Genomics and Phylogeny of Motor Proteins: Tools and Analyses

Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultäten der
Georg-August-Universität zu Göttingen

vorgelegt von
Florian Odronitz
aus Reutlingen

Göttingen 2007

D7

Referent:	Prof. Dr. Ralf Ficner
Korreferent:	Prof. Dr. Burkhard Morgenstern
Tag der mündlichen Prüfung:	23. Januar 2008

Leicht beieinander wohnen die Gedanken.
Doch hart im Raum stoßen sich die Sachen.

(Friedrich v. Schiller)

Contents

List of Figures

Part I

Introduction

1 Introduction

Genomes

Genomes are the blueprint of life. It has long been speculated that the information that organisms need to form there amazingly complex bodies are stored in a specific place inside the cells. With the discovery of DNA, this place has been identified (1). The characteristics of the DNA perfectly fit the requirements: Universal, robust, compact, mutable and open source.
With the advent of the genomic era, deciphering genomes in large numbers have become a possibility. What has been a major technical challenge before (2, 3, 4), is now only a question of resources and the number of sequenced genomes increases exponentially. Genome sequences are a very valuable resource for many types of research including molecular biology, phylogenomics, comparative genomics, functional genomics, metagenomics and pharmacogenomics.
Although a large number of genomes are technically available, there is no central authority that lists all projects and the species that have been sequenced. In order to remedy this problem we created diArk, a web application for completed sequencing projects of eukaryotic genomes (Chapter 3). diArk offers information about species and sequencing projects, alongside with literature references. It also offers sophisticated search options and provides a great number of genomes for download. diArk can be found at `http://www.diark.org`.

Genes

The most striking feature of genomes is that they encode proteins (5), the primary actors in cellular processes. Genes are the regions which code for protein. In contrast to prokaryotes, the genes of most eukaryotes are structurally complex. They comprise of regulatory elements (6), coding regions (exons) and non-coding regions (introns) (7). The length of the non-coding stretches vary considerably, ranging from dozens base pairs to many thousands.

Gene Annotation

Identifying the exact structure of a gene is important for a wide range of analyses and there have been numerous attempts to predict gene structures (8). Although undoubtedly useful on the large scale, existing programs are not well suited to optimise gene structures on the level of single bases. When searching for the gene structure, given a known protein sequence, it is desirable to have a program that does all possible refinement steps like finding small exons and optimizing splice sites since these

steps are very cumbersome when done manually and are often neglected. Numerous studies could benefit from precisely annotated genes since extensive studies were carried out based on incomplete data. In order to find the one most coherent gene structure given a protein query, we created Scipio (Chapter 6), a program that produces results that can be read by humans and computer programs alike. For users who prefer a more user-friendly way of using Scipio, we offer WebScipio (Chapter 7), a web application which enables the user to search for genes in the genomes of about 250 species on our server. The result can be viewed as clear tables and informative visualisations of the gene structure. The program has also proven very useful for cross-species annotations, which are getting increasingly useful as more and more genome sequences become available. WebScipio can accessed found at `http://www.webscipio.org`.

Splicing

As organisms got increasingly complex during evolution and adaptation, their protein repertoire got more diverse. Such diversification can be achieved by gene duplication, where one gene is copied and is then free to mutate and fill a new functional niche. But eukaryotic organisms can also increase the palette of their gene products by combining exons of a single gene in different ways by differentially splicing the pre-RNA (6). This sophisticated process enables the cell to assemble gene products in a modular way while using only minimal space in the genome. The decision on using either gene duplication or differential splicing in a certain gene family is a characteristic that is also acquired during evolution and therefore shared among closely related species. In the myosin gene family, the most extensive differential splicing is seen in the muscle myosin heavy chain genes of arthropods. The structures of these genes give interesting hints on how they may have evolved. On rare occasions, mRNAs find their way back into the genome, providing a snapshot of their momentary sequence and giving rise to pseudogenes. Close inspection of pseudogenes can reveal some details about the process of mRNA splicing (Chapter 5).

Protein Families

Genomes of different species contain homologous proteins, that are similar in sequence, structure and function. A large number of protein families have been identified (9). Since protein families evolve and diversify and can have a great number of members, they are very good subjects to study evolutionary processes (10). The myosin protein family is one of the families with representatives in virtually all eukaryote genomes. Since myosins are involved in several essential processes in the cell (11), all of them are highly conserved. On the other hand, they have seen great diversification during evolution and many classes of myosin are specific for a few taxa. These characteristics make them an ideal candidate for the reconstruction of the tree of eukaryotic life (Chapter 4). In our study we used 2269 manually annotated protein sequences and were able to greatly extend the existing classification system of the myosins and shed some light on disputed parts of the tree of life.

Apart from myosin, eukaryotic cells have other motor proteins with specific functions: Kinesin and dyneins. When combining the evolutionary information in all families of motor proteins, phylogenetic relationships can be resolved with high confidence and in great detail. In large protein families which are structured internally and can be split into classes and variants, the existence and non-existence of variants can be used to cross-check the findings of more traditional studies. We used this combined approach on the Arthropoda taxon and were able to precisely determine the phylogenetic relationship of 21 completely sequenced species (Chapter 8).
When dealing with the many members of a protein family, it is important to not only consider their sequence characteristics for categorization. Alongside information on the domain composition, the species and their taxonomy and the relevant literature should be taken into account. In order to track this information and to make it available, we created CyMoBase (Chapter 2), a web application that stores information about more that 8000 manually annotated protein sequences, more than 960 species and more than 750 publications. The database can be searched conveniently and sophisticated queries can be constructed and saved using a modular search system. CyMoBase can be found at `http://www.cymobase.org`.

Solid state NMR

Since nuclear magnetic resonance has been discovered (12), it has been performed both in liquid and in solid phase. Solution state NMR has for a longer time been successfully used for structure determination of biological macromolecules (13). However, in recent years, solid-state NMR has also been successfully used to determine protein structures (14, 15). Especially as the proteins of interest get larger, elucidating their structure becomes a very complex task that is hampered by the limited resolution, resonance overlap and chemical shift ambiguity. For this process, predicted spectra can be of great help, although existing solutions for these predictions are very limited and can hardly be adapted to changed experimental parameters.
Since this problem can be solved using the same technologies used in the other projects, we decided to create Peakr, a software program that can efficiently predict spectra for all common experimental settings that are used in protein solid state NMR and is able to handle complicated cases like different conformations and inter-molecular interactions in crystals (Chapter 9). Peakr offers an intuitive web interface and can also be used as a web service.

Note

The publications and manuscripts are ordered chronologically.

Part II

Publications

2 Pfarao: a web application for protein family analysis customized for cytoskeletal and motor proteins (CyMoBase)

Florian Odronitz,[1] and Martin Kollmar [1*]

[1]Department of NMR-based Structural Biology, Max-Planck-Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Goettingen, Germany
[*]Corresponding author.

BMC Genomics
Published: 29 November 2006
BMC Genomics 2006, 7:300 doi:10.1186/1471-2164-7-300 This article is available from: `http://www.biomedcentral.com/1471-2164/7/300`

2.1 Abstract

2.1.1 Background

Annotation of protein sequences of eukaryotic organisms is crucial for the understanding of their function in the cell. Manual annotation is still by far the most accurate way to correctly predict genes. The classification of protein sequences, their phylogenetic relation and the assignment of function involves information from various sources. This often leads to a collection of heterogeneous data, which is hard to track. Cytoskeletal and motor proteins consist of large and diverse superfamilies comprising up to several dozen members per organism. Up to date there is no integrated tool available to assist in the manual large-scale comparative genomic analysis of protein families.

2.1.2 Description

Pfarao (Protein Family Application for Retrieval, Analysis and Organisation) is a database driven online working environment for the analysis of manually annotated protein sequences and their relationship. Currently, the system can store and interrelate a wide range of information about protein

sequences, species, phylogenetic relations and sequencing projects as well as links to literature and domain predictions. Sequences can be imported from multiple sequence alignments that are generated during the annotation process. A web interface allows to conveniently browse the database and to compile tabular and graphical summaries of its content.

2.1.3 Conclusions

We implemented a protein sequence-centric web application to store, organize, interrelate, and present heterogeneous data that is generated in manual genome annotation and comparative genomics. The application has been developed for the analysis of cytoskeletal and motor proteins (CyMoBase) but can easily be adapted for any protein.

2.2 Background

The success of the genome sequencing projects have culminated in release 149 of GenBank (16) that announced two milestones: the total sequence data passed the 100 gigabases mark, and, for the first time, the number of bases derived from whole genome shotgun sequencing projects exceeded the number of bases in the traditional divisions of GenBank. However, the process of genome annotation still lags considerably behind that of genome data generation. Although many tools have been developed for the ab initio annotation of whole genomes, especially the annotation of data from higher eukaryotes yields low success rates (17). The success rates can considerably be increased by similarity searches of EST data or of annotated data from other genomes. But also these data have their drawbacks: ESTs are fragmentary and might suffer from several artefacts including contamination with genomic DNA; similarities to proteins in other species might suffer from evolutionary divergence or the orthologue-paralogue problem (18); and the presence of alternative splicing considerably complicates the interpretation of alignments between genomic DNA, cDNAs and ESTs. More seriously, however, similarity data is never complete. But it is the annotation that connects the sequence to the biology of the organism (19).

Manual annotation is still by far the most accurate and successful way to achieve correct predictions of genes. This process is best done using the possibilities of comparative genomics and multiple sequence alignments. Because a majority of the proteins are not characterized and their functions are largely unknown, the initial process involves categorizing these predicted proteins into subsets of proteins or protein families based on homology, presence of various functional domains and motifs, as well as similarity to well characterized proteins from other species.

Thus, when working with collections of protein-sequences from different species and sources, one quickly accumulates large amounts of heterogeneous data: Protein and DNA sequences, their identifiers in different databases, references to literature, information about species including taxonomy, and links to online resources like sequencing projects. Since data that can be retrieved from public

databases is often incomplete or incorrect it is very desirable to be able to combine manually edited with automatically generated content. In addition, there is often misleading and contradicting data, especially concerning the nomenclature and classification of proteins, that needs to be tracked and commented.

Cytoskeletal and motor proteins have extensively been studied in the past. They are involved in diverse processes like cell division (20), cellular transport (21), neuronal transport processes (22), or muscle contraction (23), to name a few. Especially motor proteins consist of large superfamilies. E.g. vertebrates contain up to 60 myosins and about the same number of kinesins that are spread over more that a dozen distinct classes. Since genome sequence data is rapidly accumulating it is very important to have a reference database for the nomenclature and phylogenetic relation of the proteins that allows the most accurate assignment of biological function possible.

Pfarao is a database driven web application that was written to assist researchers investigating structure, function and phylogeny of proteins. It has been developed for the analysis of cytoskeletal and motor proteins (CyMoBase), but can be adapted to any type of protein. It stores, organizes, interrelates, presents, and analyzes data of various sources. Additionally, it triggers external prediction programs, so that manually entered and automatically generated data is always synchronized.

2.3 Construction

2.3.1 Technologies

The system is running on UNIX (OS X and Linux) systems. The database management system is PostgreSQL (24). As web application framework we chose Ruby on Rails (25) since it has the advantage of rapid and agile development while keeping the code well organized. Part of this framework is an implementation of Active Record (26) which is an O/RM (Object-relational Mapping) system that makes database integration into an object oriented program considerably easier. This also allows to use the interactive ruby shell (irb) with database rows wrapped in objects for interaction with the database. This way of accessing the data often proves superior to the SQL shell. Additionally, Ruby on Rails offers XML-RPC so data can be accessed by other programs.
We implemented a service-oriented mechanism that starts specific scripts, when records in the database are added or updated. In this case, a PostgreSQL trigger starts a PL/Ruby script (27), which opens a network connection to a delegation server program written in Distributed Ruby (28) on the same machine and calls one of its functions, giving a database ID as an argument if appropriate. The server can in turn start scripts to act upon the entered or updated data and returns after completion so that the database transaction is completed. The server's state can be set from within the database or from external programs to disable certain functions during batch processing in order to avoid flooding.

The automation scripts for parsing BLAST (29) and HMMER (30) output are written in Ruby (31) making use of the BioRuby library (32). Sequences are scanned for domains using the Pfam_fs release 19.0 database (33) containing 8183 hidden markov models.

The web pages are generated as XML (XHTML with SVG (34) data islands). We used SVG (34) for charts because of the high display quality and the possibility of reuse in print. The site makes extensive use of Ajax (Asynchronous JavaScript and XML) in order to present the user with a feature rich interface while minimizing the amount of transferred data. All technologies used are freely available and open source.

2.3.2 Database

The unique requirements of the system demand a custom database schema. The schema is sequence-centric with an additional emphasis on species since these two aspects are the most important in mutual annotation and, therefore, need to be represented in high detail (Figure 2.4). Grouped around these central tables are tables for literature and sequencing projects as well as taxonomy and predicted domains.
The sequence table stores the protein sequence and the corresponding sequence as derived from the multiple sequence alignment of the protein (see Import/Export). By relating a position in the alignment to the positions in a set of protein sequences it is possible to retrieve homologous stretches from different sequences. In addition there are fields for sequence classification and nomenclature, comments, legacy names, information about the completeness of the sequence, its potential to be a pseudo-gene, and links to records in NCBI's nucleotide and protein databases (35). The comment field is one of the most important fields intended to contain information about differences of the database sequences to published sequences that may have resulted from wrong exon predictions or sequencing errors. Records in the sequence table are related to tables for proteins, species, and publications.

Several versions can be assigned to each sequence so changes and corrections can be tracked as more information becomes available. Furthermore, there are links to tables containing automatically generated protein domain predictions (see Automated processes).

Species are defined by a set of names. There are fields for the scientific name of a species, the species abbreviation as used to identify database sequences, and common names. As some species are known by different scientific names, fields containing alternatively used names are also included. To account for the different usage of the scientific names, all possible names are listed and linked to the corresponding reference record wherever species are listed or used for selection via the interface. A comment field may contain general information about the corresponding species, the specific strain used, or common and divergent features compared to closely related organisms. The taxonomy field

is converted automatically into a hierarchical representation of the taxa. (see Automated processes)

Proteins are stored with their name and abbreviation as used in the database. Furthermore, classes of a certain protein can be grouped and categorized according to aspects like cellular function or localization. The project table includes information about the sequencing centres including type of data and completeness. Publications can be related either to a sequence to provide additional links to biological information or to a sequencing project.

Data entry is done using the iiwi system (Odronitz F., Lampetsdoerfer T., Dietrich D., unpublished results (36)) allowing for remote editing and access control.

2.3.3 Automated processes

The database can trigger external programs upon insertion or update of certain fields in the database tables by contacting the delegation server program, which can in turn write computed data to the database (Figure 2.1). When a protein sequence is inserted or changed a hmmpfam (30) process is started scanning this sequence for putative domains with Pfam (33) profiles. The obtained domain identifier and the start and end positions together with the E-value are stored in a database table. Upon insertion of a new species record, the content of the taxonomy field is automatically converted into a tree-representation of interrelated taxon records. Each record contains the name of the taxon, and a reference to the parent taxon. Then the species record is connected to the common taxonomic tree. This tree representation of the taxonomy allows for convenient searching, browsing and selection of sub-trees (Figure 2.2).

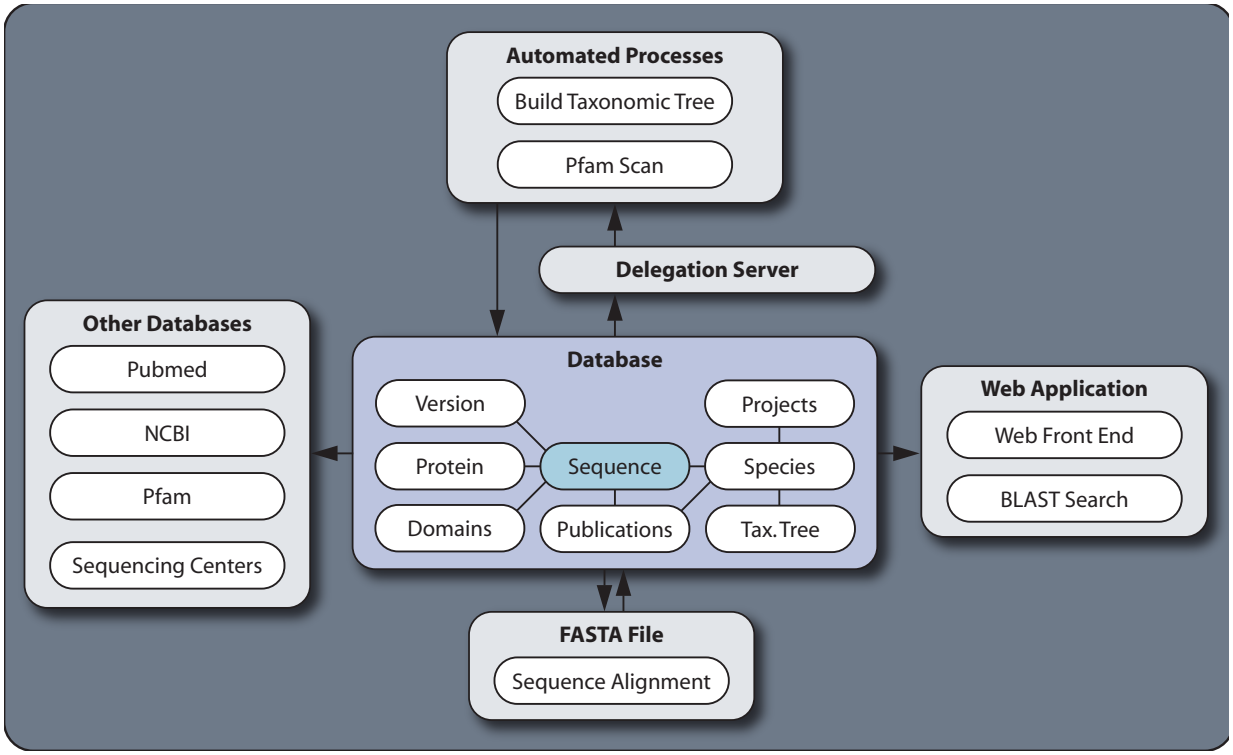

Figure 2.1: Diagram of main tables and linked resources.
The Database (blue) with the central sequence table (cyan), important associated tables (white) and connected systems are shown. Lines connecting tables depict table relations. Arrows depict flow of information. FASTA files containing sequence alignments are imported and exported using Ruby (31) scripts. The import function uses the BioRuby (32) library. Other databases are referenced via their IDs, which are used to generate hyperlinks to records on their web sites. Automated processes are started by a delegation server, which receives instructions from the database on insert or update of records. The automated processes write information into the database using Active Record. The frontend is generated using the Ruby on Rails (25) web application framework.

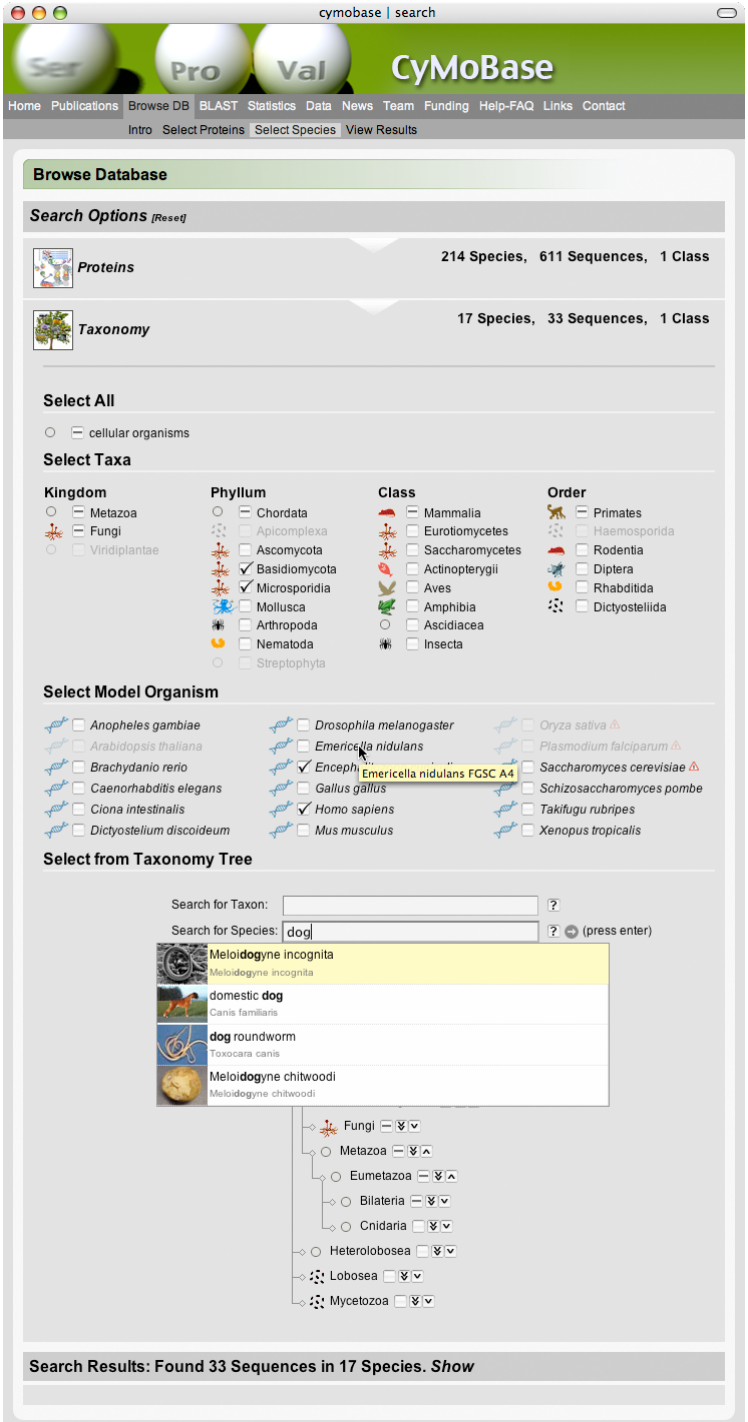

Figure 2.2: Screenshot of the species selection interface.
The user can select all species or a subset of species. Taxa and species for which no sequences for the selected proteins/protein classes exist are greyed out (taxa selection, model organisms) or are invisible (tree). Each node of the tree can be expanded and collapsed. The auto-completion fields open and highlight the tree down to the taxon/species typed. Common names like d̀ogáre also supported. All sections of the page respond to changes. Example: Nothing is selected. User selects kingdom Fungi. This selects all phyla, classes, orders, species and model organisms that belong to Fungi. Also the portion of the tree below Fungi is selected. User deselects Ascomycota. All elements react accordingly. User selects *Homo sapiens*. User clicks Śhowín the result section and is presented with a list of sequences from *Homo sapiens* and all Fungi, excluding Ascomycota (see Figure 2.3).

2.3.4 Import/export functions

Files containing protein sequences in FASTA-format can be imported into the database to update existing or insert new records in the sequence table (Figure 2.1). A naming convention at all levels ensures the correct assignment of sequences in a FASTA file to sequence records in the database. The sequence identifiers are a concatenation of species name abbreviation, protein name abbreviation, protein class and protein variant. In contrast to the usage of numerical database IDs, the naming convention thus immediately provides the user with information about the phylogenetic relation and possible functions of the protein. Sequences and sequence alignments can be exported from the database using filters to include only certain proteins, protein classes, or sequences from species in certain taxa. The resulting FASTA file also follows the naming convention and therefore can be re-imported after editing. Thus it is possible to retrieve a multiple sequence alignment from the database, edit it manually and write it back to the database. During import, sequences with identifiers that do not match any record in the database, induce the creation of a new database record according to the information included in the identifier.

2.4 Utility and discussion

The requirements for Pfarao can be summarized as follows: The key component of the database is the protein sequence that is obtained by manual annotation of genome and EST data with the help of a multiple sequence alignment. The sequence needs to be connected to data that allows the useful interpretation of the results concerning its biological function, and it needs to be linked with primary databases like GenBank or PubMed. To be useful for the specific protein community, whose members are expected to work in all biological and medical sub disciplines, the information of the database has to be presented in the most comprehensible way.

2.4.1 Web interface

Great attention has been paid to a versatile yet easy to use web interface. We think that accessibility and high quality representation is key to a productive usage of the system. Data can be entered and edited using a series of forms and lists. Relations are represented as pull-down menus.
Pfarao encompasses a live web front end that is generated from the content of the database at each request and thus always reflects the current data, eliminating the need for manual updates. To browse the content of the database, the user selects a set of proteins and protein classes, and is then guided to refine the selection by choosing a set of specific taxa or species. Taxa and species can either be selected from tables containing specific subsets, or from a tree representation of the taxa and species that is generated to match the protein and protein class selection. Taxa and species can be browsed and selected by expanding/collapsing and including/excluding subsections of the tree, or by using shortcuts or auto-completion fields (Figure 2.2). We consider the selection of specific species and taxa a key feature for comparative analyses of protein inventories and diversity (Figure 2.2).

Upon confirmation of the selection of protein and species, the system compiles a list of all sequences matching the specified criteria and presents it as a list grouped by species in taxonomic order. Additional data about the species like alternative names, links to sequencing centres and publications, as well as detailed information about the sequences including publications, comments, domain organization, and the sequence data, can selectively be shown or hidden (Figure 2.3).

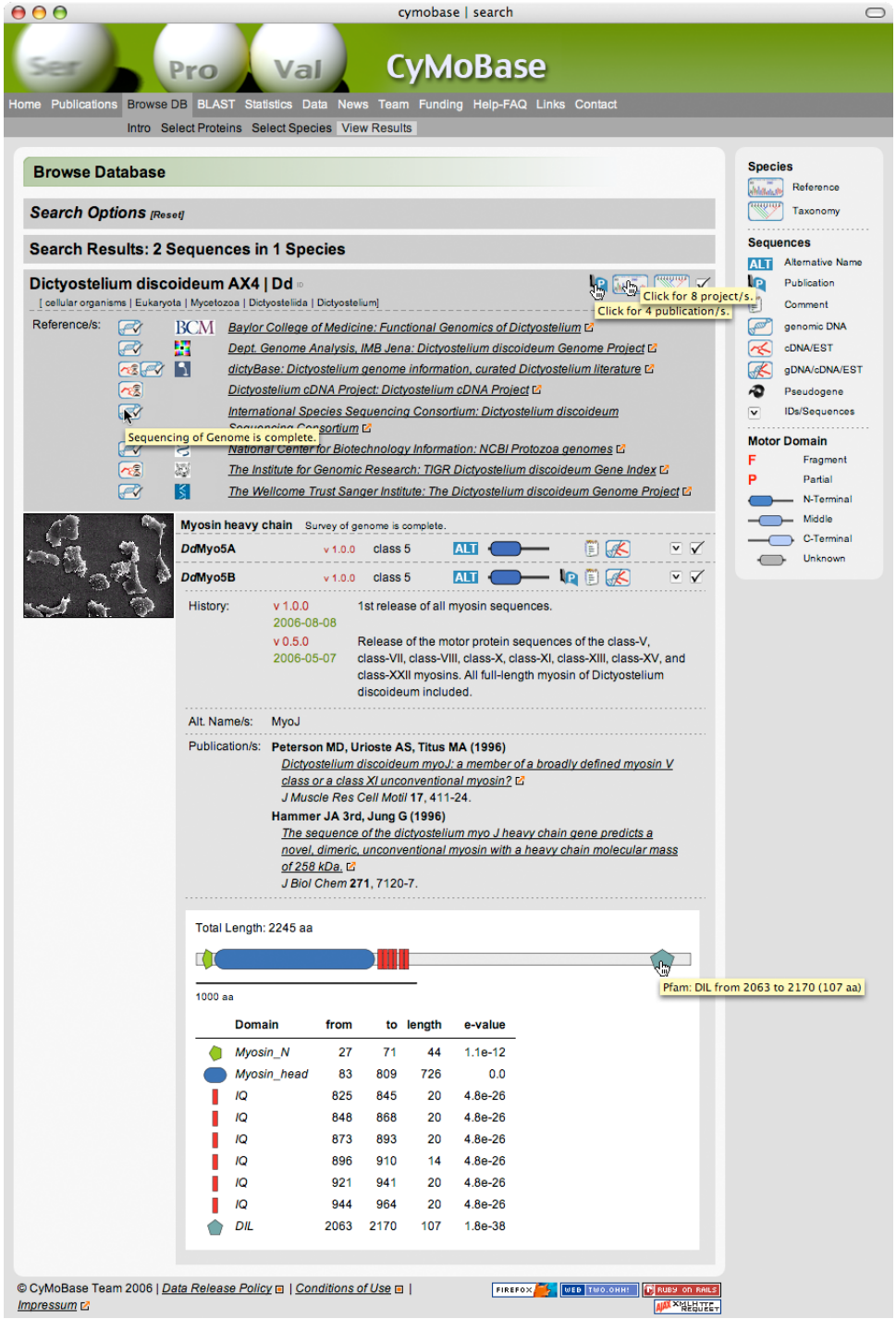

Figure 2.3: Screenshot of the protein sequence view.
The list is grouped by species. Sequences are ordered by the protein name. Different types of information are available for each species (publications, references to sequencing projects, taxonomy and name information) and each protein sequence (version history, alternative names, domain composition, publications, comments, source, amino acid sequence, links to other databases). All the details can be shown (and hidden) selectively. This way, even long lists can be viewed without cluttering the page. The data is retrieved on demand from the server via Ajax and does not have to be downloaded to the user's computer if not needed. Cursor labels provide the user with a short summary of the information behind the icons. A click shows the complete information. Additional cursors added to the figure to show cursor labels.

he system provides an integrated BLASTP (29) search and is able to link the sequences in the BLAST database with the records in the SQL database via an ID. Thus the user can, apart from the sequence, immediately access all related information. The organization of the database lends itself to different types of statistical analysis. For each protein, a set of tables and graphs can be generated. These analyses provide important information for the comparison of the protein inventory of specific taxa and species, as well as important insights into the selected protein superfamily. The protein inventory table gives an overview about the class distribution and the number of class members of all or a number of selected species (ordered by taxonomy). Color-coding of the cells helps to quickly identify characteristic patterns of specific taxa. Charts show the ratio of protein classes and the distribution of the molecular weight for a chosen set of classes. All charts are generated on the fly in resolution-independent SVG-code, so they can also be used for print.

2.4.2 Future developments

Pfarao provides a solid platform for additional features and significant future developments of the system are underway. The front end will be extended to allow the graphical representation and fast browsing of large alignments of selected sequences that will be of great value for mutational studies. The interface is also intended to support the generation of phylogenetic trees for a user-defined set of sequences. These extensions will increase the transparency of the manual annotation process, as the user will be able to look at the two basic sources of information about protein sequence relations. It is also planned to incorporate the corresponding DNA data and to track the various alternative splice forms of the proteins.

2.4.3 Case study

Pfarao has initially been developed for cytoskeletal and motor proteins but can easily been adapted to any protein. The database for cytoskeletal and motor proteins is called CyMoBase (37). Our current in house database contains 3265 Sequences (3095759 amino acids) from 666 species, 494 publications, and 385 references to 165 sequencing projects but is being extended on a daily basis. A portion of the data has been released in the publicly available CyMoBase.

2.5 Conclusion

Here, we introduce a web application for the analysis of proteins from manual annotation and their relationship. The major motivation for this work was to provide an integrated environment that organizes and relates all relevant information and presents it using a high quality interface. Pfarao is a tool that allows the researcher to constantly monitor the state of the work without having to manually aggregate data from a range of sources. It has been developed for the analysis of cytoskeletal and motor proteins (CyMoBase) but can easily be customized for any type of protein.

2.6 Availability and requirements

CyMoBase can be accessed at `http://www.cymobase.org/`.
Due to the technologies used, it requires Firefox version 1.5 or greater with cookies and JavaScript enabled. Other browsers do not have the required feature set or do not comply with the standards of the W3C (34). The database schema, the web application, the server program and all scripts can be obtained upon request and used under a Creative Commons License. Use of Pfarao by non-academics requires permission.

2.7 Authors' contributions

MK specified the requirements from a users perspective, defined the rules for data handling and participated in the design of the interface. He collected all the data and evaluated every function of the system. FO carried out the implementation of the system, designed the database scheme and did the technical design and the programming. Both authors wrote and approved the final manuscript.

2.8 Supplementary Material

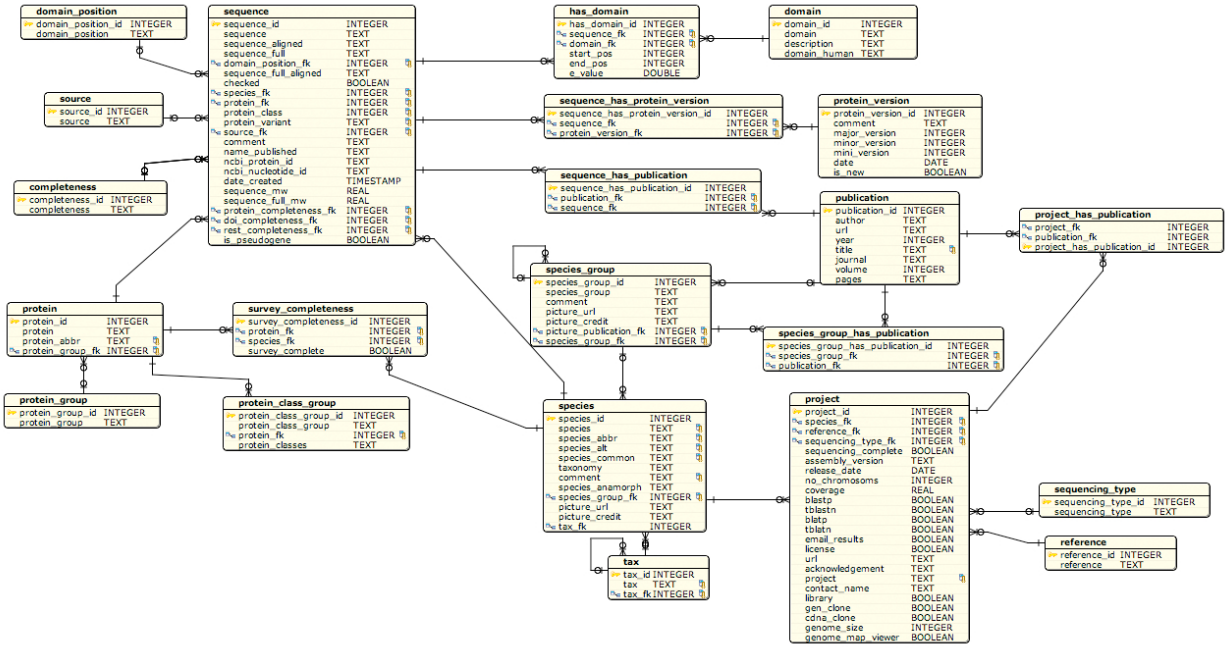

Figure 2.4: Database schema.
The schema shows the database tables and their relations. For each table the columns are listed with their name and datatype. Yellow keys in front of the names signify columns with unique identifiers. Blue window-symbols mark foreign key columns that contain values of id-columns of other tables. Symbols at the right side of the column names designate indices for better performance. Lines are relations between tables. Two unary (recursive) relationships are defined: One linking taxa to their parent taxon and one linking species groups to their parent group.

2.9 Acknowledgements

M.K. is supported by a Liebig Stipendium of the Fonds der Chemischen Industrie, which is in part financed by the BMBF. This work has been funded by grant I80798 of the VolkswagenStiftung.

3 diArk - a resource for eukaryotic genome research

Florian Odronitz,[1] Marcel Hellkamp[1] and Martin Kollmar [1*]

[1]Department of NMR-based Structural Biology, Max-Planck-Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Goettingen, Germany
[*]Corresponding author.

BMC Genomics Highly accessed
Published: 17 April 2007
BMC Genomics 2007, 8:103 doi:10.1186/1471-2164-8-103 This article is available from: `http://www.biomedcentral.com/1471-2164/8/103`

3.1 Abstract

3.1.1 Background

The number of completed eukaryotic genome sequences and cDNA projects has increased exponentially in the past few years although most of them have not been published yet. In addition, many microarray analyses yielded thousands of sequenced EST and cDNA clones. For the researcher interested in single gene analyses (from a phylogenetic, a structural biology or other perspective) it is therefore important to have up-to-date knowledge about the various resources providing primary data.

3.1.2 Description

The database is built around 3 central tables: species, sequencing projects and publications. The species table contains commonly and alternatively used scientific names, common names and the complete taxonomic information. For projects the sequence type and links to species project websites and species homepages are stored. All publications are linked to projects. The web-interface provides comprehensive search modules with detailed options and three different views of the selected

data. We have especially focused on developing an elaborate taxonomic tree search tool that allows the user to instantaneously identify e.g. the closest relative to the organism of interest.

3.1.3 Conclusions

We have developed a database, called diArk, to store, organize, and present the most relevant information about completed genome projects and EST/cDNA data from eukaryotes. Currently, diArk provides information about 415 eukaryotes, 823 sequencing projects, and 248 publications.

3.2 Background

Since the publication of the first complete genome sequence of an eukaryote, *Saccharomyces cerevisiae* (3), the genome sequencing community has produced highly advanced drafts of many other eukaryotes. The past few years have thus seen the rise of a completely new field in biology that is described as comparative genomics (38). Initial results have shown that whole genome comparisons are important to improve the annotation of genes and transcripts of a genome. It has also been demonstrated that not only genome sequences of organisms spread over all kingdoms of eukaryotic life are needed but also many of closely related organisms (39). These results have lead to the Fungi genome initiative representing the widest sampling of genomes from any eukaryotic kingdom, the mammalian genome project aimed to expand the genome coverage of mammals, and the *Drosophila* species sequencing project intended to establish methods for comparative genomics among other things. Thus, it is evident that future sequencing efforts have to include both further taxonomic sampling and closely related organisms.

In many research areas it is important to have access to DNA data and DNA samples of as many organisms as possible. For example, in structural biology there is a strong tendency to also work with homologs of other organisms to enhance the chance of obtaining structural data because cloning and protein expression are not as time consuming as they were some years ago (40). Reconstructing phylogenetic relationships between species or proteins is another expanding topic and it is clear that the addition of further sequence data improves the significance of the analyses by enhancing the statistics and therefore limiting the negative effects of outliers (41).

Two main databases provide access to lists of completed and ongoing eukaryotic genome projects. The Genomes OnLine Database (GOLD (42)) presents information on sequencing projects sorted according to the three major lineages of the tree of life. In addition, GOLD distinguishes between published and ongoing projects but lists some of the completed and not yet published genomes with the published projects. GOLD also contains some limited information about genome sizes, GC contents, and contact persons. The International Sequencing Consortium (ISC (43)) has established a web-site to provide up-to-date information about eukaryotic genome sequencing projects of member institutions. The list also provides information about the sequencing product, the strategy applied and a proposed timetable. Both databases list all funding agencies, the sequencing centers, and very basic taxonomic information about all species. However, the taxonomic information is that limited

that the user cannot identify for example the closest homolog to his organism of interest. In addition, only a very limited amount of alternative scientific names and no common names are provided, and there is also only a limited number of links to access the primary data.
Here, we present the web-interface to diArk (digital ark) providing information on eukaryotic sequencing projects that resulted either in at least preliminary assemblies of genome data or a substantial amount of EST or cDNA data. In the center of the database are extensive species-related information (commonly and alternatively used scientific names, common names, and complete taxonomies) and much information about the respective species sequencing projects. Apart from the up-to-date status of the data our focus has been on a feature rich user interface with comprehensive and easy-to-use search capabilities.

3.3 Construction and Content

3.3.1 Technologies

The system is running on UNIX (OS X and Linux) systems. The database management system is PostgreSQL (24). As web application framework we chose Ruby on Rails (25) since it has the advantage of rapid and agile development while keeping the code well organized. Part of this framework is an implementation of Active Record (26) which is an O/RM (Object-relational Mapping) system making database integration into an object oriented program considerably easier.
The web pages are generated as XML. The site makes extensive use of Ajax (Asynchronous JavaScript and XML) in order to present the user with a feature rich interface while minimizing the amount of transferred data. All technologies used are freely available and open source.

3.3.2 Database

The unique requirements of the system demand a custom database schema (Figure 3.3). At the center of the database are three interconnected tables: species, projects and publications. The species table holds all information about the different scientific and common names, so that every species can be found even when the user does not know the exact scientific name. A comment field may contain general information about the corresponding species, the specific strain used, or common and divergent features compared to closely related organisms. Each species record is linked to a tree-like data-structure representing its taxonomy. Through this hierarchical tree, it is possible to easily select sets of species in the same taxon. The maintenance of the taxonomy tree is an automated procedure, which is triggered by the database upon insertion of new species. A delegation server receives messages from the database and starts a script to update the taxonomy tree.
The projects table contains details concerning a specific sequencing effort, such as its type (genomic DNA or EST/cDNA) and a link to the web-page providing the primary data. The term completeness is intended to describe the coverage of the genome. In this respect, EST/cDNA data is always incomplete as most genes are either only partially or not at all covered. Genomic sequencing is thought to be complete if a certain quality and coverage of the assembly is reached. Genome sequences

with low assembly coverages (¡3) and/or short assembled contigs (a few kbp) do not provide enough information to reconstitute even medium sized genes and are also considered incomplete (e.g. the mammalian 2 coverage sequencing projects). Each project may be assigned to a reference, a term we use for the large-scale sequencing centers (e.g. the DOE Joint Genome Institute) or community species homepages (e.g. FlyBase). However, for many species, the sequence information is not available via a dedicated species home page but only via GenBank. Therefore the /'GenBank/' links provide BLAST search forms including the corresponding database (some data is only available from the WGS, other from the EST database) and the corresponding species name. The projects table is always linked to a species and, in case they exist, to one or more publications.
The publications table stores all relevant information about a publication like author, title, year and journal. We included publications that refer to specific cDNA datasets (e.g. the large scale cDNA sequencing of the nematodes), or that refer to the first description of the genome sequence (e.g. the publication of the Osterococcus tauri genome). These interconnected sets of species, projects and publications form the base of the search function. For example, searching for a species also returns projects and publications. Data entry is done using the iiwi system (Odronitz F., Lampetsdoerfer T., Dietrich D., unpublished results (34)) allowing for remote editing and access control.

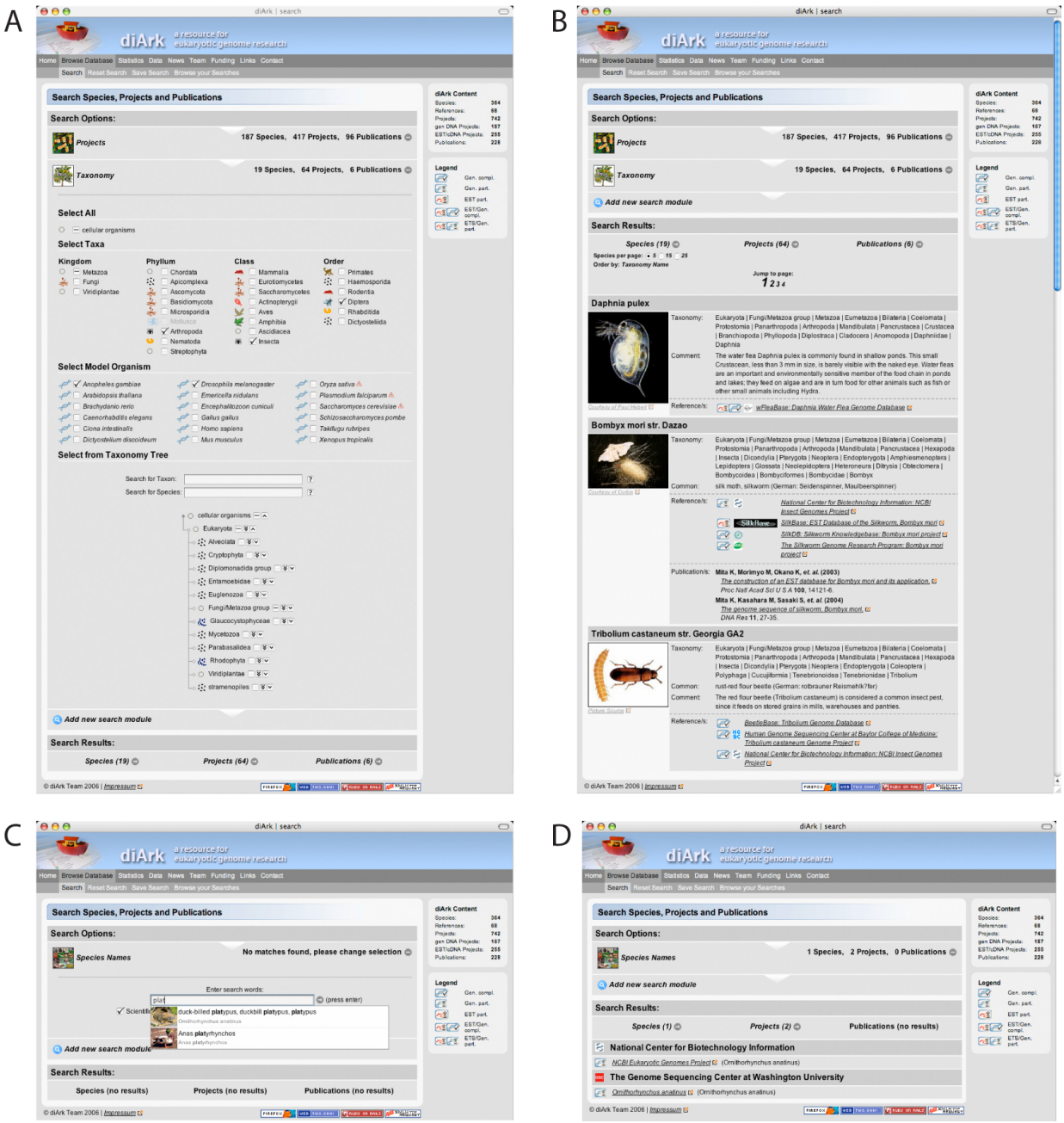

Figure 3.1: Screenshots of diArk's web-interface.
The screenshots highlight parts of the searches described in the case study.

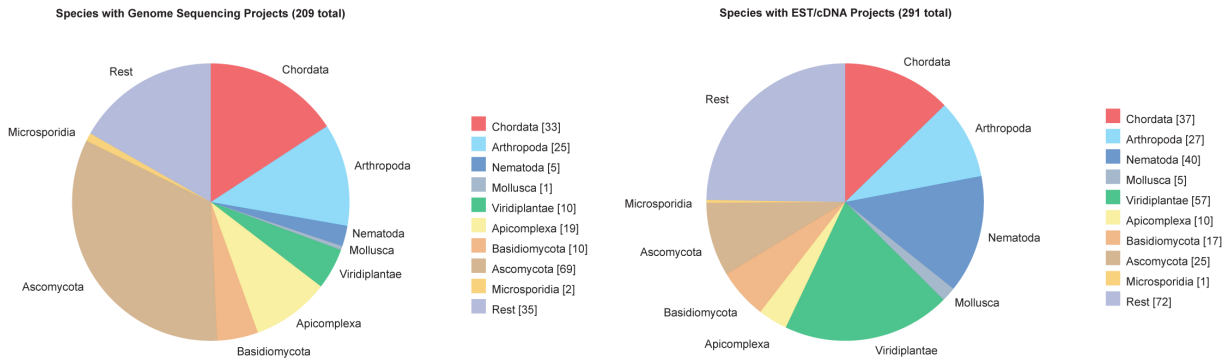

Figure 3.2: Distribution of genome sequencing and cDNA/EST projects over major branches of eukaryotic life.
The numbers of sequencing projects for some major branches of eukaryotic life are shown. The charts show the bias towards certain branches originating from the various large-scale sequencing efforts. The total number of cDNA/EST and genome projects exceeds the number of species in diArk because for some species both data are available.

3.4 Utility and Discussion

Hundreds of sequencing projects have been started in the past few years and thus the number of projects offering access to first assemblies is increasing rapidly. However, a database providing access to the primary data (genomic DNA or cDNA/EST data) of all sequenced organisms does not exist. For example, the DOE Joint Genome Institute provides access to 23 completely sequenced eukaryotes via dedicated species project pages and the data for another 3 eukaryotes via ftp server. However, the assembly data of only 9 species have already submitted to NCBI, although the data of another one has already been published. At NCBI, there are two possibilities to BLAST against genomic assembly data: directly using e.g. TBLASTN choosing the WGS database or by selecting one of the genomicBLAST tables. However, the supposedly complete table of eukaryotic genomes does not include the plant genomes. There are also strong discrepancies between the WGS database and the assemblies available via genomicBLAST. The WGS database contains 145 species while the genomicBLAST tables list only 130 organisms of which 2 are redundant. Missing species in the genomicBLAST tables comprise for example the fish *Gasterosteus aculeatus*, the plants *Ricinus communis* and *Populus trichocarpa*, and the fungus *Batrachochytrium dendrobatidis*. Even more complicating, both databases often provide different assembly versions of the genomes (e.g. v3 of the *Apis mellifera* genome in the WGS database and v4.1 via the genomicBLAST tables). These numbers show that there is a strong need for a universal database providing access to all the different sequencing projects.
diArk has been developed to store, organize and present information about sequencing projects, that have either produced preliminary or final assemblies of genome data, or that have resulted in substantial amounts of EST or cDNA data. The aim was to provide the best overview possible about the different projects so that researchers get easy access to the primary data to increase for example the taxon sampling in their phylogenetic analyses. Altogether, diArk provides links to 209 genome

assemblies and to the EST/cDNA data of 291 species (as of 12-Dec-2006). diArk does not include species for which only sequence reads are available. Given the already existing amount of completed genomes and the accumulated know-how in the sequencing centers it would not be reasonable for single researchers to build their own assemblies. We decided to not include those species until at least a draft assembly is available. Next to be up-to-date and complete, the most important requirement for diArk is a powerful and easy to use search tool.

3.4.1 Web Interface

Great attention has been paid to a versatile yet easy to use web interface. We think that accessibility and high quality representation is key to a productive usage of the system. diArk encompasses a live web front end that is generated from the content of the database at each request and thus always reflects the current data. The database is searched using modules that can be combined in chains. 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 module to select specific groups of species, a module to search sequencing project related data, and a publication search module. A search can consist of any combination of modules and their options. By adding further search modules the user can successively refine the search and narrow down the result list. For each module the resulting selection of species, projects and publications is shown, providing additional context. If a new module is added the options available will be restricted by the selection from the previous modules. At any time, the search options for every module can be changed and modifications are propagated down the chain reapplying previous user actions.

Species can be searched for in two ways. The full-text search module provides an autocompletion input field to search the list of scientific and common species names. The taxonomy search module offers tables containing specific subsets like a selection of major taxa or a range of model organisms. In addition, this module provides a taxonomic tree representation for the selection of taxa and species. Taxa and species can be browsed and selected by expanding/collapsing and including/excluding subsections of the tree, or by using shortcuts or auto-completion fields. If the dataset has been restricted by previous modules (e.g. the selection of a specific reference), excluded species and taxa are disabled in the tables.

All searches can be saved and re-run. The searches are saved purely as instructions on how to search the database. This means that if the underlying data has changed since the last run, the options set by the user will be reapplied to the data, possibly resulting in a different set of results. Based on this mechanism, we implemented an alert service that is running saved searches on a regular basis and alerts the user by email as soon as the results have changed. This enables highly customized searches to be re-run automatically in order to monitor a specific subset of the data.

