Development of novel Classical and Quantum Information Theory Based Methods for the Detection of Compensatory Mutations in MSAs

Dissertation
zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades
”Doctor rerum naturalium”
der Georg-August-Universität Göttingen

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

vorgelegt von

Mehmet Gültas
aus Kirikkale-Türkei

Göttingen, 2013

Betreuungsausschuss

Professor Dr. Stephan Waack,
Institut für Informatik, Georg-August-Universität Göttingen.

Professor Dr. Carsten Damm,
Institut für Informatik, Georg-August-Universität Göttingen.

Professor Dr. Edgar Wingender,
Institut für Bioinformatik, Universitätsmedizin, Georg-August-Universität Göttingen.

Mitglieder der Prüfungskommission

Referent:	Prof. Dr. Stephan Waack, Institut für Informatik, Georg-August-Universität Göttingen.
Korreferent:	Prof. Dr. Carsten Damm, Institut für Informatik, Georg-August-Universität Göttingen.
Korreferent:	Prof. Dr. Mario Stanke, Institut für Mathematik und Informatik, Ernst Moritz Arndt Universität Greifswald

Weitere Mitglieder der Prüfungskommission

Prof. Dr. Edgar Wingender,
Institut für Bioinformatik, Universitätsmedizin, Georg-August-Universität Göttingen.

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

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

Prof. Dr. Wolfgang May,
Institut für Informatik, Georg-August-Universität Göttingen.

Tag der mündlichen Prüfung: 18. September 2013

Abstract

Multiple sequence alignments (MSAs) of homologous proteins are useful tools to characterize compensatory mutations between non-conserved residue sites. The identification of these residue sites in MSAs is an important challenge for understanding the structural basis and molecular mechanism of protein functions. Despite the rich literature on compensatory mutations as well as sequence conservation analysis for the identification of those important residue sites, previous methods often do not take into account biochemical constraints of amino acids which are likely to be crucial for the detection of compensatory mutation signals. However, compensatory mutation signals in MSAs are often masked by noise. Thus, another challenging problem in bioinformatics is the separation of significant signals from the phylogenetic noise and unrelated pair signals.

The goal of this thesis is to develop such methods that incorporate biochemical constraints like similarities or dissimilarities of amino acids in identifying compensatory mutations and deal with the noise. Hence, we develop different methods based on classical and quantum information theory and multiple testing procedures.

Our first method is based on classical information theory. It mainly focuses on *BLOSUM62*-dissimilar amino acid pairs as a model of compensatory mutations and incorporates them in the prediction of functionally and/or structurally important sites using a doubly stochastic matrix. To complement this method, we develop our second method applying principles of quantum information theory. The new method differs from the first one by simultaneously modeling similar and dissimilar amino acid pair signals in the compensatory mutation analysis. Moreover, to separate method-based significant compensatory mutation signals from background noise, we develop an MSA-specific statistical model devised for multiple testing problems. By applying this model, we are capable of determining significant signals in MSAs as well as quantifying the error made in terms of the false discovery rate.

To demonstrate the effectiveness of our methods, we evaluate those analyzing important sites of two human proteins, namely epidermal growth factor receptor (EGFR) and glucokinase (GCK). Our results suggest that the MSA-specific statistical model is able to separate significant compensatory mutation signals from the phylogenetic noise and unrelated pair signals. Only considering the dissimilarities of amino acids, the first method successfully deals with disease-associated important sites of both proteins. In contrast, simultaneously focusing on similar and dissimilar amino acid signals, the second method is more sensible to catalytic, allosteric and binding sites. The results further show that overlaps between both methods are quite low, indicating that considerably different sets of residue sites are detected by both methods as functionally and structurally important. As a result of this, we can say that our second method complements the first method when it comes to predicting important sites, rather than replacing it.

Zusammenfassung

Multiple Sequenzalignments (MSAs) von homologen Proteinen sind nützliche Werkzeuge, um kompensatorische Mutationen zwischen nicht-konservierten Residuen zu charakterisieren. Die Identifizierung dieser Residuen in MSAs ist eine wichtige Aufgabe um die strukturellen Grundlagen und molekularen Mechanismen von Proteinfunktionen besser zu verstehen. Trotz der vielen Anzahl an Literatur über kompensatorische Mutationen sowie über die Sequenzkonservierungsanalyse für die Erkennung von wichtigen Residuen, haben vorherige Methoden meistens die biochemischen Eigenschaften von Aminosäuren nicht mit in Betracht gezogen, welche allerdings entscheidend für die Erkennung von kompensatorischen Mutationssignalen sein können. Jedoch werden kompensatorische Mutationssignale in MSAs oft durch das Rauschen verfälscht. Aus diesem Grund besteht ein weiteres Problem der Bioinformatik in der Trennung signifikanter Signale vom phylogenetischen Rauschen und beziehungslosen Paarsignalen.

Das Ziel dieser Arbeit besteht darin Methoden zu entwickeln, welche biochemische Eigenschaften wie Ähnlichkeiten und Unähnlichkeiten von Aminosäuren in der Identifizierung von kompensatorischen Mutationen integriert und sich mit dem Rauschen auseinandersetzt. Deshalb entwickeln wir unterschiedliche Methoden basierend auf klassischer- und quantum Informationstheorie sowie multiple Testverfahren.

Unsere erste Methode basiert auf der klassischen Informationstheorie. Diese Methode betrachtet hauptsächlich *BLOSUM62*-unähnliche Paare von Aminosäuren als ein Modell von kompensatorischen Mutationen und integriert sie in die Identifizierung von wichtigen Residuen. Um diese Methode zu ergänzen, entwickeln wir unsere zweite Methode unter Verwendung der Grundlagen von quantum Informationstheorie. Diese neue Methode unterscheidet sich von der ersten Methode durch gleichzeitige Modellierung ähnlicher und unähnlicher Signale in der kompensatorischen Mutationsanalyse. Des Weiteren, um signifikante Signale vom Rauschen zu trennen, entwickeln wir ein MSA-spezifisch statistisches Modell in Bezug auf multiple Testverfahren.

Wir wenden unsere Methode für zwei menschliche Proteine an, nämlich epidermal growth factor receptor (EGFR) und glucokinase (GCK). Die Ergebnisse zeigen, dass das MSA-spezifisch statistische Modell die signifikanten Signale vom phylogenetischen Rauschen und von beziehungslosen Paarsignalen trennen kann. Nur unter Berücksichtigung *BLOSUM62*-unähnlicher Paare von Aminosäuren identifiziert die erste Methode erfolgreich die krankheits-assoziierten wichtigen Residuen der beiden Proteine. Im Gegensatz dazu, durch die gleichzeitige Modellierung ähnlicher und unähnlicher Signale von Aminosäurepaare ist die zweite Methode sensibler für die Identifizierung von katalytischen und allosterischen Residuen.

Acknowledgements

I am indebted to many people for their support and encouragement which was very crucial for the successful completion of my PhD study. In the following lines I would like to use this opportunity to acknowledge and thank all.

First of all, I would like to thank my advisor Prof. Dr. Stephan Waack for giving me the opportunity to join his research group and for his excellent guidance through my study. Without his patient teaching and continuous trust, this work would not have been successful.

My sincere thanks go to my co-advisor Prof. Dr. Carsten Damm patiently guiding me through my work by sharing his experiences with me and for countless hours of fruitful discussions. Without his valuable comments and suggestions, this thesis would not have been possible. I would also like to thank Prof. Dr. Mario Stanke for being in my thesis committee and for sharing valuable ideas.

I would like to thank all my colleagues and friends for their enjoyable and scientifically inspiring environment and for our wonderful friendship: Lennart Obermann, Steffen Herbold, Martin Haubrock, Keyu Wang, Gunnar Krull, Patrick Harms, Thomas Rings, Heike Jachinke, Annette Kadziora, Cansu Topal, Nesrin Tüysüz and Pinar Öz. I would also like to thank the remaining members of my defense committee: Prof. Dr. Edgar Wingender, Prof. Dr. Burkhard Morgenstern, Prof. Dr. Wolfgang May, and Prof. Dr. Dieter Hogrefe.

I want to thank my parents: Elif and Arap Gültas, and my siblings: Dilek Aydogdu, Erdem Gültas and Bayram Gültas for all the love and support over the years.

Especially, I would like to thank my girl friend Susann Michanski who always believes in me and supports me in my research and my daily life. I wouldn't have made it without her.

Contents

List of Figures xi

List of Tables xiv

1 Introduction 2
- 1.1 Structure of the Thesis 3
- 1.2 Impact 4

2 Biological Background 7
- 2.1 Amino Acids 7
- 2.2 Proteins 9
- 2.3 Multiple Sequence Alignments 10
- 2.4 Mutations in Protein 12
- 2.5 Bioinformatics Databases and Tools 13
  - 2.5.1 Protein Data Bank 14
  - 2.5.2 HSSP Database 17
  - 2.5.3 Catalytic Site Atlas and PDBsum 21
  - 2.5.4 Calculation of Distances Between Residues using BioJava 21
  - 2.5.5 BLOSUM Matrices 23

3 Theoretical Background 25
- 3.1 Classical Information Theory 25
  - 3.1.1 Relative Entropy and Mutual Information 28
- 3.2 Quantum Information Theory 30
  - 3.2.1 Basic definitions of quantum mechanics 30
  - 3.2.2 Quantum Information and von Neumann Entropy 36
- 3.3 Mathematical Foundations 37

4 Applying Classical Information Theory for the Compensatory Mutation Analysis 39
- 4.1 Detecting compensatory mutations by the $\mathbb{U}$-metric 39
  - 4.1.1 Filtering of an MSA 43
- 4.2 Determining an MSA-specific lower bound for the significance of $\mathbb{U}$-values 44
  - 4.2.1 Step 1: Calculation of p-values 44

4.2.2 Step 2: Setting an MSA-dependent threshold τ 46
4.3 Enhancing prediction by the $\mathbb{U}_{D(\alpha)}$-metric that models dissimilar compensatory mutations 49
4.3.1 Training data set 50
4.3.2 Preparing substitution matrices 51
4.3.3 Creating the Doubly Stochastic Matrix to model dissimilar compensatory mutations 51
4.3.4 Application of doubly stochastic matrix D_{CompMut} for the computation of the new $\mathbb{U}_{D(\alpha)}$-metric 54
4.4 Discussion 56

