently, the about 6000 assemblies comprise mitochondrial, chloroplast, apicoplast, nucleolar and nuclear genomic DNA and are made available to other services such as the gene reconstruction software WebScipio (21). The quality of genome assemblies can vary significantly (22). However, approaches resulting in excellent genomes for one species might not produce assemblies of similar quality in other cases. Therefore, diArk provides access to alternative assemblies and several measures for direct comparison such as number of contigs, genome size (larger = better), N50 value (higher = better), N50 length (higher = better), contig length distributions (A50 and N50 plots), sequencing coverage (higher = better), sequencing meth-

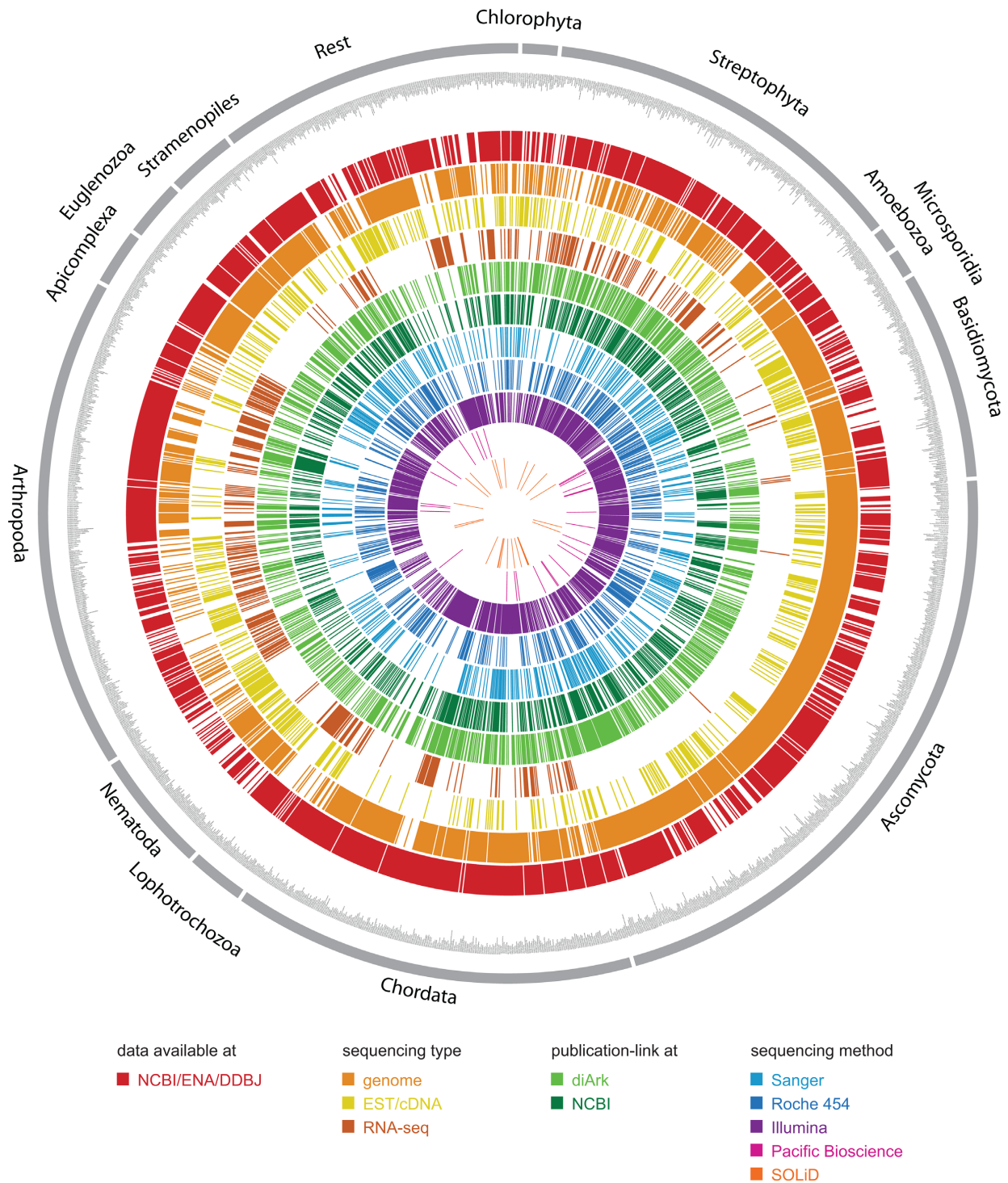


Figure 1. Representation of the nonredundant species (i.e. one strain per species) in diArk with their sequencing type and method. For comparison, all species are marked, for which transcriptome data and/or genome assemblies are available via NCBI/ENA/DDBJ. Nine hundred and eighty five of the assemblies have been published but only 784 of them are linked to the genome assemblies at NCBI.

Number of non-redundant species over the years

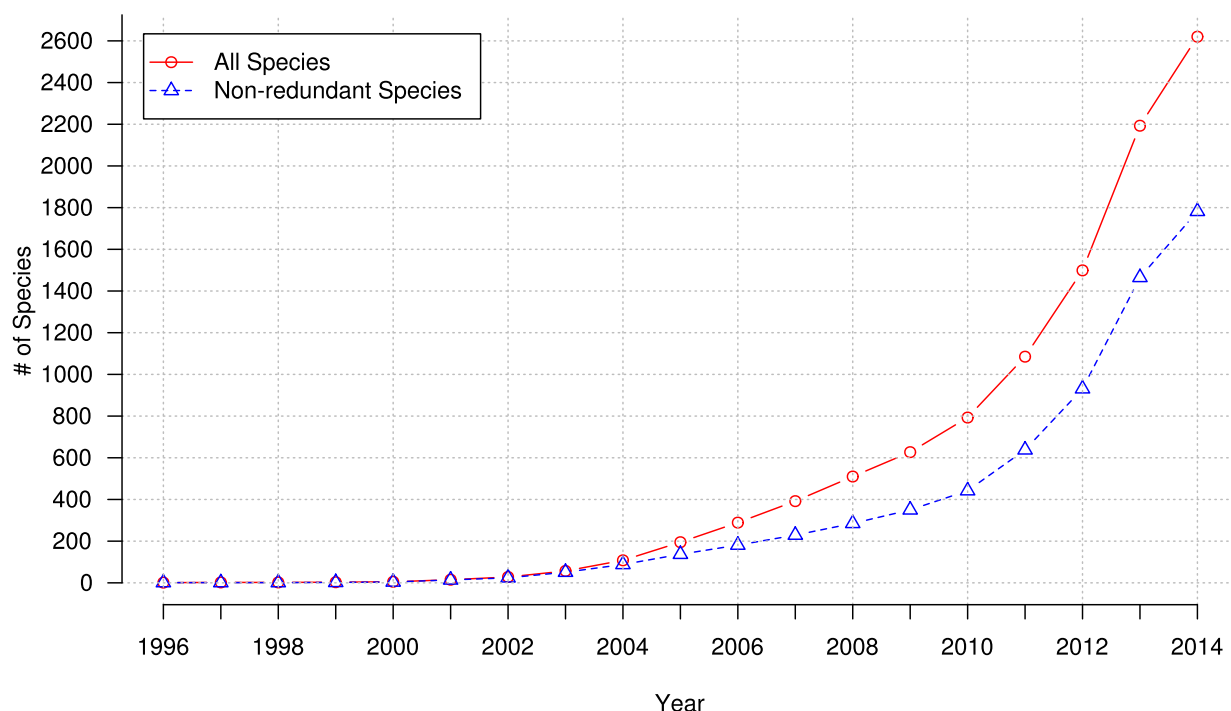


Figure 2. Evolution of the fraction of nonredundant species compared to all sequenced species over time.

ods and used assembly software. Not only the number of alternative assemblies increased in the last years, but also the number of redundant species in terms of species diversity (Figure 2) increased. Redundant species include, for example, different strains of the same fungal species, different breeds of animals, different varieties of plants and different isolates of protozoa. Within diArk, the respective genome and transcriptome assemblies can directly be compared and the most suitable for a certain research hypothesis be identified. diArk also provides chaos game representations (CGRs), which are fingerprints of genomes, and frequency chaos game representations (FCGRs) at different resolutions, which can be used, for example, for phylogenetic reconstructions (23).

Integration of RNA-seq data

The most noticeable innovation from v.2 to v.3 is diArk's integration of RNA-seq data. The first nonhuman transcriptome assemblies have been submitted to and released by NCBI in late 2012. Since then, not only the diversity of sequenced species has increased rapidly (Figure 3) but also the number of species with transcriptome assemblies generated for different developmental stages and/or organs. Given the low costs of transcriptome compared to genome sequencing, the number of species with available transcriptome assemblies will pass the number of species with sequenced genomes in the near future. Several large-scale projects have already been announced and are ex-

pected to release their data this or next year, such as The 1000 plants (oneKP or 1KP) initiative (<https://sites.google.com/a/ualberta.ca/onekp/>), the Marine Microbial Eukaryote Transcriptome Sequencing project (24) and the Fish-T1K project (<http://www.fisht1k.org/>). Interestingly, there is not much overlap between species with transcriptome and genome assemblies (Figure 3). One reason is, that RNA-seq data is still rarely generated for species, for which genome assemblies have been produced, and if generated, the RNA-seq data had been used to assist in genome annotation or to generate expression profiles but not to produce independent transcriptome assemblies. In addition, many scientific questions can be answered sufficiently and faster with transcriptome data.

CONCLUSIONS

Herein, we present an updated version of diArk, which is a central hub for all sequenced eukaryotes, for which either genome or transcriptome assemblies, or large-scale EST/cDNA data are available. diArk is unique in providing direct access to most of the sequenced eukaryotes, whose number has more than tripled compared to the previous version. The number of analysed genome and transcriptome assemblies now reaches 6000.

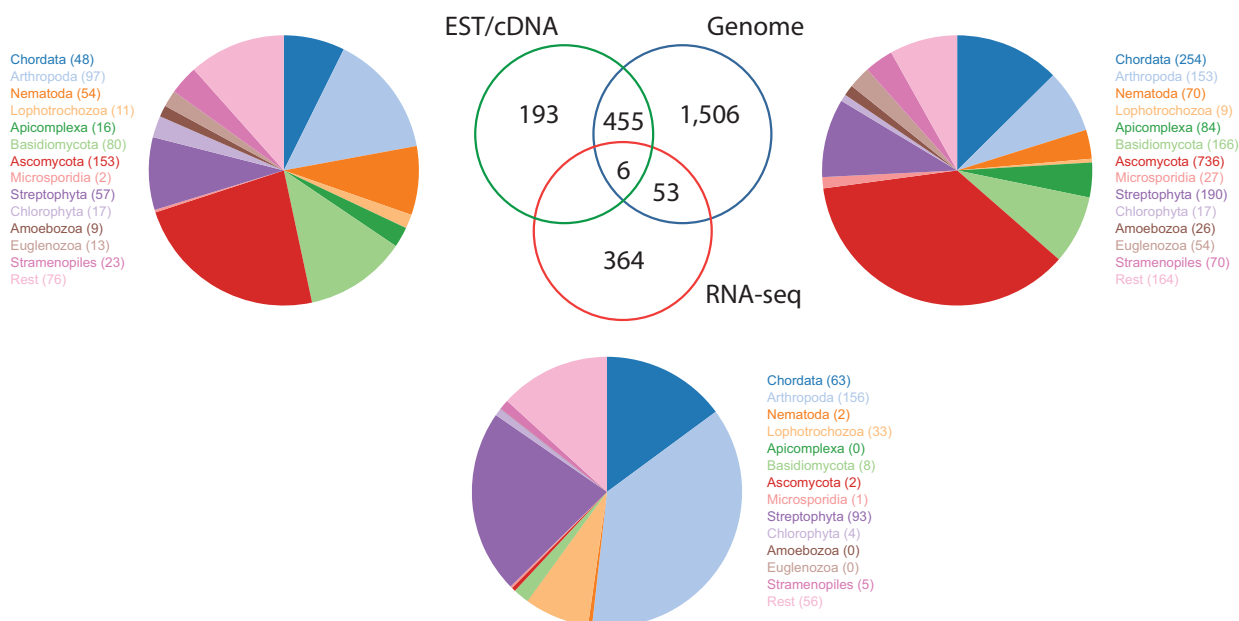


Figure 3. Distribution of species, for which EST/cDNA data, genome assemblies and transcriptome assemblies are available. For each sequencing type, the pie charts show the percentage of sequenced species for selected taxa.

ACKNOWLEDGEMENT

We would like to thank Prof. Christian Griesinger for his continuous generous support.

FUNDING

Funding for open access charge: Max-Planck-Institute for Biophysical Chemistry.

Conflict of interest statement. None declared.

REFERENCES

- Genome 10K Community of Scientists (2009) Genome 10K: a proposal to obtain whole-genome sequence for 10000 vertebrate species. *J. Hered.*, **100**, 659–674.
- i5K Consortium (2013) The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *J. Hered.*, **104**, 595–600.
- Kumar,S., Schiffer,P.H. and Blaxter,M. (2012) 959 nematode genomes: a semantic wiki for coordinating sequencing projects. *Nucleic Acids Res.*, **40**, D1295–D1300.
- Weigel,D. and Mott,R. (2009) The 1001 genomes project for *Arabidopsis thaliana*. *Genome Biol.*, **10**, 107.
- Daetwyler,H.D., Capitan,A., Pausch,H., Stothard,P., van Binsbergen,R., Brøndum,R.F., Liao,X., Djari,A., Rodriguez,S.C., Grohs,C. *et al.* (2014) Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat. Genet.*, **46**, 858–865.
- 3,000 rice genomes project (2014) The 3,000 rice genomes project. *Gigascience*, **3**, 7.
- Goodstein,D.M., Shu,S., Howson,R., Neupane,R., Hayes,R.D., Fazo,J., Mitros,T., Dirks,W., Hellsten,U., Putnam,N. *et al.* (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.*, **40**, D1178–D1186.
- Haas,B.J., Zeng,Q., Pearson,M.D., Cuomo,C.A. and Wortman,J.R. (2011) Approaches to fungal genome annotation. *Mycology*, **2**, 118–141.
- Drysdale,R. and FlyBase Consortium (2008) FlyBase: a database for the *Drosophila* research community. *Methods Mol. Biol.*, **420**, 45–59.
- Yook,K., Harris,T.W., Bieri,T., Cabunoc,A., Chan,J., Chen,W.J., Davis,P., de la Cruz,N., Duong,A., Fang,R. *et al.* (2012) WormBase 2012: more genomes, more data, new website. *Nucleic Acids Res.*, **40**, D735–D741.
- Basu,S., Fey,P., Pandit,Y., Dodson,R., Kibbe,W.A. and Chisholm,R.L. (2013) DictyBase 2013: integrating multiple Dictyostelid species. *Nucleic Acids Res.*, **41**, D676–D683.
- Aurrecochea,C., Barreto,A., Brestelli,J., Brunk,B.P., Cade,S., Doherty,R., Fischer,S., Gajria,B., Gao,X., Gingle,A. *et al.* (2013) EuPathDB: the eukaryotic pathogen database. *Nucleic Acids Res.*, **41**, D684–D691.
- Megy,K., Emrich,S.J., Lawson,D., Campbell,D., Dialynas,E., Hughes,D.S.T., Koscielny,G., Louis,C., Maccallum,R.M., Redmond,S.N. *et al.* (2012) VectorBase: improvements to a bioinformatics resource for invertebrate vector genomics. *Nucleic Acids Res.*, **40**, D729–D734.
- Stajich,J.E., Harris,T., Brunk,B.P., Brestelli,J., Fischer,S., Harb,O.S., Kissinger,J.C., Li,W., Nayak,V., Pinney,D.F. *et al.* (2012) FungiDB: an integrated functional genomics database for fungi. *Nucleic Acids Res.*, **40**, D675–D681.
- Pagani,I., Liolios,K., Jansson,J., Chen,I.-M.A., Smirnova,T., Nosrat,B., Markowitz,V.M. and Kyrpides,N.C. (2012) The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.*, **40**, D571–D579.
- Hammesfahr,B., Odronitz,F., Hellkamp,M. and Kollmar,M. (2011) diArk 2.0 provides detailed analyses of the ever increasing eukaryotic genome sequencing data. *BMC Res. Notes*, **4**, 338.
- Gallant,J.R., Traeger,L.L., Volkening,J.D., Moffett,H., Chen,P.-H., Novina,C.D., Phillips,G.N., Anand,R., Wells,G.B., Pinch,M. *et al.* (2014) Nonhuman genetics. Genomic basis for the convergent evolution of electric organs. *Science*, **344**, 1522–1525.
- Soria-Carrasco,V., Gompert,Z., Comeault,A.A., Farkas,T.E., Parchman,T.L., Johnston,J.S., Buerkle,C.A., Feder,J.L., Bast,J., Schwander,T. *et al.* (2014) Stick insect genomes reveal natural selection's role in parallel speciation. *Science*, **344**, 738–742.
- Sneddon,T.P., Li,P. and Edmunds,S.C. (2012) GigaDB: announcing the GigaScience database. *Gigascience*, **1**, 11.

D1112 *Nucleic Acids Research, 2015, Vol. 43, Database issue*

20. Nanda,S. and Kowalczyk,M.K. (2014) Unpublished genomic data-how to share? *BMC Genomics*, **15**, 5.
21. Hatje,K., Hammesfahr,B. and Kollmar,M. (2013) WebScipio: reconstructing alternative splice variants of eukaryotic proteins. *Nucleic Acids Res.*, **41**, W504–W509.
22. Bradnam,K.R., Fass,J.N., Alexandrov,A., Baranay,P., Bechner,M., Birol,I., Boisvert,S., Chapman,J.A., Chapuis,G., Chikhi,R. *et al.* (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience*, **2**, 10.
23. Hatje,K. and Kollmar,M. (2012) A phylogenetic analysis of the brassicales clade based on an alignment-free sequence comparison method. *Front. Plant. Sci.*, **3**, 192.
24. Keeling,P.J., Burki,F., Wilcox,H.M., Allam,B., Allen,E.E., Amaral-Zettler,L.A., Armbrust,E.V., Archibald,J.M., Bharti,A.K., Bell,C.J. *et al.* (2014) The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol.*, **12**, e1001889.

3.3.2. The landscape of human mutually exclusive splicing

Klas Hatje^{1,2,†}, Raza-Ur Rahman^{2,3}, Ramon O Vidal^{2,‡}, Dominic Simm^{1,4}, Björn Hammesfahr^{1,§}, Vikas Bansal^{2,3}, Ashish Rajput^{2,3}, Michel Edwar Mickael^{2,3}, Ting Sun^{2,3}, Stefan Bonn^{2,3,5,*} and Martin Kollmar^{1,*}

* Corresponding author

¹ Group Systems Biology of Motor Proteins, Department of NMR-based Structural Biology, Max-Planck-Institute for Biophysical Chemistry, Göttingen, Germany

² Group of Computational Systems Biology, German Center for Neurodegenerative Diseases, Göttingen, Germany

³ Center for Molecular Neurobiology, Institute of Medical Systems Biology, University Clinic Hamburg-Eppendorf, Hamburg, Germany

⁴ Theoretical Computer Science and Algorithmic Methods, Institute of Computer Science, Georg-August-University, Göttingen, Germany

⁵ German Center for Neurodegenerative Diseases, Tübingen, Germany

† Present address: Roche Pharmaceutical Research and Early Development, Pharmaceutical Sciences, Roche Innovation Center Basel, Basel, Switzerland

‡ Present address: Max-Delbrück-Center for Molecular Medicine, Berlin, Germany

§ Present address: Research and Development-Data Management (RD-DM), KWS SAAT SE, Einbeck, Germany

Molecular Systems Biology (2017) - Volume 13(12): 959

doi: 10.15252/msb.20177728





<https://www.embopress.org/doi/full/10.15252/msb.20177728>

Contribution

MK initiated the study and designed the analyses together with SB. KH and BH performed MXE predictions. KH integrated the human MXE data into the Kassiopeia database. KH implemented MXE candidate extraction and RNA-Seq analysis with help from ROV and SB. KH and MK did the cluster distribution, splicing mechanism and protein structure mapping analyses. KH, ROV, R-UR, VB, SB and MK performed the differential expression analysis. AR, MEM and TS performed the RT-PCR experiments. ROV, R-UR, VB and SB performed the SNP mapping and prediction analysis. The comparison of different human gene annotations was done by KH and DS. KH did the prediction of MXEs in other mammals, and their comparison together with MK. MK and SB wrote the manuscript. ROV, KH, VB and R-UR contributed to manuscript text and Supplementary Materials. All authors read and approved the final manuscript.



The landscape of human mutually exclusive splicing

Klas Hatje^{1,2,†}, Raza-Ur Rahman^{2,3}, Ramon O Vidal^{2,‡}, Dominic Simm^{1,4}, Björn Hammesfahr^{1,§}, Vikas Bansal^{2,3} , Ashish Rajput^{2,3} , Michel Edwar Mickael^{2,3}, Ting Sun^{2,3}, Stefan Bonn^{2,3,5,*}  & Martin Kollmar^{1,**} 

Abstract

Mutually exclusive splicing of exons is a mechanism of functional gene and protein diversification with pivotal roles in organismal development and diseases such as Timothy syndrome, cardiomyopathy and cancer in humans. In order to obtain a first genome-wide estimate of the extent and biological role of mutually exclusive splicing in humans, we predicted and subsequently validated mutually exclusive exons (MXEs) using 515 publically available RNA-Seq datasets. Here, we provide evidence for the expression of over 855 MXEs, 42% of which represent novel exons, increasing the annotated human mutually exclusive exome more than fivefold. The data provide strong evidence for the existence of large and multi-cluster MXEs in higher vertebrates and offer new insights into MXE evolution. More than 82% of the MXE clusters are conserved in mammals, and five clusters have homologous clusters in *Drosophila*. Finally, MXEs are significantly enriched in pathogenic mutations and their spatio-temporal expression might predict human disease pathology.

Keywords alternative splicing; differential expression; mutually exclusive splicing; splicing mechanisms

Subject Categories Chromatin, Epigenetics, Genomics & Functional Genomics; Genome-Scale & Integrative Biology; Transcription

DOI 10.15252/msb.20177728 | Received 2 May 2017 | Revised 4 November 2017 | Accepted 10 November 2017

Mol Syst Biol. (2017) **13**: 959

Introduction

Alternative splicing of pre-messenger RNAs is a mechanism common to almost all eukaryotes to generate a plethora of protein

variants out of a limited number of genes (Matlin *et al.*, 2005; Nilsen & Graveley, 2010; Lee & Rio, 2015). High-throughput studies suggested that not only 95–100% of all multi-exon genes in human are affected (Pan *et al.*, 2008; Wang *et al.*, 2008; Gerstein *et al.*, 2014) but also that alternative splicing patterns strongly diverged between vertebrate lineages implying a pronounced role in the evolution of phenotypic complexity (Barbosa-Morais *et al.*, 2012; Merkin *et al.*, 2012). Five types of alternative splicing have been identified to contribute to most mRNA isoforms, which are differential exon inclusion (exon skipping), intron retention, alternative 5' and 3' exon splicing, and mutually exclusive splicing (Blencowe, 2006; Pan *et al.*, 2008; Wang *et al.*, 2008; Nilsen & Graveley, 2010). Mutually exclusive splicing generates alternative isoforms by retaining only one exon of a cluster of neighbouring internal exons in the mature transcript and is a sophisticated way to modulate protein function (Letunic *et al.*, 2002; Meijers *et al.*, 2007; Pohl *et al.*, 2013; Tress *et al.*, 2017a). The most extreme cases known so far are the arthropod *DSCAM* genes, for which up to 99 mutually exclusive exons (MXEs) spread into four clusters were identified (Schmucker *et al.*, 2000; Lee *et al.*, 2010; Pillmann *et al.*, 2011).

Opposed to arthropods, current evidence suggests that vertebrate MXEs only occur in pairs (Matlin *et al.*, 2005; Gerstein *et al.*, 2014; Abascal *et al.*, 2015a), and genome-wide estimates in human range from 118 (Suyama, 2013) to at most 167 cases (Wang *et al.*, 2008). Despite these relatively few reported cases, mutually exclusive splicing might be far more frequent in humans than currently anticipated, as has been recently revealed in the model organism *Drosophila melanogaster* (Hatje & Kollmar, 2013). Apart from their low number, MXEs have been described in many crucial and essential human genes such as in the α -subunits of six of the 10 voltage-gated sodium channels (*SCN* genes) (Copley, 2004), in each of the glutamate receptor subunits 1–4 (*GluR1–4*) where the MXEs are called flip and flop (Sommer *et al.*, 1990), and in *SNAP-25* as part of

1 Group Systems Biology of Motor Proteins, Department of NMR-Based Structural Biology, Max-Planck-Institute for Biophysical Chemistry, Göttingen, Germany

2 Group of Computational Systems Biology, German Center for Neurodegenerative Diseases, Göttingen, Germany

3 Center for Molecular Neurobiology, Institute of Medical Systems Biology, University Clinic Hamburg-Eppendorf, Hamburg, Germany

4 Theoretical Computer Science and Algorithmic Methods, Institute of Computer Science, Georg-August-University, Göttingen, Germany

5 German Center for Neurodegenerative Diseases, Tübingen, Germany

*Corresponding author. Tel: +49 40 7410 55082; E-mail: sbonn@uke.de

**Corresponding author. Tel: +49 551 5036960; E-mail: mako@nmr.mpibpc.mpg.de

†Present address: Roche Pharmaceutical Research and Early Development, Pharmaceutical Sciences, Roche Innovation Center Basel, Basel, Switzerland

‡Present address: Max-Delbrück-Center for Molecular Medicine, Berlin, Germany

§Present address: Research and Development—Data Management (RD-DM), KWS SAAT SE, Einbeck, Germany

the neuroexocytosis machinery (Johansson *et al*, 2008). Although MXEs within a cluster often share high similarity at the sequence level, they are usually not functionally redundant, as their inclusion in the mRNAs is tightly regulated. Thus, mutations in MXEs have been shown to cause diseases such as Timothy syndrome (missense mutation in the *CACNA1C* gene) (Splawski *et al*, 2004, 2005), cardiomyopathy (defect of the mitochondrial phosphate carrier *SLC25A3*) (Mayr *et al*, 2011) or cancer (mutations in, e.g., the pyruvate kinase *PKM* and the zinc transporter *SLC39A14*) (David *et al*, 2010).

Despite the implications of mutually exclusive splicing in organismal development and disease, current knowledge on the magnitude of MXE usage and its relevance in biological processes is far from complete. In order to obtain a genome-wide, unbiased estimate of the extent and biological role of mutually exclusive splicing in humans, a set of 6,541 MXE candidates was compiled from annotated and novel predicted exons, and rigorously validated using over 15 billion reads from 515 RNA-Seq datasets.

Results

The human genome contains 855 high-confidence MXEs

Compared to other splicing mechanisms, mutually exclusive splicing in humans seems to be a rare event. MXEs are characterized by genomic vicinity, splice-site compatibility and mutually exclusive presence in protein isoforms. Accordingly, the human genome annotation (GenBank v. 37.3) contains only 158 MXEs in 79 protein-coding genes (Appendix Figs S1–S3). MXEs are often phrased “homologous exons” in the literature because they likely originated from the same ancestral exon. We refrain from using this term throughout our analysis, because several MXEs present in the genome annotation do not show any sequence homology and many neighbouring exons with high sequence similarity are not spliced in a mutually exclusive manner.

In a first attempt to chart an atlas of genome-wide mutually exclusive splicing in humans, we decided to predict potential MXE candidates and validate those using published RNA-Seq data. In a first step, we generated a set of MXE candidates in the human genome

(v. 37.3) from all annotated protein-coding exons and from novel exons predicted in intronic regions including only internal exons in the candidate list (Fig 1A, Appendix Figs S1–S4). From the annotated exons, we selected those that appeared mutually exclusive in transcripts, and neighbouring exons that show sequence similarity and are translated in the same reading frame. To generate novel exon candidates, we predicted exonic regions in neighbouring introns of annotated exons based on sequence similarity and similar lengths (Pillmann *et al*, 2011). We did not consider potential MXEs containing in-frame stop codons such as the neonatal-specific MXE reported for the sodium channel *SCN8A* (Zubović *et al*, 2012), and exons overlapping annotated terminal exons (Appendix Fig S2). The reconstruction resulted in a set of 6,541 MXE candidates in 1,542 protein-coding genes, including 1,058 (68.6%) genes for which we predicted 1,722 completely novel exons in previously intronic regions (Fig 1B). Most introns in human genes are extremely long necessitating careful and strict validation of the MXE candidates to exclude false-positive predictions (Lee & Rio, 2015).

To validate the predicted MXE candidates, we made use of over 15 billion publically available RNA-Seq reads, selecting 515 samples comprising 31 tissues and organs, 12 cell lines and seven developmental stages (Barbosa-Morais *et al*, 2012; Djebali *et al*, 2012; Tilgner *et al*, 2012; Xue *et al*, 2013; Yan *et al*, 2013; Fagerberg *et al*, 2014; Dataset EV1). The data were chosen to encompass common and rare potential splice events in a broad range of tissues, cell types and embryonic stages. Accordingly, the transcription of 6,466 (99%) of the MXE candidates is supported by RNA-Seq reads mapped to the genome (Appendix Fig S3A). To be validated as true mutually exclusive splicing event, each MXE of a cluster needed to exhibit splice junction (SJ) reads from every MXE to up- or downstream gene regions bridging the other MXE(s) of the cluster (Fig 1A). In addition, MXEs should not exhibit any SJ reads to another MXE except when the combined inclusion causes a frame shift and therefore a premature stop codon (Fig 1A, Appendix Figs S3A and D, S5, and S6). These stringent criteria define a high-confidence set of MXEs, requiring three constraints for a cluster of two MXEs and already 18 constraints for a cluster of five MXEs (Appendix Fig S7). In case of clusters with more than two MXE candidates, the validation criteria were applied to the cluster including all MXE candidates as well as to all possible sub-clusters to

Figure 1. The human genome contains 1,399 high-confidence MXEs.

- A Schematic representation of the various annotated and predicted exon types included in the MXE candidate list. For MXE validation, at least three restraints must be fulfilled: the absence of an MXE-joining read (R1), except for those leading to frame shift, and the presence of two MXE-bridging SJ reads (R2 and R3).
- B Prediction and validation of 1,399 1S] (855 3S]) human MXEs. Top: Dataset of 6,541 MXE candidates from annotated and predicted exons. Bottom left: MXE candidates for which splice junction data are currently missing hindering their annotation as MXE or other splice variant. Bottom right: Validation of the MXE candidates using over 15 billion RNA-Seq reads. The outer circles represent the validation based on at least a single read for each of the validation criteria (1S]), while the validation shown in the inner circles required at least three reads (3S]).
- C MXE saturation analysis. Whereas increasing amounts of RNA-Seq reads should lead to the confirmation of further MXE candidates, more RNA-Seq reads might also result in the rejection of previously validated MXEs. The green curves show the number of validated MXEs in relation to the percentage of total RNA-Seq reads used for validation. The orange curves indicate the number of initially “validated MXEs” that were rejected with increasing amounts of reads. Grey dashed lines indicate the point of saturation, which is defined as the point where a twofold increase in reads leads to rejection of less than 1% of the validated MXEs. Of note, whereas the rejection of validated MXEs saturates with 20% of the data, the amount of novel MXE validations is still rapidly increasing.
- D Distribution of validated MXEs in two-exon and multi-exon clusters.
- E Size and distribution of multi-cluster MXEs.
- F The *CUX1* gene (cut-like homeobox 1) contains two interleaved clusters of MXEs (clusters 1 and 2) and two standard clusters each with two MXEs (clusters 3 and 4). The exon 3 and exon 4 variants each are orthologous exons. The exon 4 variants are mutually exclusive (cluster 2). Exon 3a is a differentially included exon and only spliced together with exon 4a. The exons 3b, 3c, 3d and 3e are part of a cluster of four MXEs (cluster 1) and are only spliced together with exon 4b (Appendix Figs S16 and S17). Novel exons are labelled with an asterisk.

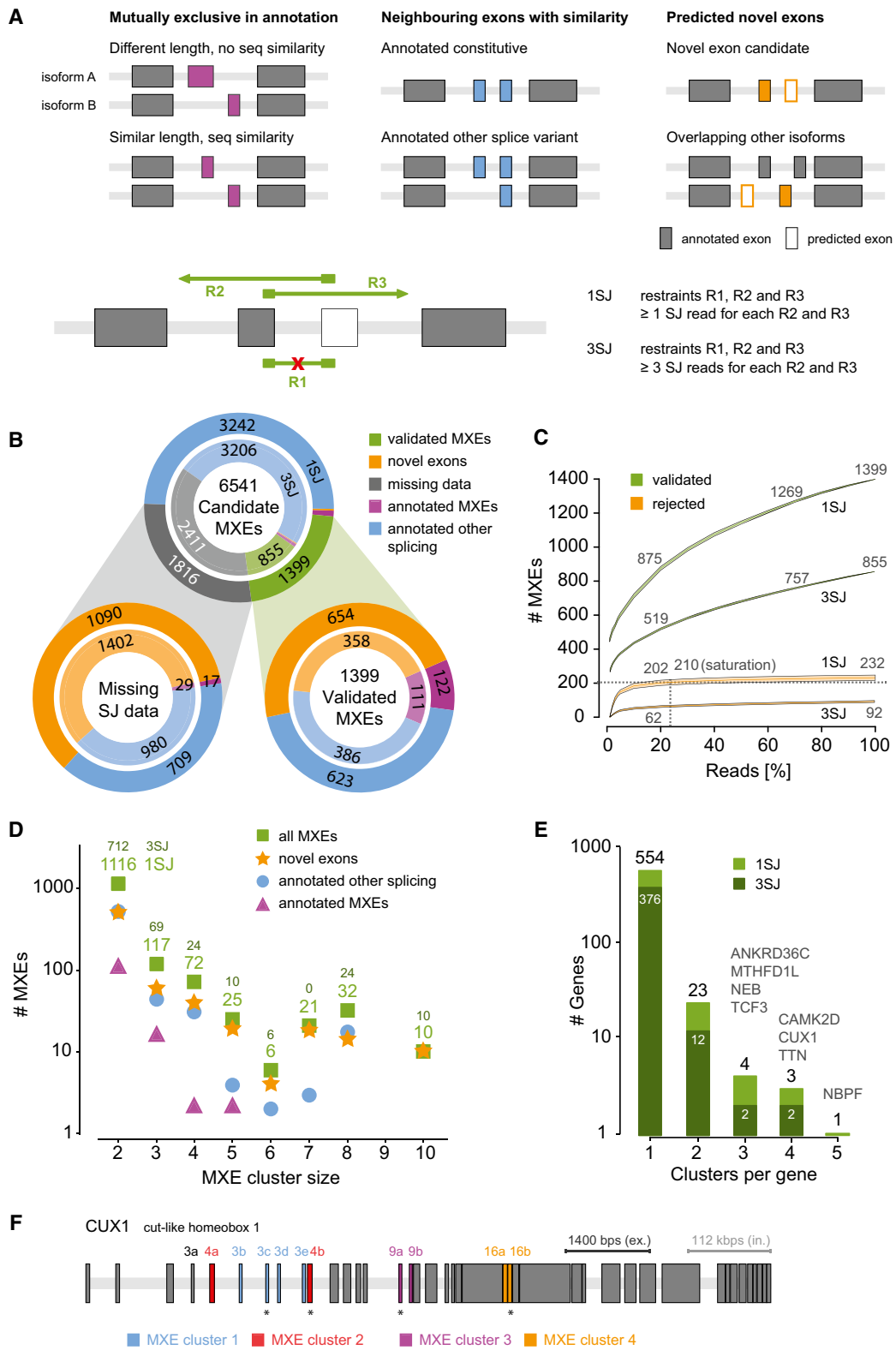


Figure 1.

identify the largest cluster fulfilling all MXE criteria. According to these criteria, 1,399 MXEs were verified with at least one SJ read per exon (1SJ), supported by 2.2 million exon mapping and 34 million SJ reads, increasing the total count of human MXEs by almost an order of magnitude (158–1,399) (Fig 1B, Dataset EV2); 855 MXEs were found to be supported by at least three splice junction reads per exon (3SJ) validated by 1.5 million exon mapping and 27 million SJ reads (Appendix Figs S3B and C, S8–S10). The 1,399 (855, numbers in brackets refer to the 3SJ validation) verified MXEs include 122 (112) annotated MXEs (Fig 1B “annotated MXE”), 623 (388) exons that were previously annotated as constitutive or differentially included (“annotated other splicing”) and 654 (358) exons newly predicted in intronic regions (“novel exon”). Our analysis also showed that 29 of the 158 annotated MXEs are in fact not mutually exclusively spliced but represent constitutively spliced exons or other types of alternative splicing (Appendix Figs S2 and S3E). Finally, 1,741 (2,336) MXE candidates including 1,090 (1,402) newly predicted exons and 17 (29) of the annotated MXEs are supported by 0.5 million exon and 13 million SJ matching reads but still have to be regarded as MXE candidates because not all annotation criteria were fulfilled (Appendix Fig S3A and E).

To estimate the dependence of MXE confirmation and rejection on data quantity, we cross-validated the MXE gain (validation) and loss (rejection) events for several subsets of the total RNA-Seq data (Fig 1C, Appendix Fig S11, Materials and Methods “Saturation analysis”). The course of the curves provides strong evidence for the validity of the MXEs because a single exon-joining read would already be sufficient to reject an MXE cluster while at least two SJ reads are needed to validate one. Whereas even 15 billion RNA-Seq reads do not achieve saturation for the amount of validated MXEs, the gain in rejected MXE candidates is virtually saturated using 25% of the data.

To further validate the list of MXEs, we compared MXE clusters that contained two “annotated other splicing” exons to splicing information from GTEEx portal (<https://www.gtportal.org/home/>). Although GTEEx portal uses an alternative aligner and different alignment settings, all MXEs that we compared showed mutually exclusive behaviour in GTEEx portal (Appendix Fig S12), substantiating our results. Lastly, we selected six brain-expressed novel MXEs for qPCR validation in human brain total RNA. All assayed MXEs showed perfect coherence with the alignment results, confirming mutually exclusive splicing of all assayed novel MXEs in human brain (Appendix Fig S13, Dataset EV3).

Many of the 1,399 (855) MXEs have roles in the cardiac and muscle function and development, while cassette exons are enriched for microtubule- and organelle localization-related terms (Appendix Fig S14).

In summary, the high-confidence set of 1,399 (855) MXEs extends current knowledge of human MXE usage by an order of magnitude, (re)-annotating over a thousand existing and predicted exons and isoforms, while suggesting the existence of further human MXEs.

The human genome contains large cluster and multi-cluster MXEs

In general, mutually exclusive splicing can be quite complex. This is best demonstrated by genes in arthropods that contain both multiple MXE clusters (“multi-cluster”) and large clusters with up to 53 MXEs

such as in the *Drosophila Dscam* genes (Graveley et al, 2004; Pillmann et al, 2011). This is in strong contrast to mutually exclusive splicing in vertebrates as there is to date no evidence of multi-cluster or higher order MXE clusters (Matlin et al, 2005; Pan et al, 2008; Wang et al, 2008; Gerstein et al, 2014; Abascal et al, 2015a,b).

The analysis of the 1,399 validated human MXEs provides first evidence for clusters of multiple MXEs in the human genome (Fig 1D, Appendix Fig S15). While most MXEs are present in clusters of two exons (1,116 MXEs), a surprisingly high number of clusters have three to 10 MXEs (283 MXEs in 71 clusters).

Interestingly, although a large part of the verified MXEs contain a single MXE cluster (554 genes, Fig 1E), we could also provide evidence for human genes containing multiple MXE clusters. Thus, *TCF3*, *NEB*, *ANKRD36C* and *MTHFD1L* contain three clusters and *TTN*, *CAMK2D* and *CUX1* four clusters of MXEs. A very interesting case of complex interleaved mutually exclusive splicing can be seen for *CUX1*, the transcription factor cut-like homeobox 1. It contains a cluster of MXEs (exons 3b–3e) that is differentially included into a set of two exons (exon 3 and exon 4), and the two sets are themselves mutually exclusive (Fig 1F, Appendix Figs S16 and S17). The identification of large clusters with multiple MXEs and many genes with multiple clusters shows that complex mutually exclusive splicing is not restricted to arthropods (Schmucker et al, 2000; Graveley, 2005; Lee et al, 2010; Hatje & Kollmar, 2013) but might be present in all bilateria.

Mutually exclusive presence of coding exons in functionally active transcripts

To understand which splicing mechanisms might be primarily responsible for the regulation of mutually exclusive splicing in humans, we investigated several mechanisms that were shown to act in some specific cases and were proposed to coordinate mutually exclusive splicing in general (Fig 2A; Letunic et al, 2002; Smith, 2005). We identified five cases (0.79% of all clusters) of U2 and U12 splice acceptor incompatibility (Appendix Fig S18) and 57 (9%) cases of potential steric interference, a too short distance between splice donor sites and branch points (< 50 bp; Fig 2B and Appendix Fig S19). Although 377 (60%) of the MXE clusters contain exons with exon lengths not divisible by three which would result in non-functional transcripts in case of combined inclusion, MXE-joining reads were found for only 83 (22%) of these clusters (Fig 2B; Appendix Figs S3B and D, and S20). Surprisingly, the majority of the annotated MXEs are of this type (91 of 122; 75%) as well as many exons previously annotated as other splice types (44 of 662), but only few of the novel MXEs predicted in intronic regions (25 of 615; Appendix Fig S3A and D). These numbers suggest that splicing of the remaining 484 MXE clusters is tightly regulated by other mechanisms (Fig 2B) such as RNA–protein interactions, interactions between small nuclear ribonucleoproteins and splicing factors (Lee & Rio, 2015), and competitive RNA secondary structural elements (Graveley, 2005; Yang et al, 2012; Lee & Rio, 2015). Competing RNA secondary structures are, however, usually not conserved across long evolutionary distances. A potential case of a docker site and selector sequences downstream of each exon variant was identified for the cluster of four MXEs in the *CD55* gene (Appendix Fig S21).

In contrast to cassette exons and micro-exons, which tend to be located in surface loops and intrinsically disordered regions instead

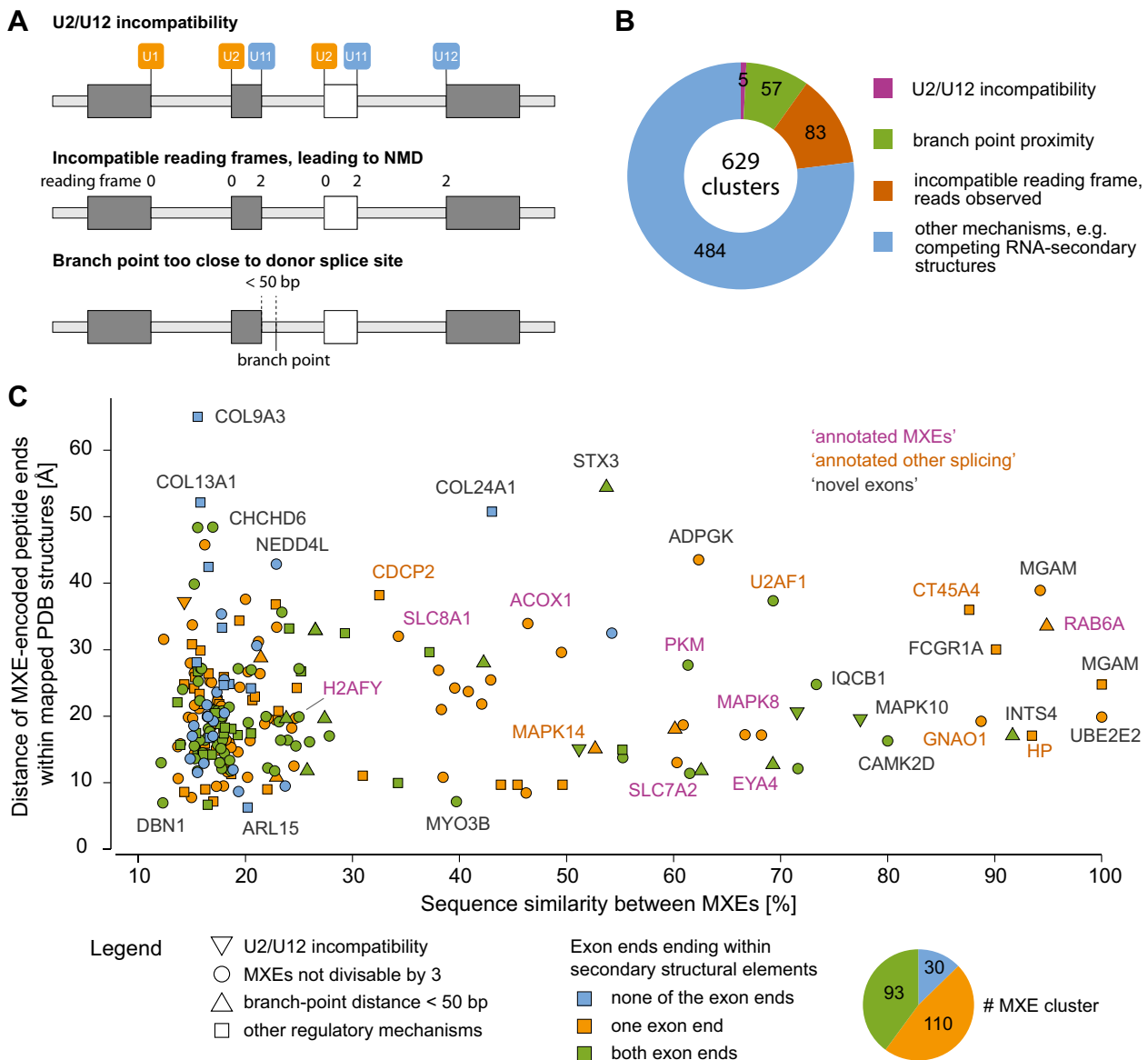


Figure 2. MXE presence is regulated at the RNA and protein folding level.