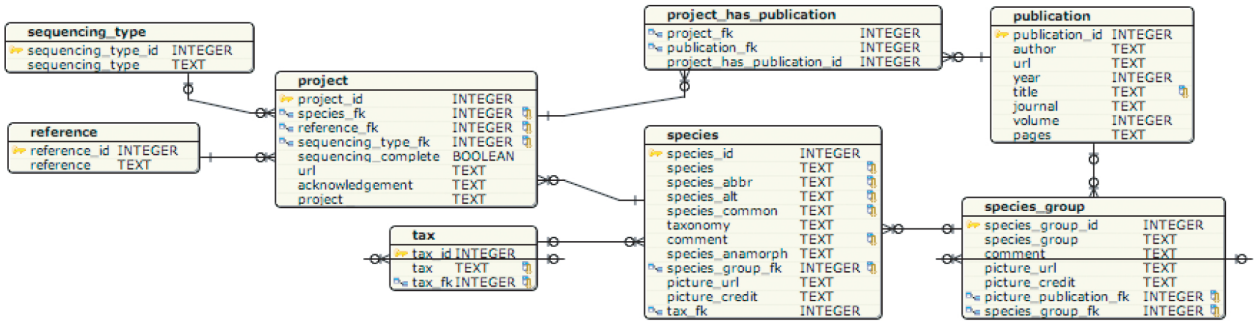

Figure 3.3: Database schema.

3.4.2 Web Services

In order to make our data available programmatically to other researchers we implemented a web service that supports XML-RPC and SOAP. The methods allow a remote program to retrieve the full data on species, publications and projects as well as the relations between different types of records. Additionally we offer a method that is equivalent to the auto-completion of the interface: When a string is given as an argument, the web service returns an array of species-IDs where the string occurs in any one of the name fields. We also make available a range of methods related to taxonomy: Taxonomy records (currently 1906), their respective children and parent as well as all species within a taxon can be retrieved. For any given species an array of taxonomy records representing their ancestry is available.

With these mechanisms we enable other programmers to conveniently construct complex queries on diArk's interconnected data without knowing about the internals.

3.4.3 Case Study

Alice wants to see which Arthropoda genomes have already been sequenced (Figure 3.1). In the taxonomic search module all species are listed regardless whether only cDNA or genomic DNA data is available. Therefore, she would have to first select /'genomic DNA data/' in the projects module. Afterwards Alice could either browse through the taxonomic tree to the Arthropoda and the underlying species or select the Arthropoda from the tax table (Figure 3.1A) and view all contents in the species result view (Figure 3.1B).

Bob wants to know whether platypus has already been sequenced, and if a genome assembly exists, to see the list of web-sites to get access to the genome data. By typing /'plat/' into the species autocompletion form of the species names search module (or the taxonomy search module) he finds that the scientific name of platypus is *Ornithorhynchus anatinus* and that there is another hit with *Anas platyrhynchos* (Figure 3.1C). Having selected *Ornithorhynchus anatinus* Bob may either choose to view the complete information connected to this organism by choosing the species view, or to view only the list of links to sequencing projects in the project view (Figure 3.1D).

3.4.4 Related Work

There is only one other serious compilation of genome sequencing projects, the GOLD database (42). GOLD comprises data of all three major lineages of life, the bacteria, the archaea and the eukaryotes. GOLD lists 674 eukaryotic sequencing projects (genome and cDNA sequencing) of which 44 are marked as published and another 13 as completed of which 4 are not publicly available. In comparison, we have found 209 genome projects (161 completed, 62 published) and included them in diArk. The major focus of GOLD seems to list all funded and ongoing sequencing projects so that researchers and sequencing consortia get an overview and help in the decision about new target species. Therefore, GOLD includes a very thorough compilation of the corresponding species sequencing centers, the funding agencies, and contact persons. On the other hand, the taxonomic information in GOLD is very limited, only a few alternative scientific names are listed and no common names are provided. In addition, only a limited number of direct links to the assembly data are given. Another major drawback of GOLD is being incomplete and not up-to-date. For example, 15 % of the links associated with eukaryotic sequencing projects do not work (397 of 2644 total). In addition, many projects are still listed as íncompleteálthough assembly data became available years ago and the genomes have been published. In contrast, the focus of diArk is to provide access to already existing genome assembly data and large cDNA/EST databases. This should enable researchers interested in comparative genomics, phylogeny, any other topic requiring taxonomic sampling, and single gene studies to get immediate access to most of the eukaryotic data available worldwide.

3.4.5 Future Developments

At the moment, it is not planed to include species data from the other two domains of life, the bacteria and the archaea, although diArk provides the framework for an easy expansion. Instead, we plan to extend diArk's current eukaryotic data content and its technical basis. From the user perspective it would be advantageous to obtain more information about the data availability and the usability of the various project web-sites. In addition, we intend to include some sequencing related data like assembly versions and coverage that will help the user to judge the different datasets. On the technical site, we plan to provide an undo function for any search as well as a general email alert for updated database content.

3.5 Conclusions

diArk is a new database to store, organize, and present the most relevant information about completed genome projects and EST/cDNA data from eukaryotes. The web-interface provides five search modules each with detailed options and three different views of the selected data. Currently, diArk provides information about 415 eukaryotic species, 823 sequencing projects, and 248 publications. cDNA/EST data is available for 291 species and genome assemblies have been released for 209 eukaryotes (13-Dec-2006; Figure 3.2). There are striking differences between the two diagrams: Due to large-scale efforts cDNA/EST data has been produced for many nematodes and plants while only a

few of these species have been sequenced on a genomic basis. In contrast, the comparative genomic programs on fungi and protozoa pathogens have resulted in many complete fungi and apicomplexa genomes.

3.6 Availability and Requirements

Project name: diArk a resource for eukaryotic genome research
Project home page: `http://www.diark.org/`
Operating system: Platform independent
Programming language: Ruby
Other requirements: The current version of diArk requires Firefox version 1.5 or higher with cookies and JavaScript enabled. Currently, other browsers do not have the required feature set or do not comply with the standards of the W3C (34).
Web-service: To use the web service via SOAP, the WSDL-file can be obtained at `http://www.diark.org/diark_backend/service.wsdl`. For using XML-RPC, users can connect to the endpoint URL `http://www.diark.org/diark_backend/api`.
Licence: The database schema, the web application and all scripts can be obtained upon request and used under a Creative Commons License.
Any restrictions to use by non-academics: Obtaining diArk by non-academics requires permission.

3.7 Authors' contributions

MK specified the requirements from a user's perspective, defined the rules for data handling, and collected all the data. FO designed the database scheme and set up the technical requirements. FO and MH did the technical design and the programming. MK and FO wrote the manuscript. All authors read and approved the final manuscript.

3.8 Supplementary Material

3.9 Acknowledgements

M.K. was supported by a Liebig Stipendium of the Fonds der Chemischen Industrie, which is in part financed by the BMBF. This work has been funded by grant I80798 of the VolkswagenStiftung and grant KO 2251/3-1 of the Deutsche Forschungsgemeinschaft.

4 Drawing the tree of eukaryotic life based on the analysis of 2269 manually annotated myosins from 328 species

Florian Odronitz,[1] and Martin Kollmar [1*]

[1]Department of NMR-based Structural Biology, Max-Planck-Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Goettingen, Germany
[*]Corresponding author.

Genome Biology
Biology for the post-genomic era
Highly accessed
Published: 18 September 2007
Genome Biology 2007, 8:R196(doi:10.1186/gb-2007-8-9-r196)

The electronic version of this article is the complete one and can be found online at `http://genomebiology.com/2007/8/9/R196`

4.1 Abstract

4.1.1 Background

The evolutionary history of organisms is expressed in phylogenetic trees. The most widely-used phylogenetic trees describing the evolution of all organisms have been constructed based on single-gene phylogenies that, however, often produce conflicting results. Incongruence between phylogenetic trees can result from the violation of the orthology assumption and stochastic and systematic errors.

4.1.2 Results

Here, we have reconstructed the tree of eukaryotic life based on the analysis of 2269 myosin motor domains from 328 organisms. All sequences were manually annotated and verified, and were grouped

into 35 myosin classes of which 16 have not been proposed previously. The resultant phylogenetic tree confirms some accepted relationships of major taxa and resolves disputed and preliminary classifications. We place the Viridiplantae after the separation of Euglenozoa, Alveolata, and Stramenopiles, we suggest a monophyletic origin of Entamoebidae, Acanthamoebidae, and Dictyosteliida, and provide evidence for the asynchronous evolution of the Mammalia and Fungi.

4.1.3 Conclusions

Our analysis of the myosins allowed combining phylogenetic information derived from class-specific trees with the information of myosin class evolution and distribution. This approach is expected to result in superior accuracy compared to single-gene or phylogenomic analyses because the orthology problem is resolved and a strong determinant not depending on any technical uncertainties is incorporated, the class distribution. Combining our analysis of the myosins with high quality analyses of other protein families, e.g. that of the kinesins, could help in resolving still questionable dependencies at the origin of eukaryotic life.

4.2 Background

Reconstructing the tree of life is one of the major challenges in biology (10). Although several attempts to derive the phylogenetic relationships among eukaryotes have been published (44, 45), many taxonomic groupings still remain heavily debated (10). The major reason for this is the fact that molecular phylogenies based on single genes often lead to apparently conflicting results (for a review, see (46)). Only recently, the application of genome-scale approaches to phylogenetic inference (phylogenomics) has been introduced to overcome this limitation ((47, 48)). In this context, large and diverse gene families are often considered unhelpful for reconstructing ancient evolutionary relationships because of the accompanying difficulties in distinguishing homologs from paralogs and orthologs (18). However, if the different homologs can be resolved, the analysis of a large gene family provides several advantages compared to a single gene analysis, because it provides additional information on the evolution of gene diversity for reconstructing organismal evolution. In addition, direct information on duplication events involving part of a genome or whole genomes can be obtained. Such an analysis requires a large and divergent gene family and sufficient taxon sampling. It is advantageous if the taxa would be closely related to provide the necessary statistical basis for subfamilies, as well as to be spread over many branches of eukaryotic life to cover the highest diversity possible. Today, sequencing of more than 300 genomes from all branches of eukaryotic life has been completed (49). In addition, many of these sequences are derived from comparative genomic sequencing efforts (e.g. the sequencing of 12 *Drosophila* species) providing the statistical basis for excluding artificial relationships.

The myosins constitute one of the largest and most divergent protein families in eukaryotes (11). They are characterized by a motor domain that binds to actin in an ATP-dependent manner, a neck domain consisting of varying numbers of IQ motifs, and amino-terminal and carboxy-terminal do-

mains of various length and function (50). Myosins are involved in many cellular tasks like organelle trafficking (21), cytokinesis (20), maintenance of cell shape (51), muscle contraction (52), and others. Myosins are typically classified based on phylogenetic analyses of the motor domain (53).
Recently, two analyses of myosin proteins describing conflicting findings have been published ((54, 55)). Both disagree with previously established models of myosin evolution (reviewed in (56)). These analyses are based on 150 myosins from 20 species grouped into 37 myosin classes (55) and 267 myosins from 67 species in 24 classes (54), respectively. However, the number of taxa and sequences included was not sufficient to provide the necessary statistical basis for myosin classification and for reconstructing the tree of eukaryotic life.
Here, we present the comparative genomic analysis of 2269 myosins found in 328 organisms. Based on the myosin class content of each organisms and the position of its single myosins in the phylogenetic tree of the myosin motor domains we reconstructed the tree of eukaryotic life.

4.3 Results

4.3.1 Identification of myosin genes

Wrongly predicted genes are the main reason for wrong results in domain predictions, multiple sequence alignments and phylogenetic analyses. Therefore, we have especially taken care in the identification and annotation of the myosin sequences. We have collected all myosin genes that have either been derived from the isolation of single genes and submitted to the nr database at NCBI, or that we obtained by manually analysing the data of whole genome sequencing and EST-sequencing projects. Gene annotation by manually inspecting the genomic DNA sequences was the only way to get the best dataset possible because the sequences derived by automatic annotation processes contained mispredicted exons in almost all genes (for an in-depth discussion of the problems and pitfalls of automatic gene annotation, gene collection, domain prediction and sequence alignment see additional data file 1). These predicted genes contain errors derived from including intronic sequence and/or leaving out exons, as well as wrong predictions of start and termination sites. Automatic gene prediction programs are also not able to recognise that parts of a gene belong together if these are spread on two or several different contigs. Often they also fail to identify all homologs in a certain organism. The only way to circumvent these problems is to perform a manual comparative genomic analysis. In addition, datasets with automatically predicted model transcripts are only available for a small part of all sequenced genomes.
The basis of our analysis was a very accurate multiple sequence alignment. In case of less conserved amino acid stretches the corresponding DNA regions of several organisms have been analysed in parallel aiming to identify coding regions and shared intron splice sites. Thus, our dataset was generated by an iterative gene identification (using TBLASTN) and gene annotation process, meaning that most of the myosin sequences have been reanalysed as soon as data from closely related organisms or further species specific data (new cDNA/EST data or a new assembly version) became available.

In addition to manually annotating the myosins from genomic data, it was also absolutely necessary to reanalyse previously published data, as these also contain many sequencing errors (especially sequences produced in the last century) and wrongly predicted translations.
The myosin dataset contains 2269 sequences from 328 organisms (Table 1), of which 1941 have been derived from 181 WGS sequencing projects. 1634 of all myosin sequences are complete (from the N-terminus to the C-terminus) while parts of the sequence are missing for 635. Sequences for which a small part is missing (up to 5%) were termed Partials while sequences for which a considerable part is missing were termed Fragments. This difference has been introduced because Partials are not expected to considerably influence the phylogenetic analysis. Indeed, even long loops like the 300 aa loop-1 of the Arthropoda variant C class-I myosins can either be included or excluded from the analysis without changing the resulting trees (data not shown). 8 of the myosins were termed pseudogenes because they contain proven single frame shifts in exons (e.g. in the HsMhc20 gene) or that many frame shifts and missing sequences that those cannot be attributed to sequencing or assembly errors.
Class-I and class-II by far comprise the most myosins (Figure 4.1A). Class-I myosins were found in almost all organisms, and class-II myosins have undergone several gene duplications (either resulting from whole genome or single gene duplications) leading to up to 22 class-II myosins per vertebrate organism. Although the total numbers of myosins per class are biased by the sequenced species we expect the class-I and class-II to remain the largest class also if many other species not containing any of these classes (e.g. the plants and Alveolata) will be sequenced in the future (Figure 4.1B). For example, the number of species of the Chordata and the Viridiplantae lineage for which myosin data is available is similar. However, the number of myosins for each of these species is very different with the Chordata species encoding up to three times more myosins. In contrast, the number of sequenced Fungi species (over 90 organisms) is almost twice as high as the number of Chordata species, but the number of Fungi myosins is only a quarter of that of the Chordata myosins.

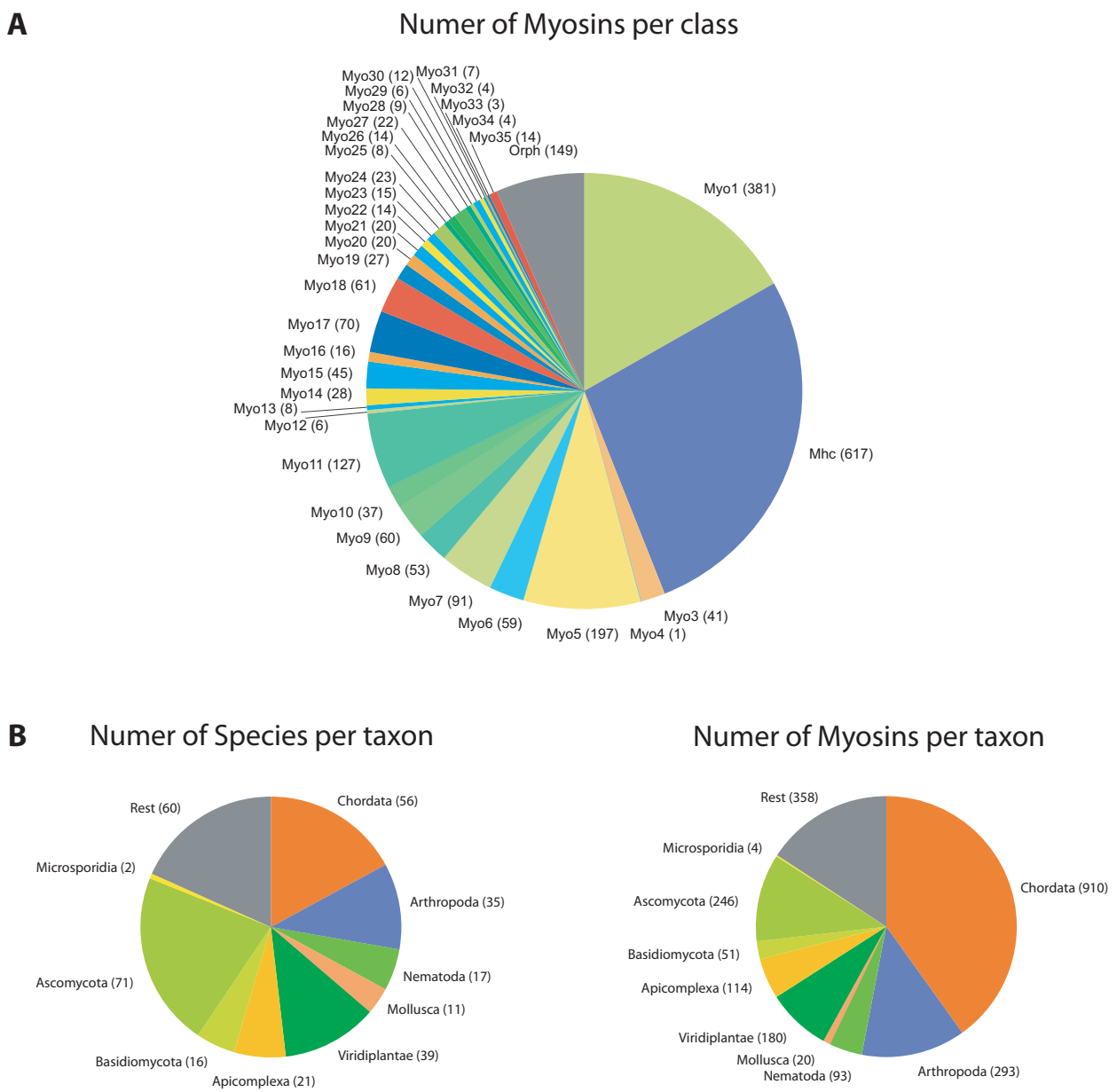

Figure 4.1: Taxon and class related statistics of the myosin dataset.
A) The pie-chart shows the number of myosins for each class. B) The charts show the number of species and the number of myosins for a set of selected taxa. Exact numbers are given in brackets.

4.3.2 Nomenclature

The amount of produced data spread over all eukaryotic kingdoms now allows and demands a consistent, systematic, and extendable nomenclature. Here, we introduce the following nomenclature that builds on the already established system (53, 56, 57, 58) and tries to keep as many of the existing names as possible. Nevertheless, it changes some of the already used names thus getting rid of sequence-specific and species-specific exceptions. We are aware of the confusion that this might introduce about the names of some sequences. But given the fact, that the annotated data known

before finishing this analysis (about 250-300 sequences) is very small compared to the data presented here, we had to introduce an appropriate nomenclature. Otherwise the number of exceptions would soon exceed the number of consistently named sequences. We are also aware that different names and classifications have recently been introduced in the literature (54, 55). However, these results were derived from analyses of small datasets based on many incorrectly assembled sequences and thus wrongly annotated myosins and we haven't found a way to incorporate the small part of matching data into our system. We also think, that even if we introduce some confusion to certain researchers in the field, there is a strong necessity to have an appropriate nomenclature to manage the existing and upcoming data. The CyMoBase, that we have developed to provide access to all myosin sequence data (37), uses the new nomenclature, provides links to previously used names, and can be used as reference.

The nomenclature is as simple as follows and in agreement with what most people in the field are already using: The names of the sequences consist of four parts: the abbreviation of the species' systematic name (A), the abbreviation of the protein (B), the class designation (C), and the variant designation (D).
A. In general, species are abbreviated by using the first letters of their systematic names (e.g. Dm for *Drosophila melanogaster*). However, there are many species, that would have the same abbreviation, and in these cases we added the second letter of the first part of the name (e.g. Drm for *Drosophila mimetica*). Different strains of the same species are differentiated by adding lowercase letters separated by an underscore (e.g. Pf_a for *Plasmodium falciparum 3D7*, Pf_b for *Plasmodium falciparum Ghanaian Isolate*, Pf_c for *Plasmodium falciparum HB3*, Pf_d for *Plasmodium falciparum Dd2*). B. The abbreviation of the protein is Myo. In the case of the class-II myosins, the abbreviations Mhc and Mys are used in the literature. As class-II comprises by far the most sequences and as numbers have very often been introduced as variant designation (e.g. human Mys1, Mys2 etc.), we decided to keep the class-II abbreviation as an exception of the proteins general abbreviation. We decided to use Mhc as protein abbreviation for class-II myosins as the abbreviation Mys has only been used for mammalian members while all other class-II myosins have been named Mhc. If the class-II myosins were named Myo2 (in accordance with the other myosin classes) we would have to also rename their variant designations to avoid confusion with other classes (e.g. Myo21 could be a class-II myosin variant 1 or a class-XXI myosin).
C. Classes are numbered according to their discovery. Thus, we keep all previously accepted class designations (56). Recent further class designations (54, 59) are based on data analyses of very small datasets of wrongly annotated myosins and will not be considered. Richards and Cavalier-Smith (55) have also used wrongly annotated myosins in their analysis and have developed a completely new classification not consistent with any previous classification. As has been agreed upon in the past new classes should only be designated if members of different organisms contribute. We have been very conservative in our analysis in designating new classes, only assigning new classes if several species contribute (e.g. class-XXI, all Arthropoda), or very divergent species (e.g. class-XXIX,

Thallassiosira pseudonana, *Phytophthora* sp. and others), or, if the species are closely related, if several homologs of each species contribute (e.g. class-XXX, *Phytophthora* sp. and *Hyaloperonospora parasitica*). It is obvious, that class separation improves as more and more divergent sequences are added. Especially the myosins of very divergent species (e.g. *Phytophthora* sp., Thallassiosira pseudonana, *Tetrahymena thermophila*, *Paramecium* tetrarelia) tend to group mainly with the homologs of the same organism. Our experience showed that if more sequences of closely related species are added (e.g. sequences of *Phytophthora ramorum*, *Phytophthora infestans*, and *Phytophthora sojae*) the class separation improves, and improves further if sequences of more divergent species are added (*Hyaloperonospora parasitica*). But in most of these cases the separation is still not good enough to distinguish between a class separation and just a variant separation. Thus, we only designated classes that are well-supported and separated. 24 classes are supported by bootstrap values higher than 985 (out of 1000, see additional data file 2) and 5 are supported by bootstrap values higher than 874. Class I has the widest taxonomic distribution and is supported by a bootstrap value of 788. Class XXVIII (bootstrap value of 750), class V (593) class XXIII (463) and class XV (305) show the lowest bootstrap values, but are well separated from any neighbouring class. We left groups of sequences (e.g. the *Tetrahymena thermophila* and *Paramecium* tetrarelia myosins) unclassified although their first node in the tree might be supported by a relatively high bootstrap value. A similar situation would exist, if only five sequences of class-VII, class-X, and class-XV myosins were known. Then these sequences would certainly group together, supported by a high bootstrap value of the first node, as they are far more similar to each other than to the other myosins. Adding more homologs showed these myosins to be separated in three classes, and we expect a similar class separation for the myosins of e.g. *Tetrahymena thermophila* and *Paramecium* tetrarelia if more sequences of closely related species will be added. D. If several myosin homologs exist for the same class, they are distinguished by a variant designation, a letter starting with A. Only for the class-II myosins variants with numbers may be used (see above).
E. If both alleles of an organism have been assembled independently providing two versions for each myosin gene, the different versions are distinguished by adding alpha and beta to the sequence name.
F. Alternative splice forms of the same gene get the same protein name.

All myosins that cannot be classified at the moment will be considered as 'orphan' myosins. If several orphans exist in a species, they get a variant designation. Orphan names are considered to be preliminary names. Thus, orphan myosins will be renamed as soon as more sequences are available that allow a well-supported classification.

4.3.3 Classification

The basis for the classification of the myosins is the phylogenetic relation of their myosin motor domains (53, 56). Now, the data for the myosins is strong enough that all designated classes are well supported. Including or excluding sets of myosins (e.g. the orphans) does not change the phylogeny

of the other classes (as has been observed for the small dataset used in previous analyses, (54)). Also, including or excluding large insertions like the loop-1 insertion of the class-I variant C myosins of Arthropoda, does not change the tree.

In contrast to other suggestions we do not agree with the idea that the tail domain architectures should also be considered in the classification process (54, 55). Our analysis shows that the motor domains and the tails coevolved in most of the assigned classes, but there are many exceptions now where the separation of organismal lineages occurred before the adaptation of further tail domains. It does not make sense to artificially 'force' sequences together, only because there is not enough sequence data for a better classification. If for example the class-XII myosins should be related to the class-XV myosins only because they also contain MyTH4 and Ferm domains (54), then they could also be grouped to the class-VII, class-X, or class-XXII myosins. Many other myosins from Stramenopiles or *Amoeba* would also have to be grouped to these classes as they also contain MyTH4 and Ferm domains. This seems very arbitrary. Also, several domains like the PH domain, Ankyrin repeats or the Pkinase domain are found on either the N-terminus or the C-terminus of the myosins. Many of the tail regions have also not been analysed in particular (domains have not been defined yet). Thus, as soon as further domains get defined other myosin classes might unexpectedly share tail regions. It is also not reasonable to consider the organismal distribution of myosins as classification helper as has been proposed (54). The species sequenced only cover an extremely small part of all organisms, and their selection has also been biased in favour of financial, medical and other interests. It is therefore not reasonable to assume that the organisms that we know are the best representatives regarding the myosin diversity of their taxa. For example even the well-studied *Drosophila melanogaster* has lost the class-XXII myosin that the closely related species *Drosophila willistoni* and other *Drosophila* species still have. Other Arthropoda (*Daphnia*, *Apis*, *Anopheles*) have additional myosins belonging to well established classes (e.g. a class-III myosins and a class-IX myosin) that all *Drosophila* species (that have been sequenced so far) have lost. The same is true for nematodes where a class-XVIII myosin is found in *Brugia malayi* and not in *Caenorhabditis* species. It is therefore very unlikely that myosins, that do not group to any of the other assigned metazoan myosins (e.g. the class-XII myosins), are closely related to one of the metazoan classes although they might share some domains in the tail regions. It is far more likely that a class-XII myosin will be found in another metazoa species (as e.g. a class-XX myosin has been found in Echinodermata in addition to Arthropoda), or that a class-XV myosin (to which the class-XII myosins have artificially been grouped to, (54)) will be found in another nematode (as e.g a class-XVIII myosin has been found in *Brugia malayi*). Both possibilities will support the current class designation. Nevertheless, at the moment it seems that all sequenced lineages have developed their own specific myosin, e.g. the class-XVI myosins in vertebrates, the class-XXI myosins in Arthropoda, and the class-XII myosins in Nematoda.

Fragments have been classified and named based on their obvious homology at the amino acid level. Those Fragments that did not obviously group to one of the assigned classes have sequentially been added to the dataset used to construct the major tree. Some of these Fragments could subsequently be

classified; others have to be considered as orphans. Note, that even very short fragments of only 100 amino acids are sufficient for proper classification. Thus, it is very unlikely that the orphan Fragments will group to one of the established 35 classes if their full-length sequences become available.

4.3.4 Renamed myosins

Change of previous classification: Class-IV contains only one myosin. According to the nomenclature guidelines outlined above this myosin would not be designated a class but be considered as orphan. To not start confusion we didn't change its classification as class-IV myosin, also expecting that more members will be added as soon as further genomes are sequenced. However, our phylogenetic tree shows that the former class-XIII myosins (of the algae *Acetabularia* cliftonii) belong to the class-XI myosins, supported by a bootstrap value of 999. Therefore, we reclassified the former *Acetabularia* class-XIII myosins as class-XI myosins, and assigned the class-XIII to a Kinetoplastida specific myosin class. The *Drosophila melanogaster* NinaC protein has previously been classified as class-III myosin. However, other Arthropoda contain real class-III myosins (or better: homologs to the mammalian class-III myosins) and NinaC as well as the NinaC homologs of the other Arthropoda form a distinct class. We decided not to rename all the mammalian class-III myosins but to rename NinaC and introduce the new class-XXI.

Change of previous names: The apicomplexan myosins have traditionally been named alphabetically (54, 60). However, even different splice forms of the same gene got different protein names. In addition, gene and genome duplication events have lead and will lead to a confuing naming. Thus, it is not possible to name these myosins consistently in an alphabetical manner and to provide consistency for the future. We renamed the apicomplexan myosins according to our nomenclature introducing some apicomplexan specific myosin classes. Nevertheless, we tried to keep the former letters as variants where possible.

The *Saccharomyces cerevisiae* myosins have previously been named numerically (61) thus leading to confusion with class numbers. In addition, several yeast species have now been sequenced that have separated before some of the gene and whole genome duplication events happened during yeast evolution. Most of the sequenced yeast species contain only one version of the class-I and class-V myosins, and *Naumovia castellii* contains one class-I but two class-V myosins. It is not possible to name the newly identified yeast myosins according to the *Saccharomyces cerevisiae* myosins. Therefore, we renamed the *Saccharomyces cerevisiae* myosins according to our nomenclature.

Some of the plant and algae myosins were given arbitrary names in the past, especially those from *Helianthus annuus* and *Arabidopsis thaliana*. This happened before genome data became available but has not been changed later on /citeReddy2001. We have renamed these few myosins. Some of the vertebrate class-II myosins have also been renamed based on their homology to myosins from closely related organisms. Especially descriptive names (e.g. 'nonmuscle myosin II' or 'fast skeletal

muscle myosin') have been disbanded in favour of numerical variant designations as suggested (56).

4.3.5 35 myosin classes

The analysis of the phylogenetic tree of the 2269 myosin motor domain sequences resulted in the definition of 35 myosin classes (Figure 4.2, Figure 4.3, additional data file 2) of which 19 classes have been assigned and described previously (56). Our analysis supports and keeps the existing classification except for the former class-XIII that consisted of two myosins from the chlorophyte *Acetabularia peniculus* (*Acetabularia* cliftonii). The former class-XIII was substituted by a Kinetoplastide specific class consisting of myosins with an N-terminal SH3-like domain, a coiled-coil region, and two tandem UBA domains. Five new classes, class-XX, class-XXI, class-XXII, class-XXVIII, and class-XXXV are specific to Metazoan species. So far, class-XX has only been found in arthropods and the sea urchin *Strongylocentrotus purpuratus* and consists of myosins with a long, coiled-coil region containing N-terminal domain, and a short neck comprised by one IQ motif. The myosins of class-XXI are very similar to the class-III myosins in their domain organisation but contain distinct motor domains. The class-XXII myosins are defined by two tandem MyTH4 and FERM domains. Most Metazoan species have lost their class-XXVIII myosin. So far, class-XXVIII myosins have only been identified in the sea anemone *Nematostella vectensis*, the frog *Xenopus tropicalis*, in *Gallus gallus*, and some fishes. From the data available it seems that the species of the Acanthopterygii branch of the fishes (including *Takifugu rubripes* and *Gasterosteus aculeatus*) have lost the class-XXVIII myosins. The tail regions of class-XXVIII myosins consist of an IQ motif, a short coiled-coil region and an SH2 domain.

Five of the new myosin classes (class-XXIII to class-XXVII) have solely members of Apicomplexan myosins. The domain organisations of these myosins have been described elsewhere (54) but classes have not been assigned yet. Another six new myosin classes were attributed to stramenopiles myosins (class-XXIX to class-XXXIV). Class-XXIX shows the highest taxonomic sampling consisting of members of all stramenopiles species. Class-XXIX myosins have very long tail domains consisting of three IQ motifs, short coiled-coil regions, up to 18 CBS domains, a PB1 domain, and a C-terminal transmembrane domain. The myosin classes XXX to XXXIV contain only members of *Phytophthora* species and the closely related *Hyaloperonospora parasitica*. Although the taxonomic sampling is quite low, these classes have distinct motor domains and unique tail domain organisations. Myosins of class-XXX are composed of an N-terminal SH3-like domain, two IQ motifs, a coiled-coil region and a PX domain. Class-XXXI myosins have a very long neck region consisting of 17 IQ motifs and two tandem Ankyrin repeats separated by a PH domain. Class-XXXII myosins do not contain any IQ motifs but a tandem MyTH4 and FERM domain. The myosins of class-XXXIII have long N-terminal regions with an N-terminal PH domain. Class-XXXIV myosins are composed of one IQ motif, a short coiled-coil region, five tandem Ankyrin repeats, and a C-terminal FYVE domain.

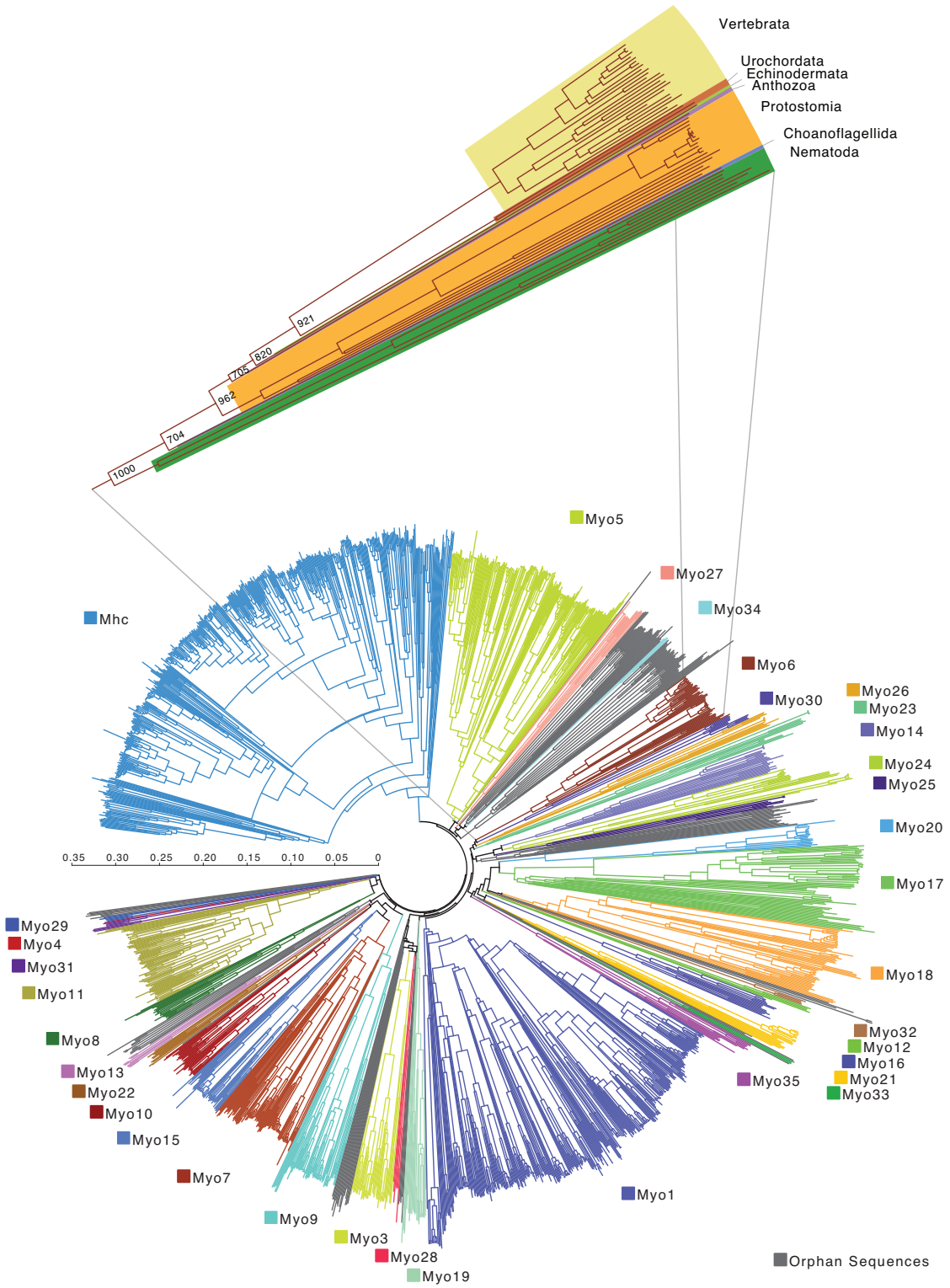

Figure 4.2: Phylogenetic tree of the myosin motor domains
The phylogenetic tree was built from the multiple sequence alignment of 1984 myosin motor domains. The complete tree with bootstrap values and sequence descriptors is available as additional data file 2. The expanded view shows the myosin sequences of class-VI and their distribution in taxa. Every other myosin class has been analysed in a similar way. Labels at branches are bootstrap values (1000 total boostraps). The scale bar corresponds to estimated amino acid substitutions per site.

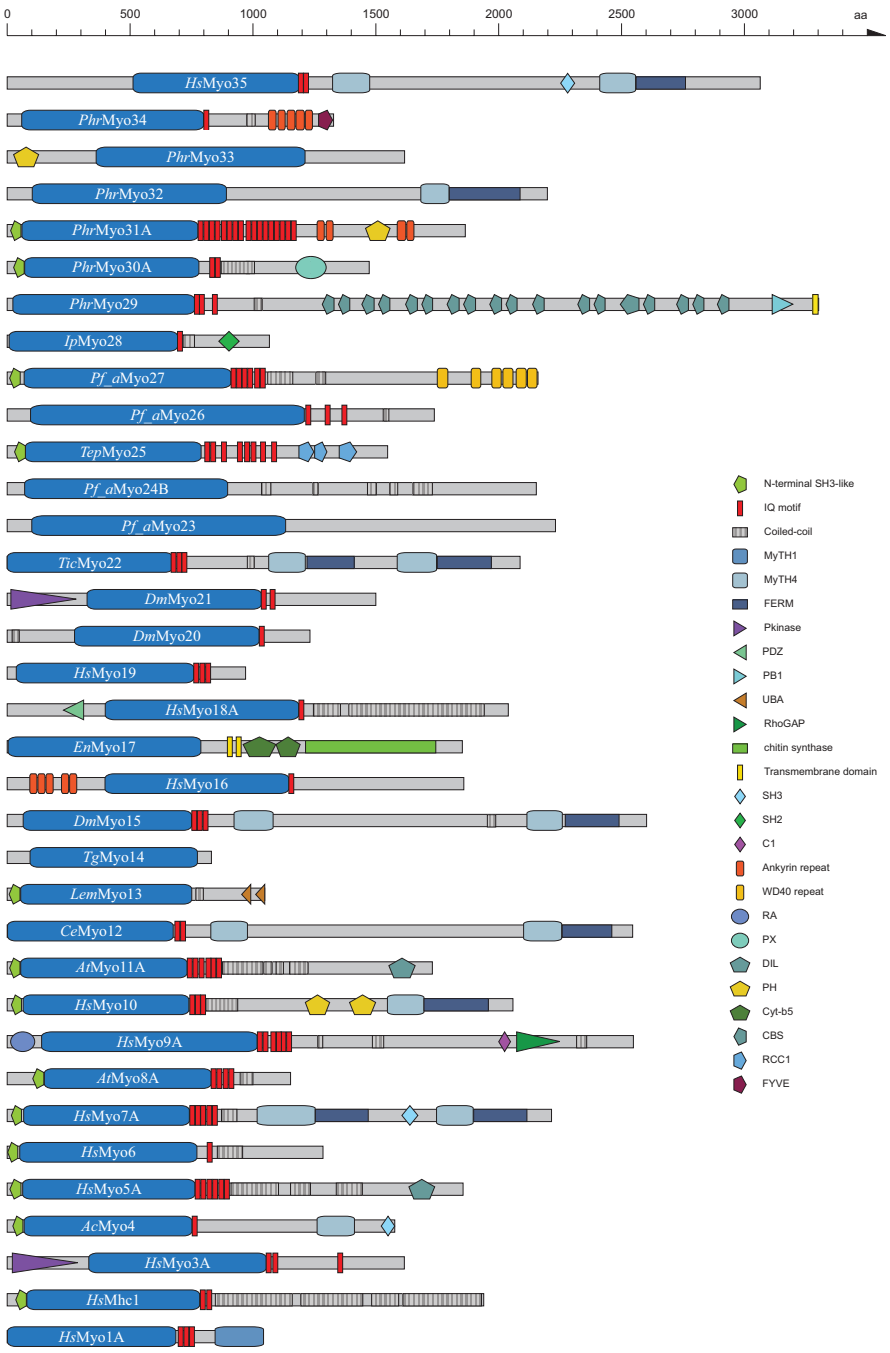

Figure 4.3: Schematic diagram of the domain structures of representative members of the 35 myosin classes.
The sequence name of the representative member is given in the motor domain of the respective myosin. A colour key to the domain names and symbols is given on the right except for the myosin domain that is coloured in blue. The abbreviations for the domains are: C1, Protein kinase C conserved region 1; CBS, cystathionine-beta-synthase; Cyt-b5, Cytochrome b5-like Heme/Steroid binding domain; DIL, dilute; FERM, band 4.1, ezrin, radixin, and moesin; FYVE, zinc finger in Fab1, YOTB/ZK632.12, Vac1, and EEA1; IQ motif, isoleucine-glutamine motif; MyTH1, myosin tail homology 1; MyTH4, myosin tail homology 4; PB1, Phox and Bem1p domain; PDZ, PDZ domain; PH, pleckstrin homology; Pkinase, Protein kinase domain; PX, phox domain; RA, Ras association (RalGDS/AF-6) domain; RCC1, Regulator of chromosome condensation; RhoGAP, Rho GTPase-activating protein; SH2, src homology 2; SH3, src homology 3; UBA, ubiquitin associated domain; WD40, WD (tryptophan-aspartate) or beta-transducin repeats.

4.3.6 Orphan myosins

Fungi/Metazoa lineage: The domain organisations of the orphan myosins of the Fungi/Metazoa lineage are shown in Figure 4.4. The Microsporida have two myosins, one class-II myosins and an orphan myosin containing a DIL domain that is also shared by class-V and class-XI myosins. In contrast to these classes, the Microsporida orphan myosins do not have any IQ motifs thus lacking the possibility to bind calmodulin-like light chains. The wasp *Nasonia vitripennis* has an orphan myosin that has a similar domain organisation to the class-V and class-XI myosins although it has less IQ motifs and its coiled-coil region is considerably shorter. This myosin is unique to all Arthropoda species sequenced so far. A myosin very similar in domain organisation to the fungal class-XVII myosins has been found in the mollusc *Atrina rigida*. It has twelve transmembrane domains separated by a chitin synthetase domain. The choanoflagellate *Monosiga brevicollis* has sixteen orphan myosins of different domain organisations. Due to missing genome sequence data of closely related species all these gene predictions are preliminary (especially the tail regions) and might change in the future. Some of the predicted orphan myosins contain domains unique to all myosins analysed so far, like the SAM and the Vicilin-N domains. Seven sequences contain SH2 domains as have been found in the class-XXVIII myosins.

Alveolata lineage: Several of the Alveolata myosins could not be classified (Figure 4.5). All *Tetrahymena thermophila* and *Paramecium tetraurelia* myosins remain ungrouped. The tails of the *Paramecium tetraurelia* myosins only contain IQ motifs, coiled-coil regions, and RCC1 domains, while some of the *Tetrahymena thermophila* myosins also contain FERM or MyTH4 domains. However, the FERM and MyTH4 domains never appear in tandem like in class-VII, class-X, or class-XXII myosins.

Orphan myosins from stramenopiles: Although they only share the class-I myosins the stramenopiles species show a similar myosin diversity as the metazoan species (Figure 4.6). So far, three *Phytophthora* species and the closely related *Hyaloperonospora parasitica* have been sequenced. All share the same set of myosins. The orphan myosins of this group have not been classified because it is not clear from the phylogenetic tree where to draw class boundaries. However, it is obvious that the Myo-A to Myo-H and the Myo-Q to Myo-U orphans form distinct groups. The domain organisations of the myosins within these groups are also very different. To resolve their classification further data from more distantly related species is needed. The genome sequences of two diatoms, *Phaeodactylum tricornutum* and *Thalassiosira pseudonana* have also been finished. Both species share several sequences, but *Thalassiosira pseudonana* has a higher myosin diversity having myosins with HEAT or Mis14 domains that do not exist in any other myosin.

Orphan myosins from other taxa (Figure 4.7): The *Dictyostelium discoideum* orphan myosins have been discussed elsewhere (62). The amoeba-flagellate *Naegleria gruberi* has three orphan myosins having only coiled-coil regions in the tail. The unicellular red alga *Galdieria sulphuraria* contains one myosin with a unique domain organisation consisting of at least nine IQ motifs followed by an

AAA domain and a DnaJ domain. Both alleles of *Trypanosoma cruzi* have been assembled independently providing two slightly different versions for each myosin gene. The seven orphan myosins of *Trypanosoma cruzi* contain N-terminal SH3-like domains, IQ motifs, or coiled-coil regions.

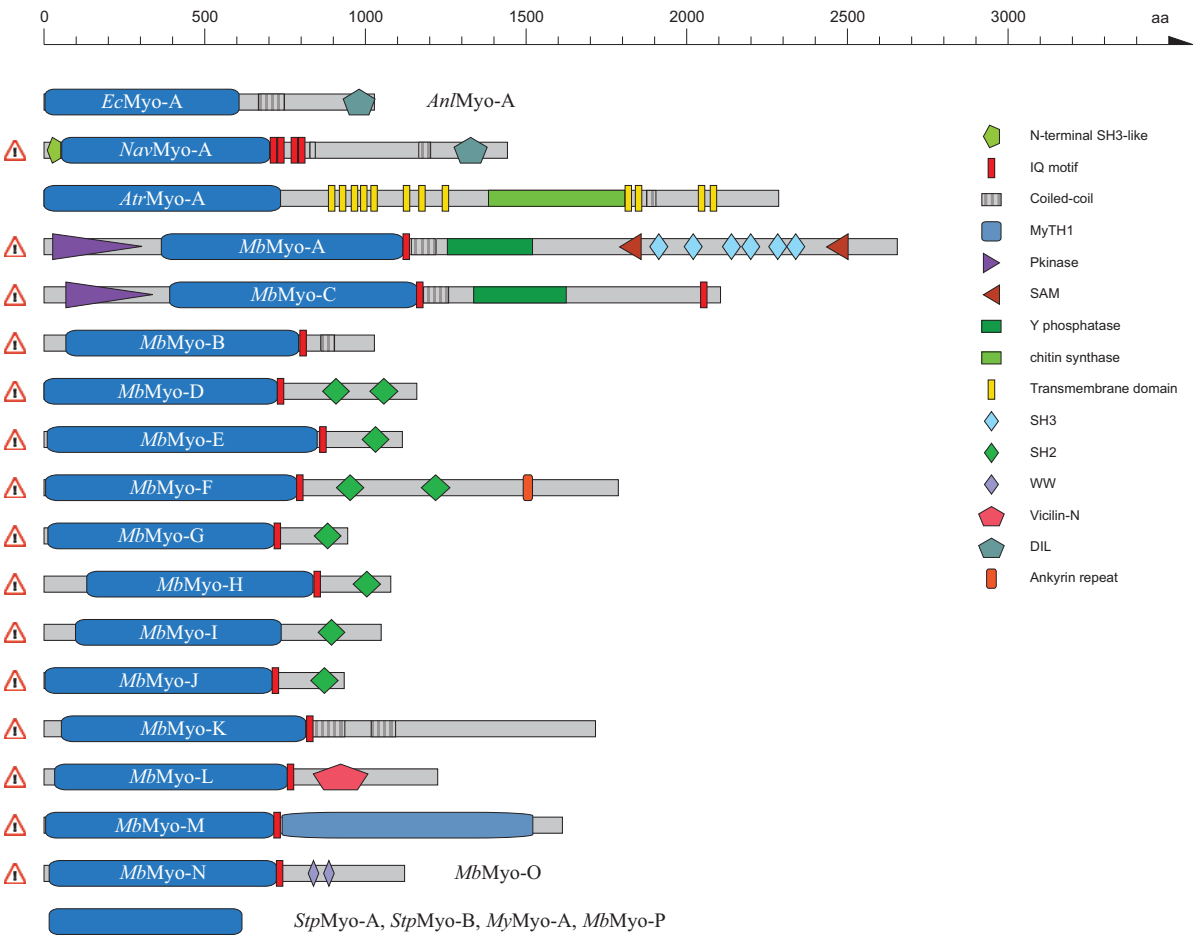

Figure 4.4: Schematic diagram of the domain structures of the orphan myosins of the Fungi/Metazoa lineage.
The sequence names of the ophan myosins are given in the motor domain of the respective myosins. Colour keys to the domain names and symbols are given on the right except for the myosin domain that is coloured in blue. Myosin names next to domain representations list orthologs from closely related species or orthologs from the same species. These sequences have a similar domain organisation. Sequences that are not ortholog and have not resulted from recent gene duplications are shown separately although their domain organisations might be very similar. The myosin domains without names on the bottom symbolize that only head fragments are available for the sequences listed on the right. The exclamation mark on the left site of some sequences signifies that the corresponding sequences (especially the tail regions) have not completely been validated because of missing comparative genome sequences. Those sequences and corresponding tail domain predictions might change with upcoming genome sequences of related species. Abbreviations for the domains are: SAM, Sterile alpha motif; Vicilin-N, Vicilin N terminal region; WW, tryptophan-tryptophan motif domain; Y phosphatase, Protein tyrosine phosphatase, catalytic domain.