5 Applying quantum information theory for the detection of functionally or structurally important sites 58
5.1 Notation mapping from quantum information theory in protein bioinformatics 58
5.1.1 Calculating the Gram matrices $\mathcal{A}_{\text{ent}}$ and $\mathcal{A}_{\text{sep}}$ to be plugged into Equation 5.1.4 62
5.1.2 Simultaneously ensuring Equations 5.1.3, 5.1.4, and 5.1.7 63
5.2 Defining two new metrics $\mathbb{Q}_{\text{ent}}$ and $\mathbb{Q}_{\text{sep}}$ 63
5.3 Discussion 64

6 Results 66
6.1 Essential sites of human EGFR and GCK protein 67
6.1.1 MSAs and 3-D structures of human EGFR and GCK protein 67
6.2 Applying the $\mathbb{U}$-metric and the $\mathbb{U}_{D(\alpha)}$-metric for the human EGFR and GCK protein 67
6.2.1 Position analysis of the Human Epidermal Growth Factor Receptor (EGFR) protein 73
6.2.2 Position analysis of the Human Glucokinase (GCK) protein 76
6.2.3 A comparison between $\mathbb{U}$-metric and $\mathbb{U}_{D(\alpha)}$-metric 79
6.3 Applying the $\mathbb{Q}_{\text{ent}}$ and the $\mathbb{Q}_{\text{sep}}$-metric for the human EGFR and GCK proteins 80
6.3.1 Position analysis of the Human EGFR protein 80
6.3.2 Position analysis of the the Human GCK protein 85
6.4 A Comparison between CMF-significant sites and QCMF-significant sites . 88

7 Discussion 92
7.1 Connectivity Degree Cut-off 92
7.2 CMF-significant residue sites 93
7.3 QCMF-significant residue sites 94

8 Conclusion 96
8.1 Summary 96

8.2 Outlook 97

Bibliography **99**

9 Appendix **110**
9.1 Appendix A: Coupled Mutation Finder 110
9.2 Appendix B: Quantum Coupled Mutation Finder 123
9.3 Appendix C: PDB enteries of training protein MSAs 141
9.4 Appendix D: Curriculum Vitae 147

List of Figures

2.1 General structure of an amino acid. 7

2.2 **Overview of the structural levels in proteins** (image from [[1], p.92]). . 9

2.3 A small section of a multiple sequence alignment. 11

2.4 **The distance calculation:** The balls show the C_α atoms at positions 14 and 15 in human glucokinase protein with PDB entry 1V4S. Based on the x,y, and z coordinates of these atoms, the Euclidean distance is calculated as 3.800095. 22

2.5 **BLOSUM62 matrix** 23

3.1 **Venn diagram visualization of entropy $\mathbb{H}(X)$, joint entropy $\mathbb{H}(X,Y)$, conditional entropy $\mathbb{H}(X|Y)$, and mutual information $\mathbb{MI}(X;Y)$.** 27

4.1 **A small part of an MSA.** The positions k and l show a strong correlation. At position k the $[A$ to $E]$ mutation cause at position l the $[V$ to $G]$ mutation or vice versa. In addition, the position m demonstrates an example for strictly conserved residues. 40

4.2 **An artificial MSA.** The positions k_1 and l_1 are slightly conserved and contain only two amino acids. The positions k_2 and l_2 contain all 20 amino acids. Since the computation of $\mathbb{MI}$ values strongly depend on the empirical amino acid distributions of the columns which is also related to the alphabet size, both pairs of columns (k_1,l_1) and (k_2,l_2) have completely different $\mathbb{MI}$ values. The more different amino acids are included in a pair of columns, the higher are their individual column entropy and related joint entropy that often causes a greater $\mathbb{MI}$ value. However, if a column pair is slightly conserved like (k_1,l_1), the alphabet size in these column pair is quite low which results in a less $\mathbb{MI}$ value. 42

4.3 **p-value distribution of $\mathbb{U}$-values for human GCK protein (PDB Entry 1V4S).** The p-values close to zero represent the significant pairs by means of which we assess the individual residue position. In contrast, p-values of unrelated pairs tend to one. As one can see, the p-values of F_0 distributed $\mathbb{U}$-values are approximately uniform. 45

4.4 **$\mathbb{U}$-value distribution for human GCK protein (PDB Entry 1V4S)**. The higher the $\mathbb{U}$-value between pairs of columns, the stronger the correlation of amino acids at these columns. 46

4.5 **p-value distribution of $\mathbb{U}$-values according to the sample mean and variance for human GCK protein (PDB Entry 1V4S).** The p-values are estimated based on the sample mean and variance which is calculated using all $\mathbb{U}$-values of M. As one can see, if we use directly the sample variance for the estimation of shape parameters α and β of the β-distribution, the p-values of $\mathbb{U}$-values do not follow the uniform distribution over $[0,1]$. . . 47

4.6 **A small part of connectivity degree network of human GCK protein (PDB-Entery 1V4S):** The nodes represent individual residue sites and each vertex indicate a significant correlation ($\mathbb{U}$-value) between residue sites k and l. The vertex count of a node indicates its connectivity degree. The red circles shows residue sites which have three highest connectivity degree in whole network and thus defined as $(\mathbb{U},M)$-*significant*. 50

4.7 **Counting pair to pair substitution:** In the column pair (k,l) occurs different pair to pair substitution such as ([A,V] $\leftrightarrow$ [E,G]), ([A,V] $\leftrightarrow$ [A,V]) and ([E,G] $\leftrightarrow$ [E,G]) or vice versa. Since the correct order of the sequences in MSAs is unknown, we count the substitutions symmetrically and store the occurrence values of substitutions in a matrix C. 52

6.1 **The sequence of the human EGFR protein (PDB-Entry 2J6M)** (image from http://www.rscb.org [2], 05.08.2013). 68

6.2 **The sequence of the human GCK protein (PDB-Entry 1V4S)** (image from http://www.rscb.org [2], 05.08.2013). 69

6.3 CMF**-significant nsSNP positions in human EGFR protein (PDB-Entry 2J6M).** The red spheres correspond to structural localization of 15 different CMF-significant nsSNP positions in the EGFR protein. 73

6.4 CMF**-significant residue positions are in contact with gefitinib binding sites in human EGFR protein (PDB-Entry 2J6M).** Yellow spheres denote positions of the gefitinib binding sites in the wild type kinase. Blue spheres show the localization of CMF-significant adjacent residue positions which are in contact with these binding sites. Additionally, the CMF-significant sites I759, Y764, M766 and K846, shown with green spheres, are already described as nsSNP positions and they are also in contact with gefitinib binding sites E762 and M793, respectively. Circles indicate clusters of gefitinib binding sites and their significant adjacent sites. 74

6.5 CMF**-significant nsSNP positions in human GCK protein (PDB-Entry 1V4S).** Red spheres show the structural localization of 16 different CMF-significant nsSNP positions in the GCK protein. 76

6.6 CMF**-significant residue positions are in contact with glucose binding site and allosteric site in human GCK protein (PDB-Entry 1V4S).** (A) Yellow spheres show the structural positions of glucose binding sites (active sites). Blue spheres correspond to localization of CMF-significant adjacent residue positions which are in contact with these active sites. (B) Orange spheres denote the allosteric sites. Blue spheres correspond to localization of just significant adjacent residue positions and green spheres indicate the significant residue positions which are already described as nsSNP positions and in contact with these allosteric sites. Additionally, the significant position R63 is allosteric site by itself and it is also in contact with another allosteric site. Circles indicate clusters of glucose binding sites (A), allosteric sites (B), and their significant adjacent sites. 77

6.7 **p-value distributions of $\mathbb{Q}_{\text{ent}}$ and $\mathbb{Q}_{\text{sep}}$-values for human EGFR protein (PDB-Entry 2J6M).** The blue bars illustrate the p-value distribution of the $\mathbb{Q}_{\text{ent}}$-values and red bars display the p-value distribution of the $\mathbb{Q}_{\text{sep}}$-values. 81

6.8 QCMF**-significant residue positions are in contact with catalytic residues in human EGFR protein (PDB-Entry 2J6M).** Red spheres denote positions of the catalytic residues. Yellow spheres show the localization of QCMF-significant adjacent residue positions which are in contact with these catalytic residues. Moreover, the QCMF-significant sites A839 and R841 are also catalytic residues by themselves. Green spheres show the structural localization of QCMF-significant nsSNP positions in the EGFR protein. Circles indicate clusters of catalytic residue sites and their significant adjacent sites. 82

6.9 QCMF**-significant residue positions are in contact with gefitinib binding sites in human EGFR protein (PDB-Entry 2J6M).** Red spheres show the structural localization of the gefitinib binding sites in the wild type kinase. Yellow spheres show significant adjacent residue positions which are in contact with these binding sites. Moreover, the QCMF-significant site T854 is also binding site by itself and interacts with gefitinib binding sites D855. Circles indicate clusters of gefitinib binding sites and their significant adjacent sites. 84

6.10 QCMF**-significant positions that are either in contact with allosteric sites or related to nsSNPs in human GCK protein (PDB-Entry 1V4S).** Yellow spheres correspond to structural localization of ten QCMF-significant residue sites which are in contact with allosteric sites where the V62, M210, Y215 are denoted as allosteric sites by themselves and they are also in contact with an other allosteric sites. Green spheres indicate eight QCMF-significant nsSNP positions in the GCK protein. Three of them (T209M, G223S and S453del) are further in contact with allosteric sites M210, I211, V452, V455 and A456. 89

6.11 QCMF**-significant residue positions are in contact with glucose binding site in human GCK protein (PDB-Entry 1V4S).** (A) Red spheres show the structural positions of the glucose binding sites (active sites) and yellow spheres show the localization of QCMF-significant adjacent residue positions which are in contact with these active sites. The circles indicate clusters of glucose binding sites and their significant adjacent sites. 91

List of Tables

2.1 **The amino acids:** Full names, three-letter abbreviations and single-letter abbreviations of primary amino acids. 8

2.2 **The standard genetic code:** Triplet combinations of nucleotides into amino acids. There are in total of 64 triplets, 61 of them encode the 20 amino acids and 3 of them are stop codons which do not code for any amino acid. 13

4.1 $\hat{P}_{20\times20}$: Observed probabilities of amino acid pairs in column pair (k,l), where $x_i, y_j \in$ 20 amino acids$_{(A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y)}$. 54

4.2 $\hat{P}^{*}{}_{20\times20}$: Transformed values of joint amino acid pair distribution 55

6.1 43 CMF**-significant residue sites found by $\mathbb{U}$ and $\mathbb{U}_{D(\alpha)}$-metric in human EGFR protein.** 71

6.2 72 CMF**-significant residue sites found by $\mathbb{U}$ and $\mathbb{U}_{D(\alpha)}$-metric in human GCK protein.** 72

6.3 CMF-significant essential sites in human EGFR protein, which are nearby either nsSNPs or strictly conserved sites. 75