A Schematic representation of MXE splicing regulation via splice-site incompatibility, branch point proximity and translational frame shift leading to NMD.
 B Observed usage of MXE splicing regulation in 629 MXE clusters.
 C By mutually exclusive inclusion into transcripts, MXEs of a cluster are supposed to encode the same region of a protein structure. If the respective regions of the protein structures are embedded within secondary structural elements (the ends of the exon-encoded peptides are part of α -helices and/or β -strands), it is highly unlikely that the translation of a transcript will result in a folded protein in case the respective exon is missing (skipped exon). If the MXEs have highly similar sequences and do not encode repeat regions, it seems unlikely that either could be present in tandem or absent at all in a folded protein. Here, we have combined protein structure features (colours) with splicing regulation information (symbols). Accordingly, 87% of the MXE-encoded protein regions are embedded in secondary structural elements (orange and green symbols), and most of the remaining MXEs can only be spliced mutually exclusive because splicing as differentially included exons would lead to frame shifts (blue circles). As examples, we labelled many MXE clusters distinguishing annotated MXEs (purple letters), known exons that we validated as MXEs (orange letters), and clusters containing novel exons (dark-grey letters).

of folded domains (Buljan *et al.*, 2012; Ellis *et al.*, 2012; Irimia *et al.*, 2014), all MXEs, whose protein structures have been analysed, are embedded within folded structural domains as has been shown for, for example, *DSCAM* (Meijers *et al.*, 2007), *H2AFY* (Kustatscher

et al., 2005), the myosin motor domain (Kollmar & Hatje, 2014) and *SLC25A3* (Tress *et al.*, 2017a). As we have shown in the beginning, there is also a subset of 73 MXEs not showing any sequence homology ("annotated no similarity"). It is unlikely that the encoded

peptides account for identical secondary structural elements. Rather, if the MXEs of this subset are true MXEs, there is a small subset (about 5%) of MXEs whose mutual inclusion leads to considerably altered protein folds or affects surface loops and disordered regions similar to cassette exons.

Because MXEs are supposed to modulate protein functions through variations and not alterations in specific restricted parts of the structure, we thought it could be possible to distinguish MXEs from cassette exons at a protein structural level. Such an analysis could provide complementary evidence for the validation as MXE in contrast to two (or more) neighbouring cassette exons. While one and only one of the exons of a cluster of MXEs has to be included in the transcript, the defining feature of a cassette exon is that it can either be present or absent. If MXEs were mis-classified and in fact neighbouring cassette exons, it would therefore be possible that all exons of the cluster were present or absent from the transcript, and accordingly the protein structure. These differences between MXEs and cassette exons impose three restrictions on their localization within protein folds (Appendix Fig S22). Thus, (i) if one or both ends of the MXE-encoded peptide end within a secondary structural element, it seems impossible that the respective peptide could be absent from the protein because this would break up multiple spatial interactions. This suggests that respective protein regions cannot be encoded by cassette exons. (ii) High sequence similarity between MXEs suggests important conserved structural interactions even if the peptide ends are not part of secondary structural elements. For example, it seems highly unlikely that a cluster of two exons encoding transmembrane helices could be spliced as cassette exons because absence or presence of both exons would switch the membrane site of all subsequent sequence. (iii) In case of cassette exons and absence of the exons, it must be possible that the remaining sequence still folds correctly. This can be assessed if a protein structure is available with the respective exon-encoded region present. Supposing the respective region was absent, the remaining ends would need to be joined to result in a correctly folded domain, which seems extremely unlikely if the peptide ends are far apart. Such regions are also more likely encoded by MXEs. To assess this model, we mapped the validated MXEs against the PDB database (Fig 2C, Appendix Fig S22, Dataset EV4; Rose *et al.*, 2015). Of the 1,399 MXEs, 273 MXEs (20%) from 233 MXE clusters (37%) matched to human or mammalian protein structures (Appendix Fig S22). For 87% of these MXEs, at least one of the exon termini is embedded within a secondary structural element, suggesting that these exons are in fact true MXEs and not mis-classified cassette exons (Fig 2C, yellow and green coloured symbols). This high level of structural conservation also strongly supports the hypothesis that MXEs modulate but do not considerably alter protein functions (Letunic *et al.*, 2002; Yura *et al.*, 2006; Abascal *et al.*, 2015a; Tress *et al.*, 2017a). Of the remaining 13% (Fig 2C, blue coloured symbols), many MXEs would lead to frame shifts if they were spliced as cassette exons (both exons present or absent in the transcript, blue circles), and in multiple cases (e.g. *COL9A3*, *COL24A1* and *COL13A1*), the peptide ends are far apart indicating strong folding problems in case the respective exons were absent in the transcripts. In total, there are only a handful cases such as the MXE cluster in *ARL15* (Fig 2C) whose mutually exclusive presence in proteins cannot be explained by the analysed splicing restrictions, by NMD targeting, or by folding constraints.

MXEs mainly consist of one ubiquitous exon and otherwise regulated exons

To modulate gene functionality, mutually exclusive splicing would need spatial and temporal splicing regulation and expression. To understand the expression patterns of MXEs, we conducted a differential inclusion analysis using the Human Protein Atlas (Fagerberg *et al.*, 2014), Embryonic Development (Yan *et al.*, 2013) and ENCODE datasets (Djebali *et al.*, 2012). Of the 1,399 MXEs, 608 MXEs (345 unique genes), 573 MXEs (389 unique genes) and 552 MXEs (330 unique genes) are differentially expressed, respectively (adjusted P -value < 0.05; Fig 3A, Appendix Figs S23–S26, Dataset EV5 and EV6). Most notably, the differentially included MXEs comprise 43.5, 40.9 and 39.5% of all MXEs indicating that MXEs are to a very large extent tissue- and developmental stage-specifically expressed.

The comparison of the genes containing differentially expressed MXEs from these three projects shows that 519 (88.7%) of all 585 MXE cluster containing genes have at least a single MXE differentially expressed in one of the covered tissues, cell types or developmental stages (Fig 3B). The 519 genes contain 942 differentially expressed MXEs (67% of the total 1,399 MXEs; Fig 3C). This number is in agreement with earlier analyses on small sets of MXEs (66 and 57%) (Wang *et al.*, 2008; Abascal *et al.*, 2015a). Expectedly, the expression of novel MXEs seems to be considerably more tissue specific than the expression of annotated MXEs and cassette exons (Appendix Fig S23). Lastly, 208 MXEs from 113 genes are preferentially expressed during embryonic development indicating that many MXEs are specific to certain developmental stages (Fig 3B and C).

The analysis of MXE specificity reveals that in many clusters one MXE dominates expression, whereas other MXEs are expressed at selected developmental time points and in specific tissues (Fig 3, Appendix Figs S23–S26). This modulation suggests crucial spatio-temporal functional roles for MXEs and can in many cases not be observed at the gene level, as gene counts can remain largely invariant. A well-known case for similar expression of MXEs in newborn heart but expression of only one MXE variant in adult heart is the ion channel *CACNA1C* (Diebold *et al.*, 1992), an example for the switch of expression are the MXEs of the *SLC25A3* gene (Wang *et al.*, 2008). We surmise that the observed specificity in combination with a generally lower expression could also explain the discovery of 654 (358) novel exons that have so far eluded annotation efforts (Fig 1A, Appendix Fig S23). In conclusion, the tight developmental and tissue-specific regulation of MXE expression suggests that changes in MXE function or expression might cause aberrant development and human disease (Xiong *et al.*, 2015). Pathogenic mutations in MXEs are known to cause Timothy syndrome, cardiomyopathy, cancer and kidney disease (Kaplan *et al.*, 2000; Splawski *et al.*, 2004, 2005; David *et al.*, 2010; Mayr *et al.*, 2011).

MXEs are high-susceptibility loci for pathogenic mutations

To obtain a comprehensive overview of MXE-mediated diseases, we annotated all MXEs with pathogenic SNPs from ClinVar (Landrum *et al.*, 2016), resulting in 35 MXEs (eight newly predicted exons) with 82 pathogenic SNPs (Fig 4A, Dataset EV7). Disease-associated MXEs show tight developmental and tissue-specific expression with

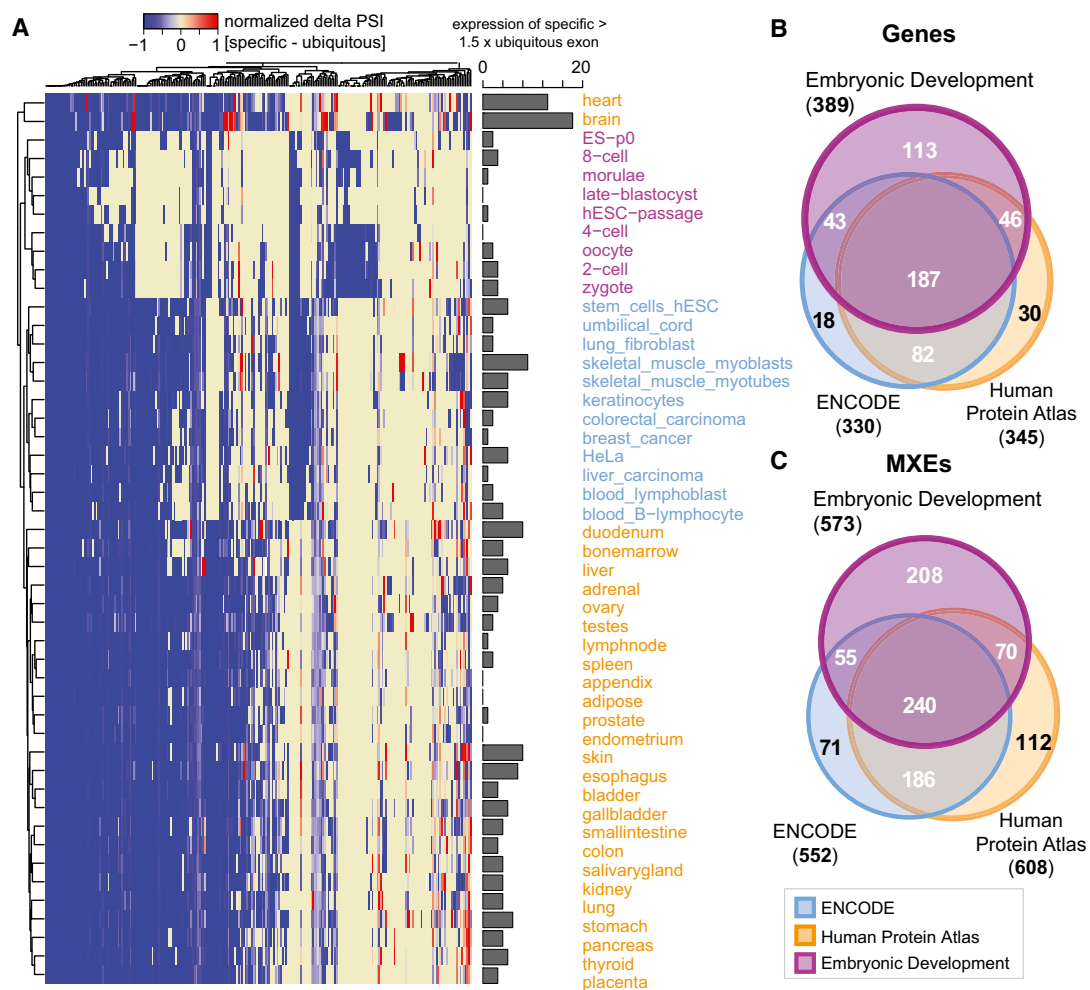


Figure 3. MXE expression is tightly regulated across tissues and development.

A Heatmap showing all differentially expressed MXE clusters with at least three RPKM. Here, we used the Gini coefficient, which is a measure of the inequality among values of a frequency distribution (Ceriani & Verme, 2012) and has successfully been used to determine tissue-enriched gene sets (Zhang *et al*, 2017), to determine highly tissue-specific MXEs (maximum normalized Gini index of cluster) and MXEs with a broad tissue expression distribution (minimum Gini index). For each MXE cluster, the per cent-spliced-in (PSI) value of the ubiquitous MXE (minimum Gini index) is subtracted from the PSI value of the specific MXE (maximum Gini index of cluster) (delta PSI value) and scaled between -1 (broad tissue distribution) and 1 (highly tissue specific). Each column represents an MXE pair, and each row represents MXE expression in a tissue, cell type or at a developmental time point. The bar graph summarizes counts where the specific MXE is 1.5-fold more spliced in than the ubiquitous MXE.

B Overview of differentially expressed genes for the Embryonic Development, ENCODE and Human Protein Atlas datasets.

C Overview of differentially expressed MXEs for the Embryonic Development, ENCODE and Human Protein Atlas datasets.

prominent selective expression in heart and brain, and cancer cell lines (Fig 4B and C, Dataset EV7). Interestingly, the percentage of pathogenic SNP-carrying MXEs is twofold higher than the percentage of all pathogenic SNP-carrying exons (Fisher's exact test, P -value = 3×10^{-11}). A similar enrichment can be found for cassette exons (Fisher's exact test, P -value = 2.2×10^{-16}) suggesting that in general alternative splicing-associated exons are susceptibility loci for pathogenic mutations. The genes with MXEs carrying pathogenic SNPs are predominantly associated with neurological disease (10), neuromuscular disorders (7), cardiomyopathies (6) and cancer (3) and are enriched in voltage-gated cation channels

(e.g. *CACNA1C* and *CACNA1D*), muscle contractile fibre genes (e.g. *TPM1*), and transmembrane receptors (e.g. *FGFR1-3*; Fig 4, Appendix Fig S27, Dataset EV7).

Disease-associated MXEs have high amino-acid identity (average 49.1%, SD 23.1%), reaching up to 89% in *ACTN4* (Appendix Fig S28), suggesting similar functional roles and in consequence similar pathogenic potential for many MXE pairs (Fig 4C, Appendix Fig S29). Four of all SNP-containing MXE clusters contain mutations in both MXEs (*FHL1*, *MAPT*, *CACNA1C* and *CACNA1D*), whereas 31 currently have pathogenic SNPs in only one MXE. The MXE expression analysis shows that many SNP-carrying MXEs are highly

3. Publications and Manuscripts

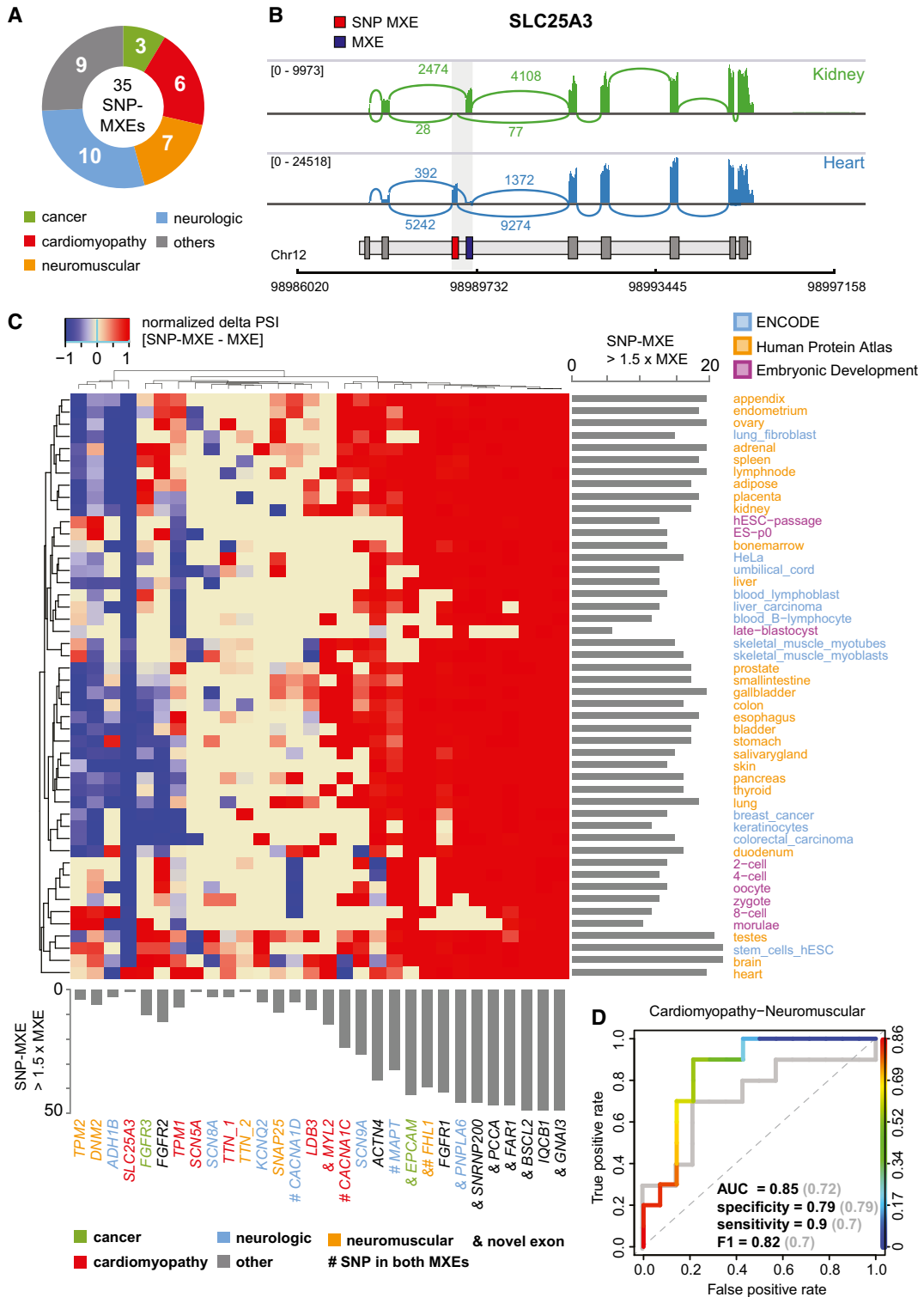


Figure 4.

Figure 4. MXE-ratio expression predicts disease pathology.

- A Thirty-five MXE clusters contain 82 pathogenic mutations causing neurologic (10), neuromuscular (7), cardiac (6), cancer (3) or other diseases (9).
- B Sashimi plots showing exon as well as splice junction reads (including number of reads) in kidney and heart for *SLC25A3*.
- C Heatmap showing the delta PSI values (PSI value of the non-SNP-containing MXE subtracted from the PSI value of the SNP-containing MXE) of MXE clusters containing pathogenic SNPs scaled between -1 and 1 (blue = high expression non-SNP-containing MXE, red = high expression SNP-containing MXE). Columns represent MXE clusters and rows tissues, cell types and developmental stages. The column bar graph summarizes counts where the SNP-containing MXE is 1.5-fold more expressed than the non-SNP-containing MXE, whereas the row bar graph shows this for each tissue, cell type and developmental stage.
- D Receiver-operating characteristic (ROC) curve showing true- and false-positive rates for cardiomyopathy-neuromuscular disease prediction based on spatio-temporal MXE (coloured lines and black text) and RPKM-based gene (grey lines and text) expression (delta PSI values).

expressed, especially in disease-associated tissues where the respective non-SNP-carrying MXEs are not or barely expressed (Fig 4B and C, Appendix Fig S29). Examples include *ACTN4*, *TPM1* and *SLC25A3* (Appendix Figs S28, S30, and S31). Moreover, MXEs with pathogenic SNPs are usually not or non-exclusively expressed at early developmental stages (Appendix Fig S28–S31), while high and exclusive expression could lead to early embryonic death or severe multi-organ phenotypes (e.g. *FAR1*, Appendix Fig S32). Conversely, several non-SNP-carrying MXEs are highly expressed in early development and are otherwise mainly expressed at equal and lower levels compared to the SNP-carrying MXEs (Appendix Figs S29E–S31). The absence of pathogenic SNPs in these MXEs suggests functional compensation of the pathogenic SNP-carrying MXEs or early lethality, both of which would result in no observable phenotype.

Of the 35 MXE clusters with pathogenic mutations eight contain novel exons (Fig 4C, Dataset EV7). A mutation in exon 9a (p.Asp365Gly) of *FAR1*, a gene of the plasmalogen-biosynthesis pathway, causes rhizomelic chondrodysplasia punctata (RCDP), a disease that is characterized by severe intellectual disability with cataracts, epilepsy and growth retardation (Buchert *et al*, 2014). Novel MXE 9b is expressed in the same tissues but at eightfold lower levels suggesting partial functional compensation of the MXE 9a mutation, which might be responsible for the “milder” form of RCDP as compared to pathogenic mutations in other genes of the pathway (*PEX7*, *GNPAT* and *AGPS*) (Appendix Fig S32). A tissue-specific compensation mechanism had already been proposed but a reasonable explanation could not be given because *FAR2* expression shows a different tissue profile and individuals with deficits in peroxisomal β -oxidation, a potential alternative supply for fatty alcohols, have normal plasmalogen levels (Buchert *et al*, 2014). Because of the young age of the affected children, it is not known yet whether a mutation in constitutive exon 4 (p.Glu165_Pro169delinsAsp), which could not be compensated in a similar way as the exon 9a mutation, leads to a strong RCDP-like phenotype (no survival of the first decade of life) or to a milder form such as the one caused by the exon 9a mutation.

In conclusion, it is tempting to speculate that MXE pathogenicity might be governed by high or exclusive expression in affected target tissues that is usually absent from early developmental processes, a pattern of expression that seems at least partially inverted for MXEs without pathogenic SNP annotations. To assess whether MXE pathogenicity follows observable rules, we trained a machine learner on MXE expression data and predicted the affected target tissue (Fig 4D, Dataset EV8). To obtain at least 10 observations per category with an expression > 3 RPKM, diseases were grouped into cardio-neuromuscular ($n = 10$) and other diseases ($n = 14$) and predicted using leave-one-out cross-validation with a Random Forest. Cardiac-neuromuscular diseases could be predicted with an accuracy of 83% (P -value < 0.01), a specificity of 79%, a sensitivity of 90% and an area under the ROC curve (AUC) of 85% (Fig 4D, Dataset EV8, Appendix Fig S29). Conversely, cardio-neuromuscular disease could be predicted with an AUC of 72% using RPKM-based gene expression values (Fig 4D). Although based on only 24 observations, our data suggest that MXE expression might predict disease pathogenicity in space and potentially also in time.

Evolutionary dynamics of MXEs in mammals and bilaterians

While tissue-specific gene expression is conserved between birds and mammals, the alternative splicing of cassette exons is conserved only in brain, heart and muscles and is mainly lineage-specific (Barbosa-Morais *et al*, 2012; Merkin *et al*, 2012). Accordingly, a core set of only ~500 exons was found with conserved alternative splicing in mammals and high sequence conservation, which was a small subset of the thousands of cassette exons identified in total. In contrast, although the total number was considerably smaller, most of the known human MXEs have been shown to be highly conserved throughout mammals if not even vertebrates (Letunic *et al*, 2002; Copley, 2004; Abascal *et al*, 2015b). In order to assess the conservation of human MXEs across mammals, we identified orthologous proteins in 18 representative species from all major sub-branches spanning 180 million years of evolution and predicted MXEs therein (Fig 5, Appendix Fig S33, Dataset EV9). Based on a

Figure 5. Evolutionary dynamics of MXEs in mammalian evolution.

Clusters of validated MXE were sorted by chromosome and chromosomal position. The names of the corresponding genes and the cluster-IDs are given in the outermost circle, and the presence of the respective MXEs (MXE clusters) in other annotations and mammals is indicated by coloured bars. Because the generation of the set of MXE candidates was based on the GenBank annotation, we analysed the presence of the validated MXEs in complementary annotations. Thus, the outer circles show whether the validated MXEs are also annotated as MXEs in Ensembl and Aceview, and whether the validated MXEs are present at all as exons in the Ensembl annotation as indicated by the legend. The lengths of the bars denote the percentage of matching exons for each cluster. For comparison, we show the annotation as MXE in two different Ensembl versions highlighting the dynamics of exon annotations over time. The comparison of the GenBank with the latest Ensembl annotation (v. 37.75) showed considerably less exons annotated as MXEs (58) in Ensembl although these include six of the “novel exons” (Appendix Fig S1). The presence of the respective validated MXEs in each of the analysed 18 mammals is shown by coloured bars. The 18 mammals, their phylogenetic relation and the total numbers of MXEs shared with human are presented at the bottom. The innermost circle represents the number of exons within each cluster of MXEs.

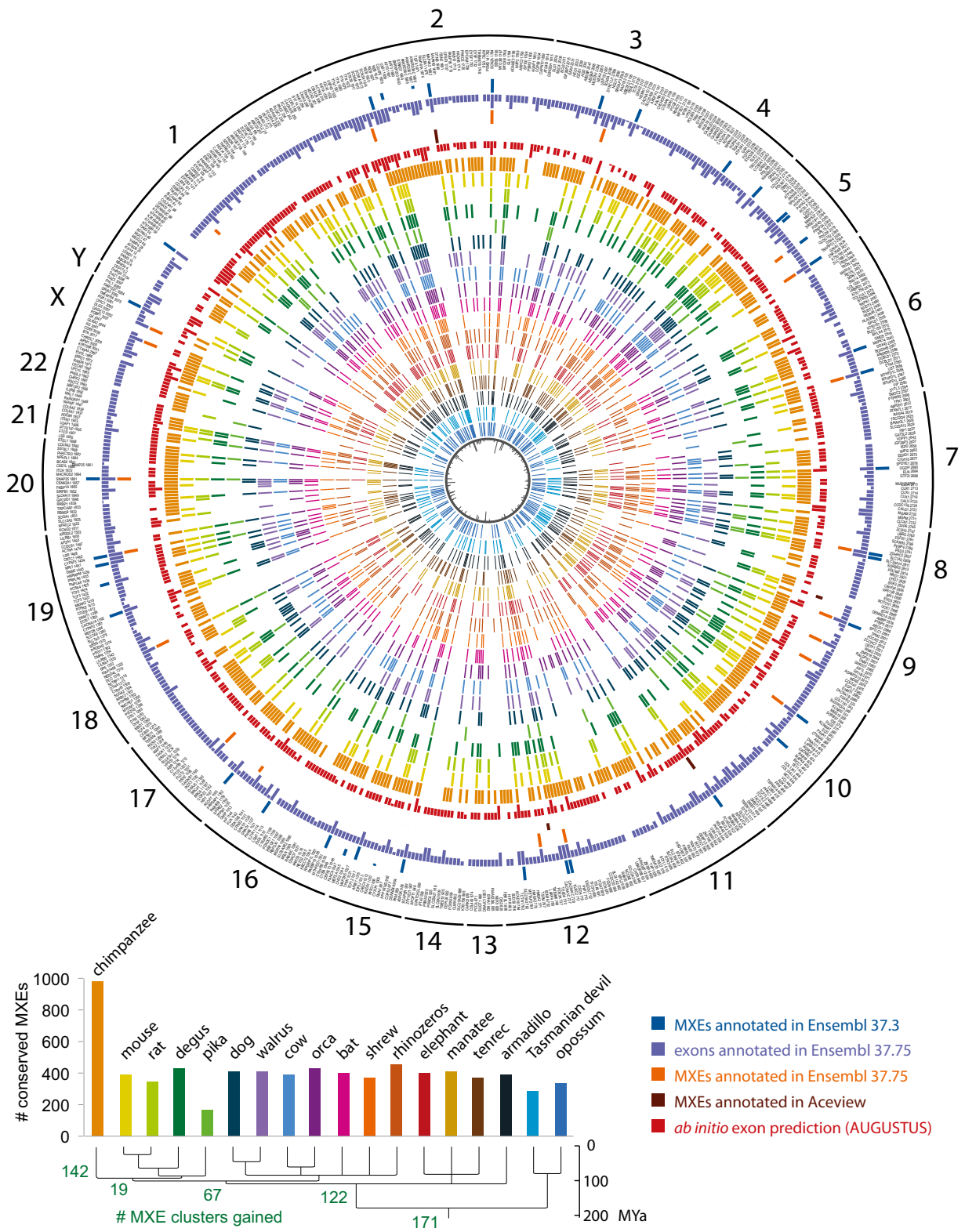


Figure 5.

simple model expecting each shared cluster to be already present in the last common ancestor of the respective species, we identified a core set of at least 173 (28%) of the human MXE clusters conserved throughout mammals (Fig 5, Appendix Fig S33). Other 122 MXE clusters were most likely present in the last common ancestor of the eutherians (16 species, placental mammals). The core set of mammalian MXE clusters includes 83 clusters shared between at least 16 of the species and 61 clusters shared between 17 species suggesting that their spurious absence in single mammals is likely due to genome assembly gaps or problems in identifying the correct orthologous genes. The remaining 29 MXE clusters of the core set have a scattered distribution across the 18 mammals indicating multiple independent branch- and species-specific cluster loss events. Such taxon-specific loss events include the MXE clusters in the *SRPK1* and *PQB1* genes, which are absent in Glires (including mouse and pika), the cluster of 10 MXEs in *ABI3BP* that has been lost in the ancestor of mouse and rat, and the MXEs in *OSTF1* and *PTPRS*, which are absent in Afrotheria. The MXE clusters in *IKZF3*, *MBD1* and *ATP10B*, for example, are present in all Eutheria but not in Metatheria (marsupials). The MXE cluster gain rate within eutherian evolution towards human is relatively constant over time with about 23 clusters per 10 million years. Interestingly, each of the 16 eutherian species also lost a similar number of MXE clusters (127 clusters on average, Appendix Fig S33). In total, 82% of the human clusters containing validated MXEs are found in at least one further mammal (Fig 5). In summary, the large core set of mammalian MXEs and the overall conservation of MXE clusters suggest that MXEs are considerably more conserved than cassette exons. This observation supports expectations from considering the encoded protein structures where MXEs are supposed to provide alternative sequences for conserved secondary structural elements, while cassette exons are on average considerably shorter and add flexibility to surface loops (Buljan et al, 2012; Ellis et al, 2012; Irimia et al, 2014).

To get a first glimpse on mutually exclusive splicing evolution across bilaterians, we identified a set of 44 orthologous genes from genes containing MXEs in *Drosophila* (Hatje & Kollmar, 2013) and human genes containing MXE candidates (Appendix Fig S34, Dataset EV10). Of these orthologous genes, 28 contain validated MXEs in human, nine were validated to be spliced differently in human, and seven could not be validated in human because read mapping data are still missing; 20 (71%) of the genes containing validated MXEs represent cases of incompatible reading frames leading to NMD in case of joined inclusion, and for 18 of these MXE clusters multiple MXE-joining reads were found (Appendix Figs S34 and S35). We further analysed the 28 orthologous genes with validated MXEs and found five genes with homologous MXE clusters (identical position in gene, identical exon phase), 13 genes with MXE clusters in human that have homologous exons in *Drosophila* and eight genes with MXEs in human where the corresponding sequence regions in the orthologous *Drosophila* genes are part of larger exons (Appendix Figs S35 and S36). The presence of orthologous MXE clusters has been attributed to convergent evolution (Copley, 2004), although the respective analysis was in part based on the comparison of non-orthologous genes (e.g. comparing human sodium channel genes [e.g. *SCN1A*] with the *Drosophila* calcium channel *cac* gene and not the orthologous sodium channel *para* gene). At least for muscle myosin heavy chain genes it could

be demonstrated that *Drosophila* already lost several MXE clusters compared to, for example, *Daphnia pulex* (crustacean) and lophotrochozoans (Kollmar & Hatje, 2014) and that the evolutionary history of the MXEs within each cluster is remarkably complex with multiple independent exon duplications and losses (Odrionitz & Kollmar, 2008). Thus, detailed studies including more bilaterian and non-bilaterian taxa would be necessary to finally conclude convergent or divergent evolution for each of the human and *Drosophila* MXE clusters. Although the overlap of MXEs in orthologous genes of human and *Drosophila* is very low, the MXE gain and loss rates are very similar (Hatje & Kollmar, 2013) indicating a conserved role of tandem exon duplication in bilaterians. Gene structures can be highly conserved between kingdoms (Rogozin et al, 2003), and certain exons therefore seem to be predisposed to undergo duplication. In summary, these findings provide strong evidence for many MXE gain and loss events during mammalian evolution, suggesting a pronounced role of these processes in speciation and establishing phenotypic differences.

Discussion

Using stringent criteria, including sequence similarity, reading frame conservation and similar lengths, and billions of RNA-Seq reads, we generated a strongly validated atlas of 1,399 human MXEs providing insights into mutually exclusive splicing mechanics, specific expression patterns, susceptibility for pathogenic mutations and deep evolutionary conservation across 18 mammals. The presented increase in human MXEs by an order of magnitude lifts MXEs into the present-day dimension of other human alternative splice types (Pan et al, 2008; Wang et al, 2008; Gerstein et al, 2014). Saturation analysis and the existence of 1,816 expressed but unconfirmed MXE candidates suggest a potential 27% increase in the MXE-ome with a twofold increase in data. Although alternative splice variants are abundant at the transcriptome level, recent mass spectrometry analyses suggested only small numbers of alternative transcripts to be translated (Abascal et al, 2015a; Ezkurdia et al, 2015; Blencowe, 2017; Tress et al, 2017a,b). Interestingly, MXEs were particularly enriched in the translated alternative transcripts, compared to other splice variants. However, ribosome profiling data showed high frequencies of ribosome engagement of cassette exons indicating that these isoforms are likely translated (Weatheritt et al, 2016). Similar results have been obtained through polyribosome profiling (Sterne-Weiler et al, 2013; Floor & Doudna, 2016). These observations suggest that most of the MXEs evaluated at the transcript level will also be found in the proteome.

About half (47%) of the 1,399 MXEs represent novel exons, which are often expressed at low levels and whose expression is restricted to few tissues and cell types, possibly explaining their absence from current genome annotations. Extrapolating these observations to all splice types and genes suggests the existence of thousands yet unannotated exons in introns. This estimation is in accordance with a recent analysis of more than 20,000 human RNA-Seq datasets that revealed over 55,000 junctions not present in annotations (Nellore et al, 2016). In this analysis, junctions found in at least 20 reads across all samples were termed “confidently called”. Although the total number of reads required for MXE validation in our analysis is lower (≥ 2 SJ reads in the 1SJ case, ≥ 6 SJ

simple model expecting each shared cluster to be already present in the last common ancestor of the respective species, we identified a core set of at least 173 (28%) of the human MXE clusters conserved throughout mammals (Fig 5, Appendix Fig S33). Other 122 MXE clusters were most likely present in the last common ancestor of the eutherians (16 species, placental mammals). The core set of mammalian MXE clusters includes 83 clusters shared between at least 16 of the species and 61 clusters shared between 17 species suggesting that their spurious absence in single mammals is likely due to genome assembly gaps or problems in identifying the correct orthologous genes. The remaining 29 MXE clusters of the core set have a scattered distribution across the 18 mammals indicating multiple independent branch- and species-specific cluster loss events. Such taxon-specific loss events include the MXE clusters in the *SRPK1* and *PQBP1* genes, which are absent in Glires (including mouse and pika), the cluster of 10 MXEs in *ABI3BP* that has been lost in the ancestor of mouse and rat, and the MXEs in *OSTF1* and *PTPRS*, which are absent in Afrotheria. The MXE clusters in *IKZF3*, *MBD1* and *ATP10B*, for example, are present in all Eutheria but not in Metatheria (marsupials). The MXE cluster gain rate within eutherian evolution towards human is relatively constant over time with about 23 clusters per 10 million years. Interestingly, each of the 16 eutherian species also lost a similar number of MXE clusters (127 clusters on average, Appendix Fig S33). In total, 82% of the human clusters containing validated MXEs are found in at least one further mammal (Fig 5). In summary, the large core set of mammalian MXEs and the overall conservation of MXE clusters suggest that MXEs are considerably more conserved than cassette exons. This observation supports expectations from considering the encoded protein structures where MXEs are supposed to provide alternative sequences for conserved secondary structural elements, while cassette exons are on average considerably shorter and add flexibility to surface loops (Buljan et al, 2012; Ellis et al, 2012; Irimia et al, 2014).

To get a first glimpse on mutually exclusive splicing evolution across bilaterians, we identified a set of 44 orthologous genes from genes containing MXEs in *Drosophila* (Hatje & Kollmar, 2013) and human genes containing MXE candidates (Appendix Fig S34, Dataset EV10). Of these orthologous genes, 28 contain validated MXEs in human, nine were validated to be spliced differently in human, and seven could not be validated in human because read mapping data are still missing; 20 (71%) of the genes containing validated MXEs represent cases of incompatible reading frames leading to NMD in case of joined inclusion, and for 18 of these MXE clusters multiple MXE-joining reads were found (Appendix Figs S34 and S35). We further analysed the 28 orthologous genes with validated MXEs and found five genes with homologous MXE clusters (identical position in gene, identical exon phase), 13 genes with MXE clusters in human that have homologous exons in *Drosophila* and eight genes with MXEs in human where the corresponding sequence regions in the orthologous *Drosophila* genes are part of larger exons (Appendix Figs S35 and S36). The presence of orthologous MXE clusters has been attributed to convergent evolution (Copley, 2004), although the respective analysis was in part based on the comparison of non-orthologous genes (e.g. comparing human sodium channel genes [e.g. *SCN1A*] with the *Drosophila* calcium channel *cac* gene and not the orthologous sodium channel *para* gene). At least for muscle myosin heavy chain genes it could

be demonstrated that *Drosophila* already lost several MXE clusters compared to, for example, *Daphnia pulex* (crustacean) and lophotrochozoans (Kollmar & Hatje, 2014) and that the evolutionary history of the MXEs within each cluster is remarkably complex with multiple independent exon duplications and losses (Odrionitz & Kollmar, 2008). Thus, detailed studies including more bilaterian and non-bilaterian taxa would be necessary to finally conclude convergent or divergent evolution for each of the human and *Drosophila* MXE clusters. Although the overlap of MXEs in orthologous genes of human and *Drosophila* is very low, the MXE gain and loss rates are very similar (Hatje & Kollmar, 2013) indicating a conserved role of tandem exon duplication in bilaterians. Gene structures can be highly conserved between kingdoms (Rogozin et al, 2003), and certain exons therefore seem to be predisposed to undergo duplication. In summary, these findings provide strong evidence for many MXE gain and loss events during mammalian evolution, suggesting a pronounced role of these processes in speciation and establishing phenotypic differences.

Discussion

Using stringent criteria, including sequence similarity, reading frame conservation and similar lengths, and billions of RNA-Seq reads, we generated a strongly validated atlas of 1,399 human MXEs providing insights into mutually exclusive splicing mechanics, specific expression patterns, susceptibility for pathogenic mutations and deep evolutionary conservation across 18 mammals. The presented increase in human MXEs by an order of magnitude lifts MXEs into the present-day dimension of other human alternative splice types (Pan et al, 2008; Wang et al, 2008; Gerstein et al, 2014). Saturation analysis and the existence of 1,816 expressed but unconfirmed MXE candidates suggest a potential 27% increase in the MXE-ome with a twofold increase in data. Although alternative splice variants are abundant at the transcriptome level, recent mass spectrometry analyses suggested only small numbers of alternative transcripts to be translated (Abascal et al, 2015a; Ezkurdia et al, 2015; Blencowe, 2017; Tress et al, 2017a,b). Interestingly, MXEs were particularly enriched in the translated alternative transcripts, compared to other splice variants. However, ribosome profiling data showed high frequencies of ribosome engagement of cassette exons indicating that these isoforms are likely translated (Weatheritt et al, 2016). Similar results have been obtained through polyribosome profiling (Sterne-Weiler et al, 2013; Floor & Doudna, 2016). These observations suggest that most of the MXEs evaluated at the transcript level will also be found in the proteome.

About half (47%) of the 1,399 MXEs represent novel exons, which are often expressed at low levels and whose expression is restricted to few tissues and cell types, possibly explaining their absence from current genome annotations. Extrapolating these observations to all splice types and genes suggests the existence of thousands yet unannotated exons in introns. This estimation is in accordance with a recent analysis of more than 20,000 human RNA-Seq datasets that revealed over 55,000 junctions not present in annotations (Nellore et al, 2016). In this analysis, junctions found in at least 20 reads across all samples were termed “confidently called”. Although the total number of reads required for MXE validation in our analysis is lower (≥ 2 SJ reads in the 1SJ case, ≥ 6 SJ

reads in the 3SJ case), the numbers seem more conservative given that we used 40 times less data for the validation.

The almost 10-fold increase in the human MXE-ome supports recent suggestions that mutually exclusive splicing might play a much more frequent role than anticipated (Pan *et al*, 2008; Wang *et al*, 2008; Ezkurdia *et al*, 2012; Abascal *et al*, 2015a). By comparing differentially expressed MXEs across cell types, tissue types and development, we could show that 14% of all genes with MXE clusters are shared between the three data sources, and 39% between any two. Most notably, however, it is almost always a different MXE from the same cluster that is differentially expressed, and only 3.3% of the MXEs are differentially expressed in all three data sources. We believe that this indicates a high spatio-temporal regulation of all MXEs in two-exon and multi-exon clusters. We rarely observed switch-like expression with only one of the MXEs of each cluster present in each cell- or tissue type or developmental stage. Rather, one of the MXEs (“default MXE”) of each cluster was present in most or all samples and the other MXEs were expressed in several selected tissues and developmental stages (“regulated MXEs”) in addition to the default MXE. Although the “regulated MXE” is usually expressed at lower level compared to the “default MXE”, there is almost always at least a single tissue or developmental stage where it is expressed at higher level. This supports previous assertions on the modulatory and compensatory effects of the regulated MXE on the enzymatic, structural or protein interaction functions of the affected protein domains (Letunic *et al*, 2002; Tress *et al*, 2017a).

The concerted annotation and splicing analysis of novel exons have deep implications for the detection and interpretation of human disease (Bamshad *et al*, 2011; Gonzaga-Jauregui *et al*, 2012; Xiong *et al*, 2015; Bowdin *et al*, 2016). For one, exome and panel sequencing remains the method of choice for the detection of genetic diseases and both methods rely on current exon annotations (Chong *et al*, 2015). Furthermore, our data suggest that MXE expression might reflect disease pathogenesis that could allow for the prediction of the affected organ(s). It is intriguing to speculate that the observed expression–disease association is a general dogma, which could be used to predict yet unseen diseases from published expression data, potentially bringing about a paradigmatic shift in (computational) disease research.

Materials and Methods

Data sources

The human genome assembly and annotated proteins (all isoforms) were obtained from GenBank (v. 37.3) (Benson *et al*, 2013). For MXE candidate validation, we selected data from 515 publically available samples comprising 31 tissues and organs, 12 cell lines and seven developmental stages (Barbosa-Morais *et al*, 2012; Djebali *et al*, 2012; Tilgner *et al*, 2012; Xue *et al*, 2013; Yan *et al*, 2013; Fagerberg *et al*, 2014) amounting to over 15 billion RNA-Seq reads. The data were chosen to encompass common and rare potential splice events in a broad range of tissues, cell types and embryonic stages. These RNA-Seq data were obtained from either GEO (NCBI) or ENA (EBI) databases (Dataset EV1). The description of the respective tissues and developmental stages is also listed in Dataset EV1.

Reconstruction of gene structures

The gene structures for the annotated proteins were reconstructed with Scipio (Keller *et al*, 2008; Hatje *et al*, 2013) using standard parameters except `-max_mismatch=7`, `-region_size=20000`, `-single_target_hits`, `-max_move_exon=10`, `-gap_to_close=0`, `-blat_oneoff=false`, `-blat_score=15`, `-blat_identity=54`, `-exhaust_align_size=20000`, and `-exhaust_gap_size=50`. We let Scipio start with `blat_tilesize=7` and, if the entire gene structure could not be reconstructed, reduced the `blat_tilesize` step by step to 4. All parameters are less stringent than default parameters to increase the chance to reconstruct all genes automatically.

Predicting mutually exclusive spliced exons

The human genome annotation does not contain specific attributes for alternative splice variants and thus does not allow extracting or obtaining lists for specific splice types. As mutually exclusive spliced exons (MXEs), we regarded those neighbouring exons of a gene locus that are present in only one of the annotated splice variants. These MXEs were termed “annotated MXEs”. However, exons appearing mutually exclusive are not necessarily spliced as MXEs. Terminal exons, for example, are included in transcripts by alternative promoter usage and by alternative cleavage and polyadenylation. MXEs were predicted in the reconstructed genes using the algorithm implemented in WebScipio (Pillmann *et al*, 2011). The minimal exon length was set to 10 aa (`-min_exon_length=10`). WebScipio determines the length of each exon (“search exon”) and generates a list of potential exonic regions with identical lengths (to preserve the reading frame) within the neighbouring up- and downstream introns. To account for potential insertions, we allowed length differences between search exon length and potential new exonic region of up to 60 nucleotides in steps of three nucleotides [`-length_difference=20` (given in aa)], thus obtaining a list of “exon candidates”. WebScipio then translates all exon candidates in the same reading frame as the search exon and removes all sequences that contain an in-frame stop codon. In case of overlapping exonic candidate regions, we modified the original WebScipio algorithm to favour exonic regions with GT–AG splice junctions over other possible splice sites (GC–AG and GG–AG). The translations of the exon candidates are then compared to the translations of the search exons, and candidates with an amino-acid similarity score of more than 10 (`-min_score=10`) are included in the final list of MXE candidates. Because the exon candidate scoring is done at the amino acid level, WebScipio expects candidates for 5′ exons of genes to start with a methionine, and candidates for 3′ exons of genes to end with a stop codon. This minor limitation is due to WebScipio’s original development as gene reconstruction software. MXE candidates for terminal exons were only searched in direction to the next/previous internal exon. The reason for looking for MXE candidates of annotated terminal exons is that we cannot exclude that further up- and downstream exons are missing in the annotation, which would turn the new MXE candidates to internal exons. Because of the described minor limitation, however, we can only propose MXE candidates if supposed additional up- and downstream exons are non-coding exons. Because terminal exons are

reads in the 3SJ case), the numbers seem more conservative given that we used 40 times less data for the validation.

The almost 10-fold increase in the human MXE-ome supports recent suggestions that mutually exclusive splicing might play a much more frequent role than anticipated (Pan *et al*, 2008; Wang *et al*, 2008; Ezkurdia *et al*, 2012; Abascal *et al*, 2015a). By comparing differentially expressed MXEs across cell types, tissue types and development, we could show that 14% of all genes with MXE clusters are shared between the three data sources, and 39% between any two. Most notably, however, it is almost always a different MXE from the same cluster that is differentially expressed, and only 3.3% of the MXEs are differentially expressed in all three data sources. We believe that this indicates a high spatio-temporal regulation of all MXEs in two-exon and multi-exon clusters. We rarely observed switch-like expression with only one of the MXEs of each cluster present in each cell- or tissue type or developmental stage. Rather, one of the MXEs (“default MXE”) of each cluster was present in most or all samples and the other MXEs were expressed in several selected tissues and developmental stages (“regulated MXEs”) in addition to the default MXE. Although the “regulated MXE” is usually expressed at lower level compared to the “default MXE”, there is almost always at least a single tissue or developmental stage where it is expressed at higher level. This supports previous assertions on the modulatory and compensatory effects of the regulated MXE on the enzymatic, structural or protein interaction functions of the affected protein domains (Letunic *et al*, 2002; Tress *et al*, 2017a).