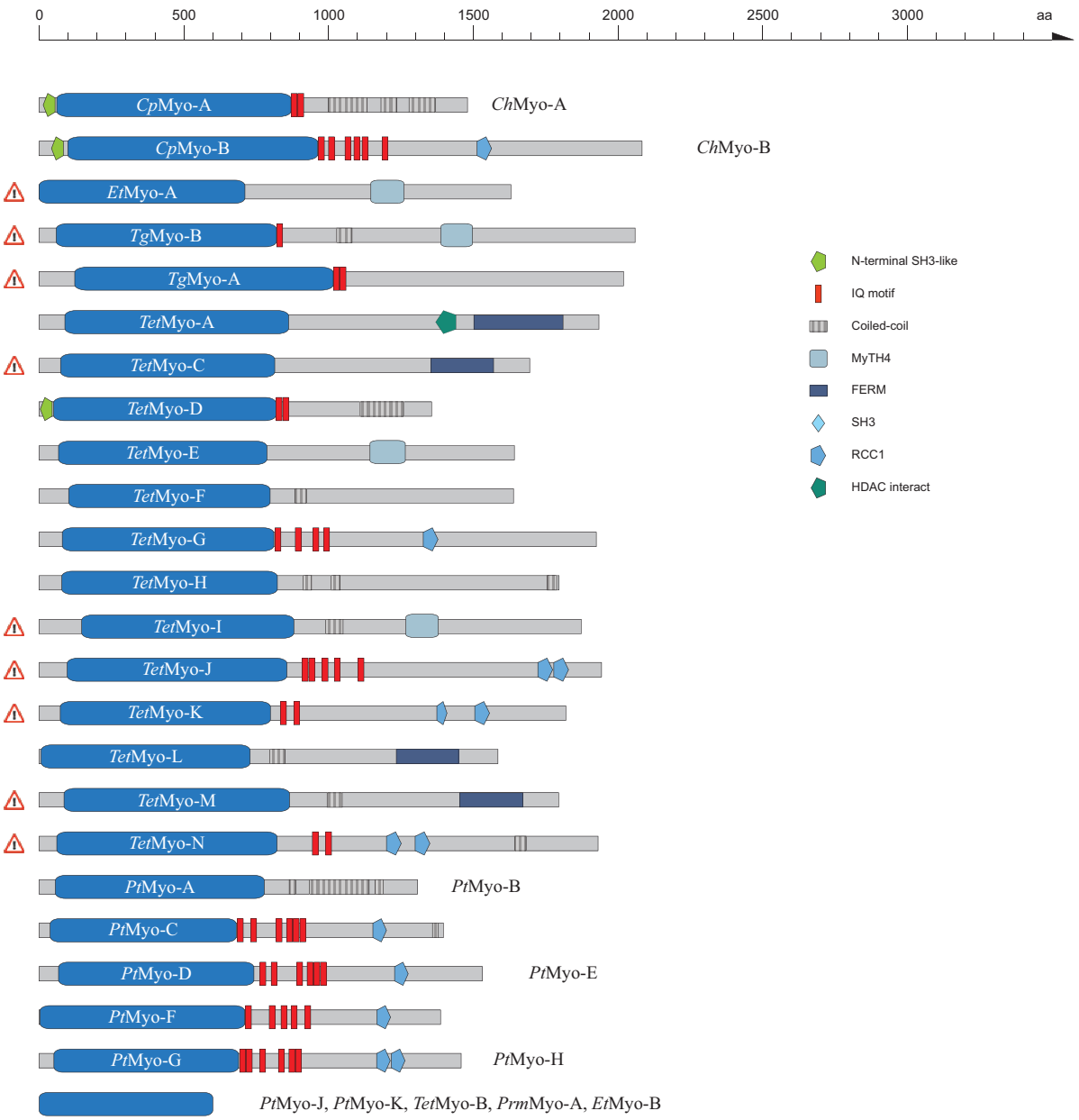

Figure 4.5: Schematic diagram of the domain structures of the orphan myosins from the Alveolata lineage.
The sequence names of the ophan myosins are given in the motor domain of the respective myosins. Colour keys to the domain names and symbols are given on the right except for the myosin domain that is coloured in blue. Myosin names next to domain representations list orthologs from closely related species or orthologs from the same species. These sequences have a similar domain organisation. Sequences that are not ortholog and have not resulted from recent gene duplications are shown separately although their domain organisations might be very similar. The myosin domains without names on the bottom symbolize that only head fragments are available for the sequences listed on the right. The exclamation mark on the left site of some sequences signifies that the corresponding sequences (especially the tail regions) have not completely been validated because of missing comparative genome sequences. Those sequences and corresponding tail domain predictions might change with upcoming genome sequences of related species. Abbreviations for the domains are: HDAC interact, Histone deacetylase (HDAC) interacting.

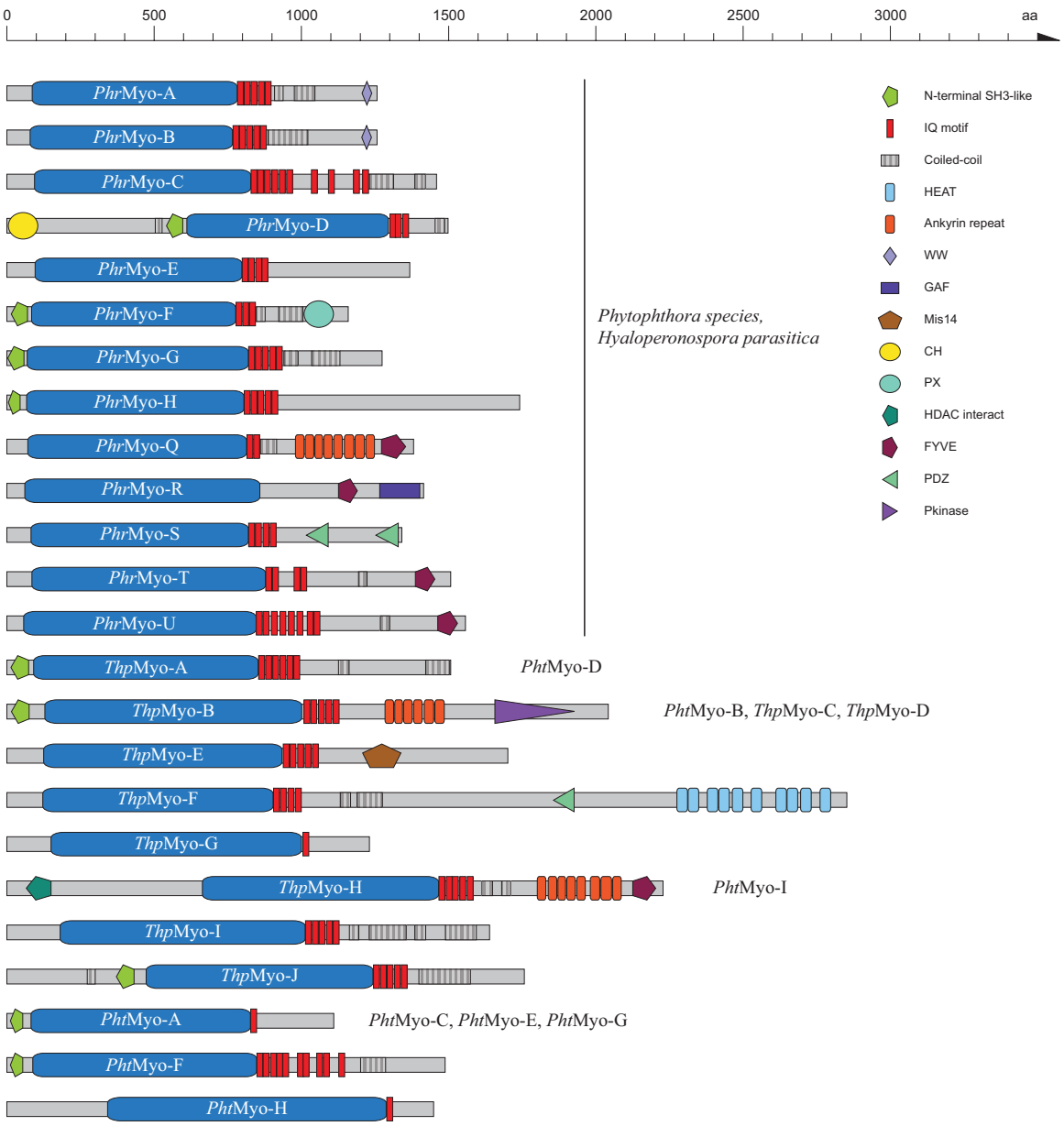

Figure 4.6: Schematic diagram of the domain structures of the orphan myosins from stramenopiles.
The sequence names of the ophan myosins are given in the motor domain of the respective myosins. Colour keys to the domain names and symbols are given on the right except for the myosin domain that is coloured in blue. Myosin names next to domain representations list orthologs from closely related species or orthologs from the same species. These sequences have a similar domain organisation. Sequences that are not ortholog and have not resulted from recent gene duplications are shown separately although their domain organisations might be very similar. Abbreviations for the domains are: CH, Calponin homology domain; GAF, domain present in phytochromes and cGMP-specific phosphodiesterases; HEAT repeat, named after the proteins huntingtin, elongation factor 3 (EF3), the 65 Kd alpha regulatory subunit of protein phosphatase 2A (PP2A) and the yeast PI3-kinase TOR1; Mis14, Kinetochore protein Mis14 like.

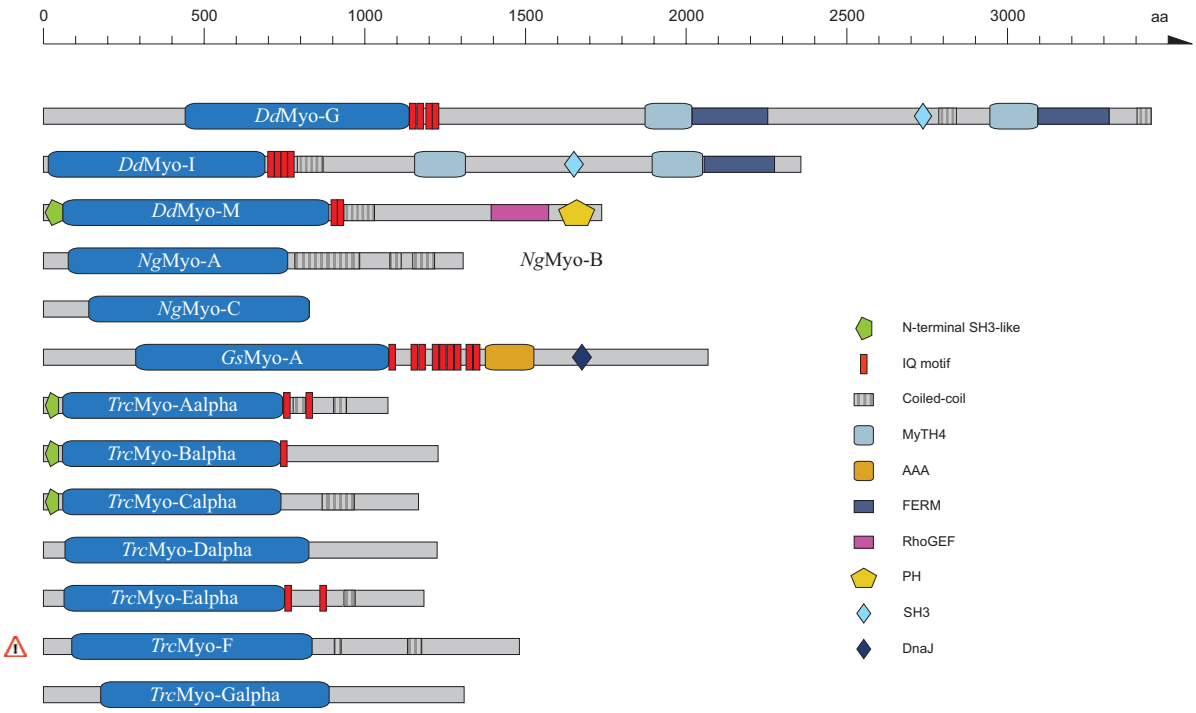

Figure 4.7: Schematic diagram of the domain structures of the orphan myosins of species not belonging to one of the other taxa.
Both alleles of *Trypanosoma cruzi* have been assembled independently providing two slightly different copies of each myosin gene. None of the Myo-F versions is complete and the presented domain organisation of Myo-F is the result of a merged version of both myosins. The sequence names of the ophan myosins are given in the motor domain of the respective myosins. Colour keys to the domain names and symbols are given on the right except for the myosin domain that is coloured in blue. Myosin names next to domain representations list orthologs from closely related species or orthologs from the same species. These sequences have a similar domain organisation. Sequences that are not ortholog and have not resulted from recent gene duplications are shown separately although their domain organisations might be very similar. Abbreviations for the domains are: AAA, ATPase family associated with various cellular activities; DnaJ domain, named after the prokaryotic heat shock protein DnaJ; RhoGEF, Rho GDP/GTP exchange factor.

4.3.7 Species that do not contain myosins

There are three species whose genome sequences are available and that do not contain any myosin: the unicellular red alga *Cyanidioschyzon merolae*, the flagellated protozoan parasite *Giardia lamblia*, and the protozoan parasite *Trichomonas vaginalis*.

4.4 Discussion

All myosin protein sequences have been derived by manually inspecting the corresponding DNA, either the published cDNA or genomic DNA, or the genomic DNA provided by the sequencing centres. Published sequences contained errors in many cases, either from sequencing or from manual annotation, while automatic annotations provided by the sequencing centers resulted in mispredicted

exons in almost all transcripts. For many sequences, the prediction of the correct exons was only possible with the help of the analysis of the homologs of related species. Thus, not only the quantity of myosin data increased as more and more genomes have been analyzed but also the quality as all ambiguous regions could be resolved for those sequences for which data from a closely related organism is available. Therefore, mispredicted exons may be limited to a few orphan myosins.

For the phylogenetic analysis of the myosin motor domains we created a structure-guided manual sequence alignment whose quality is far beyond any computer-generated alignment. It is obvious that all secondary structure elements of the class-II myosin motor domain structure remain conserved in all myosins, even in the most divergent homologs. Sequence motifs, that would have not been aligned at first glance, were placed based on the analysis of their supposed 3-dimensional counterparts that always maintained the structural integrity of the respective region. Thus, strong sequence variation and sequence insertions were limited to loop regions. Based on the phylogenetic tree constructed from 1984 myosin motor domains, 35 classes have been assigned (Figure 4.2, Figure 4.3; additional data files 2 and 3). 149 myosins still remain unclassified due to our conservative view on designating classes but it is anticipated that sequencing of further genomes will result in their classification and will substantially increase the existing number of classes. For generating the tree it does not matter whether long loop regions (e.g. the 300 aa loop-1 of the Arthropoda Myo1C proteins) are included in the alignment or not (data not shown). So far, almost all orphan myosins belong to taxa that have not undergone large-scale comparative sequencing efforts. Only short sequence fragments have been found for 277 myosins. These sequences were excluded from the phylogenetic analysis but have been classified based on their similarity in the multiple sequence alignment. Nevertheless, these data are important to define the myosin diversity in as many organisms as possible.

The highest number of myosins in a single organism has been found in *Brachydanio rerio* (61 myosins grouped into 13 classes) while the broadest class distribution is expected for the *Phytophthora* species (25 myosins grouped into at least 15 classes). The high numbers of vertebrate myosin genes in general are due to several whole genome duplications that happened after the separation from the Craniata and Urochordata (63).

Our survey of the myosin gene family now allows the reconstruction of the tree of 328 eukaryotes (Figure 4.8). The organisms of the major clades Fungi/Metazoa, Euglenozoa, Stramenopiles and Alveolata have distinct sets of myosin classes (except class I) showing that horizontal gene transfer of myosins has not happened in later stages of eukaryotic evolution. However, we cannot exclude yet that horizontal gene transfer of myosins has not happened at the origin of eukaryotic evolution. Hence, only paralogs and orthologs have to be resolved. Figure 4.8 represents a schematic reconstruction of both the phylogenetic relationships of major taxa reconstructed from class-specific trees as well as the information of myosin class evolution and distribution. For example, *Tetrahymena thermophila*, *Perkinsus marinus*, *Toxoplasma gondii*, *Plasmodium falciparum*, and *Babesia bovis* have all been classified as Alveolata. However, the relation between Ciliophora (*Tetrahymena thermophila*), Perkinsea (*Perkinsus marinus*), and Apicomplexa (*Toxoplasma gondii*, *Plasmodium falciparum*, and *Babesia bovis*) has not been resolved yet. *Tetrahymena thermophila* does not share any myosin with

the other Alveolata and should have therefore been diverged before the other species. *Perkinsus marinus* shares two myosin classes with the Apicomplexa. Thus, they must have had a common ancestor. The Apicomplexa developed three further common classes of which single classes have been lost by different species. The myosin class specific trees show that the Coccidia, the Haemosporida, and the Piroplasmida form distinct lineages. However, their relation cannot be resolved further. This principle for reconstructing the tree has been applied to all species.

The class-I myosins show the widest taxonomic distribution, are devoid of the amino-terminal SH3-like domain and thus suggested to be the first myosins evolved (see below). Only two major lineages, the Viridiplantae and the Alveolata, do not contain class-I myosins (Figure 4.8). The Alveolata have either lost the class-I myosin, or their class-I myosin diverged so far that a common ancestor could not be reconstructed. The Apicomplexa developed several specific classes, while the Ciliophora myosins cannot be classified yet. The evolutionary history of the Euglenozoa and Stramenopiles cannot be further resolved because both do not share any further myosin classes with other species, and their taxonomic sampling is not high enough for a more precise grouping.

The second myosin class developed during the evolution of the Fungi and Metazoa kingdoms was class-V. The plants have developed two kingdom-specific classes. However, the domain organization of the plant-specific class-XI is similar to that of class-V, suggesting that both had a common ancestor. In contrast to the class-I myosins, the class-V and class-XI myosins have diverged so far that a common ancestry is not visible beyond their general domain organisation. After separation of the plant lineage, the class-II myosins arose. The protists *Entamoeba* sp., *Acanthamoeba castellanii*, *Naegleria gruberi*, and *Dictyostelium discoideum* have closely related myosins suggesting that they share a common ancestor that diverged shortly before the Fungi and Metazoa split. While the Entamoebidae have lost their class-V myosin retaining only a class-I and a class-II myosin, the Acanthamoebidae, Dictyosteliida, and Heterolobosea have developed several additional specific myosins with unique domain organizations, in addition to the increase in the number of myosins genes through single gene or whole genome duplications. The Acanthamoebidae and Dictyosteliida already contain the combination of the myosin motor domain and the MyTH4 domain that is also widely found in the metazoan lineage. However, the lack of more genomic data prevents the designation of a common myosin motor domain-MyTH4 containing ancestor. The fungi developed the class-XVII myosin that consists of a functionally restricted myosin motor domain fused with a highly conserved chitin synthetase (64). While the Ascomycetes, Basidiomycetes, and Chytridiomycota have retained one member of each of the four myosin classes, the Zygomycotes *Rhizopus arrhizus* and *Phycomyces blakesleeanus* have undergone several single gene or whole genome duplications. The Saccharomycetes, Schizosaccharomycetes, and Microsporidia have lost their class-XVII myosin.

Two different models can be proposed for the further evolution of the Metazoa (Figure 4.8 and Figure 4.9). In both models a considerable boost of myosin diversity happened at the early evolution of Metazoa. The most reasonable model based on the myosin class distribution suggests an increase of the myosin diversity in three steps. After separation of the Fungi, the Metazoa developed four new classes, class-VI, class-VII, class-IX, and class-XVIII. These classes are shared by species of all

Metazoa taxa sequenced so far, except the choanoflagellate *Monosiga brevicollis* that does not contain class-IX and class-XVIII myosins. However, single species of the other taxa have also lost their members of these four classes, like the nematode *Trichinella spiralis* only contains a class-VII myosin, the *Caenorhabditis* species have lost their class-XVIII myosins, and the *Drosophila* species have lost their class-IX myosin. Our model places the choanoflagellates to the Coelomata that invented the related class-X, class-XV, and class-XXII myosins. After separation of the choanoflagellates, the Bilateria gained another three classes, class-III, class-XIX, and class-XX. The Deuterostomia, to which we placed the Cnidaria, have invented the class-XXVIII myosins and lost class-XXII. Later in evolution, the Chordata have lost the class-XX. This model proposes the continuous invention of new myosin classes over a relatively long time and the subsequent loss of single myosin classes by certain species and lineages. The placement of the Cnidaria to the Deuterostomia surprises as the Cnidaria are commonly considered to be a sister group of the Bilateria. However, the analysis of the *Nematostella vectensis* genome showed that, from a genomic perspective, *Nematostella* more closely resembles modern vertebrates than the fruit fly or nematodes (65) which is consistent with our analysis. But as long as genome sequences of further Cnidaria species are not available this placement could also be the result of long branch attraction effects in the phylogenetic tree. Sequencing of further species of the lineages Choanoflagellida, Cnidaria, and Echinodermata, which are only represented by single species yet, will provide better pictures of these taxa, as have been obtained for the nematodes, Arthropoda, and vertebrates that show a wide distribution of the myosin content between the member species. For example, during the evolution of the Arthropoda the Insecta lost the class-XIX myosin. Later in evolution the ancestor of all *Drosophila* species lost the class-III and class-IX myosins, and finally most *Drosophila* species lost the class-XXII myosin. Most of the lineages like the Nematoda, Arthropoda and Vertebrata have developed further branch-specific myosins. We propose, that sequencing of related organisms to *Strongylocentrotus purpuratus* and *Monosiga brevicollis* will result in the classification of their orphan myosins and thus also in branch-specific myosins for these lineages.

In contrast, the metazoan tree based on classical taxa and nodes shows the invention of ten myosin classes in a very short time scale (Figure 4.9). The evolution of the Metazoa would thus mainly be characterized by gene losses. While the Anthozoa *Nematostella vectensis* shares all its twelve myosin classes with vertebrates, the nematodes must have lost six of the thirteen common Metazoa myosin classes. The nematode *Trichinella spiralis* has lost another three of the remaining classes sharing only four classes with the other Metazoa. The Arthropoda must also have lost at least two of the common Metazoa myosin classes. This scenario, the invention of ten myosin classes during the evolution of only two taxa nodes and the subsequent major losses of myosin classes until the final speciation, seems very unlikely compared to the other model that proposes the invention of new myosin classes over a long period with the subsequent loss of single classes.

In both models, the tree of myosin diversity gives clear support for the classical Coelomata hypothesis that groups Arthropoda with Deuterostomia in a monophyletic class. The Nematoda sequenced so far lack four classes that the Arthropoda share with the vertebrates. It is very unlikely that the Nema-

toda have lost just these four classes and not one or more of the others. The class specific phylogenetic trees show that the Nematoda myosins always separate before the Arthropoda-Deuterostomia split, except for the class-IX myosins where the Nematoda and Arthropoda homologs group separately from the Deuterostomia homologs. These findings illustrate the advantage of analyzing the diversity of a large protein family in contrast to looking at single-gene phylogenies that have supported the monophyletic grouping of Nematoda and Arthropoda in some cases (66).

The comparative analysis of the phylogenetic relationship of the species in single myosin classes showed several incongruities. We hypothesized that the myosin genes of the corresponding organisms might have evolved asynchronously as has been observed for a number of yeast genes (67). From the phylogenetic tree we therefore determined the distances between pairs of sequences. To compensate for differences in general diversity within each class, all distances were normalized. Asynchronous evolution is visualized by the comparison of the deviation from the mean distances. As examples we analysed the myosins of completely sequenced mammalian (Figure 4.10) and fungal genomes (Figure 4.11). As expected, all Primates are very closely related, with the chimpanzee generally closer to *Homo sapiens* than to macaca. The myosin proteins from dog and cow are closer related to those of the primates than to those from rodents. The opossum *Monodelphis domestica* is in general the most divergent mammalian, although in the case of Myo1E and Myo16 it is closer related to the dog and the Primates than to the rodents. The myosin proteins from cow show the most asynchronous phylogenetic relationship of the analysed mammalian genomes. They either diverge before the split of the rodents and primates/dog, after this split, or form a monophyletic class with the corresponding dog orthologs. Hence, it is either not possible to resolve the phylogenetic grouping of the cow in general, or not by using the myosin proteins, or sequences from additional mammals have to be added for better resolving the tree.

The fungal myosins show several distinct groups that are related to the established taxa. However, the analysis resolves some so far unrecognised relationships. The Saccharomycotina do not group to the Ascomycota in all myosin classes, but have evolved asynchronously. Based on our analysis of the myosins the Saccharomycotina should be considered as an independent clade that evolved from Fungi, in parallel to the Ascomycota, the Basidiomycota, the Zygomyocota, and the Schizosaccharomycetes. These clades developed very asynchronously so that their phylogeny cannot be resolved. In addition, the species in these clades have undergone considerable asynchronous development. *Yarrowia lipolytica* that has been considered a yeast species is closer related to the Ascomycota than to the Saccharomycotina, both based on the phylogenetic relation of the respective myosin homologs and based on its myosin content containing a class-XVII myosin that all Saccharomycotina have lost.

How did the very first myosin look like? In the beginning of eukaryotic evolution, the myosin motor domain had been developed (Figure 4.12). During subsequent early evolution an extensive process of domain fusions started, during which the carboxy-terminal IQ motif was added first. After duplication of this gene, the amino-terminal SH3-like domain was fused to the motor domain. These two domain organizations are shared by myosins of all species. The class-I myosins show the widest taxonomic distribution, are devoid of the amino-terminal SH3-like domain and thus suggested to be

the first distinct myosin-class evolved. We propose that the most ancient myosin motor domain had a sequence very close to that of the class-I myosins.

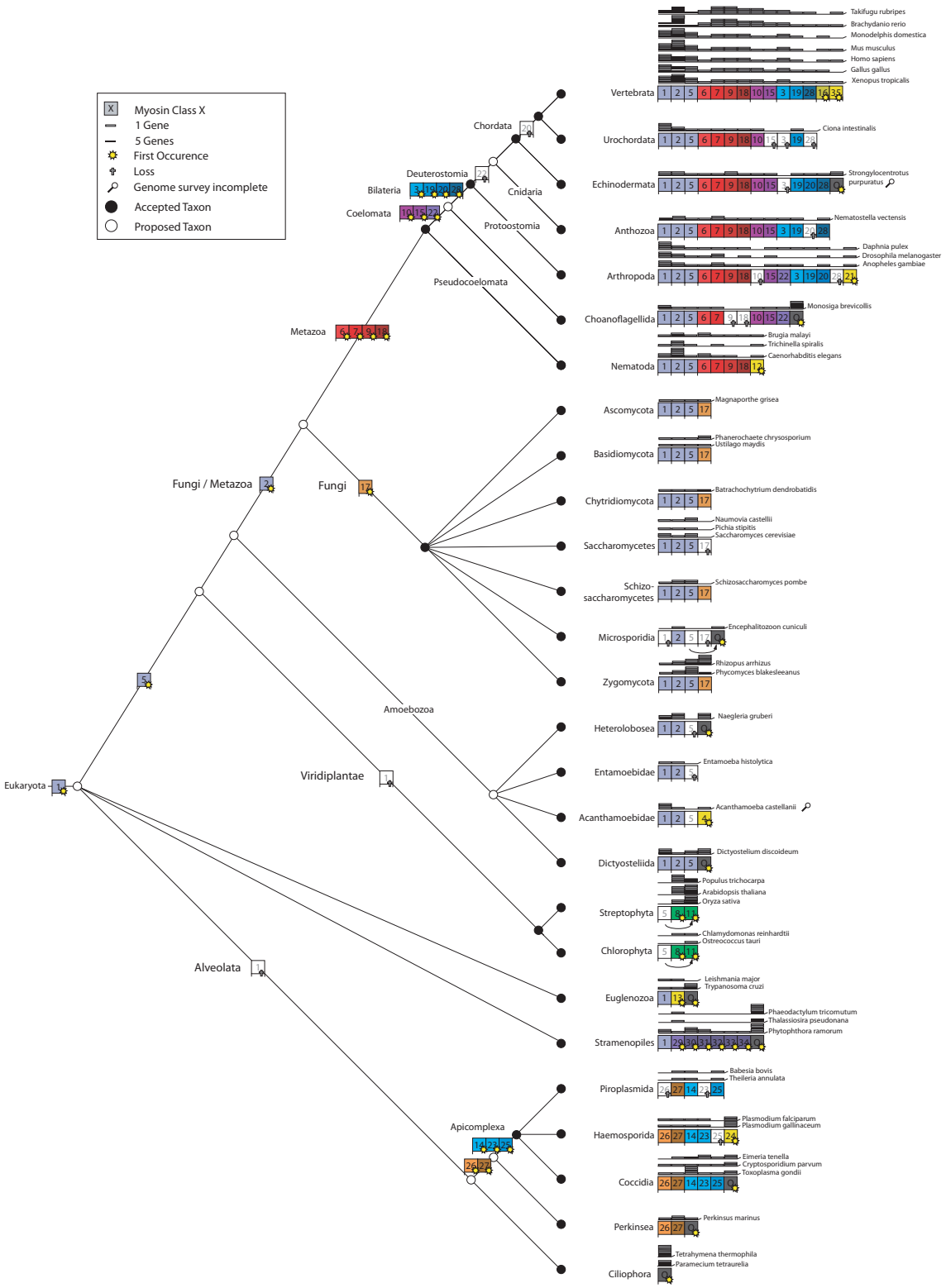

Figure 4.8: Schematic drawing of the evolution of myosin diversity.
The tree has been constructed based on the combination of the phylogenetic information obtained from the analysis of single myosin classes as well as the analysis of the class distribution of major taxa (see Materials and Methods). Thus, branch lengths do not correspond to any scale. Nodes that have already been suggested are symbolized by filled circles. Nodes that we propose base on the analysis of the myosins are represented by open circles. The exact myosin contents of several representative organisms are given. The myosin inventory of all 328 organisms is available from additional data file 3.

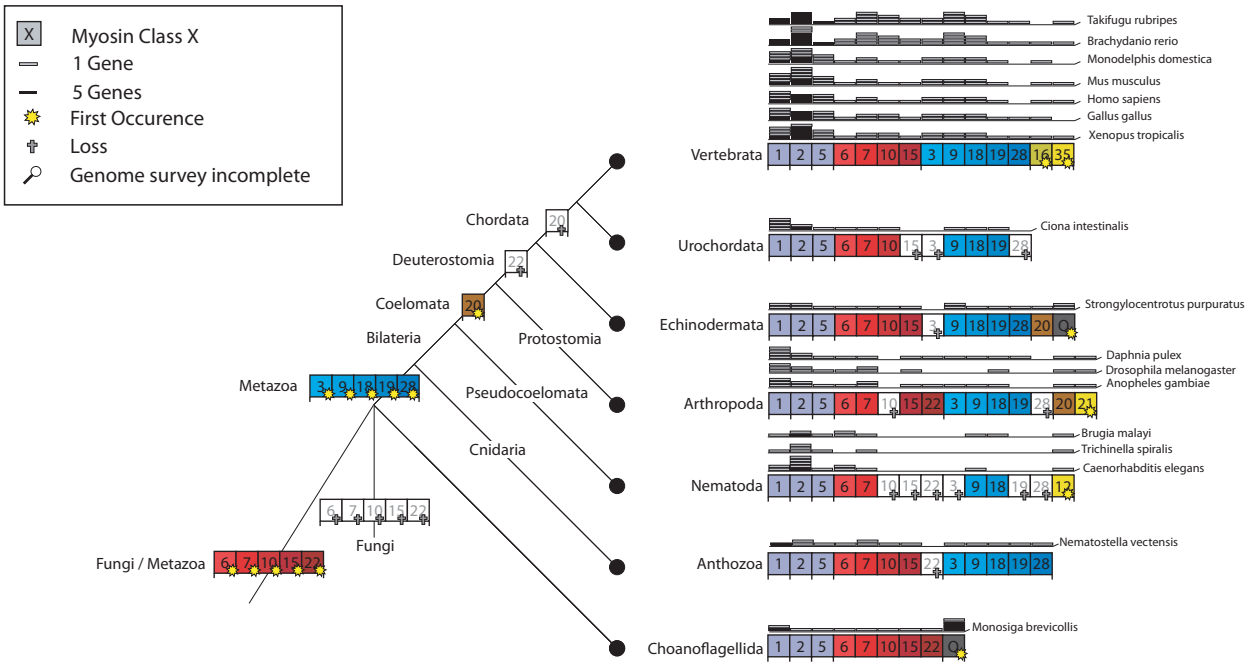

Figure 4.9: Schematic drawing of the evolution of myosin diversity in the Fungi/Metazoa lineage based on the 'accepted' taxonomy.
The inventions and losses of the myosin classes have been plotted onto the 'accepted' phylogeny of the Eukaryotes that is available at NCBI. Branch lengths do not correspond to any scale.

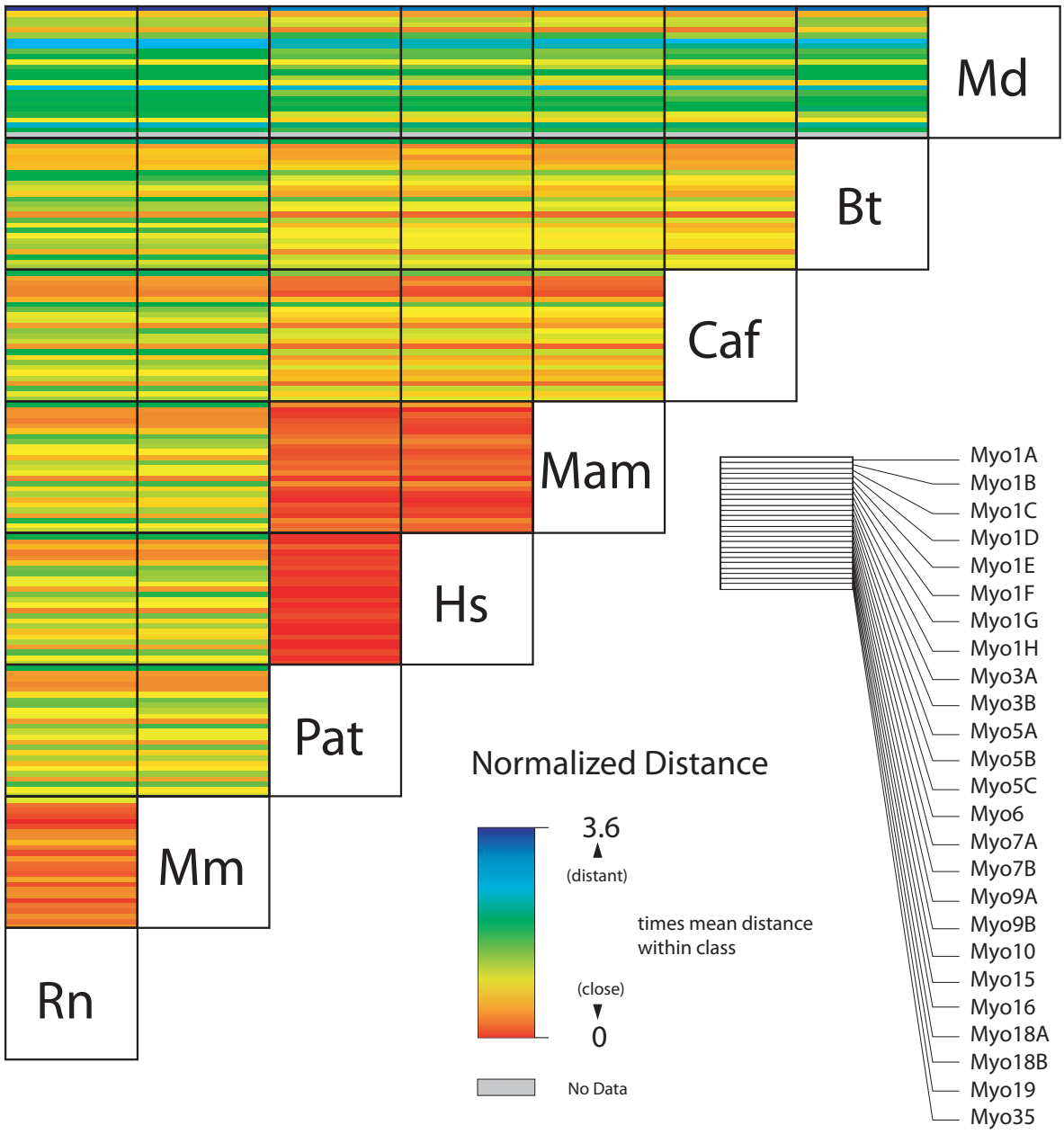

Figure 4.10: Asynchronous evolution of mammalian myosin proteins.
The matrix illustrates the normalized distances between corresponding sequences. Asynchronous evolution is observed if the pattern of the deviation from the mean is different. For example, the pattern from rat to the other mammalian species is very similar illustrating their synchronous evolution in general. However, there are differences in the patterns of some class-I myosins between rat and mouse and opossum indicating their asynchronous evolution. In contrast the sequence comparison patterns of cow and the other mammalians are very different, indicating the asynchronous evolution of all cow myosin genes. The abbreviations for the organisms are: Rn = *Rattus norvegicus*, Mm = *Mus musculus*, Pat = *Pan troglodytes*, Hs = *Homo sapiens*, Mam = *Macaca mulatta*, Caf = *Canis* familiaris, Bt = *Bos taurus*, Md = *Monodelphis domestica*.

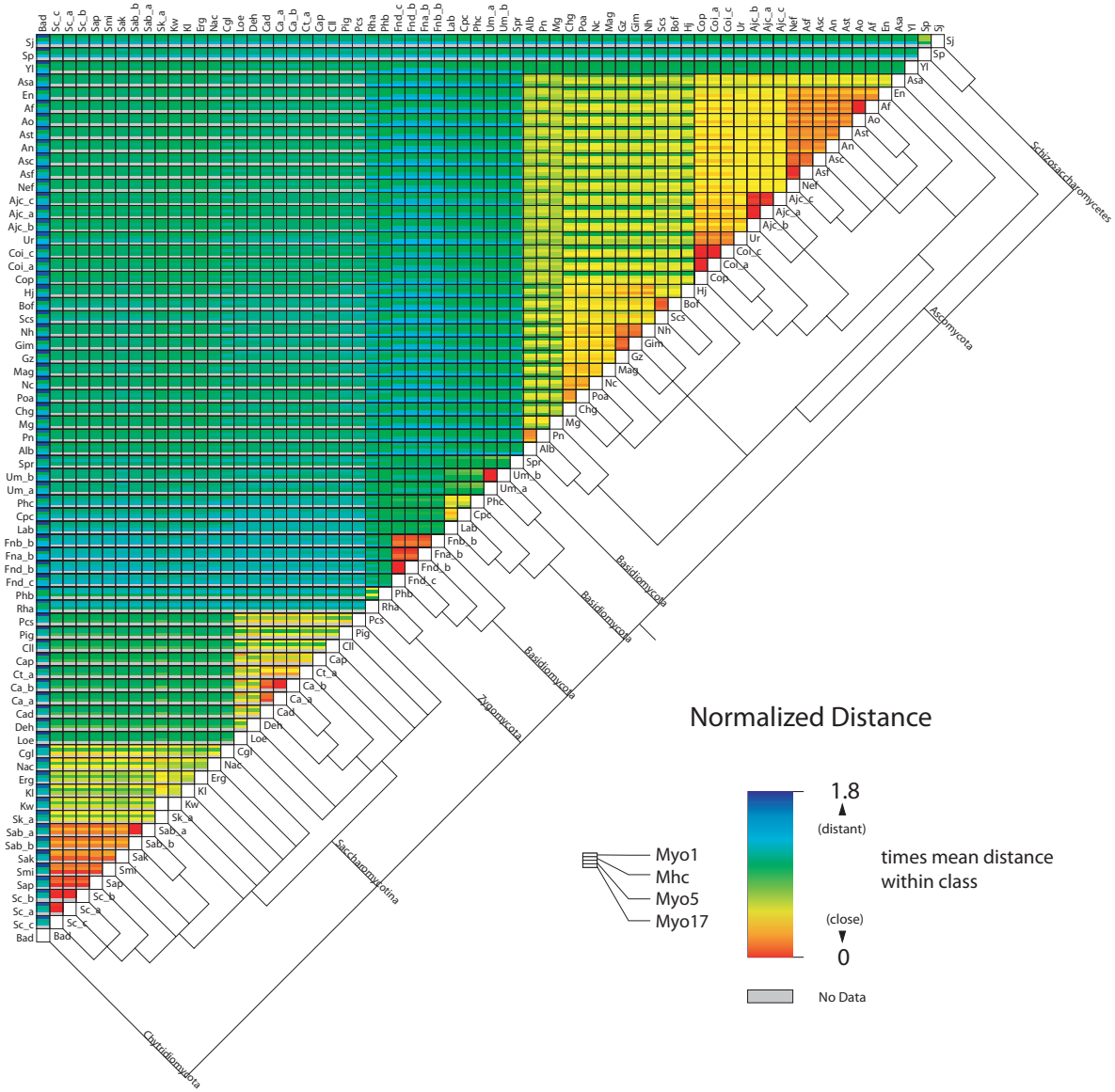

Figure 4.11: Asynchronous evolution of fungi myosin proteins.
The matrix is shown in a similar way as in Figure 4.10. The consensus tree from the analysis of the single myosin class trees is shown. The obtained polytomic tree is the result of the asynchronous evolution of the different species. The abbreviations for the organisms are listed in (10)

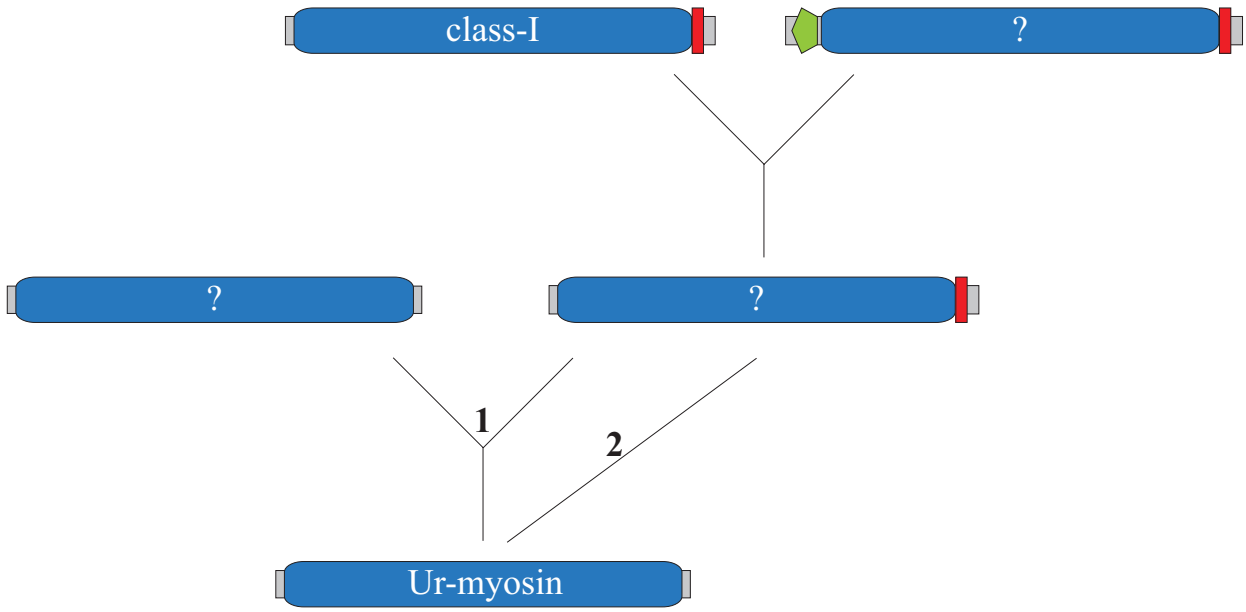

Figure 4.12: Evolution of the first myosins.
The first myosin is expected to consist only of the myosin motor domain and called urmyosin. By domain fusion it either accomplished the IQ motif directly carboxy-terminal to the motor domain (2), or after a gene duplication event (1). After a further gene duplication event, this myosin developed to the class-I myosins as well as the ancestor of most of the other myosin classes after the fusion with an SH3 domain (that developed to the N-terminal SH3-like domain).

4.5 Conclusions

Here, we presented the phylogenetic analysis of 2269 manually annotated myosin proteins. The previously assigned 19 myosin classes were confirmed and 16 new classes with unique domain organisations defined. A phylogenetic tree has been constructed including information about the class distribution and evolution in certain taxa as well as the phylogenetic information contained in class-specific subtrees. The analysis showed the Choanoflagellida as part of the Metazoa lineage and the cnidaria (*Nematostella vectensis*) to diverge after the separation of Deuterostomia and Protostomia. The myosin data shows, that several taxa have evolved asynchronously, for example the Mammalia and the Fungi.

The presented tree will increase in resolution as more organisms get sequenced. To increase the fine resolution more sequences of intermediate taxa, e.g. in the metazoan lineage, are needed. For some major taxa a significant amount of species has to be sequenced to get the resolution already obtained for the fungi and metazoan. For example, only eight species of the Viridiplantae have completely been sequenced so far. Especially sequencing of further underrepresented taxa will increase myosin diversity. The myosin data presented here will allow the correct annotation and classification of all upcoming homologs. We hope that the CyMoBase (37), that stores and presents all related information, will be an invaluable tool in all areas of myosin and motor protein research, and in classical taxonomy.

4.6 Materials and Methods

4.6.1 Identification of myosin family proteins

Myosin genes have been identified in iterated TBLASTN searches of the completed genomes of 181 organisms starting with the protein sequence of DdMhcA. All hits were manually analysed at the genomic DNA level. The correct coding sequences were identified with the help of the multiple sequence alignment of the myosins. As the amount of myosin sequences increased (especially the number of sequences in classes with few representatives), many of the initially predicted sequences were reanalysed to correctly identify all exon borders. Where possible, EST data has been analysed to help in the annotation process. Now, all designated myosin classes contain enough members to correctly predict any additional member sequence in the future. However, some of the orphan myosins (e.g. from *Tetrahymena thermophila* and *Paramecium tetraurelia*) might still contain wrongly predicted exons in the tail regions, because sufficient comparative genomic data is not yet available. For some organisms only EST data is available to date, and myosin sequences identified in these databases have been included in the analysis as long as the sequences contain at least 100 residues. These short sequences cannot be and have not been used in the phylogenetic analysis but are important to define the myosin inventory of as many organisms as possible. In addition to the analysis of these large-scale sequencing projects, all myosin sequences in the nr database at NCBI have been collected and reanalysed. Many of these sequences contain sequencing errors, mispredicted exons, and wrongly predicted gene borders.
Some of the genes contain alternative splice forms for the motor domain. The different splice forms were not considered independently in the analysis but in all cases the same splice forms were taken for homologous myosins. All sequence related data (names, corresponding species, GenBank ID's, alternative names, corresponding publications, domain predictions, and sequences) and references to genome sequencing centers are available through CyMoBase (37). The annotated sequences and the alignment of the motor domain sequences are available as AdditionalFile5 and AdditionalFile6, respectively.

4.6.2 Building trees

The phylogenetic tree was built based on a manually constructed and maintained structure-guided multiple sequence alignment. The phylogenetic tree is unrooted and was generated using neighbour joining and the Bootstrap (1,000 replicates) method as implemented in ClustalW (standard settings) (68). The phylogenetic tree presented in additional data file 2 was visualized using TreeDyn (69). The schematic tree was constructed using the following criteria. The myosins are separated in 35 classes. Thus, we assigned a class inventory to every organism. The myosin classes are not only well-separated based on their motor domains but also due to the unique composition of their tail domains. Thus, we conclude that species having myosins of the same class must have had a common ancestor. It is extremely if not completely unlikely that myosins with these distinct features have been invented independently. The other criterion is the analysis of the different trees of the myosin classes. Looking

at the tree of a single class, e.g. the class VI myosins, it is obvious that certain taxa always separated earlier than others, e.g. the Arthropoda myosins always separated before the mammalian myosins. In the next step we ordered all species according to their class inventory minimizing the number of myosin class inventions and losses. If taxa have the same class inventory we used the data from the single class trees to resolve their phylogenetic relationship. If the phylogenetic relationships of taxa in single classes contradicted each other we hypothesized asynchronous evolution and did not resolve the relationship between these taxa.

4.6.3 Distance Maps

The distances between two sequences of one class/variant where obtained from the distance matrix produced by ClustalW using default substitution matrix BLOSUM62 (70). We collected the distances between all sequences of each class/variant. The distances within each class/variant were normalized by dividing by the mean distance of the class/variant. This set the mean distance of each class/variant to 1. In the distance maps the normalized distances were visualized in blocks, each block representing the distances of all sequences from two species. Asynchronous phylogenetic relationships are visible as colour fluctuations within a block.

4.6.4 Domain and motif predictions

Protein domains were predicted using the SMART ((71, 72)) and Pfam ((73, 33)) web server. The prediction of coiled-coils is based on the coils program. The IQ motifs and N-terminal domains were predicted manually based on the homology to similar domains of the other myosins. The recognition motifs included in the SMART and Pfam databases are too restrictive, as the motifs have been created based on the small datasets available some years ago. The domain profiles of the other domains have not been revised yet.

4.7 Authors' contributions

F.O. performed data analysis. M.K. designed the study, assembled and annotated all sequences, and performed data analysis. Both authors wrote and approved the manuscript.

4.8 Supplementary Material

4.9 Acknowledgements

We would like to thank H. D. Schmitt and M. Schliwa for invaluable discussions and C. Griesinger for discussions and continuous support. M.K. was supported by a Liebig Stipendium of the Fonds der Chemischen Industrie, which is in part financed by the BMBF. This work has been funded by grant I80798 of the VolkswagenStiftung and grant KO 2251/3-1 of the Deutsche Forschungsgemeinschaft.

Part III

Manuscripts in Revision

5 Comparative genomic analysis of the arthropod muscle myosin heavy chain genes reveals a new type of partially processed pseudogenes and a model for the process of alternative splicing

Florian Odronitz[1] and Martin Kollmar [1*]

[1]Department of NMR-based Structural Biology, Max-Planck-Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Goettingen, Germany
[*]Corresponding author.

Submitted to:

BMC Molecular Biology

5.1 Abstract

5.1.1 Background

Alternative splicing of mutually exclusive exons is an important mechanism for increasing protein diversity in eukaryotes. The insect Mhc (myosin heavy chain) gene produces all different muscle myosins as a result of alternative splicing in contrast to most other organisms of the Metazoa lineage, that have a family of muscle genes with each gene coding for a protein specialized for a functional niche.

5.1.2 Results

The muscle myosin heavy chain genes of 22 Arthropoda species ranging from the waterflea to wasp and *Drosophila* have been annotated. The analysis of the gene structures allowed the reconstruction of an ancient muscle myosin heavy chain gene and showed that during Arthropoda evolution introns have been lost in these genes. Surprisingly, the genome of *Aedes aegypti* contains another and that

of *Culex pipiens quinquefasciatus* two further muscle myosin heavy chain genes, called Mhc3 and Mhc4, that contain only one variant of the corresponding alternative exons of the Mhc1 gene. Mhc3 transcription in *Aedes aegypti* is documented by EST data. Mhc3 and Mhc4 inserted in the *Aedes* and *Culex* genomes either by gene duplication followed by the loss of all but one variant of the alternative exons, or by incorporation of a transcript of which all other variants have been spliced out retaining the exon-intron structure. The second and more likely possibility represents a new type of a partially processed pseudogene.

5.1.3 Conclusions

Based on the comparative genomic analysis of the alternatively spliced arthropod muscle myosin heavy chain genes we propose a splicing process. This process consists of the splicing of the mutually exclusive exons until one exon out of the cluster remains while retaining surrounding intronic sequence. In a second step splicing of introns takes place. A related mechanism could be responsible for the splicing of other genes containing mutually exclusive exons.

5.2 Background

Alternative splicing is an important and widespread mechanism that is used by higher organisms to express molecularly distinct mRNAs in response to developmental and cellular contexts (74, 75). Mutually exclusive splicing, in which only one exon is chosen out of a cluster of alternative exons arranged in a tandem array, is the most frequent event on a genome-wide level (76, 77). Several mechanisms have been proposed that explain why only one of the two or more variants is included in the mature mRNA (78, 79, 80). Mostly, Metazoa contain mutually exclusive exons only in pairs. Extreme cases for mutually exclusive splicing are the insects Dscam genes that have arrays of up to 52 variants as observed in the *Drosophila* Dscam gene (81). A less dramatic example is the mutually exclusive spliced *Drosophila* muscle myosin heavy chain gene that can potentially produce 480 different mRNAs (82).

Myosins comprise a large superfamily of actin-based motors that fulfill a variety of cellular functions from cell division, cellular locomotion, and vesicle transport to muscle contraction (21, 50). 35 classes of myosins have been identified to date with each class being responsible for a different function (83, 56, 84). The first myosin was identified in skeletal muscle tissue over hundred years ago (for a review about the history of muscle myosin see (85)) and, since different myosins turned up, it has been referred to as conventional myosin or class-II myosin. Class-II myosins comprise the largest and most extensively studied class not only because the metazoan species are the most studied organisms but also because this class contains the most isoforms per organism.

Drosophila melanogaster contains two class-II myosin genes, one encoding the muscle isoforms (Mhc) and one the nonmuscle isoform (zipper) (86). The Mhc gene produces all different muscle myosins as a result of alternative RNA splicing (82). This is in contrast to the organisms of most other taxa of the Metazoa lineage, that have a family of muscle myosin heavy chain genes with each gene coding

for a protein specialized for a functional niche. For example, the nematode *Caenorhabditis elegans* expresses six muscle myosins (56), while the ascidian *Ciona intestinalis* genome contains five muscle myosin heavy chain genes (87) and vertebrate genomes encode up to 22 muscle myosin heavy chain isoforms (83).