6.4 CMF-significant essential sites in human GCK protein, which are nearby either nsSNPs or strictly conserved sites. 78

6.5 **Comparison between $\mathbb{U}$-metric and $\mathbb{U}_{D(\alpha)}$-metric** 79

6.6 33 QCMF**-significant residue sites found by $\mathbb{Q}_{\text{ent}}$-metric and $\mathbb{Q}_{\text{sep}}$-metric in the human EGFR protein. * denotes the residue sites are detected by both metrics $\mathbb{Q}_{\text{ent}}$ and $\mathbb{Q}_{\text{sep}}$ as significant.** 83

6.7 QCMF-significant essential sites in human EGFR protein, which are nearby either nsSNPs or strictly conserved sites. 85

6.8 64 QCMF**-significant residue sites found by $\mathbb{Q}_{\text{ent}}$-metric and $\mathbb{Q}_{\text{sep}}$-metric in human GCK protein. * denotes the residue sites are detected by both metrics $\mathbb{Q}_{\text{ent}}$ and $\mathbb{Q}_{\text{sep}}$ as significant.** 86

6.9 QCMF-significant essential sites in human GCK protein, which are nearby either nsSNPs or strictly conserved sites. 88

7.1 Statistical evaluation of CMF-significant residue sites for both human EGFR and GCK protein 93

7.2 Statistical evaluation of QCMF-significant residue sites for both human EGFR and GCK proteins. 93

9.1 Pdb enteries of training MSAs. 142

1 Introduction

Multiple sequence alignments (MSAs) of homologous protein sequences give us information about two major features of the proteins of interest. The first one consists of easily detectable highly conserved residue sites that are obviously important for the structure and/or function of the protein while the second one corresponds to compensatory (coupled) mutations between two or more non-conserved residue sites that also contain crucial information on the structural and functional basis of proteins [3]. These compensatory mutations occur according to the functional coupling of mutation positions which might be explained as one mutation in a certain site affecting a compensating mutation at another site, even if both related residue sites are distantly positioned in the protein structure [4–8]. In particular, such mutations at essential residue sites are likely to destroy the protein structure which often results in loss of the protein's function [9, 10]. Thus, for understanding the structural basis and molecular mechanism of protein functions, determination of these compensatory mutated residue sites is as important as strictly conserved sites [4,5,11,12]. These residue sites might be disease-associated, responsible for the maintenance of internal protein volume, or possibly form key sites for interactions within or between proteins [4, 8, 13–15].

Although strictly conserved residue sites are easily detectable and interpretable in MSAs, the detection of important non-conserved compensatory mutation sites needs more complex approaches. Until now, a variety of studies have employed Pearson's correlation coefficient methods [16–18], perturbation based methods [15, 19] and mutual information (MI) based methods [8, 20–23] because of their simplicity and efficiency. However, these methods strongly depend on the amino acid distributions observed in MSA columns rather than on physical or biochemical similarities or dissimilarites of amino acids that are likely to be crucial for the detection of functionally or structurally important compensatory mutations. In addition, due to background noise, all of these methods interfere with the identification of compensatory mutation signals [9,20,24]. Hence, the significant compensatory mutation signals must be separated from the background noise that might occur as a result of: i) false signals arising from insufficient data; ii) sites with low or high conservation biasing the signal; iii) phylogenetic noise. While the first two types of noise can be easily overcome by appropriately filtering the data [22], phylogenetic noise can only be eliminated to some extent by excluding highly similar sequences from the MSA [9].

Recently, several methods such as bootstrapping, simulation, or randomization methods have been utilized in order to minimize the influence of phylogenetic linkage and stochastic noise [21, 25, 26]. Dunn et al. [9] have introduced the *average product correction* (APC) to adjust MI for background effects. In their study, Merkl and Zwick [22] have used a normalized MI and focused on only the 75 residue pairs with highest normalized MI values as significant for each MSA. Gao et al. [23] have pursued a similar approach, where they have replaced the metric used in [22] with the amino acid background distribution (MIB). While the reduction of background noise in the model of Dunn et al. is not quantified, the approaches of Gao et al. and Merkl and Zwick appear to be over-conservative.

Despite the presence of a variety of different methods as mentioned above, to date, there is still need for a method that contains powerful metrics to take into account biochemical constraints of amino acids and deals with noise and background signals. As a consequence, the main goal of this thesis was to develop classical information theory based methods and quantum information theory based methods to incorporate biochemical similarities or dissimilarities of amino acids in the prediction of functionally or structurally important residue sites. Furthermore, we have developed an MSA-specific statistical model based on multiple testing procedures described in [27, 28]. In this way, unlike previous information theory based studies [22, 23], we can separate significant compensatory mutation signals from background noise with respect to our MSA-specific statistical model that quantifies the error made in terms of the false discovery rate.

To demonstrate the performance and functionality of our methods, we analyzed the structurally or functionally important positions of two human proteins, namely epidermal growth factor receptor (EGFR) and glucokinase (GCK). The proteins have been chosen because their important positions experimentally investigated in previous studies [29–40]. As a result, in both proteins our methods detect disease-associated amino acid mutations (non-synonymous single nucleotide polymorphisms (nsSNPs)), not strictly conserved catalytic or binding sites, and residues that are nearby one of these sites or in the close neighborhood of a strictly conserved positions, which are likely to affect protein stability or functionality [41–43].

1.1 Structure of the Thesis

The remainder of this thesis is organized as follows. In Chapter 2, we briefly introduce the descriptions of some biological concepts, databases, and techniques of bioinformatics which are required to motivate, and understand functionally or structurally important amino acids in proteins. In the third chapter, we describe mathematical foundations of classical information theory and quantum information theory, as well as the β-distribution, upon which we develop our models for the prediction of significant compensatory mutation signals in MSAs. In Chapters 4 and 5, we describe and develop our new classical information theory and quantum information theory based methods and MSA-specific statistical model. In addition, we present the applications of these methods and discuss them in detail at the end of the each chapter. In Chapter 6, we demonstrate the performance and functionality of our methods by applying them on the prediction of structurally or functionally important positions of two human proteins. Afterwards, we discuss our results in Chapter 7. In this chapter, we further put our work in the context of a related work which is also developed using classical information theory based method, and compare our results with this related work. Finally, we summarize our conclusions and close by highlighting some potential avenues for further work in Chapter 8.

1.2 Impact

We have published the classical information theory based method and the MSA-specific statistical model in a scientific journal article. The quantum information theory based method is also submitted to a scientific journal article and it is currently under review.

Journal Article

[1] Gültas M, Haubrock M, Tüysüz N, Waack, S: *Coupled Mutation Finder: A new entropy-based method quantifying phylogenetic noise for the detection of compensatory mutations*. BMC Bioinformatics 2012, 13:225.

[2] Gültas M, Düzgün G, Herzog S, Jäger SJ, Meckbach C, Wingender E, Waack S: *Quantum Coupled Mutation Finder: Predicting functionally or structurally important sites in proteins using quantum Jensen-Shannon divergence and CUDA programming*. BMC Bioinformatics 2014, 15:96.

Conferences, Workshops, Meetings and Student's thesis

- Statistical and dynamical models in biology and medicine (October, 2011, Göttingen): Poster presentation
- Workshop über Algorithmen und Komplexität, 63. Theorietag (Januar, 2012, Brandennburg): Oral presentation
- German Conference of Bioinformatics (September, 2012, Jena): Poster presentation
- Meeting Gene Regulation and Information Theory (April, 2013, Halle): Poster presentation
- German Conference of Bioinformatics (September, 2013, Göttingen): Poster presentation

Furthermore, the author identified the topics with Prof. Dr. Stephan Waack for and supervised the following Project Works, Bachelor and Master's thesis.

- Thomas Franke: *Identifikation von korrelierten Mutationen auf Basis von H2r*, Project Work, 2010 - 2011

- Thomas Franke: *A New Entropy Based Model for the Detection of Correlated Mutations in Multiple Sequence Alignments*, Master Thesis, 2011
- Hendrik Kemper: *Eine neue Entropie basierte Methode für die Erkennung von kompensatorischen Mutationen: Alternative zu Coupled Mutation Finder*, Bachelor Thesis, 2011 - 2012
- Linh Dang Truong Khanh: *Feature Selection for Compensatory Mutation Analysis in MSAs using Random Forest*, Project Work, 2012 - 2013
- Projekt Interdisziplinäres Lernen und Zusammenarbeiten (PILZ): *Anwendung der Quanteninformationstheorie für die Identifizierung von kompensatorischen Mutationen*, 2012 - 2013
- On going projects
  - Linh Dang Truong Khanh: *Analysis and prediction of DNA-binding proteins from MSAs using Random Forest* , Master Thesis, 2013 -
  - Hendrik Kemper: *Prediction of functionally important amino acid positions in MSAs using classical and quantum information theory based methods*, Master Thesis, 2013 -
  - Cornelia Meckbach: *Using machine learning methods to combine classical and quantum information theory based metrics for prediction of functionally and/or structurally important amino acid positions in MSAs*, Project Work, 2013 -

2 Biological Background

In this chapter, we will give a brief introduction to the key biological concepts and techniques of bioinformatics necessary to motivate, develop, and understand functionally and/or structurally important amino acids in proteins introduced in this thesis. The descriptions in this chapter are based on [1,44–47].

2.1 Amino Acids

Amino acids are organic molecules which have at least a central carbon atom (C_α) attached to a free carboxyl group ($COOH$), a free amino group (NH_2), a hydrogen atom and a side chain group (R) (Figure 2.1). R is specific to each amino acid and known as *residue*. The amino acids differ from each other according to the chemical nature of R.

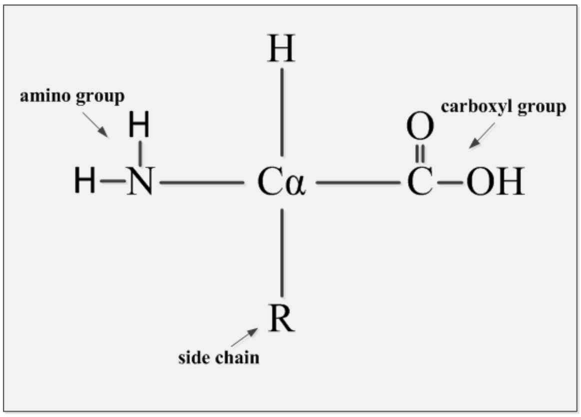


Figure 2.1: General structure of an amino acid.

There are 20 common (or primary) amino acids and each of them is found in proteins, hence they are called proteinogenic[1] amino acids. The names of all amino acids indicate their first isolated source and they are often abbreviated with three-letters and/or one-letter abbreviations (Table 2.1).