The concerted annotation and splicing analysis of novel exons have deep implications for the detection and interpretation of human disease (Bamshad *et al*, 2011; Gonzaga-Jauregui *et al*, 2012; Xiong *et al*, 2015; Bowdin *et al*, 2016). For one, exome and panel sequencing remains the method of choice for the detection of genetic diseases and both methods rely on current exon annotations (Chong *et al*, 2015). Furthermore, our data suggest that MXE expression might reflect disease pathogenesis that could allow for the prediction of the affected organ(s). It is intriguing to speculate that the observed expression–disease association is a general dogma, which could be used to predict yet unseen diseases from published expression data, potentially bringing about a paradigmatic shift in (computational) disease research.

Materials and Methods

Data sources

The human genome assembly and annotated proteins (all isoforms) were obtained from GenBank (v. 37.3) (Benson *et al*, 2013). For MXE candidate validation, we selected data from 515 publically available samples comprising 31 tissues and organs, 12 cell lines and seven developmental stages (Barbosa-Morais *et al*, 2012; Djebali *et al*, 2012; Tilgner *et al*, 2012; Xue *et al*, 2013; Yan *et al*, 2013; Fagerberg *et al*, 2014) amounting to over 15 billion RNA-Seq reads. The data were chosen to encompass common and rare potential splice events in a broad range of tissues, cell types and embryonic stages. These RNA-Seq data were obtained from either GEO (NCBI) or ENA (EBI) databases (Dataset EV1). The description of the respective tissues and developmental stages is also listed in Dataset EV1.

Reconstruction of gene structures

The gene structures for the annotated proteins were reconstructed with Scipio (Keller *et al*, 2008; Hatje *et al*, 2013) using standard parameters except `-max_mismatch=7`, `-region_size=20000`, `-single_target_hits`, `-max_move_exon=10`, `-gap_to_close=0`, `-blat_oneoff=false`, `-blat_score=15`, `-blat_identity=54`, `-exhaust_align_size=20000`, and `-exhaust_gap_size=50`. We let Scipio start with `blat_tilsize=7` and, if the entire gene structure could not be reconstructed, reduced the `blat_tilsize` step by step to 4. All parameters are less stringent than default parameters to increase the chance to reconstruct all genes automatically.

Predicting mutually exclusive spliced exons

The human genome annotation does not contain specific attributes for alternative splice variants and thus does not allow extracting or obtaining lists for specific splice types. As mutually exclusive spliced exons (MXEs), we regarded those neighbouring exons of a gene locus that are present in only one of the annotated splice variants. These MXEs were termed “annotated MXEs”. However, exons appearing mutually exclusive are not necessarily spliced as MXEs. Terminal exons, for example, are included in transcripts by alternative promoter usage and by alternative cleavage and polyadenylation. MXEs were predicted in the reconstructed genes using the algorithm implemented in WebScipio (Pillmann *et al*, 2011). The minimal exon length was set to 10 aa (`-min_exon_length=10`). WebScipio determines the length of each exon (“search exon”) and generates a list of potential exonic regions with identical lengths (to preserve the reading frame) within the neighbouring up- and downstream introns. To account for potential insertions, we allowed length differences between search exon length and potential new exonic region of up to 60 nucleotides in steps of three nucleotides [`-length_difference=20` (given in aa)], thus obtaining a list of “exon candidates”. WebScipio then translates all exon candidates in the same reading frame as the search exon and removes all sequences that contain an in-frame stop codon. In case of overlapping exonic candidate regions, we modified the original WebScipio algorithm to favour exonic regions with GT–AG splice junctions over other possible splice sites (GC–AG and GG–AG). The translations of the exon candidates are then compared to the translations of the search exons, and candidates with an amino-acid similarity score of more than 10 (`-min_score=10`) are included in the final list of MXE candidates. Because the exon candidate scoring is done at the amino acid level, WebScipio expects candidates for 5′ exons of genes to start with a methionine, and candidates for 3′ exons of genes to end with a stop codon. This minor limitation is due to WebScipio’s original development as gene reconstruction software. MXE candidates for terminal exons were only searched in direction to the next/previous internal exon. The reason for looking for MXE candidates of annotated terminal exons is that we cannot exclude that further up- and downstream exons are missing in the annotation, which would turn the new MXE candidates to internal exons. Because of the described minor limitation, however, we can only propose MXE candidates if supposed additional up- and downstream exons are non-coding exons. Because terminal exons are

included in transcripts by alternative promoter usage and by alternative cleavage and polyadenylation, we treated the list of terminal exon candidates separately (Appendix Fig S4). This list might be of interest for further investigation for other researchers. Except for this Appendix Fig S4, we entirely focused on internal MXE candidates.

Definition of criteria for RNA-Seq evaluation of the MXE candidates

While the sole mapping of RNA-Seq reads reveals the transcription of the respective genomic region, it does not prove the inclusion into functional transcripts. The mutually exclusive inclusion of the MXE candidates into functional transcripts requires at least the following splice junction (SJ) reads (Appendix Fig S5): (i) There must be SJ reads matching from every MXE to up- or downstream gene regions bridging the other MXEs of the cluster. The latter criterion takes into account that the annotated exons neighbouring the clusters of MXEs might not themselves be constitutive but alternative exons as, for example, in *NCX1* (Appendix Fig S6). (ii) SJ reads mapping from one to another MXE candidate lead to MXE candidate rejection except for those MXEs leading to a frame shift. Without this constraint, which has not been set in earlier analyses (Wang *et al*, 2008), MXEs cannot be distinguished from neighbouring differentially included exons, which are quite common in human (data not shown; see e.g. Hammesfahr & Kollmar, 2012 and Appendix Fig S6). Thus, there are three constraints for a cluster of two MXEs while clusters of three and five MXEs, for example, already require seven and 18 constraints, respectively (Appendix Figs S5 and S7). Under more stringent conditions, also SJ reads from MXEs to the neighbouring annotated exons independent of their splice type would be required giving rise to five constraints for a cluster of two MXEs (Appendix Fig S5).

Note that as a matter of principle the read coverage of MXEs and other alternative splicing events is considerably lower than that of constitutive exons due to their mutually exclusive inclusion in the transcripts. For example, each of the exons of a cluster of three MXEs is expected to only have, on average, one-third the coverage of the constitutive exons of the same gene. The number of predicted exons, of which both sites are supported by splice junction reads, is also considerably lower than the total number of supported MXE candidates (Appendix Fig S3), which we think is due to the general low coverage of the exons and not due to read mapping and exon border prediction problems (Appendix Fig S3).

Validation of the MXE candidates by RNA-Seq mapping

SRA files were converted to FASTQ files using `fastq-dump` software (v. 2.1.18). FASTQ files were mapped onto the human reference genome (hg19) using the STAR aligner (v.2.3.0e_r291) (Dobin *et al*, 2013). To this end, we first generated a reference genome index with `-sjdbGTFfeatureExon`, `-sjdbGTFtagExonParentTranscript`, a splice junction overhang size of 99 (`-sjdbOverhang`) and GTF annotation files containing all transcripts and all MXE candidates. The MXE candidate GTF file was extracted from Kassiopeia database and is available for download there (Hatje & Kollmar, 2014). The mapping was done for each sample separately. We allowed a rather stringent maximum

mismatch of 2 (`-outFilterMismatchNmax 2`; STAR default is 10) and the output was forced to SAM format (`-outStd SAM`). Otherwise, default settings were used. The resulting files with the mapped reads were sorted, converted to BAM format and indexed with SAMtools (`sort -n`) for further processing (Li *et al*, 2009).

Distinguishing MXEs from other splice variants

For the analysis of the read mapping data, we disassembled clusters with more than two MXE candidates into all possible sub-clusters. For example, a cluster with four MXE candidates [1,2,3,4] was fractionated into the following sub-cluster: [1,2], [2,3], [3,4], [1,2,3], [2,3,4], [1,2,3,4]. Each of these sub-clusters was analysed independently according to the validation criteria (splice junction reads present, exon-joining reads absent). If all criteria were satisfied for one of the sub-clusters, all MXE candidates of the respective sub-cluster were labelled “verified”. In a second analysis, each cluster of MXE candidates was analysed for exon-joining reads, which denote constitutive splicing or splicing as differentially included exons. However, MXE candidates of clusters and sub-clusters with exon-joining reads but exon lengths not divisible by three were also flagged as “verified” because their combined inclusion would lead to a frame shift in the translation of the transcript.

Limits of the MXE dataset

Similar to every genome annotation dataset, also the current dataset of RNA-Seq validated MXEs has some limitations. Some are inherent to the still incomplete human genome annotation that was used as basis for generating the list of MXE candidates. As mentioned above and shown in Appendix Fig S2C, there are genes with mis-annotated terminal exons overlapping MXEs. Also, there are “transcripts” in the GenBank dataset that combine exons from (now) different genes. The presence of these “transcripts” in the genome annotation might be the result of mis-interpreting cDNA data as coding sequence although these might be the result of some level of mis-splicing.

Similarly, mis-splicing might be an important reason for validating true MXEs as “non-MXEs”. A single exon-joining read turns MXE candidates into non-MXEs, whose mutually exclusive splicing might otherwise be supported by thousands of MXE-bridging SJ reads. Given these limitations, we expect that many of the exons, that we currently tag as constitutive or other alternative splicing, might in fact be MXEs. On the other hand, our MXE dataset might also contain some exons that are in fact non-MXEs. This is well demonstrated in the saturation analysis (Fig 1C) showing that although more data will lead to the validation of many more exons as MXEs, for which SJ reads are currently missing, there will be clusters that will be rejected as soon as more data include exon-joining reads. In addition, some MXEs with only a few supporting SJ reads might in fact be pseudoexons. However, we also did not observe any SJ reads for about 15% of the annotated exons, which are nevertheless not regarded as pseudoexons (Fig 1B, Appendix Fig S3). Finally, some MXEs determined from transcripts showing complex splicing might in fact be mutually exclusive in transcripts, but not in the sense of a cluster of uninterrupted neighbouring exons.

Saturation analysis

Theoretically, increasing the number of samples should also increase the number of validated MXEs, as the total increase in read number for different observed or novel tissues should increase the read evidence for the predicted MXEs. At the same time, increasing the number of reads also heighten the chance of rejecting an MXE candidate. This raises the question of what the expected number of validated and rejected MXEs for increasing numbers of samples is. Additionally, it would be interesting to obtain the theoretical point of saturation, the maximum expected number of MXEs in the human genome.

To obtain this information, sub-samples of STAR-aligned RNA-Seq splice junction (SJ) reads were used to estimate the expected recall and false-positive rate (Fig 1C, Appendix Fig S11). The number of verified MXEs was calculated using SJ reads for different percentages of the data. Similarly, the number of rejected MXEs was obtained. To reduce the bias from data sampling, datasets were chosen randomly and the saturation analysis was performed in 30 independent runs. To calculate the mean of validated and rejected MXEs at respective percentages of the total RNA-Seq data used for validation, we used the respective numbers from the 30 independent runs.

To estimate the potential increase in MXEs given more sequencing data, we fit the sub-sampling data to the number of expected MXEs $f(x)$ using Matlab and the optimal fits were obtained for a power function

$$f(x) = a * x^b + c$$

with the linear coefficient a , the exponential coefficient b and the error term c (Appendix Fig S11B). Given a twofold increase in the number of reads, the expected number of validated MXEs (1SJ) is $1,769 \pm 47$ (95% confidence interval), validated MXEs (3SJ) is $1,081 \pm 12$, rejected MXEs (1SJ) is 227 ± 9 , and the number of rejected MXEs (3SJ) is 95 ± 5 (Appendix Fig S11B). While the number of validated MXEs is far from saturation (a 100% increase in data results in 27% increase in the number of validations), the number of rejected MXEs seems to be saturated (a 100% increase in data results in 2% increase in the number of rejections).

qPCR validation of MXE candidates

Total RNA was purified from healthy human brain tissue (substantia nigra) using Trizol kit (Tri Reagent, Sigma T9424) following manufacturer's instructions. RNA was further purified using the RNA Clean & Concentrator © TM -5 kit (Zymo Research, cat. R1013). The RNA quality was investigated using the 6000 nano assay on a Bio-analyzer 2100 (Agilent Technologies). Reverse transcription was carried out using the iScript © cDNA Synthesis kit (cat# 1708890, Bio-Rad) using approximately 500 ng of total RNA in a volume of 20 µl.

Relative expression levels of the genes of interest as well as one housekeeping gene (glyceraldehyde 3-phosphate dehydrogenase [*Gapdh*]) were determined by qPCR using a LightCycler® 480. All qPCR experiments were performed in duplicates using SYBR™ Green PCR Master Mix (cat # 4309155). For each PCR, 20 ng cDNA

was used and negative controls contained no cDNA. The qPCR was run under the following conditions: pre-incubation at 95°C for 5 min, denaturation at 95°C for 10 s, annealing 60°C for 15 s, extension at 72°C for 10 s repeated for 40 cycles (Sybr green standard protocol II). Detailed information on the primers and qPCR results can be found in Dataset EV3.

Analysis of the splice mechanism

To determine the distance between intron donor site and branch point, we analysed all introns smaller than 500 bp using the standalone version of SVM-BPfinder (beta) (Corvelo *et al*, 2010) to predict branch point locations. Longer introns harbour high numbers of branch point candidates, and the accuracy of the branch point prediction considerably decreases. Longer introns also often contain multiple branch points with different splicing kinetics (Corvelo *et al*, 2010) so that a steric hindrance criterion for splicing multiple MXEs into the same transcript might not apply anymore. Branch points are usually located in the 3' regions of the introns and it seems highly unlikely to identify only a single potential branch point within an, for example, > 2,000-bp intron, which would in addition be located within the 5' 50 bps. Thus, the highest-scoring location within the < 500-bp introns was taken as best guess for the branch point and the distance to the intron donor site determined.

In order to identify U12-type introns, we analysed all donor splice sites of the introns preceding the clusters of MXEs and those subsequent to all MXEs using the consensus pattern described by Sharp and Burge (Sharp & Burge, 1997). The acceptor splice sites of U12-type introns do not show conserved patterns and were therefore not used here for verification.

Binding windows for competing intron RNA secondary structures were predicted for all candidate clusters of MXEs using the SeqAn package (Döring *et al*, 2008). The identified binding windows of all homologous genes were aligned using MUSCLE (Edgar, 2004) and the RNA secondary structures predicted by RNAalifold (ViennaRNA package) (Lorenz *et al*, 2011).

Mapping MXE sequences onto protein structures

To identify the best structural models for the sequences encoded by the MXEs, we mapped the protein sequences of the respective genes against available protein structure data. To this end, we made use of a recently developed database, called Allora (<http://allora.motorprotein.de>), in which genomic information is mapped onto protein structures. Allora currently contains 94,148 PDB entries (derived from the RCSB Protein Data Bank, <http://www.rcsb.org>, Rose *et al*, 2015) with 247,959 chains, of which 120,665 represent unique sequences. Based on the database references in the PDB entries, the full-length proteins were fetched from UniProt KB (UniProt Consortium, 2015) or GenBank (Benson *et al*, 2013) and the corresponding gene structures of the eukaryotic proteins reconstructed with WebScipio (Hatje *et al*, 2013). In Allora, all PDBs belonging to the same UniProt or GenBank entries are connected. BLAST+ (Camacho *et al*, 2009) was used to search for the most similar UniProt/GenBank protein sequence compared to the human proteins containing MXEs. The hit with the lowest E-value was taken, and the associated PDB chains were aligned to the human protein using m-coffee (Wallace *et al*, 2006). The MXE part of the alignment was

extracted for further analysis (\Rightarrow “MXE structure”). As “intron distances”, we determined the distances between the CA atoms of the first and the last residues of the MXE structures.

Evaluating the differential inclusion of MXEs into transcripts

Splice junction read counts were extracted from STAR output “SJ.out.tab” files. For each MXE in a cluster, the per cent-spliced-in (PSI) value was calculated by dividing the number of junction reads of the MXE by the sum of junction reads for all MXEs in the same cluster. Differential inclusion analysis on the Human Protein Atlas, Embryonic Development and ENCODE datasets was performed using a Kruskal–Wallis rank sum test with a Benjamini–Hochberg (BH) multiple testing correction. Values were computed using the “kruskal.test” and “p.adjust” functions in R. For each project, we created a design matrix with sample name and experimental condition and replicate numbers. The results of the differential inclusion analysis are summarized in Dataset EV5.

Differential expression of pairs of annotated and novel MXEs

For each sample (tissue, cell type and developmental stage), we calculated the median RPKM (reads per kilobase of transcript per million mapped reads) from the replicates for each MXE. To compile a set of MXEs with significant expression, only pairs of MXEs were selected of which either the annotated or the novel exon had a median expression of more than 3. The number of MXEs for this analysis would not considerably decrease if a cut-off of 30 were chosen (252 MXEs at a cut-off of 3 versus 240 MXEs at a cut-off of 30). For each pair of MXEs, we subtracted the PSI value of the ubiquitous/known/non-SNP-containing MXE from the PSI value of the respective specific/novel/SNP-containing MXE (delta PSI values) and scaled those values between -1 (high PSI for ubiquitous/known/non-SNP-containing MXE) and 1 (high PSI for specific/novel/SNP-containing MXE) (see also Figs 3A and 4C, Appendix Fig S23). In case an MXE pair was not expressed in a certain tissues (NA or 0), the value was set to 0.

Inequality analysis

The mean PSI values of each MXE were calculated for each tissue in the Human Protein Atlas project, each developmental stage in the embryonic development (Peking University) project, and each cell type in the ENCODE (Caltech) project. For each MXE, the Gini index (Ceriani & Verme, 2012) was calculated independently for each project based on the mean PSI values using the Gini function with standard parameters from the *ineq* R package version 0.2-13 (Achim Zeileis, Christian Kleiber, <https://CRAN.R-project.org/package=ineq>; Cowell, 2011). For the analysis of MXE clusters, only those clusters were taken into account that include at least two MXEs with an RPKM ≥ 10 in at least one dataset within each project. Furthermore, we excluded clusters where all MXEs have “NA” PSI values within each project (244, 96 and 225 clusters, respectively).

Identification of pathogenic SNPs in MXEs

To identify potentially pathogenic SNPs in MXEs, the MXEs were compared to the ClinVar SNP database (ClinVar VCF file

downloaded on 11 Aug 2016, version updated at 30 Jun 2016, Landrum *et al*, 2016). The ClinVar variant summary file (VCF file) was converted into a BED file keeping all original information. Positions overlapping between MXEs and ClinVar-SNPs were accessed using the BEDTools feature intersection software (Quinlan & Hall, 2010). SNPs are classified as pathogenic or non-pathogenic according to ClinVar’s “ClinicalSignificance” field annotation. All entries containing “benign” and all structural variations were removed. All ClinVar-SNPs overlapping with MXEs were manually verified in order to keep only potentially pathogenic variations.

To access the statistical significance of disease enrichment in MXEs and cassette exons, we compared the amount of pathogenic SNP-containing to non-SNP-containing exons. Of 615,410 annotated exons, 21,030 (3.4%) contain pathogenic SNPs; of 1,399 MXEs, 99 (7.1%) contain pathogenic SNPs; and of 31,745 cassette exons, 2,143 (6.8%) contain pathogenic SNPs. The ~ 2 -fold enrichment of alternative splicing-associated exons (MXEs and cassette exons) is highly significant (Fisher’s exact test, P -value MXE = 3×10^{-11} , P -value cassette = 2.2×10^{-16}).

Disease prediction using pathogenic SNPs in MXEs

In order to predict disease from MXE expression, we first filtered for MXEs that had a minimal RPKM value of 3 and then subtracted the expression of the non-SNP-containing MXE from the SNP-containing MXE for all MXE pairs with mutations, across all developmental stages, tissues and cell types (49 features per MXE pair). Delta PSI values (PSI for SNP-containing MXE—PSI for non-SNP-containing MXE) were subsequently scaled and centred, and the MXE pairs were annotated to two disease classes, cardiomyopathy-neuromuscular disease ($n = 10$) or other diseases ($n = 14$). We regrouped genes into these categories to obtain relatively balanced categories while keeping a minimum of 10 observations per category.

Classification with limited observations needs careful execution, as over-fitting (high variance) and under-fitting (high bias) are common problems. To avoid high variance or bias, several crucial steps were taken. First, we did not optimize hyperparameters, using a Random Forest with 250 trees and a maximum tree depth of 16 (number of predictors/3). Second, we used leave-one-out cross-validation to avoid sampling bias and model instability. Third, diseases were grouped into two categories of relatively even size (see above). Models were built using the R packages *caret* (Kuhn, 2008) and *randomForest*, and ROC curves were generated with *ROCR* (Sing *et al*, 2005).

Of note, models trained on PSI values (considering only the PSI value of the SNP-containing MXE, data not shown) or RPKM values (Appendix Fig S29) obtained similar accuracies as the model trained on delta PSI values, indicating the stability of the prediction across slight variations in feature pre-processing.

Gene ontology enrichment analysis

We used WebGestalt for Gene Ontology enrichment analyses (Wang *et al*, 2013). The lists of unique genes in gene symbol format were uploaded to WebGestalt and the GO Enrichment Analysis selected. The entire human genome annotation was set as background and 0.05 as threshold for the P -value for the significance test using the

default statistical method “hypergeometric”. Categorical enrichment of MXEs and cassette exons was summarized in a heatmap.

Protein–protein interaction analysis

The protein–protein interaction network was built by using GeneMANIA webservice (Warde-Farley *et al*, 2010). The list of unique genes containing a pathogen SNP was submitted to GeneMANIA’s webservice, and we downloaded the resulting network in SVG format and manually included disease and ontology information.

Assessing the dynamics of MXE annotations over time

MXEs might have already been annotated/described although not been included in the NCBI reference dataset. This might especially account for newer annotations based on the recently published ENCODE project data. Therefore, we obtained alternative protein sequence datasets from Aceview (Thierry-Mieg & Thierry-Mieg, 2006) and Ensembl (Yates *et al*, 2016). Further datasets like the VEGA and GENCODE annotations are continuously integrated into Ensembl and were therefore not considered separately. The Aceview database has been built in the year 2000 to represent comprehensive and non-redundant sequences of all public mRNA sequences. The human dataset has last been updated in November 2011, thus before the availability of the ENCODE data.

To assess the novelty of our MXE assignments with respect to the timely updates and changes of the human annotations, we compared our data with that of Aceview and with the latest annotation from Ensembl (Fig 5, Appendix Fig S1). As at the beginning of the project, only a few MXEs are annotated as such in other databases. Surprisingly, however, many of the previously annotated exons (independent of their splicing status) were removed from the latest Ensembl annotation, although our RNA-Seq mapping not only strongly supports their inclusion into transcripts but also their splicing as MXEs. This shows that further collaborative efforts are needed to reveal a stable and persistent human gene annotation.

Ab initio exon prediction

Exon prediction by *ab initio* gene finding software is another means of generating a database of potential coding sequences. *Ab initio* exon prediction was done with AUGUSTUS (Stanke & Waack, 2003) using default parameters to find alternative splice forms and the feature set for *Homo sapiens*.

Identifying orthologous proteins in 18 mammals

Cross-species searches in 18 mammals (Dataset EV9) were done with WebScipio (Hatje *et al*, 2013) with same parameters as for gene reconstructions except `-min_identity=60`, `-max_mismatch=0` (allowing any number of mismatches), `-gap_to_close=10`, `-min_intron_length=35`, `-blat_tilesize=6` and `-blat_oneoff=true`. MXE candidates in cross-species gene reconstructions were searched with `-length_difference=20`, `-min_score=15` and `-min_exon_length=15`, for all exons in all introns but not in up- and downstream regions. Reasons for not detecting clusters of MXEs might be gene and MXE loss events, sequence divergence precluding ortholog identification, and

assembly gaps. For determining the origin of a conserved MXE cluster, we used a simple model expecting each shared cluster to be already present in the last common ancestor of the respective species. This approach is equivalent to inferring ancestral character states with Dollo parsimony (Farris, 1977).

Comparing human genes with MXEs to orthologous genes in *Drosophila melanogaster*

Orthologous genes in *D. melanogaster* for all human genes containing MXE candidates were obtained with the Ensemble BioMart service (Yates *et al*, 2016). This list of orthologous genes was filtered with the list of *D. melanogaster* genes containing MXEs, which was obtained from Hatje and Kollmar (2013), to obtain a list of genes with both types of exons, (i) MXEs in human and MXEs in *D. melanogaster*, and (ii) MXE candidates in human but validated to be spliced differently and MXEs in *D. melanogaster*. Several of the human and *D. melanogaster* genes contain multiple clusters of MXEs. Thus, we compared all genes manually to determine whether MXEs are orthologous in both species, whether MXEs in human have orthologous exons in *D. melanogaster*, and whether MXEs in human do not correspond to exons in *D. melanogaster* genes.

Data availability

All generated data can be searched, filtered and browsed at Kassiopeia (www.motorprotein.de/kassiopeia; Hatje & Kollmar, 2014). The primary RNA-Seq datasets used in this study are available in the following databases:

<http://www.ebi.ac.uk/ena/data/view/ERP003613>
<http://www.ebi.ac.uk/ena/data/view/ERP000546>
<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE36552>
<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE44183>
<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE33480>
<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30567>

Expanded View for this article is available online.

Acknowledgements

We would like to thank Prof. Paul Lingor and Lucas Araujo Caldi Gomes of the University Medicine Göttingen for providing total RNA isolated from human brain. We would like to thank Daniel Sumner Magruder from the Bonn group for critical suggestions. In particular, we would like to thank André Ahrens from the Kollmar group for his tremendous help in implementing the Allora database and performing the mapping of MXEs onto protein structures. The Kollmar group would like to thank Prof. Christian Griesinger for his continuous generous support. We would like to thank Dr. Robert P. Zinzen and Dr. Carla Margulies for critical reading of the manuscript.

Author contributions

MK initiated the study and designed the analyses together with SB. KH and BH performed MXE predictions. KH integrated the human MXE data into the Kassiopeia database. KH implemented MXE candidate extraction and RNA-Seq analysis with help from ROV and SB. KH and MK did the cluster distribution, splicing mechanism and protein structure mapping analyses. KH, ROV, R-UR, VB, SB and MK performed the differential expression analysis. AR, MEM and TS performed the RT-PCR experiments. ROV, R-UR, VB and SB performed the SNP

default statistical method “hypergeometric”. Categorical enrichment of MXEs and cassette exons was summarized in a heatmap.

Protein–protein interaction analysis

The protein–protein interaction network was built by using GeneMANIA webservice (Warde-Farley *et al*, 2010). The list of unique genes containing a pathogen SNP was submitted to GeneMANIA’s webservice, and we downloaded the resulting network in SVG format and manually included disease and ontology information.

Assessing the dynamics of MXE annotations over time

MXEs might have already been annotated/described although not been included in the NCBI reference dataset. This might especially account for newer annotations based on the recently published ENCODE project data. Therefore, we obtained alternative protein sequence datasets from Aceview (Thierry-Mieg & Thierry-Mieg, 2006) and Ensembl (Yates *et al*, 2016). Further datasets like the VEGA and GENCODE annotations are continuously integrated into Ensembl and were therefore not considered separately. The Aceview database has been built in the year 2000 to represent comprehensive and non-redundant sequences of all public mRNA sequences. The human dataset has last been updated in November 2011, thus before the availability of the ENCODE data.

To assess the novelty of our MXE assignments with respect to the timely updates and changes of the human annotations, we compared our data with that of Aceview and with the latest annotation from Ensembl (Fig 5, Appendix Fig S1). As at the beginning of the project, only a few MXEs are annotated as such in other databases. Surprisingly, however, many of the previously annotated exons (independent of their splicing status) were removed from the latest Ensembl annotation, although our RNA-Seq mapping not only strongly supports their inclusion into transcripts but also their splicing as MXEs. This shows that further collaborative efforts are needed to reveal a stable and persistent human gene annotation.

Ab initio exon prediction

Exon prediction by *ab initio* gene finding software is another means of generating a database of potential coding sequences. *Ab initio* exon prediction was done with AUGUSTUS (Stanke & Waack, 2003) using default parameters to find alternative splice forms and the feature set for *Homo sapiens*.

Identifying orthologous proteins in 18 mammals

Cross-species searches in 18 mammals (Dataset EV9) were done with WebScipio (Hatje *et al*, 2013) with same parameters as for gene reconstructions except `-min_identity=60`, `-max_mismatch=0` (allowing any number of mismatches), `-gap_to_close=10`, `-min_intron_length=35`, `-blat_tilesize=6` and `-blat_oneoff=true`. MXE candidates in cross-species gene reconstructions were searched with `-length_difference=20`, `-min_score=15` and `-min_exon_length=15`, for all exons in all introns but not in up- and downstream regions. Reasons for not detecting clusters of MXEs might be gene and MXE loss events, sequence divergence precluding ortholog identification, and

assembly gaps. For determining the origin of a conserved MXE cluster, we used a simple model expecting each shared cluster to be already present in the last common ancestor of the respective species. This approach is equivalent to inferring ancestral character states with Dollo parsimony (Farris, 1977).

Comparing human genes with MXEs to orthologous genes in *Drosophila melanogaster*

Orthologous genes in *D. melanogaster* for all human genes containing MXE candidates were obtained with the Ensemble BioMart service (Yates *et al*, 2016). This list of orthologous genes was filtered with the list of *D. melanogaster* genes containing MXEs, which was obtained from Hatje and Kollmar (2013), to obtain a list of genes with both types of exons, (i) MXEs in human and MXEs in *D. melanogaster*, and (ii) MXE candidates in human but validated to be spliced differently and MXEs in *D. melanogaster*. Several of the human and *D. melanogaster* genes contain multiple clusters of MXEs. Thus, we compared all genes manually to determine whether MXEs are orthologous in both species, whether MXEs in human have orthologous exons in *D. melanogaster*, and whether MXEs in human do not correspond to exons in *D. melanogaster* genes.

Data availability

All generated data can be searched, filtered and browsed at Kassiopeia (www.motorprotein.de/kassiopeia; Hatje & Kollmar, 2014). The primary RNA-Seq datasets used in this study are available in the following databases:

<http://www.ebi.ac.uk/ena/data/view/ERP003613>
<http://www.ebi.ac.uk/ena/data/view/ERP000546>
<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE36552>
<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE44183>
<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE33480>
<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30567>

Expanded View for this article is available online.

Acknowledgements

We would like to thank Prof. Paul Lingor and Lucas Araujo Caldi Gomes of the University Medicine Göttingen for providing total RNA isolated from human brain. We would like to thank Daniel Sumner Magruder from the Bonn group for critical suggestions. In particular, we would like to thank André Ahrens from the Kollmar group for his tremendous help in implementing the Allora database and performing the mapping of MXEs onto protein structures. The Kollmar group would like to thank Prof. Christian Griesinger for his continuous generous support. We would like to thank Dr. Robert P. Zinzen and Dr. Carla Margulies for critical reading of the manuscript.

Author contributions

MK initiated the study and designed the analyses together with SB. KH and BH performed MXE predictions. KH integrated the human MXE data into the Kassiopeia database. KH implemented MXE candidate extraction and RNA-Seq analysis with help from ROV and SB. KH and MK did the cluster distribution, splicing mechanism and protein structure mapping analyses. KH, ROV, R-UR, VB, SB and MK performed the differential expression analysis. AR, MEM and TS performed the RT-PCR experiments. ROV, R-UR, VB and SB performed the SNP

mapping and prediction analysis. The comparison of different human gene annotations was done by KH and DS. KH did the prediction of MXEs in other mammals, and their comparison together with MK. MK and SB wrote the manuscript. ROV, KH, VB and R-UR contributed to manuscript text and Supplementary Materials. All authors read and approved the final manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

References

- Abascal F, Ezkurdia I, Rodriguez-Rivas J, Rodriguez JM, del Pozo A, Vázquez J, Valencia A, Tress ML (2015a) Alternatively spliced homologous exons have ancient origins and are highly expressed at the protein level. *PLoS Comput Biol* 11: e1004325
- Abascal F, Tress ML, Valencia A (2015b) The evolutionary fate of alternatively spliced homologous exons after gene duplication. *Genome Biol Evol* 7: 1392–1403
- Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J (2011) Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* 12: 745–755
- Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Çolak R, Kim T, Misquitta-Ali CM, Wilson MD, Kim PM, Odom DT, Frey BJ, Blencowe BJ (2012) The evolutionary landscape of alternative splicing in vertebrate species. *Science* 338: 1587–1593
- Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2013) GenBank. *Nucleic Acids Res* 41: D36–D42
- Blencowe BJ (2006) Alternative splicing: new insights from global analyses. *Cell* 126: 37–47
- Blencowe BJ (2017) The relationship between alternative splicing and proteomic complexity. *Trends Biochem Sci* 42: 407–408
- Bowdin S, Gilbert A, Bedoukian E, Carew C, Adam MP, Belmont J, Bernhardt B, Biesecker L, Bjornsson HT, Blitzner M, D'Alessandro LCA, Deardorff MA, Demmer L, Elliott A, Feldman GL, Glass IA, Herman G, Hindorff L, Hisama F, Hudgins L et al (2016) Recommendations for the integration of genomics into clinical practice. *Genet Med* 18: 1075–1084
- Buchert R, Tawamie H, Smith C, Uebe S, Innes AM, Al Hallak B, Ekici AB, Sticht H, Schwarze B, Lamont RE, Parboosingh JS, Bernier FP, Abou Jamra R (2014) A peroxisomal disorder of severe intellectual disability, epilepsy, and cataracts due to fatty acyl-CoA reductase 1 deficiency. *Am J Hum Genet* 95: 602–610
- Buljan M, Chalancon G, Eustermann S, Wagner GP, Fuxreiter M, Bateman A, Babu MM (2012) Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Mol Cell* 46: 871–883
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10: 421
- Ceriani L, Verme P (2012) The origins of the Gini index: extracts from *Variabilità e Mutabilità* (1912) by Corrado Gini. *J Econ Inequal* 10: 421–443
- Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, Harrell TM, McMillin MJ, Wiszniewski W, Gambin T, Coban Akdemir ZH, Doheny K, Scott AF, Avramopoulos D, Chakravarti A, Hoover-Fong J, Mathews D, Witmer PD, Ling H, Hetrick K et al (2015) The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. *Am J Hum Genet* 97: 199–215
- Copley RR (2004) Evolutionary convergence of alternative splicing in ion channels. *Trends Genet* 20: 171–176
- Corvelo A, Hallegger M, Smith CWJ, Eyraas E (2010) Genome-wide association between branch point properties and alternative splicing. *PLoS Comput Biol* 6: e1001016
- Cowell F (2011) *Measuring inequality*. Oxford: Oxford University Press
- David CJ, Chen M, Assanah M, Canoll P, Manley JL (2010) HnRNP proteins controlled by c-Myc deregulate pyruvate kinase mRNA splicing in cancer. *Nature* 463: 364–368
- Diebold RJ, Koch WJ, Ellinor PT, Wang JJ, Muthuchamy M, Wieczorek DF, Schwartz A (1992) Mutually exclusive exon splicing of the cardiac calcium channel alpha 1 subunit gene generates developmentally regulated isoforms in the rat heart. *Proc Natl Acad Sci USA* 89: 1497–1501
- Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T et al (2012) Landscape of transcription in human cells. *Nature* 489: 101–108
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29: 15–21
- Döring A, Weese D, Rausch T, Reinert K (2008) SeqAn: an efficient, generic C++ library for sequence analysis. *BMC Bioinformatics* 9: 11
- Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5: 113
- Ellis JD, Barrios-Rodiles M, Colak R, Irimia M, Kim T, Calarco JA, Wang X, Pan Q, O'Hanlon D, Kim PM, Wrana JL, Blencowe BJ (2012) Tissue-specific alternative splicing remodels protein-protein interaction networks. *Mol Cell* 46: 884–892
- Ezkurdia I, del Pozo A, Frankish A, Rodriguez JM, Harrow J, Ashman K, Valencia A, Tress ML (2012) Comparative proteomics reveals a significant bias toward alternative protein isoforms with conserved structure and function. *Mol Biol Evol* 29: 2265–2283
- Ezkurdia I, Rodriguez JM, Carrillo-de Santa Pau E, Vázquez J, Valencia A, Tress ML (2015) Most highly expressed protein-coding genes have a single dominant isoform. *J Proteome Res* 14: 1880–1887
- Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, Habuka M, Tahmasebpoor S, Danielsson A, Edlund K, Asplund A, Sjöstedt E, Lundberg E, Szijarto CA-K, Skogs M, Takanan JO, Berling H, Tegel H, Mulder J, Nilsson P et al (2014) Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics* 13: 397–406
- Farris JS (1977) Phylogenetic analysis under Dollo's law. *Syst Zool* 26: 77–88
- Floor SN, Doudna JA (2016) Tunable protein synthesis by transcript isoforms in human cells. *Elife* 5: e10921
- Gerstein MB, Rozowsky J, Yan K-K, Wang D, Cheng C, Brown JB, Davis CA, Hillier L, Sisu C, Li JJ, Pei B, Harmanci AO, Duff MO, Djebali S, Alexander RP, Alver BH, Auerbach R, Bell K, Bickel PJ, Boeck ME et al (2014) Comparative analysis of the transcriptome across distant species. *Nature* 512: 445–448
- Gonzaga-Jauregui C, Lupski JR, Gibbs RA (2012) Human genome sequencing in health and disease. *Annu Rev Med* 63: 35–61
- Graveley BR, Kaur A, Gunning D, Zipursky SL, Rowen L, Clemens JC (2004) The organization and evolution of the dipteran and hymenopteran Down syndrome cell adhesion molecule (Dscam) genes. *RNA* 10: 1499–1506
- Graveley BR (2005) Mutually exclusive splicing of the insect Dscam Pre-mRNA directed by competing intronic RNA secondary structures. *Cell* 123: 65–73

- Hammesfahr B, Kollmar M (2012) Evolution of the eukaryotic dyactin complex, the activator of cytoplasmic dynein. *BMC Evol Biol* 12: 95
- Hatje K, Hammesfahr B, Kollmar M (2013) WebScipio: reconstructing alternative splice variants of eukaryotic proteins. *Nucleic Acids Res* 41: W504–W509
- Hatje K, Kollmar M (2013) Expansion of the mutually exclusive spliced exome in *Drosophila*. *Nat Commun* 4: 2460
- Hatje K, Kollmar M (2014) Kassiopeia: a database and web application for the analysis of mutually exclusive exomes of eukaryotes. *BMC Genom* 15: 115
- Irimia M, Weatheritt RJ, Ellis JD, Parikshak NN, Gonatopoulos-Pournatzis T, Babor M, Quesnel-Vallières M, Tapial J, Raj B, O'Hanlon D, Barrios-Rodiles M, Sternberg MJE, Cordes SP, Roth FP, Wrana JL, Geschwind DH, Blencowe BJ (2014) A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* 159: 1511–1523
- Johansson JU, Ericsson J, Janson J, Beraki S, Stanić D, Mandić SA, Wikström MA, Hökfelt T, Ögren SO, Rozell B, Berggren P-O, Bark C (2008) An ancient duplication of exon 5 in the Snap25 gene is required for complex neuronal development/function. *PLoS Genet* 4: e1000278
- Kaplan JM, Kim SH, North KN, Renne H, Correia LA, Tong HQ, Mathis BJ, Rodríguez-Pérez JC, Allen PG, Beggs AH, Pollak MR (2000) Mutations in ACTN4, encoding alpha-actinin-4, cause familial focal segmental glomerulosclerosis. *Nat Genet* 24: 251–256
- Keller O, Odrionitz F, Stanke M, Kollmar M, Waack S (2008) Scipio: using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. *BMC Bioinformatics* 9: 278
- Kollmar M, Hatje K (2014) Shared gene structures and clusters of mutually exclusive spliced exons within the metazoan muscle myosin heavy chain genes. *PLoS One* 9: e88111
- Kuhn M (2008) Building predictive models in R using the caret package. *J Stat Softw* 28: 1–26
- Kustatscher G, Hothorn M, Pugieux C, Scheffzek K, Ladurner AG (2005) Splicing regulates NAD metabolite binding to histone macroH2A. *Nat Struct Mol Biol* 12: 624–625
- Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, Jang W, Katz K, Ovetsky M, Riley G, Sethi A, Tully R, Villamarín-Salomon R, Rubinstein W, Maglott DR (2016) ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* 44: D862–D868
- Lee C, Kim N, Roy M, Graveley BR (2010) Massive expansions of Dscam splicing diversity via staggered homologous recombination during arthropod evolution. *RNA* 16: 91–105
- Lee Y, Rio DC (2015) Mechanisms and regulation of alternative pre-mRNA splicing. *Annu Rev Biochem* 84: 291–323
- Letunic I, Copley RR, Bork P (2002) Common exon duplication in animals and its role in alternative splicing. *Hum Mol Genet* 11: 1561–1567
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25: 2078–2079
- Lorenz R, Bernhart SH, Höner Zu Siederdisen CH, Tafer H, Flamm C, Stadler PF, Hofacker IL (2011) ViennaRNA package 2.0. *Algorithms Mol Biol* 6: 26
- Matlin AJ, Clark F, Smith CWJ (2005) Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol* 6: 386–398
- Mayr JA, Zimmermann FA, Horváth R, Schneider H-C, Schoser B, Holinski-Feder E, Czermin B, Freisinger P, Sperl W (2011) Deficiency of the mitochondrial phosphate carrier presenting as myopathy and cardiomyopathy in a family with three affected children. *Neuromuscul Disord* 21: 803–808
- Meijers R, Puettmann-Holgado R, Skiniotis G, Liu J, Walz T, Wang J, Schmucker D (2007) Structural basis of Dscam isoform specificity. *Nature* 449: 487–491
- Merkin J, Russell C, Chen P, Burge CB (2012) Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science* 338: 1593–1599
- Nellore A, Jaffe AE, Fortin J-P, Alquicira-Hernández J, Collado-Torres L, Wang S, Phillips III RA, Karbhari N, Hansen KD, Langmead B, Leek JT (2016) Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the sequence read archive. *Genome Biol* 17: 266
- Nilsen TW, Graveley BR (2010) Expansion of the eukaryotic proteome by alternative splicing. *Nature* 463: 457–463
- Odrionitz F, Kollmar M (2008) Comparative genomic analysis of the arthropod muscle myosin heavy chain genes allows ancestral gene reconstruction and reveals a new type of 'partially' processed pseudogene. *BMC Mol Biol* 9: 21
- Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 40: 1413–1415
- Pillmann H, Hatje K, Odrionitz F, Hammesfahr B, Kollmar M (2011) Predicting mutually exclusive spliced exons based on exon length, splice site and reading frame conservation, and exon sequence homology. *BMC Bioinformatics* 12: 270
- Pohl M, Bortfeldt RH, Grützmann K, Schuster S (2013) Alternative splicing of mutually exclusive exons—a review. *Biosystems* 114: 31–38
- Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26: 841–842
- Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV (2003) Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr Biol* 13: 1512–1517
- Rose PW, Prlić A, Bi C, Bluhm WF, Christie CH, Dutta S, Green RK, Goodsell DS, Westbrook JD, Woo J, Young J, Zardecki C, Berman HM, Bourne PE, Burley SK (2015) The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res* 43: D345–D356
- Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, Dixon JE, Zipursky SL (2000) *Drosophila Dscam* is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* 101: 671–684
- Sharp PA, Burge CB (1997) Classification of introns: U2-type or U12-type. *Cell* 91: 875–879
- Sing T, Sander O, Beerwinkel N, Lengauer T (2005) ROCr: visualizing classifier performance in R. *Bioinformatics* 21: 3940–3941
- Smith CWJ (2005) Alternative splicing—when two's a crowd. *Cell* 123: 1–3
- Sommer B, Keinänen K, Verdoorn TA, Wisden W, Burnashev N, Herb A, Köhler M, Takagi T, Sakmann B, Seeburg PH (1990) Flip and flop: a cell-specific functional switch in glutamate-operated channels of the CNS. *Science* 249: 1580–1585
- Splawski I, Timothy KW, Sharpe LM, Decher N, Kumar P, Bloise R, Napolitano C, Schwartz PJ, Joseph RM, Condouris K, Tager-Flusberg H, Priori SG, Sanguinetti MC, Keating MT (2004) Ca_v1.2 calcium channel dysfunction causes a multisystem disorder including arrhythmia and autism. *Cell* 119: 19–31
- Splawski I, Timothy KW, Decher N, Kumar P, Sachse FB, Beggs AH, Sanguinetti MC, Keating MT (2005) Severe arrhythmia disorder caused by

- cardiac L-type calcium channel mutations. *Proc Natl Acad Sci USA* 102: 8089–8096; discussion 8086–8088
- Stanke M, Waack S (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19: ii215–ii225
- Sterne-Weiler T, Martinez-Nunez RT, Howard JM, Cvitovik I, Katzman S, Tariq MA, Pourmand N, Sanford JR (2013) Frac-seq reveals isoform-specific recruitment to polyribosomes. *Genome Res* 23: 1615–1623
- Suyama M (2013) Mechanistic insights into mutually exclusive splicing in dynamin 1. *Bioinformatics* 29: 2084–2087
- Thierry-Mieg D, Thierry-Mieg J (2006) AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol* 7(Suppl 1): S12.1–S12.14
- Tilgner H, Knowles DG, Johnson R, Davis CA, Chakraborty S, Djebali S, Curado J, Snyder M, Gingeras TR, Guigó R (2012) Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res* 22: 1616–1625
- Tress ML, Abascal F, Valencia A (2017a) Alternative splicing may not be the key to proteome complexity. *Trends Biochem Sci* 42: 98–110
- Tress ML, Abascal F, Valencia A (2017b) Most alternative isoforms are not functionally important. *Trends Biochem Sci* 42: 408–410
- UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res* 43: D204–D212
- Wallace IM, O'Sullivan O, Higgins DG, Notredame C (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res* 34: 1692–1699
- Wang ET, Sandberg R, Luo S, Khrebtkova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* 456: 470–476
- Wang J, Duncan D, Shi Z, Zhang B (2013) WEB-based GEne SeT Analysis toolkit (WebGestalt): update 2013. *Nucleic Acids Res* 41: W77–W83
- Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, Maitland A, Mostafavi S, Montojo J, Shao Q, Wright G, Bader GD, Morris Q (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res* 38: W214–W220
- Weatheritt RJ, Sterne-Weiler T, Blencowe BJ (2016) The ribosome-engaged landscape of alternative splicing. *Nat Struct Mol Biol* 23: 1117–1123
- Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, Hua Y, Gueroussov S, Najafabadi HS, Hughes TR, Morris Q, Barash Y, Krainer AR, Jojic N, Scherer SW, Blencowe BJ, Frey BJ (2015) RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science* 347: 1254806
- Xue Z, Huang K, Cai C, Cai L, Jiang C, Feng Y, Liu Z, Zeng Q, Cheng L, Sun YE, Liu J, Horvath S, Fan G (2013) Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* 500: 593–597
- Yan L, Yang M, Guo H, Yang L, Wu J, Li R, Liu P, Lian Y, Zheng X, Yan J, Huang J, Li M, Wu X, Wen L, Lao K, Li R, Qiao J, Tang F (2013) Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat Struct Mol Biol* 20: 1131–1139
- Yang Y, Sun F, Wang X, Yue Y, Wang W, Zhang W, Zhan L, Tian N, Shi F, Jin Y (2012) Conservation and regulation of alternative splicing by dynamic inter- and intra-intron base pairings in Lepidoptera 14-3-3 ξ pre-mRNAs. *RNA Biol* 9: 691–700
- Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, Girón CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Johnson N, Juettemann T, Keenan S, Lavidas I, Martin FJ et al (2016) Ensembl 2016. *Nucleic Acids Res* 44: D710–D716
- Yura K, Shionyu M, Hagino K, Hijikata A, Hirashima Y, Nakahara T, Eguchi T, Shinoda K, Yamaguchi A, Takahashi K-I, Itoh T, Imanishi T, Gojobori T, Go M (2006) Alternative splicing in human transcriptome: functional and structural influence on proteins. *Gene* 380: 63–71
- Zhang JD, Hatje K, Sturm G, Broger C, Ebeling M, Burtin M, Terzi F, Pomposiello SI, Badi L (2017) Detect tissue heterogeneity in gene expression data with BioQC. *BMC Genom* 18: 277
- Zubović L, Baralle M, Baralle FE (2012) Mutually exclusive splicing regulates the Nav 1.6 sodium channel function through a combinatorial mechanism that involves three distinct splicing regulatory elements and their ligands. *Nucleic Acids Res* 40: 6255–6269



License: This is an open access article under the terms of the Creative Commons Attribution 4.0 License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

3.3.3. Identifying Sequenced Eukaryotic Genomes and Transcriptomes with diArk

Martin Kollmar^{1,*} and Dominic Simm^{1,2}

* Corresponding author

¹ Group Systems Biology of Motor Proteins, Department of NMR-based Structural Biology, Max-Planck-Institute for Biophysical Chemistry, Göttingen, Germany

² Theoretical Computer Science and Algorithmic Methods, Institute of Computer Science, Georg-August-University, Göttingen, Germany

Eukaryotic Genomic Databases. Methods in Molecular Biology (2018) - Volume 1757

doi: 10.1007/978-1-4939-7737-6_1

https://link.springer.com/protocol/10.1007/978-1-4939-7737-6_1

Contribution

MK drafted the manuscript. DS contributed to manuscript text. All authors read and approved the final manuscript.