The *Drosophila* Mhc gene consists of 30 exons including five clusters of alternatively spliced exons and one differentially included penultimate exon. Thus, 480 combinations of alternative exons are possible. The four alternative exons in the motor domain part of the gene code for 120 different variations of the motor domain. In contrast to the muscle myosins of the other metazoa species, changes modulating myosin function are thus limited to four regions in the head domain. These discrete regions of sequence variation have been shown to produce physiological differences among the various muscle types (88). Although many variations are possible and all alternative exons get expressed at some point in *Drosophila*'s life, only a limited number of combinations seem to be employed. For example, during *Drosophila* embryogenesis only seven Mhc transcripts have been found to be expressed (88).

The genome of *Drosophila melanogaster* was the third eukaryotic genome to be completely sequenced (89). Since then, the number of sequenced organisms has increased rapidly. Of the Arthropoda phylum, the genomes of the mosquitos *Anopheles gambiae* (90) and *Aedes aegypti* (91) and the silkworm *Bombyx mori* (92) have been published, and 17 further insect genomes have been finished of which eleven belong to the *Drosophila* species group (93).

Originally, pseudogenes have been defined as DNA sequences that are derived from functional genes, but acquired such degenerative features as premature stop codons and frameshift mutations, which make them unable to produce functional proteins (94,95,96). Non-processed pseudogenes are thought to result from tandem duplications of genes with subsequent accumulation of disabling mutations. Processed pseudogenes lack introns and presumably arise by retrotransposition of a mature messenger RNA (mRNA). While non-processed pseudogenes are commonly found near the functional original gene, processed pseudogenes are randomly inserted into the genome. Also, partially processed pseudogenes have been reported that sometimes contain the complete coding region (97, 98). Recent studies have shown, that pseudogenes are not just 'Junk' DNA but often exhibit functional roles (for a review see (95)).

Here, we report the comparative genomic analysis of the muscle myosin heavy chain genes of all Arthropoda species that have completely been sequenced so far. On this basis we propose a model for the process of alternative splicing that involves the splicing of all unwanted alternative versions of an exon while retaining intronic sequence around the remaining variant.

5.3 Results

5.3.1 Identification and annotation of the muscle myosin heavy chains

The arthropod muscle myosin heavy chain genes were identified by TBLASTN searches against the corresponding genome data of the different species using the *Drosophila melanogaster* protein as

query (Figure 5.1). The species analysed were the mosquitos *Aedes aegypti*, *Culex pipiens quinquefasciatus* and *Anopheles gambiae*, the silkworm *Bombyx mori*, the honeybee *Apis mellifera*, the jewel wasp *Nasonia vitripennis*, the waterflea *Daphnia pulex*, the rust-red flour beetle *Tribolium castaneum*, the body louse *Pediculus humanus corporis*, and thirteen *Drosophila* species (Table 5.1). According to the general nomenclature for myosin sequences (83) the alternatively spliced muscle myosin heavy chain genes are named Mhc1, and the non-muscle myosin heavy chain genes are denoted Mhc2. The sequences were assigned by manual inspection of the genomic DNA sequences. Exons have been confirmed by the identification of flanking consensus intron-exon splice junction donor and acceptor sequences (Figure 5.1) (99). Because of the five to nine mutually exclusive exons and the included or excluded penultimate exon, automatic identification of all exons failed. The genomic sequences of *Apis mellifera* and *Bombyx mori* contain several gaps that at least in one case must have contained missing exons. The cellular expression of the myosin genes including the transcription of some of the mutually exclusive exons has been confirmed by analysis of corresponding EST data.
The untranslated first exons of the genes have been assigned by analysing EST data, if possible. Because untranslated 5' exons were found for all those species for which EST data covering the amino-termini of the genes is available, it is expected that the other arthropod myosin genes also contain untranslated first exons. Accordingly, the unambiguously identified exons have been numbered starting with exon two. Duplicated exons were named in alphabetical order according to the direction of transcription, the exception being the alternatively spliced exon 11 of the *Drosophila* Mhc1 of which the first of the mutually exclusive exons was named 11e for historical reasons (82). The differentially included penultimate exons of the *Drosophila* species have been predicted based on their similarity at the DNA level. Although this exon mainly consists of untranslated bases and its identity between the *Drosophila* species is almost as low as that found in intron regions, the exon borders are conserved enough to be recognised. The carboxy-terminal exons of the other arthropod Mhc1 genes have been confirmed by analysing EST data, if possible. For TicMhc1 and DapMhc1 only one carboxy-terminal exon could be confirmed by EST data. However, given the exon conservation between all arthropod Mhc1 genes it is expected that both genes contain another carboxy-terminal exon. For *Nasonia*, EST data is not available. The carboxy-terminal exon of the NavMhc1 gene was identified based on its homology to the other Mhc1 exons. An exon corresponding to the penultimate exon of the other genes could not be identified. The *Drosophila* sp. Mhc1 genes, the AeaMhc1 and the CpqMhc1 gene contain consensus polyadenylation signals AATAAA, while the Mhc1 genes of *Ang*, *Am*, *Dap*, *Nav*, *Pdc*, and *Tic* contain polyadenylation signals of type AAAAAA. For the DmMhc1 gene it has been shown that the use of either polyadenylation site is not regulated (100, 101) and the same might be true for the two or multiple polyadenylation sites of the other arthropod genes.

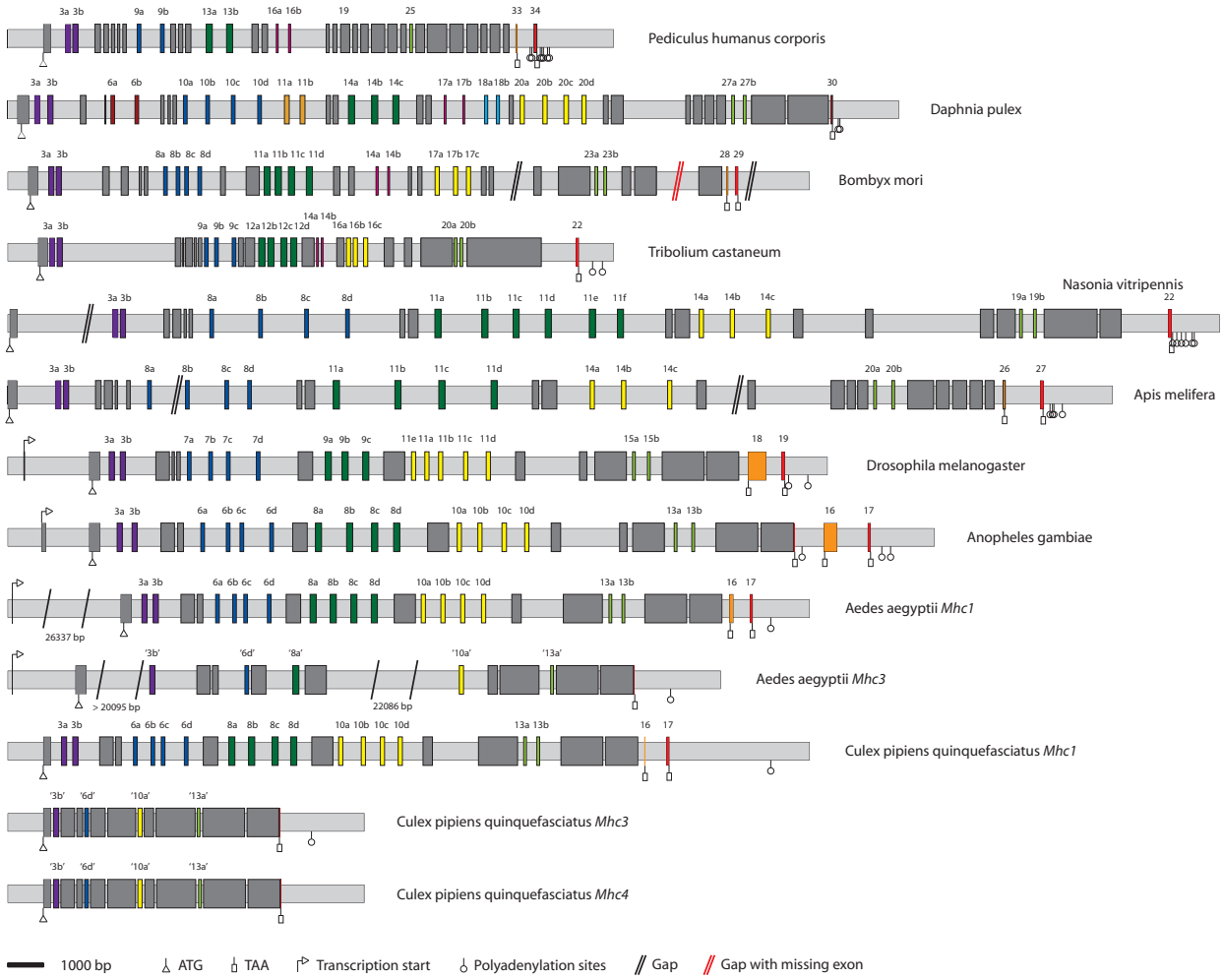

Figure 5.1: Diagram of the arthropod Mhc1 genes with exon-intron structure
The gene structures of the arthropod muscle myosins genes are shown using the following color code: light-gray: intron sequences; dark-gray: common exons; colored: alternatively spliced exons. The Drosophila melanogaster Mhc1 gene is shown as representative for all Drosophila sp. Mhc1 genes, because their gene structures only differ in the length of the introns. The transcriptional and translational start sites, the stop codons and polyadenylation sites are shown if they have been determined. Some genes are spread on several contigs. The corresponding gap positions are shown in black, if further exons are not expected, and in red, if exons are definitively missing. The genes are drawn to scale except for the Aedes aegypti genes where the extremely long introns have been shortened. Gaps have been filled with 100 bp although their exact length is unknown.

5.3.2 Identification of further muscle myosin heavy chain genes in Aedes aegypti and Culex pipiens quinquefasciatus

Surprisingly, a second muscle myosin heavy chain gene has been identified in *Aedes aegypti* (Figure 5.1) and named Mhc3. The Mhc3 gene contains the same exon organisation as Mhc1 except that it does not have any alternatively spliced exons and misses the two carboxy-terminal exons (Figure 5.1). Many EST clones provide supporting evidence for the deduced carboxy-terminus, the amino-

terminal untranslated exon1, and other parts of the gene. The exons related to the alternatively spliced exons of Mhc1 are either identical ("exon3b") or very similar to one of the Mhc1 exons. The protein sequence of Mhc3 has an overall sequence identity of 91.4 % to Mhc1. Besides the different carboxy-termini, the largest differences are in loop-1, which is three residues shorter in Mhc3, and in loop-2, which has only six instead of ten glycines and might therefore be structurally more restricted. The *Culex pipiens quinquefasciatus* genome decodes another two muscle myosin heavy chain genes that are very similar to each other and have been named Mhc3 and Mhc4 (Figure 5.1). Both have the same exon organisation as the CpqMhc1 gene except that they do not have any alternatively spliced exons and miss the two carboxy-terminal exons. Another difference is that alternative exons 8 fused to the following constitutive exons in the Mhc3 and Mhc4 genes. The protein sequence identity between CpqMhc3 and CpqMhc4 is 92.0 %, the identity to CpqMhc1 is 84.4 % and 90.4 %, respectively. Surprisingly, AeaMhc3, CpqMhc3 and CpqMhc4 retained the same variants of the alternatively spliced exons of the corresponding Mhc1 genes.

5.3.3 The BmMhc1, TicMhc1, PdcMhc1 and DapMhc1 genes contain further alternatively spliced exons

The analysis of the BmMhc1, TicMhc1, PdcMhc1, and DapMhc1 genes revealed further alternatively spliced exons compared to the DmMhc1 gene. All further alternative exons encode for sequence that is part of the motor domain. The additional alternative exon of *Bm*, *Pdc* and *Tic* is conserved between these three organisms, and also shared with *Dap*. It is located between the alternatively spliced exons 11 and 17 (*Bm*), alternative exon 13 and constitutive exon 19 (*Pdc*), and alternative exons 12 and 16 (*Tic*), respectively, and separated from the neighbouring alternatively spliced exons by constitutively expressed exons (Figure 5.1). In contrast to the other alternatively spliced exons, these alternatively spliced exons are different in length and amino acid conservation. The first part of the exon encodes part of loop-2 (see below), that is a very flexible loop involved in actin-binding. In the arthropod genes it mainly consists of glycines, arginines, and lysines. Thus, the alternatively spliced exons of *Bm*, *Tic*, *Pdc*, and *Dap* encode different numbers and compositions of these residues. The second part of the alternatively spliced exon is part of the following alpha-helix and hence completely conserved in length and strongly conserved in composition. In addition to this alternatively spliced exon, the DapMhc1 gene contains three further alternatively spliced exons extending its number of alternatively spliced exons to nine (compared to five in *Drosophila*). Alternative exon 6 encodes an alternative P-loop to loop-1 sequence, alternative exon 11 directly follows the alternative exon encoding a structural part near the ATP-binding site, and alternative exon 18 encodes an alternative version of the sequence after loop-2 (Figure 5.1).

5.3.4 The PdcMhc1 gene encodes a strongly reduced set of possible transcripts

The *Pediculus humanus corporis* Mhc1 gene contains the most reduced set of alternative exons (Figure 5.1). It has four sets of alternative exons each comprising two variants. However, the sequence encoding part of the converter domain, which is encoded by sets of three to five alternative

exons in the other arthropod genes, has been fused to the following exon forming one constitutive exon in the PdcMhc1 gene (exon 19, Figure 5.1). Also, the part in the tail domain encoded by a set of two alternative exons in all other arthropod genes is represented by only one exon in the PdcMhc1 gene (exon 25). Altogether, the alternative exons decode for 16 different versions of the motor domain and 32 different mRNAs of the PdcMhc1 gene, compared to 120 different combinations of alternative exons for only the motor domain of the *Drosophila* Mhc1 gene.

5.3.5 Conservation of alternatively spliced exons

The number of variants differs between the Arthropoda species for many of the alternatively spliced exons (Figures 5.1 and 5.2). For the first alternatively spliced exon two variants have been found in all Mhc1 genes. Both differ by two absolutely conserved residues, namely the amino acids alanine and aspartate at positions 25 and 26 in the 'a' variants of the exon that are substituted by serine and asparagine in the 'b' variants (Figure 5.3). A slightly less conserved marker for the 'b' variants is a cysteine at position 21. Variant 3a of the DapMhc1 is an exception as it has an additional residue at the N-terminus compared to the other Mhc1 variant 'a' exons. The DapMhc1 gene encodes three alternatively spliced exons not found in the other arthropod Mhc1 genes. For all three exons variant 'b' is more homologous to the corresponding amino acid sequences of the other Mhc1 proteins than variant 'a' (see Figures 5.2 and 5.4). The alternatively spliced exons of BmMhc1, DapMhc1, PdcMhc1 and TicMhc1 covering loop-2 are different in length and starting position. However, the 'a' variants are more similar to each other than to the 'b' variants and the corresponding amino acid sequences of the other Mhc1 proteins. Thus, the common ancestor of *Bm*, *Dap*, and *Tic* has in all probability already contained an 'a' and a 'b' variant. Completely conserved residues characterizing the 'a' variant are a serine at the end of loop-2, a glutamate at position 3 and a leucine at position 8 of the following helix (s[G/K/R 8-9]**S**[G/A]F[**Q**/M]TVS[S/A]**L**YR). Except for PdcMhc1, all arthropod Mhc1 genes have two variants of the mutually exclusively spliced exon in the tail (Figure 5.2). The most conserved differences between the two variants are an aspartate at position 14 in variant 'b' (either an asparagine or a glutamine in variant 'a') and an asparagine at position 24 (an arginine in variant 'a'). In addition, at position 15 the 'b' variants have a large hydrophobic residue (leucine, methionine, or phenylalanine) while the 'a' variants have a small polar residue (serine or threonine). In contrast to the other Mhc1 genes, the 'a' variant of DapMhc1 is closer related to the 'b' variants than to the other 'a' variants.

The situation is more complex for the remaining mutually exclusive exons that contain three to six variants. The exon encoding a loop-helix motif adjacent to the ATP-binding site (blue color in Figure 5.1 and Figure 5.6) is not as conserved as the other alternatively spliced exons (Figure 5.2). Therefore, it is difficult to identify characteristic residues/motifs for the respective variants. Except for the PdcMhc1 and TicMhc1 genes all genes contain four variants. The variant with the most characteristic residues is variant 'c'. It is characterized by a positively charged residue at position 8 (arginine or histidine), a conserved arginine at position 21, and a conserved asparagine at position 26. None of these residues appear in any of the other variants at the respective positions. The TicMhc1,

PdcMhc1, and DapMhc1 genes have lost this variant. The only strong characteristic of variant 'd' is a conserved isoleucine or valine at position 20 that is found in all Mhc1 genes. Variants 'a' and 'b' do not contain any distinguishing residues. The alternatively spliced exon spanning the relay helix and the relay loop is the longest and most conserved of the mutually exclusive exons. The variability ranges from two variants in the *Pediculus* Mhc1 gene to six variants in the *Nasonia* gene (Figures 5.1 and 5.2). The least conserved part of the exon is the relay loop that is not embedded in the motor domain. In this region, characteristic residues for certain variants are found. Variant 'c' is characterized by a conserved glutamine at position 49 and either a glutamine or an asparagine at position 50. A copy of this variant is present in all Mhc1 genes except that of *Tic*. Another conserved variant is variant 'd' characterized by a glutamine at position 49 followed by a proline at position 50. This variant appears in the Mhc1 genes of *Aea*, *Ang*, *Cpq*, *Tic*, and *Bm*. Similar to the situation for the alternatively spliced exon at the ATP-binding site, the other variants are not conserved enough to define characteristic residues. It is thus not clear which were present in the ancient arthropod gene and which arose through exon duplication in the individual genes. Again, the DapMhc1 is the exception because its first two variants, characterized by two conserved methionines at positions 42 and 55, differ from all other variants.

The variants of the alternative exon encoding part of the converter domain also show a high degree of variability (Figure 5.2). Two of the variants have characteristic features. Variant 'a' is the most conserved of the variants at the protein level having a conserved methionine at position 9 and a conserved cysteine at position 26. These residues do not appear in any of the other variants. Variant 'a' is conserved in the Mhc1 genes of all species and therefore must have been present in their common ancestor. The last of the variants has a characteristic feature at the DNA level. While all other variants of this exon have a GC 5' splice site at the following intron, the intron following the last variant always has a GT 5' splice site. At the amino acid level this variant is characterized by a lysine at position 2, a cysteine at position 5 and a glutamate at position 20.

Wherever EST and/or cDNA data was available a differentially excluded penultimate exon could be identified. These exons are very short (one to thirteen residues) and not conserved, and therefore similar exons have not been predicted for the species for which EST data is not available. For *Ang* three carboxy-termini have been identified. Based on EST data the AngMhc1 transcript may also end with a short extension to the antepenultimate exon. This C-terminus is similar to that found for AeaMhc3 and CpqMhc4 and might be used in a similar combination of the other alternatively spliced exons.

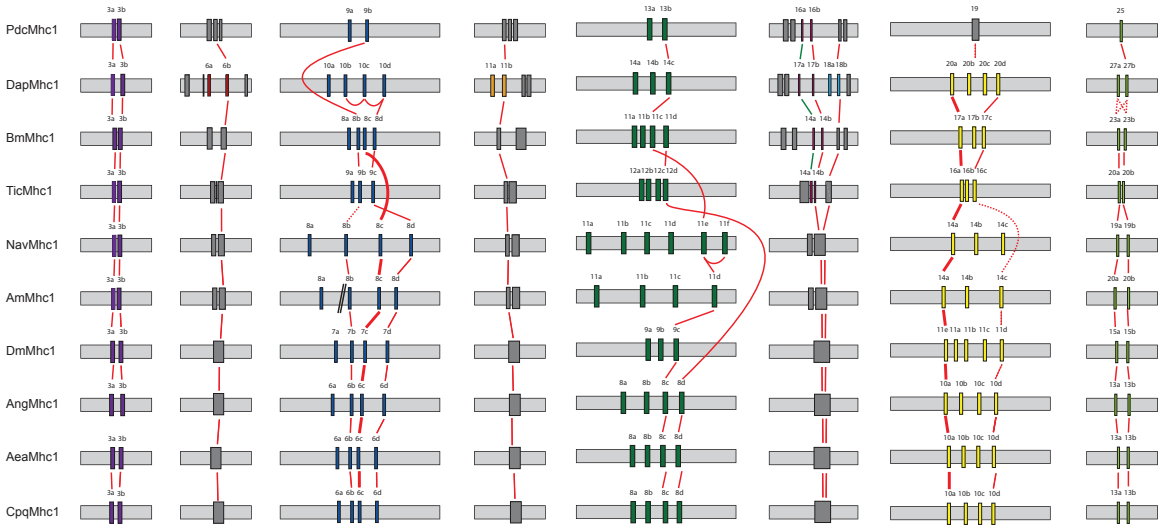

Figure 5.2: Relationships between alternatively spliced exon
Sections of the Mhc1 genes of Figure 5.1 have been aligned showing the relationship between the exon-intron structures of the regions containing alternatively spliced exons. Continuous lines connect variants that are almost identical and thus expected to be derived from a common ancestor. Bold lines connecting alternative exons in regions containing multiple variants per Mhc1 gene highlight particularly conserved exons in these sets. Dotted lines represent putative connections between certain variants although their identity is not very strong on the protein level.

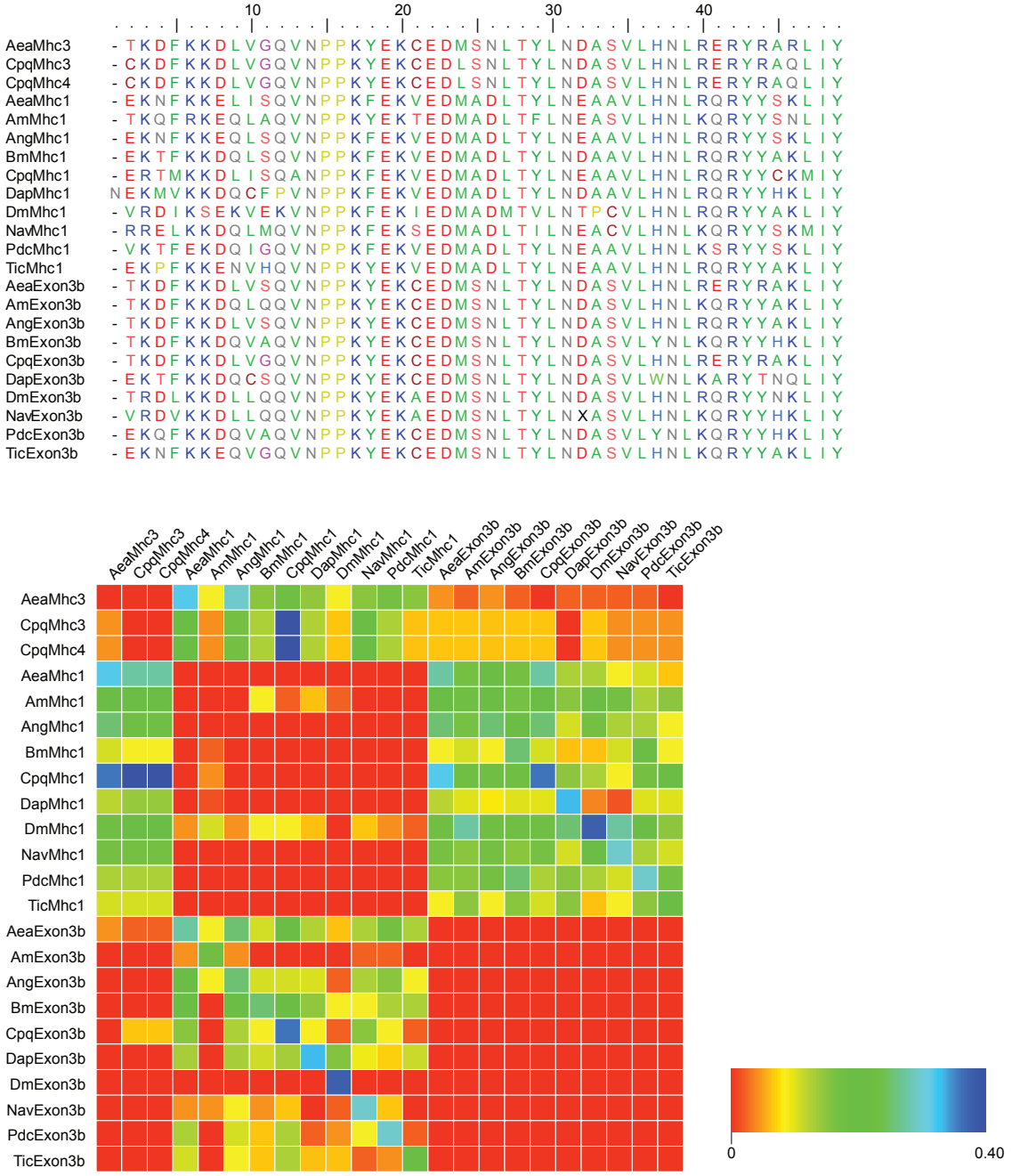

Figure 5.3: Sequence conservation in the first set of the alternatively spliced exons
On top, the protein sequence alignment of the alternative exons is shown. The upper sequences, termed Mhc1, Mhc3, and Mhc4, respectively, represent the variant a exons. Below, the comparison of the sequence identity between each exon and variant 'a' and 'b' of every other Mhc1 protein is shown. The graphic has to be read in columns. The higher identity between an exon listed on top and variant 'a' or 'b' of a certain Mhc1 protein listed on the left side has been set to 1 (red color) while the difference of the lower identity to the value of the higher identity is plotted for the other combination of exons. Thus, in every column the higher identity of the named exon to one of the variants of the other Mhc1 proteins is visualized.

5.3.6 Phylogenetic analysis of the arthropod muscle myosin heavy chain genes

A phylogenetic tree of all arthropod Mhc1 protein sequences, always incorporating the first of the alternatively spliced exons and excluding the differentially included penultimate exon, has been generated (Figure 5.4). In general, the tree reflects the phylogenetic relationship between the species. The AeaMhc3 sequence is most closely related to the CpqMhc3 and the CpqMhc4 sequence implicating that the last common ancestor of *Aedes* and *Culex* already had one of these genes. The phylogeny of the *Drosophila* species slightly differs compared to other analyses (93). Thus, the DaMhc1 sequence would have been expected to separate after the divergence of the DpMhc1 sequence. Similarly, the DseMhc1 gene would have been expected to be the closest relative of the DssMhc1 sequence. Overall, the sequence identity is very high. Between DapMhc1 and the other sequences the identity is 70.6 - 77.9 %, while it is between 77.0 % and 99.7 % between the other species.

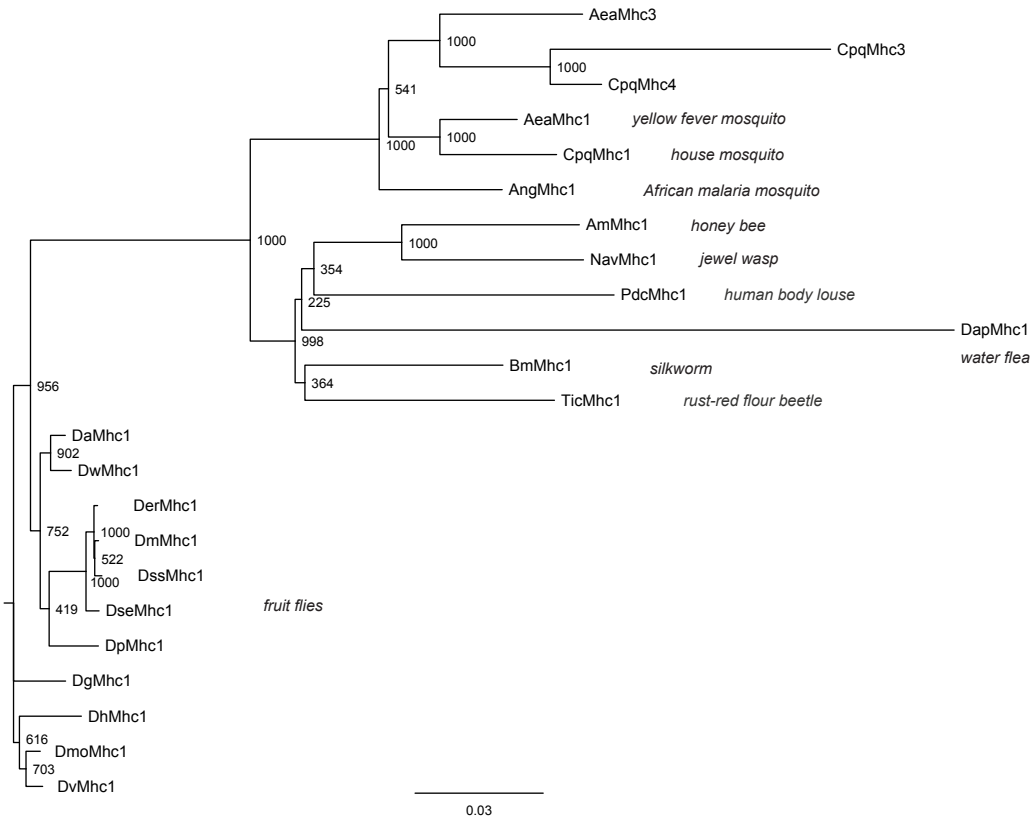

Figure 5.4: Phylogenetic tree of the arthropod muscle myosin heavy chain proteins
The amino acid sequences of the full-length proteins were aligned manually. Because of their incompleteness the sequences of Drosophila persimilis and Drosophila yakuba have been omitted from the tree calculation. Support values for each internal branch were obtained by 1,000 bootstrap steps. The scale bar corresponds to 0.1 estimated amino acid substitutions per site.

5.3.7 Predicting the gene structure of an ancient Mhc1 gene

Whenever intron positions are shared between the genes, the corresponding type of splice site is conserved, with the exception of the shared exon 9 (AmMhc1), exon 10 (TicMhc1), exon 9 (BmMhc1), and the alternatively spliced exon 11 of DapMhc1 (Figure 5.5). All introns have consensus dinucleotide borders except those downstream of the last alternative exons encoding part of the motor domain (exon 11 in DmMhc1), which have a GC dinucleotide at the 5' donor site instead of the consensus GT. The 3' exons of these alternatively spliced exons again have a consensus GT site. *As* exon '10a' of AeaMhc3 is almost identical to exon 10a of AeaMhc1 the following intron also has a GC dinucleotide at the 5' donor site. In contrast to the introns following the exons 9 of AmMhc1, NavMhc1, and BmMhc1, and the intron following exon 10 of PdcMhc1 that have a consensus GT site, exon 10 of TicMhc1 has a GC 5' donor site. The intron following exon 11a of DapMhc1 starts with a consensus GT site, while the intron following exon 11b starts with the absolutely rare GA dinucleotide. Also, all split codons are shared between the genes.
In the part encoding the motor and the neck domain, all intron positions are shared by at least two genes (Figure 5.5). In the coiled-coil tail domain, all genes have lost several introns so that the exons are considerably longer and the intron positions in many cases are not identical. Assuming, that introns have in most cases been lost and were not gained during evolution (102), an ancient arthropod Mhc1 gene can be reconstructed (Figure 5.5). The ancient Mhc1 gene is expected to contain all intron positions that appear in at least one of the analysed Mhc1 genes. In the motor domain, the proposed ancient Mhc1 gene structure completely resembles the DapMhc1 gene. The exon lengths are between 30 and 210 bp. The exons in the tail domain are considerably longer (up to 480 bp).

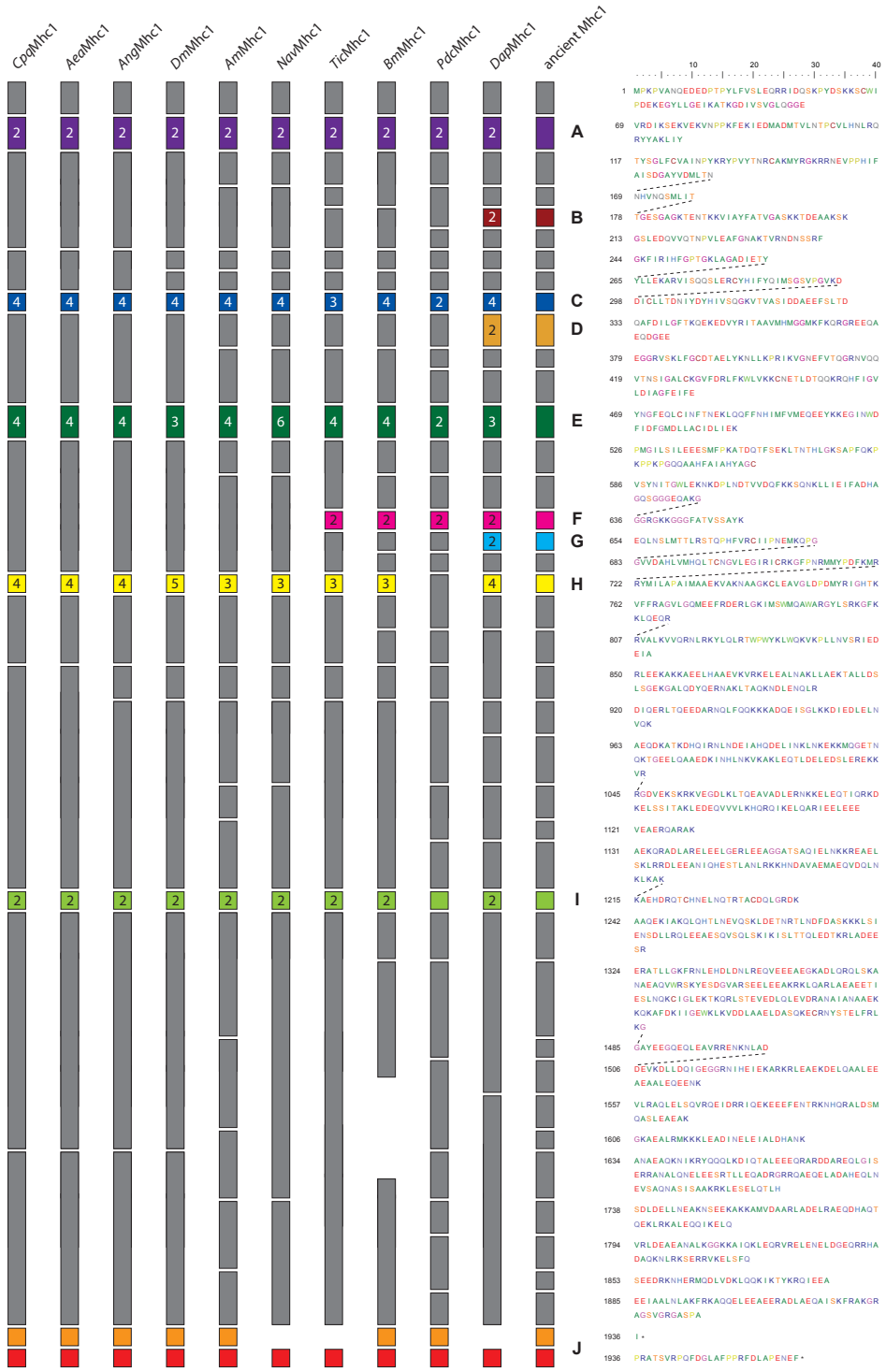

Figure 5.5: Diagram of the arthropod Mhc1 proteins
The exon-intron structure of the Mhc1 genes is shown based on the protein sequence. Exons are shown as boxes while introns are represented by spaces. The same colour scheme has been used as in Figure 5.1. Numbers on alternative exons denote the number of variants. The exons are drawn that the intron positions align between the different Mhc1 genes. Thus, the exon lengths are not drawn to scale (e.g. the exons encoding the variable loop-2 are different in lengths). On the right side, the protein sequence of Drosophila melanogaster Mhc1 is shown as reference. Dotted lines connect amino acids that are derived from split codons.

5.3.8 Structural implications of the alternatively spliced exons

The locations of the alternatively spliced exons of DmMhc1 in the motor domain have been discussed in detail elsewhere (103). The position of the additional alternatively spliced exons of the BmMhc1, TicMhc1, PdcMhc1, and DapMhc1 genes in the structure of the motor domain are shown in Figure 5.6. The alternative exons of DapMhc1 encoding the structural part from the P-loop to loop-1 have identical P-loop sequences. The loop-1 sequences are identical in length but differ significantly in composition. Studies have shown that the flexibility of this loop affects the rate of ADP and phosphate release, with greater flexibility leading to an enhancement in the rate of product release (104). Although the amino acid composition is different between the alternative variants, both contain two glycines and a similar overall charge. Potential ATPase modulating properties must therefore rely on a different mechanism. The alternative exons of DapMhc1 including loop-4 are similar in length and composition. This region of the motor domain has not been investigated so far and therefore functional consequences of differences in the two variants cannot be drawn. Loop-4 has been postulated to be important for the proper localization of class-I myosins that contain elongated loops that sterically interact with actin-binding proteins (105) but the loop-4 sequences are almost identical between the two DapMhc1 variants and the two variants must therefore modulate a different property of the motor domain. The loop-2 sequence is modulated by alternative exons in the BmMhc1, DapMhc1, PdcMhc1, and TicMhc1 genes. By studies of *Dictyostelium* myosin with its loop-2 replaced with the analogous loop from four other myosins with different enzymatic activities, loop-2 was shown to be involved in the weak and the strong binding interactions with actin (106). It also plays an important role in the rate-limiting step of Pi release (107, 108). The exon variants of all three Mhc1 genes have identical numbers of lysine and arginine residues. The 'a' variants are always one residue shorter and have only four instead of five glycines. These differences are, however, very subtle and their influence on actin binding is expected to be very small. The variants of the alternative exon in DapMhc1 following loop-2 are very similar. This part of the motor domain has also not been investigated so far.

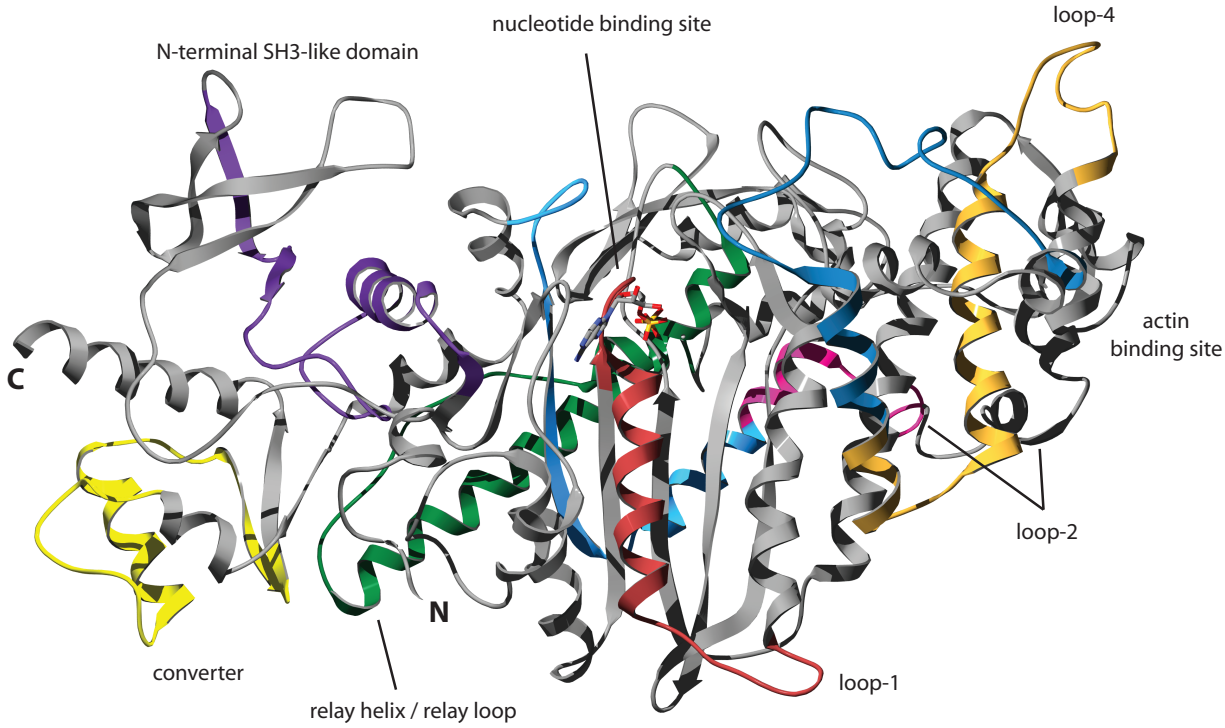

Figure 5.6: Structure of the myosin motor domain
The structure of the motor domain of the class-II myosin of Dictyostelium discoideum has been used to highlight the regions encoded by alternatively spliced exons in arthropod Mhc1 genes. The color-coding is the same as in Figure 5.1 allowing the identification of corresponding regions.

5.4 Discussion

25 muscle myosin heavy chain genes have been identified in 22 Arthropoda species. All sequences share strong homology to the alternatively spliced Mhc1 gene that has first been described in *Drosophila melanogaster* (82). The genes contain five to nine mutually exclusive exons and an penultimate exon that might either be included or excluded in the mRNA, and were assigned by manual inspection of the genomic DNA sequences (Figure 5.1). Because of the many alternatively spliced exons automatic identification of all exons failed. This is probably also the main reason for the wrong prediction of the exon organisation of the *Anopheles* Mhc1 gene (supplementary material of (90)).

Altogether, the transcription of the Mhc1 genes may result in several hundred differently spliced mRNAs (Table 5.1). The *Pediculus* Mhc1 gene has the least alternatives for its alternatively spliced exons resulting in a maximum of 32 different mRNAs, while the water flea gene might result in at least 3072 different mRNAs. Thus, except for *Pediculus*, *Nasonia*, and *Apis mellifera* all arthropod Mhc1 genes, for which all exons could be identified, outscore the 480 mRNA possibilities of *Drosophila melanogaster*. Although the number of variations seems vast compared to the number

of different muscle myosin heavy chain genes in other metazoa species, the regions for changes are limited to five to nine. In *Drosophila melanogaster*, all alternative exons are expressed depending on the developmental stage, but only a limited number of combinations seem to be employed (88). Whether all alternative exons are expressed in the other Arthropoda species and which combinations are used has to be determined.

The phylogenetic analysis of the Mhc1 protein sequences agrees with the expected phylogenetic relationship between the species. There are two notable exceptions in the *Drosophila* species section of the tree. The DseMhc1 sequence would have been expected to be the closest relative of the DssMhc1 sequence, and the DaMhc1 sequence would have been expected to separate after the split of the DpMhc1 and DrpMhc1 sequences. There are two possible ways to explain this observation. Either, the Mhc1 genes have evolved asynchronously as has been found for many yeast genes (67) or the genes might have incorporated back-mutations. The sequence identities of 96.1 to 99.7 % are very high, and thus only a few mutations would lead to a different phylogenetic classification.

The *Tribolium castaneum*, *Pediculus humanus corporis*, and *Bombyx mori* Mhc1 genes contain one additional and the *Daphnia pulex* Mhc1 gene contains four additional alternatively spliced exons compared to the *Drosophila melanogaster* gene (Figure 5.1, Figure 5.2). All additional alternatively spliced exons are mutually exclusive and encode parts of the motor domain. The additional exons of the *Tic*, *Pdc*, and *Bm* Mhc1 genes encode alternative versions of the loop-2 sequence while the additional exons of the *Dap* Mhc1 gene are spread over the entire motor domain. In each case, the 3 variant is more homologous to the corresponding sequences in the other Mhc1 genes than the 5 variant (Figure 5.2).

A similar conservation is found for alternative exons with multiple variants (Figure 5.2). In almost all cases, the most 3' variant is the most conserved one. For the alternative exon encoding part of the motor domain near the ATP-binding site (exon 7 in DmMhc1), the last of the variants is the only variant that is conserved in all species. The other variants are either missing in certain species, or are very similar to each other as well as to those of other species, so that it is not clear whether they have been derived from independent variant duplications or whether they have been present in a common ancestor. Thus, all variants except for the most 3' variant have been evolved after the separation of *Daphnia* from the other species. The variants encoding the relay-helix and the relay-loop are highly conserved. Therefore, conserved differences confine to only one or two residues. The second-last of the variants seems to be the most conserved, although mutation of one residue might change this. The exon encoding part of the converter domain has two highly conserved variants, the most 5' and the most 3' variants. The most 3' variant distinguishes from all other variants at the DNA level because the following intron starts with a GT donor site. The most 5' exon is the most important, though not the only, determinant for flight capabilities (109, 110).

Based on the exon-intron patterns of the 21 Mhc1 genes the gene structure of the ancient arthropod Mhc1 gene can be predicted. In the first half of the genes encoding the motor and the neck domain, all except one intron position are shared by at least two genes (Figure 5.5). The exons encoding the coiled-coil tail domain starting at amino acid 850 are considerably longer and the intron positions in

almost all genes are not identical. This is due to the fact that all genes have lost different introns. It is highly probable that further sequencing of arthropod Mhc1 genes will reveal different exon-intron patterns in the tail region while intron positions with one or more of the already analysed genes will be shared. The reconstructed ancient arthropod Mhc1 gene supports the idea that introns have been lost in most cases and not gained during evolution (102). It is very unlikely that the different species, distributed over a broad taxonomic range, invented introns at the same positions independently from each other. Thus, the ancient Mhc1 gene is expected to contain all intron positions that appear in at least one of the analysed Mhc1 genes. Analysis of Mhc1 genes of further species might add additional intron positions especially in the tail region. The exon lengths of the ancient Mhc1 gene are between 30 and 210 bp in the motor domain and up to 480 bp in the tail region. These short exons (compared to e.g. the *Drosophila* Mhc1 gene) resemble exon lengths in vertebrates and further comparative analysis with vertebrate muscle myosin heavy chain genes will reveal the gene structure of the ancient Metazoa gene.

In addition to the Mhc1 gene, *Aedes aegypti* encodes a further muscle myosin heavy chain gene, named Mhc3, that encodes only one variant of each of the alternatively spliced exons of the Mhc1 gene. The presence of this gene is not an artefact from sequencing or the assembly process. Although the translated exons show high identities, both genes are very different at the DNA level, and both are confirmed by several EST clones. That also means, that the Mhc3 gene, that does not encode any alternatively spliced exons, is expressed during the life cycle of *Aedes aegypti*. Note that the combination of alternatively spliced exons does not correspond to any of the tissue-specific combinations found in *Drosophila* (summarized in (88)). *Culex pipiens quinquefasciatus* contain another two muscle myosin heavy chain genes in addition to the Mhc1 gene, named Mhc3 and Mhc4, that, similarly to AeaMhc3, encode only one variant of most of the alternatively spliced exons of the Mhc1 gene. In one case, the intron between the presumed variant of the alternatively spliced exons and the following constitutive exon disappeared. Unfortunately, there is not enough EST data available for *Culex pipiens quinquefasciatus* to support any of the myosin heavy chain genes. AeaMhc3, CpqMhc3, and CpqMhc4 retained the same variants of the alternative exons of the corresponding Mhc1 genes. The presence of these further muscle myosin heavy chain genes is very surprising because the number of alternatively spliced exons in the Mhc1 genes already allows for the transcription of several hundred different muscle myosin isoforms. How could it happen that the genomes of *Aedes aegypti* and *Culex pipiens quinquefasciatus* encode such genes? According to the phylogenetic tree of the myosin heavy chain genes, the Mhc3 and Mhc4 genes obviously appeared in the common ancester of *Aedes* and *Culex* after the divergence from *Anopheles gambiae*. In addition, there is no evidence for a (partial) second muscle myosin heavy chain gene in the *Anopheles gambiae* genome. Also, the carboxy-terminal ends of AeaMhc3 and CpqMhc4, that are 3 elongations of the last constitutive exon, do not exist in the AeaMhc1 and CpqMhc1 genes but have identical counterparts in the AngMhc1 gene that is also supported by several EST clones. It is unlikely that these three organisms have developed such a carboxy-terminal end of the myosin gene independently from each other. Instead, it is more probable that the ancient AeaMhc1 and CpqMhc1 genes have lost this specific carboxy-

terminus after incorporation of the Mhc3 and Mhc4 genes into the genome. This would mean that this carboxy-terminus is only used in the specific combination of alternatively spliced exons as found in the AeaMhc3 and CpqMhc4 genes. Whether this is also true for the AngMhc1 gene has to be verified. Based on their identity in sequence and gene structure it is most probable that CpqMhc3 and CpqMhc4 have been derived by duplication of one of the other.

There are two possibilities how the Mhc3 and Mhc4 genes could have appeared in the common ancestor of *Aedes* and *Culex*. The genes have either been derived from a duplication of the Mhc1 gene as part of a single gene or chromosomal region duplication event. Or, a partially spliced transcript of Mhc1 has been reincorporated into the genome (Figure 5.7). If the Mhc3 and Mhc4 genes had been derived from duplication, then all variants except one of the alternative exons of only one of the (then) two Mhc genes had to be lost in addition to the loss of both terminal exons in Mhc3. Given the amount of possible transcripts of the Mhc1 gene and the possibility to duplicate alternative exons, it is very unlikely that there would be a need for a second gene with the same set of alternative exons. If it were advantageous to keep two almost identical genes, it would be very unlikely that only one of the genes has lost all except one of its alternative exons. In addition, there must have been a very strong evolutionary pressure to keep exactly this special combination of alternative exons. The second possibility would mean that in the first step during the splicing process all alternatively spliced exons, which are not needed, are removed leaving introns between the remaining alternatively spliced and constitutive exons (Figure 5.7). In the second step, all introns are spliced to yield the mRNA for translation. In the case of the Mhc3 and Mhc4 genes, the transcript containing one combination of alternative exons but all introns would have been integrated into the genome, probably after retrotranscription. How should these type of genes be called? At least the AeaMhc3 gene is completely transcribed, and also CpqMhc3 and CpqMhc4 do not contain any premature stop codons or frameshift mutations. However, compared to the corresponding Mhc1 genes they retained only one variant exon of each of the alternative exons. Thus, they do not belong to the non-processed pseudogenes. We would rather regard them as a new type of partially processed pseudogenes.