---

[1] Proteinogenic means protein building.

Full name	Three-letter abbreviation	Single-letter abbreviation
Alanine	Ala	A
Cysteine	Cys	C
Aspartic acid	Asp	D
Glutamic acid	Glu	E
Phenylalanine	Phe	F
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Lysine	Lys	K
Leucine	Leu	L
Methionine	Met	M
Asparagine	Asn	N
Proline	Pro	P
Glutamine	Gln	Q
Arginine	Arg	R
Serine	Ser	S
Threonine	Thr	T
Valine	Val	V
Tryptophan	Trp	W
Tyrosine	Tyr	Y

Table 2.1: **The amino acids:** Full names, three-letter abbreviations and single-letter abbreviations of primary amino acids.

All of the proteinogenic amino acids are vital for protein synthesis due to the optimal maintenance of body growth and function. Eleven out of the 20 amino acids: alanine, arginine, asparagine, aspartic acid, cysteine, glutamic acid, glutamine, glycine, proline, serine, and tyrosine are called *non-essential amino acids* since they are synthesized by the human body. The remaining nine amino acids, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, threonine, tryptophan, and valine are called *essential amino acids* which cannot be synthesized by the body and thus they must be taken from dietary sources [48].

Amino acids can be classified into hydrophobic and hydrophilic groups based on the physical and chemical characteristics of their residues as follows:

- Hydrophobic groups:
  - nonpolar-aliphatic residues: glycine, alanine, proline, valine, leucine, isoleucine, methionine
  - aromatic residues: phenylalanine, tyrosine, tryptophan
- Hydrophilic groups:
  - polar-charged residues: serine, threonine, cysteine, asparagine, glutamine
  - positively charged residues: lysine, arginine, histidine

- negatively charged residues: aspartate, glutamate

2.2 Proteins

Proteins are linear macromolecules made of one or more chains of amino acids arranged in a specific order. Proteins are found in all living cells and are comprised of different combinations of 20 amino acids. The individual side chain of amino acids has a protein-specific spatial arrangement which defines structure and function of the protein.

According to the structural features, proteins are usually described at four levels of complexity: primary structure, secondary structure, tertiary structure, quaternary structure.

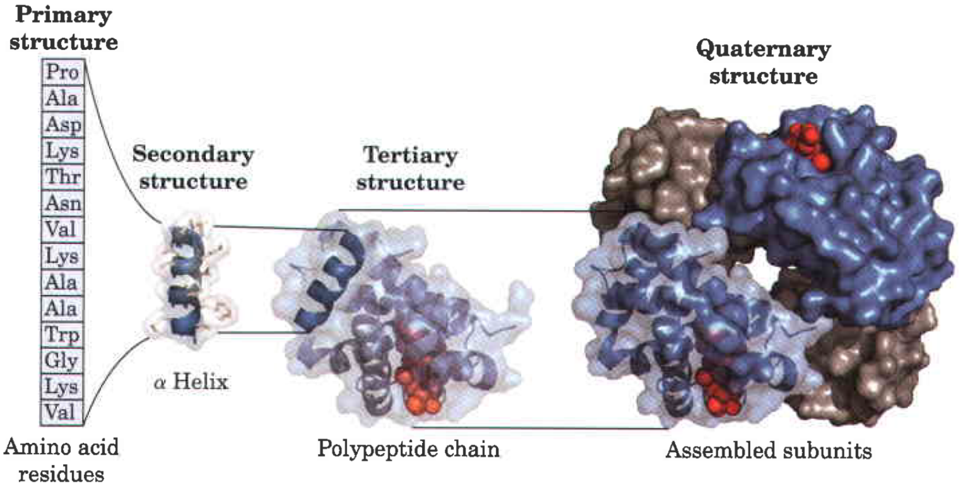


Figure 2.2: **Overview of the structural levels in proteins** (image from [[1], p.92]).

Primary Structure: The linear sequence of amino acid residues is referred to as the primary structure of the protein.

Secondary Structure: is the local conformation of the primary structure which is mainly stabilized by hydrogen bonds. There are two types of stable secondary structures: alpha helices (α-helices) and beta-sheets (β-sheets) that are preferably located at the core of the protein.

Tertiary Structure: is the final three-dimensional (3-D) structure of a protein which consists of secondary structure elements. For stabilization of the protein, very favorable residue interactions occur in this structure. Tertiary structure is unique for each protein and reflects its chemical or structural function.

Quaternary Structure: is the arrangement of two or more 3-D structures in a protein complex. The quaternary structure is also stabilized by similar interactions as in tertiary structures.

Protein Structure and Function

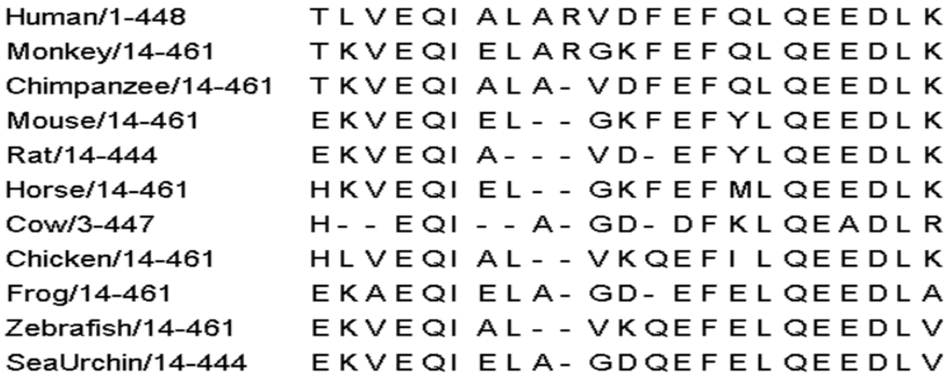
Although almost all proteins share the same structural levels, they also contain various chemical environments. Because of this, the behavior of their amino acids can be completely different. The most significant difference occurs between soluble proteins and membrane proteins. While soluble proteins tend to be surrounded by water molecules and have polar or hydrophilic amino acids on their surface, membrane proteins are surrounded by lipids and they contain on their surface the hydrophobic amino acids which are interacting with membrane [49].

The tertiary structure of soluble, membrane, and all other proteins provides essential information for the understanding of their functions. These functions often consist of the reversible binding between protein molecules and other molecules. Such molecules which are usually bound reversibly by proteins are called *ligands*. Ligands interact with proteins to bind at their specific sites hence they are called *binding sites*, which are complementary to the ligand size, shape, charge and character [49].

Another important factor for the protein function are enzymes which catalyze the reactions of molecules. After these catalysis, the molecules are called reaction substrates rather than ligands. In addition, ligand binding sites are called *catalytic sites or active sites* that are crucial for the function and activity of a protein [1,49].

2.3 Multiple Sequence Alignments

A multiple sequence alignment (MSA) of proteins contains a set of aligned amino acid sequences in which homologous residues of different sequences are placed in same columns (Figure 2.3). Therefore, functionally or structurally important amino acid positions are often strictly conserved in MSAs [4]. These conserved positions often correspond to catalytic sites, binding sites, sequence family signature, or possibly key sites for interactions within or between proteins. Such conserved positions occur because all living organisms evolved from a common ancestor. According to the common ancestor, it is assumed that all living organism and their proteins are related to each other through evolution [50]. As a result of this, the sequences of proteins are aligned in MSAs very well. In contrast, if the evolutionary relationship between aligned amino acid sequences is distant, their sequences can be aligned poorly. Consequently, for the aligning of these sequences a lot of insertions, deletions, and substitutions (see Section 2.4) are needed in order to create the corresponding MSA. Because of such insertions or deletions, beside of the common 20 amino acid a new element, called gap ('-'), is required for these process.



```
Human/1-448           TLVEQIALARVDFEFQLQEEDLK
Monkey/14-461         TKVEQIELARGKFEFQLQEEDLK
Chimpanzee/14-461     TKVEQIALA-VDFEFQLQEEDLK
Mouse/14-461          EKVEQIEL--GKFEFYLQEEDLK
Rat/14-444            EKVEQIA---VD-EFYLQEEDLK
Horse/14-461          HKVEQIEL--GKFEFMLQEEDLK
Cow/3-447             H--EQI--A-GD-DFKLQEADLR
Chicken/14-461        HLVEQIAL--VKQEFILQEEDLK
Frog/14-461           EKAEQIELA-GD-EFELQEEDLA
Zebrafish/14-461      EKVEQIAL--VKQEFELQEEDLV
SeaUrchin/14-444      EKVEQIELA-GDQEFELQEEDLV
```

Figure 2.3: A small section of a multiple sequence alignment.

A more precisely definition of an MSA is given in [51] as follows;

Definition 2.1 *Let* $A_1, \cdots, A_r$ *be* r *sequences over the alphabet of residues* Σ. *A multiple sequence alignment (MSA) of* A *is a matrix which is obtained by inserting gaps ('-') into the original sequences such that all resulting sequences* A_i^* *have equal length* L, $A_i^* = A_i$ *after removal of all gaps from* A_i^*, *and no column consist of only gaps*

$$A = \begin{cases} A_1 & = (a_{11}, a_{12}, \ldots, a_{1n_1}) \\ A_2 & = (a_{21}, a_{22}, \ldots, a_{2n_2}) \\ \quad \vdots & \\ A_r & = (a_{r1}, a_{r2}, \ldots, a_{rn_r}) \end{cases} \quad \Longrightarrow \quad MSA(A) = \begin{cases} A_1^* & = (a_{11}^*, a_{12}^*, \ldots, a_{1L}^*) \\ A_2^* & = (a_{21}^*, a_{22}^*, \ldots, a_{2L}^*) \\ \quad \vdots & \\ A_r^* & = (a_{r1}^*, a_{r2}^*, \ldots, a_{rL}^*) \end{cases}$$

MSAs are essential and one of the most important computational tools in bioinformatics because they are used almost in every application of bioinformatics, e.g.:

- to visualize and reveal the degree of evolutionary relationship between amino acid sequences,
- to determine the protein family of a newly sequenced protein with unknown structure, function, or evolutionary history in order to get more crucial information about the structure or function of this new protein,
- to predict the secondary structure of proteins, sometimes it even helps for the determination of the 3-D structure,
- to identify functional or structural sites of proteins,
- to identify domains by extracting profiles using them against corresponding databases,

- to locate DNA regulatory elements such as binding sites,
- to construct a tree for the phylogenetic analysis,
- to cluster proteins according to similar regions.