Chapter 1

Identifying Sequenced Eukaryotic Genomes and Transcriptomes with diArk

Martin Kollmar and Dominic Simm

Abstract

The diArk Eukaryotic Genome Database is a manually curated and updated repository of available eukaryotic genome and transcriptome assemblies. diArk is a key resource for researchers interested in comparative eukaryotic genomics, and the entry point to browsing sequenced eukaryotes in general and to find the most closely related species to the own organism of interest in particular. The exponentially increasing number of sequenced species demands sophisticated search and data presentation tools. In this chapter we describe how to navigate the diArk database keeping a first-time user in mind.

Key words Eukaryotes, Sequenced genomes, Genome assembly, Transcriptome assembly

1 Introduction

The diArk Eukaryotic Genome Database (<http://www.diark.org>) was started in 2005 as a central repository for eukaryotes with genome assembly data available [1]. The species comprised human, the most widely used model organisms such as *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Arabidopsis thaliana*, and multiple fungi. While dedicated databases were available for most of these species connecting the genomes with additional data and analyses, diArk filled the gap as central entry point allowing browsing through the sequenced eukaryotome and providing links to species databases and primary data repositories. Given the progress in sequencing strategy and technology development it was clear that genome assemblies will become available with increasing pace. Several genome sequencing initiatives had been started aiming to obtain genome assemblies of multiple related species. Large scale at that time meant the sequencing of 12 *Drosophila* species [2] or the sequencing of 29 low- and high-coverage mammalian genomes [3]. Soon after, size and speed of the projects increased by two

orders of magnitude. Current large scale projects cover all major branches of eukaryotic life and new genomes are released on a daily basis. The most important ongoing whole-genome sequencing projects are the sequencing of vertebrates within Genome 10K [4], the sequencing of all 10,500 bird species within B10K [5], the sequencing of 5,000 arthropods (i5k) [6], the 959 Nematodes project [7, 8], the 1000 Fungal Genomes Project (1FKG; <http://1000.fungalgenomes.org>), and the sequencing of 1000 yeasts (y1000+; <https://y1000plus.wei.wisc.edu/>). In addition, transcriptomes have been sequenced and assembled from 1000 plants (1KP project) [9], more than 1200 insects (1KITE project; <http://www.1kite.org>), and hundreds of marine microorganisms within MMETSP (Marine Microbial Eukaryote Transcriptome Sequencing Project) [10]. Pilot data have been obtained for two further ambitious projects, the 1000 Plant & Animal Reference Genomes Project and Fish-T1K [11], but it is not clear whether these projects still continue. Most recently, plans to sequence the genomes of at least 10,000 plants have been announced (10KP) [12]. In addition to these massive sequencing efforts, there are dozens of small-scale sequencing projects dedicated to a few species, and hundreds of groups generate genome and transcriptome assemblies of just single species. Because most journals require only the raw sequencing data to be submitted to NCBI/ENA/DDBJ, genome and transcriptome assemblies and annotations are often stored at university servers or digital repositories.

At diArk, we manually track the availability of genome and transcriptome assemblies by browsing publications, genome databases of species and taxa from large-scale efforts and small research communities, and the NCBI genome assembly database [13]. While diArk includes almost all available genome assemblies, manually keeping track with the pace of transcriptome assembly generation is more difficult. In case of the 1KP and MMETSP projects hundreds of transcriptome assemblies were made available at once, and these assemblies have not yet been integrated into diArk. At diArk, we not only provide links to assembly data but also provide the genome and transcriptome data directly [14, 15]. For many species, updated and/or alternative genome assemblies are available. Alternative genome assemblies have been generated by using different software on the same sequencing read data, or by independent sequencing and assembling efforts. diArk provides meta-data for each assembly such as date of generation, version, sequencing method, and assembly software used, and reports simple statistics such as number of entries (contigs, supercontigs, etc.), total sequence length, GC content, and N50. This chapter explains how to use diArk's search and filter tools, and how genome and transcriptome data are displayed.

2 Methods

2.1 *diArk* Website Navigation

diArk's home page offers several entry points to search for sequenced genomes and transcriptomes (Fig. 1). The main access to tools and information is via the dark-grey menu bar near the top of the page which is present on all diArk pages and facilitates quick navigation to the search tools, to extensive statistical analyses of the data, to help pages and other information. In addition to the menu items, the website provides two autocompletion search forms. An autocompletion search box, which can be added to the search plugins of the Firefox browser, is located on the right side of the menu bar and provides quick access to species summary pages (Fig. 1A; *see* Subheading 2.2). Right in the middle of the website is an autocompletion search field by which a search using the Fast Search functionalities is started (Fig. 1B; *see* Subheading 2.3). The home page also includes a quick view on a randomly selected species, a short list of the latest species and their genome/transcriptome assembly data added to diArk, a short list of the most recent publications about species' genome/transcriptome analyses, and a small box to the right with some metrics of diArk's content (Fig. 1C). More statistics can be obtained by clicking the info-icon in the header of this box, and by browsing the statistics subpage, which is accessible from the menu bar.

Because many species (especially fungi and yeasts) are known under multiple, synonymously used scientific names, and because many species were renamed recently because of better classification based on now available whole-genome data, we adhered to a few conventions from the beginning of the diArk project. First, diArk's reference scientific species names are the main species entries as given in the NCBI-Taxonomy database [16]. There is no notification by NCBI-Taxonomy, and thus we only adjust names as soon as we observe changes. Second, as an exception to this convention, we consistently use the fungi's teleomorph names as diArk's scientific reference names. Teleomorph and anamorph are the sexual and asexual states, respectively, of fungi, and both states were often given different scientific names in times of pure morphological classification. Many of these states were later shown to belong to the same holomorph (the whole fungus, including anamorph and teleomorph) but all scientific names remained in use. NCBI-Taxonomy does not have, to our knowledge, a convention on how to assign a main species name to these fungi. In most cases, the name of the most widely analysed state seems to be taken. Accordingly, in case the anamorph is the main experimental research state for a fungus, the corresponding genome/transcriptome assemblies at diArk refer to the less used teleomorph name. Third, different sequenced strains/breeds/cultivars/isolates of the same species (summarized as "strains" from here on) get different

diArk a resource for eukaryotic genome research

Home Search Database Statistics-History Data News Team Funding Help-FAQ Links Contact

homo sapiens

Welcome to diArk 3.0.

A database for eukaryotic genome and EST sequencing projects.

Enter your Search here Search

If you use diArk, please cite:
B. Hammesfahr, F. Odronitz, M. Hellkamp & M. Kollmar (2011) diArk 2.0 provides detailed analyses of the ever increasing eukaryotic genome sequencing data. *BMC Research Notes* 4, 338. **Highly accessed** **Open Access**

Do you know?

Echinostoma paraense

Taxonomy: Eukaryota | Opisthokonta | Metazoa | Eumetazoa | Bilateria | Platyhelminthes | Trematoda | Digenea | Plagiorchiida | Echinostomata | Echinostomatoidea | Echinostomatidae | Echinostoma

Go to NCBI Taxonomy (48215) | Encyclopedia of life | Wikipedia

Project: GenBank - NIH genetic sequence database: GenBank species TBLASTN ?

Publication: Nowak TS, Loker ES (2005) *Echinostoma paraense*: differential gene transcription in the sporocyst stage. *Exp Parasitol* 109, 94-105. 10.1016/j.exppara.2004.11.005

Newest Species and Genome Assemblies in diArk

2017-08-30

Saccharomyces cerevisiae DBVPG6044 v.1.0.0 (chromosome assembly3) | Taxonomy: Opisthokonta | Fungi | Dikarya | ... | Saccharomycetaceae | Saccharomyces Coverage: 315 | Contigs: 15 | GC-Content: 38.3% | Bases: 10.7Mbp | Filesize: 10.34MB

Saccharomyces cerevisiae Y12 v.1.0.0 (chromosome assembly2) | Taxonomy: Opisthokonta | Fungi | Dikarya | ... | Saccharomyces | Saccharomyces cerevisiae Coverage: 145 | Contigs: 14 | GC-Content: 38.2% | Bases: 9.8Mbp | Filesize: 9.45MB

2017-08-29

Blifiguratus adelaidae AZ0501 v.1.0.0 (contigs) | Taxonomy: Opisthokonta | Fungi | Mucoromycota | ... | Mucoromycotina incertae sedis | Blifiguratus Coverage: 28 | Contigs: 560 | GC-Content: 48.3% | Bases: 19.2Mbp | Filesize: 18.85MB

diArk 3.0 Content

Species: 5699
References: 167
Projects: 8087
gen DNA Projects: 4486
EST/cDNA Projects: 849
Publications: 2177

CymbaBase

Scipio eukaryotic gene identification

Systems Biology of Motor Proteins

MAX-PLANCK-GESELLSCHAFT

Recommend
Tweet
+1

Fig. 1 diArk home page. The dark-grey menu bar below the page header can be used to quickly navigate to search tools, statistical representations of sequenced genome/transcriptome data, help pages and more. (A) Use the autocompletion search box to access species summary pages. The menu bar and the search box are visible from all diArk pages. (B) Entering a search text in the autocompletion search form and selecting a species from the list of hits starts the Fast Search. (C) Some metrics on diArk's content are shown in the right box. More numbers will be shown upon clicking the info-icon

entries in diArk to facilitate distinguishing multiple genome/transcriptome assemblies of the same species. Sometimes, this information cannot be revealed and the respective assemblies are combined under a common species entry without strain information. These assemblies (and also multiple assemblies of identical strains) are numbered consecutively.

To facilitate the search for sequenced genomes/transcriptomes in case researchers only know a common name but not the scientific name (or one of the multiple used scientific names) the species

are tagged by “alternative” and synonymously used scientific names and by common names, which are all fully available for the keyword search.

2.2 The Fastest Access to Sequenced Genome and Transcriptome Data

Species summary pages provide the most basic information about available genome and transcriptome assembly data. These pages are accessed by using the autocompletion search box in the menu bar which is present on all diArk pages (Fig. 1A). The query is automatically performed in all species name categories, diArk’s main scientific names, alternative/synonymous scientific names, names of anamorphs, and common names. The search box can be added to the list of search plugins of the Firefox browser by clicking on the Firefox browser icon on the right side of the search box. This allows accessing diArk’s species summary pages from anywhere.


1. To access a species summary page, first open the diArk home page (<http://www.diark.org>) (Fig. 1). Locate the autocompletion search box in the right top corner of the page.
2. To find a sequenced eukaryote of interest, start typing your query into the search box (e.g., type “dict”). The autocompletion function will list the first 10 hits for your query string in alphabetical order and summarize the remaining if more hits were found. Direct matches of the query string to diArk’s main species names are shown in bold. In case the query string matched to a common name, a name of an anamorph or an alternative scientific name, diArk’s main species name is shown in the list (e.g., type “dolphin,” “chicken,” or “duck”).
3. Move the mouse to the name of your eukaryote of interest. For faster orientation, the currently activated species name is highlighted in yellow. Click on the name of the eukaryote of interest to open the species summary page (e.g., click on “Dictyostelium discoideum AX4”).
4. On top of the species summary page is a header with diArk’s main species name in bold and a brief taxonomic overview with the three most ancient and the three most recent taxa (Fig. 2A). The page lists species-related information such as full taxonomy, alternative and common species names, comments about specific strains sequenced, and provides links to NCBI taxonomy and Encyclopedia of Life pages (Fig. 2B). The project-related part of the page lists the type of data (genomic DNA or transcribed DNA) and the sequencing centers where the data have been obtained from (Fig. 2C). Via the links you can directly go to the sequencing centers’ reference pages for the respective species. The projects also list and provide links to community-driven species home pages such as dictyBase, XenBase, or ZFIN. Using these links you can directly access more

A

Dictyostelium discoideum AX4

[Amoebozoa | Mycetozoa | Dictyosteliida | ...
... | Dictyosteliida | Dictyostelium | Dictyostelium discoideum]

B



Picture Source [↗](#)

Taxonomy: Eukaryota | Amoebozoa | Mycetozoa | Dictyosteliida | Dictyostelium | Dictyostelium discoideum

Comment: cellular slime mold

[↻](#) [Go to NCBI Taxonomy \(352472\)](#) [↗](#)

[📖](#) [Encyclopedia of life](#) [↗](#)

[📖](#) [Wikipedia](#) [↗](#)

C

Projects:

- [📁](#) [Dept. Genome Analysis, IMB Jena: Dictyostelium discoideum Genome Project](#) [↗](#)
- [📁](#) [dictyBase: Dictyostelium genome information, curated Dictyostelium literature](#) [↗](#)
- [📁](#) [? Dictyostelium cDNA Project: Dictyostelium cDNA Project](#) [↗](#)
- [📁](#) [e! Ensembl: Dictyostelium discoideum](#) [↗](#)
- [📁](#) [HGSC Human Genome Sequencing Center at Baylor College of Medicine: Functional Genomics of Dictyostelium](#) [↗](#)
- [📁](#) [? International Species Sequencing Consortium: Dictyostelium discoideum Sequencing Consortium](#) [↗](#)
- [📁](#) [National Center for Biotechnology Information: NCBI Protozoa genomes](#) [↗](#)
- [📁](#) [? The Gene Index Project: DFCI Dictyostelium discoideum Gene Index](#) [↗](#)
- [📁](#) [The Wellcome Trust Sanger Institute: The Dictyostelium discoideum Genome Project](#) [↗](#)

Publications:

Eichinger L, Pachebat JA, Glockner G et. al. |> (2005)
[The genome of the social amoeba Dictyostelium discoideum.](#) [↗](#)
Nature **435**, 43-57. [v](#)
[doi: 10.1038/nature03481](#) [↗](#)

Urushihara H, Morio T, Saito T et. al. |> (2004)
[Analyses of cDNAs from growth and slug stages of Dictyostelium discoideum.](#) [↗](#)
Nucleic Acids Res **32**, 1647-53. [v](#)
[doi: 10.1093/nar/gkh262](#) [↗](#)

Glockner G, Eichinger L, Szafranski K et. al. |> (2002)
[Sequence and analysis of chromosome 2 of Dictyostelium discoideum.](#) [↗](#)
Nature **418**, 79-85. [v](#)
[doi: 10.1038/nature00847](#) [↗](#)

[Show all 4 Publications](#) [📄](#)

D

Genome files

[📁](#) National Center for Biotechnology Information

| Type | Version | Date | Compl | Cov | GC % | Size (Mbp) | Contigs | CGR | Illegal Chars | N50 (kbp) | Acc | File | Seq info |
|---------------------|---------|------------|-------------------|-----|------|------------|-----------------------|-----|---------------|-----------|-------------------|-------------------|-------------------|
| Chromosome | v 2.0.0 | 2008-01-22 | 📁 | - | 22.4 | 33.9 | 6 i | | - | 5450 | v | 📄 | 📄 |
| Chromosome | v 2.1.0 | 2009-08-12 | 📁 | 8.3 | 22.4 | 33.9 | 6 i | | - | 5450 | v | 📄 | 📄 |
| Contigs | v 2.0.0 | 2008-01-22 | 📁 | - | 22.4 | 34.2 | 261 i | | - | 341 | v | 📄 | 📄 |
| Contigs | v 2.1.0 | 2009-08-12 | 📁 | 8.3 | 22.4 | 34.1 | 259 i | | - | 341 | v | 📄 | 📄 |
| Mitochondrion | v 2.0.0 | 2008-01-22 | 📁 | - | 27.4 | 0.1 | 1 i | | - | 55 | v | 📄 | 📄 |
| Unplaced chromosome | v 2.0.0 | 2008-01-22 | 📁 | - | 26.1 | 0.3 | 36 i | | - | 11 | v | 📄 | 📄 |
| Unplaced chromosome | v 2.1.0 | 2009-08-12 | 📁 | 8.3 | 27.5 | 0.2 | 33 i | | - | 7 | v | 📄 | 📄 |

[📁](#) e! Ensembl

| Type | Version | Date | Compl | Cov | GC % | Size (Mbp) | Contigs | CGR | Illegal Chars | N50 (kbp) | Acc | File | Seq info |
|-------------------|-----------------------------|------|-------|-----|------|------------|---------|-----|---------------|-----------|-----|------|----------|
| 📄 | Chromosome v 2.1.0 | | | | | | | | | | | | |
| 📄 | Unplaced chromosome v 2.1.0 | | | | | | | | | | | | |

Fig. 2 Species summary page. (A) Header with species name and basic taxonomy. (B) Complete species information including full taxonomy and alternative scientific, common and anamorph names. (C) List of all external resources that provide access to genome/transcriptome assembly data for the particular species.

species-related data and analysis tools. If the genome and/or transcriptome assemblies of the species have already been published, these data are listed in the Publications section.

5. The lower part of the species summary page contains the Genome files section which is sorted by projects (Fig. 2D). For each project the sequence data available is tabulated. The table provides a quick overview about assembly version, assembly release date, genome assembly completeness, sequencing coverage (if known), GC content, assembly size, contig number, presence of illegal characters in the data (not being g/G, a/A, t/T, c/C, or n/N), and the N50 of the dataset. These data allow an easy comparison if multiple versions of the same data, different assembly states (e.g., contigs vs. supercontigs vs. chromosome; Table 1), and alternative assemblies (indicated by “_assembly”) are available. In addition, the Genome files tables contain icons that can be clicked for more detailed information. Click on the info-icon in the contigs column to view N50 (contigs sorted by lengths) and A50 (cumulated assembly length by contig number) plots (Fig. 3A). Clicking the Chaos Game Representation (CGR) fingerprint icon provides the CGR of the assembly and Frequency Chaos Game Representations (FCGRs) of the data in all resolutions up to 1024×1024 (Fig. 3B). Click the arrow-down button in the Acc column to see accession numbers, if available. The chromosome icon in the File column allows access to the assembly data in archived fasta format. Click the Seq info icon to get further information about the sequencing method (e.g., Sanger, Roche/454, Illumina, and PacBio) and the assembly software. The background color of the Seq info icon indicates the sequencing method with, for example, blue being used for Sanger, red for Roche/454, and yellow for Illumina sequencing.
6. In case multiple strains have been sequenced for the same species, only one can be selected from the search box list. The Genome files sections are collapsed if multiple species are shown, but can be opened by clicking *Show genomes*. If you want to get a quick overview about all sequenced strains, use the Taxonomy Search Module (*see* Subheading 2.4).

2.3 The Fast Search

The Fast Search is enabled by using the autocompletion search field in the middle of the home page or by selecting Search Database from the menu bar. The autocompletion function in the search

←

Fig. 2 (continued) Some of these external resources provide extensive additional data and multiple analyses tools. (D) List of assembly files available at diArk sorted by project. In case several projects host exactly the same data, these data are linked. The table contains multiple metrics for comparing different assemblies and evaluating the quality of each assembly. Multiple icons provide further detailed information and plots upon clicking

Table 1
Genome and transcriptome assembly file types

| | |
|---------------|--|
| Chromosome | Every fasta-entry in this file is a chromosome. There are a few exceptions where chromosomes are split in two or more fasta-entries. |
| Uchromosome | These files contain contigs/supercontigs which could not be mapped to any (unknown chr.) or anchored (random chr.) to a certain chromosome. |
| Supercontigs | Every fasta-entry represents a supercontig which consists of sorted contigs separated by estimated or fixed numbers of “N” bases. |
| Usupercontigs | These files contain contigs which could not be mapped to supercontigs. |
| Ultracontigs | Every fasta-entry represents a number of supercontigs assembled to an ultracontig. |
| Contigs | Contigs (from “contiguous sequence”) are the smallest pieces of an assembly and consist of overlapping sequence reads. |
| Ureads | These files contain the unplaced reads, reads that could not be assembled to contigs. These files are especially important for low-coverage genomes that in most cases end up with very short contigs. In these cases, small proteins or some exons can be reconstructed from the ureads-files. |
| Apicoplast | These files contain the apicoplast DNA. The apicoplast is a relict, nonphotosynthetic plastid found in Apicomplexa. It is proposed that it evolved via secondary endosymbiosis. The apicoplast is surrounded by four membranes within the outermost part of the endomembrane system. |
| Chloroplast | These files contain the chloroplast DNA. Chloroplasts are organelles found in plant cells and eukaryotic algae that conduct photosynthesis. |
| Kinetoplast | These files contain the kinetoplast DNA. A Kinetoplast is a disk-shaped mass of circular DNAs inside a large mitochondrion that contains many copies of the mitochondrial genome. Kinetoplasts are only found in protozoa of the class kinetoplastea. Kinetoplasts are usually adjacent to the organisms’ flagellar basal body leading to the thought that they are tightly bound to the cytoskeleton. |
| Plastid | These files contain plastid DNA. Plastids are major double-membrane organelles found in the cells of plants, algae, and some other eukaryotic organisms. |

(continued)

Table 1
(continued)

| | |
|-------------|---|
| Mito | These files contain the mitochondrial DNA. Mitochondria are membrane-enclosed organelles found in most eukaryotic cells. |
| Nucleomorph | These files contain nucleomorph DNA. Nucleomorphs are small, vestigial eukaryotic nuclei found between the inner and outer pairs of membranes in certain plastids. So far, nucleomorphs have only been identified in cryptomonads, which belong to the Chromista supergroup, and in chlorarachniophytes, which are a subphylum of Rhizaria. |
| TSA | Transcriptome shotgun assembly. |
| _assembly | This extension is used to distinguish multiple assemblies of the same strain/breed/cultivar/isolate. Different assemblies can be based on using the same sequencing data but different assembly software and protocols, or are the result of sequencing and assembling the same strain/breed/cultivar/isolate multiple times. |
| _haploid | This extension to an assembly type indicates that the data (e.g., contigs, supercontigs) have been merged to generate an assembly representing a haploid state. |
| _diploid | This extension to an assembly type indicates that both alleles were assembled independently. |

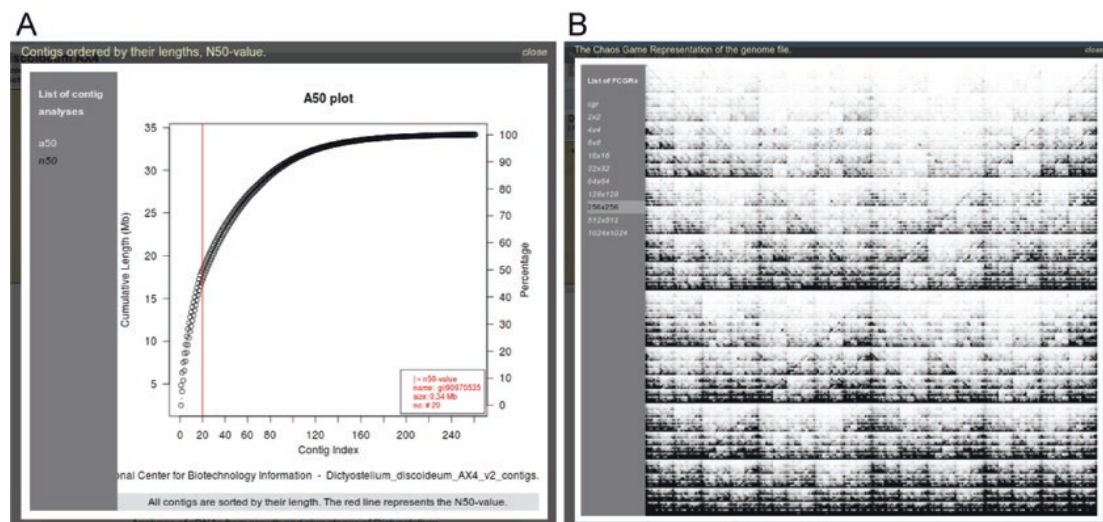


Fig. 3 Genome assembly analyses. (A) By clicking the info-icon in the *Contigs* column of the Genome files section of the species summary page, the user can inspect N50 and A50 plots of the assemblies. (B) A Chaos Game Representation (CGR) is a unique fingerprint for each genome dataset and the number of pixels in the graph is exactly the number of nucleotides in the dataset. Therefore, for comparing CGRs usually Frequency Chaos Game Representations (FCGRs) are taken, in which the pixels in a certain resolution-dependent section of the graph are summed


field works similar to that in the autocompletion search box in the menu bar (*see* Subheading 2.2). However, small pictures of the species, if available, allow easier identification of the species of interest, and additional information is provided for each species in terms of linked project pages and available genome assembly files. In contrast to the species summary pages, the Fast Search allows easy extension or filtering of the search space by selecting/deselecting species, and the Results tabs provide full information about species and assembly file related data.

1. Go to diArk's home page and enter "dict" in the autocompletion search field in the middle of the page.
2. Use the up- and down-keys and press enter, or use the mouse to select "**D**ictyostelium discoideum AX4 | Dd" and start the Fast Search.
3. The Fast Search contains a headline summarizing filter parameters (Fig. 4), which were preselected and combined from the extended Search Modules from the complex search (*see* Subheading 2.4). Below the headline are short sections separating filter for species names, several taxa and model organisms, and a few options to filter by sequencing type, completeness of sequencing, and availability of genome assembly files for download in diArk (Fig. 4A). The small exclamation mark icons next to model organism names indicate, that genome/transcriptome assemblies are available for multiple strains/breeds/cultivars/isolates and that only the most commonly used is selected (e.g., in the case of *Caenorhabditis elegans* the Bristol N2 strains, and for *Saccharomyces cerevisiae* the S288c strain). The Sequencing type filter allows selecting those species for which only EST, genome or RNAseq data or for which multiple data types are available.
4. Move further down the page to inspect the Search Results which are organized in multiple tabs (Fig. 4B). By default the **Species result** view is opened summarizing species related information, links to species sequencing centers and other resources with assembly data, and publications. This result tab is organized species by species. The **Projects result** view is organized by sequencing center allowing a resource-centric view on the data in case multiple species were selected. The **Publications result** view summarizes all publications related to the selected species. The **Genome Stats result** view provides a

←

Fig. 4 (continued) Below the species selection section there are three options to select certain sequencing types (genome, EST, transcriptome, and combinations of these), complete or incomplete genomes, and genome data available for download at diArk. (B) Data for the selected species can be browsed in seven Result tabs. By default, the Species result tab is opened which is identical to the species section on the species summary page (*see* Fig. 2B). Here, the References result tab is opened providing information about available analysis tools and data types at species home pages and repositories

Fast search

 **Species, taxonomy and more** A

Search for species












Enter search words:

Scientific name
 Abbreviation
 Alternative name
 Common name
 Anamorph
 Search in all names

or Select Taxa

| | | | |
|--|--|--|---|
| Kingdom | Phylum | Class | Order |
| <input type="radio"/> Metazoa <input checked="" type="radio"/> Fungi <input type="radio"/> Viridiplantae | <input type="checkbox"/> Apicomplexa <input type="checkbox"/> Ascomycota <input type="checkbox"/> Basidiomycota <input type="checkbox"/> Microsporidia <input type="checkbox"/> Chordata <input type="checkbox"/> Mollusca <input type="checkbox"/> Arthropoda <input type="checkbox"/> Nematoda <input type="checkbox"/> Streptophyta | <input type="checkbox"/> Eurotiomycetes <input type="checkbox"/> Saccharomycetes <input type="checkbox"/> Actinopterygii <input type="checkbox"/> Mammalia <input type="checkbox"/> Aves <input type="checkbox"/> Amphibia <input type="checkbox"/> Ascidiacea <input type="checkbox"/> Insecta | <input type="checkbox"/> Haemosporida <input type="checkbox"/> Dictyosteliida <input type="checkbox"/> Primates <input type="checkbox"/> Rodentia <input type="checkbox"/> Diptera <input type="checkbox"/> Rhabditida |

or Select Model Organism

| | | |
|--|--|--|
| <input type="checkbox"/> <i>Anopheles gambiae</i>  | <input type="checkbox"/> <i>Drosophila melanogaster</i> | <input type="checkbox"/> <i>Oryza sativa</i>  |
| <input type="checkbox"/> <i>Arabidopsis thaliana</i>  | <input type="checkbox"/> <i>Emmericella nidulans</i> | <input type="checkbox"/> <i>Plasmodium falciparum</i>  |
| <input type="checkbox"/> <i>Brachydanio rerio</i>  | <input type="checkbox"/> <i>Encephalitozoon cuniculi</i>  | <input type="checkbox"/> <i>Saccharomyces cerevisiae</i>  |
| <input type="checkbox"/> <i>Caenorhabditis elegans</i>  | <input type="checkbox"/> <i>Gallus gallus</i>  | <input type="checkbox"/> <i>Schizosaccharomyces pombe</i> |
| <input type="checkbox"/> <i>Ciona intestinalis</i> | <input type="checkbox"/> <i>Homo sapiens</i>  | <input type="checkbox"/> <i>Takifugu rubripes</i> |
| <input checked="" type="checkbox"/> <i>Dictyostelium discoideum</i> | <input type="checkbox"/> <i>Mus musculus</i>  | <input type="checkbox"/> <i>Xenopus tropicalis</i> |

Further settings


Sequencing type:

Completed sequencing: ignore yes no

Genome provided by diArk: ignore yes no

4 Publications (0% total) 9 Projects (0% total) 1 Species (0% total)

Search Results

| | | |
|------------------|------------------|--|
| Species (?) | Projects (?) | Publications (?) |
| Genome Stats (?) | Genome Files (?) | References (?)  |
| | | Sequencing Stats (?) |

Dictyostelium discoideum AX4


















| Project | Completion | Release Date | Assembly Version | Genome Map Viewer | cDNA Clone | TBLASTN | BLASTP |
|--|---|--------------|------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|
| DFCI Dictyostelium discoideum Gene Index  |  | 2004-09-16 | 5.0 | | | <input checked="" type="checkbox"/> | |
| The Dictyostelium discoideum Genome Project  |  | | | | | <input checked="" type="checkbox"/> | |
| NCBI Protozoa genomes  |  | 2009-08-12 | | | | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> |
| Dictyostelium discoideum Genome Project  |  | | | | | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> |
| Functional Genomics of Dictyostelium  |  | | | <input checked="" type="checkbox"/> | | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> |
| Dictyostelium genome information, curated Dictyostelium literature  |  | 2005-10-29 | | <input checked="" type="checkbox"/> | | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> |
| Dictyostelium cDNA Project  |  | | | | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | |
| Dictyostelium discoideum Sequencing Consortium |  | | | | | <input checked="" type="checkbox"/> | |
| Dictyostelium discoideum  |  | 2009-08-12 | | <input checked="" type="checkbox"/> | | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> |

Fig. 4 The Fast Search. (A) The Fast Search offers a collection of the most commonly used search options. Species can be searched using an autocompletion search field and are selected by pressing the enter key or the left mouse button. Species and taxa can also be selected from the list of model organisms and taxa.

quick overview about chromosome numbers (if known), genome assembly size, GC content, and contig numbers. This view is especially useful for comparing multiple strains or closely related species. By default, the analysis data for the suspected most complete assembly (size of the largest available assembly) are shown. The **Genome Files result** view provides the same information as the Genome files section of the species summary pages (Fig. 2D). The **References result** view provides further information about the availability of genome map viewers (GBrowse or JBrowse) and BLAST servers (TBLASTN and BLASTP, which also indicates availability of protein annotations), and the possibility to obtain cDNA clones for further research (Fig. 4C). The **Sequencing Stats result** view provides multiple plots, which are obviously only available and useful if multiple species were selected.

5. Move up the page again and reset the Fast Search by clicking on “Fast search” in the menu bar. Select “Mammalia” from the taxa. Automatically, “Primates” and “Rodentia” are also selected, as well as human and mouse from the Model Organisms. Dashes in the selection fields from “Metazoa” and “Chordata” indicate selection of a subset of the respective taxa. At the bottom of the Search section, the number of selected species and related publications/projects are given. Select/deselect “Primates” and/or “Rodentia” to see the interplay between the various selection fields and the influence on the filtered results. In the default Species result view, the selected species are sorted by “Taxonomy.” The order can be changed to sorting by “Name” and the number of visible species adjusted. Browse through the other Result tabs to see how the information is organized in case multiple species are selected. Go to the Genome Stats tab to compare the genome assembly sizes (Fig. 5).
6. Move up again to the Search for species input field, enter “danio” and select the “Brachydanio rerio str. Tuebingen” for comparison.

2.4 Complex Searches by Combining Search Modules

In the extended search, available by Search Database and then Search from the menu bar, the database is searched using modules that can be combined in any order. These modules work as selection baskets, meaning that by default nothing is selected when choosing a search module. Users are usually interested in only a small subset of the data, and this is faster to select from the available options than to deselect the rest. There are five different modules each providing specific options: a module for the full-text search in all species names, a taxonomy search module, a publication search module, a module to search sequencing project related data, and a module to filter by parameters related to genome/transcriptome assembly files. A search operation can consist of any combination of modules and their options. By adding

Search Results

Species (?)
Projects (?)
Publications (?)

Genome Stats (?)
Genome Files (?)
References (?)
Sequencing Stats (?)

Genome Stats per page: 15 25 50 100 1000
 Order by: Taxonomy Name

Jump to page:

1 2 3 4 5 13 14 15 16 17

Genomes

| Species | Chr No | Size (MBp) | GC Content | Contig No |
|--|--------|------------|------------|-----------|
| Mammalia | | | | |
| Ornithorhynchus anatinus duck-billed platypus, duckbill platypus, platypus (German: Schnabeltier) | - | 1842.2 | 45.5 % | 205536 |
| Chrysochloris asiatica Cape golden mole | | 3363.5 | 40.0 % | 20499 |
| Procavia capensis rock dassie, large-toothed rock hyrax, cape hyrax, rock hyrax cape rock hyrax (German: Klippschliefer) | - | 3121.8 | 41.0 % | 31463 |
| Elephantulus edwardii Cape long-eared elephant shrew, Cape elephant shrew (German: Kap-Elefantenspitzmaus) | | 3315.9 | 40.3 % | 8768 |
| Loxodonta africana African savannah elephant, African bush elephant, African savanna elephant, elephant (German: Afrikanischer Elefant) | - | 3118.5 | 40.8 % | 2886 |
| Mammuthus primigenius mammoth, woolly mammoth (German: Mammut) | - | - | - | - |
| Trichechus manatus latirostris | | 2762.5 | 40.7 % | 3145 |
| Echinops telfairi lesser hedgehog tenrec, small Madagascar hedgehog (German: Kleiner Igeltenrec) | - | 2605.2 | 43.0 % | 8401 |
| Orycteropus afer afer aardvark (German: Erdferkel) | | 3415.3 | 40.1 % | 22508 |
| Galeopterus variegatus Malayan flying lemur, Sunda flying lemur, Malayan colugo (German: Malaien-Gleitflieger, Temminck-Gleitflieger) | | 2657.0 | 40.2 % | 54928 |
| Oryctolagus cuniculus domestic rabbit, rabbits, Japanese white rabbit, European rabbit (German: Kaninchen, Wildkaninchen) | - | 2604.0 | 43.7 % | 3690 |
| Ochotona princeps southern American pika, American pika (German: Amerikanischer Pfleifhase) | - | 1944.0 | 43.3 % | 10420 |
| Rodentia | | | | |
| Castor canadensis American beaver, North American beaver (German: Kanadischer Biber, Amerikanischer Biber) | | 2518.0 | 39.7 % | 21157 |
| Fukomys damarensis Damaraland mole rat, Damara mole rat, Damaraland blesmol (German: Damara-Graumull) | | 2286.0 | 40.3 % | 162545 |
| Heterocephalus glaber naked mole-rat (German: Nacktmull) | - | 2430.0 | 40.1 % | 39266 |

Fig. 5 The Genome Stats result view. The Genome Stats result view shows the genome size and GC content for the selected species, here all mammals. To better evaluate these metrics the number of contigs for the underlying dataset is given. In most cases the sizes of the contig datasets are bigger than those of supercontigs and chromosomes because the latter do not contain those contigs that cannot be ordered into supercontigs or mapped to chromosomes

further search modules the user can successively refine the search and narrow down the result list. For each module the resulting set of species, projects and publications is shown below the modules, providing additional context. When a new module is added the options available are restricted by the selection from the previous module(s). At any time, the search options for every module can be changed and modifications are propagated down the chain reapplying previous user actions.

1. Click on Search Database and then Search in the menu bar to move to the extended search by Search Modules (Fig. 6). By default, all data in the database are selected and can be browsed using the Search Result tabs as described above (*see* Subheading 2.3).
2. The Species Names search module is identical to the species autocompletion search field available in the Fast Search (Fig. 4A). Click on the Taxonomy search module picture. The headline contains a search module icon and the search module name on the left, and a number of icons on the right (Fig. 7A).

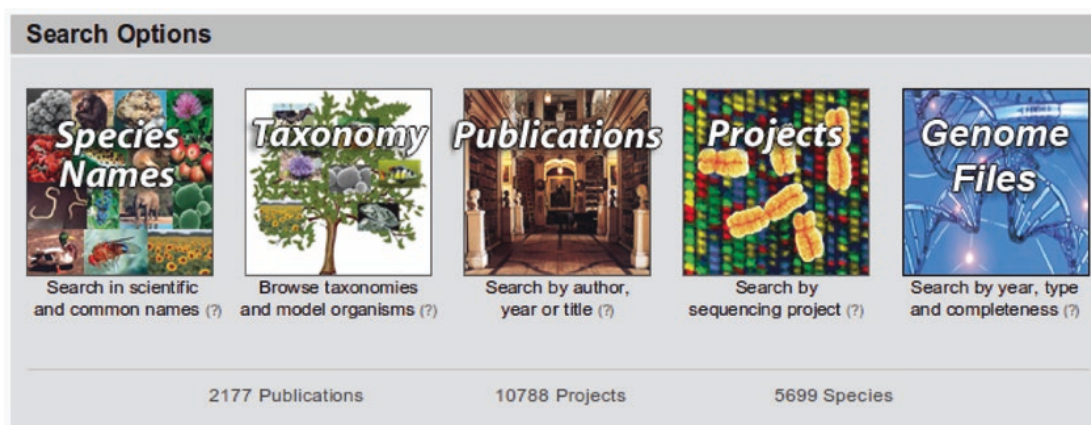


Fig. 6 Search modules for the extended search. The extended search allows any combination of the available five search modules to select specific subsets of the species and to filter for certain assembly characteristics. When opening the search from the menu bar, all species are selected by default. By selecting any search module all previous data (either all or the selection from previous search modules) are blocked. By selecting species/publications/projects from the search modules these data immediately become available in the Result views

Fig. 7 (continued) contracting the module (arrow-up/arrow-down icon). (B) Section of the Taxonomy search module that allows selecting any combination of species from a full taxonomic tree. Here, “*Dictyostelium discoideum* AX4” was searched and selected in the autocompletion field, which opened the corresponding part of the tree up to the searched species. Species connected to internal nodes are automatically shown. The tree can be expanded and contracted using the arrow-down/arrow-up icons, and species and taxa are selected using the check boxes

A

Taxonomy

B

Select from Taxonomy Tree

Search for Taxon:

Search for Species:

- cellular organisms
- Eukaryota
 - Alveolata
 - Amoebozoa
 - Mycetozoa
 - Dictyosteliida
 - Dictyostelium
 - Dictyostelium lacteum TK
 - Dictyostelium firmibasis
 - Dictyostelium citrinum strain OH494
 - Dictyostelium deminutivum
 - Dictyostelium polycephalum
 - Dictyostelium polycarpum
 - Dictyostelium fasciculatum
 - Dictyostelium purpureum
 - Dictyostelium intermedium
 - Dictyostelium discoideum
 - Dictyostelium discoideum AX4
 - Apusozoa
 - Breviatea
 - Centrohelioczoa
 - Cryptophyta
 - Euglenozoa
 - Fornicata
 - Glaucocystophyceae
 - Haptophyceae
 - Heterolobosea
 - Jakobida
 - Malawimonadidae
 - Opisthokonta
 - Oxymonadida
 - Parabasalia
 - Rhizaria
 - Rhodophyta
 - Stramenopiles
 - unclassified eukaryotes
 - Viridiplantae

Fig. 7 The Taxonomy search module. (A) The header of the Taxonomy search module. The icons in the right top corner of the header are available in all search modules and allow removing the module (minus icon), getting explanations for the search options (help icon), activating/deactivating the module (checkbox), and expanding/

By clicking the minus icon, the respective search module and the respective applied search filter can be removed. The question mark icon allows opening a short help as separate window on top of the webinterface. The checkbox allows deselecting/selecting the respective search module from the list of filters while the selected options within the search module remain unchanged. The arrow-up/arrow-down icon to the right allows closing and opening, respectively, of the search module view.

3. You should be familiar with the Taxa and Model Organism selection sections (Fig. 4A; *see* Subheading 2.3). Below these sections, you can browse the entire eukaryotic taxonomic tree to select any combination of species (Fig. 7B). On top of the tree representation, there are two autocompletion input fields, one for taxa and one for species names. As usual, the species name input field accepts all different types of names, the scientific names, the alternatively used names, common names, and the anamorphs, while it selects the scientific reference name. After submission of the taxon/species name search, the taxonomy tree is reloaded with the specific branch(es) containing the selected taxon/species opened. Instead of searching for taxa or species names, taxa and species can be browsed and selected by expanding/collapsing and including/excluding subsections of the tree.
 - Use the checkboxes to select/deselect single species, or all species within a taxon.
 - Clicking the double arrow-down icon will expand the respective subsection of the tree by up to 5 nodes (levels). On mouse-over the icon the number of subtaxa and species will be given.
 - Clicking the arrow-down icon will expand the tree by 1 taxon.
 - Clicking the arrow-up icon will collapse the subtree.
 - On mouse-over a species name an image of the species will be shown.
 - It is important to know that by selecting a taxon you will select the complete subtree and all species included. Thus you will also select all taxa/species that are not shown based on restrictions in previous modules. The rationale is that you might want to change your restrictions in a previous module but still want to keep your taxonomic selection. For example, if you restricted your search to genomic sequences in the Projects module and selected the Arthropoda you will only see Arthropoda species with sequenced genomes. If you then change your selection in the Projects module to also include the cDNA/EST projects, your species selection will automatically expand to include all Arthropoda for which genomic sequences and/or cDNA/EST data are available.