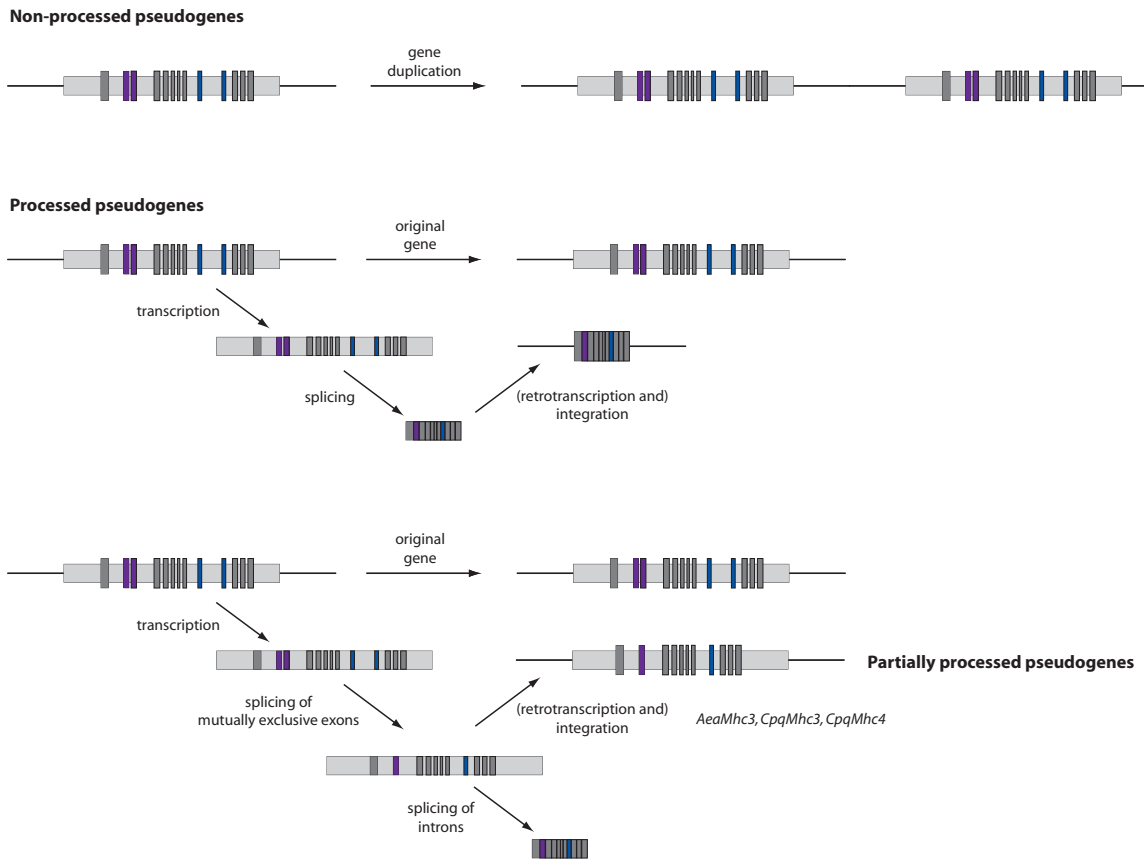

Figure 5.7: Model for the process of alternative splicing
The model describes the three different origins of pseudogenes. Non-processed pseudogenes are often found adjacent to their paralogous functional gene and retain the same exon-intron structure. Processed pseudogenes are marked by the absence of both 5' promotor sequence and introns, the presence of flanking direct repeats, and are randomly integrated into the genome. In the case of the arthropod Mhc genes, these get in the first step transcribed. In a second step, the alternative exons get spliced resulting in a certain combination of alternative exons and retaining the exon-intron structure. In the case of AeaMhc3, CpqMhc3, and CpqMhc4, these transcripts have been integrated into the genome. Normally in a third step, the introns get spliced revealing the final mRNA ready for translation. Dark grey bars represent constitutive and coloured bars alternatively spliced exons. Light grey bars represent non-coding sequence.

5.5 Conclusions

25 arthropod muscle myosin heavy chain genes have been identified and analysed. Compared to the well-studied gene of *Drosophila melanogaster* other arthropod genes might contain up to four additional alternatively spliced exons encoding part of the motor domain. This considerably extends the possibilities of other Arthropoda species to fine-tune myosin and thus muscle characteristics. An ancient arthropod muscle myosin heavy chain gene could be reconstructed whose gene structure can only be explained if introns are lost and not gained during evolution of this gene. *Aedes aegypti* and *Culex pipiens quinquefasciatus* even encode further muscle myosin heavy chain genes that, however,

Species	Species Abbr.	Nucleotide IDs GenBank:	Motor domain	Full-length protein
Daphnia pulex	*Dap*		1536	> 3072
Bombyx mori str. Dazao	*Bm*	AADK01001734, BAAB01137479, BAAB01017092, AV404226, AADK01040535, AADK01049792	192	768
Tribolium castaneum str. Georgia GA2	*Tic*	AAJJ01000118	192	> 384
Nasonia vitripennis str. SymAX	*Nav*	AAZX01008059, AAZX01007288	144	> 288
Apis mellifera str. DH4	*Am*	AADG05005753, AADG05005754, AADG05005757	96	384
Drosophila ananassae TSC#14024-0371.13	*Da*	AAPP01015693	120	480
Drosophila erecta TSC#14021-0224.01	*Der*	AAPQ01007075	120	480
Drosophila grimshawi TSC#15287-2541.00	*Dg*	AAPT01021775	120	480
Drosophila hydei	*Dh*	X77570	120	480
Drosophila melanogaster	*Dm*	NM_165190	120	480
Drosophila mojavensis TSC#15081-1352.22	*Dmo*	AAPU01010481	120	480
Drosophila persimilis MSH-3	*Drp*	AAIZ01000908, AAIZ01000907, AAIZ01000906, AAIZ01000905, AAIZ01000904, AAIZ01024863, AAIZ01000903	120	480
Drosophila pseudoobscura MV2-25	*Dp*	AAFS01000199	120	480
Drosophila sechellia Rob3c	*Dse*	AAKO01001629	120	480
Drosophila simulans str. white501	*Dss*		120	480
Drosophila virilis TSC#15010-1051.87	*Dv*	AANI01016210, AANI01016211	120	480
Drosophila yakuba Tai18E2	*Dy*	AAEU01002444, AAEU01002445, AAEU01002446	120	480
Drosophila willistoni TSC#14030-0811.24	*Dw*	AAQB01006734	120	480
Anopheles gambiae str. PEST	*Ang*	AAAB01008980	128	768
Aedes aegypti str. Liverpool Mhc1	*Aea*	AAGE02009209	128	512
Aedes aegypti str. Liverpool Mhc3	*Aea*	AAGE02009019, AAGE02009018	1	1
Pediculus humanus corporis str. USDA	*Pdc*	AAZO01001178	16	32
Culex pipiens quinquefasciatus JHB Mhc1	*Cpq*	AAWU01000999	128	512
Culex pipiens quinquefasciatus JHB Mhc3	*Cpq*	AAWU01000999	1	1
Culex pipiens quinquefasciatus JHB Mhc4	*Cpq*	AAWU01000999	1	1

Table 5.1: Nucleotide ID's and number of combinations of alternative exons for the motor domains and the full-length proteins.

have lost all except one variant of the alternatively spliced exons. These genes most probably entered the genome by reincorporating a certain processed transcript and not via a gene or genomic region duplication event. If the gene has been derived from a processed transcript then splicing of alternative exons must involve a first step, in which all other variants are spliced out leaving intronic sequence around the variant of choice. In a second step, all introns are spliced.

5.6 Methods

5.6.1 Identification and annotation of the arthropod muscle myosin heavy chains

The genes for *Aea*, *Ang*, *Am*, *Bm*, *Cpq*, *Dm*, *Drp*, *Dp*, *Dse*, *Dss*, *Dy*, *Dw*, *Pdc*, and *Tic* Mhc1 and Mhc3 have been obtained by TBLASTN searches against the insects section of the NCBI wgs database (Table 5.1) (111). The genes for the *Da*, *Der*, *Dg*, *Dmo*, and *Dv* Mhc1 have been obtained using the BLAT alignment tool (112) against the UCSC Genome Browser database (113, 114). The DhMhc1 sequence was derived from the NCBI nonredundant database. The DapMhc1 sequence has been obtained by a TBLASTN search against the 9x assembly of the *Daphnia pulex* genome provided by the DOE Joint Genome Institute (115) and the *Daphnia* Genomics Consortium (116). The NavMhc1 gene was derived from version 1.0 of the *Nasonia vitripennis* assembly provided by the Human Genome Sequencing Center at Baylor College of Medicine (117). The exons of the genes were predicted by manual inspection of the nucleotide sequences. For the correct prediction of the transcriptional start and the 3' terminal exons, the analysis of cDNA and EST data, that has been obtained from the EST section of NCBI's nucleotide database, was necessary. In particular, the following data has been obtained: For TicMhc1, only a small amount of EST data is available, confirming the prediction of exon2. There is not enough data to exclude a further untranslated 5' exon, as well as further C-terminal exons. For AngMhc1, several EST and cDNA clones support exon1 and the different C-termini. The C-termini of AeaMhc1 are also supported by several EST clones (e.g. GenBank ID DV384821). Exon1 of AeaMhc3 is supported by EST data. Exon1 of AeaMhc3 has been used for the identification of exon1 of AeaMhc1, as there is no direct evidence by EST data. Surprisingly, it is found 26,432 bp before the translation start codon ATG. For AmMhc1, the N-terminus is not supported by EST or cDNA data. Therefore it is not clear whether there might be an additional 5' untranslated exon. The C-termini are supported by several EST and cDNA clones (e.g. GenBank ID CK629939). The C-terminus of DapMhc1 is supported by EST data (e.g. GenBank ID BJ927473), while there is no EST data for the N-terminus. For BmMhc1, exon2 is supported by EST data. However, the corresponding EST clones are not long enough to exclude a further 5' untranslated exon. Both C-termini of BmMhc1 are supported by EST clones (e.g. GenBank ID BP179837). The genomic DNA of the BmMhc1 gene contains a gap in the coiled-coil tail region. The missing amino acid sequence has been derived from EST data. However, the exon/intron structure in the corresponding region remains unresolved.

5.6.2 Analysis of the relationship of the alternatively spliced exons

All alternatively spliced exons have been aligned manually. Some kind of relationship is already obvious from these sequence alignments. To get a more quantitative description, sequence identity matrices have been calculated for each set of aligned exons. Subsequently, sets of homologous exons from all Mhc1 genes have been clustered by sequence similarity. We have visualized the results in graphs that have to be read in columns. The highest identity between an exon listed on top and any variant of a certain Mhc1 protein listed on the left side has been set to 1 (red colour) while the differences between the values of the lower identity exons and the value of the highest identity have been plotted for the other combinations of exons. Thus, in every column the highest identity of the named exon to one of the variants of the other Mhc1 proteins is visualized.

5.6.3 Building trees

The phylogenetic tree was generated using neighbour joining and the Bootstrap (1,000 replicates) method as implemented in ClustalW (standard settings) (68) and drawn by using TreeView (118). The sequence of DapMhc1 has been used as outgroup.

5.6.4 Authors' contributions

F.O. performed data analysis. M.K. assembled all sequences, performed data analysis and wrote the manuscript. Both authors read and approved the manuscript.

5.6.5 Acknowledgements

M.K. was supported by a Liebig Stipendium of the Fonds der Chemischen Industrie, which was in part financed by the BMBF. This work has been funded by grant I80798 of the VolkswagenStiftung and grant KO 2251/3-1 of the Deutsche Forschungsgemeinschaft. We thank the DOE Joint Genome Institute (115) and the *Daphnia* Genomics Consortium (116) for providing access to the assembly of the *Daphnia pulex* genome, and the Human Genome Sequencing Center at Baylor College of Medicine for providing access to the assembly of the *Nasonia vitripennis* genome preliminary to publication.

Part IV

Manuscripts

6 Scipio: Using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species

Oliver Keller[1], Florian Odronitz[2],Mario Stanke[1], Martin Kollmar[2] and Stephan Waack[1*]

[1]Universität Göttingen, Institut für Informatik, Lotzestr. 16-18, 37083 Göttingen, Germany
[2]Department of NMR-based Structural Biology, Max-Planck-Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Goettingen, Germany
[*]Corresponding author.

6.1 Abstract

6.1.1 Background

For many types of analyses, data about gene structure and non-coding regions of genes are required. Although a vast amount of genomic sequence data is available, precise annotation of genes is lacking behind. Finding the corresponding gene of a given protein sequence by means of conventional tools is time consuming and error prone, and requires considerable experience.

6.1.2 Results

Scipio is a tool to determine the precise gene structure given a protein sequence and a genome. It identifies intron-exon borders and splice sites and is able to cope with sequencing errors and genes spanning several contigs in genomes that have not yet been assembled to supercontigs or chromosomes. Instead of producing a set of hits with varying confidence, Scipio gives the user a coherent summary of locations on the genome that code for the query protein. The output contains information about discrepancies resulting from sequencing errors. Scipio has also successfully been used to find homologous genes in closely related species. Scipio was tested with 979 protein queries against 16 arthropoda genomes (intra species search). For cross-species annotation, Scipio was used to annotate 40 genes from *Homo sapiens* in *Pongo pygmaeus abelii* and *Callithrix jacchus*.

6.1.3 Conclusions

Scipio is able to precisely map a protein query onto a genome. Even under less than ideal circumstances like sequencing errors and incomplete genome assemblies, it most often provides the user with correct determination of intron-exon borders and splice-sites. Apart from being able to find genes in the genome that encode the query protein, Scipio can also be used to annotate genes in closely related species.

6.2 Background

In the post-genome era, sequence data is the entry point for many studies. Often, it is essential to obtain the correct genomic DNA sequences of eukaryotic genes because of the information contained in non-coding regions. For example, the intron regions contain important sites for the regulation of gene transcription like enhancers, repressors, and silencers (119). Transcription initiator sequences are located upstream from the target gene (120). The determination of the exon/intron structures of genes is also important in comparative genomic analyses like the identification of ancient exons (121). Today, over 300 eukaryotic genome sequencing projects have resulted in genome assemblies (49). For most of the eukaryotic genomes genome sequences of genes are only available for ab-initio derived gene predictions, if at all. However, it has been shown that computer derived sequences are often wrong because of sequencing and assembly errors, and mispredictions (83). Correct protein sequences have in many cases been derived from manual annotation of the genes of interest or from full-length cDNAs. But experimentally obtained cDNA sequences often do not completely correspond to annotated genes because so far undescribed alternatively spliced forms have been isolated. In many cases, it might also be interesting to look at the genes of evolutionary closely related species. If these species have not been annotated yet, it is, however, very time-consuming to identify and manually annotate the corresponding homologous genes.

Currently, two programs are available for the retrieval of non-coding sequence. The Java application Retrieval of Regulative Regions (RRE) parses annotation and homology data from NCBI (122). RRE requires local installation and a local copy of the desired genomes and annotation files. The web application of RRE only hosts a small number of eukaryotic genomes and only annotation data from NCBI. Recently, the non-coding sequences retrieval system (NCSRS) has been published (123) that has 16 genomes and annotation data from both NCBI and Ensembl. In summary, both tools rely on annotation files provided by NCBI and Ensembl, with all possible errors, for only a few organisms. Furthermore, programs have been published that use a cDNA query to perform a spliced alignment for the determination of exact splice site locations; examples are Splign (124) and SPA (125).

We have developed Scipio for the retrieval of the genome sequence corresponding to a protein query. The tool does not require any annotation data, and is able to correctly identify the gene even if it spans several genome contigs and contains mismatches and frameshifts. Because of these capabilities, Scipio is not only able to correctly identify the gene in the genome corresponding to the protein query but also to correctly identify the homologous genes in the genomes of closely related organisms.

6.3 Implementation

The task of determining the gene structure of a query protein within a DNA target sequence can in most instances be accomplished with the search for a spliced alignment. Since a large number of tools performing this task has already been available for a long time (the tblastn variant of BLAST (126), PROCRUSTES (127), and BLAT (112)), writing another one would mean reinventing the wheel. However, in the example of BLAT, when performing a search for the protein in the translated DNA, the output does not coincide with the exon structure of a single gene. Usually, multiple hits are found for each query, varying in accuracy, and exon boundaries are given only on amino acid level, missing those codons that are split by introns. Hence, manual processing was still needed in the majority of cases to determine the exact location of the query. In cases where the genomic sequence is in an early stage of the assembly process, several parts of one particular gene are often found on different target sequences (contigs), making this task very tedious and time consuming.

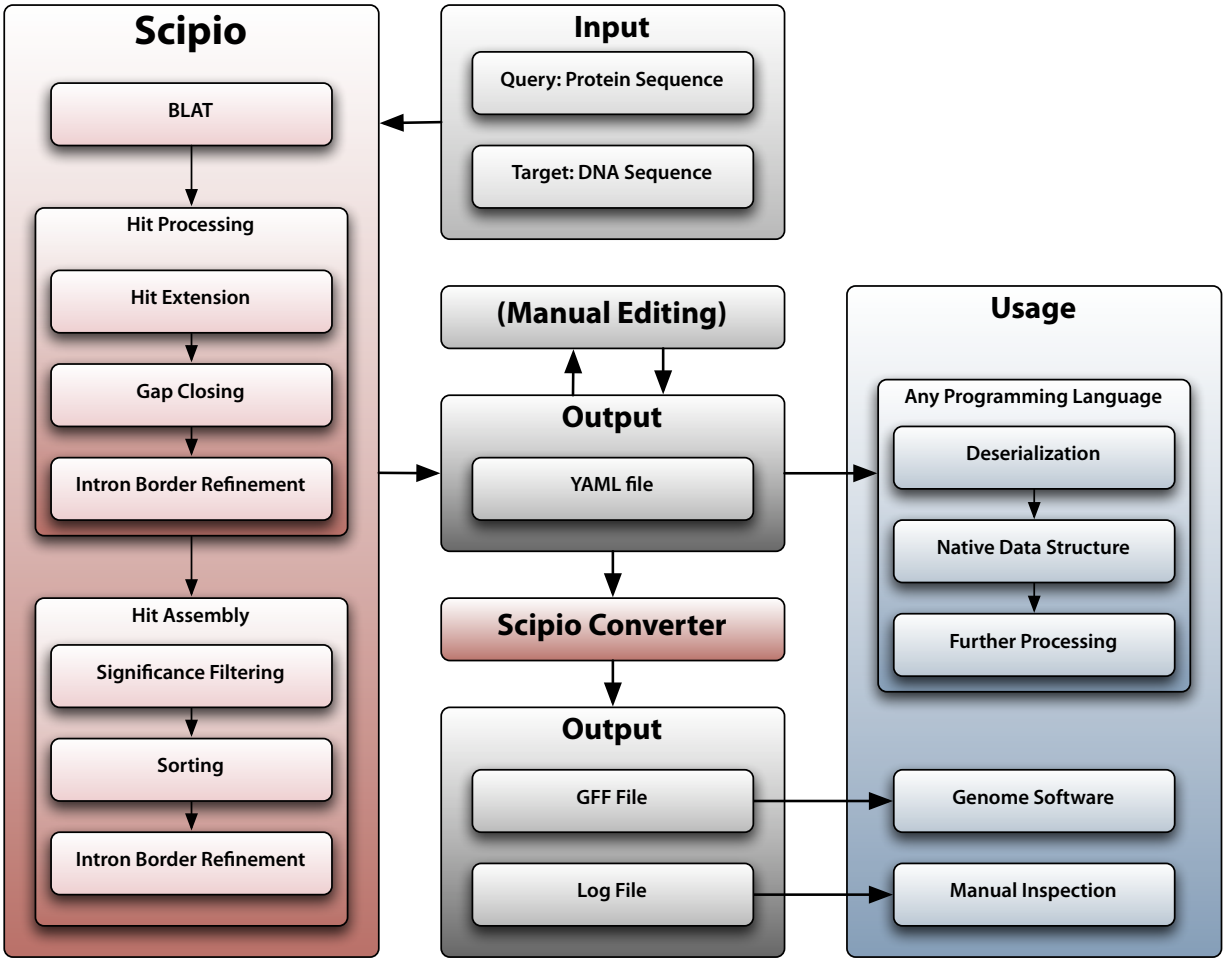

Figure 6.1: The Scipio Workflow
This diagram depicts the data flow of a Scipio run. Scipio needs two FASTA files as input, one containing the protein query and one with the genome sequence. Scipio starts BLAT and processes the results in a series of steps, successively refining and assembling the hits. Scipio's output is a YAML file which can further be converted into a GFF file or a log file. YAML files can also be manually edited and read by a parser of which many exist for all modern programming languages. The resulting data structure can then be further processed.

6.3.1 The Scipio script

We designed the perl script Scipio to automate this process and output the results in both human- and machine-readable output formats. The summary of the process is depicted in the diagram in Fig. 1. We chose to run BLAT to provide us with the spliced alignments because it is specialized for the case of high sequence identity, which is obviously the case when locating genes of the same species (where mismatches are mainly due to sequencing errors), but it turned out to be very applicable also for the case of closely related species.

Stage one: hit refinement

After running BLAT, Scipio processes the query protein and target DNA sequences, and the file containing the BLAT hits. In the first stage, each hit is then "refined" by a number of steps. A BLAT hit is a collection of consecutive matchings of the protein sequence aligned to the translated DNA. We do not want to include hits with low quality, so everything with an accuracy below a given threshold is discarded at this stage. The refinement consists of the following steps:

- Unaligned parts of the target sequence between the matchings that form a BLAT hit are analysed. In a significant hit, they consist of at most one residue of the query corresponding to a longer piece of DNA, so they will be considered introns. Scipio tries to determine the exact location of the splice site by looking for a splice site pattern (GT–AG, GC–AG, and other candidates). This way, codons that are split by an intron, and are only joined after splicing, can be revealed. In cases where all residues are aligned by BLAT but a splice site pattern is missing, Scipio tries to improve the prediction by shifting the splice sites in single nucleotide steps. If an exact location can not be found, a heuristic is used to determine a trade-off between the number of additional mismatches and the presence of the splice site pattern.

- In addition, two more types of unaligned target sequence are distinguished: First, actual gaps with significant parts of the query sequence unaligned (mostly due to low coverage of sequencing resulting in gaps between contigs represented by contiguous N's in supercontigs/chromosomes). Second, short gaps resulting from sequencing and assembly errors leading to additional or missing bases or codons, with or without a frameshift. Additional DNA in this case is not interpreted as intron an but as a sequence shift of the query against the target.

- Scipio tries to locate very short exons where the BLAT hit misses parts of the query sequence. This is done by simple pattern matching. Thus only pieces with full identity are added. Terminal exons are added only when an intact splice site is found.

The filtering during the first stage ensures that nothing will be shown that cannot be regarded as a good match. If no hit is left after filtering, Scipio simply considers the gene non-existent in the target sequence, and no further processing is done.

Stage two: hit filtering and assembly

All BLAT hits that survive the first stage are subsequently filtered in the second stage to determine those that form the gene corresponding to the protein query. If only complete chromosomes were considered, one could expect a single optimal BLAT hit coinciding with that gene; however, in cases without a complete assembly, partial hits on different targets need to be taken into account.
First, all hits are sorted by a score proportional to the number of matches, with a penalty subtracted for each mismatch. Second, all incompatible hits are discarded in the order just determined. Hits are incompatible if their queries overlap but their targets do not. (An exception is the complete identity on DNA level at the ends of two contigs. This could result from an incomplete assembly, and the

possibility of an overlap is taken into account.) At the end of this step, we come out with a small number (usually just one) of non-overlapping hits forming the best gene candidate.
The final part of stage two is another refinement step: by assembling multiple hits, sequence parts may have been identified as parts of an intron that is split on different targets, the first half at the end of one target, the second at the beginning of the next. After the assembly Scipio uses the same method as in stage one to determine the exact splice site locations.

6.3.2 Output

The output contains target names, and location coordinates (genomic and protein) of all features: introns, exons, and gaps; exons can have sub-features: sequence shifts, mismatches, or undetermined positions. In addition, it contains the genomic DNA for all regions (including up- and downstream of the hit) and the translation of the coding sequence.

For the output format we defined two essential requirements: Human readability and machine readability. We chose YAML as it is a format that is complex enough to express our data structures and at the same time simple enough to be human readable and editable. YAML can easily be parsed and there are numerous bindings for any modern programming language. The resulting native data structures can be used to further process the data generated by Scipio.

Conversion tools

Scipio provides two tools to convert YAML files:

- yaml2log: Converts YAML files into an easily readable log file with summary information about the results and clearly arranged sequence alignments.
- yaml2gff: Converts YAML files into GFF Format which can be read by a wide range of genome-related software packages.

```
1: mismatch
gDNA        AAA ttt GGG
translation  K   F   G
             |   X   |
query        K   A   G

2: undetermined query
gDNA        AAA ttt GGG
translation  K   F   G
             |       |
query        K   X   G

3: undetermined target
gDNA        AAA nnn GGG
translation  K   X   G
             |       |
query        K   A   G

4: additional codon in target
gDNA        AAA ttt GGG
translation  K   F   G
             |       |
query        K   -   G

5: unmatched query
gDNA        AAA --- GGG
translation  K   -   G
             |       |
query        K   A   G

6: frameshift (+1) target only
gDNA        AAA t-- GGG
translation  K   X   G
             |       |
query        K   -   G

7: frameshift (+1) target/query
gDNA        AAA t-- GGG
translation  K   X   G
                     |
query        X   -   G

8: frameshift (+2) target only
gDNA        AAA tt- GGG
translation  K   X   G
             |       |
query        K   -   G

9: frameshift (+2) target/query
gDNA        AAA tt- GGG
translation  K   X   G
                     |
query        X   -   G

10: frameshift (-2) target only
gDNA        AAA t-- GGG
translation  K   X   G
             |       |
query        K   A   G

11: frameshift (-2) target/query
gDNA        AAA t-- GGG
translation  K   X   G
             |       |
query        K   X   G

12: frameshift (-1) target only
gDNA        AAA tt- GGG
translation  K   X   G
             |       |
query        K   A   G

13: frameshift (-1) target/query
gDNA        AAA tt- GGG
translation  K   X   G
             |       |
query        K   X   G

14: stopcodon target/query
gDNA        AAA tag GGG
translation  K   *   G
             |   |   |
query        K   *   G

15: stopcodon, target only
gDNA        AAA tag GGG
translation  K   *   G
             |   X   |
query        K   D   G

16: stopcodon, undetermined query
gDNA        AAA tag GGG
translation  K   *   G
             |       |
query        K   X   G

17: additional stopcodon
gDNA        AAA tag GGG
translation  K   *   G
             |       |
query        K   -   G
```

Figure 6.2: Types of discrepancies
This chart lists all types of discrepancies between protein query and target translation/DNA that are known to Scipio. The identifiers as written into the log files are given.

6.4 Results and discussion

In many biological studies, protein sequences have been obtained by isolating mRNAs and translating them into the corresponding cDNAs. Also, large-scale cDNA sequencing projects resulted in thousands of supposed-to-be full-length cDNA sequences for some eukaryotes (128), (129). Protein sequences might have also been obtained by manual annotation. Sets of genomic DNA sequences of genes exist for some annotation projects. However, for many eukaryotic sequencing projects, the annotation process is lacking years behind the sequencing and assembly. In addition, experimentally obtained cDNA sequences often differ from annotated sequences because new alternatively spliced forms have been isolated. Therefore, for subsequent studies it might be useful or crucial to obtain the genomic DNA and the gene structure corresponding to the protein of interest.
Scipio has been designed for this task, and based on its differentiated processing capabilities it is able to cope with genes spanning multiple contigs as well as various kinds of sequencing and assembly errors. Scipio has been developed for the correct identification of eukaryotic genes. It can also be used for bacterial and archaeal genes although these genes are easily identified manually based on their simple single-exon structure. Depending on the similarity of the protein sequences, Scipio is also often able to correctly identify homologous genes in closely related organisms. We have implemented the following features:

A. If the query is distributed on several targets, the target contigs will be assembled according to the query. Untranslated regions from the last exon on a contig to the contig end and from the beginning of the next contig to the next exon are regarded as intronic. Scipio is also able to resolve overlapping contig ends if they consist of coding sequence, hence contributing to an improvement of the assembly.

B. The yaml2log script identifies cases from a list of alignment discrepancies and mismatches between query and target sequence that can result from sequencing/assembly errors (Figure 2). The simplest case is that amino acids differ (cases 1 to 3), or that they are missing in either the target or the query (cases 4 and 5). Sequencing/assembly errors may lead to additional or missing bases. These frameshifts are represented by an X in the translation corresponding to one or two nucleotides. The query sequence might have either been obtained from cDNA sources thus leading to a mismatch between query and translated target (cases 6, 8, 10, and 12), or the sequencing errors might have already been interpreted represented by an X in the query (cases 7, 9, 11, and 13). The target sequence might also contain in-frame stop codons (cases 14 to 17). These can be the result of sequencing errors or real stop codons as they appear in pseudogenes. In all these cases, the stop codon is shown as an asterisk ('*') in the translation.

C. Scipio interprets splice site patterns to determine intron locations. Exons borders are chosen so that the splice sites belong to one of the following classes, in decreasing priority: GT–AG, GC–AG, AT–AC, GA–AG, GG–AG. In cases where the translation of the adjacent intronic sequence was identical to the query, it was necessary to shift the intron location predicted by

BLAT by several codons to determine the splice site location.

D. Scipio searches for stop codons at the end of genes. This helps evaluating the completeness of the query sequence.

E. Scipio tries to locate very short exons that are not recognized by BLAT. These short exons might either appear in-between longer exons or at the ends of the gene. For example, very often genes start with an N-terminal methionine that is the only translated codon in the first exon. Scipio locates N-terminal methionines only if matching splice sites are found.

6.4.1 Insect genomes

To develop and test Scipio we used a test set of 16 arthropod species encoding 979 proteins (Figure 3). The genome sequences (the newest collections of contigs as submitted to NCBI) differ in quality and completeness and are thus representative for straight-forward and difficult identifications of the genes. *Drosophila melanogaster* is an example of a perfect genome sequence with all reads assembled to chromosomes and almost all gaps closed. *Bombyx mori p50T* was used as example for a very preliminary assembly with many short contigs. The other genome sequences represent all stages in-between these extreme cases. For example, the genomes of *Drosophila persimilis* and *Drosophila sechellia* are quite complete which is visible from their number of contigs, but they have a low sequence coverage and/or contain many sequencing errors leading to high numbers of mismatches and frameshifts in the identified genes (Figure 3). In total, almost all query sequences have been identified correctly by Scipio (90.9 %), although many are spread on several target contigs (e.g. see *Aedes aegypti* and *Culex pipiens*). 4.7 % of the genes have correctly been identified but the target DNA sequence contains sequencing or assembly errors. Another 1.7 % has not completely been found with the standard BLAT settings (BLAT tilesize of 5) because these genes contain very short exons. After changing the BLAT tilesize to 3 or 4 these genes have also completely been identified. Further 1.7 % of the genes could not be identified correctly, because the query sequence has been derived from manual annotation thus having incorporated EST data, data from other genome assemblies (e.g. newer data from the sequencing centers), or errors in the manual annotation process. E.g., the *Bombyx mori p50T* genome data is very incomplete but a lot of EST data is available. Thus, the query protein sequences have been built to a large part on these EST data. EST data has also been used to close gaps in the *Apis mellifera* and *Drosophila virilis* genomes. But these errors are not due to problems in the implementation of Scipio. 2 sequences (0.2 %) could not be identified correctly, because the genome sequences are that bad, that several frame shifts happened next to each other. The query protein sequences have correctly been identified based on EST data, but the corresponding genome regions contain successions of sequencing errors, so Scipio could not resolve the mismatches and frameshifts. The remaining 7 sequences (0.7 %) contain very long overlapping regions due to problems in the genome assemblies. Currently, Scipio handles overlapping hits by choosing the one with the higher overall score, in some cases discarding the one with fewer missmatches in the overlap region. The other cases that Scipio did not resolve are those, where a frame shift exists very close to

an intron border. BLAT does not include the stretches past the frame shifts since they are smaller than the tile size used for searching. Scipio was not able to place the missing residues between the exons.

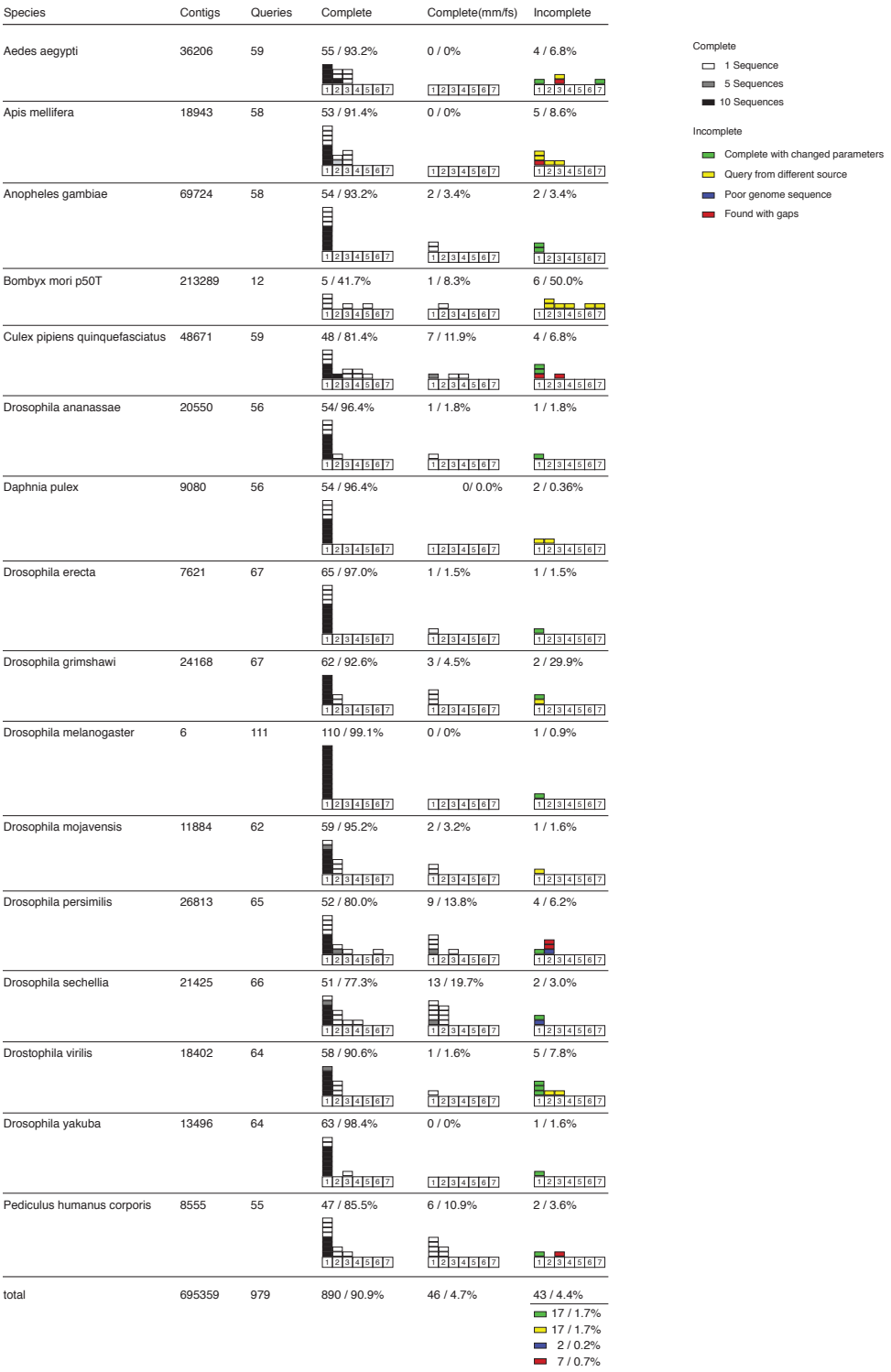

Figure 6.3: Performance
This chart shows Scipio's performance. The charts shown are histograms depicting how many sequences where found on a particular number of contigs in the genome. Black rectangles represent ten, grey ones five and white ones single sequences. Complete means the queries where found without discrepancies. Complete(mm/fs) means Scipio found the complete gene without gaps but with discrepancies like mismatches or framshifts. Incomplete means Scipio could not determine the complete gene structure with standard parameters. All searches where carried out with a BLAT-tilesize of 5.

6.4.2 Cross species search

To test the ability of Scipio to correctly predict orthologous genes in closely related organisms we have annotated the myosins in the recently assembled primates *Pongo pygmaeus abelii* and *Callithrix jacchus* (130). As a query, we used the 40 manually annotated myosins from *Homo sapiens* (83), which can be found in CyMoBase (`www.cymobase.org`, (37)). Although the genome assembly is not complete and most of the sequencing/assembly errors as described above have been seen, Scipio correctly predicted and identified all orthologs of the human myosins in the two primates. Scipio located all parts of the genes if they were distributed on several target contigs, and it correctly identified a rare splice site (GG–AG) that is specific for vertebrate sequences in a certain myosin class. Only in the tails of the class-15 and class-35 myosins very small and divergent gaps had to be filled manually.

6.4.3 Future developments

Eukaryotic genes contain far more information than is encoded in the sequence of one expressed protein. Most of this information is contained in the untranslated regions. Therefore, our future developments will focus on analyzing the untranslated regions to provide the user with additional gene-related information. Thus, Scipio will be developed to identify mutually exclusive exons, to determine other alternatively spliced exons, and to identify untranslated exons.
We will also implement a web interface for Scipio to address a wider audience and to make Scipio more user-friendly.

6.5 Conclusions

Scipio is a tool for the determination of gene structure and annotation of genes for a given protein sequence. Based on the widely used program BLAT, it performs exhaustive processing to ensure the best possible mapping of the protein onto the genome. Thereby Scipio goes beyond the scope of present spliced alignment tools and presents the user with a coherent set of matches that are often accurate to the level of single bases. Having a certain level of tolerance, Scipio can handle mismatches and frameshifts that often result from sequencing errors in genomes and cDNA. The same tolerance can be used to track down homologous genes in closely related species, allowing for cross-species annotation.

6.6 Availability and requirements

Project name: Scipio

Project home page: `http://www.webscipio.org`

Operating system: Platform independent
Programming language: Perl

Software requirements: Installation of BLAT and BioPerl.
Hardware requirements: BLAT may demand several times the genome size in RAM.

License: Scipio may be obtained upon request and used under a Creative Commons License.
Any restrictions to use by non-academics: Using Scipio by non-academics requires permission.

6.7 Authors contributions

OK, FO and MK set the requirements for the system, performed testing, and wrote the manuscript. OK wrote the source code for the software. FO and MK did the analysis of the insect genomes and the cross-species searches. MS and SW supervised the implementation of the software. All authors read and approved the final version of the manuscript.

6.8 Acknowledgements

MK has been funded by grant KO 2251/3-1 of the Deutsche Forschungsgemeinschaft. The sequence data for *Pongo pygmaeus abelii* and *Callithrix jacchus* were produced by the Genome Sequencing Center at Washington University School of Medicine in St. Louis and can be obtained from `ftp://genome.wustl.edu/pub/organism/Primates`.

7 WebScipio: An online tool for the determination of gene structures using protein sequences.

Florian Odronitz[1], Holger Pillmann[1], Oliver Keller[2],Stephan Waack[2], Martin Kollmar [1*]

[1]Department of NMR-based Structural Biology, Max-Planck-Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Goettingen, Germany
[2]Universität Göttingen, Institut für Informatik, Lotzestr. 16-18, 37083 Göttingen, Germany
[*]Corresponding author.

7.1 Abstract

7.1.1 Background

Obtaining the gene structure for a given protein is an important step in many analyses. A software suited for this task should be readily accessible, easy to handle and should provide the user with a coherent representation of the most probable gene structure. It should be rigorous enough to optimise features on the level of single bases and at the same time flexible enough to allow for cross-species searches.

7.1.2 Results

WebScipio, a software program based on Scipio (131), allows to obtains the corresponding gene structure of any query protein sequence that belongs to one of the already assembled eukaryotic genomes. The resulting gene structure is presented in various human readable formats like a schematic representation, and a detailed alignment of the query and the target sequence highlighting any discrepancies. WebScipio can also be used to identify and characterise the gene structures of homologs in related organisms. In addition, it offers a web service for integration with other programs.

7.1.3 Conclusions

WebScipio is a tool that allows users to get a high-quality gene structure prediction from a protein query. It offers a more than 250 eukaryotic genomes that can be searched and produces predictions that are close to what can be achieved by manual annotation, for in-species and cross-species searches alike.

7.2 Background

Cross-species DNA sequence comparisons are increasingly being used to identify both coding regions and functional DNA elements that often lie in the untranslated regions of the genes (132,133). These functional elements might be promotor sequences, transcription factor binding sites, termination signals or other regulatory elements. Comparisons of sequences of multiple species have either been performed at a genomic level (e.g. (134) (135)) or at the single gene and gene family scale (e.g. (136)). One important aim of most of the large-scale comparative studies has been to improve the annotation of the genomes, like the identification of new genes (137) or new constitutive and alternative exons (138). These studies have also resulted in the prediction of regulatory regions (139). However, only a limited number of the conserved noncoding sequences that have been identified by these studies have been characterised functionally.

Cross-species DNA sequence alignments of entire genomes are available for several eukaryotes (140, 141). These genomes, however, cover only a small part of the about 330 eukaryotic genomes for which genome assemblies are available (`wwww.diark.org`, as of Dec. 2007 (49)). Thus, a comparison of the genomic DNA sequences of a specific gene or gene family of a certain set of species would require a lot of time consuming manual steps. These involve obtaining the desired eukaryotic genome assemblies, identifying all homologous genes, and predicting their gene structures.

Here, we have developed WebScipio, a web interface to the Scipio software ((131), submitted). WebScipio provides access to a continuously updated list of almost all eukaryotic genome assemblies that are available worldwide (for a comprehensive list see www.diark.org). Additionally , the user can obtain all relevant data in human readable format in a very convenient way. For the integration with other programs, WebScipio provides a webservice.

7.3 Implementation

The web application is implemented in the Ruby programming language. As a framework we used Ruby on Rails (25). Ruby on Rails generates the pages for the website and handles the requests for the web service. After entering a protein query and selecting a genome, WebScipio starts Scipio on the server and searches the genome file for matches. Scipio produces a YAML file, which is parsed by WebScipio. WebScipio then presents the hits on the web interface in clear schemes and tables.

For visualisation, we create SVG drawings, which can also be downloaded to produce high-quality figures (142). WebScipio determines the optimal ratio for the scaling of exons and introns so that large intron sequences do not render the visualisation useless. It can also convert SVG to PNG if the browser does not support SVG viewing.
The species search function is enriched by additional metadata about the species. This information is retrieved from an index file which is synchronized with our sequencing projects database diArk (49). The genomes that are provided for searching have been collected from various sequencing centres and consortiums.

7.4 Results and discussion

7.4.1 Web interface

WebScipio offers a clean and simple web interface that can easily be used by inexperienced users. At the same time expert users have enough options to adjust the underlying algorithm to get the best possible results, even in difficult cases.

Species selection

Species are selected using an auto completion field. The user starts typing and a selection of species matching the search term is shown. Apart from searching for the scientific name of a species, many different types of information can be searched for: Alternative scientific names, common names, anamorph names (for fungi), and taxonomy (Figure 1). Users can also search for abbreviations of sequencing centers (e.g. 'JGI' for Joint Genome Institute) or type of genome files (e.g. 'chromosome').

Genomes

WebScipio offers 820 genome files from 249 eukaryotes (as of Dec. 2007) for searching, which amounts to more than 360 GB of sequence data. Genome data is kept up to date, but at the same time older versions are offered since newer assemblies sometimes lack some parts of the sequence. Many different types of genome data can be searched: Chromosomes, supercontigs, contigs, unplaced reads/contigs as well as genomes from mitochondria, chloroplasts and apicoplasts.

Protein query

The query for the search is one or several protein sequences, plain or in in FASTA format.

Search options

The search options define how tolerant the algorithm is regarding contigs and exons (Figure 2). 'Best Size' defines the minimum fraction of the query that has to be found on one single contig. 'Min

Identity' defines the minimal identity within a stretch of DNA in order to be taken into account by WebScipio. If, for example, the genome sequence is in an early stage of assembly and highly fragmented, the largest part that is found on one contig might only be 20% of the query. 'Max Mismatch' defines the maximum number of mismatches on a contig in order to be included in the results. The Value "0" allows an unlimited number of mismatches. The values for these parameters largely depend on the quality of the genome. 'Region Size' defines the length of the up- and downstream regions that can be retrieved. 'BLAT Tilesize' determines the width of the search window used to scan the genome. Decreasing this value makes it more likely that small exons are found but also slows the search process.

Visualization

A characteristic of conventional spliced alignment tools is that they produce lists of hits, maybe alongside with basic graphics, but most of the time the user does not see at a glance what the gene structure might be. WebScipio generates a graphical representation of the gene that clearly indicates the length and position of exons and introns and shows where discrepancies are located. It also shows the identifiers of the target sequences (Figure 3). In order not make small exons vanish when very large intronic stretches are found, the scaling of introns end exons in automatically balanced to make the picture visually meaningful. Tooltips show additional information.

Alignments, DNA and target translation

For detailed inspection of the hits, WebScipio generates an easy to read alignment of the query and the genome. It is grouped by exons, and mismatches and frame shifts are highlighted. Different stretches of DNA can be viewed: Up- and downstream DNA, genomic DNA from the first to the last exon including introns, or the coding DNA. The translation of the coding DNA as determined by the algorithm can also be viewed.

File download

Five types of files can be downloaded: A FASTA file containing all types of DNA sequences as described above, a FASTA file containing the protein translation, a log file with alignments, a GFF file for use with genome software, and a YAML file which contains all information generated by WebScipio.

File upload

WebScipio can also be used as a viewer for Scipio result files. When a YAML file of a previous search is uploaded, all result views are available. This way, users can store the results of their searches locally and can look at them any time, instead of repeating extensive searches. WebScipio can thus also be used as a viewer for results obtained from Scipio, the commandline version of the program.

7.4.2 Web service

All functions of WebScipio can also be used remotely as a web service. This allows for seamless integration with existing programs. Many modern programming languages offer built-in support for the required protocols. This frees programmers from the need to locally install software and to download and store large genome files. By using this service, it is easy to augment existing data with information produced by WebScipio. In-house, we use WebScipio's web service to determine the gene structure of thousands of motor proteins stored in CyMoBase (37). Storing the YAML data produced by WebScipio in a database and parsing it on demand is a powerful way of using this information. Ruby classes for conveniently handling the data structures can be obtained upon request.

7.4.3 Cross-species analysis of myosins in Human, Pongo, Callithrix, and Mouse myosins.

To test the capability of WebScipio when searching in species other than the origin of the query, we performed searches in four species, Human, as a reference, and *Pongo pygmaeus*, *Callithrix jacchus*, and Mouse (ordered by increasing phylogenetic distance to Human). As queries we used a set of 40 manually annotated myosin protein sequences as described in (83). For each species two searches were performed, one with the myosins from the species itself and one with the myosins from Human, giving a total of 280 searches.

We are confident that the manually annotated sequences we used as queries contain the least possible number of errors, since we compared them to EST data and dozens of homologue sequences from other species. Thus, most discrepancies with their source genomes are due to sequencing errors and low coverage. For each search we provide two percentages: The first and most significant number is the percentage of protein stretches that could be mapped onto the genome, allowing for mismatches that naturally occur when doing cross-species searches. The second number is the percentage of individual amino acids that could be aligned with codons on the genome, counting all discrepancies. As expected the agreement is very high when searching with queries from the target genome itself. But also when queries from humans are used to search genomes from other species, WebScipio is able to map most of the genes correctly. For *Pongo* and *Callithrix*, on average, more than 94% percent of the Human query sequence were successfully found in the genomes. Even in Mouse, which is much more diverged, the difference between searching with a native query and searching with a query from Human is below 10%, meaning than in most cases, the structure of genes can be predicted with only minor gaps and inaccuracies.

Figures 7.4, 7.5, 7.6 shows typical examples of in-species searches and cross-species searches for Myosin Class I proteins. The searches against the source genome are all almost perfect matches. Only in the *Pongo* and mouse genomes, three genes could only be mapped with gaps (PpMyo1B, MmMyo1A, MmMyo1F). Cross-species searches are, apart from the expected mismatches, almost as complete as the in-species searches.

For *Pongo*, three cross-species searches resulted in a reduction of the matching rate of less than five

Species	**vs self**	**vs Human**	**difference**
Human	99.815 % / 99.808 %	n.a.	n.a.
Pongo	97.975 % / 97.945 %	94.660 % / 94.125 %	3.315 % / 3.82 %
Callithrix	98.780 % / 98.685 %	96.558 % / 95.530 %	3.250 % / 3.155 %
Mouse	96.862 % / 96.692 %	87.825 % / 85.850 %	9.037 % / 10.842 %

Table 7.1: Average matching percentages for 40 myosin protein sequences from Human. Percentages are (percentage of protein, not subtracting mismatches) / (percentage of amino acids found, subtracting mismatches).

percent (MyoA, MyoB, MyoE), three stayed the same (MyoC, MyoD, MyoH), one got considerable worse, which can be attributed to the poor genome sequence in this region which contains stretches of Ns. HsMyo1G was found with better agreement since in this case WebScipio found a perfect 27bp match on another contig, which was not present in the search results for PpMyo1G.
In *Callithrix*, six out of the eight Human sequences where found with the same percentage as the *Callithrix* sequences (MyoA, MyoB, MyoD, MyoE, MyoF, MyoH) and two with minor losses (MyoC, MyoG).
In the Mouse genome, three sequences where found with the same (MyoB, MyoC, MyoD) or very similar (MyoE, MyoG) agreement. For Myo1H, the percentage decreased considerable. Myo1F was not found, instead, it was matched with the gene of Myo1E, a close homologue. The reason for this probably is the high degree of fragmentation or the occurrence of large gaps in the region of the Myo1F gene. The observation that Human Myo1A can be slightly better mapped then the ones from Mouse can be attributed to noise, since both hits have a low percentage of agreement (less than 40 %).

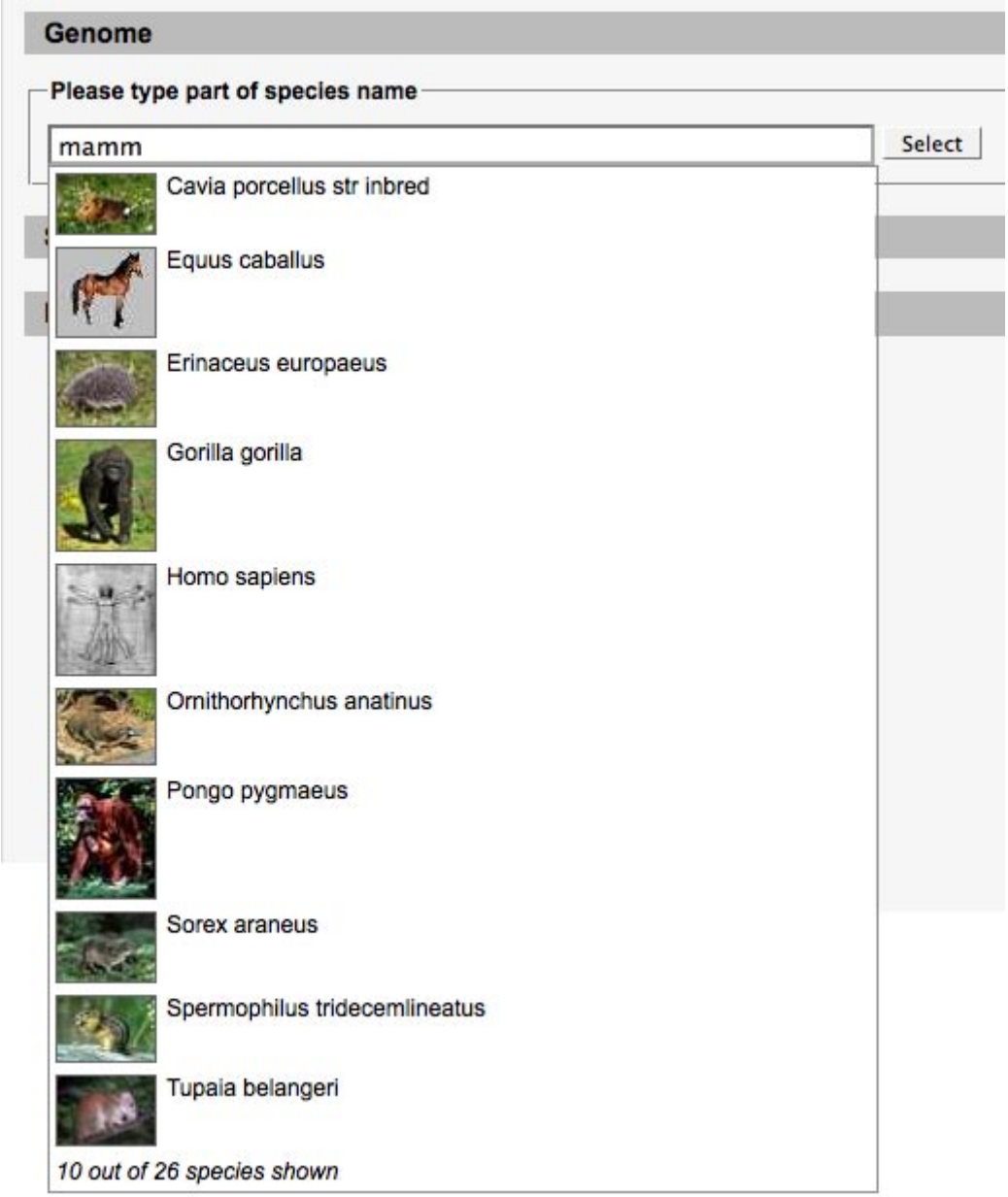

Figure 7.1: Species selection
The screenshot shows the species selection auto completion field. As the user types, species matching his query appear. Different types of information are taken into account when searching. In the example the user types 'mamm' and all Mammalia are listed.