2.4 Mutations in Protein

A protein is a linear sequence over the alphabet of the 20 amino acids that are encoded by the successive triplets of letters from the DNA sequence (Table 2.2). Therefore, mutations in the DNA sequence can lead to variations or substitutions in the structure of corresponding encoded proteins. These variations in the amino acid sequences may lead to a drastic change in the functionality or structural stability of the protein and they can be responsible for diseases. For instance, substitutions at certain positions like $L858R$, $T790M$, or $G719S$ of the human epidermal growth factor receptor (EGFR) protein result in misregulation of its activity or functionality [4, 52]. EGFR is a famous example in which a small alteration affects protein function, also damages the structural stability of the EGFR protein and increases susceptibility to diseases.

Mutations are generally assigned into three main categories: substitutions, insertions, and deletions. The detailed explanation of these mutations can found in [46] which we highly recommend to interested readers. However, reasons of mutations and how they in detail occur, are out of scope of this thesis, thus we will briefly mention their short definitions.

Substitution: is also known as point mutation, which involve amino acid variations at a certain position. These variations of amino acids in the sequence may produce one of three types of mutation:

1. *Missense mutation* is also referred as non-synonymous single nucleotide polymorphisms (nsSNPs) that changes one amino acid into another amino acid in proteins. As a result of missense mutation, the amino acid sequence of encoded protein is changed and thus a residue substitution often impacts on the protein structure and also can result in a change or loss of the protein function.
2. *Nonsense mutation* changes an amino acid into a STOP codon.
3. *Frameshift mutation* is a genetic mutation caused by a deletion or insertion in a DNA. This mutation causes a change in the reading frame, leading to introduction of unrelated amino acids into the protein, generally followed by a STOP codon.

Deletions: occur if an amino acid is removed from a certain position. As a result, the positions of all the surrounding amino acids are changed.

Insertions: occur if an amino acid is added between two existing ones in sequence which have similar effects like deletions.

	Second base								
First base	**T**		**C**		**A**		**G**		**Third base**
T	TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys	T
	TTC	Phe	TCC	Ser	TAC	Tyr	TGC	Cys	C
	TTA	Leu	TCA	Ser	TAA	Stop	TGA	Stop	A
	TTG	Leu	TCG	Ser	TAG	Stop	TGG	Trp	G
C	CTT	Leu	CCT	Pro	CAT	His	CGT	Arg	T
	CTC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
	CTA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
	CTG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
A	ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser	T
	ATC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
	ATA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	A
	ATG	Met	ACG	Thr	AAG	Lys	AGG	Arg	G
G	GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly	T
	GTC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
	GTA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
	GTG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G

Table 2.2: **The standard genetic code:** Triplet combinations of nucleotides into amino acids. There are in total of 64 triplets, 61 of them encode the 20 amino acids and 3 of them are stop codons which do not code for any amino acid.

Besides of these three main categories of mutations there is another type of mutation, called *correlated mutation*, in proteins. Correlated mutations in proteins can be basically explained if one mutation in a certain position occurs within a protein, this mutation can affect a compensating mutation at another position, even if both related residue positions are distantly positioned in the protein structure. For detection of correlated mutations, the basic approach is based on functional coupling of mutation positions in protein multiple sequence alignments which often display correlations between columns [4]. This coupling between residue positions can result from spatial, physical, or chemical restrictions or signaling of allostery [4, 7, 47]. Thus, determination of these positions is as crucial as the recognition of strictly conserved positions for the understanding the structural basis of protein's function, and for the identification of functionally important residue regions which might be disease associated, responsible for the maintenance of internal protein volume, or possibly form key sites for interactions within or between proteins [4].

2.5 Bioinformatics Databases and Tools

In this section, we introduce the databases which are used in this thesis to get 3-D structures of proteins and MSAs.

2.5.1 Protein Data Bank

Protein Data Bank (PDB), at the Research Collaboratory for Structural Bioinformatics (RCSB), is the major database for the structural information of biological macromolecules such as proteins, nucleic acids, or carbohydrates. The PDB database was established by Walter Hamilton at the Brookhaven National Laboratory in 1971 and is freely accessible at http://www.rscb.org. The PDB database contains all publicly available and experimentally determined 3-D structures of approximately 90000 proteins, nucleic acids, protein/nucleic acids complexes, and other biological macromolecules. Although, a large number of database entries are 3-D structures of proteins, the protein structures are often redundant indicating same protein structures which are often observed under different conditions or experiments. In order to present structural information of proteins, PDB has a standard file format in which:

- authors who solved the structure,
- atomic coordinates,
- literature references,
- experimental details about the structure determination,
- primary and secondary structure information such as disulfide bonds, helices, sheets,
- information about binding sites, active sites, as well as hyperlinks to many other scientific databases

are included [2]. In PDB files, each of these information is presented in one line therefore each line of information is called a *record*. In addition, each PDB file of proteins possesses a name of four characters, known as PDB entry. The PDB entry can consist of either letters A to Z or digits 0 to 9 like ”1V4S.pdb” for the human glucokinase (GCK) protein.

2.5.1.1 PDB file format

There are several different types of records in PDB files, which are arranged in a specific order to describe a structure. We will briefly describe some of the most important records below. The following descriptions are based on *Protein Data Bank Contents Guide* [53] where the explanations of all records can be found in detail.

HEADER and TITLE record

```
HEADER   TRANSFERASE                                          19-NOV-03   1V4S
TITLE    CRYSTAL STRUCTURE OF HUMAN GLUCOKINASE
```

The HEADER record describes the molecule, provides the deposition date of the PDB file and repeats the PDB entry. The TITLE record describes the title of the experiment represented in the entry.

COMPND record

```
COMPND   MOL-ID: 1;
COMPND   2 MOLECULE: GLUCOKINASE ISOFORM 2
COMPND   3 CHAIN: A;
...      ...;
```

The COMPND record contains the description of macromolecular contents of an entry. This record sometimes provides information that may also be found in the TITLE record.

JRNL and REMARK record

```
JRNL     AUTH             K.KAMATA,M.MITSUYA,T.NISHIMURA,J.EIKI,Y.NAGATA
JRNL     TITL             STRUCTURAL BASIS FOR ALLOSTERIC REGULATION OF THE
JRNL     TITL 2           MONOMERIC ALLOSTERIC ENZYME HUMAN GLUCOKINASE
JRNL     REF              STRUCTURE V. 12 429 2004
JRNL     REFN             ISSN 0969-2126
JRNL     PMID             15016359
JRNL     DOI              10.1016/J.STR.2004.02.005
REMARK   1
REMARK   2 RESOLUTION.    2.30 ANGSTROMS.
REMARK   3 REFINEMENT.
REMARK   3 PROGRAM        : CNX 2002
...      ...;
```

The JRNL record contains the literature reference or publication in which the experiment has been described. However, for more than two references the REMARK record is needed.

The REMARK record was initially meant for various comments and annotations about the structure of the entry but they are currently used for all general remarks.

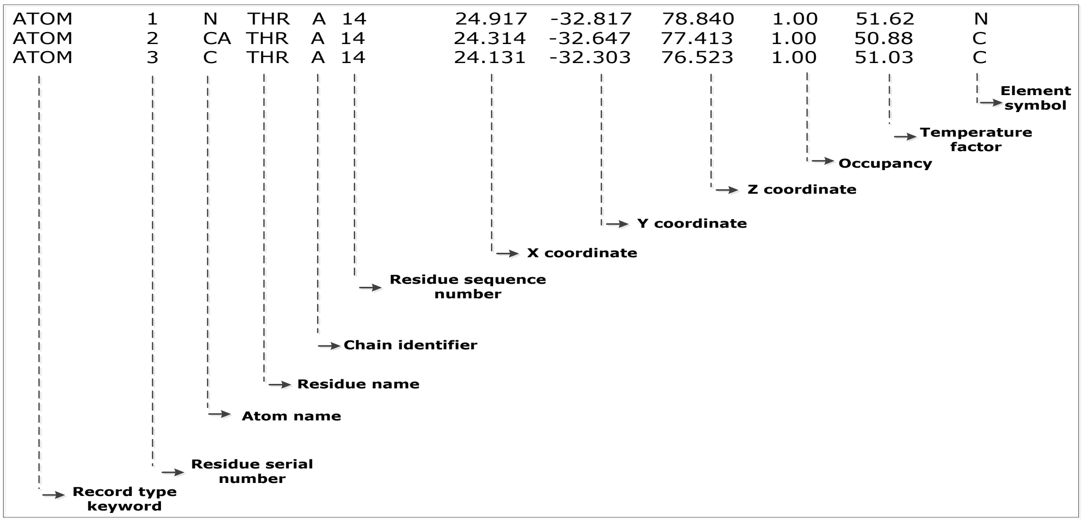
HELIX and SHEET record

The HELIX and SHEET records describe the secondary structure of the protein and polypeptide structures. The HELIX record indicates the location and type (right-handed alpha, etc.) of helices in the molecule. The SHEET record is used to identify the location, sense (anti-parallel, etc.) of the sheet in the molecule.

HELIX	1 1 THR A	14 ALA A	21 1 8
HELIX	2 2 GLU A	22 GLN A	24 5 3
…	…;		
SHEET	1 A 6 LEU A	58 ARG A	63 0
SHEET	2 A 6 ARG A	250 ASN A	254 -1
…	…		

ATOM record

The ATOM record describes the atomic coordinates containing the x,y,z orthogonal Angstrom[2] coordinates for atoms in amino acids and nucleic acids.



```
ATOM      1   N   THR  A  14       24.917  -32.817   78.840   1.00    51.62     N
ATOM      2   CA  THR  A  14       24.314  -32.647   77.413   1.00    50.88     C
ATOM      3   C   THR  A  14       24.131  -32.303   76.523   1.00    51.03     C
```

---

[2] An angstrom is a unit of measurement for very small distances. One Angstrom is equal to 10^{-10} meter and it is shown with the symbol of Å.

The third item in the ATOM record shows the atom name that consists of the chemical symbol for the atom type. The atom names begins either with ”C”, ”N”, or ”O” which indicate carbon, nitrogen, or oxygen atoms, respectively. The next character is the remoteness indicator code which is transliterated according to:

α	β	γ	δ	ε	η	ζ
A	B	G	D	E	H	Z

2.5.2 HSSP Database

The HSSP (homology-derived secondary structure of proteins) was created by Chris Sander and Reinhard Schneider which is a derived database to merge primary structure information and secondary/tertiary structure (2-D/3-D) information of proteins [54]. The HSSP database is tightly coupled with the PDB and Swiss-Prot database[3]. While the former database is used to get structural information of proteins, the latter is used to get primary structure information of proteins. According to the structure information, HSSP database provides a special HSPP file like ”1V4S.hssp” for each protein from PDB database. The HSSP file contains a multiple sequence alignment of all available homologous that are very likely to share the same 3-D structures [54].