4. Go to the header of the Taxonomy search module and click the minus icon to remove it from the current search, or click the Search menu item to reset the search. From the Search Modules select the Publications search module. The publications related to the species and sequencing projects can be searched in several ways. Full-text searches are provided for titles, authors, abstracts, and journals. In addition, the journal search input field is supplemented with an autocompletion function. Searches can be restricted by dates. By default, searches are unrestricted and include the full time span of publication dates. The option to select “All Publications” is useful as filter for selecting only species with published genome/transcriptome because unpublished genome data are often under embargo and should be used with respect to data sharing policies [17].
5. Remove the Publications search module and select the Projects search module. Here, you can select all species sequenced by a certain sequencing center or available from a specific species home page. The selection of species can be filtered by Sequencing type (*see* Subheading 2.3) and by analysis tools available at species home pages and data repositories, e.g., only those data resources can be selected which provide access to a BLASTP server or which allow ordering cDNA clones. Further down in the search module, there is a section to select one or more specific references. Reference is referred to sequencing centers, data repositories, and species home pages, in diArk. For example, the Joint Genome Institute is a sequencing center (= *reference* in diArk) and hosts many species-dedicated websites (= *projects* in diArk). Selection of a certain option in the table at the top of the module will automatically disable those references that do not provide the selected data and tools. If you are particularly interested in for example filtering for a certain Sequencing type, select the respective type and select “All Projects.” Keep in mind that all search modules work as selection baskets. Thus, whatever filter you apply from the table at the top of the module, you have to “select” one (or all) of the references to include any data.
6. Reset the search, or close the Projects search module, and click on the Genome Files search module picture (Fig. 8). The Genome File search module allows filtering for assembly-related metadata and metrics. By default, all assemblies are included. By restricting the time span of the assembly release dates, for example, only assemblies of the current year can be selected. For several assemblies, release dates could not be revealed, and these assemblies can either be included or excluded in total. The data can further be filtered by assembly completeness, presence or absence of illegal characters, sequencing coverage, GC-content,



Genome Files

Genome released

From to (press enter)

Include genome files with no release date

Select/exclude

Completed sequencing ignore yes no

Illegal characters ignore yes no

Genome provided by diArk ignore yes no

Coverage

From to (press enter)

Include genome files with no coverage data in the database

GC-content

From % to % (press enter)

Select all genome types

All genome types

Select specific genome types

show/hide all

Chloroplast

Chromosome

Contigs

Mitochondrium

Reads

Supercontigs

...

Select sequencing methods

10x Genomics ignore and or

Chromium

Bionano Genomics ignore and or

Fig. 8 The Genome Files search module. The Genome Files search module allows filtering the genome/transcriptome data for release dates, complete/incomplete sequencing, sequencing coverage, and GC content. These parameters can only restrict the space of selectable genome/transcriptome assembly files. Thus, either “all genome types” or at least one of the listed genome types need to be selected to inspect the respective genome/transcriptome assembly data via the Result views

genome type, sequencing method, and assembly software. In the Genome File search module, selection of any data is done via choosing one or more genome types. Thus, whatever filter you apply from the other options, you have to “select” one (or all) of the genome types.

7. Browse through the result tabs as explained above (*see* Subheading 2.3).

References

1. Odronitz F, Hellkamp M, Kollmar M (2007) diArk--a resource for eukaryotic genome research. *BMC Genomics* 8:103. <https://doi.org/10.1186/1471-2164-8-103>
2. Clark AG, Eisen MB, Smith DR et al (2007) Evolution of genes and genomes on the drosophila phylogeny. *Nature* 450:203–218. <https://doi.org/10.1038/nature06341>
3. Lindblad-Toh K, Garber M, Zuk O et al (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478:476–482. <https://doi.org/10.1038/nature10530>
4. Genome 10K Community of Scientists (2009) Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. *J Hered* 100:659–674. <https://doi.org/10.1093/jhered/esp086>
5. Zhang G, Rahbek C, Graves GR et al (2015) Genomics: bird sequencing project takes off. *Nature* 522:34. <https://doi.org/10.1038/522034d>
6. i5K Consortium (2013) The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *J Hered* 104:595–600. <https://doi.org/10.1093/jhered/est050>
7. Kumar S, Schiffer PH, Blaxter M (2012) 959 nematode genomes: a semantic wiki for coordinating sequencing projects. *Nucleic Acids Res* 40:D1295–D1300. <https://doi.org/10.1093/nar/gkr826>
8. Kumar S, Koutsovoulos G, Kaur G, Blaxter M (2012) Toward 959 nematode genomes. *WormBook* 1:42–50. <https://doi.org/10.4161/worm.19046>
9. Matasci N, Hung L-H, Yan Z et al (2014) Data access for the 1,000 plants (1KP) project. *GigaScience* 3:17. <https://doi.org/10.1186/2047-217X-3-17>
10. Keeling PJ, Burki F, Wilcox HM et al (2014) The marine microbial eukaryote Transcriptome sequencing project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol* 12:e1001889. <https://doi.org/10.1371/journal.pbio.1001889>
11. Sun Y, Huang Y, Li X et al (2016) Fish-T1K (Transcriptomes of 1,000 fishes) project: large-scale transcriptome data for fish evolution studies. *GigaScience* 5:18. <https://doi.org/10.1186/s13742-016-0124-7>
12. Normile D (2017) Plant scientists plan massive effort to sequence 10,000 genomes. In: *Sci. AAAS*. <http://www.sciencemag.org/news/2017/07/plant-scientists-plan-massive-effort-sequence-10000-genomes>. Accessed 28 Aug 2017
13. Kitts PA, Church DM, Thibaud-Nissen F et al (2016) Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res* 44:D73–D80. <https://doi.org/10.1093/nar/gkv1226>
14. Hammesfahr B, Odronitz F, Hellkamp M, Kollmar M (2011) diArk 2.0 provides detailed analyses of the ever increasing eukaryotic genome sequencing data. *BMC Res Notes* 4:338. <https://doi.org/10.1186/1756-0500-4-338>
15. Kollmar M, Kollmar L, Hammesfahr B, Simm D (2015) diArk – the database for eukaryotic genome and transcriptome assemblies in 2014. *Nucleic Acids Res* 43:D1107–D1112. <https://doi.org/10.1093/nar/gku990>
16. Federhen S (2012) The NCBI taxonomy database. *Nucleic Acids Res* 40:D136–D143. <https://doi.org/10.1093/nar/gkr1178>
17. Kaye J, Heeney C, Hawkins N et al (2009) Data sharing in genomics — re-shaping scientific practice. *Nat Rev Genet* 10:331–335. <https://doi.org/10.1038/nrg2573>

3.3.4. The Protein-Coding Human Genome: Annotating High-Hanging Fruits

Klas Hatje¹, Stefanie Mühlhausen², Dominic Simm^{2,3} and Martin Kollmar^{2,*}

* Corresponding author

¹ Roche Pharmaceutical Research and Early Development, Pharmaceutical Sciences, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd., Basel, Switzerland

² Group Systems Biology of Motor Proteins, Department of NMR-based Structural Biology, Max-Planck-Institute for Biophysical Chemistry, Göttingen, Germany

³ Theoretical Computer Science and Algorithmic Methods, Institute of Computer Science, Georg-August University, Göttingen, Germany

BioEssays (2019) - Volume 10(19): 959

doi: 10.1002/bies.201900066

<https://doi.org/10.1002/bies.201900066>

Contribution

KH and MK drafted the manuscript. SM and DS contributed to manuscript text. All authors read and approved the final manuscript.

The Protein-Coding Human Genome: Annotating High-Hanging Fruits

Klas Hatje, Stefanie Mühlhausen, Dominic Simm, and Martin Kollmar*

The major transcript variants of human protein-coding genes are annotated to a certain degree of accuracy combining manual curation, transcript data, and proteomics evidence. However, there is considerable disagreement on the annotation of about 2000 genes—they can be protein-coding, noncoding, or pseudogenes—and on the annotation of most of the predicted alternative transcripts. Pure transcriptome mapping approaches seem to be limited in discriminating functional expression from noise. These limitations have partially been overcome by dedicated algorithms to detect alternative spliced micro-exons and wobble splice variants. Recently, knowledge about splice mechanism and protein structure are incorporated into an algorithm to predict neighboring homologous exons, often spliced in a mutually exclusive manner. Predicted exons are evaluated by transcript data, structural compatibility, and evolutionary conservation, revealing hundreds of novel coding exons and splice mechanism re-assignments. The emerging human pan-genome is necessitating distinctive annotations incorporating differences between individuals and between populations.

and is therefore of high impact for biological and medical interpretation of experimental results. Usually, the more complete the gene annotation is, the more accurate are downstream analyses, such as short read mapping for spliced reads^[1] or identification of genetic variations from whole genome sequencing studies.^[2] In contrast, overestimated numbers of annotated exons and transcripts lead to less robust gene abundance and differential gene expression estimates.^[3,4] Gene annotation impacts the choice of genetic regions included in exome sequencing, targeted sequencing approaches, and precise genome editing utilizing CRISPR/Cas9. Genomic experiments are more and more applied in clinical practice and, therefore, the accuracy of these methods is highly relevant for diagnosis and treatment of patients. Accurate annotation of genes also has an enormous influence on drug discovery. Gene therapies

1. Introduction

Accuracy of human genome annotation significantly impacts whole genome, transcriptome, and exome sequencing studies

already allow direct manipulation of the genome, transcripts can be targeted with oligonucleotides, and exon splicing events can be adjusted by small molecules.^[5,6]

Dr. K. Hatje
Roche Pharmaceutical Research and Early Development
Pharmaceutical Sciences
Roche Innovation Center Basel
F. Hoffmann-La Roche Ltd.
Grenzacherstr. 124, 4070 Basel, Switzerland
Dr. S. Mühlhausen, D. Simm, Dr. M. Kollmar
Group Systems Biology of Motor Proteins
Department of NMR-based Structural Biology
Max-Planck-Institute for Biophysical Chemistry
Am Fassberg 11, 37077 Göttingen, Germany
E-mail: mako@nmr.mpibpc.mpg.de
D. Simm
Theoretical Computer Science and Algorithmic Methods
Institute of Computer Science
Georg-August-University Göttingen
Goldschmidtstr. 7, 37077 Göttingen, Germany

In contrast to the centralized worldwide effort to assemble the human reference genome led by the Genome Reference Consortium (GRC), efforts to annotate the human genome are split between several major consortia and innumerable smaller groups. Consequently, after decades of identifying human genes a consistent and standardized catalogue of human protein-coding genes is still missing. In addition, many human multi-exon genes undergo alternative splicing and differences between annotated isoforms are even larger between databases. In this review, we will report on recent studies that performed human gene, transcript, and protein identification, while focusing on efforts to investigate those exonic regions that are difficult to detect.

2. Do Current Approaches Annotate Functional Regions, Junk, or Both?

Annotating a genome comprises all efforts to assign biological functions, mechanistic and structural roles, and observations linked to genomic positions to every nucleotide in the genome. Thus, a set of publications reporting on the final data from the Encyclopedia of DNA Elements project (ENCODE) and claiming that most of the human genome consists of functional regions^[7] initiated a lively, highly controversial and ongoing debate about “function” versus “junk.”^[8–16] In short, ENCODE generated

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/bies.201900066>

© 2019 The Authors. *BioEssays* Published by Wiley Periodicals, Inc. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/bies.201900066

Box 1**What is a gene?**

At the time when the word “gene” was coined, the term was looked at from the phenotype perspective as a distinct region, a “locus,” on a chromosome explaining mechanisms of heredity, development, and physiological function. Later, with the discovery of DNA and the publication of the “Central Dogma” of molecular biology, a gene became a physical entity that is transcribed and finally translated into protein. While the phenotype-based view (also called functionalist approach) did not substantially change over time, the genocentric view and molecular understanding was significantly refined including RNA-genes, “genes in pieces,” and regulatory aspects. When the human genome sequence became finished, the gene still was a genomic region with clear structural boundaries. However, alternative splicing already challenged the genotype-phenotype relationship, because it generates different protein isoforms implying different physiological functions derived from the same gene. Subsequent further discoveries seem to have completely dissolved this relationship. While “gene” is one of the major terms in biology, there is no unifying definition for different purposes and disciplines.^[121] The “instrumental gene” comprises the concept in which a phenotype (a disease, a Mendelian trait or another observable characteristic) is related to a locus that is not necessarily a molecular gene but could be any other functional DNA element. While being clear on the phenotype, this concept remains vague on the genotype. In contrast, the “nominal gene” describes the molecular entity encoded by DNA and transcribed into RNA. This concept constitutes the definition used by the Encyclopedia of DNA Elements project (ENCODE) project: “A gene is a union of genomic sequences encoding a coherent

set of potentially overlapping functional products.”^[19] While this clearly defines a genotype, it is completely oblivious of physiological consequence. The ENCODE definition is more restricted than a previous definition by the Sequence Ontology consortium: “A locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions, and/or other functional sequence regions.”^[122] Although both concepts are highly used and especially the molecular gene is a critical practical tool for communication between bioscientists, both concepts did not keep pace with the data generated and knowledge acquired in the past 15 years. Given the transcription of almost the entire genome, genes hardly have defined boundaries,^[123] exons from different genes can be part of the same transcript,^[124] some microbial genomes contain thousands of scrambled genes that need to be decrypted during development,^[125] the functional status of a gene can be passed down to a daughter cell, and at least mammals and plants can rewrite their DNA based on RNA inherited from past generations.^[126] Accordingly, the traditional reductionist way of thinking is continuously changing toward a more systemic concept, which is summarized by a so-called “relational or systemic gene” or “postgenomics gene.”^[127–129] Advancing current human genome annotation to such a holistic view will be an enormous step forward allowing the readout of DNA/RNA/protein sequences based on biological questions and not predefined computational categories. In such a system, pieces of DNA, for example, could be assembled independent of their (current) status (exon, regulatory, protein-binding, etc.) and their (current) assignment to a dedicated gene.

massive data from high-throughput RNA-sequencing, transcription-factor-binding, chromatin structure, and histone modification experiments and concluded that 80.4% of the genome shows some sort of biochemical activity. Their conclusion that this activity is completely functional was strongly criticized for ignoring the “affirming the consequent”^[11] and for showing only “causal role” but no “selected-effect” functionality of the respective genomic regions. According to the critique, only those regions showing “selected-effect” functionality should be termed functional, and regions having a “causal role” should better be described to have an “effect,” “role,” “consequence,” or “activity.” This and most other points of critique such as the size of detected regions and the missing explanation of the C-value paradox^[17] boiled down to the question what function is and what it is not. It has been argued that this is not a semantic question because the answer would extend deeply into many biological disciplines, such as biochemistry, molecular and evolutionary biology, and genetics.^[10,18] While ENCODE provided a very detailed definition of the term “gene”^[19]—which is another fundamental concept with changing meaning across disciplines and time (**Box 1**)—ENCODE missed to explain their use of the term “function.”^[14,18,20]

Unfortunately, currently available guidelines such as those outlined in the “evolutionary classification of genomic function”^[15] are unsuitable when it comes to such complex problems as human genome annotations. For example, many genes are nonessential, about 75% of yeast genes,^[21,22] and while the exact number is impossible to be determined for humans, a good estimate might be that at least 6% of genes are non-functional in a small part of the Iceland population.^[23] Should the nonessential genes be annotated, therefore, as “indifferent DNA [...] whose main function is being there, but whose exact sequence is not important?”^[15] Should the same loci, that are nonfunctional in the Iceland population, be termed “junk DNA” (the suggested term for pseudogenes) in Icelander genomes, but functional, “literal DNA” in genomes of other populations?

To resolve the complexity of the concept “function” with respect to genome annotation we suggest using the term in combination with the respective annotation layer. For example, according to the latest Ensembl genome annotation, 41.56% of the human genome is made up of protein-coding genes (40.78% according to NCBI). Therefore, at the layer of “genes,” 41.56% of the human genome is functional. However, only 1.15% of the genome represents protein-coding sequence (sum of all constitutive and

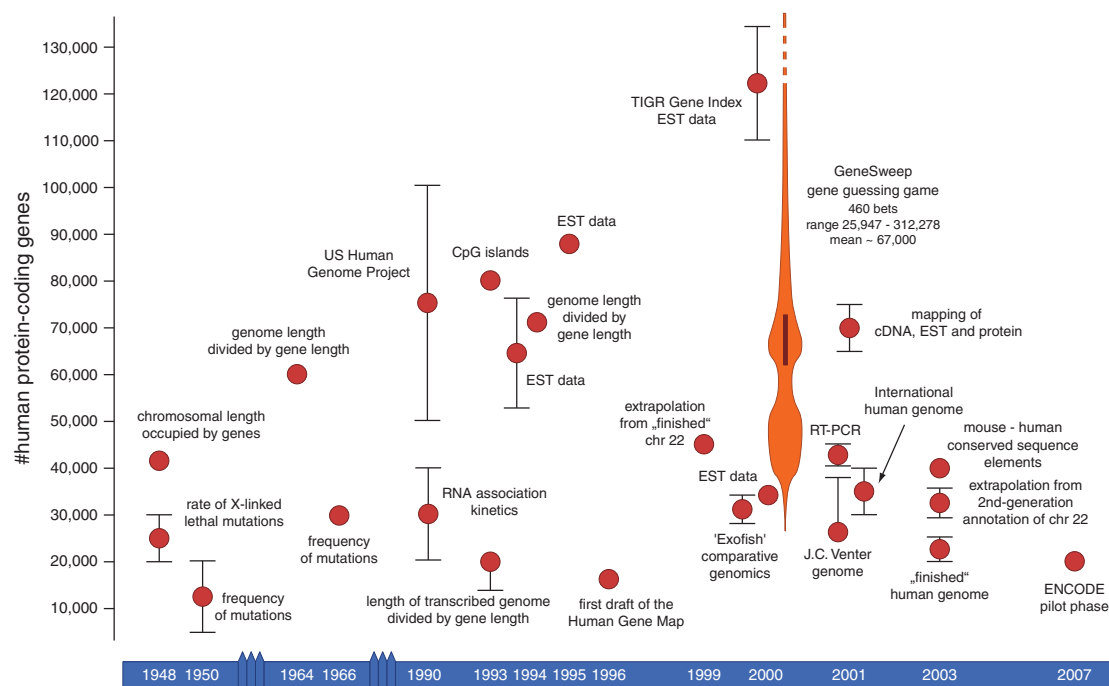


Figure 1. History of the prediction of the number of human genes. Early estimates of the number of human genes in the 1940s to 1960s were based on extrapolations of the few known genetic markers and sequence lengths. In 1990, the U. S. Human Genome Project proposed a range of 50 000 to 100 000 human genes without specifying the data underlying this estimation. Subsequent numbers were derived by various experimental methods and the numbers reported in the publications of the draft genome assemblies were derived from ab initio gene predictions. The latter numbers were subsequently refined by mapping extensive experimental data.

alternative exons; 1.17% according to NCBI) and thus determines the functional genome at the layer of protein-coding nucleotides. Similarly, at the resolution of the assays employed by ENCODE, 80.4% of the human genome might be considered “biochemically functional,” while at the resolution of nucleotides, the number is much lower. In case of the chromatin immunoprecipitation assay, for example, the resolution of the assay with about 600 nucleotides results in 8.5% of the human genome representing transcription factor binding sites, while at the layer of nucleotides (transcription factor binding sites are usually not longer than 3 to 15 nucleotides), in fact only about 0.14% might bind transcription factors.^[11] To be consistent with the “selected-effect” functionality concept, only annotations at the nucleotide level could be termed functional (still, there is the population problem), while annotations at all other layers could be termed “role” or “activity.” This would, however, not be consistent with the use of the term “function” in genetics, cell biology, and most other biological disciplines.^[24] The various functional annotations of the human genome are usually available as separate data tracks in genome browsers and gene reference entries.

3. The Number of Human Protein-Coding Genes Went Up and Down in the Pre-Genome Era

Estimating the number of human genes dates back to the 1940s when the genetic code and even the structure of DNA were un-

known. In 1948, James N. Spuhler estimated the number of genes a) to 42 000 by assuming human genes occupying the same mean chromosomal length than fruit fly genes and b) to 19 890–30 420 by extrapolating the number of loci in the nonhomologous segment of the sex chromosome derived from X-linked lethal mutations (Figure 1 and Box 2).^[25] At about the same time, Hermann J. Muller estimated the frequency of mutations in humans resulting in 5000 to 20 000 human genes,^[26] which is, on the upper limit, very close to the numbers discussed today. Later, Friedrich Vogel for the first time used physical entities estimating the number of genes based on the weight of genes of average length extrapolated to the weight of one human haploid chromosome set.^[27] He calculated two numbers: one number he based on the length of haemoglobin genes resulting in 6.7 Mio genes, which he already dismissed as disturbingly high. Unfortunately, this number was nevertheless presented by Pertea and Salzberg as Vogel’s number of genes,^[28] which likely caused others to later cite this wrong number as well.^[11,29] The number Friedrich Vogel considered more reliable, 60 000 human genes, he derived by dividing the length of the human genome by the gene-length of 50 000 nucleotides inferred from the length of genes in Dipteran giant chromosomes.^[27] In 1966, Muller revised his earlier estimate to “not much more than 30 000” genes.^[30] These early estimates of 20 000–40 000 human genes based on genetic load arguments became the reference in textbooks and publications for the next 25 years, although the respective primary publications did not receive the citations they would have deserved.

Box 2**Early estimates of the number of human protein-coding genes**

At the time when Spuhler and Muller first estimated the number of human genes,^[25,26] the entity “gene” was thought to be a certain part, a “locus”, on the physical chromosome that harbors a trait that is affected by mutations. Muller estimated not only the number of human genes close to today’s values (5000–20 000 genes), but, interestingly, he also estimated the number of *Drosophila* genes to a minimum of 5000–10 000 in the same article, which is also very close to the number of genes known nowadays from genome sequence analysis. The discovery of the genetic code and the amino acid sequences of the α - and β -chains of human hemoglobin allowed Vogel in 1964 to first estimate the number of human genes based on physical entities.^[27] However, the number of genes based on the lengths of these proteins was disturbingly high (6.7 Mio genes). Vogel thus calculated another number based on gene lengths in Diptera, which he regarded more reliable (60 000 human genes). Although introns and repetitive regions were not known at that time, it was already well established that the amount of DNA in closely related species, with likely very similar numbers of genes, can differ by two orders of magnitude,^[130] and that, at least in bacteria, parts of gene material work as regulators of other gene material^[131] providing reasonable explanations for genes being longer than their actual coding sequences. Shortly thereafter, Muller revised his earlier estimate to “not much more than 30 000” genes based on newer data on spontaneous mutations and frequencies of X-ray induced mutations.^[30] Geneticists referred to these numbers, 30 000 genes and an upper limit of 40 000, in the following 25 years albeit without properly citing the original studies. In 1990, the U. S. Human Genome Project proclaimed to sequence the human genome and to locate the suspected 50 000–100 000 human genes without providing any data or reference for this estimate.^[31] Fast progress in

sequencing genomic DNA led to human gene lengths bridging more than three orders of magnitude, and, depending on the methods applied to generating an average, the extrapolated number of human genes ranged from 20 000^[35] to 71 000.^[33] Application of high-throughput techniques provided further numbers, including 20 000–40 000 genes implied by the measurement of RNA re-association kinetics,^[132] 80 000 genes implied by determining and extrapolating CpG island coverage,^[32] and 64 000 genes implied by expressed sequence tag (EST) sequencing followed by clustering and extrapolation.^[33] From today’s perspective, it seems weird that the authors of the CpG island-based estimate (80 000 genes) strongly insisted against reduction of their estimate to 67 000 genes by others.^[33,133] In 1996, a first human gene map was constructed to complement the human genetic map by mapping gene-based sequence tagged site markers resulting in 16 354 distinct loci.^[134] The authors of this study did not present an own estimate of the total number of human genes, as referenced and repeated by others (e.g., [28,29]), but repeated the expectation from the announcement of the Human Genome Project. The publication of the human draft genomes in early 2001 did not stop speculations on higher gene numbers. Extrapolation of RT-PCR data of chromosome 22 predicted 41 000–45 000 genes^[135] and mapping of available cDNA, EST, and protein data combined with gene predictions suggested 65 000–75 000 genes.^[136] Still in 2003, when the “completion” of the human genome sequence was announced, researchers predicted 29 000–36 000 genes based on the extrapolation of a refined annotation of chromosome 22^[137] and up to 40 000 protein-coding genes based on analysis of conserved sequence elements between human and mouse.^[138]

In 1990, the U. S. Human Genome Project proclaimed 50 000–100 000 human genes,^[31] a number without any basis but which subsequently became the cited reference (Figure 1). The disruptive success of sequencing and high-throughput technologies soon superimposed the older genetic load-based estimates and it became more popular to estimate higher numbers. Accordingly, research groups generating large-scale CpG island coverage and expressed sequence tag (EST) data predicted 64 000–87 983 human genes.^[32–34] All these estimates were, however, extrapolations from nonexhaustive data or data from single chromosomes and there were always groups favoring more conservative assumptions. Different assumptions about average gene length and subsequent extrapolation of gene density, for example, resulted in 14 000–20 000^[35] and 71 000 human genes^[33] (Figure 1).

While approaching the release of the first draft of the human genome, researchers from The Institute for Genomic Research (TIGR) predicted 110 000 to 134 000 genes made available in the TIGR Gene Index based on massive EST data^[36] (erroneously, this estimate has been referenced as “57 000” genes in later

reviews^[28,29]). In the same journal issue, other researchers predicted 33 630 and 34 700 genes based on similar EST data^[37] and 28 000–34 000 genes by comparison with pufferfish^[38] (Figure 1). The latter estimates were close to the ranges predicted from the draft genome assemblies, 26 588–38 588 genes^[39] and 30 000–40 000 genes.^[40] Still, in the following months and years, many scientists betted for considerably higher numbers of up to 312 278 genes with a mean of 67 006 in the GeneSweep gene guessing game.^[41,42] Accordingly, the number of predicted genes became cited as “surprisingly low” in the introduction of almost every subsequent paper, despite the many previous studies presenting similar numbers and multiple further arguments that the genome-based numbers were not surprising at all.^[43] Today, the exact number of human protein-coding genes is still unknown and given as 17 694 in neXtProt release 01/2019, 19 033 in Consensus Coding Sequence [CCDS] database release 22 (06/2018), 19 975 in GENCODE release 31 (06/2019), 20 203 in NCBI Homo sapiens Annotation Release 109 (03/2018), and 20 465 in Ensembl release 96 (04/2019).

All efforts differ by the criteria used for gene counting and annotation.

4. Evidence for Protein-Coding Genes Comes from Multiple Sources

A genome annotation is usually performed in two steps termed as structural and functional annotations. Structural annotation comprises the identification of protein-coding genes, RNA-coding genes, regulatory regions, protein-binding sequences, pseudogenes, noncoding RNA, transposons, and other repeats, while functional annotation assigns specific functions to each of the identified regions. In many eukaryotes, especially in humans, another level of complexity is reached through alternative combination of genomic regions (“alternative splicing”) leading to different and overlapping transcripts that encode for proteins with often at least slightly different functions. In essence, a genome annotation does not prove the function of a genomic region, but is a structured collection of predictions and observations that can be used as reference.

There are multiple layers of evidence for genes, transcripts, and exons and support can be inferred from computational and experimental data. In a most basic approach an initial gene annotation is derived from genomic sequence alone using (generalized) hidden Markov models, conditional random fields, or support vector machines.^[44] This approach is based on simple assumptions such as “protein-coding genes should contain start and stop codons” and “protein-coding genes should consist of concatenated exons without internal stop codons.” Coding regions are then identified based on sequence patterns distinctive for exons, introns, and intergenic sequences. These statistical features are enhanced by specific, probabilistic patterns for, e.g., transcript splice sites and polyadenylation sequences.^[45] Pure *ab initio* gene predictions are of little use for the identification of novel human genes and transcripts, because human genome annotation is one of the best genome annotations available. Still, unguided gene predictions have a use in evaluating the predictive power of sequence motifs for gene encoding. Used as such, unguided transcript reconstructions yield valid isoforms for about 41% of expressed genes including both completely and partially overlapping transcripts.^[46]

Considerably better gene annotations are obtained if gene prediction software is supported by transcript data for feature training and gene model evaluation. Extensive evaluation of available gene prediction software and protocols on human gene annotation within RGASP (RNA-seq Genome Annotation Assessment Project) demonstrated that there is, at least currently, an upper limit of about 20% to 40% completely recovered transcripts plus an additional 20% to 30% transcripts recovered missing one exon.^[46] Unfortunately, the fraction of false positive exon and transcript predictions (i.e., the number of predicted exons that are highly likely to be noncoding) has not extensively been evaluated yet. In a small-scale attempt, only 3.2% of a selection of 221 computationally predicted exons could experimentally be validated.^[47]

Information from transcript data alone, from ESTs in earlier times to high-throughput RNA-sequencing (RNA-seq) today (Box 3), is used to both identify and evaluate transcribed regions and to refine annotation models.^[48–50] More sequencing

data also means more transcriptional noise and, therefore, various types of filters (e.g., presence of reads from multiple individuals and different tissues) are applied to collect transcripts of strong evidence only. Although protein-coding genes have, in general, specific nucleotide patterns and transcripts of these genes have polyadenylation tails allowing their physical enrichment for EST/cDNA data generation, there is also a low level of pseudo-gene transcription.^[51,52] For many candidate genes their status as protein-coding gene or non-coding RNA gene is not resolved yet.^[53–55] At this level, proteogenomics provides protein-level evidence combining genomics data and mass spectrometry-based proteomics. In these experiments, gene predictions and translations of transcript data are combined into protein sequence databases, which are used to compute theoretical peptide mass spectra, which in turn are compared to the experimental mass spectra for peptide identification.^[54,56–60] The proteomics approach unfolds its full potential when tissue-specific transcript data are available allowing to prepare tissue-specific sequence databases to avoid mis-assignment of spectra.^[60–63] Complementing direct evidence from transcriptional and translational data, comparative genomics allows finding functional genomic elements that are conserved between species. Genes, exons, and other DNA elements under natural selection are often highly conserved between close relatives and the core set of human genes are conserved in all bilateria.^[64–67]

5. The Genome Is Annotated at Multiple Overlapping Layers

Functional annotations can be derived by *in silico* predictions from sequence alone or through comparison with knowledge bases. The latter means transferring functional assignments such as a protein domain or a tRNA isoacceptor type from homologous regions of one protein/RNA to another, which is essentially a prediction by homology. Functional annotations are done at multiple layers. Single nucleotides varying in a population are annotated as single nucleotide polymorphisms (SNPs) and receive a functional annotation through association with phenotypic characteristics or dysfunction as found in genetic diseases. If nucleotides happen to be part of a continuous stretch they might get a common function or role assignment as part of a larger region (global assignment; regions may overlap) and at the same time a functional assignment as single nucleotide (a local assignment). For example, the adenine nucleotide involved in lariat formation during eukaryotic transcript splicing is part of the branchpoint (local assignment), which is an essential part of a spliceosomal intron (first level global assignment). This, in turn, is part of a eukaryotic multi-exon gene (second level global assignment), which again might contribute to a phenotype-associated locus (third level global assignment). However, in the human genome, the vast majority of nucleotides do not have a local function but only a role as part of larger genomic regions. This is the case, for example, for most of the nucleotides within the extended inter- and intragenic regions. Their global role is, among others, being a spacer (or “ballast”)^[68] between genic and regulatory regions or having impact on chromosome structure, while the actual number and specific order of those nucleotides do not matter. Naturally, the density of nucleotides with local function

Box 3**Technological improvements impacting human genome annotation**

The first generation of sequencing technologies utilizing Sanger's method^[139,140] provided expressed sequence tags (ESTs), which are short cDNA snippets representing regions of expressed mRNA. EST data expanded the ab initio gene predictions with genome-wide transcriptional information.^[141–143] ESTs were not only used to identify expressed regions in the genome, but also to determine exact splice junctions that define exons and provide information on how these are concatenated to transcripts.^[144] Similarly, short reads from second generation sequencers yielded deep sampling information across the whole genome and are used to evaluate gene annotation models. The major technology advancements were introduced by the 454, SOLiD, and Illumina sequencing platforms^[140,145] leading to a massive increase in RNA sequence information. These data had high impact on

the accuracy of genome annotations, although there seems to be a limit upon which accuracy and precision cannot be further improved with current methods.^[46] The main challenge in genome annotation thus shifted from identifying expressed regions with high sensitivity to annotating functional genes, alternative transcripts, and exons with high precision. In contrast to the EST/cDNA data, the short read data provides evidence for short exons and exon junctions only. Here, high-throughput long read sequencing allows to sequence transcripts in full-length and to validate combinations of alternative exons. Two technologies are most established for this purpose, Pacific Biosciences^[146,147] and Nanopore^[148–150] Isoform expression can nowadays also be analyzed at single-cell level^[151] using and combining short reads and long-read sequencing methods.^[149,152]

highly correlates with the size of a genome, meaning genomes having less and shorter inter- and intragenic regions show higher density.

6. Is There Consensus on the Low-Hanging Fruits?

Two major databases, NCBI (RefSeq) and Ensembl (GENCODE), dominated human gene annotations since release of the initial human genome assembly. While both efforts (and any other initiative) agreed on using the assembly provided by the GRC as reference, every annotation effort had and continues to have its own rules and definitions and these rules and definitions even change over time.^[69,70] Even for the term “gene” no definition has agreed on (Box 1) so that not only numbers of named categories highly disagree but also categories and subcategories themselves. For this reason, the CCDS project started in 2005 to generate a reliable set of consensus sequences as reference for the scientific community. In its first release, this set contained 12 950 protein-coding genes (Figure 2). “Consensus” was defined as “protein-coding regions that agree at the start codon, stop codon, and splice junctions and for which the prediction meets quality assurance benchmarks.”^[71] The consensus increased rapidly to 18 407 protein-coding genes in 2011 (CCDS version 8) and reached 19 033 genes in the current version (v. 22). Although both the RefSeq and the GENCODE annotations heavily rely on manual gene annotation efforts, there is still a gap of 1200 and 950 genes comparing CCDS to RefSeq and GENCODE, respectively (Figure 2).

In 2011, the Human Proteome Project (HPP) started to map the entire human proteome in a systematic effort.^[72] The two main efforts within HPP, neXtProt, and PeptideAtlas, try to match proteomics data with available gene and transcript datasets. They differ mainly in neXtProt including not only evidence from mass spectrometry data but also from Edman sequencing, biochemical studies, posttranslational modifications, protein–protein interactions, antibody-based techniques, 3D structures, and disease mutations.^[73] The current releases

of neXtProt and PeptideAtlas claim to contain unambiguous evidence for 17 694 and 15 798 proteins, respectively. Although both approaches steadily close the gap to the number of human genes known from transcript data, gene prediction, and evolutionary conservation, they are technically limited in detecting low-abundance proteins, sequences lacking proteolytic cleavage sites and proteins expressed in tissues unavailable for studies.

In summary, major annotation databases currently reach a consensus on about 94% of all protein-coding genes and agree that about 86% of these genes are in fact translated and present in at least one human tissue. These numbers are also in accordance with The Human Protein Atlas project, which combines antibody-based imaging, mass spectrometry-based proteomics, transcriptomics, and systems biology. Currently, these data support 18 899 protein-coding genes in human.^[61] Given this wealth of evidence, claims that “one in five human genes still have unresolved coding status”^[55] and might rather be noncoding genes or pseudogenes than protein-coding genes seem rather exaggerated.

Apart from consortia-guided gene annotation efforts, there are a few small-scale efforts that have provided impressive results. In one project, ab initio gene prediction was used to generate a set of 8 million candidate transcripts. Subsequent filtering and validation by 26 RNA-seq data sets and shotgun proteomics revealed 36 novel proteins and more than 31 000 new transcripts.^[74] For generating the Intropolis resource, this approach was taken even further.^[48] There, 21 504 RNA-seq samples from the SRA archive were aligned against the human genome. About 57 000 new exon junctions have been identified that are present in at least 1000 samples.^[48] In another approach, unstranded and stranded RNA-seq data were not directly mapped to the human genome but first assembled into transcript assemblies and subsequently improved by predicting the orientation of unstranded reads and by integrating information about transcription start, cleavage, and polyadenylation sites.^[75] With this pipeline, called CAFE, the BIGTranscriptome dataset was generated adding thousands of potential transcripts

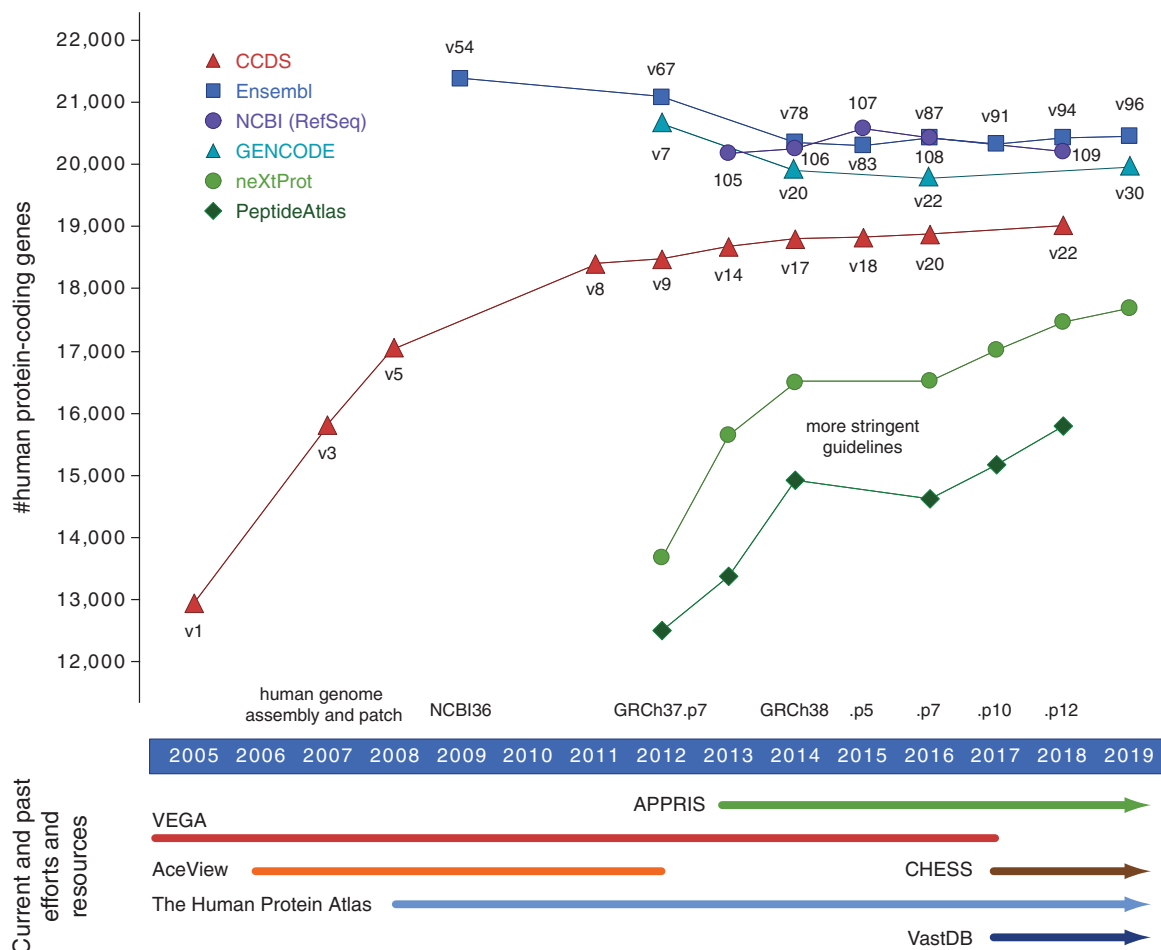


Figure 2. Efforts to annotate the human protein-coding genes. The scheme compiles the number of protein-coding genes as given by the respective data releases of the largest human genome annotation efforts. Numbers are different because definitions of terms and categories vary from effort to effort and even from time to time within efforts. New releases not only contain novel protein-coding genes but also drop genes. Thus, more genes from one release to the next are not directly related to the addition of the respective number of genes and likewise fewer genes do not represent the number of genes removed. The Consensus Coding Sequence project (CCDS) is an effort to mark all protein-coding genes with identical genomic coordinates in both the Ensembl and the NCBI annotations and contains 19 033 genes in the latest release. As example for a state-of-the-art analysis integrating transcriptomics and MS data, a recent tissue-specific expression study by Wang et al. detected 18 072 transcripts and 13 640 proteins.^[60] Abbreviations of genome annotation efforts and resources: VEGA (The Vertebrate Genome Annotation), CHES (Comprehensive Human Expressed Sequences), and APPRIS (annotation of principal and alternative splice isoforms).

to GENCODE and RefSeq annotations (Figure 3). Similarly, in the most recent effort termed CHES (Comprehensive Human Expressed Sequences), 9795 RNA-seq data sets generated by the genotype-tissue expression (GTEx) consortium were assembled resulting in 224 novel protein-coding genes and more than 116 000 novel transcripts (protein-coding and noncoding; Figure 3).^[50] Many of the novel protein-coding genes were shown to be conserved in other mammals adding evolutionary evidence. Remarkably, 30 million of the assembled transcripts were annotated as nonfunctional and transcriptional noise.^[50]

In summary, the number of protein-coding genes with support at protein, transcript, and homology level seems to stabilize around 20 000.^[73,76] In addition, there are about 500 candidate genes whose existence or classification is unclear. While the

uncertainty about the number of genes strongly decreased in the past 20 years (compare Figure 1 and Figure 2), the number of generated transcripts remains largely unclear (Figure 3B). It steadily increased in RefSeq, GENCODE, and other databases over the past ten years, from about 60 000 in 2009 to 210 000 in 2018, but it is not clear yet, to which extent these transcripts result from erroneous splicing^[50,77–80] or are translated to a significant level, if at all.^[62,81–85]

7. Finding All Protein-Coding Segments Remains Difficult

Alternative processing of primary RNA transcripts has been found across all eukaryotes and is a characteristic ranging back

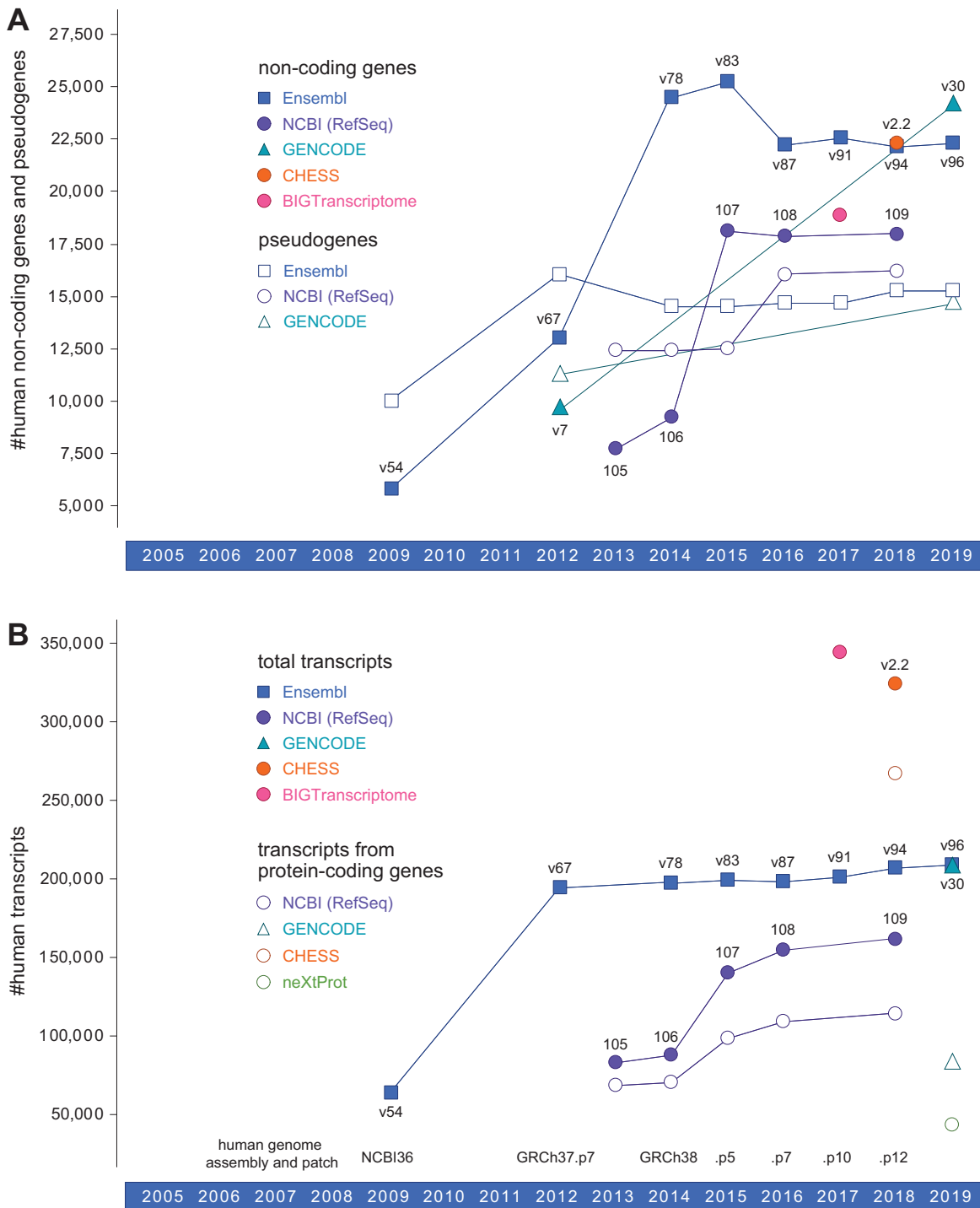


Figure 3. Annotation of human noncoding genes, pseudogenes, and transcripts. A) This scheme compiles the number of noncoding genes and pseudogenes as given by respective databases. B) The number of human transcripts, both the total number of transcripts and the fraction generated from protein-coding genes, considerably vary over time and across databases.

to the last eukaryotic common ancestor. It is used to increase proteome diversity and has been shown to be highly regulated in many species. There are many different types of alternative splicing such as differential inclusion of exons, intron retention, or alternative 5'- and 3'-splicing of exons. A particularly interesting case is mutually exclusive splicing, in which neighboring exons are spliced in a mutually exclusive manner into the mature transcript.