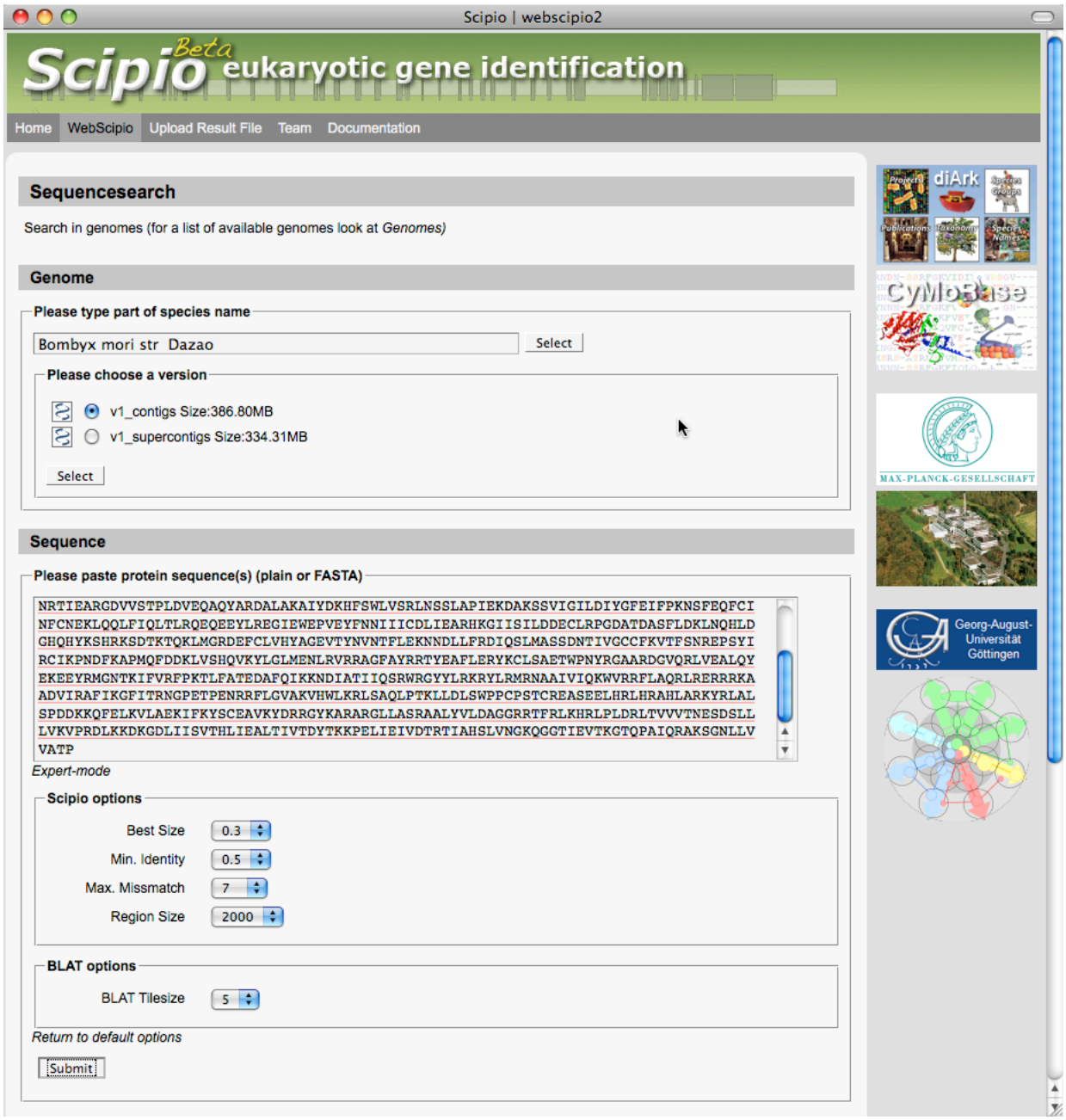

Figure 7.2: Input interface
The screenshot shows the input interface of WebScipio. First the user choses a species, then a genome, enters the query sequence and then specifies optional search parameters.

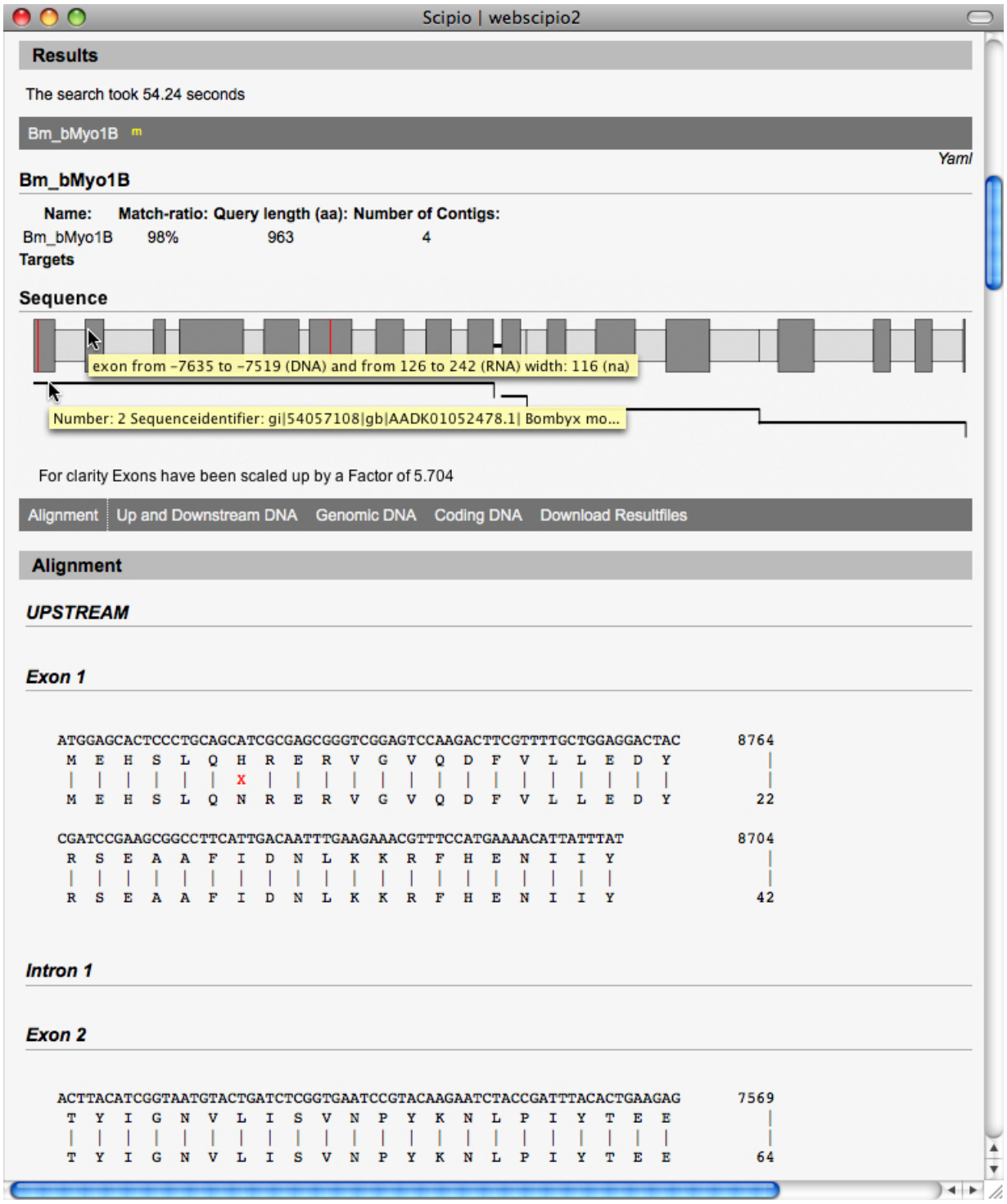

Figure 7.3: Result view
The screenshots shows the result view for a query. Basic statistics are provided along with an visualisation of the gene structure showing introns, exons, mismatches and frameshifts. It also shows which part of the gene was found on which contig. Tooltips provide further detail. Below, the alignment view is shown, clearly highlighting sites of disagreement.

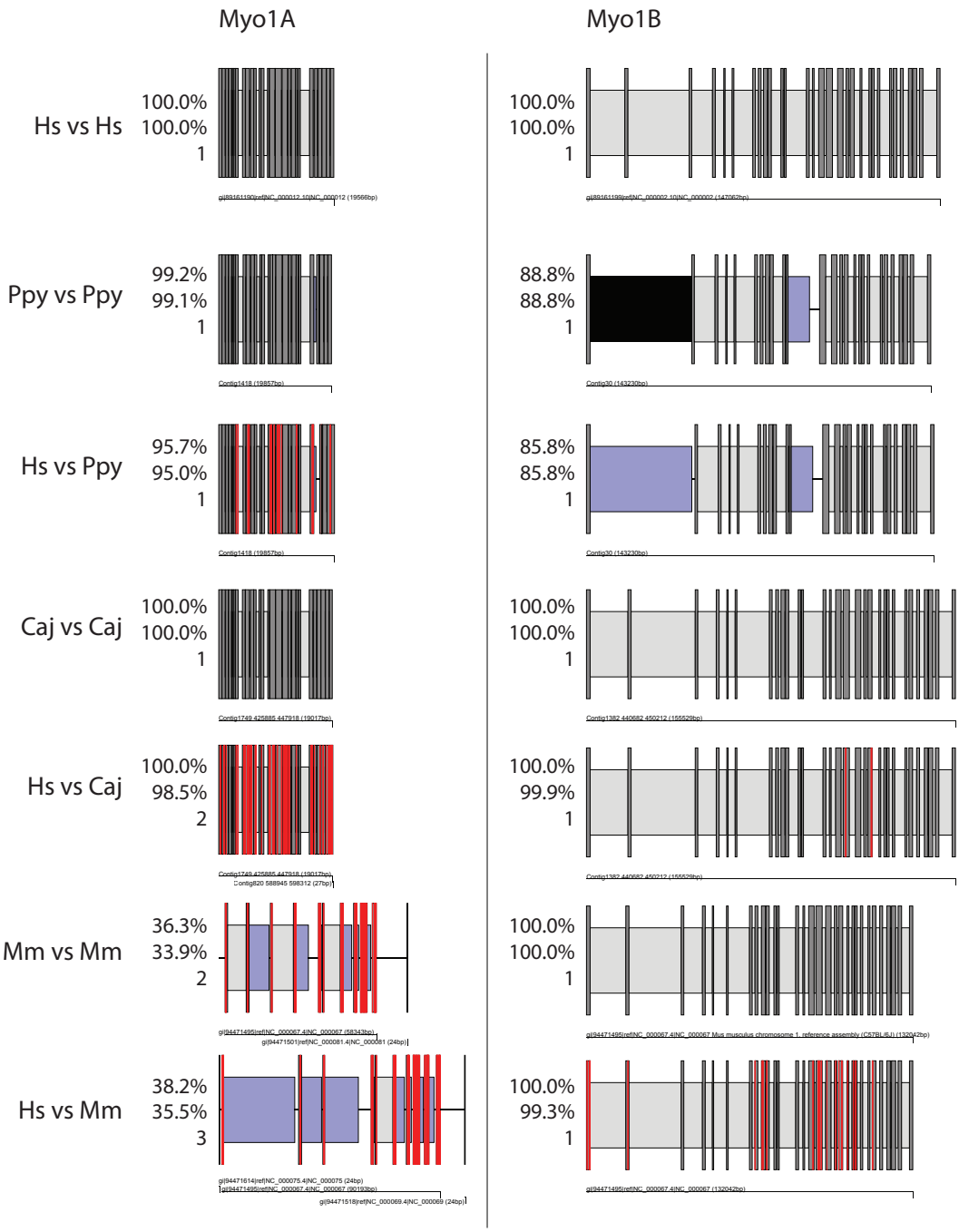

Figure 7.4: Gene structures of Myo1A and Myo1B as determined by WebScipio. Columns are the the different variants of Myosin 1. Rows are either in-species or cross-species searches. Hs: *Homo sapiens*, Ppy: *Pongo pygmaeus*, Caj: *Callithrix jacchus*, Mm: *Mus musculus*. Numbers are: top: percentage of protein that could be mapped, middle: percentage of amino acids that could be mapped, buttom: number of contigs the predicted gene structure has been found on. Dark grey bars are introns, red bars are mismatches or frame shifts, light grey bars are introns with correctly determined splice sites, blue bars are introns without correctly determined splice sites, black bars are regions where amino acids could not be mapped onto the genome although there are nucleotides between the matching regions, central lines are amino acids that have no corresponding nucleotides. Thin lines beneath the gene structure depict the contigs on which the nuleotides have been found. For clarity, intron sequences have been scaled by a factor of 15.

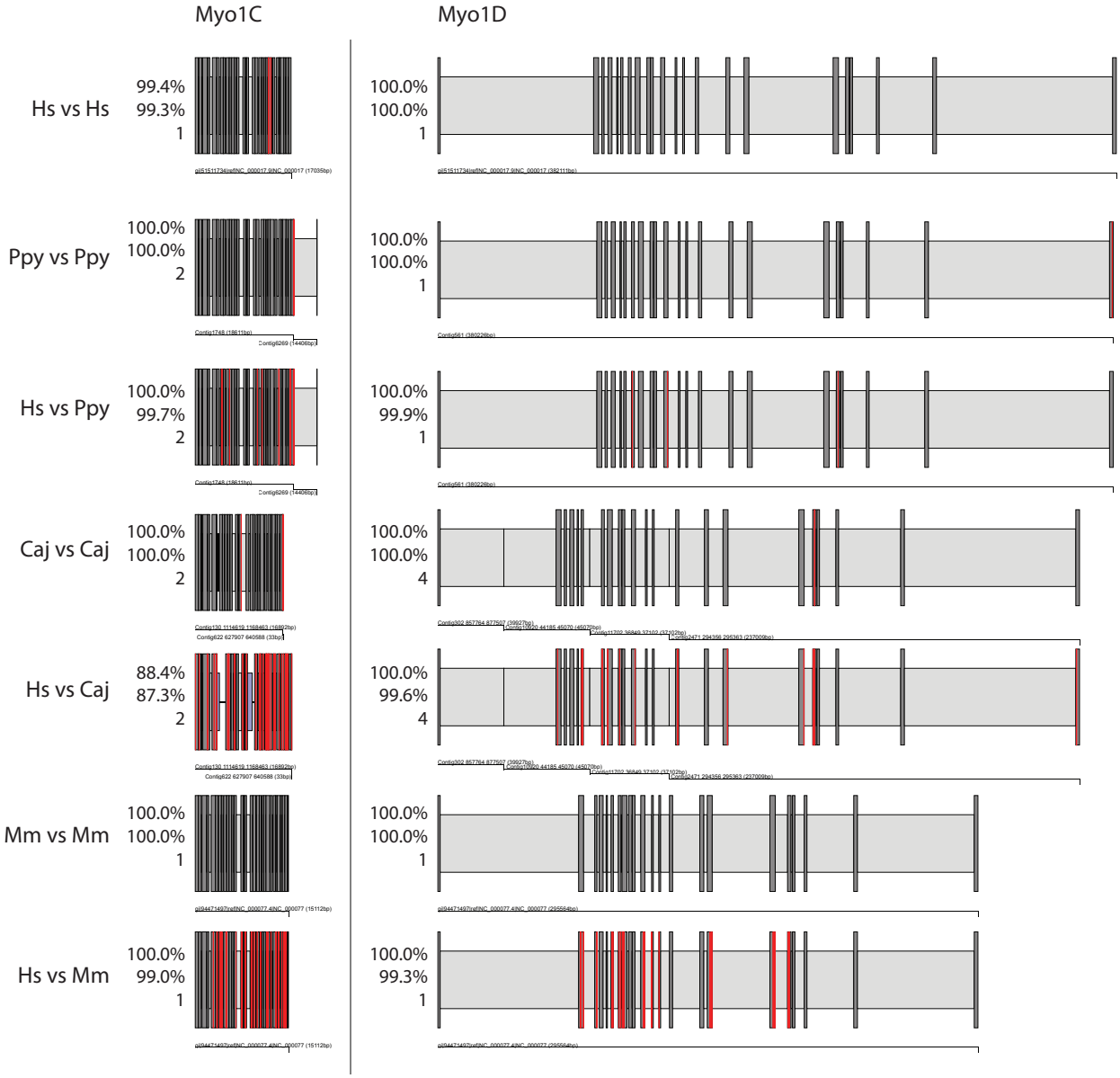

Figure 7.5: Gene structures of Myo1C and Myo1D as determined by WebScipio. Columns are the the different variants of Myosin 1. Rows are either in-species or cross-species searches. Hs: *Homo sapiens*, Ppy: *Pongo pygmaeus*, Caj: *Callithrix jacchus*, Mm: *Mus musculus*. Numbers are: top: percentage of protein that could be mapped, middle: percentage of amino acids that could be mapped, buttom: number of contigs the predicted gene structure has been found on. Dark grey bars are introns, red bars are mismatches or frame shifts, light grey bars are introns with correctly determined splice sites, blue bars are introns without correctly determined splice sites, black bars are regions where amino acids could not be mapped onto the genome although there are nucleotides between the matching regions, central lines are amino acids that have no corresponding nucleotides. Thin lines beneath the gene structure depict the contigs on which the nuleotides have been found. For clarity, intron sequences have been scaled by a factor of 15.

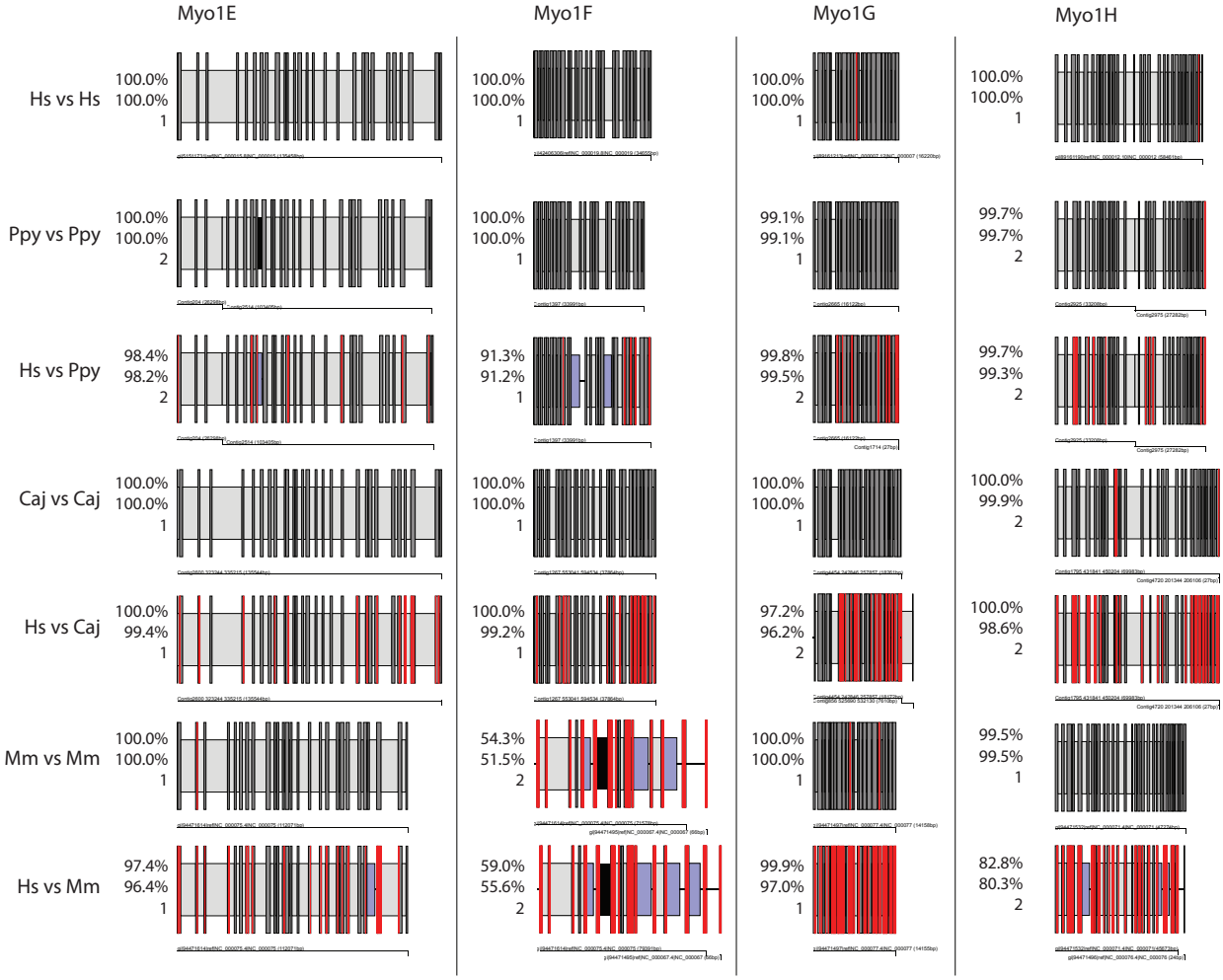

Figure 7.6: Gene structures of Myo1E, MyoF, MyoG and Myo1H as determined by WebScipio.
Columns are the the different variants of Myosin 1. Rows are either in-species or cross-species searches. Hs: *Homo sapiens*, Ppy: *Pongo pygmaeus*, Caj: *Callithrix jacchus*, Mm: *Mus musculus*. Numbers are: top: percentage of protein that could be mapped, middle: percentage of amino acids that could be mapped, buttom: number of contigs the predicted gene structure has been found on. Dark grey bars are introns, red bars are mismatches or frame shifts, light grey bars are introns with correctly determined splice sites, blue bars are introns without correctly determined splice sites, black bars are regions where amino acids could not be mapped onto the genome although there are nucleotides between the matching regions, central lines are amino acids that have no corresponding nucleotides. Thin lines beneath the gene structure depict the contigs on which the nuleotides have been found. For clarity, intron sequences have been scaled by a factor of 15.

7.4.4 Future developments

For many applications it is useful to have information about the structures of genes in closely related species. Therefore, we plan on implementing a feature to select species based on a taxonomic tree. When working with gene families, one might be interested not only in the orthologs in a related

species but also in the paralogs. This could be achieved by displaying not only the best set of hits but also the second and third best.

7.5 Conclusions

WebScipio is a service that maps protein queries onto a genome. All functionality and data resides on the server, so it is not required that the user installs software or downloads large files. WebScipio can be used through its webinterface or as a webservice, allowing for automated querying from within other software programs. The result of a search is a coherent prediction of the gene structure, consisting of a plausible combination of DNA stretches. Since WebScipio combines hits on different contigs, searches in genomes that are in an early stage of assembly are possible. The success rate of in-species searches is very high and the quality approaches the one of manual annotation. For cross-species searches, the tolerance of WebScipio makes it possible to find gene structures even in species with considerable phylogenetic distance to the source organism of the protein sequence.
We think that WebScipio can in many cases provide even non-specialists with gene structure predictions that are plausible and precise, therefore leading to more meaningful analyses.

7.6 Availability and requirements

Project name: WebScipio

Project home page: `http://www.webscipio.org`

Operating system: Platform independent
Programming language: Ruby

Software requirements: WebScipio has been tested with IE6, IE7, Firefox (¿2.0), and Safari.

License: WebScipio may be obtained upon request and used under a Creative Commons License.
Any restrictions to use by non-academics: Using WebScipio by non-academics requires permission.

7.7 Authors contributions

FO and MK set the requirements for the system. FO and HP wrote the software. FO and MK performed testing, and wrote the manuscript. OK improved the Scipio source code. SW supervised the implementation of Scipio. All authors read and approved the final version of the manuscript.

7.8 Acknowledgements

MK has been funded by grant KO 2251/3-1 and KO 2251/6-1 of the Deutsche Forschungsgemeinschaft. We thank all the known and unknown users of WebScipio for their testing and feedback.

8 Reconstructing the phylogeny of 21 completely sequenced arthropod species based on their motor proteins

Florian Odronitz[1], Sebastian Becker[1] and Martin Kollmar[1*]

[1]Department of NMR-based Structural Biology, Max-Planck-Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Goettingen, Germany
[*]Corresponding author.

8.1 Abstract

8.1.1 Background

Motor proteins have extensively been studied in the past and consist of large superfamilies. They are involved in diverse processes like cell division, cellular transport, neuronal transport processes, or muscle contraction, to name a few. E.g. vertebrates contain up to 60 myosins and about the same number of kinesins that are spread over more that a dozen distinct classes.

8.1.2 Results

Here, we present the comparative genomic analysis of the motor protein repertoire of 21 completely sequenced arthropod species using the owl limpet *Lottia gigantea* as outgroup. Arthropoda contain up to 17 myosins each, grouped into 13 classes. The myosins are in almost all cases clear paralogs, and thus the evolution of the arthropod myosin inventory is mainly determined by gene losses. Arthropod species contain up to 29 kinesins each, spread over 13 classes. In contrast to the myosins, the evolution of the arthropod kinesin inventory is not only determined by gene losses but also by many taxon- and species-specific gene duplications. All arthropods contain each of the subunits of the cytoplasmic dynein/dynactin complex. Except for the dynein light chains and the p150 dynactin subunit they contain single gene copies of the other subunits. Especially the roadblock light chain repertoire is very species specific.

8.1.3 Conclusions

Every of the 21 completely sequenced arthropods, including the twelve sequenced *Drosophila* species, contains a species-specific set of motor proteins. The phylogenetic analysis of all genes as well as the protein repertoire placed *Daphnia pulex* closest to the root of the Arthropoda. The louse *Pediculus humanus corporis* groups to the honeybee *Apis mellifera* and the jewel wasp *Nasonia vitripennis*. After this group the rust-red flour beetle *Tribolium castaneum* and the silkworm *Bombyx mori* diverged very closely from the lineage leading to the *Drosophila* species.

8.2 Background

Nearly each single cell in eukaryotes hosts particular proteins which are responsible for intracellular transport. These molecular motor molecules are highly conserved among the different species of eukaryotes and evolved slowly over time (143). This property grant them the role of an appropriate candidate to carry out evolutionary studies. The three superfamilies of transporting motor proteins are the myosins, kinesins and dyneins. Attached to the cytoskeletal networks (microtubules and actin) they transport all kinds of organelles and vesicles, remodel the cytoskeleton and organize developmental processes in eukaryotes (144). Energy for their unidirectional cargo transport on one of the filamentous cytoskeletal tracks is derived from ATP hydrolysis (145). Out of the three superfamilies only the members of the kinesin superfamily are found in all eukaryotes, whereas not all members of the dynein (146) and myosin (145) superfamilies has been found in particular eukaryotic lineages.

The members of the actin-based myosin family have their origin early in eukaryotic evolution. Based on the latest analysis, the myosins are grouped into 35 classes (83). Their domain structures consist of three regions, the motor (or head) domain, a neck domain, and the tail, which comprises all C-terminal domains as well as domains N-terminal to the motor domain. The motor domain is highly conserved and contains both the ATP and actin binding site where the force generation resides. This energy-transducing motor domain is coupled to a regulatory neck region (helical region) which is able to bind calmodulin or calmodulin-like light chains. Linked to the neck region most myosins have tail domains. Contrary to the head domains the tail domains show high variability in sequence and length, thus reflecting their functional diversity. This diversity ranges from eukaryotic cytokinesis, organellar and intracellular transport, cell polarization to signal transduction. Some of the myosin classes also contain large domains at the N-terminus of the motor domains.

The second molecular motor protein family is kinesin (members also known as KRPs, KLPs, or KIFs) (147, 148). The members of this superfamily are microtubule-based and provide movement in both directions (either plus or minus end-directed) (149). For their movement along the microtubules they utilize ATP similarly to the other motor proteins. The classical kinesin forms a tetramer with two kinesin heavy chains (KHCs) and two kinesin light chains (KLCs). The structural parts of kinesins comprise the motor domain, the neck, and the stalk. Like in myosins the head domain is well conserved and responsible for the movement by hydrolysis of ATP, whereas the stalk and

tail domains play fundamental roles in the interaction with other subunits of the holoenzyme or with cargo molecules such as proteins, lipids or nucleic acids. The tail region between the head and the stalk varies most which is due to family-specific features e.g. direction of motility as well as regulation of activity. Kinesin binds a variety of cargoes and perform force-generating tasks such as transport of vesicles and organelles, spindle formation and elongation, chromosome segregation, and MT organization (149) (150).
The members of the dynein superfamily are minus end-directed motor proteins (151). Thus, they are responsible for the retrograde transport of cargos along microtubules toward the centrosome. They are involved in many processes like spindle formation as well as chromosome segregation, and the transport of a variety of cargoes like viruses, RNA, signaling molecules, and organelles (152). Dyneins are multi-subunit protein complexes with two or three heavy chains (DHCs), light chains, light intermediate, and intermediate chains (153). Supported by an activator protein called dynactin which consists of 11 subunits dynein is able to move and bind to membranes or further cargos (154).
The genome of *Drosophila melanogaster* was the third eukaryotic genome to be completely sequenced (90). Since then, the number of sequenced organisms has increased rapidly. Of the Arthropoda phylum, the genomes of the mosquitos *Anopheles gambiae* (90) and *Aedes aegypti* (91) and the silkworm *Bombyx mori* (155) have been published, and 17 further insect genomes have been finished of which eleven belong to the *Drosophila* species group (93, 156).
Here, we present the analysis of the phylogenetic relationship of 21 completely sequenced arthropods based on their motor protein inventory.

8.3 Results

8.3.1 Identification and annotation of the motor proteins

The arthropod motor protein genes were identified by TBLASTN searches against the corresponding genome data of the different species. Species, that missed certain orthologs in the first instance, were searched again with supposed-to-be orthologs of the other species. In this iterative process all motor proteins have been identified or their loss in certain species was confirmed. The species analyzed were the mosquitos *Aedes aegypti*, *Culex pipiens quinquefasciatus* and *Anopheles gambiae*, the silkworm *Bombyx mori*, the honeybee *Apis mellifera*, the jewel wasp *Nasonia vitripennis*, the waterflea *Daphnia pulex*, the rust-red flour beetle *Tribolium castaneum*, the body louse *Pediculus humanus corporis*, twelve *Drosophila* species, and the mollusc *Lottia gigantea* which we used as outgroup. The sequences were assigned by manual inspection of the genomic DNA sequences. Exons have been confirmed by the identification of flanking consensus intron-exon splice junction donor and acceptor sequences (99). The genomic sequences of *Drosophila virilis*, *Apis mellifera*, and especially *Bombyx mori* contain several gaps. Many of the gaps have been filled by analyzing EST data.

8.3.2 Analysis of the arthropod myosins

All myosins have been classified based on the phylogenetic analysis of their motor domains together with the motor domains of the already grouped myosins (83). All myosins belong to previously defined classes except one myosin of *Nasonia* that has a very similar domain organization to the class-V myosins but a considerably different motor domain (Figure 8.2). During their evolution the Arthropoda decreased their myosin diversity. *Daphnia* which is closest to the origin of the Arthropoda has the larges repertoire still containing a class-XIX myosins, that all other analyzed arthropods have lost, and four class-I myosins. Class-XIX myosins have else been found in Deuterostomia and Cnidaria. Also, all other arthropods have lost at least one of the class-I myosins. However, the retained variants between the analyzed species are different which means, that they lost the class-I variants after separating from the next closest species. For example, *Apis* and *Nasonia* both have lost the class-I myosin variant C, that the closest relative *Pediculus* still has, but in addition specifically lost the variant D and variant B, respectively. All arthropods contain a non-muscle as well as a muscle myosin heavy chain gene (class-II myosins). The alternatively spliced muscle myosin heavy chain genes have been described elsewhere (Odronitz and Kollmar, submitted). The *Drosophila* species and *Tribolium* have lost the class-3 myosin. The *Drosophila melanogaster* NinaC protein has previously been classified as a class-III myosin. Based on the analysis of the more than 2000 myosin the NinaC protein does not group to the vertebrate class-III myosins and all arthropod homologs of NinaC have been grouped into a new class, the class-XXI (83). Surprisingly, *Nasonia* does not contain a class-VI myosin, that all other Metazoa, that have been analyzed so far, contain. The lack of the class-VI myosin might be a specific characteristic of *Nasonia vitripennis*, or due to sequencing and assembly problems which are, however, unlikely given the high coverage of the *Nasonia* genome sequencing. Finishing of the genomes of the other two *Nasonia* species, whose genome sequences are in progress at the Baylor College of Medicine, will either confirm the lineage specific loss of the class-VI myosin or reveal sequencing problems of the *Nasonia vitripennis* genome. *Daphnia*, *Pediculus*, and *Apis* have lost the variant B of the class-VII myosin. The class-VII myosin which they contain is a clear homolog of the class-VII variant A myosins of the other arthropods. The *Drosophila* lineage has also completely lost the class-IX myosin. All arthropod genomes contain a class-XV, a class-XVIII, a class-XX, and a class-XXI myosin. The class-XXII myosin has independently been lost by several sub-lineages of the *Drosophila* species. The *Drosophila* species, that have been marked as having their class-XXII myosin lost all still contain some of the exons of the ancient class-XXII myosin but spread over several hundred thousands of base pairs so that it is almost impossible that these pieces might belong to still functional genes.

The domain organizations of the arthropod myosins are identical to those found for other members of the respective classes (83). Figure 8.2 shows diagrams of the *Daphnia* myosins that have the largest diversity of the arthropod myosins. The class-XXI myosins have an identical domain organization as the class-III myosin, although the phylogenetic analysis of their motor domain reveals two distinct classes. It is highly probable, that the class-XXI myosins are the result of an arthropod specific gene duplication of the ancient class-III myosin followed by the divergence and specialization of the

new duplicate. The class-XXII myosins have a similar domain organization as the class-VII myosin. In contrast to the class-VII myosins they lack the N-terminal SH3-like domain, they contain three instead of five IQ-motifs for the binding of calmodulin-like light chains, they have a longer coiled-coil regions containing domain till the first MyTH4 domain, and they lack the SH3 domain of the C-terminal tail.

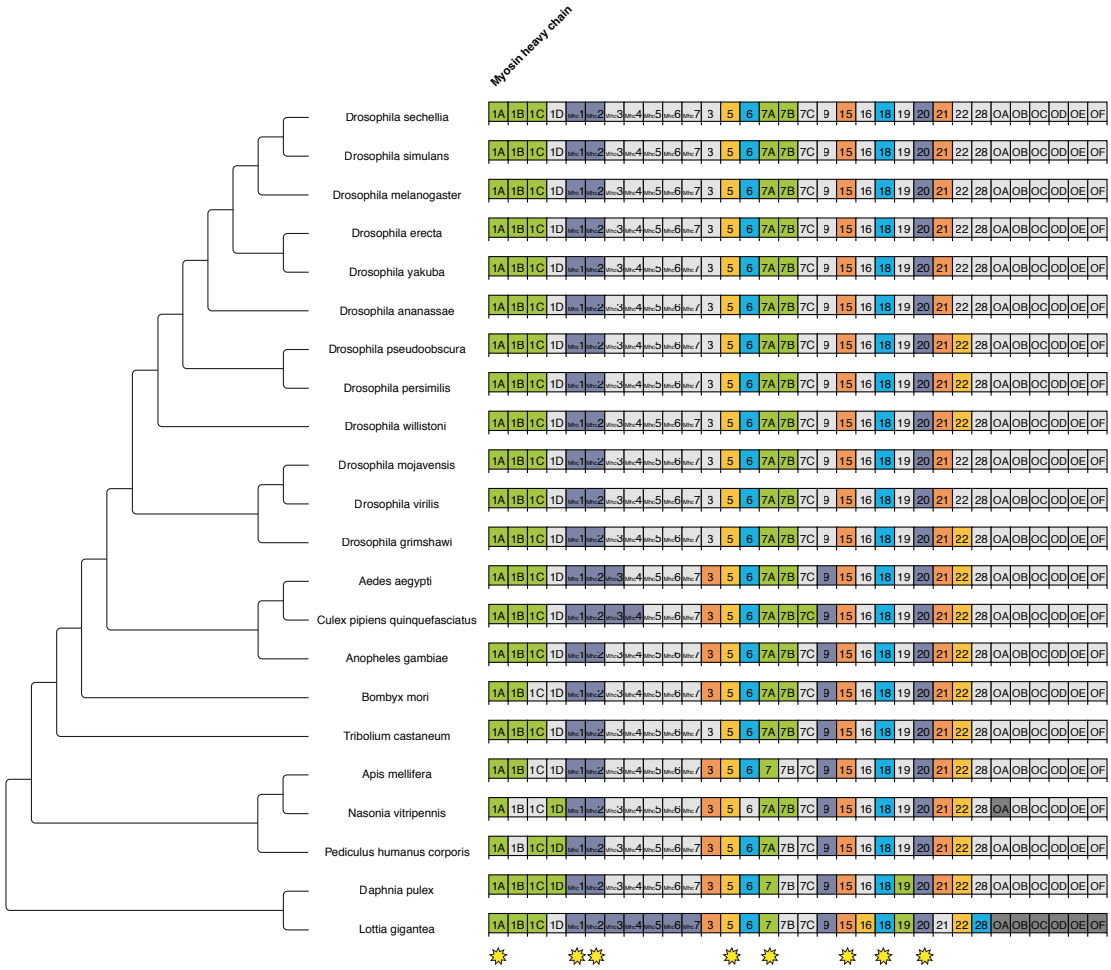

Figure 8.1: Protein Inventory: Myosins
This chart shows the protein inventory of myosins for all species in the analysis. To the left is a schematic phylogenetic tree, roughly depicting the relationships (no scale). The identifiers in the boxes indicate protein classes/variants. "O" means orphan class. Colored boxes mean the class/variant exists in this species. Grey boxes mean the class/variant is not found. Columns marked with stars were included in the phylogenomics analysis.

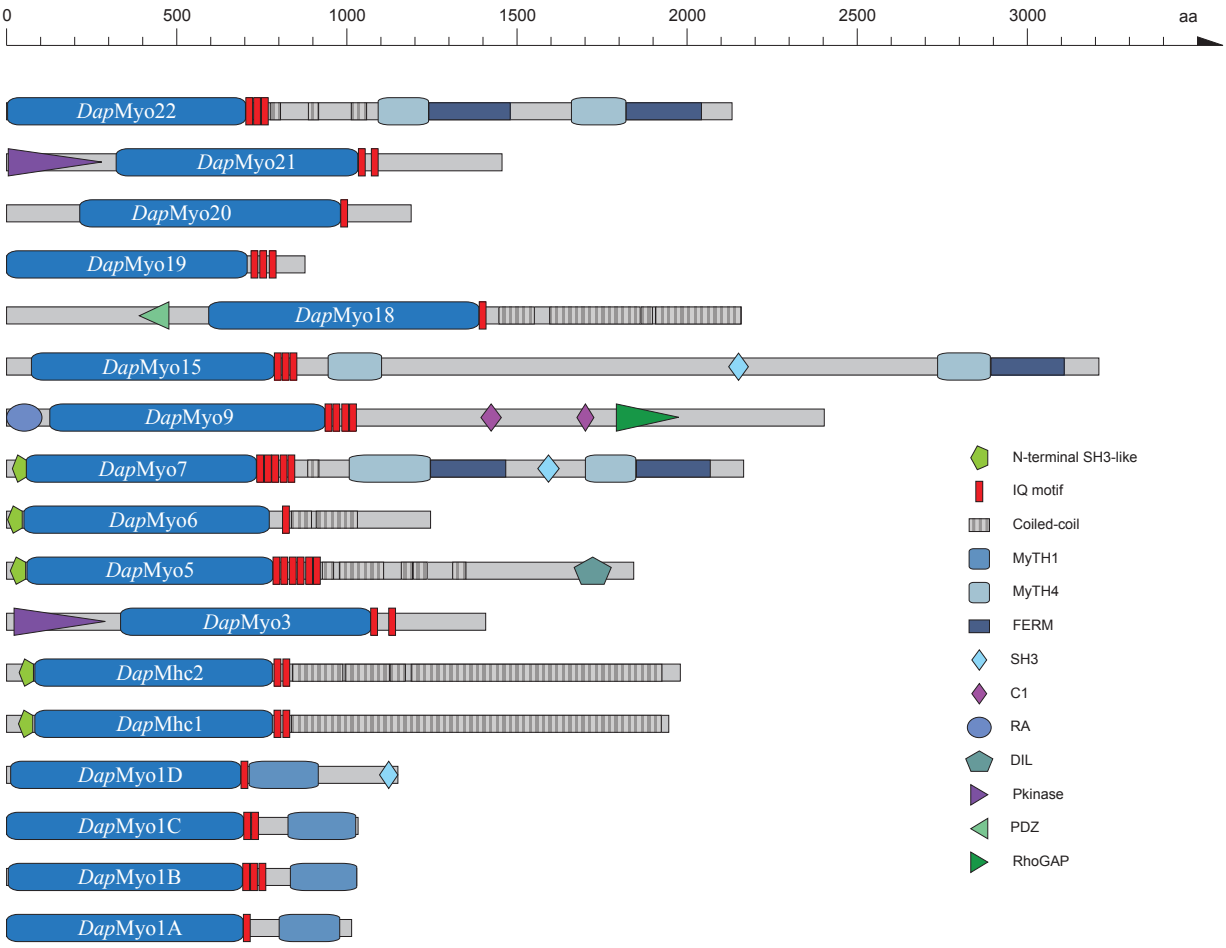

Figure 8.2: Domain organisation of the *Daphnia pulex* myosins
The sequence name is given in the motor domain of the respective myosin. A colour key to the domain names and symbols is given on the right except for the myosin domain that is coloured in blue. The abbreviations for the domains are: C1, Protein kinase C conserved region 1; DIL, dilute; FERM, band 4.1, ezrin, radixin, and moesin; IQ motif, isoleucine-glutamine motif; MyTH1, myosin tail homology 1; MyTH4, myosin tail homology 4; PDZ, PDZ domain; Pkinase, Protein kinase domain; RA, Ras association (RalGDS/AF-6) domain; RhoGAP, Rho GTPase-activating protein; SH3, src homology 3.

8.3.3 Analysis of the arthropod kinesins

For their classification, the kinesin motor domains have been used in a phylogenetic analysis together with the motor domains of the human kinesins (157, 158). The sequences have been named according to the standardized kinesin nomenclature (146) leaving some kinesins unclassified (Figure 8.3). Orphan kinesins, that are clear homologs, got the same variant designation to allow for a better comparison. In general, all analyzed species contain species-specific sets of kinesins. Except for *Drosophila pseudoobscura* and *Drosophila persimilis*, that have the same set of kinesins, even closest related species have different kinesin inventories. Thus, the evolution of the kinesin inventories of the arthropods is strongly determined by species specific gene duplications and gene losses. It is impossible to identify lineage-specific duplication and loss events. Some gene duplications and gene losses

are especially interesting. In this respect, we will not consider the kinesin inventory of *Bombyx mori* because the genome has not been sequenced with high coverage and is strongly fragmented. Thus, further *Bombyx mori* kinesins will certainly be identified. *Drosophila ananassae* does not contain a kinesin-2C that all other arthropods have. *Drosophila willistoni* does not contain a kinesin-4A, but two class-VI kinesins and two species-specific kinesins that have not been classified yet, kinesin-D and kinesin-E. While most arthropods contain only one kinesin-5, *Tribolium* contains a set of four class-V kinesins. The *Pediculus* genome does not encode a class-VII kinesin, but encodes a kinesin-9, that is otherwise only found in *Apis*. None of the analyzed arthropods contains a kinesin-10. *Nasonia* does not contain a kinesin-12, that all other arthropods have. The set of class-XIII kinesins in the arthropods ranges from one to four homologs. *Tribolium*, *Apis*, *Nasonia*, *Pediculus*, and *Daphnia* contain one or two further kinesins that could not be grouped to any of the known classes. The *Daphnia* kinesins mainly consist of the kinesin motor domain and long coiled-coil regions in the tail (Figure 8.4). Only the class-III kinesins contain further domains that have been characterised and named. A characteristic of almost all class-III kinesins is an FHA domain following C-terminal to the motor domain. The class-III variant A kinesins also contain a CAP-Gly domain at the C-terminus, while the variant B kinesins contain a PH domain.

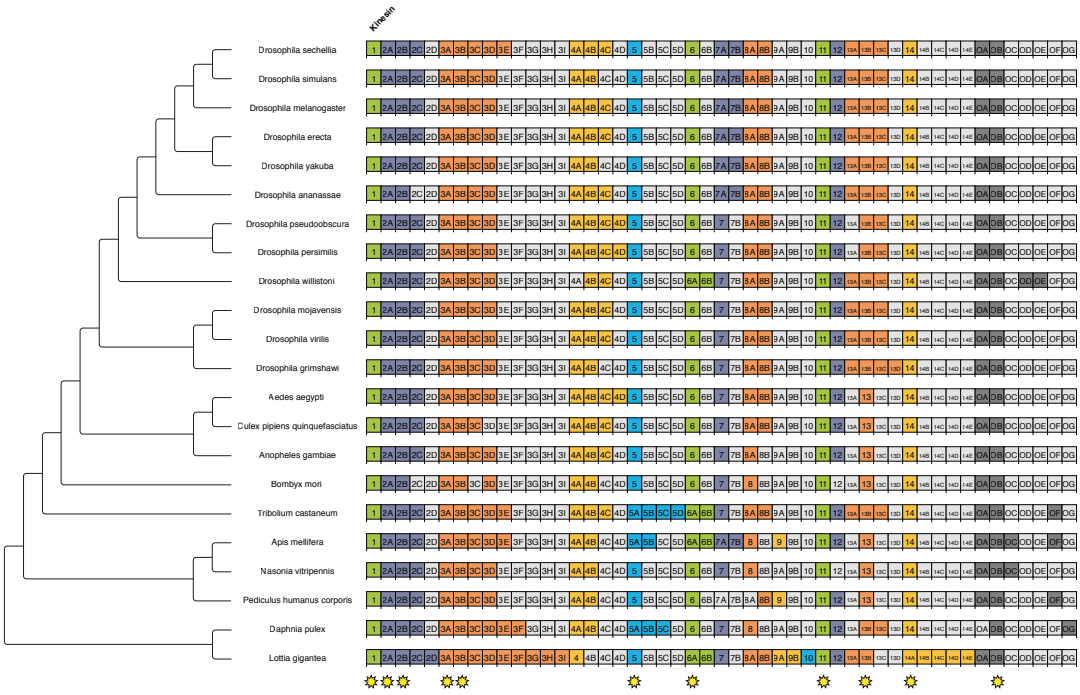

Figure 8.3: Protein Inventory: Kinesins
This chart shows the protein inventory of kinesins for all species in the analysis. To the left is a schematic phylogenetic tree, roughly depicting the relationships (no scale). The identifiers in the boxes indicate protein classes/variants. "O" means orphan class. Colored boxes mean the class/variant exists in this species. Grey boxes mean the class/variant is not found. Columns marked with stars were included in the phylogenomics analysis.

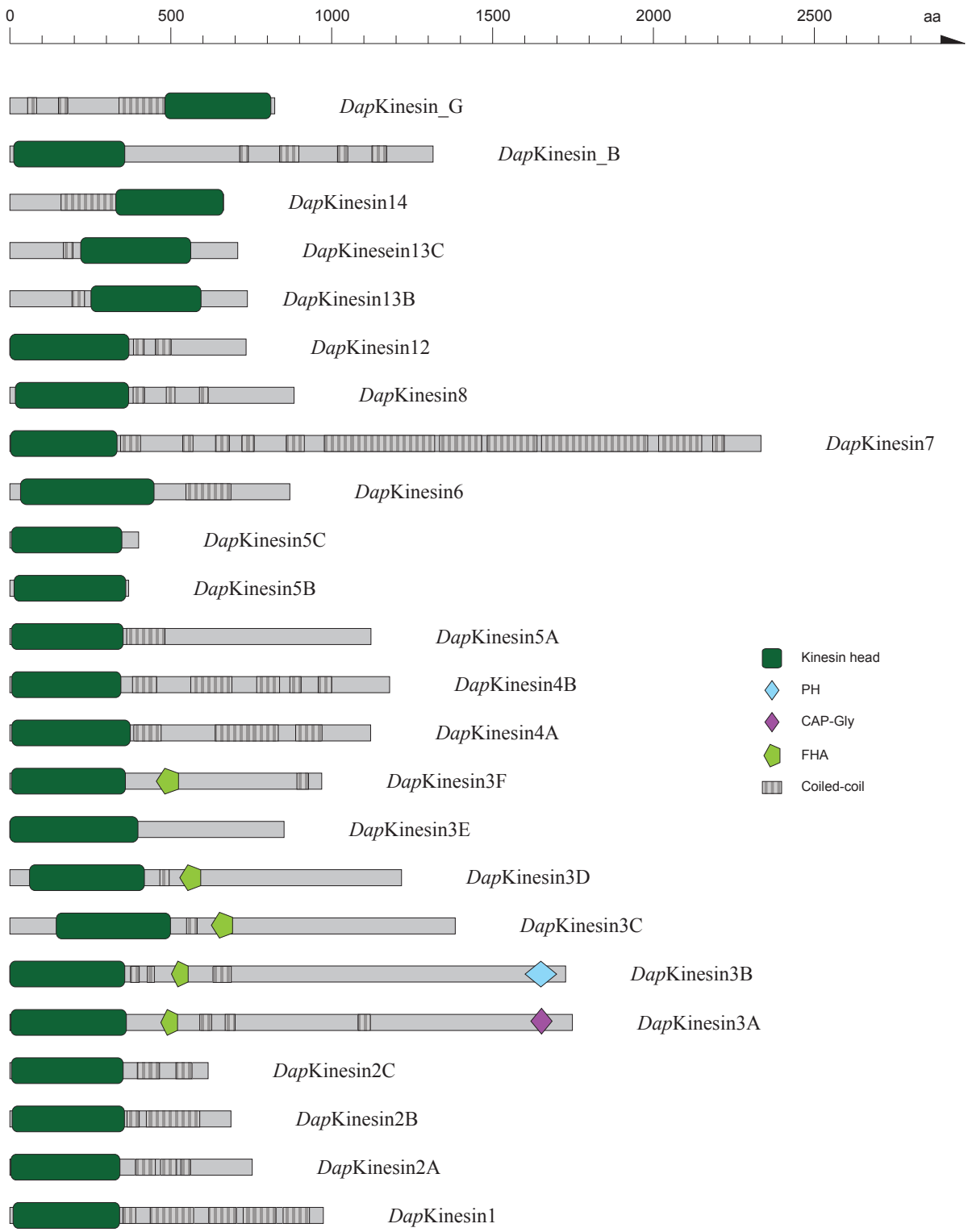

Figure 8.4: Domain organisation of the *Daphnia pulex* kinesins
The sequence name is given next to the respective kinesin. A colour key to the domain names and symbols is given on the right except for the kinesin domains that are coloured in dark-green. The abbreviations for the domains are: CAP-Gly, Cytoskeleton-associated protein-Gly; FHA, forkhead homology associated; PH, pleckstrin homology.