2.5.2.1 HSSP file format

Each HSSP file includes the following four blocks: HEADERS, PROTEINS, ALIGNMENTS, and SEQUENCE PROFILE. While the HEADERS block is obligatory in every HSSP file, the other three blocks are only found if there is at least one homologous alignment. Further, the blocks are separated from one others with the string ”##” in HSSP files. Similar to the PDB files, the block information in HSSP files are presented in lines. Below, we will briefly describe these four blocks according to [54].

HEADERS block

The HEADERS block begins with the version number of HSSP software (program MaxHom) that is used to generate this file. The PDBID (Protein Data Bank Identifier) stands for the PDB entry of the test protein on which the HSSP file is based. The THRESHOLD line describes the homology threshold curve used. The REFERENCE, CONTACT, and DATE lines are already self-explanatory. The header section goes on with information about the PDB lines, including the HEADER, COMPND, SOURCE, and AUTHOR each of them have been explained in the Section 2.5.1. Further, SEQLENGTH, NCHAIN,

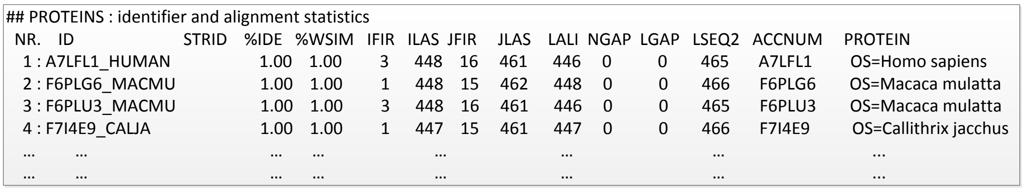
---

[3] SWISS-PROT: A protein database that stores only the primary structures of proteins [55].

HSSP	Homology derived secondary structure of proteins, version 2.0 2011
PDBID	1V4S
THRESHOLD	according to: t(L)=(290.15 * L ** -0.562) + 5
REFERENCE	Sander C., Schneider R. : Database of homology-derived protein structures. Proteins
CONTACT	Maintained at http://www.cmbi.ru.nl/ by Maarten L. Hekkelman
DATE	file generated on 2012-11-25
HEADER	TRANSFERASE 19-NOV-03 1V4S
COMPND	MOLECULE: GLUCOKINASE ISOFORM 2
SOURCE	ORGANISM-SCIENTIFIC: HOMO SAPIENS
AUTHOR	K.KAMATA,M.MITSUYA,T.NISHIMURA,J.EIKI,Y.NAGATA
DBREF	1V4S A 12 465 UNP P35557 HXK4-HUMAN 13 466
SEQLENGTH	448
NCHAIN	1 chain(s) in 1V4S data set
NALIGN	308
NOTATION :	ID: EMBL/SWISSPROT identifier of the aligned (homologous) protein

and NALIGN present the length of the sequence, number of distinct chains, and number of aligned sequences in the MSA, respectively. Finally, the NOTATION lines contain some general information about the header description of other blocks and names of the sequence database from which the aligned sequences were obtained, e.g, EMBL/SWISS-PROT or PIR/NBRF.

PROTEINS block



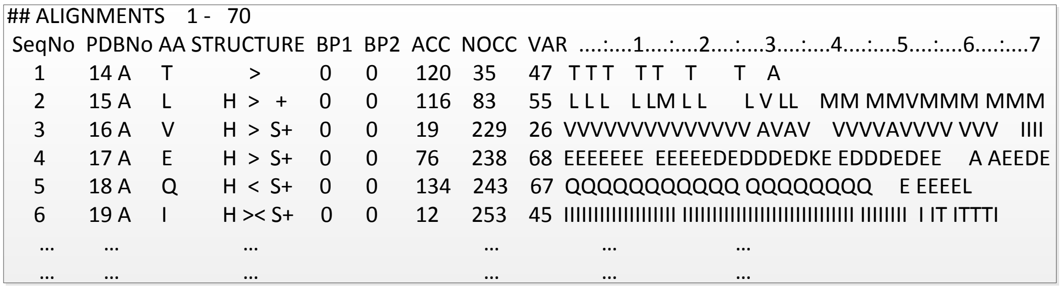
```
## PROTEINS : identifier and alignment statistics
 NR.   ID                  STRID   %IDE  %WSIM  IFIR  ILAS  JFIR   JLAS   LALI  NGAP  LGAP  LSEQ2  ACCNUM    PROTEIN
  1 : A7LFL1_HUMAN                  1.00   1.00    3    448   16    461    446   0      0    465    A7LFL1     OS=Homo sapiens
  2 : F6PLG6_MACMU                  1.00   1.00    1    448   15    462    448   0      0    466    F6PLG6     OS=Macaca mulatta
  3 : F6PLU3_MACMU                  1.00   1.00    3    448   16    461    446   0      0    465    F6PLU3     OS=Macaca mulatta
  4 : F7I4E9_CALJA                  1.00   1.00    1    447   15    461    447   0      0    466    F7I4E9     OS=Callithrix jacchus
  ...    ...                         ...    ...               ...                 ...                ...                     ...
  ...    ...                         ...    ...               ...                 ...                ...                     ...
```

The second block of an HSSP file is the PROTEINS block. This block describes the pairwise aligned data for each protein to the structurally homologous test protein. It begins with "## PROTEINS:". The explanation of headers of columns, which is already described in the NOTATION lines in HEADERS block, is following:

- NR: the line identifier,
- ID: the EMBL/SWISSPROT identifier of the aligned (homologous) protein,
- STRID: the PDB identifier of proteins with known 3-D structure,
- %IDE: the percentage of the alignment's residue identity,

- %SIM (%WSIM): the (weighted) similarity of the alignment,
- IFIR/ILAS: first and last residue of the alignment in the test sequence,
- JFIR/JLAS: first and last residue of the alignment in the aligned protein,
- LALI: length of the alignment excluding insertions and deletions,
- NGAP: number of insertions and deletions in the alignment,
- LGAP: total length of all insertions and deletions,
- LSEQ2: length of the entire sequence of the aligned protein,
- ACCNUM: Swiss-Prot accession number,
- PROTEIN: one-line description of aligned protein.

ALIGNMENTS block



```
## ALIGNMENTS    1 -   70
 SeqNo  PDBNo AA STRUCTURE  BP1  BP2  ACC  NOCC  VAR  ....:....1....:....2....:....3....:....4....:....5....:....6....:....7
    1      14 A    T          >      0    0    120   35    47  T T T    T T    T       T    A
    2      15 A    L       H  >   +  0    0    116   83    55  L L L    L LM L L      L V LL    MM MMVMMM MMM
    3      16 A    V       H  >  S+  0    0    19    229   26 VVVVVVVVVVVVVV AVAV    VVVVAVVVV VVV    IIII
    4      17 A    E       H  >  S+  0    0    76    238   68 EEEEEEE  EEEEEDEDDDEDKE EDDDEDEE     A AEEDE
    5      18 A    Q       H  <  S+  0    0    134   243   67 QQQQQQQQQQQ QQQQQQQQ     E EEEEL
    6      19 A    I       H >< S+  0    0    12    253   45 IIIIIIIIIIIIIIIIIII IIIIIIIIIIIIIIIIIIIIIIIIIIIII IIIIIIII  I IT ITTTI
   ...    ...             ...                    ...              ...              ...
   ...    ...             ...                    ...              ...              ...
```

The third section of an HSSP file is the ALIGNMENTS block which begins with "## ALIGNMENTS". In this block, family alignment details are presented residue by residue. The sequences of the test protein and the aligned database proteins are listed vertically, with the leftmost entry at the top. The descriptions of the column headers are explained in the NOTATION lines as following:

- SeqNo: the sequential residue number of the PDB protein as in DSSP [4] file,
- PDBNo: the residue number followed by the name as in the PDB file,
- AA: the amino acid type in the one-letter code,
- STRUCTURE: the summary of the secondary structure as in the DSSP file,
- BP1 and BP2: the β bridge partners as obtained from the DSSP file,
- ACC: the surface area of the residue in Å^2,

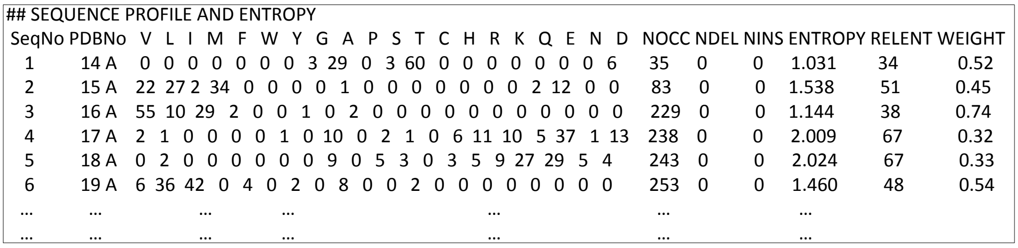
---

[4] DSSP: A database of secondary structure, solvent accessibility and other information derived from 3-D structures in the Protein Data Bank [56].

- NOCC: the number of aligned sequences at this position,
- VAR: the sequence variability derived from the NALIGN alignments.

In this block, while the pair of lower case characters in columns indicates an insertion at corresponding position in the aligned sequence, the dots $(\cdots)$ indicate deletions. In addition, each line of this block contains at most 70 residues of aligned protein sequences. If the total number of aligned sequences (NALIGN, see HEADERS block) is higher than 70 in HSSP files, the ALIGNMENTS block is divided into several intervals like $[1 \ldots 70]$, $[71 \ldots 140]$ etc. until the number of aligned sequences is reached.

SEQUENCE PROFILE block



```
## SEQUENCE PROFILE AND ENTROPY
SeqNo PDBNo  V  L  I  M  F  W  Y  G  A  P  S  T  C  H  R  K  Q  E  N  D  NOCC NDEL NINS ENTROPY RELENT WEIGHT
  1     14 A  0  0  0  0  0  0  0  3 29  0  3 60  0  0  0  0  0  0  0  6   35   0    0   1.031    34     0.52
  2     15 A 22 27 2 34  0  0  0  0  1  0  0  0  0  0  0  0  2 12  0  0    83   0    0   1.538    51     0.45
  3     16 A 55 10 29  2  0  0  1  0  2  0  0  0  0  0  0  0  0  0  0  0  229   0    0   1.144    38     0.74
  4     17 A  2  1  0  0  0  0  1  0 10  0  2  1  0  6 11 10  5 37  1 13  238   0    0   2.009    67     0.32
  5     18 A  0  2  0  0  0  0  0  0  9  0  5  3  0  3  5  9 27 29  5  4   243   0    0   2.024    67     0.33
  6     19 A  6 36 42  0  4  0  2  0  8  0  0  2  0  0  0  0  0  0  0  0   253   0    0   1.460    48     0.54
 ...    ...       ...      ...                ...                 ...          ...
 ...    ...       ...      ...                ...                 ...          ...
```

The last section of an HSSP file is the SEQUENCE PROFILE block which begins as usual with "## SEQUENCE PROFILE AND ENTROPY". In this block, the relative frequency of all 20 amino acids at each position is summarized. These frequencies are obtained by counting the amino acids in that position for all aligned sequences, including the test sequence. A frequency of 100 indicates that this position is conserved and consists of only one amino acid. The headers are also explained in the NOTATION lines.