7.1. Micro-Exons

Micro-exons are very short coding exons and therefore difficult to detect. The maximum length of micro-exons differs between publications, further complicating systematic comparison.^[86] Volfovsky et al. defined micro-exons by a maximum length of 25 nucleotides,^[87] Wen et al. looked for short alternative splicing events of maximal 51 nucleotides,^[88] Irimia et al. characterized micro-exons with lengths of 3–27 nucleotides^[89] and Li et al. referred to a length between six and 51 nucleotides.^[90] The latter, most exhaustive approach detected unknown micro-exons by mapping RNA-seq data from diverse datasets to the human reference transcriptome from Ensembl.^[90] Mapped reads were filtered for insertions, which were defined as short, additional RNA stretches with a maximum length of 51 nucleotides. When occurring at the exon boundaries of a transcript, such insertions were considered candidate micro-exons. Subsequently, corresponding introns were scanned for the canonical splice site pattern GT-AG. Candidate micro-exons were approved if their sequences exactly matched the intronic sequences between AG and GT.^[90] With this approach, 310 predicted novel micro-exons were added to the 12 835 Ensembl-annotated micro-exons.^[90]

Micro-exons are considered to preserve the reading frame and, if alternatively spliced, to modulate protein structure. Interestingly, micro-exons are highly enriched in transcripts in brain compared to other tissues.^[91] About 2500 neural-regulated micro-exons have been identified in each human and mouse^[89] and their splicing was shown to be mis-regulated in autism.^[89,92,93] Many of the very short micro-exons are well conserved from fish to human. Of the about 150 mammalian, neural-regulated micro-exons with lengths of 3–15 nucleotides at least 55 are deeply conserved in vertebrate species spanning 400–450 million years of evolution.^[89] Micro-exon splicing is suggested to be promoted by a specialized domain in an ancestral splicing factor that originated in a common bilaterian ancestor.^[94]

7.2. Wobble Splicing

Another mechanism introducing small variations to protein isoforms is wobble splicing. Here, a GYN repeat at the donor splice site (5' splice site; Y stands for C or T and N stands for A, C, G, or T) or an NAG repeat at the acceptor splice site (3' splice site) leads to subtle length variations in the spliced transcripts and finally to alternative isoforms differing in few amino acids. The most frequent wobble splicing element is the NAGNAG tandem repeat, which was systematically identified in the human genome using EST data.^[95–97] In a first attempt to characterize NAGNAG splic-

ing, the RefSeq annotation was searched, and 7326 candidate acceptor splice sites were discovered, corresponding to NAGNAG splicing in 30% of all RefSeq transcripts.^[95] Subsequent EST data mapping confirmed 878 sites. In addition, NAGNAG acceptors show high conservation^[98] with 73% being conserved between human and mouse.^[95] Interestingly, NAGNAG repeats mainly alternate the protein sequence by one amino acid, but rarely introduce premature stop codons. Later, the same approach was used to identify and verify GYNGYN tandem repeats at donor splice sites in the human genome.^[99]

In another approach to identify wobble splice sites the SwissProt database was searched for pairs of protein isoforms with single amino acid differences. The candidate list was then extended by reports in the literature about subtle protein differences.^[96] Identified alternative donor splice sites were verified through presence in the human genome and through comparison with cDNA available from NCBI and EBI's AltSplice database. RNA-seq data were also used to investigate the regulation of NAGNAG splicing.^[100] Here, sequences flanking the NAGNAG acceptor splice sites identified in the human genome were extracted and mapped with RNA-seq reads requiring at least six nucleotides overhang on each side. Evidence for tissue regulated splicing was found for 73% of NAGNAG acceptor splicing events.^[100,101] Still, some events might better be explained by stochastic splicing alone.^[102–105] This stochasticity has been described as a physiologically triggered, concerted shift in alternative splicing.^[106] A specialized tool to detect and quantify NAGNAG splicing events identifies NAGNAG motifs in splice sites and counts RNA-seq reads mapping to those sites.^[107]

7.3. Mutually Exclusive Exons

Mutually exclusive splicing means that exactly one exon of a cluster of neighboring exons is spliced into the mature transcript. Although mutually exclusive exons (MXEs) of a cluster are relatively similar, they cannot substitute each other if one is damaged. MXEs have been described in many crucial and essential human genes such as in the α -subunits of six of the ten voltage-gated sodium channels (*SCN* genes), in each of the glutamate receptor subunits 1–4 (*GluR1-4*) in which the MXEs are called flip and flop and in *SNAP-25* as part of the neuroexocytosis machinery. Mutations in MXEs have been shown to cause diseases such as Timothy syndrome (missense mutation in the *CACNA1C* gene), cardiomyopathy (defect of the mitochondrial phosphate carrier *SLC25A3*), or cancer (mutations in, e.g., the pyruvate kinase *PKM* and the zinc transporter *SLC39A14*).

To explore the extent of mutually exclusive splicing in humans, we recently predicted 1722 completely novel exons in previously intronic regions in the human genome.^[108] Our prediction algorithm is based on criteria derived from biological knowledge and has successfully been applied to plants, worm, and fruit fly before.^[109,110] MXEs must be translated in same reading frames and splice sites must be compatible. We expect MXEs to have about the same length, because they code for the same structural region in the resulting protein, and length differences should only be possible in loop regions. Finally, the protein sequences coded by MXEs are supposed to be similar, because they code for the same region in the protein and evolved most

probably through exon duplication. Together with already annotated exons matching the aforementioned criteria, the mentioned 1722 newly predicted MXEs became part of a list of 6541 MXE candidates. By mapping 15 billion RNA-seq reads, representing 515 samples comprising 31 tissues and organs, 12 cell lines, and seven developmental stages, we could show that each novel exon is covered by at least one read. Applying strict criteria requiring reads bridging the respective other MXE and absence of reads joining MXEs, 1399 of the 6541 MXE candidates comprise high-confidence MXEs. This number is about tenfold higher than previous MXE estimates in human (ENCODE, e.g., reported only 14 MXEs in human and other analyses showed a maximum of 147). Mapping high-confidence MXEs onto known protein structures revealed further support for their annotation. The ends of MXE-encoded sequences preferentially match within secondary structural elements excluding that these exons can be spliced as constitutive or differentially included exons.

In order to assess the conservation and evolution of human MXEs across mammals, we identified orthologous proteins in 18 representative species from all major sub-branches spanning 180 million years and predicted MXEs therein.^[108] Of the 554 human MXE clusters, 100 (18%) and 86 (15%) are shared between at least 15 and 16 of the 18 species, respectively. Conserved clusters include the annotated MXEs of sodium-channel *SCN* genes, *MAPK8* and *MAPK14*, glutamate-receptors *GRIA* genes, and *KC-NMA1*. Dozens of genes contained novel exons such as collagen genes, members of the *SLIT* family of secreted glycoproteins, calcium-channel *CACNA1E*, and many genes of the solute-carrier family. MXEs are of high relevance in human diseases as an overlay of the set of high-confidence MXEs with the ClinVar database showed (35 of the MXEs contained 82 pathogenic SNPs). Disease-associated MXEs show tight developmental and tissue-specific expression with prominent selective expression in heart and brain and in cancer cell lines.

8. To Each Individual Its Own Annotation

Every individual has its own genome including parts conserved in families, differences at population level, and a collection of sequences distinguishing us from related hominins. The first two human genome assemblies represented examples of two extremes: a private genome sequencing effort reported the genome of an individual, John Craig Venter,^[39] while the international human genome project generated a reference genome representing a mix of cell-lines and various donors.^[40] The latest reference genome, GRCh38, provides representations for alternative loci that are alternative sequences found in largely haploid assemblies (earlier also termed “alternative alleles” and “alternative haplotypes”).^[111,112] However, this reference genome and corresponding reference genome annotation represent a rather artificial consensus of the “human genome.” Genomic drift causes random duplication and deletion of genes, which means that every individual genome contains a significant amount of copy number variations (CNV). The first analysis of genomic drift in humans identified CNVs of sensory receptor genes among 270 individuals from the HapMap data demonstrating a difference of at least eleven olfactory receptor genes between randomly chosen

individuals.^[113] At the level of nations, sequencing and de novo assembly of 150 genomes from Denmark showed large differences at the chromosome scale.^[114] Recently, the first pan-genome of a human population, the population of 910 humans of African descent, revealed a collection of sequences totaling about 300 million nucleotides that are not present in the human reference genome assembly but shared among multiple individuals of the African population.^[115] Similarly, building and analyzing the pan-genome of Han Chinese detected 29.5 million novel nucleotides and at least 188 novel protein-coding genes.^[116]

Human reference genome assembly and annotation are undoubtedly invaluable tools with respect to comparison and evaluation of annotations, tools, approaches, and conclusions. However, the available sequencing tool set should now allow combining analyses of genome, transcriptome, and proteome data at the individual level. Such an approach will also shed light on the reported discrepancies of alternatively spliced isoforms found in transcript and proteomics studies. We also anticipate that some of the loci annotated as pseudogenes in the reference annotation will become protein-coding genes in other annotations. Annotation of “alternate sequences” placed on alternative assembly units in RefSeq and Ensembl is a huge step forward (e.g., Ensembl release 96 contains 2960 coding genes on alternative sequences), but still the annotations are mixed and do not represent different populations, or even individuals.

9. Conclusions and Outlook: Are We Approaching Completeness of Human Genome Annotation?

Although the human genome assembly was declared to be completed in 2003, it has seen substantial changes since then. The current reference assembly has representations for alternative regions, from allelic differences and copy number variations and still there are gaps that need to be closed. These gaps (some of which have recently been closed) have a considerable influence on the annotation. For example, a gap of about 56 000 nucleotides within the dynein heavy chain gene *DHC7C* covering about 860 of the 3800 amino acids was closed only in the latest genome version, GRCh38, although contigs spanning the entire gene region were already present and in correct order in the J. C. Venter assembly.^[117] Accordingly, the old gene annotations contained separate genes for *DHC7C* N- and C-terminus. These are still present as alternative transcripts in latest genome annotations, although they are artificial and will never result in folded proteins. A big step forward would be to not only search for novel transcripts but also to clean up old annotations. Therefore, it is highly likely that similar to finishing the genome assembly finishing reference genome annotation will also last for many years to come.

In addition to data provided by large consortia the scientific community highly profits from complementary efforts by, for example, CHES in determining a set of consensus transcript sequences or by the Human Proteome Project and the Human Protein Atlas, which provide strong evidence for tissue-specific presence of proteins. Integrating many layers of evidence from ab initio predictions, RNA-seq data, splicing mechanism, protein structure, and evolutionary conservation is promising in case experimental data generation is limited by, e.g., protein abundance

or due to highly tissue- and developmental stage-specific expression. Such a multi-facet analysis revealed a detailed landscape of the mutually exclusive exome including not only the identification of hundreds of novel exons but also a re-evaluation of the splice type of hundreds of already known exons^[108] and could be used as blueprint for the comprehensive analysis of other splice variants.

It has been pointed out already that the “human reference genome” generated from mostly a single individual is as bad a reference for human genomes as a genome of every other individual would be.^[118] We suppose that the same will become true for the “human reference annotation”, which will be substituted by annotations for individuals, groups, and populations. Genome annotations have many layers and differ from individual to individual and in many cases also from cell to cell. Mapping transcriptional, gene regulatory, and genetic variant data on a reference genome alone considerably limits the use and application of these data.^[118] Deep learning approaches, which currently still incorporate steady genome annotations,^[119,120] might resolve the issues of such individual annotations in the future.

Conflict of Interest

The authors declare no conflict of interest. KH is employed by F. Hoffmann-La Roche Ltd.

Keywords

alternative splicing, human genome annotation, human pan-genome, micro-exon, mutually exclusive exons, protein-coding genes, wobble splicing

Received: April 15, 2019

Revised: August 7, 2019

Published online:

- [1] S. Zhao, *PLoS One* **2014**, *9*, e101374.
- [2] G. Chen, C. Wang, L. Shi, X. Qu, J. Chen, J. Yang, C. Shi, L. Chen, P. Zhou, B. Ning, W. Tong, T. Shi, *RNA* **2013**, *19*, 479.
- [3] P.-Y. Wu, J. H. Phan, M. D. Wang, *BMC Bioinf.* **2013**, *14*, S8.
- [4] S. Zhao, B. Zhang, *BMC Genomics* **2015**, *16*, 97.
- [5] N. A. Naryshkin, M. Weetall, A. Dakka, J. Narasimhan, X. Zhao, Z. Feng, K. K. Y. Ling, G. M. Karp, H. Qi, M. G. Woll, G. Chen, N. Zhang, V. Gabbeta, P. Vazirani, A. Bhattacharyya, B. Furia, N. Risher, J. Sheedy, R. Kong, J. Ma, A. Turpoff, C.-S. Lee, X. Zhang, Y.-C. Moon, P. Trifillis, E. M. Welch, J. M. Colacino, J. Babiak, N. G. Almstead, S. W. Peltz, et al., *Science* **2014**, *345*, 688.
- [6] M. Sivaramakrishnan, K. D. McCarthy, S. Campagne, S. Huber, S. Meier, A. Augustin, T. Heckel, H. Meistermann, M. N. Hug, P. Birrer, A. Moursy, S. Khawaja, R. Schmucki, N. Berntenis, N. Giroud, S. Golling, M. Tzouros, B. Banfai, G. Duran-Pacheco, J. Lamerz, Y. Hsiu Liu, T. Luebbbers, H. Ratni, M. Ebeling, A. Cléry, S. Paushkin, A. R. Krainer, F. H.-T. Allain, F. Metzger, *Nat. Commun.* **2017**, *8*, 1476.
- [7] ENCODE Project Consortium, *Nature* **2012**, *489*, 57.
- [8] S. R. Eddy, *Curr. Biol.* **2012**, *22*, R898.
- [9] W. F. Doolittle, *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 5294.
- [10] S. R. Eddy, *Curr. Biol.* **2013**, *23*, R259.
- [11] D. Graur, Y. Zheng, N. Price, R. B. R. Azevedo, R. A. Zufall, E. Elhaik, *Genome Biol. Evol.* **2013**, *5*, 578.
- [12] D.-K. Niu, L. Jiang, *Biochem. Biophys. Res. Commun.* **2013**, *430*, 1340.
- [13] W. F. Doolittle, T. D. P. Brunet, S. Linquist, T. R. Gregory, *Genome Biol. Evol.* **2014**, *6*, 1234.
- [14] M. Kellis, B. Wold, M. P. Snyder, B. E. Bernstein, A. Kundaje, G. K. Marinov, L. D. Ward, E. Birney, G. E. Crawford, J. Dekker, I. Dunham, L. L. Elnitski, P. J. Farnham, E. A. Feingold, M. Gerstein, M. C. Giddings, D. M. Gilbert, T. R. Gingeras, E. D. Green, R. Guigo, T. Hubbard, J. Kent, J. D. Lieb, R. M. Myers, M. J. Pazin, B. Ren, J. A. Stamatoyannopoulos, Z. Weng, K. P. White, R. C. Hardison, *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 6131.
- [15] D. Graur, Y. Zheng, R. B. R. Azevedo, *Genome Biol. Evol.* **2015**, *7*, 642.
- [16] W. F. Doolittle, T. D. P. Brunet, *BMC Biol.* **2017**, *15*, 116.
- [17] C. A. Thomas, *Annu. Rev. Genet.* **1971**, *5*, 237.
- [18] T. D. P. Brunet, W. F. Doolittle, *Proc. Natl. Acad. Sci. USA* **2014**, *111*, E3365.
- [19] M. B. Gerstein, C. Bruce, J. S. Rozowsky, D. Zheng, J. Du, J. O. Korbel, O. Emanuelsson, Z. D. Zhang, S. Weissman, M. Snyder, *Genome Res.* **2007**, *17*, 669.
- [20] P.-L. Germain, E. Ratti, F. Boem, *Biol. Philos.* **2014**, *29*, 807.
- [21] G. Giaever, A. M. Chu, L. Ni, C. Connelly, L. Riles, S. Véronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. André, A. P. Arkin, A. Astromoff, M. El-Bakkoury, R. Bangham, R. Benito, S. Brachat, S. Campanaro, M. Curtiss, K. Davis, A. Deutschbauer, K.-D. Entian, P. Flaherty, F. Foury, D. J. Garfinkel, M. Gerstein, D. Gotte, U. Güldener, J. H. Hegemann, S. Hempel, Z. Herman, et al., *Nature* **2002**, *418*, 387.
- [22] D.-U. Kim, J. Hayles, D. Kim, V. Wood, H.-O. Park, M. Won, H.-S. Yoo, T. Duhig, M. Nam, G. Palmer, S. Han, L. Jeffery, S.-T. Baek, H. Lee, Y. S. Shim, M. Lee, L. Kim, K.-S. Heo, E. J. Noh, A.-R. Lee, Y.-J. Jang, K.-S. Chung, S.-J. Choi, J.-Y. Park, Y. Park, H. M. Kim, S.-K. Park, H.-J. Park, E.-J. Kang, H. B. Kim, et al., *Nat. Biotechnol.* **2010**, *28*, 617.
- [23] P. Sulem, H. Helgason, A. Oddson, H. Stefansson, S. A. Gudjonsson, F. Zink, E. Hjartarson, G. T. Sigurdsson, A. Jonasdottir, A. Jonasdottir, A. Sigurdsson, O. T. Magnusson, A. Kong, A. Helgason, H. Holm, U. Thorsteinsdottir, G. Masson, D. F. Gudbjartsson, K. Stefansson, *Nat. Genet.* **2015**, *47*, 448.
- [24] A. G. Wouters, *Stud. Hist. Philos. Sci. Part C: Stud. Hist. Philos. Biol. Biomed. Sci.* **2003**, *34*, 633.
- [25] J. N. Spuhler, *Science* **1948**, *108*, 279.
- [26] H. J. Muller, *Am. J. Hum. Genet.* **1950**, *2*, 111.
- [27] F. Vogel, *Nature* **1964**, *201*, 847.
- [28] M. Pertea, S. L. Salzberg, *Genome Biol.* **2010**, *11*, 206.
- [29] C. Willyard, *Nature* **2018**, *558*, 354.
- [30] H. J. Muller, *Am. Nat.* **1966**, *100*, 493.
- [31] *Understanding Our Genetic Inheritance: The U.S. Human Genome Project: The First Five Years: Fiscal Years 1991–1995*, U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, National Center for Human Genome Research; U.S. Department of Energy, Office of Energy Research, Office of Health and Environmental Research, Human Genome Program.
- [32] F. Antequera, A. Bird, *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 11995.
- [33] C. Fields, M. D. Adams, O. White, J. C. Venter, *Nat. Genet.* **1994**, *7*, 345.
- [34] M. D. Adams, A. R. Kerlavage, R. D. Fleischmann, R. A. Fuldner, C. J. Bult, N. H. Lee, E. F. Kirkness, K. G. Weinstock, J. D. Gocayne, O. White, *Nature* **1995**, *377*, 3.
- [35] R. P. Wagner, M. P. Maguire, R. L. Stallings, *Chromosomes: A Synthesis*, Wiley, Hoboken, NJ **1993**.
- [36] F. Liang, I. Holt, G. Pertea, S. Karamycheva, S. L. Salzberg, J. Quackenbush, *Nat. Genet.* **2000**, *25*, 239.

- [37] B. Ewing, P. Green, *Nat. Genet.* **2000**, *25*, 232.
- [38] H. Roest Crolius, O. Jaillon, A. Bernot, C. Dasilva, L. Bouneau, C. Fischer, C. Fizames, P. Wincker, P. Brottier, F. Quétier, W. Saurin, J. Weissenbach, *Nat. Genet.* **2000**, *25*, 235.
- [39] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, et al., *Science* **2001**, *291*, 1304.
- [40] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczy, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, *Nature* **2001**, *409*, 860.
- [41] Gambling on the genome. *J. Natl. Cancer Inst.* **2000**, *92*, 1373.
- [42] P. Smaglik, *Nature* **2000**, *405*, 264.
- [43] J. M. Claverie, *Science* **2001**, *291*, 1255.
- [44] R. D. Sleator, *Gene* **2010**, *461*, 1.
- [45] Y. Huang, S.-Y. Chen, F. Deng, *Comput. Struct. Biotechnol. J* **2016**, *14*, 298.
- [46] T. Steijger, J. F. Abril, P. G. Engström, F. Kokocinski, The RGASP Consortium, T. J. Hubbard, R. Guigó, J. Harrow, P. Bertone, *Nat. Methods* **2013**, *10*, 1177.
- [47] R. Guigó, P. Flicek, J. F. Abril, A. Reymond, J. Lagarde, F. Denoeud, S. Antonarakis, M. Ashburner, V. B. Bajic, E. Birney, R. Castelo, E. Eyras, C. Ucla, T. R. Gingeras, J. Harrow, T. Hubbard, S. E. Lewis, M. G. Reese, *Genome Biol.* **2006**, *7*, S2.
- [48] A. Nellore, A. E. Jaffe, J.-P. Fortin, J. Alquicira-Hernández, L. Collado-Torres, S. Wang, R. A. Phillips Iii, N. Karbhari, K. D. Hansen, B. Langmead, J. T. Leek, *Genome Biol.* **2016**, *17*, 266.
- [49] C. Wilks, P. Gaddipati, A. Nellore, B. Langmead, *Bioinformatics* **2018**, *34*, 114.
- [50] M. Pertea, A. Shumate, G. Pertea, A. Varabyou, F. P. Breitwieser, Y.-C. Chang, A. K. Madugundu, A. Pandey, S. L. Salzberg, *Genome Biol.* **2018**, *19*, 208.
- [51] S. Kalyana-Sundaram, C. Kumar-Sinha, S. Shankar, D. R. Robinson, Y.-M. Wu, X. Cao, I. A. Asangani, V. Kothari, J. R. Prensner, R. J. Lonigro, M. K. Iyer, T. Barrette, A. Shanmugam, S. M. Dhanasekaran, N. Palanisamy, A. M. Chinnaiyan, *Cell* **2012**, *149*, 1622.
- [52] X. Guo, M. Lin, S. Rockowitz, H. M. Lachman, D. Zheng, *PLoS One* **2014**, *9*, e93972.
- [53] M. Clamp, B. Fry, M. Kamal, X. Xie, J. Cuff, M. F. Lin, M. Kellis, K. Lindblad-Toh, E. S. Lander, *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 19428.
- [54] I. Ezkurdia, D. Juan, J. M. Rodriguez, A. Frankish, M. Diekhans, J. Harrow, J. Vazquez, A. Valencia, M. L. Tress, *Hum. Mol. Genet.* **2014**, *23*, 5866.
- [55] F. Abascal, D. Juan, I. Jungreis, L. Martinez, M. Rigau, J. M. Rodriguez, J. Vazquez, M. L. Tress, *Nucleic Acids Res.* **2018**, *46*, 7070.
- [56] A. I. Nesvizhskii, *Nat. Methods* **2014**, *11*, 1114.
- [57] J. C. Wright, J. Mudge, H. Weisser, M. P. Barzine, J. M. Gonzalez, A. Brazma, J. S. Choudhary, J. Harrow, *Nat. Commun.* **2016**, *7*, 11778.
- [58] K. V. Ruggles, K. Krug, X. Wang, K. R. Clauser, J. Wang, S. H. Payne, D. Fenyö, B. Zhang, D. R. Mani, *Mol. Cell. Proteomics* **2017**, *16*, 959.
- [59] B. Eraslan, D. Wang, M. Gusic, H. Prokisch, B. M. Hallström, M. Uhlen, A. Asplund, F. Pontén, T. Wieland, T. Hopf, H. Hahne, B. Kuster, J. Gagneur, *Mol. Syst. Biol.* **2019**, *15*, e8513.
- [60] D. Wang, B. Eraslan, T. Wieland, B. Hallström, T. Hopf, D. P. Zolg, J. Zecha, A. Asplund, L.-H. Li, C. Meng, M. Frejno, T. Schmidt, K. Schnatbaum, M. Wilhelm, F. Ponten, M. Uhlen, J. Gagneur, H. Hahne, B. Kuster, *Mol. Syst. Biol.* **2019**, *15*, e8503.
- [61] M. Uhlen, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A.-K. Szgyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P.-H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, et al., *Science* **2015**, *347*, 1260419.
- [62] M. L. Tress, F. Abascal, A. Valencia, *Trends Biochem. Sci.* **2017**, *42*, 98.
- [63] Y. Zhu, L. M. Orre, H. J. Johansson, M. Huss, J. Boekel, M. Vesterlund, A. Fernandez-Woodbridge, R. M. M. Branca, J. Lehtiö, *Nat. Commun.* **2018**, *9*, 903.
- [64] K. Lindblad-Toh, M. Garber, O. Zuk, M. F. Lin, B. J. Parker, S. Washietl, P. Kheradpour, J. Ernst, G. Jordan, E. Mauceli, L. D. Ward, C. B. Lowe, A. K. Holloway, M. Clamp, S. Gnerre, J. Alföldi, K. Beal, J. Chang, H. Clawson, J. Cuff, F. Di Palma, S. Fitzgerald, P. Flicek, M. Guttman, M. J. Hubisz, D. B. Jaffe, I. Jungreis, W. J. Kent, D. Kostka, M. Lara, et al., *Nature* **2011**, *478*, 476.
- [65] M. E. Dolan, R. M. Baldarelli, S. M. Bello, L. Ni, M. S. McAndrews, C. J. Bult, J. A. Kadin, J. E. Richardson, M. Ringwald, J. T. Eppig, J. A. Blake, *Mamm. Genome* **2015**, *26*, 305.
- [66] S. König, L. W. Romoth, L. Gerischer, M. Stanke, *Bioinformatics* **2016**, *32*, 3388.
- [67] J. Armstrong, I. T. Fiddes, M. Diekhans, B. Paten, *Annu. Rev. Anim. Biosci.* **2019**, *7*, 41.
- [68] M. Freeling, J. Xu, M. Woodhouse, D. Lisch, *Mol. Plant* **2015**, *8*, 899.
- [69] A. Frankish, B. Uszczyńska, G. R. S. Ritchie, J. M. Gonzalez, D. Perouchine, R. Petryszak, J. M. Mudge, N. Fonseca, A. Brazma, R. Guigo, J. Harrow, *BMC Genomics* **2015**, *16*, S2.
- [70] T. Weirick, D. John, S. Uchida, *Brief. Bioinform.* **2017**, *18*, 226.
- [71] K. D. Pruitt, J. Harrow, R. A. Harte, C. Wallin, M. Diekhans, D. R. Maglott, S. Searle, C. M. Farrell, J. E. Loveland, B. J. Ruff, E. Hart, M.-M. Suner, M. J. Landrum, B. Aken, S. Ayling, R. Baertsch, J. Fernandez-Banet, J. L. Cherry, V. Curwen, M. Dicuccio, M. Kellis, J. Lee, M. F. Lin, M. Schuster, A. Shkeda, C. Amid, G. Brown, O. Dukhanina, A. Frankish, J. Hart, et al., *Genome Res.* **2009**, *19*, 1316.
- [72] P. Legrain, R. Aebbersold, A. Archakov, A. Bairoch, K. Bala, L. Beretta, J. Bergeron, C. H. Borchers, G. L. Corthals, C. E. Costello, E. W. Deutsch, B. Domon, W. Hancock, F. He, D. Hochstrasser, G. Markovarga, G. H. Salekdeh, S. Sechi, M. Snyder, S. Srivastava, M. Uhlen, C. H. Wu, T. Yamamoto, Y.-K. Paik, G. S. Omenn, *Mol. Cell. Proteomics* **2011**, *10*, M111.009993.
- [73] G. S. Omenn, L. Lane, C. M. Overall, F. J. Corrales, J. M. Schwenk, Y.-K. Paik, J. E. Van Eyk, S. Liu, M. Snyder, M. S. Baker, E. W. Deutsch, *J. Proteome Res.* **2018**, *17*, 4031.
- [74] Z. Hu, H. S. Scott, G. Qin, G. Zheng, X. Chu, L. Xie, D. L. Adelson, B. E. Oftedal, P. Venugopal, M. Babic, C. N. Hahn, B. Zhang, X. Wang, N. Li, C. Wei, *Sci. Rep.* **2015**, *5*, 10940.
- [75] B.-H. You, S.-H. Yoon, J.-W. Nam, *Genome Res.* **2017**, *27*, 1050.
- [76] A. Frankish, M. Diekhans, A.-M. Ferreira, R. Johnson, I. Jungreis, J. Loveland, J. M. Mudge, C. Sisu, J. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, S. Carbonell Sala, J. Christ, F. Cunningham, T. Di Domenico, S. Donaldson, I. T. Fiddes, C. Garcia Girón, J. M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, T. Hunt, O. G. Izuogu, J. Lagarde, F. J. Martin, L. Martínez, S. Mohanan, et al., *Nucleic Acids Res.* **2019**, *47*, D766.
- [77] S. Light, A. Elofsson, *Curr. Opin. Struct. Biol.* **2013**, *23*, 451.
- [78] B. Saudemont, A. Popa, J. L. Parmley, V. Rocher, C. Blugeon, A. Necsulea, E. Meyer, L. Duret, *Genome Biol.* **2017**, *18*, 208.
- [79] Y. Wan, D. R. Larson, *Genome Biol.* **2018**, *19*, 86.
- [80] W. F. Doolittle, *Genome Biol.* **2018**, *19*, 223.

- [81] F. Abascal, I. Ezkurdia, J. Rodriguez-Rivas, J. M. Rodriguez, A. del Pozo, J. Vázquez, A. Valencia, M. L. Tress, *PLoS Comput. Biol.* **2015**, *11*, e1004325.
- [82] M. González-Porta, A. Frankish, J. Rung, J. Harrow, A. Brazma, *Genome Biol.* **2013**, *14*, R70.
- [83] M. L. Tress, F. Abascal, A. Valencia, *Trends Biochem. Sci.* **2017**, *42*, 408.
- [84] B. J. Blencowe, *Trends Biochem. Sci.* **2017**, *42*, 407.
- [85] S. A. Bhuiyan, S. Ly, M. Phan, B. Huntington, E. Hogan, C. C. Liu, J. Liu, P. Pavlidis, *BMC Genomics* **2018**, *19*, 637.
- [86] D. Ustianenko, S. M. Weyn-Vanhentenyck, C. Zhang, *Wiley Interdiscip. Rev.: RNA* **2017**, *8*, e1418.
- [87] N. Volfovsky, B. J. Haas, S. L. Salzberg, *Genome Res.* **2003**, *13*, 1216.
- [88] F. Wen, F. Li, H. Xia, X. Lu, X. Zhang, Y. Li, *Trends Genet.* **2004**, *20*, 232.
- [89] M. Irimia, R. J. Weatheritt, J. D. Ellis, N. N. Parikhshak, T. Gonatopoulos-Pournatzis, M. Babor, M. Quesnel-Vallières, J. Tapial, B. Raj, D. O'Hanlon, M. Barrios-Rodiles, M. J. E. Sternberg, S. P. Cordes, F. P. Roth, J. L. Wrana, D. H. Geschwind, B. J. Blencowe, *Cell* **2014**, *159*, 1511.
- [90] Y. I. Li, L. Sanchez-Pulido, W. Haerty, C. P. Ponting, *Genome Res.* **2015**, *25*, 1.
- [91] M. Melé, P. G. Ferreira, F. Reverter, D. S. DeLuca, J. Monlong, M. Sammeth, T. R. Young, J. M. Goldmann, D. D. Pervouchine, T. J. Sullivan, R. Johnson, A. V. Segrè, S. Djebali, A. Niar-chou, GTEx Consortium, F. A. Wright, T. Lappalainen, M. Calvo, G. Getz, E. T. Dermitzakis, K. G. Ardlie, R. Guigó, *Science* **2015**, *348*, 660.
- [92] T. Gonatopoulos-Pournatzis, M. Wu, U. Braunschweig, J. Roth, H. Han, A. J. Best, B. Raj, M. Aregger, D. O'Hanlon, J. D. Ellis, J. A. Calarco, J. Moffat, A.-C. Gingras, B. J. Blencowe, *Mol. Cell* **2018**, *72*, 510.
- [93] A. Parras, H. Anta, M. Santos-Galindo, V. Swarup, A. Elorza, J. L. Nieto-González, S. Picó, I. H. Hernández, J. I. Díaz-Hernández, E. Belloc, A. Rodolosse, N. N. Parikhshak, O. Peñagarikano, R. Fernández-Chacón, M. Irimia, P. Navarro, D. H. Geschwind, R. Méndez, J. J. Lucas, *Nature* **2018**, *560*, 441.
- [94] A. Torres-Méndez, S. Bonnal, Y. Marquez, J. Roth, M. Iglesias, J. Permanyer, I. Almudí, D. O'Hanlon, T. Guitart, M. Soller, A.-C. Gingras, F. Gebauer, F. Rentzsch, B. J. Blencowe, J. Valcárcel, M. Irimia, *Nat. Ecol. Evol.* **2019**, *3*, 691.
- [95] M. Hiller, K. Huse, K. Szafranski, N. Jahn, J. Hampe, S. Schreiber, R. Backofen, M. Platzer, *Nat. Genet.* **2004**, *36*, 1255.
- [96] K. Tadokoro, M. Yamazaki-Inoue, M. Tachibana, M. Fujishiro, K. Nagao, M. Toyoda, M. Ozaki, M. Ono, N. Miki, T. Miyashita, M. Yamada, *J. Hum. Genet.* **2005**, *50*, 382.
- [97] C.-H. Lai, L.-Y. Hu, W. Lin, *Biochem. Biophys. Res. Commun.* **2006**, *342*, 197.
- [98] M. Akerman, Y. Mandel-Gutfreund, *Nucleic Acids Res.* **2006**, *34*, 23.
- [99] M. Hiller, K. Huse, K. Szafranski, P. Rosenstiel, S. Schreiber, R. Backofen, M. Platzer, *Genome Biol.* **2006**, *7*, R65.
- [100] R. K. Bradley, J. Merkin, N. J. Lambert, C. B. Burge, *PLoS Biol.* **2012**, *10*, e1001229.
- [101] A. Busch, K. J. Hertel, *Genome Biol.* **2012**, *13*, 143.
- [102] T.-M. Chern, E. van Nimwegen, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki, M. Zavolan, *PLoS Genet.* **2006**, *2*, e45.
- [103] M. Hiller, K. Szafranski, R. Backofen, M. Platzer, *PLoS Genet.* **2006**, *2*, e207; author reply e208.
- [104] K.-W. Tsai, W.-C. Lin, *Genomics* **2006**, *88*, 855.
- [105] R. Sinha, S. Nikolajewa, K. Szafranski, M. Hiller, N. Jahn, K. Huse, M. Platzer, R. Backofen, *Nucleic Acids Res.* **2009**, *37*, 3569.
- [106] K. Szafranski, M. Kramer, *RNA Biol.* **2015**, *12*, 115.
- [107] X. Yan, G. Sablok, G. Feng, J. Ma, H. Zhao, X. Sun, *FEBS Lett.* **2015**, *589*, 1766.
- [108] K. Hatje, R.-U. Rahman, R. O. Vidal, D. Simm, B. Hammesfahr, V. Bansal, A. Rajput, M. E. Mickael, T. Sun, S. Bonn, M. Kollmar, *Mol. Syst. Biol.* **2017**, *13*, 959.
- [109] H. Pillmann, K. Hatje, F. Odrionitz, B. Hammesfahr, M. Kollmar, *BMC Bioinf.* **2011**, *12*, 270.
- [110] K. Hatje, M. Kollmar, *Nat. Commun.* **2013**, *4*, 2460.
- [111] V. A. Schneider, T. Graves-Lindsay, K. Howe, N. Bouk, H.-C. Chen, P. A. Kitts, T. D. Murphy, K. D. Pruitt, F. Thibaud-Nissen, D. Albracht, R. S. Fulton, M. Kremitzki, V. Magrini, C. Markovic, S. McGrath, K. M. Steinberg, K. Auger, W. Chow, J. Collins, G. Harden, T. Hubbard, S. Pelan, J. T. Simpson, G. Threadgold, J. Torrance, J. M. Wood, L. Clarke, S. Koren, M. Boitano, P. Peluso, et al., *Genome Res.* **2017**, *27*, 849.
- [112] M. Jain, S. Koren, K. H. Miga, J. Quick, A. C. Rand, T. A. Sasani, J. R. Tyson, A. D. Beggs, A. T. Dilthey, I. T. Fiddes, S. Malla, H. Marriott, T. Nieto, J. O'Grady, H. E. Olsen, B. S. Pedersen, A. Rhie, H. Richardson, A. R. Quinlan, T. P. Snutch, L. Tee, B. Paten, A. M. Phillippy, J. T. Simpson, N. J. Loman, M. Loose, *Nat. Biotechnol.* **2018**, *36*, 338.
- [113] M. Nozawa, Y. Kawahara, M. Nei, *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 20421.
- [114] L. Maretty, J. M. Jensen, B. Petersen, J. A. Sibbesen, S. Liu, P. Villesen, L. Skov, K. Belling, C. Theil Have, J. M. G. Izarzugaza, M. Grosjean, J. Bork-Jensen, J. Grove, T. D. Als, S. Huang, Y. Chang, R. Xu, W. Ye, J. Rao, X. Guo, J. Sun, H. Cao, C. Ye, J. van Beusekom, T. Espeseth, E. Flindt, R. M. Friborg, A. E. Halager, S. Le Hellard, C. M. Hultman, et al., *Nature* **2017**, *548*, 87.
- [115] R. M. Sherman, J. Forman, V. Antonescu, D. Puiu, M. Daya, N. Rafaels, M. P. Boorgula, S. Chavan, C. Vergara, V. E. Ortega, A. M. Levin, C. Eng, M. Yazdanbakhsh, J. G. Wilson, J. Marrugo, L. A. Lange, L. K. Williams, H. Watson, L. B. Ware, C. O. Olopade, O. Olopade, R. R. Oliveira, C. Ober, D. L. Nicolae, D. A. Meyers, A. Mayorga, J. Knight-Madden, T. Hartert, N. N. Hansel, M. G. Foreman, et al., *Nat. Genet.* **2019**, *51*, 30.
- [116] Z. Duan, Y. Qiao, J. Lu, H. Lu, W. Zhang, F. Yan, C. Sun, Z. Hu, Z. Zhang, G. Li, H. Chen, Z. Xiang, Z. Zhu, H. Zhao, Y. Yu, C. Wei, *Genome Biol.* **2019**, *20*, 149.
- [117] M. Kollmar, *Mol. Biol. Evol.* **2016**, *33*, 3249.
- [118] X. Yang, W.-P. Lee, K. Ye, C. Lee, *Genome Biol.* **2019**, *20*, 104.
- [119] Z. Zhang, Z. Pan, Y. Ying, Z. Xie, S. Adhikari, J. Phillips, R. P. Carstens, D. L. Black, Y. Wu, Y. Xing, *Nat. Methods* **2019**, *16*, 307.
- [120] K. Jaganathan, S. Kyriazopoulou Panagiotopoulou, J. F. McRae, S. F. Darbandi, D. Knowles, Y. I. Li, J. A. Kosmicki, J. Arbelaez, W. Cui, G. B. Schwartz, E. D. Chow, E. Kanterakis, H. Gao, A. Kia, S. Batzoglu, S. J. Sanders, K. K.-H. Farh, *Cell* **2019**, *176*, 535.
- [121] P. E. Griffiths, K. Stotz, *Theor. Med. Bioethics* **2006**, *27*, 499.
- [122] H. Pearson, *Nature* **2006**, *441*, 398.
- [123] ENCODE Project Consortium, E. Birney, J. A. Stamatoyannopoulos, A. Dutta, R. Guigó, T. R. Gingeras, E. H. Margulies, Z. Weng, M. Snyder, E. T. Dermitzakis, R. E. Thurman, M. S. Kuehn, C. M. Taylor, S. Neph, C. M. Koch, S. Asthana, A. Malhotra, I. Adzhubei, J. A. Greenbaum, R. M. Andrews, P. Flicek, P. J. Boyle, H. Cao, N. P. Carter, G. K. Clelland, S. Davis, N. Day, P. Dhami, S. L. C. Dillon, M. O. Dorschner, H. Fiegler, et al., *Nature* **2007**, *447*, 799.
- [124] Y. He, C. Yuan, L. Chen, M. Lei, L. Zellmer, H. Huang, D. J. Liao, *Genes* **2018**, *9*, 40.
- [125] X. Chen, J. R. Bracht, A. D. Goldman, E. Dolzhenko, D. M. Clay, E. C. Swart, D. H. Perlman, T. G. Doak, A. Stuart, C. T. Amemiya, R. P. Sebra, L. F. Landweber, *Cell* **2014**, *158*, 1187.
- [126] M. Rassoulzadegan, V. Grandjean, P. Gounon, S. Vincent, I. Gillot, F. Cuzin, *Nature* **2006**, *441*, 469.
- [127] P. Portin, *Hereditas* **2009**, *146*, 112.
- [128] L. Perbal, *EMBO Rep.* **2015**, *16*, 777.
- [129] P. Portin, A. Wilkins, *Genetics* **2017**, *205*, 1353.
- [130] A. E. Mirsky, H. Ris, *J. Gen. Physiol.* **1951**, *34*, 451.

- [131] F. Jacob, J. Monod, *J. Mol. Biol.* **1961**, *3*, 318.
- [132] B. Lewin, *Genes IV*, Oxford University Press, Oxford, UK **1990**.
- [133] F. Antequera, A. Bird, *Nat. Genet.* **1994**, *8*, 114.
- [134] G. D. Schuler, M. S. Boguski, E. A. Stewart, L. D. Stein, G. Gyapay, K. Rice, R. E. White, P. Rodriguez-Tomé, A. Aggarwal, E. Bajorek, S. Bentolila, B. B. Birren, A. Butler, A. B. Castle, N. Chiannikulchai, A. Chu, C. Clee, S. Cowles, P. J. Day, T. Dibling, N. Drouot, I. Dunham, S. Duprat, C. East, C. Edwards, J. B. Fan, N. Fang, C. Fizames, C. Garrett, L. Green, et al., *Science* **1996**, *274*, 540.
- [135] M. Das, C. B. Burge, E. Park, J. Colinas, J. Pelletier, *Genomics* **2001**, *77*, 71.
- [136] F. A. Wright, W. J. Lemon, W. D. Zhao, R. Sears, D. Zhuo, J. P. Wang, H. Y. Yang, T. Baer, D. Stredney, J. Spitzner, A. Stutz, R. Krahe, B. Yuan, *Genome Biol.* **2001**, *2*, research0025.1.
- [137] J. E. Collins, M. E. Goward, C. G. Cole, L. J. Smink, E. J. Huckle, S. Knowles, J. M. Bye, D. M. Beare, I. Dunham, *Genome Res.* **2003**, *13*, 27.
- [138] Z. Xuan, J. Wang, M. Q. Zhang, *Genome Biol.* **2002**, *4*, R1.
- [139] F. Sanger, S. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. USA* **1977**, *74*, 5463.
- [140] J. M. Heather, B. Chain, *Genomics* **2016**, *107*, 1.
- [141] M. R. Brent, *Genome Res.* **2005**, *15*, 1777.
- [142] M. Stanke, O. Schöffmann, B. Morgenstern, S. Waack, *BMC Bioinf.* **2006**, *7*, 62.
- [143] M. R. Brent, *Nat. Biotechnol.* **2007**, *25*, 883.
- [144] K. Hatje, B. Hammesfahr, M. Kollmar, *Nucleic Acids Res.* **2013**, *41*, W504.
- [145] M. L. Metzker, *Nat. Rev. Genet.* **2010**, *11*, 31.
- [146] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, et al., *Science* **2009**, *323*, 133.
- [147] D. Sharon, H. Tilgner, F. Grubert, M. Snyder, *Nat. Biotechnol.* **2013**, *31*, 1009.
- [148] Y. Feng, Y. Zhang, C. Ying, D. Wang, C. Du, *Genomics, Proteomics Bioinf.* **2015**, *13*, 4.
- [149] A. Byrne, A. E. Beaudin, H. E. Olsen, M. Jain, C. Cole, T. Palmer, R. M. DuBois, E. C. Forsberg, M. Akeson, C. Vollmers, *Nat. Commun.* **2017**, *8*, 16027.
- [150] D. R. Garalde, E. A. Snell, D. Jachimowicz, B. Sipo, J. H. Lloyd, M. Bruce, N. Pantic, T. Admassu, P. James, A. Warland, M. Jordan, J. Ciccone, S. Serra, J. Keenan, S. Martin, L. McNeill, E. J. Wallace, L. Jayasinghe, C. Wright, J. Blasco, S. Young, D. Brocklebank, S. Juul, J. Clarke, A. J. Heron, D. J. Turner, *Nat. Methods* **2018**, *15*, 201.
- [151] Á. Arzalluz-Luque, A. Conesa, *Genome Biol.* **2018**, *19*, 110.
- [152] K. Karlsson, S. Linnarsson, *BMC Genomics* **2017**, *18*, 126.

4

Conclusions

The present work incorporates the results of five studies that investigate two different topics from the research fields »coiled-coil prediction« and »heterologous gene expression«. At the centre of these studies stands the elucidation of biological questions that have become feasible and even more necessary in the first place through the ever-increasing amount of biological information. More and more genomic data is accumulated each year due to the advances in sequencing techniques. These developments led to an exponential growth of sequenced genome data and subsequently to further types of biological data such as sequential information of genes and proteins. By now, there are billions of generated protein sequences available in public databases to a large extent without experimentally proven structure and function. Computational approaches, like the ones presented here, complement and assist experimentally intensive research efforts by processing, analyzing and interpreting the produced amounts of biological data and by giving first forecasts to uncover hidden knowledge from the data and help answering these questions.

The prediction of coiled-coil domains as part of the computational protein structure and function prediction aims to contribute to one of the major challenges in biology, the gain of insight into the molecular structure, its mechanics and supported cellular processes of proteins to achieve a deeper understanding of their functions and roles. Coiled-coil prediction software has an important role in one of the first steps of the structural annotation of newly generated protein sequences with unknown molecule structure. In focus of this type of annotation stands therefore the generation of precise and reliable predictions of coiled-coil domains as well as high demands on the performance. Especially since these applications are not only used for single proteins, but even on a much larger scale for the annotation of entire genomes. Unfortunately established software have been shown to have a rather limited applicability especially in terms of the prediction quality with regard to large-scale coiled-coil analyses run against diverse types of protein sequence information. Difficulties lie not only in the prediction of the correct sequence regions, but also in whether regions with coiled-coils are recognized at all and, if so, are not mistaken for SAH domains. In addition, the quality of the predictions depends significantly on the software being used, because these often give different, and in part contradictory, results. This is either due to technical limitations in the chosen underlying model or deficiencies in the choice of reference data for training and fitting model parameters of the respective software. For this reason, the field of coiled-coil domain detection providing accurate and reliable predictions remains a challenge.