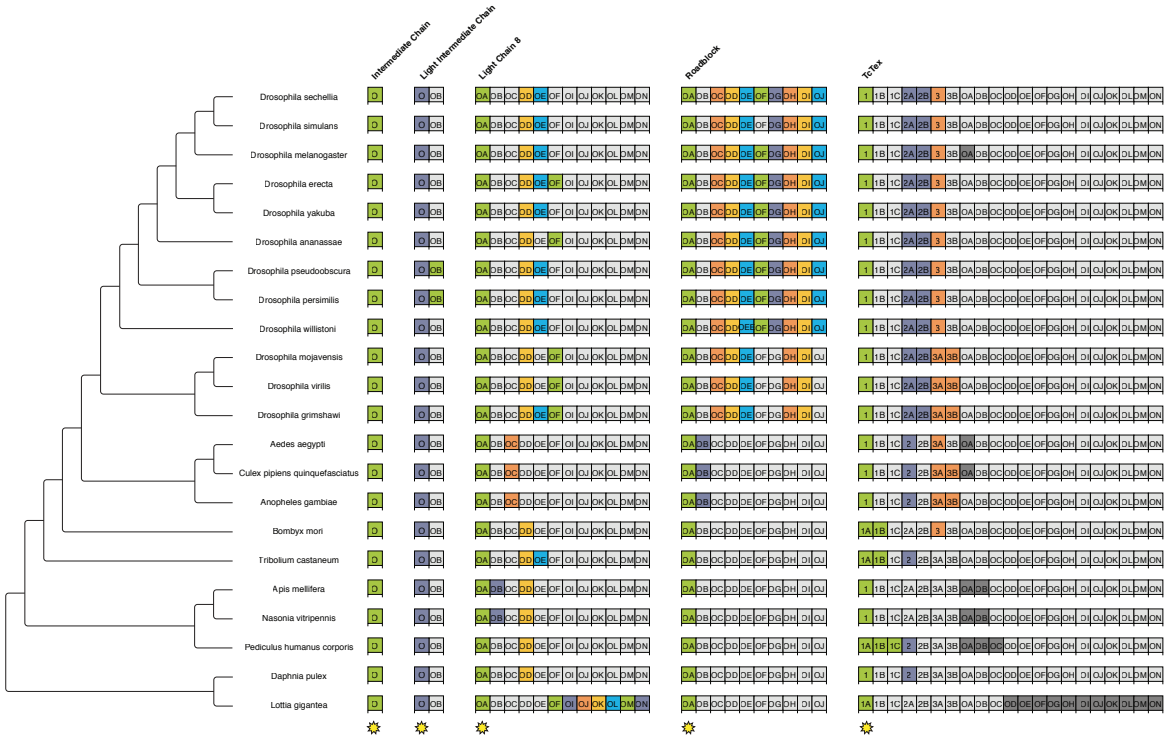

Figure 8.5: Protein Inventory: Dyneins
This chart shows the protein inventory of dyneins for all species in the analysis. To the left is a schematic phylogenetic tree, roughly depicting the relationships (no scale). The identifiers in the boxes indicate protein classes/variants. "O" means orphan class. Colored boxes mean the class/variant exists in this species. Grey boxes mean the class/variant is not found. Columns marked with stars were included in the phylogenomics analysis.

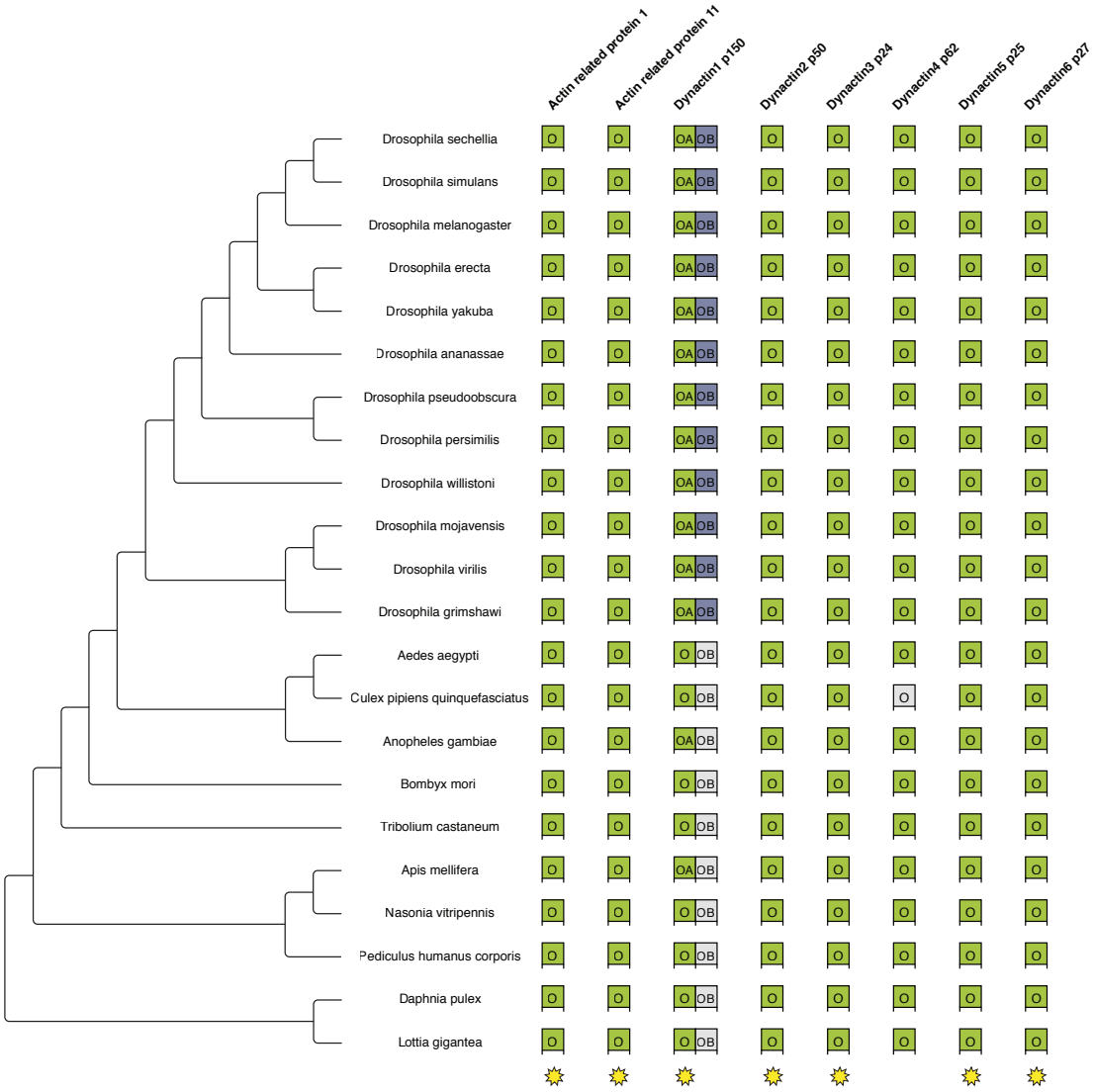

Figure 8.6: Protein Inventory: Actin related proteins and dynactins
This chart shows the protein inventory of actin related proteins and dynactins for all species in the analysis. To the left is a schematic phylogenetic tree, roughly depicting the relationships (no scale). The identifiers in the boxes indicate protein classes/variants. "O" means orphan class. Colored boxes mean the class/variant exists in this species. Grey boxes mean the class/variant is not found. Columns marked with stars were included in the phylogenomics analysis.

8.3.4 Arthropod phylogeny

When analysing the phylogeny of several homologs in a set of species, each homolog might results in a different phylogeny. This is a result of the different rates of evolutionary change for different genes. In order to compensate for this asynchronous evolution, a phylogenomics approach was used to infer the phylogeny. For each protein, the variants/classes for which a homolog exists in every species were concatenated resulting in sequences that are more representative. For the dynein, dynactin and ARP

proteins, only one homolog was found in all species, whereas for myosin and kinesin eight and ten were found, respectively (marked with stars in Figures 8.1, 8.3, 8.5 and 8.6). When inspecting the trees from all proteins, it can be stated that three clades and their internal topologies are constant: The *Drosophila* clade, a clade of *Apis mellifera* and *Nasonia vitripennis* and the clade of *Aedes aegypti*, *Culex pipiens quinquefasciatus* and *Anopheles gambiae*. Only in the tree of LC8, the clade of *Anopheles*, *Aedes* and *Culex* is placed within the *Drosophila* clade. All other species were placed at different branches, where the discrepancy among dynein, dynactin and ARP was higher when compared to myosin and kinesin. The trees calculated from myosin and kinesin only disagree in the position of *Bombyx mori*, *Tribolium castaneum* and *Pediculus humanus corporis*. In order to obtain trees with greater fidelity, all common homologues from each species where concatenated. For each of the 22 species, 31 homologs were used, amounting to 682 motor protein sequences. The resulting trees are shown in Figure 8.7. Except for *Tribolium*, all four phylogenomics trees show identical phylogeny. All branches are supported with very high bootstrap values and are therefore reliable within the limits of the method. The difference in placement of *Pediculus* depends on which method is used. In the trees generated with neighbour joining, *Pediculus* forms a clade with *Nasonia* and *Apis*, whereas with maximum likelihood, only *Nasonia* and *Apis* are monophyletic and *Pediculus* is more closely related to *Daphnia*. The phylogenetic tree inferred from the occurrence of classes/variants has a limited resolution and agrees only in some respects with the maximum likelihood tree: *Drosophila* form a clade, *Drosophila pseudoobscura* and *Drosophila persimilis* are monophyletic, *Drosophila virilis*, *Drosophila mojavensis* and *Drosophila grimshawi* are monophyletic and *Culex*, *Aedes* and *Anopheles* are monophyletic. In all trees *Lottia gigantea* is the most divergent species.

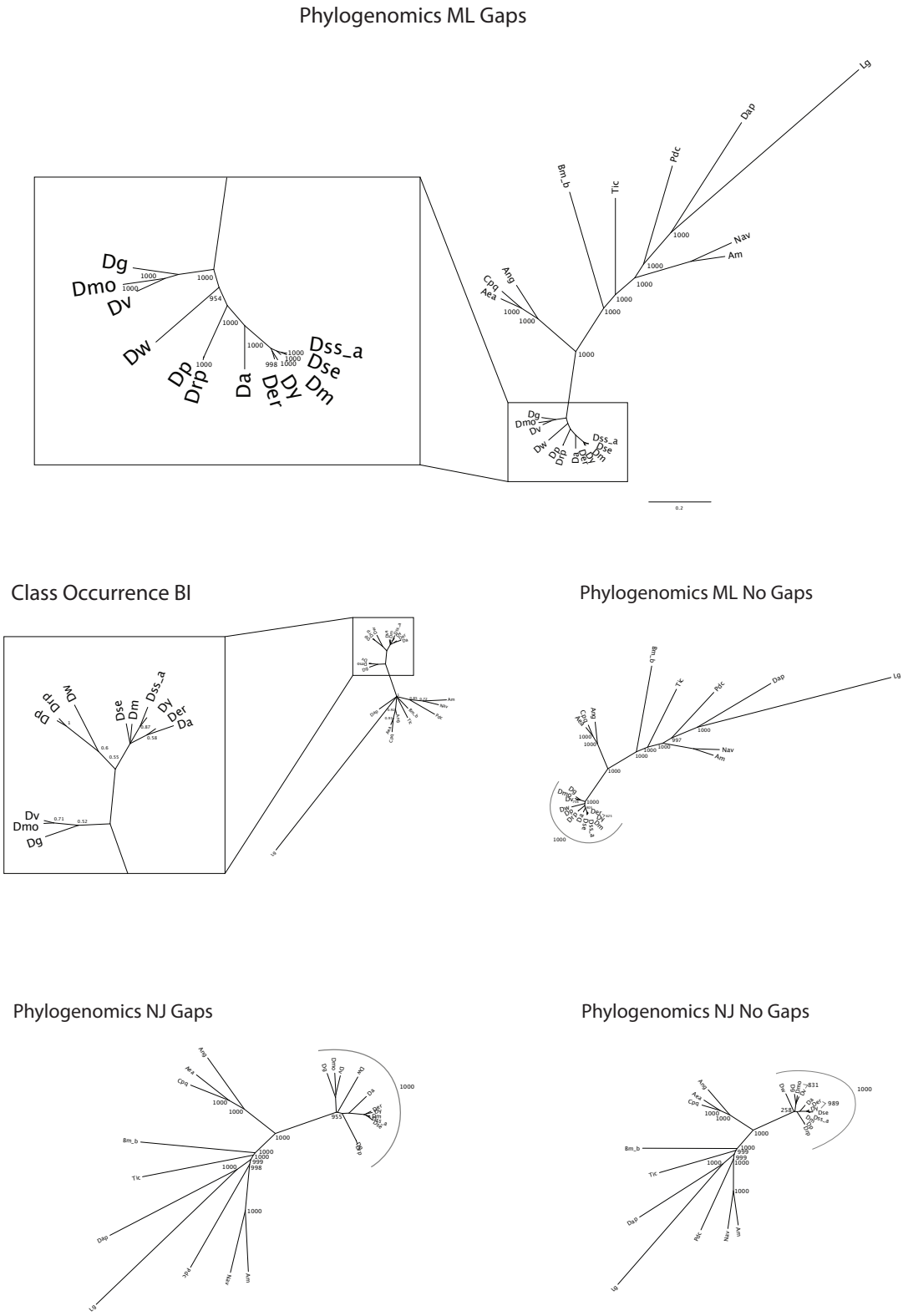

Figure 8.7: Phylogenomics and Class Occupation
The trees illustrate the phylogenetic relationship between the arthropod species. The phylogenomics trees are based on a total of 660 cancatenated protein sequences. Methods as indicated. The class occupation tree was constructed using Bayesian inference based on the presence or absence of protein classes/variants as indicated in the inventory.

8.4 Discussion

Most of the myosins that we discuss here have been identified and annotated in the course of the annotation of over 2000 myosins from more than 300 organisms (83). Since then, the genome sequences of the arthropod species *Culex pipiens quinquefasciatus* and *Pediculus humanus corporis* have been finished as well as that of the mollusc *Lottia gigantea* which we used as outgroup.
It has been observed, given heterogeneous evolutionary rates, that the results of the maximum likelihood method are statistically more robust than the ones produced by neighbour joining (159). Therefore we conclude that *Apis*, *Nasonia*, and *Pediculus* are not monophyletic, but that *Pediculus* is more closely related to *Daphnia*.
The class occurrence tree shows that the classification system we used for the protein families does not contradict the finding of the sequence-based phylogenetic inference.
The phylogeny of *Drosophila* is in exact agreement with what has been found in an analysis based on the complete genome sequences of the twelve species (160). In our analysis, this clade is the one with the most closely related species, and therefore the one that is the hardest to resolve. Therefore it can be regarded as a benchmark of the quality of the whole tree.
Our study suggests the following phylogeny: The *Drosophila* clade is composed of the *Drosophila simulans*/ *Drosophila sechella* clade which forms a clade with *Drosophila melanogaster*. This clade together with the *Drosophila yakuba*/*Drosophila* erecta clade forms the melanogaster subgroup. This subgroup together with *Drosophila ananassae* forms the melanogaster group. The melanogaster group is most closely related to the obscura group, a clade that consists of *Drosophila pseudoobscura* and *Drosophila persimilis*. The closest relative to the obscura group is *Drosophila willistone*. All of the before mentioned species form the subgenus Sophophora. Its sister subgenus is *Drosophila*, consisting of the clade of *Drosophila virilis*/*Drosophila mojavensis* and *Drosophila grimshawi* (taxonomy as in (160)).
The closest relatives to the *Drosophila* clade are *Aedes aegypti* and *Culex pipiens*, forming one clade and *Anopheles gambiae*. The next closest relatives are *Bombyx mori*, *Tribolium castaneum*, followed by the clade *Nasonia vitripennis*/*Apis mellifera*, *Daphnia pulex* and *Lottia gigantea*.

8.5 Conclusions

In this analysis, we were able to resolve the phylogenetic relationship of 21 completely sequenced arthropod species based in their motor proteins. A large number of sequences that have been checked manually were used. We were able to systematically analyse the protein inventory of all species as well as the domain composition of all memebers of the four protein families in *Daphnia pulex*. When inferring phylogenetic trees from the sequence data, variations in evolutionary speed were accounted for by using a phylogenomics approach. This analysis produces a phylogenetic tree that is highly resolved and that has statistically well supported branching. Our findings are in accordance which results from studies based on whole genome sequences. We can conclude that from all arthropods analysed, *Daphnia pulex* is the most basal one. Pediculis humanus and corpiris form one group, as

well as *Apis mellifera* and nasonia vitripennis. Next, the group of *Tribolium castaneum* and Bambyx mori deverged, followed by the *Drosophila* clade.

8.6 Materials and Methods

8.6.1 Identification and annotation of the arthropod myosin, kinesins, and dynein/dynactin subunits

The genes for Aea, Ang, Am, Bm, Cpq, Da, Der, Dg, Dm, Dmo, Drp, Dp, Dse, Dss, Dv, Dy, Dw, Nav, Pdc, and Tic have been obtained by TBLASTN searches against the insects section of the NCBI wgs database. The Dap sequences have been obtained by TBLASTN searches against the 9x assembly of the *Daphnia pulex* genome provided by the DOE Joint Genome Institute and the *Daphnia* Genomics Consortium. All hits were manually analysed at the genomic DNA level. The correct coding sequences were identified with the help of the multiple sequence alignments of the corresponding proteins. In this process, the sequence alignments of all proteins contained in our in-house version of CyMoBase have been used. As the amount of protein sequences increased (especially the number of sequences in classes with few representatives), many of the initially predicted sequences were reanalysed to correctly identify all exon borders. Where possible, EST data available from the NCBI est database has been analysed to help in the annotation process. All sequence related data (names, corresponding species, GenBank ID's, alternative names, corresponding publications, domain predictions, and sequences) and references to genome sequencing centers are available through the CyMoBase (37).

8.6.2 Building trees

The phylogenetic trees based on protein sequence where generated using two different methods: 1. Neighbour joining using the GONNET substitution matrix with bootstrapping (1,000 replicates) using ClustalW 2.0 (161). 2. Maximum likelihood (ML) (162) using a JTT model with estimated proportion of invariable sites and bootstrapping (1,000 replicates) using PHYML (163).
The sequence data that was used for the analyses were multiple sequence alignments consisting either of single homologous sequences from each species or multiple concatenated homologous sequences from each species (phylogenomis approach). For comparison, multiple sequence alignments were used including columns with gaps or with columns containing gaps removed.
The class occurrence tree was generated using Bayesian inference with a binary model using MrBayes 3.1.2 (164). For each species the existence/non-existence of a protein class/variant was used as a binary character as depicted in Figures 8.1, 8.3, 8.5 and 8.6. Using this encoding, each species is represented by a series of binary characters, one for each protein class/variant. Constant rates were used whereas gamma-distributed rates gave very similar results. The tree was generated using 1.000.000 generations and standard settings.

8.6.3 Domain and motif prediction

Protein domains were predicted using the SMART (165) and Pfam (33) web server. The prediction of protein motifs (coiled coils, leucine zipper, etc.) is mainly based on the results of the predict-protein server (166). The IQ-motifs and N-terminal domains of the myosins were predicted manually based on the homology to similar domains of other myosins included in the multiple sequence alignment of the myosins. The recognition motifs included in the SMART and Pfam databases are too restrictive, as the motifs have been created based on the small datasets available some years ago.

8.7 Authors contributions

FO performed the data analysis of the myosins, kinesins, and dynein subunits and the combined analysis. SB assembled all dynactin sequences and performed their analysis. MK assembled all myosin, kinesin, and dynein sequences. All authors wrote and approved the manuscript.

8.8 Acknowledgements

This work has been funded by grant I80798 of the VolkswagenStiftung and grant KO 2251/3-1 of the Deutsche Forschungsgemeinschaft. The sequencing and portions of the analyses were performed at the DOE Joint Genome Institute under the auspices of the U.S. Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under Contract No. DE-AC02-05CH11231, Los Alamos National Laboratory under Contract No. W-7405-ENG-36 and in collaboration with the *Daphnia* Genomics Consortium (DGC) http://daphnia.cgb.indiana.edu. Additional analyses were performed by wFleaBase, developed at the Genome Informatics Lab of Indiana University with support to Don Gilbert from the National Science Foundation and the National Institutes of Health. Coordination infrastructure for the DGC is provided by The Center for Genomics and Bioinformatics at Indiana University, which is supported in part by the METACyt Initiative of Indiana University, funded in part through a major grant from the Lilly Endowment, Inc. Our work benefits from, and contributes to the *Daphnia* Genomics Consortium.

9 Peakr: Predicting solid state NMR spectra of Proteins

Florian Odronitz [1*], Robert Schneider [1*] and Martin Kollmar [1]

[1]Department of NMR-based Structural Biology, Max-Planck-Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Goettingen, Germany
[*]These authors contributed equally.

9.1 Abstract

When analyzing spectra of proteins in solid state NMR, the assignment of resonances to atoms and the derivation of constraints for 3D structure calculations are challenging and time-consuming processes. Predicted spectra that have been calculated based on assumptions about the physical couplings between atoms during measurement can be of considerable help. Existing solutions are very limited in the type of experiment they can consider and can not be easily adapted to different settings. Here we present a software program that is able to predict solid state NMR spectra of proteins. It can take into account all types of correlations between atoms usually relevant for assignment and structure elucidation and is able to produce lists and visualizations that can be useful when analyzing measured spectra. Compared to other solutions it is fast, versatile and user friendly. Peakr is available through a web site and a web service.

9.2 Background

In recent years, solid-state NMR has made significant progress in structural and functional studies of biomolecules such as membrane proteins and protein fibrils. However, especially in larger proteins, assignment of resonances to atoms and derivation of constraints for 3D structure calculations is still a difficult and time-consuming process due to limited resolution, resonance overlap and chemical shift ambiguity. This especially applies to through-space correlations that can not be traced along bonds. For these tasks, it has proven very helpful to have predictions for spectra based on amino acid

sequences, known or modelled 3D structures, chemical shift assignments or predictions from tools such as ShiftX (167), and the type of correlation probed in the respective experiment. This way, one can make suggestions for crosspeak assignments and, for example, investigate whether a measured spectrum can be explained by one of several alternative structural models. Spectral predictions can also be used in an iterative process of obtaining shift assignments and refining molecular structure at the same time (168).

Existing commercial software packages for the prediction of NMR spectra are usually tailored to calculating 1D spectra of small molecules in solution (169). At present, prediction of solid-state NMR 2D spectra of proteins is usually carried out using general purpose tools like spreadsheet software or custom-made programs limited in flexibility and usability. It would be desirable to be able to predict a wide range of solid-state NMR experiments commonly used for signal assignment and structure elucidation with a single software tool that is flexible enough to swiftly handle changes in input data such as shift values, structural models or labeling patterns. We implemented Peakr, a software package that fulfills those requirements. Spectra for 2D N-C and C-C intra- and inter-residue as well as through-space correlations can be computed quickly and are easily adapted to specific needs. Using the calculated spectra, visual and computational comparisons between predicted and measured data are possible. Peakr is available for installation on a local system, as a website and as a webservice.

9.3 Implementation

Peakr is implemented using a twofold approach. The higher abstracted parts are implemented in the object oriented programming language Ruby. The more data intensive parts and the complex book keeping is done using a PostgreSQL database. Ruby on Rails is used for the web application which drives the web site as well as the web service.

9.4 Concept

In Peakr different concepts are represented as different classes of objects. When using Peakr, the user creates and combines these objects in order to arrive at a virtual experimental setup 9.1.

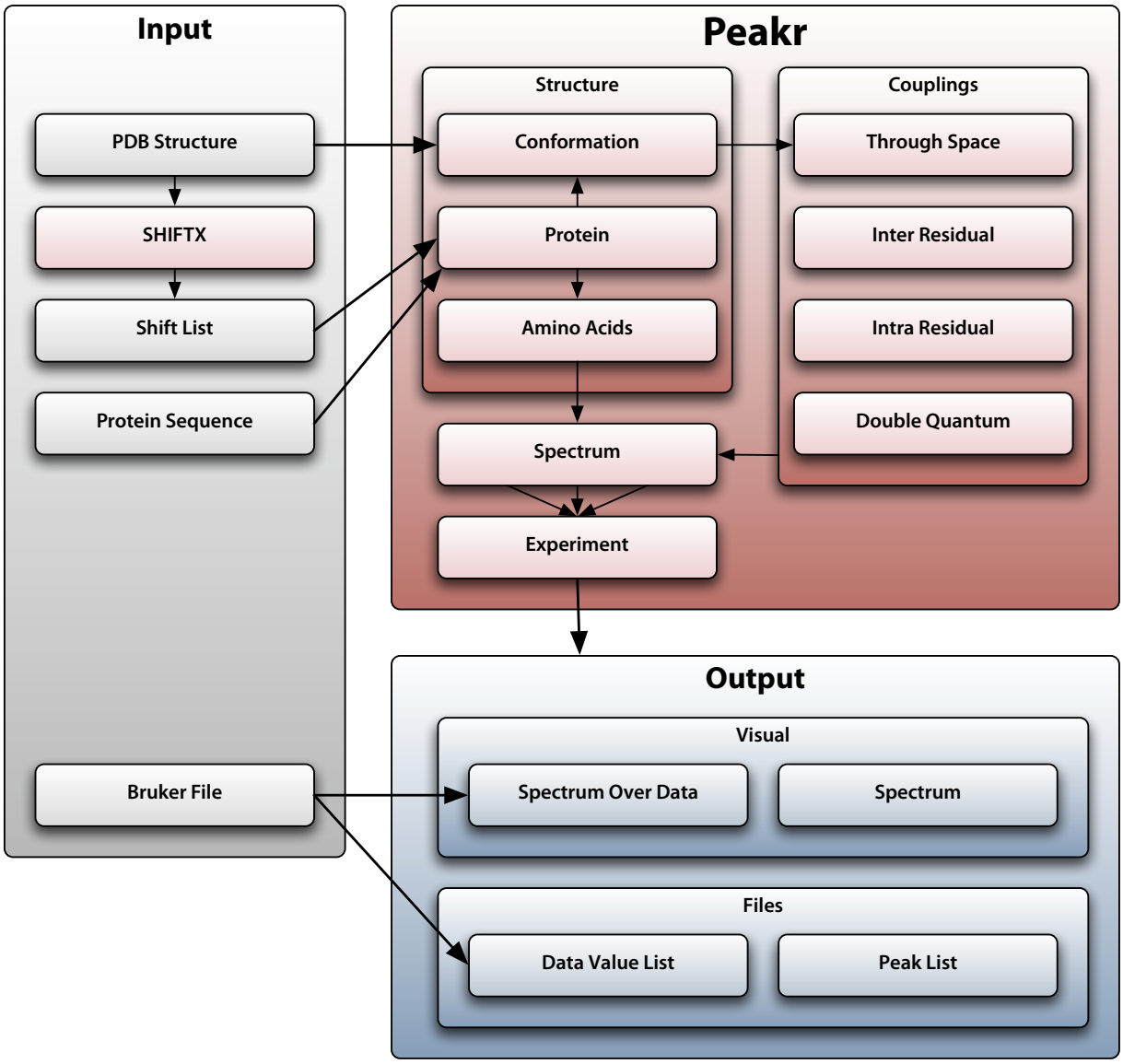

Figure 9.1: The Peakr Workflow
The diagram depicts the data flow of a Peakr run (red) and shows used input data (red) and generated output files (blue).

9.4.1 Proteins

Proteins represent the amino acid sequence, the molecular topology and the chemical shifts of atoms. Each atom can have multiple shifts to account for possible cases of conformational polymorphism sometimes seen in solid-state preparations of proteins.

Protein objects can be created either from a protein sequence or a PDB (170) file (Figure 9.2 A). Chemical shifts are added to the atoms by providing a list of shifts. Formats that can be read are Sparky (171) and ShiftX (167). Also, average chemical shifts as provided by the Biological Magnetic Resonance Database (BMRB) (172) can be assigned to atoms by Peakr. When a protein object is

created from a PDB file, ShiftX is used to predict chemical shifts
It is possible to assign priorities to the different ways of shift determination, for example: "given in user-provided list" > "predicted by ShiftX" > "provided by BMRB".

9.4.2 Conformations

Conformations represent a set of coordinates of the atoms of a given protein. Each protein can have several conformations. This way an ensemble of structures can be represented. Conformations are created from PDB files.

9.4.3 Couplings

Couplings represent the type of experiment conducted to obtain specific inter-atom correlations. They are applied to a set of residues, i.e. a protein sequence, to yield a list of crosspeaks. There are 5 types of Couplings:

Intra-residual C-C couplings

C-C couplings within residues can be defined by specifying the maximum number of bonds the carbons can be away from the backbone and the maximum number of bonds that may be between two carbon atoms that the coupling encompasses. This way, the user can select correlations of interest and avoid overcrowding of the predicted spectrum with peaks that might not be present in a measured spectrum due to, e.g., increased sidechain mobility or short mixing time.

Inter-residual couplings

C-C couplings between residues can be defined by specifying the maximum number of bonds the carbons can be away from the backbone and the maximum range of neighbouring residues that are taken into account (Figure 9.2 B).

N-C couplings

N-C couplings can be defined as intra-residue N(i)-CA(i)-CX(i) or sequential N(i)-CO(i-1)-CX(i-1) correlations (with CX representing any sidechain carbon atom). The maximum number of bonds the carbons can be away from the backbone can be specified (yielding, e.g., only N(i)-CA(i) or also N(i)-CA(i)-CB(i) etc. correlations).

Double quantum couplings

C-C double quantum couplings represent correlations seen in 2D double quantum-single quantum correlation spectra where the shift of a crosspeak in the indirect dimension corresponds to the sum of the chemical shifts of the two interacting atoms. Double quantum couplings are created the same way as intra-residual couplings.

Through space couplings

Through space couplings are created by specifying a set of residues, a set of conformations with atom coordinates and a pairwise distance threshold up to which correlations should be taken into account. This type of coupling can act in different modes, allowing either the direct distances between heteroatoms to be considered (yielding C-C or N-C through-space correlations) or the distances between protons directly attached to heteroatoms (yielding N-H-H-C and C-H-H-C correlations).
In the solid state, crystal packing interactions can lead to crosspeaks arising from intermolecular correlations. In order to distinguish between those and intramolecular crosspeaks, Peakr can selectively predict both types of correlation.

9.4.4 Spectra

Spectra represent a set of crosspeaks that are generated by applying a coupling object to a protein (Figure 9.2 C).

9.4.5 Experiment

Experiments are containers for any number of spectra which can be combined for visual comparison. Or, for example, to yield a C-C spectrum containing intra- and inter-residue correlations that are defined as separate couplings in Peakr but may appear together in an experimentally measured spectrum. An example of spectrum creation is given in Figure 9.2 D.

9.4.6 Measured spectra

Measured spectra can be used to produce overlays with predicted spectra (Figure 9.2 E). Currently, Peakr can read spectra in Bruker format (Figure 9.2 F).

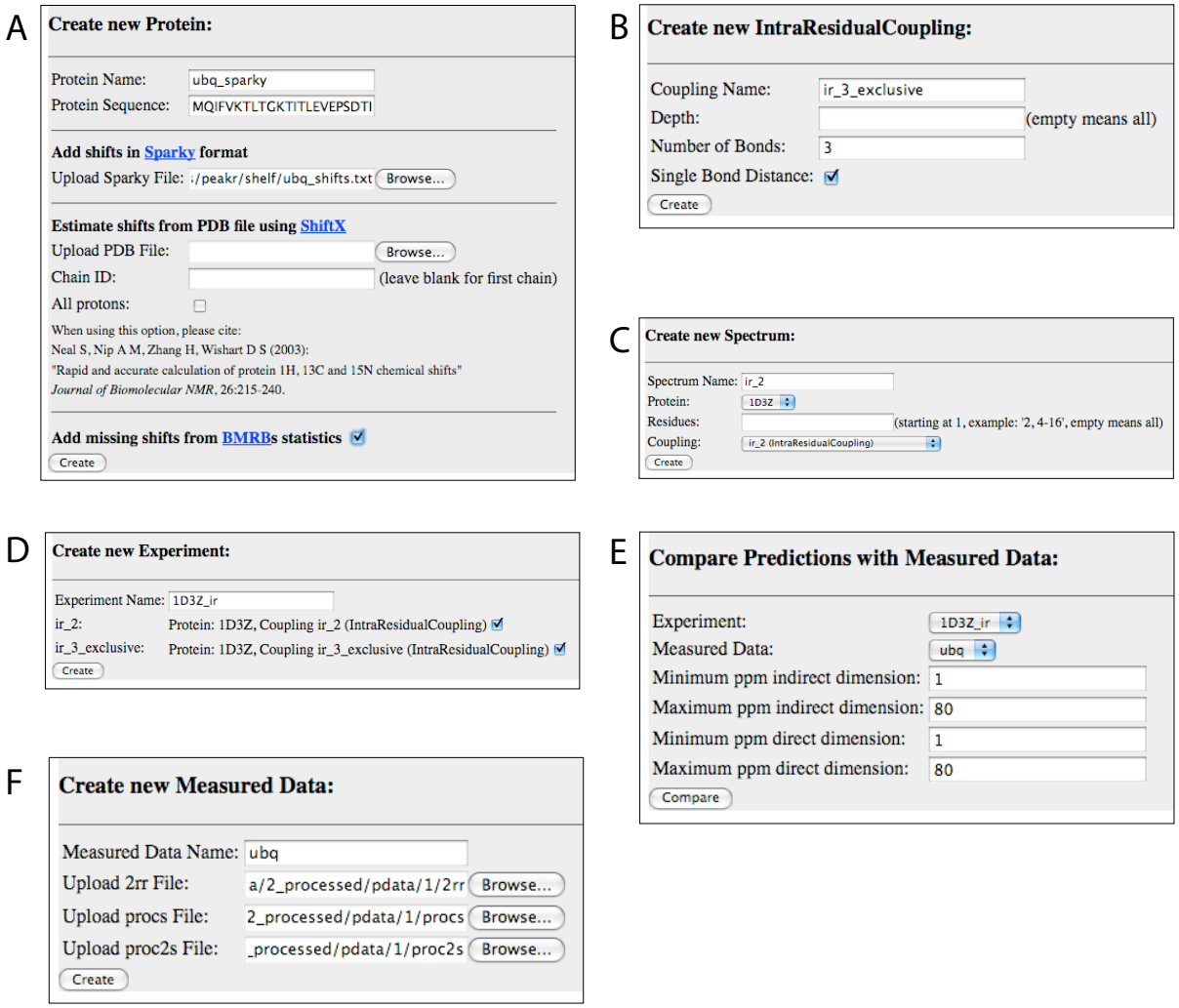

Figure 9.2: Peakr Web Interface
Forms in Peakr's web interface. A: Protein creation. B: Coupling creation. C: Spectrum creation. D: Experiment creation. E: Comparison setup, F: Upload of measured data.

9.4.7 Web Service

Peakr's functionality can be used remotely by other software programs through its web service. Since the protocols (XML-RPC and SOAP) are supported by any modern programming language, the service can be seamlessly integrated. Since the objects that are handled by Peakr can be considerably large, they are not sent over the network. Instead, a checksum is passed to the client program, identifying the object on the server. The only data that is actually sent from the server are results like crosspeak lists or pictures of spectra.

9.4.8 Data Persistence

Every object that is generated through the web interface or the web service is stored on the server and can be referenced by a small checksum string. Using this approach, the user can download a file from the website, containing references to all objects s/he created. This session file can be uploaded later by the same person or can be passed by email or other means. When a session file is uploaded, Peakr re-instantiates all objects and the user can pick up where s/he left off. Since no login is necessary and the checksums can hardly be guessed, the service can be used privately and anonymously.

9.5 Output

9.5.1 Lists

The crosspeak lists that are stored in a spectrum object can be retrieved as tab-delimited files. When comparing with measured data, the intensity of the measured spectrum at the positions of the predicted crosspeaks can be included.

9.5.2 Graphics

Peakr can generate plots with crosspeaks as SVG or PNG. Crosspeaks from different spectra can be combined in one plot and can be distinguished by color. Labels can be plotted into the picture or can be shown as tooltips in SVGs. When plotting crosspeaks from several proteins with identical sequence but different sets of chemical shifts in one plot (arising, e.g., from different conformations), crosspeaks that originate from the same correlation can be connected so the effect of the changed shifts becomes apparent.

9.6 Case study

As an example for using Peakr, we demonstrate the prediction of an intra-residue C-C correlation spectrum of solid ubiquitin and compare it with an experimental spectrum measured on a 700 MHz magnet using the DARR (dipolar assisted rotational resonance) (173) pulse sequence using 7.8 ms mixing time. This measured spectrum should display chiefly intra-residue correlations. Chemical shifts for ubiquitin in the solid state were obtained from (174). We generated predictions for intra-residue C-C correlations with varying numbers of bonds allowed between correlated carbon atoms in order to evaluate the extent to which different correlations (with different numbers of bonds between the carbon atoms) contribute to the experimental spectrum.
Visual inspection (Figure 9.3) shows that almost all experimental peaks are accounted for by the Peakr prediction, confirming the assumption that the measured spectrum displays mostly intra-residue correlations. Some differences between prediction and experimental data are apparent, usually in the form of predicted peaks which are close to measured peaks, but outside of spectral regions above the selected intensity threshold. Such differences can come about if the assigments used for the

prediction were made on a protein sample whose solid-phase preparation method was different from that used for the measured protein (174). For this reason, the percentage of predicted crosspeaks that correspond to a region with measured intensity above the selected threshold is rather low. However, this can be explained by the strict intensity cutoff we have used here and by the fact that we did not allow for chemical shift tolerance. For one-bond correlations, we find 46% of the predicted peaks in regions of intensity above threshold, while two- and three-bond predictions fit with the experimental spectrum in 28% and 18% of all cases, respectively. This is to be expected since the physical couplings between atoms spaced further apart are weaker, leading to the attenuation of signals arising from these correlations.

Based on this comparison, one could, for example, investigate which residues are less well represented in the measured spectrum compared with the prediction, or which residues only exhibit short-range correlations in experimental data. This way, one can identify regions of the protein sequence where elevated molecular mobility might occur (which would attenuate signals from the affected residues in a spectrum based on dipolar transfer such as DARR).

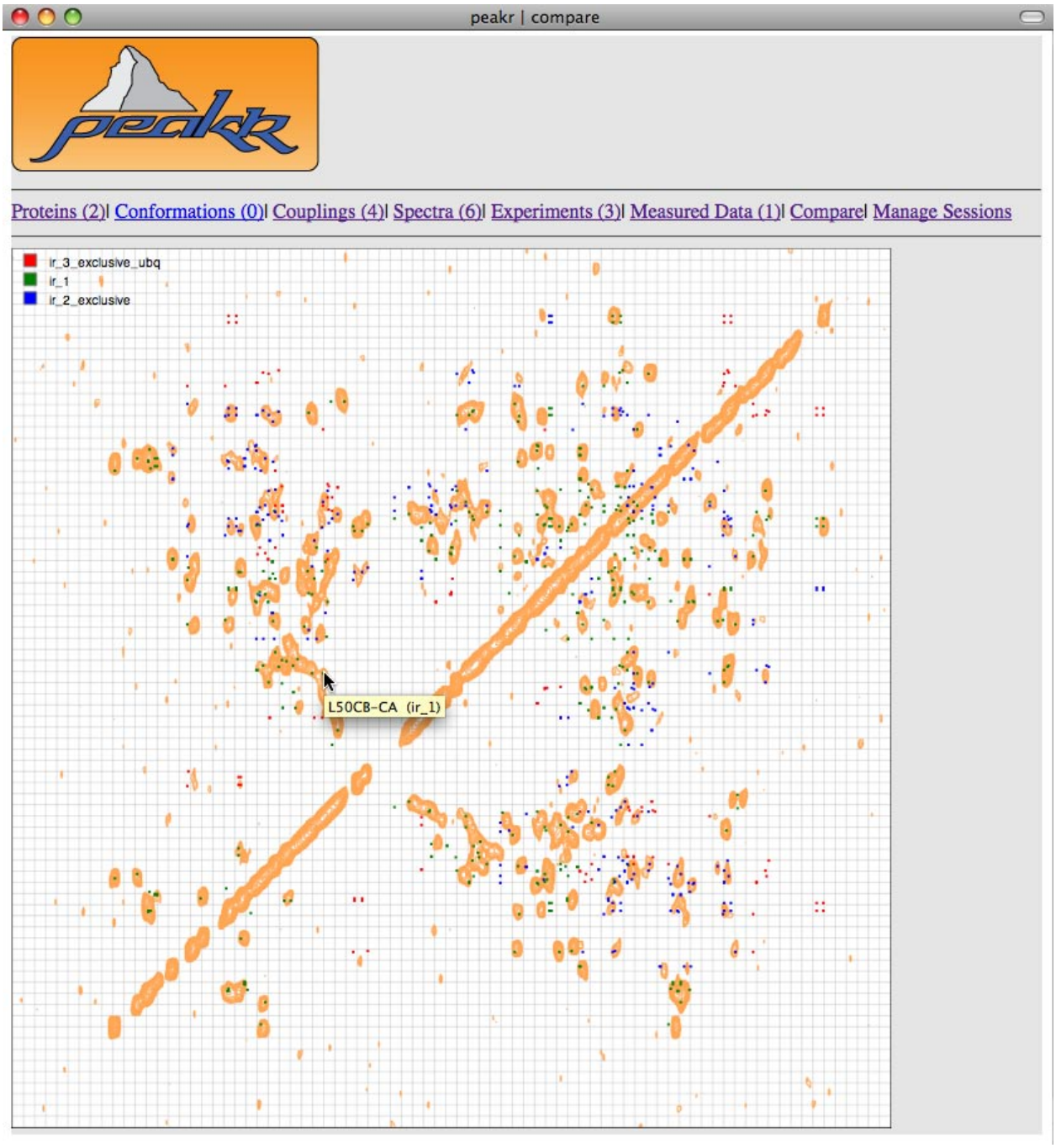

Figure 9.3: Comparison of Predicted and Measured Cross Peaks
This screen shot shows the comparison of a measured spectrum from Ubiquitin with three sets of predicted crosspeaks.
ir_1 : Intra-residual correlation between directly bonded carbons.
ir_2_exclusive: Intra-residual correlation between carbons with a two-bond distance.
ir_3_exclusive_ubq: Intra-residual correlation between carbons with a three-bond distance.

9.7 Discussion and Conclusions

The software program Peakr presented here can be of considerable help when analyzing measured solid state NMR spectra of proteins. It is able to predict spectra for all common experimental setups. The predicted spectra can be helpful when assigning resonances to atoms and when deriving constraints for 3D structure calculations. As demonstrated in the case study, basic assumptions about a measured spectrum can be made in a matter of minutes. In contrast to existing solutions, Peakr is very flexible and uses criteria like residue numbers and amino acid types to define spectra. This is especially valuable when reverse labeling (175) is used or when only a portion of the protein, e.g. the N-terminus, is of interest. The same applies to handling several conformations. With existing solutions, this is either time-consuming and error-prone or impossible. With Peakr, it can be achieved fast and efficiently.
The ability to compare predicted spectra with measured spectra allows to estimate the degree of agreement between the prediction and the measurement. The percentage of predicted crosspeaks with a measured intensity above a given threshold can be seen as a simple figure of merit and can be used to optimize the shifts and/or structure that is used in the prediction. Through the output of tab delimited lists and the availability as a web service, Peakr can easily be integrated into complex analysis pipelines.

9.8 Authors' contributions

RS specified the requirements, FO designed and wrote the software, RS reviewed the code for the core functionality. FO and RS performed tests. Both authors wrote and approved the final manuscript.

9.9 Acknowledgments

The ubiquitin spectrum was kindly provided by Dr. H. Förster (Bruker Biospin, Karlsruhe).

Part V

Conclusions & Acknowledgements

10 Conclusions

With a solid foundation of manually curated data and a growing suite of software tools, we were able to make significant contributions to the field of phylogenetic research. The myosin study reveals a fascinating perspective on how the evolution of organisms is accompanied by a fanning out of a rich molecular diversity and how the succession of evolution can be traced back by looking at the protein repertoire and the sequences. The result is a closely sampled tree of eukaryotic life. Furthermore we were able to greatly extend the existing categorization system of the myosin protein family and are confident that members discovered in the future can be integrated.

While analyzing the gene structure of myosins it became apparent that the Arthropods have used differential splicing as a strategy to greatly increase the diversity of their gene products. Furthermore, the structure of one of three of the genes hints at a peculiar origin, being the reincorporation of a partially processed mRNA into the genome. This sheds new light on the order of steps involved in the process of splicing.

After learning from our daily work when manually annotating and handling protein sequences and related data, we created CyMoBase. Been implemented in this bottom-up fashion, it has been a very helpful tool for our projects but soon grew to a level where it became apparent that it would be useful for the motor protein community as a whole. Since our database also includes information about sequenced genomes that are of interest to a larger audience, we adapted the technological base to create diArk. This web application acts as a source of information about species, finished sequencing projects and related literature, something that surprisingly did not exist before. Both CyMoBase and diArk apply current technologies in unconventional ways in order to provide the user with an intuitive way of searching the diverse content of the database.

The motivation for Scipio was similar to the one for CyMoBase since it grew from the observations how manual annotation was best carried out. This experience was then used to write a software program that is able to do a large part of this tedious work. The result is a tool that answers the simple question which part of the genome encodes for a given protein. Instead of returning a long list of hits, Scipio gives the user one coherent gene structure which is optimized on the level of single base pairs. This kind of response is what we feel most users want.

Having collected a large number of genome files for our own annotation efforts, it was a logical step to make them searchable by others using Scipio. With WebScipio, we offer a web application with which one can search for the gene of a given protein in hundreds of genomes. Combined with the flexible visualization of gene structures, the possibility to download the result files and the accessibility by other software programs, WebScipio is a unique service.

With the experience from the myosin project, a grown number of annotated sequences and some new

approaches to phylogenetic inference, we turned to analyze the phylum Arthropoda. We were able to resolve the phylogeny of 21 species with high confidence and in great detail, providing insights into relations of organisms such as Daphnia, Anopheles, Bombyx and Drosophila. Our findings are in accordance with the results of a high-profile study of the Drosophila genus.
Been thematically unrelated, the Peakr project is an in-house cooperation with another PhD student. We realized that the technology that drives CyMoBase and diArk can be employed to solve common problems in data analysis as carried out in solid state NMR. The result is a software program that can predict spectra of proteins. It removes a bottleneck in the structure elucidating process and is much more user-friendly and flexible than existing solutions.

11 Acknowledgements

First of all, I like to thank Martin Kollmar for his excellent supervision, openness to new ideas and generosity.

I also like to express my gratitude to Prof. Griesinger for continuous support and to Prof. Ficner and Prof. Morgenstern for being members of my thesis committee.

I like to thank Robert Schneider for many enjoyable hours of climbing and programming and Peter Haberz for some good times on the dark side of Göttingen. I also like to thank my collaborators and coworkers Oliver Keller, Holger Pillmann and Marcel Hellkamp for their team spirit. Kudos to Matsumoto-san for creating such a beautiful programming language and many thanks to the myosins for being such a well-behaved protein family.

I might not have turned this way ten years ago without the inspiration from one of my school teachers, Dr. Werner Bils, whom I like to thank for this.

A special thanks goes to my mother and my father for supporting and sponsoring me.
Finally, I wish to thank my girlfriend Moira. Her understanding and sense of humor were a great help.

Part VI

Appendix

A Bibliography

References

[1] Avery OT, MacLeod CM, McCarty M: **Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Inductions of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III.** *J Exp Med* 1979, **149**(2):297–326.

[2] Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM: **Whole-genome random sequencing and assembly of Haemophilus influenzae Rd.** *Science* 1995, **269**(5223):496–512.

[3] Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG: **Life with 6000 genes.** *Science* 1996, **274**(5287):546, 563–7.

[4] **Genome sequence of the nematode C. elegans: a platform for investigating biology.** *Science* 1998, **282**(5396):2012–2018.

[5] Beadle GW, Tatum EL: **Genetic Control of Biochemical Reactions in Neurospora.** *Proc Natl Acad Sci U S A* 1941, **27**(11):499–506.

[6] Jacob F, Monod J: **Genetic regulatory mechanisms in the synthesis of proteins.** *J Mol Biol* 1961, **3**:318–356.

[7] Berget SM, Moore C, Sharp PA: **Spliced segments at the 5' terminus of adenovirus 2 late mRNA.** *Proc Natl Acad Sci U S A* 1977, **74**(8):3171–3175.

[8] R BM, R G: **Recent advances in gene structure prediction.** *Curr Opin Struct Biol* 2004, **14**(3):264–272.

[9] V K, I C, J EA, de Lorenzo V, A OC: **Myriads of protein families, and still counting.** *Genome Biol* 2003, **4**(2):401.

[10] Embley TM, Martin W: **Eukaryotic evolution, changes and challenges**. *Nature* 2006, **440**:623–30.

[11] Krendel M, Mooseker MS: **Myosins: tails (and heads) of functional diversity**. *Physiology (Bethesda)* 2005, **20**:239–51.

[12] Rabi II, Zacharias JR, Millman S, Kusch P: **Milestones in magnetic resonance: 'a new method of measuring nuclear magnetic moment' . 1938.** *J Magn Reson Imaging* 1992, **2**(2):131–133.

[13] Wuthrich K: **Protein structure determination in solution by NMR spectroscopy**. *J. Biol. Chem.* 1990, **265**(36):22059–22062, [http://www.jbc.org/cgi/content/abstract/265/36/22059].

[14] Castellani F, van Rossum B, Diehl A, Schubert M, Rehbein K, Oschkinat H: **Structure of a protein determined by solid-state magic-angle-spinning NMR spectroscopy.** *Nature* 2002, **420**(6911):98–102.

[15] Lange A, Becker S, Seidel K, Giller K, Pongs O, Baldus M: **A concept for rapid protein-structure determination by solid-state NMR spectroscopy.** *Angew Chem Int Ed Engl* 2005, **44**(14):2089–2092.

[16] **GenBank**. *http://www.ncbi.nih.gov/Genbank/index.html* 2006.

[17] Reese MG, Hartzell G, Harris NL, Ohler U, Abril JF, Lewis SE: **Genome annotation assessment in Drosophila melanogaster**. *Genome Res***10**(4):483–501.

[18] Koonin EV: **Orthologs, paralogs, and evolutionary genomics**. *Annu Rev Genet* 2005, **39**:309–38.

[19] Stein L: **Genome annotation: from sequence to biology**. *Nat Rev Genet***2**(7):493–503.

[20] Scholey JM, Brust-Mascher I, Mogilner A: **Cell division**. *Nature* 2003, **422**:746–52.

[21] Vale RD: **The molecular motor toolbox for intracellular transport**. *Cell* 2003, **112**:467–80.

[22] Hirokawa N, Takemura R: **Molecular motors and mechanisms of directional transport in neurons**. *Nat Rev Neurosci***6**(3):201–14.