- SeqNo: residue number that is defined in the DSSP file for the test protein (sequence),
- PDBNo: residue number that is defined in the PDB file,
- Columns 3 to 22: the relative frequencies of 20 amino acids under the one letter code,
- NOCC: number of aligned sequences with amino acids in corresponding position (including test sequence),
- NDEL: number of sequences with a deletion, relative to the test protein, at corresponding position,
- NINS: number of sequences with an insertion, relative to the test protein, at corresponding position,
- ENTROPY: measure of sequence variability at corresponding position,

- RELENT: relative entropy scaled from 0-100,
- WEIGHT: conservation weight.

2.5.2.2 HSSP to MSA

Although an HSSP file contains all of the required information about the protein of interest and its homologous, the format of HSSP files is not suitable to use it directly as an MSA. Therefore, we use the MView algorithm [57] to rewrite an HSSP file in a common MSA file format (fasta format). The MView algorithm is implemented in Perl, Version 5 for UNIX to reformat the results of a sequence database search or a multiple alignment. However, it is important to note that the MView is not a multiple alignment program [57]. For the remainder of this thesis, we use these reformatted MSAs as input for our methods.

2.5.3 Catalytic Site Atlas and PDBsum

Catalytic Site Atlas (CSA) is a database which stores experimentally validated catalytic residues in proteins. The database contains two types of validated catalytic residues. While the first one consists of hand-annotated descriptions of catalytic residues, the second type corresponds to equivalent sites in homologous proteins which are found subsequently by a sequence alignment with the original set of hand annotated entries [58]. In the database, there are currently $\sim$ 6262 annotated catalytic reside sites according to 968 literature entries. The CSA database is freely available via http://www.ebi.ac.uk/thornton-srv/databases/CSA/.

Like CSA, the PDBsum database is a part of the European Bioinformatics Institute (*EBI*)[5]. The PDBsum includes more detailed information about each experimentally determined 3-D structure of proteins in the PDB database. Moreover, the PDBsum illustrates protein structures thereby creates annotated plots for secondary structure of proteins, schematic diagrams of protein-ligand bindings and protein-DNA interactions [59, 60]. The PDBsum database is also freely accessible via http://www.ebi.ac.uk/pdbsum/ where a variety of related bionformatics databases are further presented.

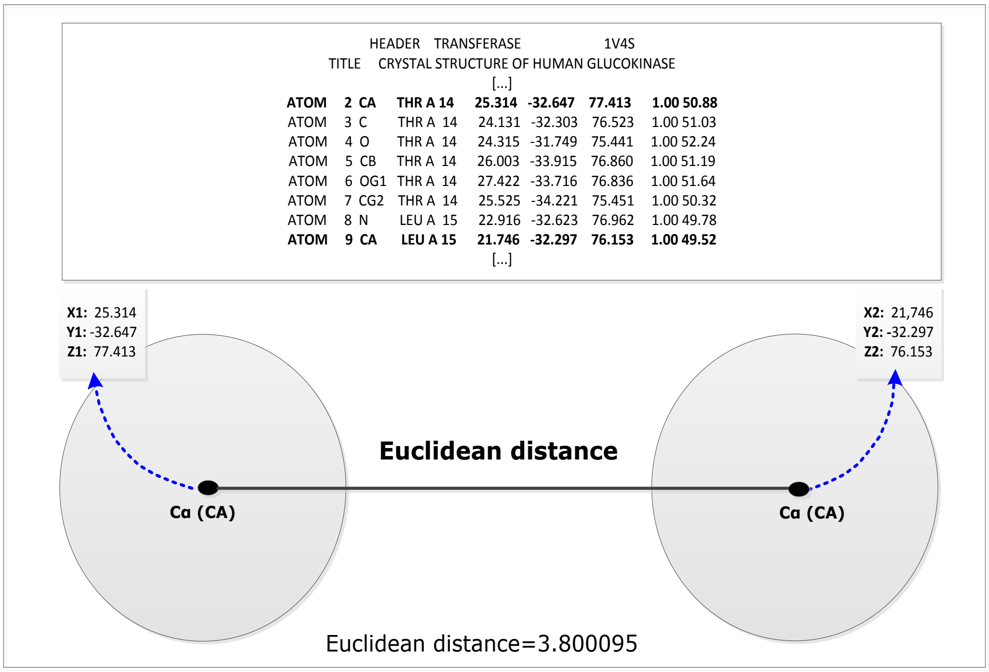
In this thesis, we have used both CSA and PDBsum databases in order to observe functionally important catalytic sites and ligand binding sites of proteins which are necessary for the biological evaluation of our methods.

2.5.4 Calculation of Distances Between Residues using BioJava

BioJava is one of the most useful open-source bioinformatics library implemented in JAVA. The BioJava library contains a variety of methods and packages each of them are very help-

---

[5] EBI is a center for research and services in bioinformatics.



```
HEADER    TRANSFERASE                 1V4S
TITLE    CRYSTAL STRUCTURE OF HUMAN GLUCOKINASE
[...]
ATOM    2  CA    THR A 14    25.314  -32.647   77.413     1.00 50.88
ATOM    3  C     THR A  14    24.131  -32.303   76.523    1.00 51.03
ATOM    4  O     THR A  14    24.315  -31.749   75.441    1.00 52.24
ATOM    5  CB    THR A  14    26.003  -33.915   76.860    1.00 51.19
ATOM    6  OG1  THR A  14    27.422  -33.716   76.836    1.00 51.64
ATOM    7  CG2  THR A  14    25.525  -34.221   75.451    1.00 50.32
ATOM    8  N      LEU A  15    22.916  -32.623   76.962    1.00 49.78
ATOM    9  CA    LEU A 15     21.746  -32.297   76.153    1.00 49.52
[...]
```

Figure 2.4: **The distance calculation:** The balls show the C_α atoms at positions 14 and 15 in human glucokinase protein with PDB entry 1V4S. Based on the x,y, and z coordinates of these atoms, the Euclidean distance is calculated as 3.800095.

ful to process biological data, to parse common file formats, and to manipulate sequences as well as 3-D structures [61].

In this thesis, we have used BioJava for parsing PDB files in order to get three dimensional coordinates of atoms in amino acids. Using these coordinates, we calculate the Euclidean distance between major carbon atoms (C_α) of different amino acids based on the ”nearby” definition of Nussinov et al. [62].

Definition 2.2 ***(Nearby amino acids)*** *Two amino acids (residues) are defined to be in contact or nearby when the distance between their (C_α) atoms is less than* $6\AA$.

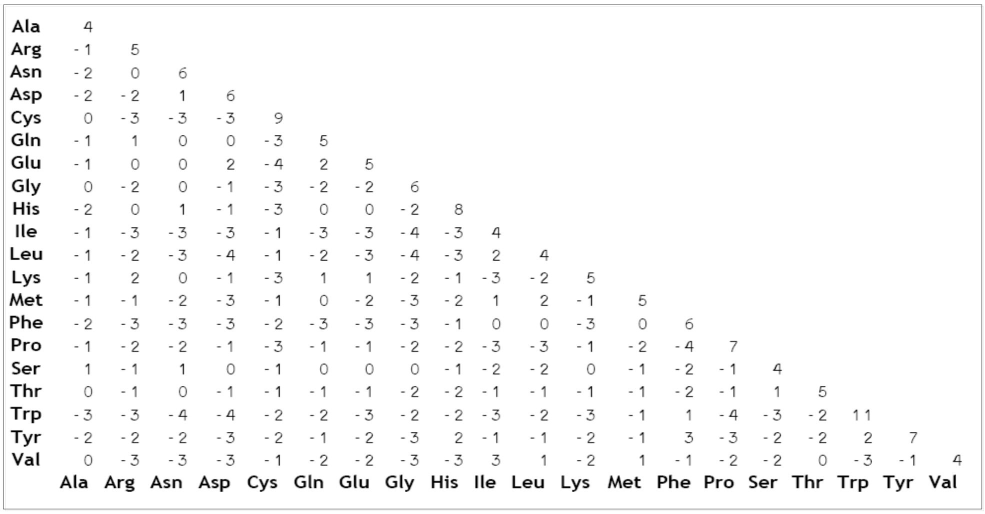
It can be assumed that an atom seems likely to be a ball with its three dimensional x,y, and z coordinates and according to these coordinates, the Euclidean distance is calculated.

Definition 2.3 ***(Euclidean distance)*** *Euclidean distance between two atoms in 3-D space with Cartesian coordinates is defined as*

$$Euclidean\ distance = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}, \tag{2.5.1}$$

where (x_1, y_1, z_1) *and* (x_2, y_2, z_2) *are the x, y and z coordinates of both atoms under study [63].*

Figure 2.4 shows the distance calculation between two C_α atoms in different amino acids of human GCK protein. However, it is important to note that in this distance calculation we do not consider the *van der Waals radius*[6].



Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
	Ala	**Arg**	**Asn**	**Asp**	**Cys**	**Gln**	**Glu**	**Gly**	**His**	**Ile**	**Leu**	**Lys**	**Met**	**Phe**	**Pro**	**Ser**	**Thr**	**Trp**	**Tyr**	**Val**

Figure 2.5: **BLOSUM62 matrix**

2.5.5 BLOSUM Matrices

The BLOSUM (BLOcks SUbstitution Matrix) is a widely used scoring matrix for sequence alignments of proteins which were first introduced by Henikoff et. al in 1992 [64]. In order to construct BLOSUM matrices, Henikoff et al. have used the BLOCKS Database [65] which contains a set of aligned and un-gapped regions from protein families. In BLOSUM

---

[6] van der Waals radius: indicates the radius of an imaginary hard sphere around an atom that is used for spatial measurements between atoms.

matrices, each substitution between amino acids is assigned a score according to their observed frequencies in the data set of homologous sequences of proteins. The substitution scores can be negative, zero, or positive. A positive score indicates that the occurrence of substitutions between similar or identical amino acids is greater than that would have been observed by random chance in the data set. In contrast, a negative score indicates that the occurrence of substitutions between dissimilar amino acids is less than that would have been observed by random chance. A score of zero means that the occurrence of weakly similar amino acids is equal to that expected by chance.