4.1. Contributions to prediction of coiled-coil and SAH domains

With our approach, we try to address improvements in protein structure prediction of coiled-coil domains on two levels, on the one side the prediction and interpretation for individual cases in very detail and on the other for large scale investigations as the annotation of whole genomes. One fo-

cus of this work has therefore been placed on the development of a web-based application, termed »Waggawagga«, for the comparative visualization of coiled-coil predictions from individual protein sequences generated by different softwares (Simm et al., 2015). A solution to provide end-users an easy-to-use tool, that provides access to the most widely used prediction softwares of the field in one place and helps to make a profound decision about obtained coiled-coil predictions for a single protein sequence. Unique features of Waggawagga lie in the structural overview representation of multiple aligned predictions in direct comparison and in detailed views of the respective predictions down to single amino acid interactions. As a basis for decision-making over a coiled-coil domain in question, the strength of the majority consensus of multiple and freely combinable prediction tools builds the central aspect of this comparative approach to overcome the limitations of single applications. It can be extended by the activation of additional data from oligomerization state predictors. Each of the predicted coiled-coil domains can be inspected in detail with the support of the *helical wheel* and the *helical net* structure representations. Both are interactive views, that allow a biological assessment of predicted coiled-coil domains based on interactions on the residue level. The helical wheel represents the predicted oligomeric state and its position-dependent amino acid interactions based on the heptad assignment in a defined region. Similarly, the helical net view visualizes interacting amino acids within a single stable α -helix according to the occupied heptad positions. Furthermore, supportive hints are provided by the specifically developed SAH prediction algorithm, a window-based net score, which assesses the occurrence of helix-stabilizing interactions based on single and network interactions between amino acids. It enables a distinct discrimination between putative coiled coils and actual stable single α -helix domains (SAH) and helps in the identification of real coiled-coil domains. This type of confusion is made by many tools through false-positive predictions, and is one of the side-effects of a too specialized training of the prediction models. Our approach follows the principle to use the strength of combining different information to build a broader basis for the decision-making and help to prevent these mispredictions in the future.

As a »proof of concept«, the developed SAH prediction algorithm has been calibrated, tested and evaluated as an initial benchmark against the largest available myosin sequence dataset. It consists of 7919 manually annotated myosin sequences from 938 species representing all major eukaryotic branches (Simm et al., 2017). A small part of the dataset based on well-known coiled-coil and SAH domains served initially for the calibration of the SAH scoring threshold. Although the function of the SAH-domains in human class-6 and class-10 myosins has well been characterized, the distribution and evaluation of the SAH-domain in all myosin subfamilies and across the eukaryotic tree of life remained elusive. With our approach we identified SAH-domains in more than one third of the supposed 79 myosin subfamilies. Depending on the myosin class, the presence of SAH-domains can range from a few to almost all class members indicating complex patterns of independent and taxon-specific SAH-domain gain and loss.

For studies at a larger scale as for genome-wide annotations, the SAH prediction algorithm to analyze SAH-domains in protein sequence datasets has been implemented as a stand-alone version for the command line, termed »Waggawagga-CLI« (Simm and Kollmar, 2018). Although the number of reported SAH-domains is steadily increasing, genome-wide analyses of SAH-domains in eukaryotic genomes are still missing. We took this as an opportunity to demonstrate and evaluate the function of the SAH prediction algorithm on large datasets. Using Waggawagga-CLI we predicted SAH-domains in 24 datasets from eukaryotes across the tree of life. SAH-domains were predicted in 0.5 to 3.5% of the protein-coding content per species. SAH-domains are particularly present in longer

proteins supporting their function as structural building block in multi-domain proteins. In human, SAH-domains are mainly used as alternative building blocks not being present in all transcripts of a gene. Gene ontology analysis showed that yeast proteins with SAH-domains are particularly enriched in macromolecular complex subunit organization, cellular component biogenesis and RNA metabolic processes, and that they have a strong nuclear and ribonucleoprotein complex localization and function in ribosome and nucleic acid binding. Human proteins with SAH-domains have roles in all types of RNA processing and cytoskeleton organization, and are predicted to function in RNA binding, protein binding involved in cell and cell-cell adhesion, and cytoskeletal protein binding. Waggawagga-CLI allows the user to adjust the stabilizing and destabilizing contribution of amino acid interactions in $i, i + 3$ and $i, i + 4$ spacings, and provides extensive flexibility for user-designed analyses.

As concluding study, the »status quo« of the current coiled-coil prediction options was re-evaluated and critically assessed. We performed a large-scale in-depth prediction analysis by testing the most relevant softwares of the field against the most comprehensive reference data set available, the entire Protein Data Bank, and tracked down the results to each amino acid and its secondary structure (Simm et al., 2021). Apart from the 30-fold difference in minimum and maximum number of coiled coils predicted the tools strongly vary in where they predict coiled-coil regions. Accordingly, there is a high number of false predictions and missed, true coiled-coil regions. The evaluation of the binary classification metrics in comparison with naïve coin-flip models and the calculation of the Matthews correlation coefficient, the performance metric for imbalanced data sets, suggests that the tested tools' performance is close to random. This implicates that the tools' predictions have only limited informative value. Coiled-coil predictions are often used to interpret biochemical data and are part of in-silico functional genome annotation. The observed results indicate that these predictions should be treated very cautiously and need to be supported and validated by experimental evidence. In the course of this study, the complete prediction results for all the 144,270 analyzed PDB structures were processed separately for transparency purposes and made available to the general public. The individual predictions for each PDB accession can be inspected and reviewed in detail in the web-application Waggawagga using the provided WebGL viewers on the 3D molecule structures and in direct comparison in the comparative visualization.

Outlook

Although many studies and developments have been made in this field over long time, some approaches to contribute to the improvement of the overall quality of coiled-coil predictions are still conceivable. One of the obvious and directly feasible solutions to address this challenge, referring here to the results of our performed studies on comparative coiled-coil prediction and the discrimination SAH domains, should lie in the combined consideration of multiple predictions generated from different detection softwares. These predictions differ in parts strongly in their extend of coiled-coil harbouring regions and with a high proportion of false positive predictions, but deliver at same time helpful information to identify the actual coiled coils, compare our findings from Simm et al. (2021). Using several predictions stemming from a well-chosen set of different softwares creates a type of consensus of the predicted domains and regions within the analyzed protein sequences, that provides profound information about conserved and hence likely correct predicted regions and delivers reliable evidence.

A more elaborate way could be the development of a next-generation of coiled-coil prediction softwares by combining well-performing traits of previous developments such as the usage of windowless models, adopting efficient approaches from other fields and extensively exploiting additional biological information based on public structural data of proteins. As demonstrated by the pioneering development of the protein structure prediction software Alphafold (Jumper et al., 2021), crucial key factors lie in combining the incorporation of large amounts of diverse biological data information with the novel techniques of deep learning for interpretation. At the same time, new developments should address coiled-coil types deviating globally from the heptad periodicity, because currently most of the published coiled-coil detection tools are designed to analyze unbroken heptad repeats and have therefore limited applicability to coiled coils with other periodicities.

4.2. Contributions to heterologous expression of genes

Equally essential for a deeper understanding of the cellular functions of proteins are the underlying mechanisms from the encoding and transcription of genetic information through biosynthetic production by the ribosome to the correctly folded and functional protein. Heterologous protein expression is an important method for analyzing these mechanisms and processes in the cells to identify regulating factors. It is therefore often applied in the investigation of cellular functions, in genetic circuit engineering and in overexpressing proteins for biopharmaceutical applications and structural biology research. One of the key factors for heterologous expression represents the degeneracy of the genetic code, which enables a single protein to be encoded by a multitude of synonymous gene sequences and simultaneously allows adjusting gene sequences without changing the protein sequences, but substantial uncertainty exists concerning the details of this phenomenon.

The focus of this work was the development of the software »Odysseus«, that implements a new approach for the design of typical protein-coding genes for heterologous expression applications in common model organisms. The software realizes a probabilistic approach based on Markov models, that is highly configurable and can be operated with pre-trained genome profiles to control protein expression levels by the codon usage adaptation of genes. Specially composed genomic data sets build the basis for the computed profiles from which characteristic genetic information of the expression systems are derived. In contrast to most other tools, which intend to optimize the codon usage by selecting only codons from a few highly expressed proteins or by selecting only the codon with the highest relative codon usage from each codon box, Odysseus generates genes resembling the codon usage of any conceivable subset of endogenous genes, for instance a selected group of proteins. Such groups can be the highest or lowest expressed proteins of a species (with the cut-off free to choose), or even a subset of proteins with a certain function. Odysseus is made available to the general public as a web-application for performing host-specific protein adaptations to a set of the most commonly used model organisms.

As a »proof of concept« we analyze in the performed study "Design of typical genes for heterologous gene expression", referring to the recently submitted manuscript from Chapter 3.2, the influence of the profiled codon usage adaptation approach on protein expression levels in the eukaryotic model organism *Saccharomyces cerevisiae*. We selected green fluorescent protein (GFP) and human α -synuclein (α Syn) as representatives for stable and intrinsically disordered proteins and representing a benchmark and a challenging test case. Using the new approach, synthetic genes for GFP and

α Syn were generated, heterologously expressed and evaluated in yeast. To this end, we first started to generate synthetic genes of the non-toxic, highly structured protein GFP and to evaluate their expression level. The expression level strongly increased from the gene based on the lowest expressed proteins to the gene based on the highest expressed proteins. This supports the general finding that protein expression is stronger when adapting a heterologous gene to the most used codons. Such a strong expression is, however, often not wanted and disfavoured when trying to express a toxic protein. To test our software for its use for expressing proteins at low endogenous protein expression levels, we designed synthetic genes for the toxic, non-structured protein α -synuclein and showed that human α Syn can be adapted to low expression levels. These initially conducted tests on the two representatives of the protein classes suggest that Odysseus is a valuable approach for designing typical genes for heterologous protein expression with which insightful results can be achieved in the regulation of expression levels.

Outlook and future work

However, our approach is still at the stage of development and more experimental tests with several different proteins are needed. At this point in time, our approach can be used to roughly influence the direction of the expression levels. In order to make more targeted adaptations in the gene design process, that have the adjusted effect on the expression level, a much broader experimental evaluation of modified genes is required to better analyze and understand the inferences between synonymous sequence modifications and effects on expressed proteins. Furthermore, in our study only the model organism *Saccharomyces cerevisia* was used as a test expression system. To be able to make more reliable gene adaptations in general the effects need also to be investigated in other common model organisms such as the in the web-application provided genome profiles of *E. coli* and *A. thaliana*.

References

- Armstrong, C. T., Vincent, T. L., Green, P. J., and Woolfson, D. N. (2011). SCORER 2.0: an algorithm for distinguishing parallel dimeric and trimeric coiled-coil sequences. *Bioinformatics*, 27(14):1908–1914.
- Barbara, K. E., Willis, K. A., Haley, T. M., Deminoff, S. J., and Santangelo, G. M. (2007). Coiled coil structures and transcription: an analysis of the *S. cerevisiae* coilome. *Mol. Genet. Genomics*, 278(2):135–147.
- Berger, B., Wilson, D. B., Wolf, E., Tonchev, T., Milla, M., and Kim, P. S. (1995). Predicting coiled coils by use of pairwise residue correlations. *Proc Natl Acad Sci U S A*, 92(18):8259–8263.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242.
- Boël, G., Letso, R., Neely, H., Price, W. N., Wong, K.-H., Su, M., Luff, J., Valecha, M., Everett, J. K., Acton, T. B., Xiao, R., Montelione, G. T., Aalberts, D. P., and Hunt, J. F. (2016). Codon influence on protein expression in *E. coli* correlates with mRNA levels. *Nature*, 529(7586):358–363.
- Brase, L. R. and Ridge, P. G. (2019). ExtRamp: a novel algorithm for extracting the ramp sequence based on the tRNA adaptation index or relative codon adaptiveness. *Nucleic Acids Res*, pages 1–9.
- Burgess-Brown, N. A., Sharma, S., Sobott, F., Loenarz, C., Oppermann, U., and Gileadi, O. (2008). Codon optimization can improve expression of human genes in *Escherichia coli*: A multi-gene study. *Protein Expr. Purif*, 59(1):94–102.
- Chalfie, M., Tu, Y., Euskirchen, G., Ward, W. W., and Prasher, D. C. (1994). Green fluorescent protein as a marker for gene expression. *Science (New York, NY)*, 263(5148):802–805.
- Chandra, S., Chen, X., Rizo, J., Jahn, R., and Südhof, T. C. (2003). A broken alpha-helix in folded alpha-Synuclein. *Journal of Biological Chemistry*, 278(17):15313–15318.
- Cheng, S., Melkonian, M., Smith, S. A., Brockington, S., Archibald, J. M., Delaux, P.-M., Li, F.-W., Melkonian, B., Mavrodiev, E. V., Sun, W., Fu, Y., Yang, H., Soltis, D. E., Graham, S. W., Soltis, P. S., Liu, X., Xu, X., and Wong, G. K.-S. (2018). 10KP: A phylodiverse genome sequencing plan. *Giga-Science*, 7(3):1–9.
- Church, G. M., Gao, Y., and Kosuri, S. (2012). Next-generation digital information storage in DNA. *Science*, 337(6102):1628–1628.
- Coleman, J. R., Papamichail, D., Skiena, S., Futcher, B., Wimmer, E., and Mueller, S. (2008). Virus attenuation by genome-scale changes in codon pair bias. *Science*, 320(5884):1784–1787.
- Crick, F. H. C. (1953). The packing of α -helices: simple coiled-coils. *Acta crystallographica*, 6(8):689–697.

- Delorenzi, M. and Speed, T. (2002). An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. *Bioinformatics*, 18(4):617–625.
- Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis*. Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press.
- Gáspári, Z., Süveges, D., Perczel, A., Nyitray, L., and Tóth, G. (2012). Charged single alpha-helices in proteomes revealed by a consensus prediction approach. *Biochim Biophys Acta*, 1824(4):637–646.
- Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E. M., Sipos, B., and Birney, E. (2013). Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature*, 494(7435):77–80.
- Goodman, S. R. (2007). Medical Cell Biology. In *Chapter 1 - Tools of the Cell Biologist: Green Fluorescent Protein*, page 336. Academic Press.
- Gorochowski, T. E., Ignatova, Z., Bovenberg, R. A. L., and Roubos, J. A. (2015). Trade-offs between tRNA abundance and mRNA secondary structure support smoothing of translation elongation rate. *Nucleic Acids Research*, 43(6):gkv199–3032.
- Grantham, R., Gautier, C., Gouy, M., Mercier, R., and Pavé, A. (1980). Codon catalog usage and the genome hypothesis. *Nucleic Acids Research*, 8(1):r49–r62.
- Gruber, M., Söding, J., and Lupas, A. N. (2005). REPPER—repeats and their periodicities in fibrous proteins. *Nucleic Acids Res*, 33(Web Server issue):W239–43.
- Hershberg, R. and Petrov, D. A. (2008). Selection on codon bias. *Annu. Rev. Genet.*, 42(1):287–299.
- Hershberg, R. and Petrov, D. A. (2009). General rules for optimal codon choice. *PLoS Genet*, 5(7):e1000556.
- Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. (2012). A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science*.
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, pages 1–11.
- Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637.
- Levinthal, C. (1969). How to fold graciously. In *Mossbauer Spectroscopy in Biological Systems Proceedings of a meeting held at Allerton House, Monticello, Illinois*, pages 22–24. University of Illinois Press Urbana.
- Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., Durbin, R., Edwards, S. V., Forest, F., Gilbert, M. T. P., Goldstein, M. M., Grigoriev, I. V., Hackett, K. J., Hausler, D., Jarvis, E. D., Johnson, W. E., Patrinos, A., Richards, S., Castilla-Rubio, J. C., van Sluys, M.-A., Soltis, P. S., Xu, X., Yang, H., and Zhang, G. (2018). Earth BioGenome Project: Sequencing life

- for the future of life. *Proceedings of the National Academy of Sciences of the United States of America*, 115(17):4325–4333.
- Liu, H., Wei, J., Yang, T., Mu, W., Song, B., Yang, T., Fu, Y., Wang, X., Hu, G., Li, W., Zhou, H., Chang, Y., Chen, X., Chen, H., Cheng, L., He, X., Cai, H., Cai, X., Wang, M., Li, Y., Sahu, S. K., Yang, J., Wang, Y., Mu, R., Liu, J., Zhao, J., Huang, Z., Xu, X., and Liu, X. (2019). Molecular digitization of a botanical garden: high-depth whole-genome sequencing of 689 vascular plant species from the Ruili Botanical Garden. *GigaScience*, 8(4):1372.
- Liu, J. and Rost, B. (2001). Comparing function and structure between entire proteomes. *Protein Sci*, 10(10):1970–1979.
- Ludwiczak, J., Winski, A., Szczepaniak, K., Alva, V., and Dunin-Horkawicz, S. (2019). DeepCoil - a fast and accurate prediction of coiled-coil domains in protein sequences. *Bioinformatics*, 25:3389.
- Lupas, A., Van Dyke, M., and Stock, J. (1991). Predicting coiled coils from protein sequences. *Science (New York, NY)*, 252(5009):1162–1164.
- Lupas, A. N., Bassler, J., and Dunin-Horkawicz, S. (2017). The Structure and Topology of α -Helical Coiled Coils. *Subcellular Biochemistry*, 82(6813):95–129.
- Mahrenholz, C., Abfalter, I., Bodenhofer, U., Volkmer, R., and Hochreiter, S. (2010). ProCoil - Advances in predicting two- and three-stranded coiled coils. *Nature Precedings*, pages 1–1.
- McDonnell, A. V., Jiang, T., Keating, A. E., and Berger, B. (2006). Paircoil2: improved prediction of coiled coils from sequence. *Bioinformatics*, 22(3):356–358.
- McFarlane, A. A., Orriss, G. L., and Stetefeld, J. (2009). The use of coiled-coil proteins in drug delivery systems. *European journal of pharmacology*, 625(1-3):101–107.
- Meade, R. M., Fairlie, D. P., and Mason, J. M. (2019). Alpha-synuclein structure and Parkinson’s disease – lessons and emerging principles. *Molecular Neurodegeneration*, 14(1):1–14.
- Merkl, R. and Waack, S. (2003). *Bioinformatik Interaktiv. Algorithmen und Praxis*. Wiley-VCH Verlag GmbH.
- Mount, D. W. (2004). *Bioinformatics. Sequence and Genome Analysis*. CSHL Press.
- Moutevelis, E. and Woolfson, D. N. (2009). A periodic table of coiled-coil protein structures. *J Mol Biol*, 385(3):726–732.
- Newman, J. R., Wolf, E., and Kim, P. S. (2000). A computationally directed screen identifying interacting coiled coils from *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A*, 97(24):13203–13208.
- Nierhaus, K. H., Wadzack, J., Burkhardt, N., Jünemann, R., Meerwinck, W., Willumeit, R., and Stuhmann, H. B. (1998). Structure of the elongating ribosome: arrangement of the two tRNAs before and after translocation. *Proc Natl Acad Sci U S A*, 95(3):945–950.
- Ormö, M., Cubitt, A. B., Kallio, K., Gross, L. A., Tsien, R. Y., and Remington, S. J. (1996). Crystal structure of the *Aequorea victoria* green fluorescent protein. *Science (New York, NY)*, 273(5280):1392–1395.

- Parry, D. A. (1982). Coiled-coils in alpha-helix-containing proteins: analysis of the residue types within the heptad repeat and the use of these data in the prediction of coiled-coils in other proteins. *Bioscience reports*, 2(12):1017–1024.
- Peckham, M. and Knight, P. J. (2009). When a predicted coiled coil is really a single α -helix, in myosins and other proteins. *Soft Matter*, 5(13):2493–2503.
- Plotkin, J. B. and Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet*, 12(1):32–42.
- Rackham, O. J. L., Madera, M., Armstrong, C. T., Vincent, T. L., Woolfson, D. N., and Gough, J. (2010). The evolution and structure prediction of coiled coils across all genomes. *J Mol Biol*, 403(3):480–493.
- Roberts, R. J., Vincze, T., Posfai, J., and Macelis, D. (2010). REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res*, 38(Database issue):D234–6.
- Rose, A. S., Bradley, A. R., Valasatava, Y., Duarte, J. M., Prlić, A., and Rose, P. W. (2018). NGL viewer: web-based molecular graphics for large complexes. *Bioinformatics*, 34(21):3755–3758.
- Sehnal, D., Bittrich, S., Deshpande, M., Svobodová, R., Berka, K., Bazgier, V., Velankar, S., Burley, S. K., Koča, J., and Rose, A. S. (2021). Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res*, 49(W1):W431–W437.
- Shankland, S. (2019). Startup Catalog has jammed all 16GB of Wikipedia’s text onto DNA strands. *CNET*.
- Sharp, P. M., Cowe, E., Higgins, D. G., Shields, D. C., Wolfe, K. H., and Wright, F. (1988). Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. *Nucleic Acids Research*, 16(17):8207–8211.
- Simm, D., Hatje, K., and Kollmar, M. (2015). Waggawagga - comparative visualization of coiled-coil predictions and detection of stable single α -helices (SAH domains). *Bioinformatics*, 31(5):767–769.
- Simm, D., Hatje, K., and Kollmar, M. (2017). Distribution and evolution of stable single α -helices (SAH domains) in myosin motor proteins. *PLoS ONE*, 12(4):e0174639.
- Simm, D., Hatje, K., Waack, S., and Kollmar, M. (2021). Critical assessment of coiled-coil predictions based on protein structure data. *Sci Rep*, 11(1):12439–18.
- Simm, D. and Kollmar, M. (2018). Waggawagga-CLI: A command-line tool for predicting stable single α -helices (SAH-domains), and the SAH-domain distribution across eukaryotes. *PLoS ONE*, 13(2):e0191924.
- Stark, H., Orlova, E. V., Rinke-Appel, J., Jünke, N., Mueller, F., Rodnina, M., Wintermeyer, W., Brimacombe, R., and van Heel, M. (1997). Arrangement of tRNAs in pre- and posttranslocational ribosomes revealed by electron cryomicroscopy. *Cell*, 88(1):19–28.
- Testa, O. D., Moutevelis, E., and Woolfson, D. N. (2009). CC+: a relational database of coiled-coil structures. *Nucleic Acids Res*, 37(Database issue):D315–22.

- Trigg, J., Gutwin, K., Keating, A. E., and Berger, B. (2011). Multicoil2: Predicting Coiled Coils and Their Oligomerization States from Sequence in the Twilight Zone. *PLoS ONE*, 6(8):e23519.
- Tsien, R. Y. (1998). The green fluorescent protein. *Annu Rev Biochem*, 67:509–544.
- Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., Potapenko, A., Ballard, A. J., Romera-Paredes, B., Nikolov, S., Jain, R., Clancy, E., Reiman, D., Petersen, S., Senior, A. W., Kavukcuoglu, K., Birney, E., Kohli, P., Jumper, J., and Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature*, pages 1–9.
- Turnbull, C., Scott, R. H., Thomas, E., Jones, L., Murugaesu, N., Pretty, F. B., Halai, D., Baple, E., Craig, C., Hamblin, A., Henderson, S., Patch, C., O'Neill, A., Devereau, A., Smith, K., Martin, A. R., Sosinsky, A., McDonagh, E. M., Sultana, R., Mueller, M., Smedley, D., Toms, A., Dinh, L., Fowler, T., Bale, M., Hubbard, T., Rendon, A., Hill, S., Caulfield, M. J., and 100000 Genomes Project (2018). The 100000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ (Clinical research ed.)*, 361:k1687.
- UniProt Consortium, T. (2018). UniProt: the universal protein knowledgebase. *Nucleic Acids Res*, 46(5):2699.
- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferreira, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigo, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B.,

- Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The sequence of the human genome. *Science (New York, NY)*, 291(5507):1304–1351.
- Vincent, T. L., Green, P. J., and Woolfson, D. N. (2013). LOGICOIL—multi-state prediction of coiled-coil oligomeric state. *Bioinformatics*, 29(1):69–76.
- Walshaw, J. and Woolfson, D. N. (2001). Socket: a program for identifying and analysing coiled-coil motifs within protein structures. *Journal of Molecular Biology*, 307(5):1427–1450.
- Wang, M., Herrmann, C. J., Simonovic, M., Szklarczyk, D., and von Mering, C. (2015). Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics*, 15(18):3163–3168.
- Watson, J. D., Baker, T. A., and Bell, S. P. (2014). *Molecular Biology of the Gene*. Benjamin-Cummings Publishing Company, 7 edition.
- Weinreb, P. H., Zhen, W., Poon, A. W., Conway, K. A., and Lansbury, P. T. (1996). NACP, a protein implicated in Alzheimer’s disease and learning, is natively unfolded. *Biochemistry*, 35(43):13709–13715.
- Wolf, E., Kim, P. S., and Berger, B. (1997). MultiCoil: a program for predicting two- and three-stranded coiled coils. *Protein Sci*, 6(6):1179–1189.
- Wood, C. W. and Woolfson, D. N. (2018). CCBuilder 2.0: Powerful and accessible coiled-coil modeling. *Protein Science*, 27(1):103–111.
- Woolfson, D. N. (2017). Coiled-Coil Design: Updated and Upgraded. In *Fibrous Proteins: Structures and Mechanisms*, pages 35–61. Springer, Cham.
- Woolfson, D. N. and Alber, T. (1995). Predicting oligomerization states of coiled coils. *Protein Science*, 4(8):1596–1607.
- wwPDB consortium (2019). Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res*, 47(D1):D520–D528.
- Xia, Y., Saitoh, T., Ueda, K., Tanaka, S., Chen, X., Hashimoto, M., Hsu, L., Conrad, C., Sundsmo, M., Yoshimoto, M., Thal, L., Katzman, R., and Masliah, E. (2001). Characterization of the human α -synuclein gene: Genomic structure, transcription start site, promoter region and polymorphisms. *Journal of Alzheimer’s Disease*, 3(5):485–494.
- Zimmer, M. (2002). Green fluorescent protein (GFP): applications, structure, and related photophysical behavior. *Chemical reviews*, 102(3):759–781.

Acknowledgements

First and foremost, I would like to thank my supervisors PD Dr. Martin Kollmar and Prof. Dr. Stephan Waack. They were always available to me for valuable discussions and with an open ear for questions, and they provided ideas when one of the projects stalled. I would like to thank Martin for his unwavering commitment and time. Through the countless emails and hours of scientific discussion, ideas and help with analyses and his eye for detail, he was always a good mentor and role model for conscientious scientific work.

Another special thanks goes to Stephan, who, after I was able to start my PhD project at the Max-Planck-Institute for Biophysical Chemistry but was unable to continue it due to our research group leaving at the time, accepted me into his department and has since supported me in many instructive discussions as his PhD student.

I would like to thank Prof. Dr. Gerhard Braus for his generous support for the experimental evaluation of our gene expression project, as well as Dr. Blagovesta Popova for the great implementation of the experiments in the lab.

Furthermore, I would like to thank my colleagues and fellow students Dr. Klas Hatje, Dr. Daniel Honsel, Fabian Korte, Dr. Benjamin Heuer and Dr. Fabian Telschow for questions, suggestions and exchanges in the office or on the way to the canteen. Special thanks also go to my family and friends, of course, who have always supported me with enthusiasm and motivating words. Last but not least, my thanks go to my dear Ricarda, who supported me in all phases of my doctorate.

Dominic Simm,
Göttingen, January 3rd, 2022

List of Figures

| | | |
|-------|--|-----|
| 2.1. | Growth of the protein sequence knowledge base. | 9 |
| 2.2. | Genetic code. | 13 |
| 2.3. | The four structural levels of protein folding. | 14 |
| 2.4. | Representatives of stable and intrinsically disordered proteins. | 15 |
| 2.5. | Representations of a parallel dimeric coiled coil. | 18 |
| 2.6. | Oligomeric states of coiled coils: antiparallel dimer, parallel dimer, trimer and tetramer. | 18 |
| 2.7. | The periodic table of coiled coils. | 19 |
| 2.8. | Heptad net representation of a CC and a stable SAH. | 21 |
| 2.9. | Illustration of the state graph of the Markovian adapter. | 22 |
| 2.10. | Illustration of the extended state graph of the Markovian adapter. | 24 |
| 2.11. | Overview of the available coiled-coil prediction programs. | 26 |
| 2.12. | MTK and MTIDK matrices of the coiled coil prediction program COILS. | 27 |
| 2.13. | Pairwise residue correlations: illustration of the maximum window score calculation. | 28 |
| 2.14. | Marcoil: representation of the HMM-based approach. | 29 |
| 2.15. | Growth of the RCSB Protein Data Bank. | 33 |
| 2.16. | Extracts from a PDB flat file. | 34 |
| 2.17. | Protein abundance histogram. | 35 |
| | | |
| 3.1. | Example of a Markov chain. | 102 |
| 3.2. | Odysseus flowchart. | 104 |
| 3.3. | Plasmids used in this study. | 106 |
| 3.4. | Yeast strains used in this study. | 106 |
| 3.5. | GPome-plots (genome versus proteome plots). | 109 |
| 3.6. | Steady-state protein levels of GFP. | 112 |
| 3.7. | Expression of designed and human α -synuclein. | 114 |
| 3.8. | The "inverted" codon usage. | 115 |
| | | |
| A.1. | Prediction of coiled-coils in human MyoVI with CCHMM- PROF, and in mouse MyoX with DeepCoil and DeepCoil2. | 203 |
| A.2. | Prediction of coiled-coils in human muscle and non-muscle myosin heavy chain proteins (class-2 myosins) using DeepCoil2 (v. 2.0.1. 30 Nov 2020). | 204 |
| A.3. | Upset plot of the intersection of SOCKET hits and coiled-coil predictions. | 205 |
| A.4. | Performance of coiled-coil prediction tools in dependence of reference sequence length. | 206 |
| A.5. | Coiled-coil prediction on human adult skeletal myosin heavy chain protein. | 207 |
| A.6. | Overlap of coiled-coil predictions with SOCKET hit regions. | 208 |
| A.7. | Performance of coiled-coil prediction tools in dependence of overlap with SOCKET reference. | 209 |
| A.8. | Performance of coiled-coil prediction tools in dependence of overlap with SOCKET reference. | 210 |

| | |
|--|-----|
| A.9. Amino acid preferences at heptad positions abcdefg. | 211 |
| A.10. Simplified database scheme. | 212 |
| A.11. SOCKET coiled coils not present in the benchmark data due to SOCKET output problems and missing multimers. | 213 |
| A.12. Southern hybridization and copy number determination. | 214 |
| A.13. Full size blots corresponding to Figure 3.6A. (Manuscript 3.2) | 215 |
| A.14. Full size blots corresponding to Figure 3.7A. | 216 |
| A.15. Full size blots corresponding to Figure 3.7C. | 217 |
| A.16. The "inverted" di-codon usage. | 218 |

A

Appendix

A.1. Supplementary information

Limits of previous evaluations of coiled-coil prediction tools

Earlier assessments of coiled-coil tool performance used user-defined subsets of the PDB. However, instead of mining the PDB directly, another database was used, the SCOP Structural Classification of Proteins database (Andreeva et al., 2004), to extract super families with coiled-coil domains for the positives and superfamilies of the alpha and beta protein class for the negatives (Gruber et al., 2006). Both categories were enriched by homology searches in the nonredundant database followed by sequence alignment and redundancy reduction to result in a final data set of about 1% positives, a highly imbalanced data set. This approach generated hardly verifiable reference data as it involved manual removal of SCOP coiled-coil superfamilies from the positives, which according to SOCKET did not contain coiled-coils, manual removal of SCOP superfamilies from the negatives, which obviously contained coiled-coils, and the sequence search and alignment of the notoriously difficult to align coiled-coil domains. The performance of coiled-coil prediction tools was assessed at the residue level by sensitivity and specificity, which were termed coverage and reliability for unknown reasons (Gruber et al., 2006).

In a subsequent analysis, all PDB structures deposited in an 18-month period were downloaded, filtered to 95% sequence identity, and split into positives and negatives using SOCKET hits as criterion (Rackham et al., 2010). With about 2.6% positives this data set was also quite imbalanced. Coiled-coils were evaluated at the protein sequence level, albeit only by counting hits on the same sequence without verifying the precise location of reference and prediction (e.g. overlap was not confirmed).

A later comparative study combined both approaches (Li et al., 2016): The positives were determined by mining the PDB with SOCKET. The negatives were compiled in a difficult to follow procedure of selecting alpha and beta classes from SCOPe, removing classes that contain annotated coiled coils, removing all available training data from prediction tools, recombining the resulting set again with the training data of three tools, and finally applying a 30% sequence identity cut-off using CD-HIT. The final data set contained 1643 sequences, of which 601 did not contain any coiled coil (negatives) and of which 1042 contained 2176 coiled coils (positives). Performance was evaluated as sensitivity and false positive rate ($1 - \text{specificity}$) at the sequence level, e.g. a SOCKET hit and a coiled-coil prediction in the same protein sequence were regarded as true positive hit. It is obvious, that this data set is highly imbalanced towards positives (63%), and that chances to obtain positive hits are additionally increased by the fact that most of the sequences of the positive class contain multiple coiled coils.

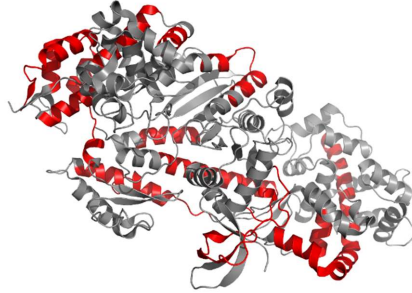
Most recently, test data were compiled by selecting the PDB data clustered by 50% sequence identity, running SOCKET and preferentially selecting structures with coiled coils as representatives of the clusters (Ludwiczak et al., 2019). Further redundancy in the selected structures was removed by applying CD-HIT to 50% sequence identity. The data were further filtered by structure resolution and minimum and maximum protein sequence length cut-offs. To increase the proportion of the positives, half of the negatives (sequences without coiled coil) were randomly removed. The prediction tools' performance was evaluated using multiple binary classification metrics (except the MCC) at the sequence and residue level.

References

- Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J. P., Chothia, C., and Murzin, A. G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, 32(Database issue):D226–229.
- Gruber, M., Söding, J., and Lupas, A. N. (2006). Comparative analysis of coiled-coil prediction methods. *J. Struct. Biol.*, 155(2):140–145.
- Li, C., Ching Han Chang, C., Nagel, J., Porebski, B. T., Hayashida, M., Akutsu, T., Song, J., and Buckle, A. M. (2016). Critical evaluation of in silico methods for prediction of coiled-coil domains in proteins. *Brief. Bioinformatics*, 17(2):270–282.
- Ludwiczak, J., Winski, A., Szczepaniak, K., Alva, V., and Dunin-Horkawicz, S. (2019). DeepCoil—a fast and accurate prediction of coiled-coil domains in protein sequences. *Bioinformatics*, 35(16):2790–2795.
- Rackham, O. J. L., Madera, M., Armstrong, C. T., Vincent, T. L., Woolfson, D. N., and Gough, J. (2010). The evolution and structure prediction of coiled coils across all genomes. *J. Mol. Biol.*, 403(3):480–493.
- Simm, D., Hatje, K., and Kollmar, M. (2015). Waggawagga: comparative visualization of coiled-coil predictions and detection of stable single α -helices (SAH domains). *Bioinformatics*, 31(5):767–769.

A.2. Supplementary figures

1A



1B

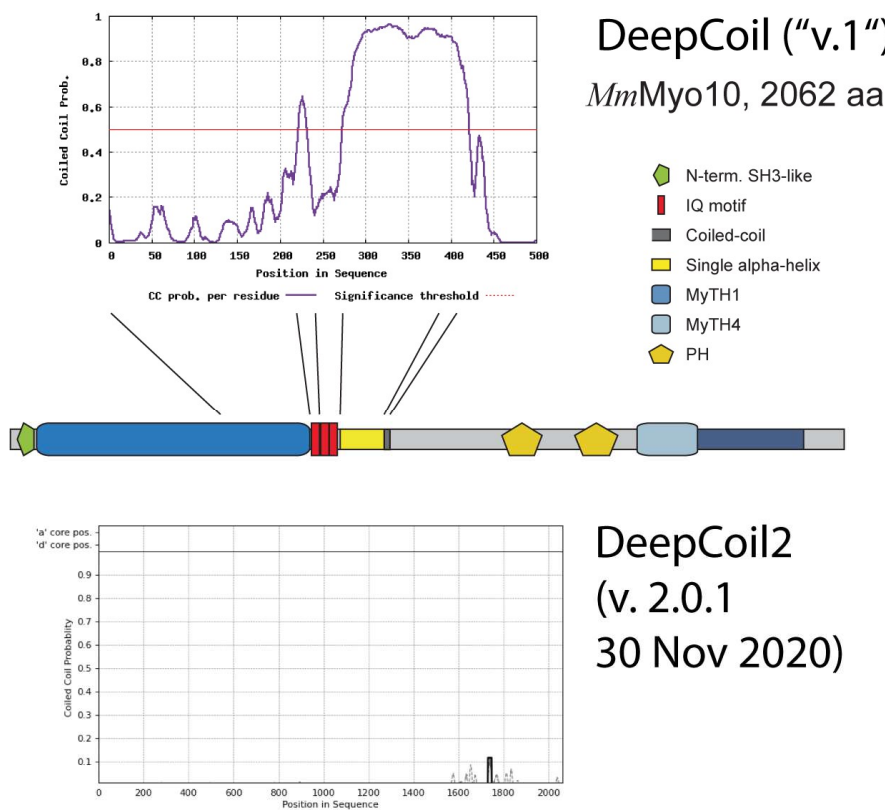


Figure A.1. | Prediction of coiled-coils in human MyoVI with CCHMM- PROF, and in mouse MyoX with DeepCoil and DeepCoil2. A) To test the performance of CCHMM-PROF, which is only accessible via a web interface, the motor domain sequence of human MyoVI was used. MyoVI is a backwards walking myosin motor, and was chosen by chance from the list of available protein crystal structures of myosin motor domains. Predicted coiled-coil regions were mapped onto the crystal structure, PDB ID 2BKH, and are shown in red. **B)** To test the performance of DeepCoil and DeepCoil2, which are accessible via a web interface, the mouse MyoX was used. The output of DeepCoil was mapped onto the domain architecture for better orientation. Surprisingly, while there is coiled-coil probability for a large region from DeepCoil prediction, DeepCoil2 does not predict any coiled-coil region.

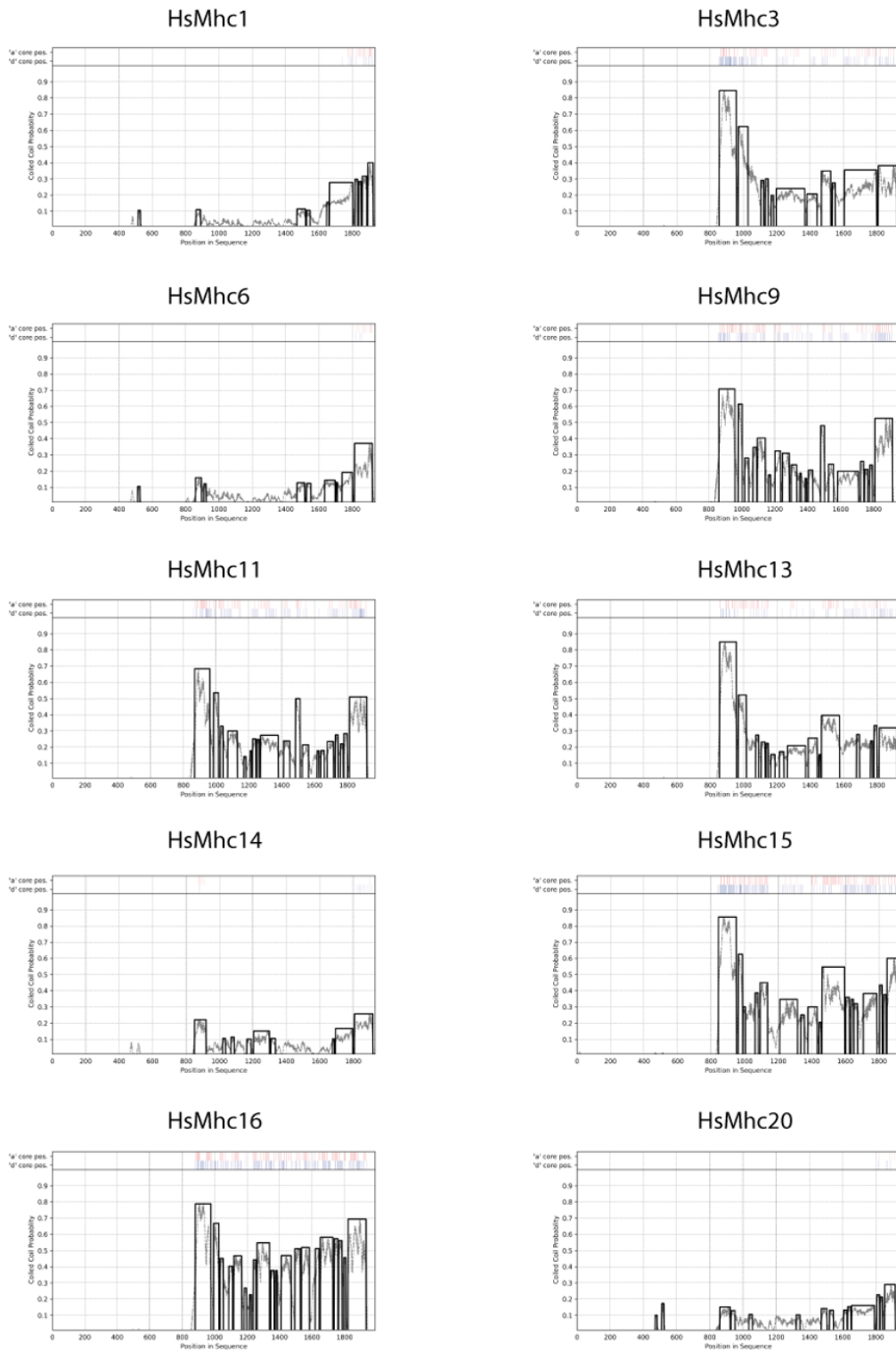


Figure A.2. | Prediction of coiled-coils in human muscle and non-muscle myosin heavy chain proteins (class-2 myosins) using DeepCoil2 (v. 2.0.1. 30 Nov 2020). To test the performance of DeepCoil2 we predicted coiled-coil regions in several of the classic coiled-coil forming proteins, the muscle and non-muscle myosin heavy chain proteins from human. For protein naming please check (Kollmar and Muhlhausen, 2017b). Predictions by the tools analysed in this benchmark study are shown in figure S5.

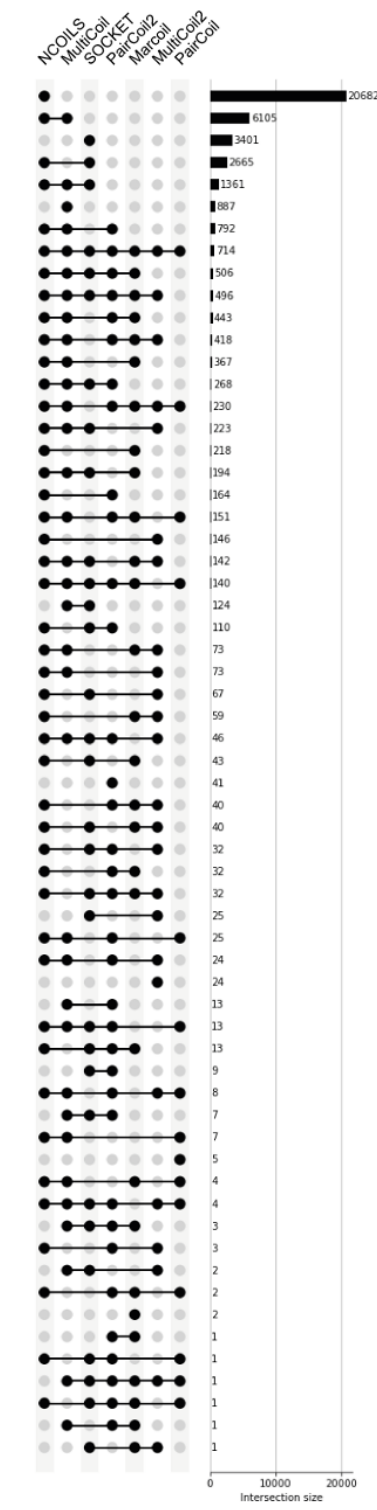


Figure A.3. | Upset plot of the intersection of SOCKET hits and coiled-coil predictions. The upset plot is based on the same data presented in main Fig 1C.

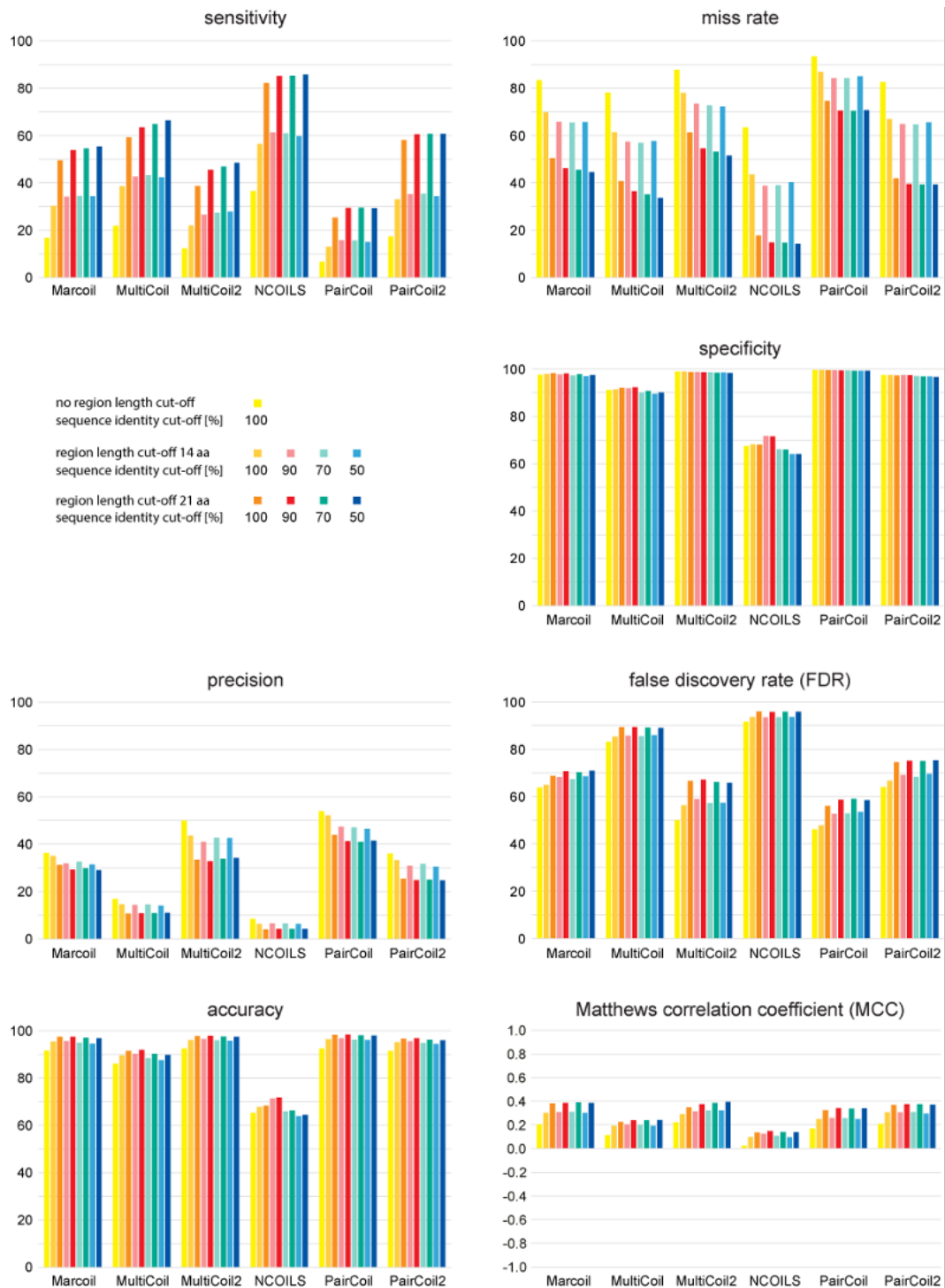


Figure A.4. | Performance of coiled-coil prediction tools in dependence of reference sequence length. Several classification metrics are shown for the six coiled-coil prediction tools with respect to SOCKET coiled-coil identifications. Classification as true positive hit requires overlap of at least a single amino acid between prediction and SOCKET. The performance was analysed for data sets with increasing minimum length of coiled-coil region (SOCKET hit or prediction) and decreasing levels of sequence redundancy.