[23] Geeves MA, Holmes KC: **The molecular mechanism of muscle contraction**. *Adv Protein Chem***71**:161–93.

[24] **PostgreSQL**. *http://www.postgresql.org* 2006.

[25] **Ruby on Rails**. *http://www.rubyonrails.com* 2006.

[26] Fowler M: **Patterns of Enterprise Application Architecture** 2002.

[27] **PostgreSQL Procedure Language**. *http://raa.ruby-lang.org/project/pl-ruby* 2006.

[28] **Distributed Ruby**. *http://raa.ruby-lang.org/project/druby/* 2004.

[29] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res***25**(17):3389–402.

[30] **HMMER: profile HMMs for protein sequence analysis**. *http://hmmer.wustl.edu/* 2003.

[31] **Ruby**. *http://www.ruby-lang.org/* 2005.

[32] **BioRuby**. *http://www.bioruby.org* 2006.

[33] Finn R, Tate J, Mistry J, Coggill P, Sammut S, Hotz H, Ceric G, Forslund K, Eddy S, Sonnhammer E, Bateman A: **The Pfam protein families database.** *Nucleic Acids Res* 2007.

[34] **World Wide Web Consortium**. *http://www.w3c.org* 2006.

[35] Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Helmberg W, Kapustin Y, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Suzek TO, Tatusov R, Tatusova TA, Wagner L, Yaschenko E: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res***34**(Database issue):D173–80.

[36] **iiwi**. *http://www.iiwi.de* 2006.

[37] Odronitz F, Kollmar M: **Pfarao: a web application for protein family analysis customized for cytoskeletal and motor proteins (CyMoBase)**. *BMC Genomics* 2006, **7**:300.

[38] Binnewies TT, Motro Y, Hallin PF, Lund O, Dunn D, La T, Hampson DJ, Bellgard M, Wassenaar TM, Ussery DW: **Ten years of bacterial genome sequencing: comparative-genomics-based discoveries**. *Funct Integr Genomics***6**(3):165–85.

[39] Galagan JE, Henn MR, Ma LJ, Cuomo CA, Birren B: **Genomics of the fungal kingdom: insights into eukaryotic biology**. *Genome Res***15**(12):1620–31.

[40] Surade S, Klein M, Stolt-Bergner PC, Muenke C, Roy A, Michel H: **Comparative analysis and "expression space" coverage of the production of prokaryotic membrane proteins for structural genomics**. *Protein Sci***15**(9):2178–89.

[41] Snel B, Huynen MA, Dutilh BE: **Genome trees and the nature of genome evolution**. *Annu Rev Microbiol***59**:191–209.

[42] Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides NC: **The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide**. *Nucleic Acids Res***34**(Database issue):D332–4.

[43] **International Sequencing Consortium**.

[44] Yang S, Doolittle RF, Bourne PE: **Phylogeny determined by protein domain content**. *Proc Natl Acad Sci U S A* 2005, **102**:373–8.

[45] Doolittle RF: **Evolutionary aspects of whole-genome biology**. *Curr Opin Struct Biol* 2005, **15**:248–53.

[46] Jeffroy O, Brinkmann H, Delsuc F, Philippe H: **Phylogenomics: the beginning of incongruence?** *Trends Genet* 2006, **22**:225–31.

[47] Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P: **Toward automatic reconstruction of a highly resolved tree of life**. *Science* 2006, **311**:1283–7.

[48] Delsuc F, Brinkmann H, Philippe H: **Phylogenomics and the reconstruction of the tree of life**. *Nat Rev Genet* 2005, **6**:361–75.

[49] Odronitz F, Hellkamp M, Kollmar M: **diArk–a resource for eukaryotic genome research.** *BMC Genomics* 2007, **8**:103.

[50] Schliwa M, Woehlke G: **Molecular motors**. *Nature* 2003, **422**:759–65.

[51] Yumura S, Uyeda TQ: **Myosins and cell dynamics in cellular slime molds**. *Int Rev Cytol* 2003, **224**:173–225.

[52] Geeves MA, Holmes KC: **The molecular mechanism of muscle contraction**. *Adv Protein Chem* 2005, **71**:161–93.

[53] Cheney RE, Riley MA, Mooseker MS: **Phylogenetic analysis of the myosin superfamily**. *Cell Motil Cytoskeleton* 1993, **24**:215–23.

[54] Foth BJ, Goedecke MC, Soldati D: **New insights into myosin evolution and classification**. *Proc Natl Acad Sci U S A* 2006, **103**:3681–6.

[55] Richards TA, Cavalier-Smith T: **Myosin domain evolution and the primary divergence of eukaryotes**. *Nature* 2005, **436**:1113–8.

[56] Berg JS, Powell BC, Cheney RE: **A millennial myosin census**. *Mol Biol Cell* 2001, **12**:780–94.

[57] Gillespie PG, Albanesi JP, Bahler M, Bement WM, Berg JS, Burgess DR, Burnside B, Cheney RE, Corey DP, Coudrier E, de Lanerolle P, Hammer JA, Hasson T, Holt JR, Hudspeth AJ, Ikebe M, Kendrick-Jones J, Korn ED, Li R, Mercer JA, Milligan RA, Mooseker MS, Ostap EM, Petit C, Pollard TD, Sellers JR, Soldati T, Titus MA: **Myosin-I nomenclature**. *J Cell Biol* 2001, **155**:703–4.

[58] Hodge T, Cope MJ: **A myosin family tree**. *J Cell Sci* 2000, **113 Pt 19**:3353–4.

[59] Williams SA, Gavin RH: **Myosin genes in Tetrahymena**. *Cell Motil Cytoskeleton* 2005, **61**:237–43.

[60] Heintzelman MB, Schwartzman JD: **Myosin diversity in Apicomplexa**. *J Parasitol* 2001, **87**:429–32.

[61] Brown SS: **Myosins in yeast**. *Curr Opin Cell Biol* 1997, **9**:44–8.

[62] Kollmar M: **Thirteen is enough: the myosins of Dictyostelium discoideum and their light chains**. *BMC Genomics* 2006, **7**:183.

[63] Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, Nicaud S, Jaffe D, Fisher S, Lutfalla G, Dossat C, Segurens B, Dasilva C, Salanoubat M, Levy M, Boudet N, Castellano S, Anthouard V, Jubin C, Castelli V, Katinka M, Vacherie B, Biemont C, Skalli Z, Cattolico L, Poulain J, De Berardinis V, Cruaud C, Duprat S, Brottier P, Coutanceau JP, Gouzy J, Parra G, Lardier G, Chapple C, McKernan KJ, McEwan P, Bosak S, Kellis M, Volff JN, Guigo R, Zody MC, Mesirov J, Lindblad-Toh K, Birren B, Nusbaum C, Kahn D, Robinson-Rechavi M, Laudet V, Schachter V, Quetier F, Saurin W, Scarpelli C, Wincker P, Lander ES, Weissenbach J, Roest Crollius H: **Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype**. *Nature* 2004, **431**:946–57.

[64] Fujiwara M, Horiuchi H, Ohta A, Takagi M: **A novel fungal gene encoding chitin synthase with a myosin motor-like domain**. *Biochem Biophys Res Commun* 1997, **236**:75–8.

[65] Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, Terry A, Shapiro H, Lindquist E, Kapitonov VV, Jurka J, Genikhovich G, Grigoriev IV, Lucas SM, Steele RE, Finnerty JR, Technau U, Martindale MQ, Rokhsar DS: **Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization**. *Science* 2007, **317**:86–94.

[66] Adoutte A, Balavoine G, Lartillot N, Lespinet O, Prud'homme B, de Rosa R: **The new animal phylogeny: reliability and implications**. *Proc Natl Acad Sci U S A* 2000, **97**:4453–6.

[67] Langkjaer RB, Cliften PF, Johnston M, Piskur J: **Yeast genome duplication was followed by asynchronous differentiation of duplicated genes**. *Nature* 2003, **421**:848–52.

[68] Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD: **Multiple sequence alignment with the Clustal series of programs**. *Nucleic Acids Res* 2003, **31**:3497–500.

[69] Chevenet F, Brun C, Banuls AL, Jacq B, Christen R: **TreeDyn: towards dynamic graphics and annotations for analyses of trees**. *BMC Bioinformatics* 2006, **7**:439.

[70] Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks**. *Proc Natl Acad Sci U S A* 1992, **89**:10915–9.

[71] Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP, Bork P: **SMART 4.0: towards genomic data integration**. *Nucleic Acids Res* 2004, **32**:D142–4.

[72] **Simple Modular Architecture Research Tool**[`http://smart.embl-heidelberg.de/`].

[73] Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR: **The Pfam protein families database**. *Nucleic Acids Res* 2004, **32**:D138–41.

[74] Graveley BR: **Alternative splicing: increasing diversity in the proteomic world**. *Trends Genet* 2001, **17**:100–7.

[75] Black DL: **Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology**. *Cell* 2000, **103**:367–70.

[76] Thanaraj TA, Stamm S, Clark F, Riethoven JJ, Le Texier V, Muilu J: **ASD: the Alternative Splicing Database**. *Nucleic Acids Res* 2004, **32**:D64–9.

[77] Kondrashov FA, Koonin EV: **Origin of alternative splicing by tandem exon duplication**. *Hum Mol Genet* 2001, **10**:2661–9.

[78] Anastassiou D, Liu H, Varadan V: **Variable window binding for mutually exclusive alternative splicing**. *Genome Biol* 2006, **7**:R2.

[79] Graveley BR: **Mutually exclusive splicing of the insect Dscam pre-mRNA directed by competing intronic RNA secondary structures**. *Cell* 2005, **123**:65–73.

[80] Letunic I, Copley RR, Bork P: **Common exon duplication in animals and its role in alternative splicing**. *Hum Mol Genet* 2002, **11**:1561–7.

[81] Graveley BR, Kaur A, Gunning D, Zipursky SL, Rowen L, Clemens JC: **The organization and evolution of the dipteran and hymenopteran Down syndrome cell adhesion molecule (Dscam) genes**. *Rna* 2004, **10**:1499–506.

[82] George EL, Ober MB, Emerson CPJ: **Functional domains of the Drosophila melanogaster muscle myosin heavy-chain gene are encoded by alternatively spliced exons**. *Mol Cell Biol* 1989, **9**:2957–74.

[83] Odronitz 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**(9):R196.

[84] Oliver TN, Berg JS, Cheney RE: **Tails of unconventional myosins**. *Cell Mol Life Sci* 1999, **56**:243–57.

[85] Holmes KC: **Introduction**. *Philos Trans R Soc Lond B Biol Sci* 2004, **359**:1813–8.

[86] Yamashita RA, Sellers JR, Anderson JB: **Identification and analysis of the myosin superfamily in Drosophila: a database approach**. *J Muscle Res Cell Motil* 2000, **21**:491–505.

[87] Chiba S, Awazu S, Itoh M, Chin-Bow ST, Satoh N, Satou Y, Hastings KE: **A genomewide survey of developmentally relevant genes in Ciona intestinalis. IX. Genes for muscle structural proteins**. *Dev Genes Evol* 2003, **213**:291–302.

[88] Zhang S, Bernstein SI: **Spatially and temporally regulated expression of myosin heavy chain alternative exons during Drosophila embryogenesis**. *Mech Dev* 2001, **101**:35–45.

[89] Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA, Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, McPherson D, Merkulov G, Milshina NV, Mobarry C, Morris J, Moshrefi A, Mount SM, Moy M, Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RD, Scheeler F, Shen H, Shue BC, Siden-Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM, Weissenbach J, Williams SM, WoodageT, Worley KC, Wu D, Yang S, Yao QA, Ye J, Yeh RF, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO, Gibbs RA, Myers EW, Rubin GM, Venter JC: **The genome sequence of Drosophila melanogaster**. *Science* 2000, **287**:2185–95.

[90] Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JM, Wides R, Salzberg SL, Loftus B, Yandell M, Majoros WH, Rusch DB, Lai Z, Kraft CL, Abril JF, Anthouard V, Arensburger P, Atkinson PW, Baden H, de Berardinis V, Baldwin D, Benes V, Biedler J, Blass C, Bolanos R, Boscus D, Barnstead M, Cai S, Center A, Chaturverdi K, Christophides GK, Chrystal MA, Clamp M, Cravchik A, Curwen V, Dana A, Delcher A, Dew I, Evans CA, Flanigan M, Grundschober-Freimoser A, Friedli L, Gu Z, Guan P, Guigo R, Hillenmeyer ME, Hladun SL, Hogan JR, Hong YS, Hoover J, Jaillon O, Ke Z, Kodira C, Kokoza E, Koutsos A, Letunic I, Levitsky A, Liang Y, Lin JJ, Lobo NF, Lopez JR, Malek JA, McIntosh TC, Meister S, Miller J, Mobarry C, Mongin E, Murphy SD, O'Brochta DA, Pfannkoch C, Qi R, Regier MA, Remington K, Shao H, Sharakhova MV, Sitter CD, Shetty J, Smith TJ, Strong R, Sun J, Thomasova D, Ton LQ, Topalis P, Tu Z, Unger MF, Walenz B, Wang A, Wang J, Wang M, Wang X, Woodford KJ, Wortman JR, Wu M, Yao A, Zdobnov EM, Zhang H, Zhao Q, Zhao S, Zhu SC, Zhimulev I, Coluzzi M, della Torre A, Roth CW, Louis C, Kalush F, Mural RJ, Myers EW, Adams MD, Smith HO, Broder S, Gardner MJ, Fraser CM, Birney E, Bork P, Brey PT, Venter JC, Weissenbach J, Kafatos FC, Collins FH, Hoffman SL: **The genome sequence of the malaria mosquito Anopheles gambiae**. *Science* 2002, **298**:129–49.

[91] Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ, Loftus B, Xi Z, Megy K, Grabherr M, Ren Q, Zdobnov EM, Lobo NF, Campbell KS, Brown SE, Bonaldo MF, Zhu J, Sinkins SP, Hogenkamp DG, Amedeo P, Arensburger P, Atkinson PW, Bidwell S, Biedler J, Birney E, Bruggner RV, Costas J, Coy MR, Crabtree J, Crawford M, Debruyn B, Decaprio D, Eiglmeier K, Eisenstadt E, El-Dorry H, Gelbart

WM, Gomes SL, Hammond M, Hannick LI, Hogan JR, Holmes MH, Jaffe D, Johnston JS, Kennedy RC, Koo H, Kravitz S, Kriventseva EV, Kulp D, Labutti K, Lee E, Li S, Lovin DD, Mao C, Mauceli E, Menck CF, Miller JR, Montgomery P, Mori A, Nascimento AL, Naveira HF, Nusbaum C, O'Leary S, Orvis J, Pertea M, Quesneville H, Reidenbach KR, Rogers YH, Roth CW, Schneider JR, Schatz M, Shumway M, Stanke M, Stinson EO, Tubio JM, Vanzee JP, Verjovski-Almeida S, Werner D, White O, Wyder S, Zeng Q, Zhao Q, Zhao Y, Hill CA, Raikhel AS, Soares MB, Knudson DL, Lee NH, Galagan J, Salzberg SL, Paulsen IT, Dimopoulos G, Collins FH, Birren B, Fraser-Liggett CM, Severson DW: **Genome sequence of Aedes aegypti, a major arbovirus vector**. *Science* 2007, **316**:1718–23.

[92] Mita K, Kasahara M, Sasaki S, Nagayasu Y, Yamada T, Kanamori H, Namiki N, Kitagawa M, Yamashita H, Yasukochi Y, Kadono-Okuda K, Yamamoto K, Ajimura M, Ravikumar G, Shimomura M, Nagamura Y, Shin IT, Abe H, Shimada T, Morishita S, Sasaki T: **The genome sequence of silkworm, Bombyx mori**. *DNA Res* 2004, **11**:27–35.

[93] Clark A, Gibson G, Kaufman T, Myers E, O'Grady P: **Proposal for Drosophila as a Model System for Comparative Genomics.** Tech. rep. 2003.

[94] D'Errico I, Gadaleta G, Saccone C: **Pseudogenes in metazoa: origin and features.** *Brief Funct Genomic Proteomic* 2004, **3**(2):157–167.

[95] Balakirev ES, Ayala FJ: **Pseudogenes: are they "junk" or functional DNA?** *Annual review of genetics* 2003, **37**:123–51.

[96] Proudfoot N: **Pseudogenes**. *Nature* 1980, **286**:840–1.

[97] Dhawan P, Yang E, Kumar A, Mehta KD: **Genetic complexity of the human geranylgeranyltransferase I beta-subunit gene: a multigene family of pseudogenes derived from mis-spliced transcripts**. *Gene* 1998, **210**:9–15.

[98] Suzuki E, Lowry J, Sonoda G, Testa JR, Walsh K: **Structures and chromosome locations of the human MEF2A gene and a pseudogene MEF2AP**. *Cytogenetics and cell genetics* 1996, **73**:244–9.

[99] Breathnach R, Chambon P: **Organization and expression of eucaryotic split genes coding for proteins**. *Annu Rev Biochem* 1981, **50**:349–83.

[100] Rozek CE, Davidson N: **Differential processing of RNA transcribed from the single-copy Drosophila myosin heavy chain gene produces four mRNAs that encode two polypeptides**. *Proc Natl Acad Sci U S A* 1986, **83**:2128–32.

[101] Bernstein SI, Hansen CJ, Becker KD, Wassenberg DRn, Roche ES, Donady JJ, Emerson CPJ: **Alternative RNA splicing generates transcripts encoding a thorax-specific isoform of Drosophila melanogaster myosin heavy chain**. *Mol Cell Biol* 1986, **6**:2511–9.

[102] Raible F, Tessmar-Raible K, Osoegawa K, Wincker P, Jubin C, Balavoine G, Ferrier D, Benes V, de Jong P, Weissenbach J, Bork P, Arendt D: **Vertebrate-type intron-rich genes in the marine annelid Platynereis dumerilii**. *Science* 2005, **310**:1325–6.

[103] Bernstein SI, Milligan RA: **Fine tuning a molecular motor: the location of alternative domains in the Drosophila myosin head**. *J Mol Biol* 1997, **271**:1–6.

[104] Sweeney HL, Rosenfeld SS, Brown F, Faust L, Smith J, Xing J, Stein LA, Sellers JR: **Kinetic tuning of myosin via a flexible loop adjacent to the nucleotide binding pocket**. *J Biol Chem* 1998, **273**:6262–70.

[105] Kollmar M, Durrwang U, Kliche W, Manstein DJ, Kull FJ: **Crystal structure of the motor domain of a class-I myosin**. *Embo J* 2002, **21**:2517–25.

[106] Uyeda TQ, Ruppel KM, Spudich JA: **Enzymatic activities correlate with chimaeric substitutions at the actin-binding face of myosin**. *Nature* 1994, **368**:567–9.

[107] Joel PB, Trybus KM, Sweeney HL: **Two conserved lysines at the 50/20-kDa junction of myosin are necessary for triggering actin activation**. *J Biol Chem* 2001, **276**:2998–3003.

[108] Furch M, Geeves MA, Manstein DJ: **Modulation of actin affinity and actomyosin adenosine triphosphatase by charge changes in the myosin motor domain**. *Biochemistry* 1998, **37**:6317–26.

[109] Littlefield KP, Swank DM, Sanchez BM, Knowles AF, Warshaw DM, Bernstein SI: **The converter domain modulates kinetic properties of Drosophila myosin**. *Am J Physiol Cell Physiol* 2003, **284**:C1031–8.

[110] Swank DM, Knowles AF, Suggs JA, Sarsoza F, Lee A, Maughan DW, Bernstein SI: **The myosin converter domain modulates muscle performance**. *Nat Cell Biol* 2002, **4**:312–6.

[111] **NCBI BLAST with arthropoda genomes**[http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi?organism=insects].

[112] Kent WJ: **BLAT–the BLAST-like alignment tool**. *Genome Res* 2002, **12**:656–64.

[113] Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, Hillman-Jackson J, Kuhn RM, Pedersen JS, Pohl A, Raney BJ, Rosenbloom KR, Siepel A, Smith KE, Sugnet CW, Sultan-Qurraie A, Thomas DJ, Trumbower H, Weber RJ, Weirauch M, Zweig AS, Haussler D, Kent WJ: **The UCSC Genome Browser Database: update 2006**. *Nucleic Acids Res* 2006, **34**:D590–8.

[114] **UCSC Genome Bioinformatics**[http://genome.cse.ucsc.edu/].

[115] **DOE Joint Genome Institute**[http://www.jgi.doe.gov/].

[116] **Daphnia Genomics Consortium**[http://daphnia.cgb.indiana.edu/wfleabase/].

[117] **Human Genome Sequencing Center at Baylor College of Medicine**[http://www.hgsc.bcm.tmc.edu/projects/nasonia/].

[118] Page RD: **TreeView: an application to display phylogenetic trees on personal computers**. *Comput Appl Biosci* 1996, **12**:357–8.

[119] Fedorova L, Fedorov A: **Introns in gene evolution**. *Genetica* 2003, **118**(2-3):123–31.

[120] Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y, Hume DA: **Mammalian RNA polymerase II core promoters: insights from genome-wide studies**. *Nature reviews* 2007, **8**(6):424–36.

[121] Irimia M, Rukov J, Penny D, Roy S: **Functional and evolutionary analysis of alternatively spliced genes is consistent with an early eukaryotic origin of alternative splicing.** *BMC Evol Biol* 2007, **7**:188, [http://dx.doi.org/10.1186/1471-2148-7-188].

[122] Lazzarato F, Franceschinis G, Botta M, Cordero F, Calogero RA: **RRE: a tool for the extraction of non-coding regions surrounding annotated genes from genomic datasets**. *Bioinformatics (Oxford, England)* 2004, **20**(16):2848–50.

[123] Doh ST, Zhang Y, Temple MH, Cai L: **Non-coding sequence retrieval system for comparative genomic analysis of gene regulatory elements**. *BMC bioinformatics* 2007, **8**:94.

[124] Kapustin Y, Souvorov A, Tatusova T: **Splign - a Hybrid Approach To Spliced Alignments**. In *RECOMB 2004 – Currents in Computational Biology* 2004:741.

[125] van Nimwegen E, Paul N, Sheridan R, Zavolan M: **SPA: a probabilistic algorithm for spliced alignment**. *PLoS Genet* 2006, **2**(4), [http://www.hubmed.org/display.cgi?uids=16683023].

[126] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**(3):403–410.

[127] Gelfand MS, Mironov AA, Pevzner PA: **Gene recognition via spliced sequence alignment**. *Proc Natl Acad Sci U S A* 1996, **93**(17):9061–9066, [http://www.hubmed.org/display.cgi?uids=8799154].

[128] Gerhard DS, Wagner L, Feingold EA, Shenmen CM, Grouse LH, Schuler G, Klein SL, Old S, Rasooly R, Good P, Guyer M, Peck AM, Derge JG, Lipman D, Collins FS, Jang W, Sherry S, Feolo M, Misquitta L, Lee E, Rotmistrovsky K, Greenhut SF, Schaefer CF, Buetow K, Bonner TI, Haussler D, Kent J, Kiekhaus M, Furey T, Brent M, Prange C, Schreiber K, Shapiro N, Bhat NK, Hopkins RF, Hsie F, Driscoll T, Soares MB, Casavant TL, Scheetz TE, Brown-stein MJ, Usdin TB, Toshiyuki S, Carninci P, Piao Y, Dudekula DB, Ko MSH, Kawakami K, Suzuki Y, Sugano S, Gruber CE, Smith MR, Simmons B, Moore T, Waterman R, Johnson SL, Ruan Y, Wei CL, Mathavan S, Gunaratne PH, Wu J, Garcia AM, Hulyk SW, Fuh E, Yuan Y, Sneed A, Kowis C, Hodgson A, Muzny DM, McPherson J, Gibbs RA, Fahey J, Helton E, Ketteman M, Madan A, Rodrigues S, Sanchez A, Whiting M, Madari A, Young AC, Wetherby KD, Granite SJ, Kwong PN, Brinkley CP, Pearson RL, Bouffard GG, Blakesly RW, Green ED, Dickson MC, Rodriguez AC, Grimwood J, Schmutz J, Myers RM, Butterfield YSN, Griffith M, Griffith OL, Krzywinski MI, Liao N, Morin R, Palmquist D, Petrescu AS, Skalska U, Smailus DE, Stott JM, Schnerch A, Schein JE, Jones SJM, Holt RA, Baross A, Marra MA, Clifton S, Makowski KA, Bosak S, Malek J: **The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC).** *Genome Res* 2004, **14**(10B):2121–2127.

[129] Ota T, Suzuki Y, Nishikawa T, Otsuki T, Sugiyama T, Irie R, Wakamatsu A, Hayashi K, Sato H, Nagai K, Kimura K, Makita H, Sekine M, Obayashi M, Nishi T, Shibahara T, Tanaka T, Ishii S, Yamamoto Ji, Saito K, Kawai Y, Isono Y, Nakamura Y, Nagahari K, Murakami K, Yasuda T, Iwayanagi T, Wagatsuma M, Shiratori A, Sudo H, Hosoiri T, Kaku Y, Kodaira H, Kondo H, Sugawara M, Takahashi M, Kanda K, Yokoi T, Furuya T, Kikkawa E, Omura Y, Abe K, Kamihara K, Katsuta N, Sato K, Tanikawa M, Yamazaki M, Ninomiya K, Ishibashi T, Yamashita H, Murakawa K, Fujimori K, Tanai H, Kimata M, Watanabe M, Hiraoka S, Chiba Y, Ishida S, Ono Y, Takiguchi S, Watanabe S, Yosida M, Hotuta T, Kusano J, Kanehori K, Takahashi-Fujii A, Hara H, Tanase To, Nomura Y, Togiya S, Komai F, Hara R, Takeuchi K, Arita M, Imose N, Musashino K, Yuuki H, Oshima A, Sasaki N, Aotsuka S, Yoshikawa Y, Matsunawa H, Ichihara T, Shiohata N, Sano S, Moriya S, Momiyama H, Satoh N, Takami S, Terashima Y, Suzuki O, Nakagawa S, Senoh A, Mizoguchi H, Goto Y, Shimizu F, Wakebe H, Hishigaki H, Watanabe T, Sugiyama A, Takemoto M, Kawakami B, Yamazaki M, Watanabe K, Kumagai A, Itakura S, Fukuzumi Y, Fujimori Y, Komiyama M, Tashiro H, Tanigami A, Fujiwara T, Ono T, Yamada K, Fujii Y, Ozaki K, Hirao M, Ohmori Y, Kawabata A, Hikiji T, Kobatake N, Inagaki H, Ikema Y, Okamoto S, Okitani R, Kawakami T, Noguchi S, Itoh T, Shigeta K, Senba T, Matsumura K, Nakajima Y, Mizuno T, Morinaga M, Sasaki M, Togashi T, Oyama M, Hata H, Watanabe M, Komatsu T, Mizushima-Sugano J, Satoh T, Shirai Y, Takahashi Y, Nakagawa K, Okumura K, Nagase T, Nomura N, Kikuchi H, Masuho Y, Yamashita R, Nakai K, Yada T, Nakamura Y, Ohara O, Isogai T, Sugano S: **Complete sequencing and characterization of 21,243 full-length human cDNAs.** *Nat Genet* 2004, **36**:40–45.

[130] **Genome sequencing centre at the Washington University School of Medicine**. *http://genome.wustl.edu* 2007.

[131] Keller O, Odronitz F, Stanke M, Kollmar M, Waack S: **Scipio: Using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species** 2007.

[132] Dubchak I, Frazer K: **Multi-species sequence comparison: the next frontier in genome annotation**. *Genome biology* 2003, **4**(12):122.

[133] Bird CP, Stranger BE, Dermitzakis ET: **Functional variation and evolution of non-coding DNA**. *Current opinion in genetics I& development* 2006, **16**(6):559–64.

[134] Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, Giresi PG, Goldy J, Hawrylycz M, Haydock A, Humbert R, James KD, Johnson BE, Johnson EM, Frum TT, Rosenzweig ER, Karnani N, Lee K, Lefebvre GC, Navas PA, Neri F, Parker SC, Sabo PJ, Sandstrom R, Shafer A, Vetrie D, Weaver M, Wilcox S, Yu M, Collins FS, Dekker J, Lieb JD, Tullius TD, Crawford GE, Sunyaev S, Noble WS, Dunham I, Denoeud F, Reymond A, Kapranov P, Rozowsky J, Zheng D, Castelo R, Frankish A, Harrow

J, Ghosh S, Sandelin A, Hofacker IL, Baertsch R, Keefe D, Dike S, Cheng J, Hirsch HA, Sekinger EA, Lagarde J, Abril JF, Shahab A, Flamm C, Fried C, Hackermuller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, Washietl S, Korbel J, Emanuelsson O, Pedersen JS, Holroyd N, Taylor R, Swarbreck D, Matthews N, Dickson MC, Thomas DJ, Weirauch MT, Gilbert J, et al.: **Identification and analysis of functional elements in 1project**. *Nature* 2007, **447**(7146):799–816.

[135] Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, et al.: **Initial sequencing and comparative analysis of the mouse genome**. *Nature* 2002, **420**(6915):520–62.

[136] Fischer DF, Backendorf C: **Identification of regulatory elements by gene family footprinting and in vivo analysis**. *Advances in biochemical engineering/biotechnology* 2007, **104**:37–64.

[137] Guigo R, Dermitzakis ET, Agarwal P, Ponting CP, Parra G, Reymond A, Abril JF, Keibler E, Lyle R, Ucla C, Antonarakis SE, Brent MR: **Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(3):1140–5.

[138] Ner-Gaon H, Leviatan N, Rubin E, Fluhr R: **Comparative cross-species alternative splicing in plants**. *Plant physiology* 2007, **144**(3):1632–41.

[139] Ureta-Vidal A, Ettwiller L, Birney E: **Comparative genomics: genome-wide analysis in metazoan eukaryotes**. *Nature reviews* 2003, **4**(4):251–62.

[140] Kuhn RM, Karolchik D, Zweig AS, Trumbower H, Thomas DJ, Thakkapallayil A, Sugnet CW, Stanke M, Smith KE, Siepel A, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pedersen JS, Hsu F, Hinrichs AS, Harte RA, Diekhans M, Clawson H, Bejerano G, Barber GP, Baertsch R, Haussler D, Kent WJ: **The UCSC genome browser database: update 2007**. *Nucleic acids research* 2007, **35**(Database issue):D668–73.

[141] Elnitski LL, Shah P, Moreland RT, Umayam L, Wolfsberg TG, Baxevanis AD: **The ENCODEdb portal: simplified access to ENCODE Consortium data**. *Genome research* 2007, **17**(6):954–9.

[142] Pillmann H: **WebScipio: Implementierung eines Webservice und eines Webinterfaces zur Exonsuche in eukaryotischen Genomen** 2007, [`http://www.swe.informatik.uni-goettingen.de/pubs/index.php?lang=en`].

[143] Bezanilla M, Horton AC, Sevener HC, Quatrano RS: **Phylogenetic analysis of new plant myosin sequences.** *J Mol Evol* 2003, **57**(2):229–239.

[144] Reck-Peterson SL, Yildiz A, Carter AP, Gennerich A, Zhang N, Vale RD: **Single-molecule analysis of dynein processivity and stepping behavior.** *Cell* 2006, **126**(2):335–348.

[145] Hirokawa N: **Kinesin and dynein superfamily proteins and the mechanism of organelle transport.** *Science* 1998, **279**(5350):519–526.

[146] Lawrence CJ, Dawe RK, Christie KR, Cleveland DW, Dawson SC, Endow SA, Goldstein LSB, Goodson HV, Hirokawa N, Howard J, Malmberg RL, McIntosh JR, Miki H, Mitchison TJ, Okada Y, Reddy ASN, Saxton WM, Schliwa M, Scholey JM, Vale RD, Walczak CE, Wordeman L: **A standardized kinesin nomenclature.** *J Cell Biol* 2004, **167**:19–22.

[147] Alphey L, Parker L, Hawcroft G, Guo Y, Kaiser K, Morgan G: **KLP38B: a mitotic kinesin-related protein that binds PP1.** *J Cell Biol* 1997, **138**(2):395–409.

[148] Schoch CL, Aist JR, Yoder OC, Gillian Turgeon B: **A complete inventory of fungal kinesins in representative filamentous ascomycetes.** *Fungal Genet Biol* 2003, **39**:1–15.

[149] Reddy AS, Day IS: **Analysis of the myosins encoded in the recently completed Arabidopsis thaliana genome sequence.** *Genome Biol* 2001, **2**(7):RESEARCH0024.

[150] Leopold PL, McDowall AW, Pfister KK, Bloom GS, Brady ST: **Association of kinesin with characterized membrane-bounded organelles.** *Cell Motil Cytoskeleton* 1992, **23**:19–33.

[151] King SJ, Brown CL, Maier KC, Quintyne NJ, Schroer TA: **Analysis of the dynein-dynactin interaction in vitro and in vivo.** *Mol Biol Cell* 2003, **14**(12):5089–5097.

[152] Vallee RB, Williams JC, Varma D, Barnhart LE: **Dynein: An ancient motor protein involved in multiple modes of transport.** *J Neurobiol* 2004, **58**(2):189–200.

[153] Muresan V, Stankewich MC, Steffen W, Morrow JS, Holzbaur EL, Schnapp BJ: **Dynactin-dependent, dynein-driven vesicle transport in the absence of membrane proteins: a role for spectrin and acidic phospholipids.** *Mol Cell* 2001, **7**:173–183.

[154] Karki S, Holzbaur EL: **Cytoplasmic dynein and dynactin in cell division and intracellular transport.** *Curr Opin Cell Biol* 1999, **11**:45–53.

[155] Xia Q, Zhou Z, Lu C, Cheng D, Dai F, Li B, Zhao P, Zha X, Cheng T, Chai C, Pan G, Xu J, Liu C, Lin Y, Qian J, Hou Y, Wu Z, Li G, Pan M, Li C, Shen Y, Lan X, Yuan L, Li T, Xu H, Yang G, Wan Y, Zhu Y, Yu M, Shen W, Wu D, Xiang Z, Yu J, Wang J, Li R, Shi J, Li H, Li G, Su J, Wang X, Li G, Zhang Z, Wu Q, Li J, Zhang Q, Wei N, Xu J, Sun H, Dong L, Liu D, Zhao S, Zhao X, Meng Q, Lan F, Huang X, Li Y, Fang L, Li C, Li D, Sun Y, Zhang Z, Yang Z, Huang Y, Xi Y, Qi Q, He D, Huang H, Zhang X, Wang Z, Li W, Cao Y, Yu Y, Yu H, Li J, Ye J, Chen H, Zhou Y, Liu B, Wang J, Ye J, Ji H, Li S, Ni P, Zhang J, Zhang Y, Zheng H, Mao B, Wang W, Ye C, Li S, Wang J, Wong GKS, Yang H: **A draft sequence for the genome of the domesticated silkworm (Bombyx mori).** *Science* 2004, **306**(5703):1937–1940.

[156] Richards S, Liu Y, Bettencourt BR, Hradecky P, Letovsky S, Nielsen R, Thornton K, Hubisz MJ, Chen R, Meisel RP, Couronne O, Hua S, Smith MA, Zhang P, Liu J, Bussemaker HJ, van Batenburg MF, Howells SL, Scherer SE, Sodergren E, Matthews BB, Crosby MA, Schroeder AJ, Ortiz-Barrientos D, Rives CM, Metzker ML, Muzny DM, Scott G, Steffen D, Wheeler DA, Worley KC, Havlak P, Durbin KJ, Egan A, Gill R, Hume J, Morgan MB, Miner G, Hamilton C, Huang Y, Waldron L, Verduzco D, Clerc-Blankenburg KP, Dubchak I, Noor MAF, Anderson W, White KP, Clark AG, Schaeffer SW, Gelbart W, Weinstock GM, Gibbs RA: **Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution.** *Genome Res* 2005, **15**:1–18.

[157] Miki H, Okada Y, Hirokawa N: **Analysis of the kinesin superfamily: insights into structure and function.** *Trends Cell Biol* 2005, **15**(9):467–476.

[158] Miki H, Setou M, Kaneshiro K, Hirokawa N: **All kinesin superfamily protein, KIF, genes in mouse and human.** *Proc Natl Acad Sci U S A* 2001, **98**(13):7004–7011.

[159] Hasegawa M, Fujiwara M: **Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny.** *Mol Phylogenet Evol* 1993, **2**:1–5.

[160] Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, Iyer VN, Pollard DA, Sackton TB, Larracuente AM, Singh ND, Abad JP, Abt DN, Adryan B, Aguade M, Akashi H, Anderson WW, Aquadro CF, Ardell DH, Arguello R, Artieri CG, Barbash DA, Barker D, Barsanti P, Batterham P, Batzoglou S, Begun D, Bhutkar A, Blanco E, Bosak SA, Bradley RK, Brand AD, Brent MR, Brooks AN, Brown RH, Butlin RK, Caggese C, Calvi BR, Bernardo de Carvalho A, Caspi A, Castrezana S, Celniker SE, Chang JL, Chapple C, Chatterji S, Chinwalla A, Civetta A, Clifton SW, Comeron JM, Costello JC, Coyne JA, Daub J, David RG, Delcher AL, Delehaunty K, Do CB,

Ebling H, Edwards K, Eickbush T, Evans JD, Filipski A, Findeiss S, Freyhult E, Fulton L, Fulton R, Garcia ACL, Gardiner A, Garfield DA, Garvin BE, Gibson G, Gilbert D, Gnerre S, Godfrey J, Good R, Gotea V, Gravely B, Greenberg AJ, Griffiths-Jones S, Gross S, Guigo R, Gustafson EA, Haerty W, Hahn MW, Halligan DL, Halpern AL, Halter GM, Han MV, Heger A, Hillier L, Hinrichs AS, Holmes I, Hoskins RA, Hubisz MJ, Hultmark D, Huntley MA, Jaffe DB, Jagadeeshan S, Jeck WR, Johnson J, Jones CD, Jordan WC, Karpen GH, Kataoka E, Keightley PD, Kheradpour P, Kirkness EF, Koerich LB, Kristiansen K, Kudrna D, Kulathinal RJ, Kumar S, Kwok R, Lander E, Langley CH, Lapoint R, Lazzaro BP, Lee SJ, Levesque L, Li R, Lin CF, Lin MF, Lindblad-Toh K, Llopart A, Long M, Low L, Lozovsky E, Lu J, Luo M, Machado CA, Makalowski W, Marzo M, Matsuda M, Matzkin L, McAllister B, McBride CS, McKernan B, McKernan K, Mendez-Lago M, Minx P, Mollenhauer MU, Montooth K, Mount SM, Mu X, Myers E, Negre B, Newfeld S, Nielsen R, Noor MAF, O'Grady P, Pachter L, Papaceit M, Parisi MJ, Parisi M, Parts L, Pedersen JS, Pesole G, Phillippy AM, Ponting CP, Pop M, Porcelli D, Powell JR, Prohaska S, Pruitt K, Puig M, Quesneville H, Ram KR, Rand D, Rasmussen MD, Reed LK, Reenan R, Reily A, Remington KA, Rieger TT, Ritchie MG, Robin C, Rogers YH, Rohde C, Rozas J, Rubenfield MJ, Ruiz A, Russo S, Salzberg SL, Sanchez-Gracia A, Saranga DJ, Sato H, Schaeffer SW, Schatz MC, Schlenke T, Schwartz R, Segarra C, Singh RS, Sirot L, Sirota M, Sisneros NB, Smith CD, Smith TF, Spieth J, Stage DE, Stark A, Stephan W, Strausberg RL, Strempel S, Sturgill D, Sutton G, Sutton GG, Tao W, Teichmann S, Tobari YN, Tomimura Y, Tsolas JM, Valente VLS, Venter E, Venter JC, Vicario S, Vieira FG, Vilella AJ, Villasante A, Walenz B, Wang J, Wasserman M, Watts T, Wilson D, Wilson RK, Wing RA, Wolfner MF, Wong A, Wong GKS, Wu CI, Wu G, Yamamoto D, Yang HP, Yang SP, Yorke JA, Yoshida K, Zdobnov E, Zhang P, Zhang Y, Zimin AV, Baldwin J, Abdouelleil A, Abdulkadir J, Abebe A, Abera B, Abreu J, Acer SC, Aftuck L, Alexander A, An P, Anderson E, Anderson S, Arachi H, Azer M, Bachantsang P, Barry A, Bayul T, Berlin A, Bessette D, Bloom T, Blye J, Boguslavskiy L, Bonnet C, Boukhgalter B, Bourzgui I, Brown A, Cahill P, Channer S, Cheshatsang Y, Chuda L, Citroen M, Collymore A, Cooke P, Costello M, D'Aco K, Daza R, De Haan G, DeGray S, DeMaso C, Dhargay N, Dooley K, Dooley E, Doricent M, Dorje P, Dorjee K, Dupes A, Elong R, Falk J, Farina A, Faro S, Ferguson D, Fisher S, Foley CD, Franke A, Friedrich D, Gadbois L, Gearin G, Gearin CR, Giannoukos G, Goode T, Graham J, Grandbois E, Grewal S, Gyaltsen K, Hafez N, Hagos B, Hall J, Henson C, Hollinger A, Honan T, Huard MD, Hughes L, Hurhula B, Husby ME, Kamat A, Kanga B, Kashin S, Khazanovich D, Kisner P, Lance K, Lara M, Lee W, Lennon N, Letendre F, LeVine R, Lipovsky A, Liu X, Liu J, Liu S, Lokyitsang T, Lokyitsang Y, Lubonja R, Lui A, MacDonald P, Magnisalis V, Maru K, Matthews C, McCusker W, McDonough S, Mehta T, Meldrim J, Meneus L, Mihai O, Mihalev A, Mihova T, Mittelman R, Mlenga V, Montmayeur A, Mulrain L, Navidi A, Naylor J, Negash T, Nguyen T, Nguyen N, Nicol R, Norbu C, Norbu N, Novod N, O'Neill B, Osman S, Markiewicz E, Oyono OL, Patti C, Phunkhang P, Pierre F, Priest M, Raghuraman S, Rege F, Reyes R, Rise C, Rogov P, Ross K, Ryan E, Settipalli S, Shea T, Sherpa N, Shi L, Shih D, Sparrow T, Spaulding J, Stalker J, Stange-Thomann N, Stavropoulos S, Stone C, Strader C, Tesfaye S, Thomson T, Thoulutsang Y, Thoulutsang D, Topham K, Topping I, Tsamla T, Vassiliev H, Vo A, Wangchuk T, Wangdi T, Weiand M, Wilkinson J, Wilson A, Yadav S, Young G, Yu Q, Zembek L, Zhong D, Zimmer A, Zwirko Z, Jaffe DB, Alvarez P, Brockman W, Butler J, Chin C, Gnerre S, Grabherr M, Kleber M, Mauceli E, MacCallum I: **Evolution of genes and genomes on the Drosophila phylogeny.** *Nature* 2007, **450**(7167):203–218.

[161] Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD: **Multiple sequence alignment with the Clustal series of programs.** *Nucleic Acids Res* 2003, **31**(13):3497–3500.

[162] Felsenstein J: **Evolutionary trees from DNA sequences: a maximum likelihood approach.** *J Mol Evol* 1981, **17**(6):368–376.

[163] Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Syst Biol* 2003, **52**(5):696–704.

[164] Ronquist F, Huelsenbeck JP: **MrBayes 3: Bayesian phylogenetic inference under mixed models.** *Bioinformatics* 2003, **19**(12):1572–1574.

[165] Letunic I, Copley RR, Bork P: **Common exon duplication in animals and its role in alternative splicing.** *Hum Mol Genet* 2002, **11**(13):1561–1567.

[166] Rost B, Yachdav G, Liu J: **The PredictProtein server.** *Nucleic Acids Res* 2004, **32**(Web Server issue):W321–6.

[167] Neal S, Nip AM, Zhang H, Wishart DS: **Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts.** *J Biomol NMR* 2003, **26**(3):215–240.

[168] Matsuki Y, Akutsu H, Fujiwara T: **Spectral fitting for signal assignment and structural analysis of uniformly 13C-labeled solid proteins by simulated annealing based on chemical shifts and spin dynamics.** *J Biomol NMR* 2007, **38**(4):325–339.

[169] Golotvin SS, Vodopianov E, Lefebvre BA, Williams AJ, Spitzer TD: **Automated structure verification based on 1H NMR prediction.** *Magn Reson Chem* 2006, **44**(5):524–538.

[170] [http://www.wwpdb.org/docs.html].

[171] Goddard TD, Kneller DG: **SPARKY 3**[http://www.cgl.ucsf.edu/home/sparky/].

[172] [http://www.bmrb.wisc.edu/].

[173] Takegoshi K, Nakamura S, T T: **13C–1H dipolar-assisted rotational resonance in magic-angle spinning NMR**. *Chemical Physics Letters* 2001, **344**.

[174] Seidel K, Etzkorn M, Heise H, Becker S, Baldus M: **High-resolution solid-state NMR studies on uniformly [13C,15N]-labeled ubiquitin.** *Chembiochem* 2005, **6**(9):1638–1647.

[175] Vuister W, Kim SJ, Wu C, Bax A: **2D and 3D NMR Study of Phenylalanine Residues in Proteins by Reverse Isotopic Labeling**. *J. Am. Chem. Soc* 1994.

B Abbreviations

Abbr.	Meaning
Abbr.	Abbreviation
ADP	Adenosine diphosphate
ATP	Adenosine triphosphate
BLAST	Basic local alignment search tool
BLASTP	Protein-protein BLAST
BLAT	BLAST like alignment tool
BLOSUM	Blocks of amino acid substitution matrix
BMBF	Bundesministerium fr Bildung und Forschung
BMRB	Biological magnetic resonance bank
CBS	Cystathionine beta-synthase
cDNA	copy DNA
cGMP	Cyclic guanosine monophosphate
DARR	Dipolar assisted rotational resonance
DGC	Daphnia Genomics Consortium
DNA	Deoxyribonucleic acid
EST	Expressed sequence tag
GFF	General feature format
GTP	Guanosine triphosphate
ID	Identification Number
ISC	International Sequencing Consortium
JGI	Joint Genome Institute
KHC	Kinesin heavy chain
KIF	Kinesin family member
KLC	Kinesin light chains
KLP	Kinesin-like protein
KRP	Kinesin-related protein
Mhc	Myosin heavy chain
NCBI	National Center for Biotechnology Information
NCSRS	Non-Coding Sequence Retrieval System
NMR	Nuclear magnetic resonance
PDB	Protein Data Bank
PNG	Portable network graphics
RAM	Random access memory
RNA	Ribonucleic acid
RPC	Remote procedure call
RRE	Retrieval of Regulative Regions
SOAP	Simple object access protocol
SQL	Structured english query language
SVG	Scalable vector graphics
TBLASTN	Protein-nucleotide BLAST
UCSC	University of California, Santa Cruz
URL	Uniform resource locator
WGS	Whole-genome shotgun
WSDL	Web services description language
XHTML	Extensible hyper text markup language
XML	Extensible markup language
XML-RPC	XML remote procedure call
YAML	YAML ain't markup language

Species Abbr.	Species
Aea	*Aedes aegypti str. Liverpool*
Am	*Apis mellifera str. DH4*
Ang	*Anopheles gambiae str. PEST*
As	*Anemonia sulcata*
Bos	*Botryllus schlosseri*
Bt	*Bos taurus*
Caf	*Canis lupus familiaris str. boxer*
Cpq	*Culex pipiens quinquefasciatus str. JHB*
Da	*Drosophila ananassae TSC#14024-0371.13*
Dap	*Daphnia pulex*
Der	*Drosophila erecta TSC#14021-0224.01*
Dg	*Drosophila grimshawi TSC#15287-2541.00*
Dh	*Drosophila hydei*
Dm	*Drosophila melanogaster*
Dmo	*Drosophila mojavensis TSC#15081-1352.22*
Dp	*Drosophila pseudoobscura MV2-25*
Drm	*Drosophila mimetica*
Drp	*Drosophila persimilis MSH-3*
Dse	*Drosophila sechellia Rob3c*
Dss	*Drosophila simulans str. Mosaic*
Dv	*Drosophila virilis TSC#15010-1051.87*
Dw	*Drosophila willistoni TSC#14030-0811.24*
Dy	*Drosophila yakuba Tai18E2*
He	*Heliconius erato*
Hs	*Homo sapiens*
Mam	*Macaca mulatta*
Md	*Monodelphis domestica*
Mm	*Mus musculus C57BL/6J*
Mus	*Mougeotia scalaris*
Nav	*Nasonia vitripennis str. SymAX*
Pat	*Pan troglodytes*
Pdc	*Pediculus humanus corporis str. USDA*
Ras	*Radopholus similis*
Rn	*Rattus norvegicus BN/SsNHsdMCW*
So	*Saccharum officinarum*
Tic	*Tribolium castaneum str. Georgia GA2*

C Curriculum vitae

Name	Florian Odronitz
Address	Keplerstr. 18 37085 Göttingen
Date of Birth	06/03/1976
Place of Birth	Reutlingen, Germany
Education	1982-1987 Matthäus Beger-Schule Reutlingen 1987-1996 Isolde Kurz-Gymnasium Reutlingen
Community Service	July 1996 to August 1997 Paramedics, Deutsches Rotes Kreuz
Working Experience	August to October 1997 Digitalstudio Lange, Rottenburg Conception, Webdesign, Video Compositing November 1997 to May 1998 Welsch & Partner, Tübingen 3D-Animation, Scripting
PhD Studies	May 2005 - present PhD Study in the Kollmar Group at the Max Planck Institute for Biophysical Chemistry, Göttingen
Studies	1998 - 2004 Diploma Programme in Biology at the University of Konstanz
Study Emphasis	Molecular Genetics Biophysics Information Visualization Neuroethology Evolutionary Biology
Hands-on Training	August 2000 Byk Gulden (now Nycomed), Konstanz Department of Bioinformatics March and April 2001 Fraunhofer Gesellschaft, St. Augustin Department of Biomolecular Information Processing
Studies Abroad	August 2002 to June 2003 at the University of Western Ontario, London, Canada
Student Assistance	Database Design and Database Programming Conference Organisation IT Support Network Security
Degree	Diploma in Biology,

Supervision	2005 - present Several months of hands-on-training with computer science students August - September 2008 Bachelor project in computer science: "WebScipio: Implementierung eines Webservice und eines Webinterfaces zur Exonsuche in eukaryotischen Genomen"
Conferences	September 2005 Motors and Cytoskeleton Hamburg, Germany May 2006 8th International School on the Crystallography of Biological Macromolecules Como, Italy January 2007 5th European Conference on Computational Biology - ECCB 06 Eilat, Israel June 2007 NETTAB 2007: A Semantic Web for Bioinformatics: Goals, Tools, Systems, Applications Pisa, Italy
Publications	F Odronitz and M Kollmar. Drawing the tree of eukaryotic life based on the analysis of 2,269 manually annotated myosins from 328 species. *Genome Biol*, 8(9):R196, 2007. (marked as "highly accessed") F. Odronitz, M. Hellkamp, and M. Kollmar. diArk–a resource for eukaryotic genome research. *BMC Genomics*, 8:103, 2007. 1471-2164 (marked as "highly accessed") F. Odronitz and M. Kollmar. Pfarao: a web application for protein family analysis customized for cytoskeletal and motor proteins (cymobase). *BMC Genomics*, 7:300, 2006. 1471-2164 Florian Odronitz, Jens Baltin, Sandra Leist, Hans-Peter Wollscheid, Martina Baack, Thomas Kapitza, Daniel Schaarschmidt, and Rolf Knippers. Dna replication in protein extracts from human cells requires orc and mcm proteins. *J BiolChem*, 281(18):12428–12435, 2006.