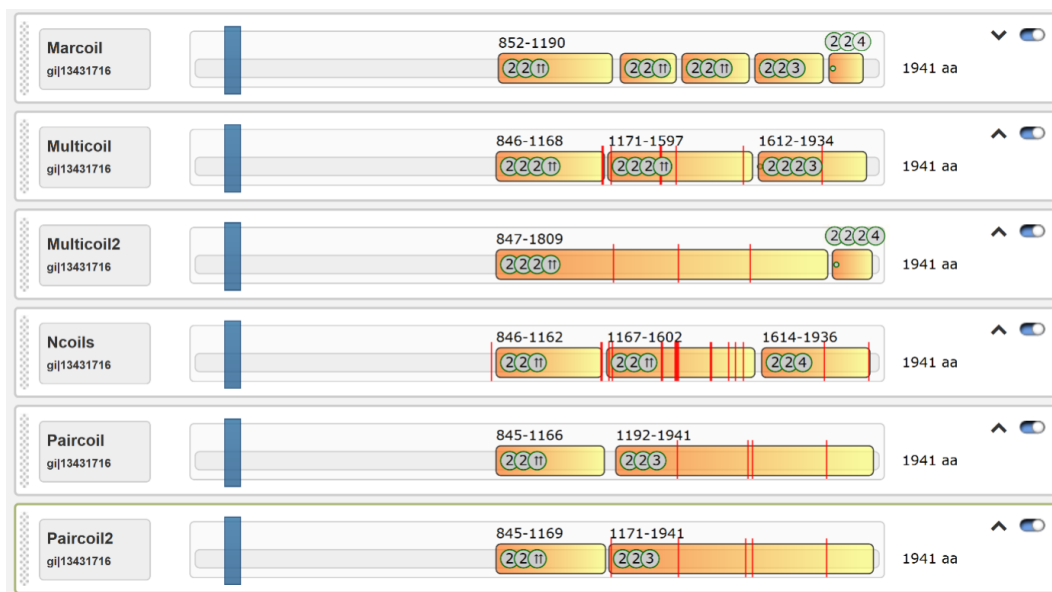


Figure A.5. | Coiled-coil prediction on human adult skeletal myosin heavy chain protein. The coiled-coil tools of interest were run on the human myosin heavy chain 2a protein sequence (skeletal muscle, adult 2 variant; GenBank identifier 13431716). The orange bars denote the predicted coiled-coil regions with their range given in numbers on top of the bars. Red lines indicate interruptions of heptads, regular breaks such as stutters and stammers as well as any other breaks. The grey circles show the predictions of oligomeric states with from left to right: MultiCoil or MultiCoil2 (only for the MultiCoil predictions), SCORER 2.0, ProCoil and LOGICOIL (LOGICOIL also predicts parallel or anti-parallel dimeric coiled coils, which are indicated by arrows here). The blue bars are sequence sliders for detailed helical wheel and helical net views and can be ignored here. The image has been generated with the Waggawagga tool (Simm et al., 2015).

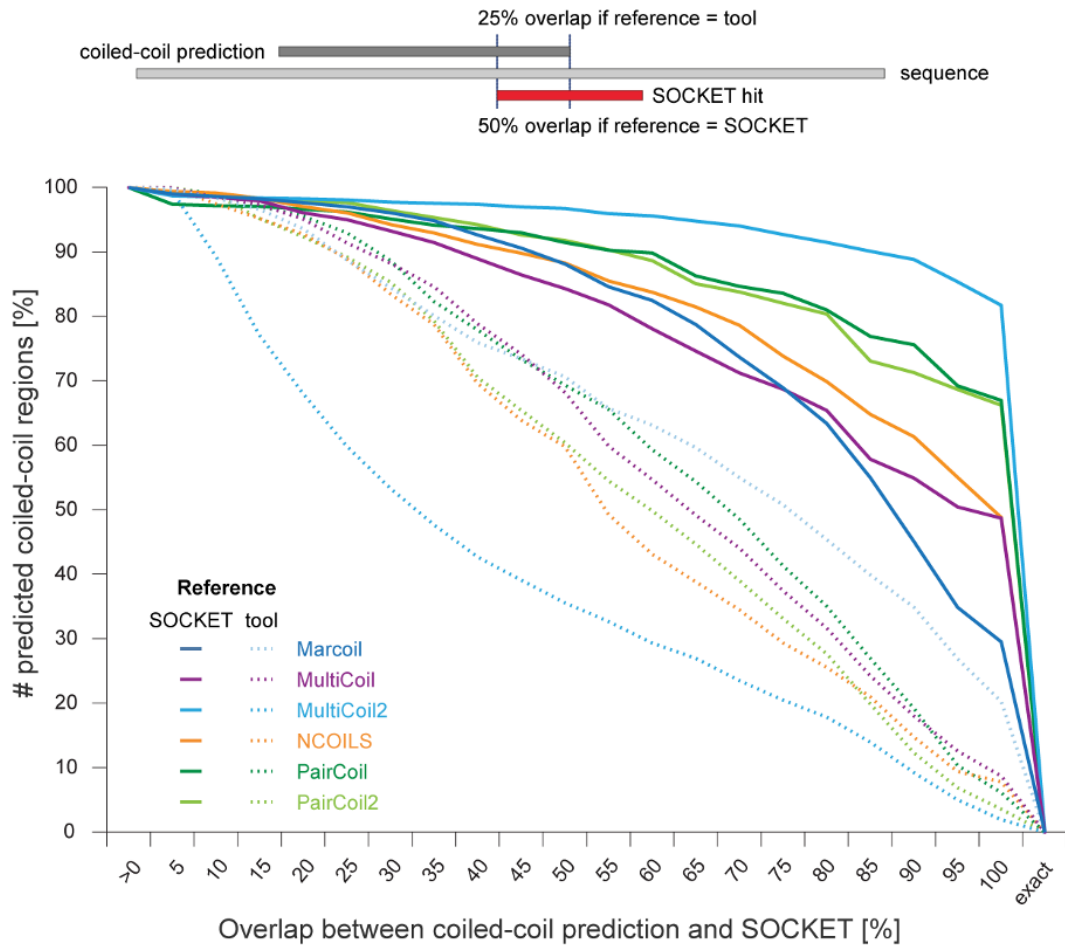


Figure A.6. | Overlap of coiled-coil predictions with SOCKET hit regions. Top: Schematic drawing of a coiled-coil prediction overlapping a SOCKET hit. The ratio of overlap between prediction and SOCKET hit is different depending on whether the prediction or the SOCKET hit is taken as reference. **Bottom:** The plot shows the percentage of coiled-coil predictions in dependence of the degree of overlap with the SOCKET hits. For each tool, the percentage of overlapping hits is counted once with taking the coiled-coil prediction as reference (solid lines) and once with the SOCKET hits as reference (dashed lines). The percentage of PDB files with predictions overlapping a SOCKET hit with at least a single amino acid was set to 100%.

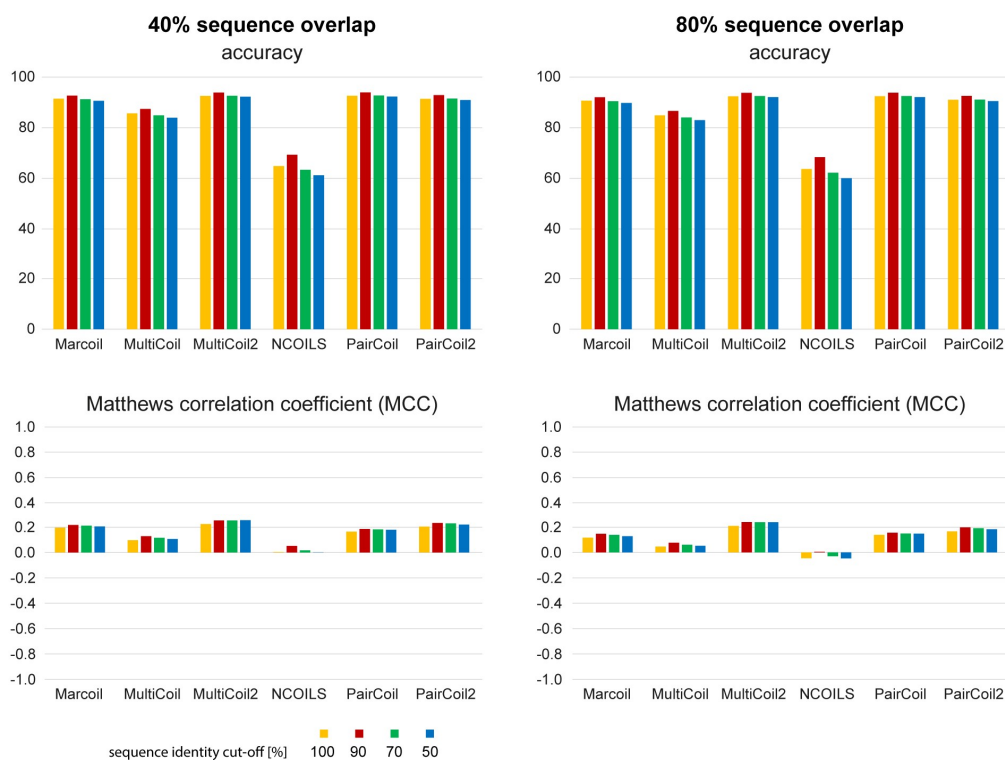


Figure A.7. | Performance of coiled-coil prediction tools in dependence of overlap with SOCKET reference. Accuracy and MCC are shown for the six coiled-coil prediction tools with respect to SOCKET coiled-coil identifications. Classification as true positive hit requires 40% (left plots) and 80% (right plots) overlap between SOCKET and prediction. Percentage overlap was determined with respect to the SOCKET hit.

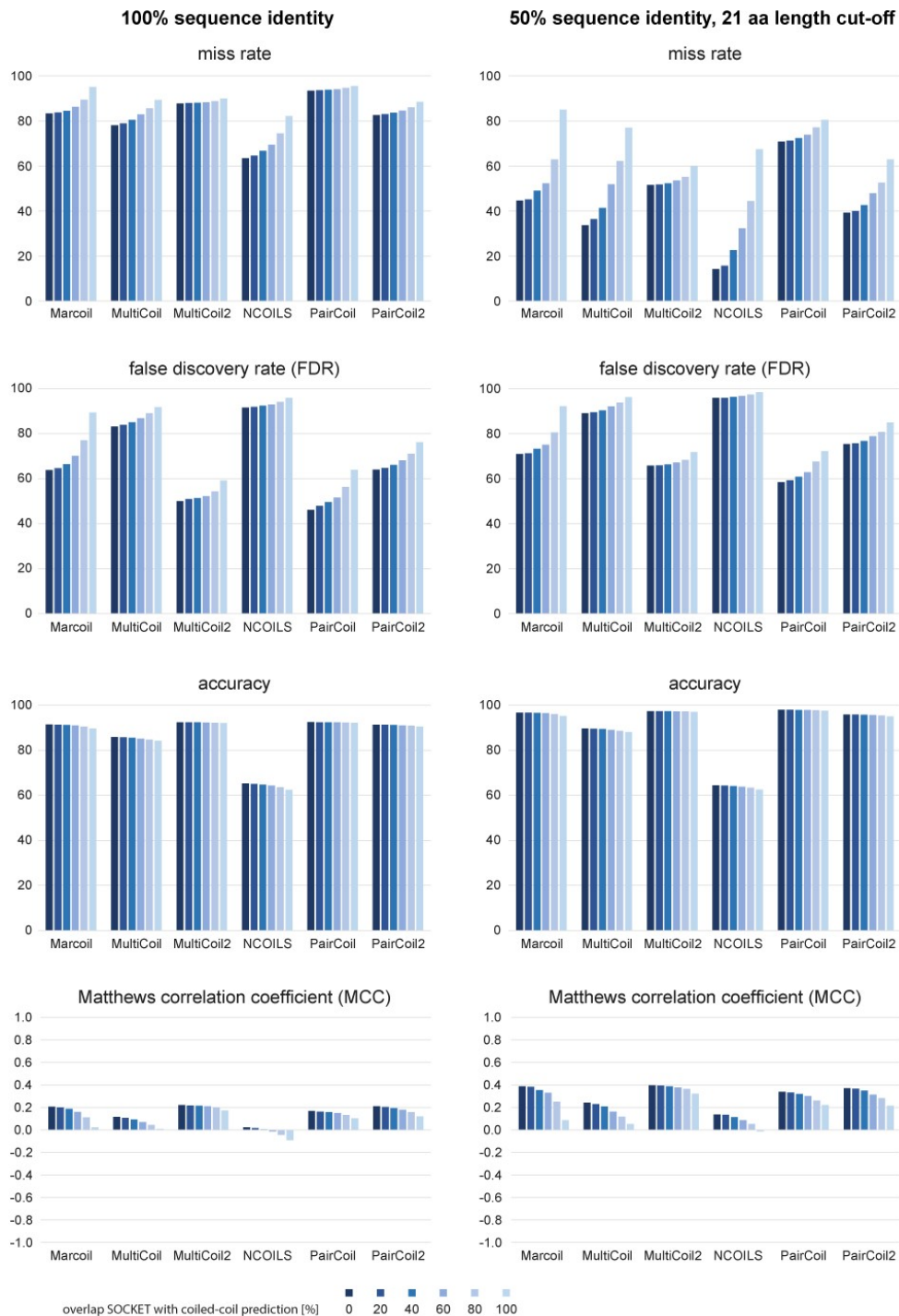


Figure A.8. | Performance of coiled-coil prediction tools in dependence of overlap with SOCKET reference. Several classification metrics are shown for six coiled-coil prediction tools with respect to SOCKET coiled-coil identifications. Classification as true positive hit requires overlap between prediction and SOCKET. The performance was analysed for two data sets. Plots on the left column: no filter. Plots on the right column: minimum length of 21 amino acid for coiled-coil regions and a sequence redundancy cut-off of 50%. For each data set, the metrics were computed for increasing percentages of overlap between prediction and reference. Percentage overlap was determined with respect to the SOCKET hit.

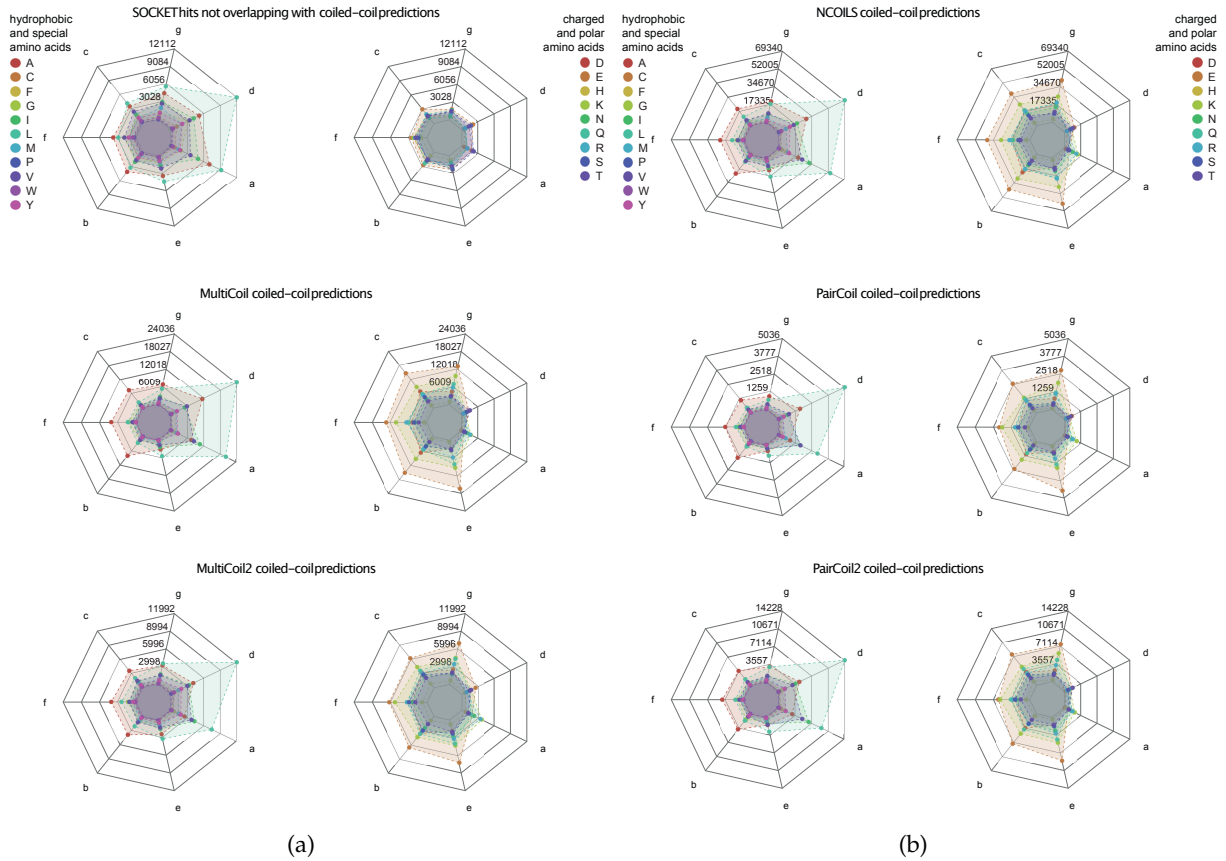


Figure A.9. | Amino acid preferences at heptad positions *abcdefg*. It is well known that hydrophobic amino acids are preferred at the interface between coiled α -helices, at positions *a* and *d*, and that charged and polar amino acids are preferred at the outside, especially at positions *e* and *g*. Because SOCKET not only detects "classical" coiled-coils but interacting α -helices within globular protein structures, the distribution of hydrophobic and charged amino acids is slightly less biased in the latter structures. The heptad patterns of the coiled-coils predictions show strong bias for leucine and isoleucine at the interior positions *a* and *d*, and for glutamate at all other positions, independently of the tool. The letters at the axes denote the heptad register positions. Data values at grid lines refer to amino acid counts at each heptad position over all heptads.

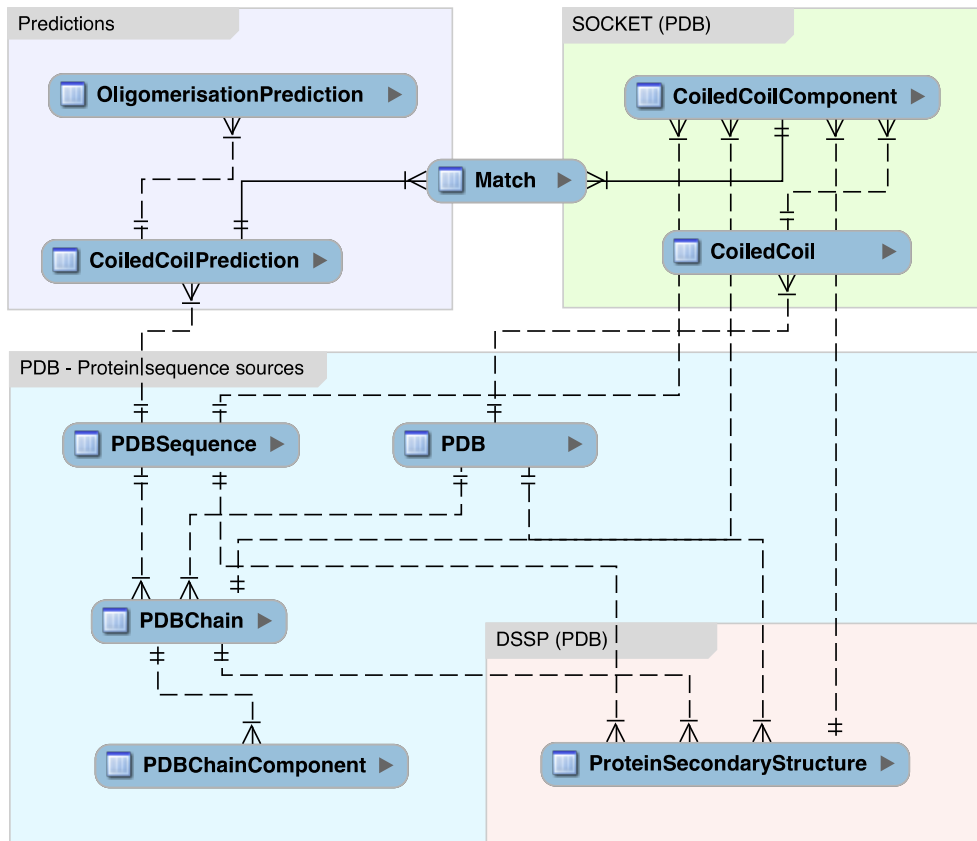


Figure A.10. | Simplified database scheme. Only the table names and connections are shown for clarity.

A. 4AU6, 3ZRY, 5J7V, 4N5C, 2R17, 4B2Q, 5HIU, 2WPD, 6B2Z, 6B8H, 6BA1, 6FF6, 4CFE, 4CBJ, 6AP4, 2XOK, 1I5N

B. Incomplete output

```

assigning heptad to helix 17 (Y) 60-77:D
extent of coiled coil packing: 63- 77:D
sequence TILQETHLXENLLDEAR
register abcdefgabcdfga
partner --!-!X--X--X----
knobtype --2--44--4--4----
repeats 0 non-canonical interrupts in 15 residues: 7,7,1

coiled coil 15:
angle between helices 0 and 1 is 167.145 antiparallel
angle between helices 0 and 3 is 161.145 antiparallel
angle between helices 1 and 2 is 162.695 antiparallel
angle between helices 2 and 1 is 162.695 antiparallel
angle between helices 2 and 3 is 146.836 antiparallel
angle between helices 2 and 4 is 23.195 parallel
angle between helices 3 and 2 is 146.836 antiparallel
angle between helices 3 and 4 is 149.257 antiparallel
this coiled coil is antiparallel
could not find a complementary knob to knob 23 in daisy chain 0
    
```

Correct output

```

assigning heptad to helix 2 (Y) 60-77:A
extent of coiled coil packing: 63- 77:A
sequence TILQETHLXENLLDEAR
register abcdefgabcdfga
partner --!-!X--X--X----
knobtype --2--44--4--4----
repeats 0 non-canonical interrupts in 15 residues: 7,7,1
1I5N_mm4.pdb c 7.00 e 0 result 1 COILED COILS PRESENT (+ 2 helix groups are either
pairs with too few complementary kno
b in hole interactions or are subsets of larger coiled coils)
Finished
    
```

C.

| Chain-B | aa 13-23 | aa 40-54 | aa 62-77 | aa 78-87 | aa 95-106 |
|------------|----------|----------|-------------------------|----------|-----------|
| SOCKET | Yes | Yes | Yes | No | Yes |
| MarCoil | FN | FN | FN | FN | FN |
| MultiCoil | FN | FN | FN | FP | FN |
| MultiCoil2 | FN | FN | FN | FN | FN |
| NCOILS | FN | FN | single long coiled coil | | |
| PairCoil | FN | FN | FN | FN | FN |
| PairCoil2 | FN | FN | FN | FN | FN |

Supplementary Figure S11. SOCKET coiled coils not present in the benchmark data due to SOCKET output problems and missing multimers.

A) PDB files containing SOCKET coiled coils where the output file of the SOCKET run is incomplete. B) Incomplete output of SOCKET run on 1I5N.pdb, and SOCKET run on one of

Figure A.11. | SOCKET coiled coils not present in the benchmark data due to SOCKET output problems and missing multimers. A) PDB files containing SOCKET coiled coils where the output file of the SOCKET run is incomplete. B) Incomplete output of SOCKET run on 1I5N.pdb, and SOCKET run on one of the generated "multimers" of 1I5N.pdb. C) Example of the performance of the prediction tools on a coiled coil missed in the benchmark data set, chain-B of 1I5N.pdb. 1I5N.pdb contains 4 chains of identical sequence but slightly different structure. For chain-A, a 5- stranded coiled coil is predicted by SOCKET, but the output is incomplete not providing the detected registers. Chain-B is shown in the figure. Chain-C is identical to chain-B. Chain-D is predicted by SOCKET to contain a 2-stranded coiled coil (aa 40-54 and aa 62-77).

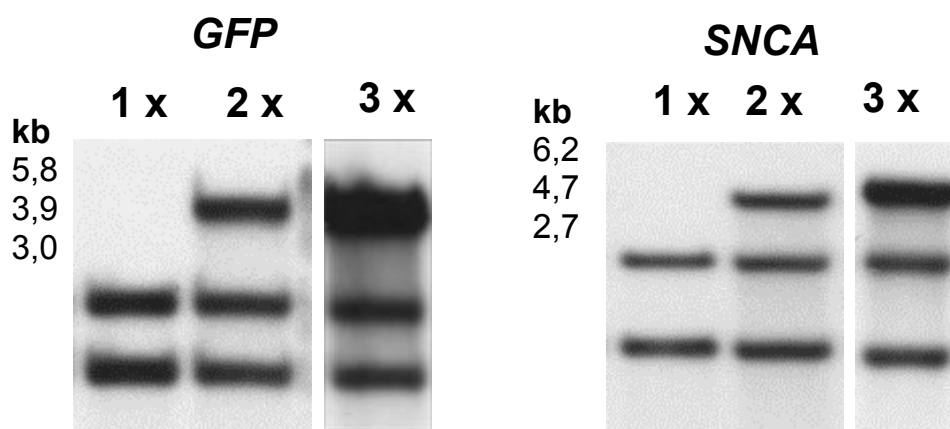


Figure A.12. | Southern hybridization and copy number determination. Yeast cells were transformed with integrative plasmids, harboring *GFP*- or α Syn-encoding genes, respectively. Multiple transformants were analyzed by Southern hybridization for verification of the integration of *GFP* or α Syn genes into the mutated genomic *ura3-52* locus using labeled *URA3* as a probe. Left plots: One copy (1x) of integrated *GFP* corresponds to 3.0 kb + 3.9 kb; two copies (2x) to 3.0 kb + 3.9 kb + 5.8 kb, and three copies (3x) to 3.0 kb + 3.9 kb + 5.8 kb (higher intensity). Right plots: One copy (1x) of integrated α Syn corresponds to 2.7 kb + 4.7 kb; two copies (2x) to 2.7 kb + 4.7 kb + 6.2 kb, and three copies (3x) to 2.7 kb + 4.7 kb + 6.2 kb (higher intensity), as indicated. Copy numbers were determined with the ImageJ software.

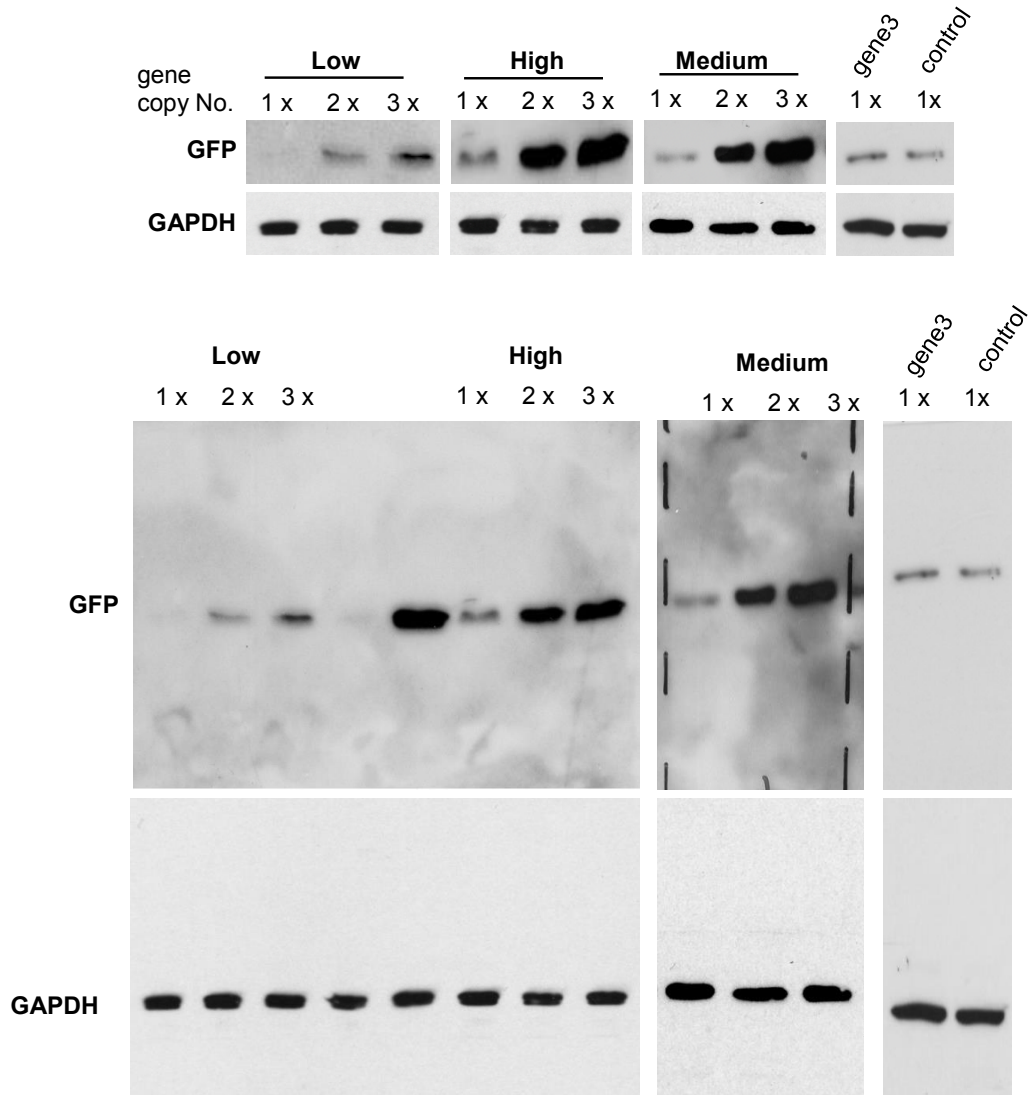


Figure A.13. | Full size blots corresponding to Figure 3.6A. On top, the part shown in Figure 3.6A. is presented for comparison. Western blot analysis of crude protein extracts from yeast strains, expressing *GAL1*-driven GFP from one, two and three copies. Protein expression was induced for 6 h in galactose-containing medium, crude protein extracts were prepared and equal protein amounts from all samples were used for Western blotting. The membrane was probed with anti-GFP antibody. GAPDH antibody was used as a loading control.

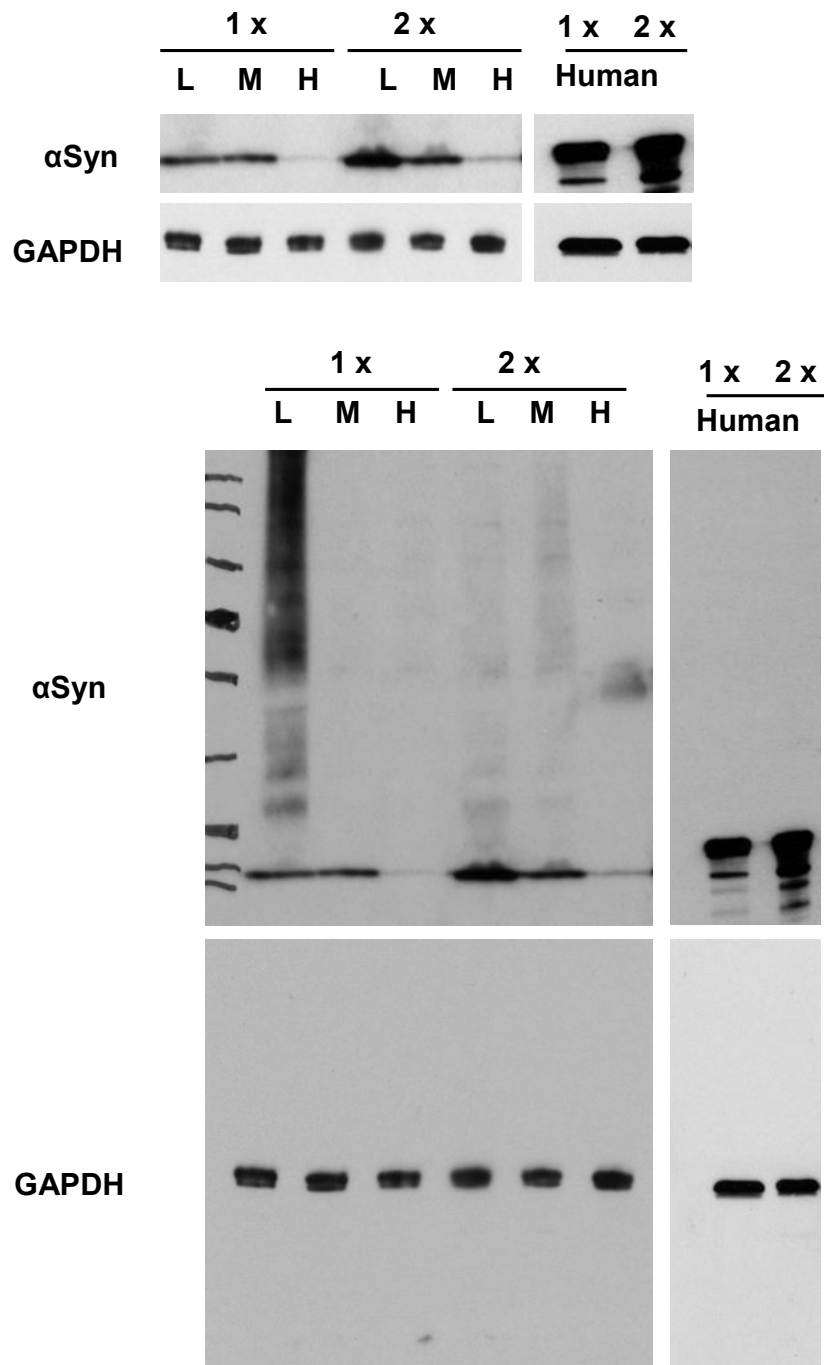


Figure A.14. | Full size blots corresponding to Figure 3.7A. On top, the part shown in Figure 3.7A is presented for comparison. Western blot analysis for determination of the protein level of α Syn. Protein expression was induced for 6 h, crude protein extracts were prepared and the protein concentrations were determined with a Bradford assay. 160 μ g crude protein extract from samples gene4 (L), gene5 (M) and gene6 (H), and 40 μ g from samples "human" were used for Western blotting. The membrane was probed with anti α Syn antibody. GAPDH antibody was used as a loading control.

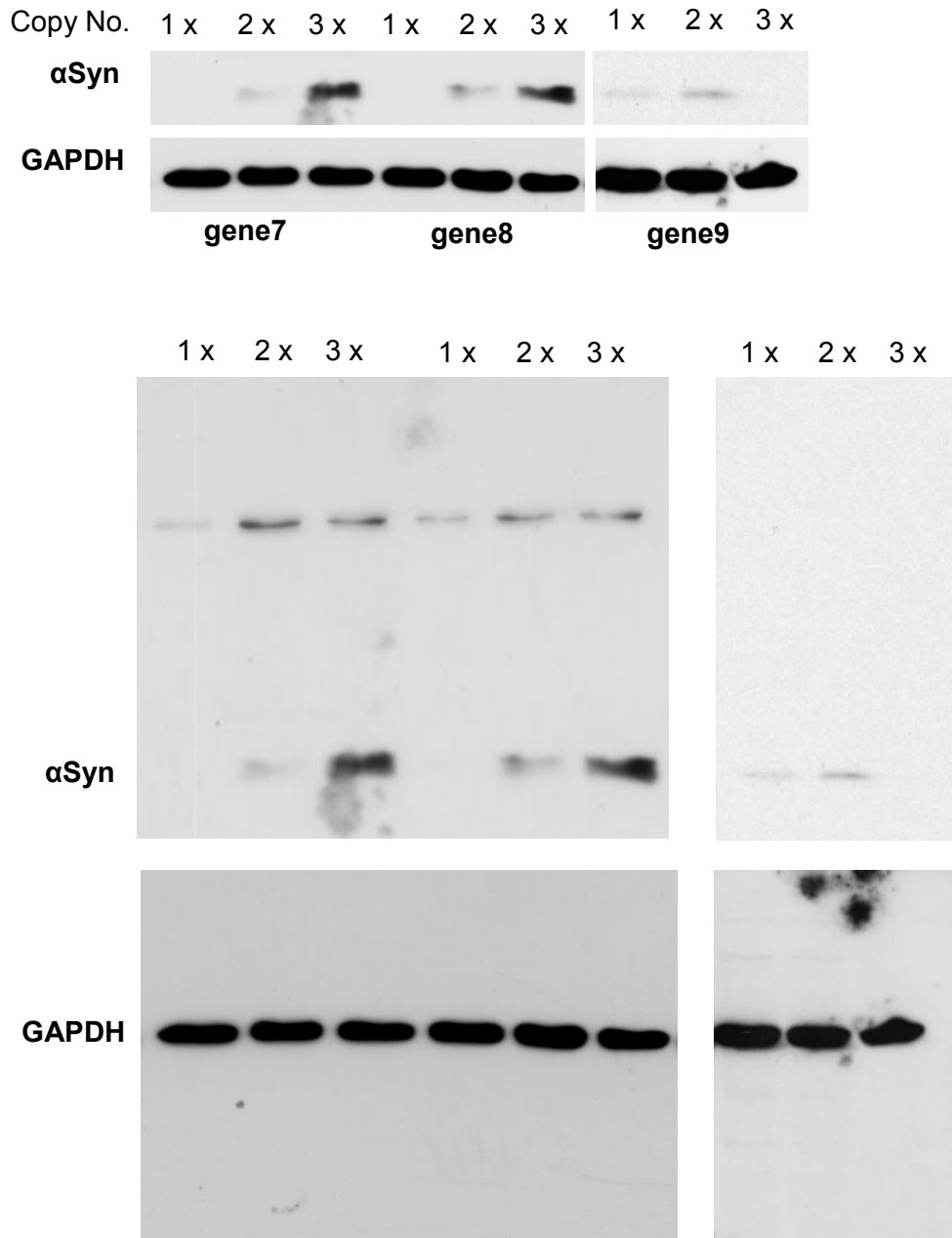
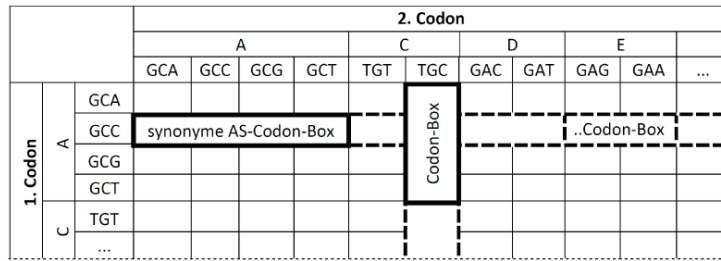
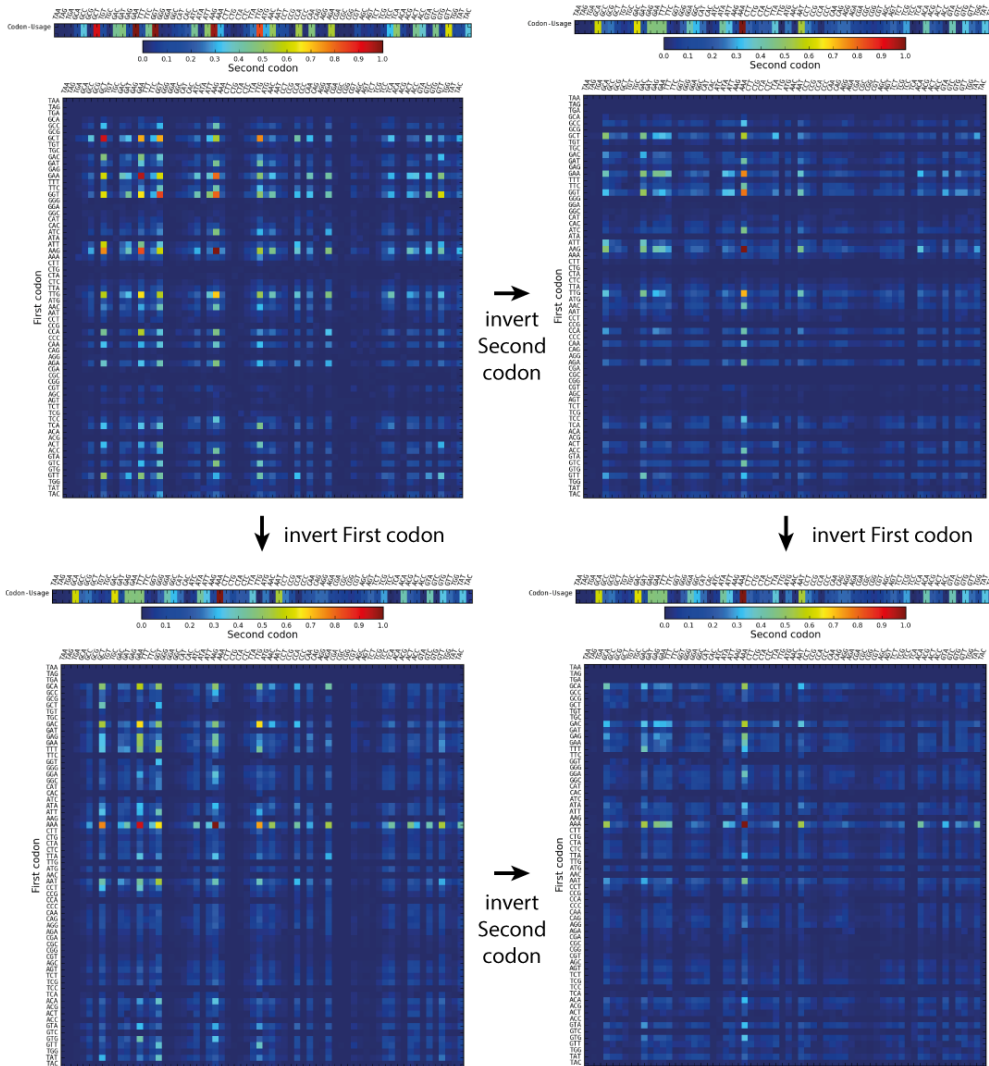


Figure A.15. | Full size blots corresponding to Figure 3.7C. On top, the part shown in Figure 3.7C is presented for comparison. Western blot analysis of crude protein extracts from yeast strains, expressing *GAL1*-driven α Syn from one, two and three copies. Protein expression was induced for 6 h, crude protein extracts were prepared and the protein concentrations were determined with a Bradford assay. 160 μ g crude protein extract from samples gene7, gene8 and gene9, and 40 μ g from samples "human" were used for Western blotting. The membrane was probed with anti- α Syn antibody. GAPDH antibody was used as a loading control.



RCU - highly expressed proteins, weighted



RCU - highly expressed proteins, weighted, inverted

Figure A.16. | The "inverted" di-codon usage. The schematic view at the top shows example codon boxes within the di-codon usage matrix. The heatmap plots at the bottom show the relative codon usage of the 308 highest expressed proteins of *S. cerevisiae*, when weighted (top-left), and the two ways to get the di-codon usage inverted (bottom-right).

A.3. Abbreviations and acronyms

| | |
|--|----|
| 3DEM 3D electron microscopy | 32 |
| AD Alzheimer disease | 16 |
| CC coiled coil | 17 |
| CCD coiled-coil domain | 17 |
| CNN convolutional neural network | 26 |
| cDNA complementary DNA | |
| DSSP Database of secondary structure assignments | |
| EMBL European Molecular Biology Laboratory | 9 |
| ET electron tomography | 32 |
| FAIR findability-accessibility-interoperability-reusability | |
| GFP green fluorescent protein | 12 |
| HMM hidden Markov model | 26 |
| HPC high performance computing | |
| IDP intrinsically disordered protein | 15 |
| KIH knobs-into-holes | 17 |
| MA Markovian adapter | 22 |
| MS mass spectrometry | 35 |

| | |
|--|----|
| MX macromolecular crystallography | 32 |
| NMR nuclear magnetic resonance spectroscopy | 32 |
| OS oligomerization state | 17 |
| PDB Protein Data Bank | 10 |
| PSSM position-specific scoring matrix..... | 26 |
| REBASE Restriction enzyme database..... | 36 |
| RCSB Research Collaboratory for Structural Bioinformatics | 33 |
| SAH stable single-alpha helix | 8 |
| SVM support vector machine | 26 |