The BLOSUM matrices are 20×20 symmetric matrices and its columns and rows are indexed by the elements of amino acids (see Figure 2.5). Further, each BLOSUM matrix is denoted by BLOSUMr, where $0 < r < 100$, like BLOSUM50 or BLOSUM62 which indicates the percent identity level of sequences in the alignment. For example, for BLOSUM50, the sequences with approximately 50% identity were counted.

3 Theoretical Background

In this chapter, we provide a basic introduction to information theory. In the first section we start with the definitions of entropy, mutual information, Kullback Leibler divergence and Jensen-Shannon divergence. Most of this section is based on Chapter 2 in [66] which provides a friendly introduction to the concept of entropy. In the next section, we present some of the fundamental principles of quantum mechanics which are further required to explain the definitions of the quantum information theory. The introduced definitions and notations in this section are based on [67–71].

3.1 Classical Information Theory

The Shannon entropy is the key concept of the classical information theory and it is a measure of the average uncertainty of a random variable. Assume that we have a random variable X. The entropy of X quantifies the amount of information that we gain after we learn the value of X. Alternatively, it can be explained that the entropy of X provides a measure for the amount of uncertainty about X before we learn its value.

Definition 3.1 ***(Entropy)*** *Let X be a discrete random variable with alphabet $\mathfrak{X}$ and probability distribution $p(x) = Pr\{X = x\}$, $x \in \mathfrak{X}$, where probabilities satisfy* $0 \leq p(x) \leq 1$ *and* $\sum_{x\in\mathfrak{X}} p(x) = 1$. *The entropy of X is denoted by $\mathbb{H}(X)$ and defined as*

$$\mathbb{H}(X) = -\sum_{x\in\mathfrak{X}} p(x)\log p(x). \tag{3.1.1}$$

In this definition, the $\log(x)$ indicates the log in base two. Further, we adopt the convention that $p(x)\log p(x) = 0$, if $p(x) = 0$ for realizations with zero probability. In the Equation 3.1.1, one can easily see that the computation of $H(X)$ only depends on the probability distribution of X and it takes its maximum value when all probabilities $p(x_i)$ are equal.

Definition 3.2 ***(Joint Entropy)*** *Let X and Y two discrete random variables with alphabet $x \in \mathfrak{X}$ and $y \in \mathfrak{Y}$. Further, let $p(x) = Pr\{X = x\}$ and $p(y) = Pr\{Y = y\}$ be their individual probability distributions. The joint entropy $\mathbb{H}(X,Y)$ of X and Y with a joint probability distribution $p(x,y)$ is defined as*

$$\mathbb{H}(X,Y) = -\sum_{x}\sum_{y} p(x,y)\log p(x,y). \tag{3.1.2}$$

Definition 3.3 ***(Conditional Entropy)*** *Let X and Y be discrete random variables with joint probability distribution $p(x,y)$ and conditional distribution $p(x|y)$, then the conditional entropy $\mathbb{H}(X|Y)$ is defined as*

$$\mathbb{H}(X|Y=y) = -\sum_{x} p(x|y) \log p(x|y). \tag{3.1.3}$$

The conditional entropy $\mathbb{H}(X|Y)$ (Equation 3.1.3) can be written as

$$\begin{aligned} \mathbb{H}(X|Y) &= \sum_{y} p(y)\, \mathbb{H}(X|Y=y) \\ &= -\sum_{y} p(y) \sum_{x} p(x|y) \log p(x|y) \\ &= -\sum_{x,y} p(x,y) \log p(x|y). \end{aligned}$$

If X and Y are independent then

$$\mathbb{H}(X|Y) = \mathbb{H}(X).$$

It is important to note that the conditional entropy is not symmetric, i.e., $\exists\, X,Y$ such that $\mathbb{H}(X|Y) \neq \mathbb{H}(Y|X)$.

Theorem 3.1 ***(Chain rule for entropy:)*** *The joint entropy $\mathbb{H}(X,Y)$ of a pair of discrete random variables X and Y can be defined according to $\mathbb{H}(X)$, $\mathbb{H}(Y)$, $\mathbb{H}(X|Y)$, and $\mathbb{H}(Y|X)$ as follows*

$$\mathbb{H}(X,Y) = \mathbb{H}(X) + H(Y|X) \;\; or \;\; \mathbb{H}(Y,X) = \mathbb{H}(Y) + \mathbb{H}(X|Y).$$

The additive property of Shannon entropy can be applied as

$$\mathbb{H}(X,Y) = \mathbb{H}(X) + \mathbb{H}(Y),$$

if X and Y are completely independent.

According to the Equation 3.1.2, we can show that $\mathbb{H}(X,Y) = \mathbb{H}(Y,X)$, i.e.,

$$\mathbb{H}(X,Y) = \mathbb{H}(X) + \mathbb{H}(Y|X) = \mathbb{H}(Y) + \mathbb{H}(X|Y) = \mathbb{H}(Y,X),$$

from which the following equation is obtained as

$$\mathbb{H}(X) - \mathbb{H}(X|Y) = \mathbb{H}(Y) - \mathbb{H}(Y|X). \tag{3.1.4}$$

The Equation 3.1.4 is different but equivalent forms of mutual information which we will introduce in the next section. The venn diagram in Figure 3.1 illustrates the relationships between $\mathbb{H}(X)$, $\mathbb{H}(X)$, $\mathbb{H}(X,Y)$, $\mathbb{H}(X|Y)$ and $\mathbb{H}(Y|X)$.

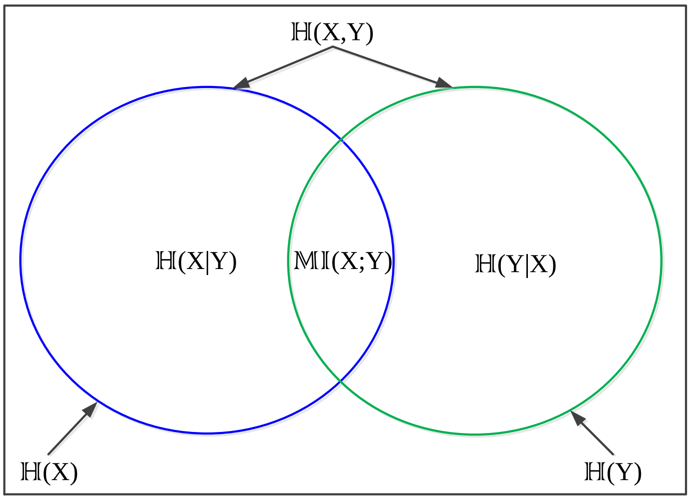


Figure 3.1: **Venn diagram visualization of entropy $\mathbb{H}(X)$, joint entropy $\mathbb{H}(X,Y)$, conditional entropy $\mathbb{H}(X|Y)$, and mutual information $\mathbb{MI}(X;Y)$.**

Theorem 3.2 *(Basic properties of Shannon entropy)*

1. *Non-negativity: The entropy is always non-negative for any $p(x)$:*

$$\mathbb{H}(X) \geq 0.$$

2. *Upper bound: The maximum value of $\mathbb{H}(X)$ for random variable X with alphabet size n is $\log(n)$:*

$$\mathbb{H}(X) \leq \log(n).$$

3. *Conditioning reduces the entropy:*

$$\mathbb{H}(X) \geq \mathbb{H}(X|Y).$$

4. *Subadditivity:*

$$\mathbb{H}(X,Y) \leq \mathbb{H}(X) + \mathbb{H}(Y).$$

   The equality is only possible if and only if the random variables X and Y are independent.

5. $\mathbb{H}(X) \leq \mathbb{H}(X,Y)$*, the equality is only possible if and only if the random variable Y is a function of X.*

6. *The entropy* $\mathbb{H}(X)$ *is concave in the probability density function* $p(x)$.

3.1.1 Relative Entropy and Mutual Information

The relative entropy is another entropic measure that quantifies the difference between two probability distributions over the same alphabet.

Definition 3.4 ***(Kullback Leibler divergence)*** *Relative entropy or Kullback Leibler (KL) divergence between two probability distributions* $p(x)$ *and* $q(x)$ *with alphabet* $\mathfrak{X}$ *is defined as*

$$KL(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = \mathbb{H}(X) - \sum_x p(x) \log q(x). \tag{3.1.5}$$

In the Equation 3.1.5, we adopt the convention that $0 \log 0 = 0$ and $-p(x) \log 0 = \infty$ if $p(x) > 0$.

Theorem 3.3 ***(Basic properties of the Kullback Leibler divergence)***

1. *Non-negativity:* $KL(p||q) \geq 0$.
2. $KL(p||q) = 0$ *if and only if* $p(x) = q(x)$.
3. *The KL divergence is not symmetric:* $KL(p||q) \neq KL(q||p)$ *if* $p(x) \neq q(x)$.
4. *The KL divergence does not satisfy the triangle inequality.*

Definition 3.5 ***(Jensen-Shannon divergence)*** *Jensen-Shannon (JS) divergence is a symmetrized, smoothed, and bounded version of the KL divergence between two (or more) probability distributions. The JS divergence between two probability distributions* $p(x)$ *and* $q(x)$ *with weights* π_1, π_2 *defined by [72] and [73] independently as*

$$JS(p(x)||q(x)) = \pi_1 KL\left(p(x)||\frac{p(x)+q(x)}{2}\right) + \pi_2 KL\left(q(x)||\frac{p(x)+q(x)}{2}\right), \tag{3.1.6}$$

where π_1 *and* π_2 *satisfy the constraints* $\pi_1 + \pi_2 = 1$, *and* $0 \leq \pi_i \leq 1$.

The *JS* divergence can also be written in the following form

$$JS(p||q) = \mathbb{H}\left(\frac{p+q}{2}\right) - \frac{1}{2}\mathbb{H}(p) - \frac{1}{2}\mathbb{H}(q) \tag{3.1.7}$$

where $\mathbb{H}(p) = -\sum_i p_i \log p_i$ is the Shannon entropy and $\pi_1 = \pi_2 = \frac{1}{2}$.

Theorem 3.4 ***(Basic properties of Jensen-Shannon divergence)***

1. *Non-negativity:* $JS(p||q) \geq 0$.

2. *$JS(p||q) = 0$ if and only if $p = q$.*

3. *The JS divergence is symmetric and always well defined: $JS(p||q) = JS(q||p)$.*

4. *The JS divergence is bounded $0 \leq JS(p||q) \leq 1$.*

5. *The square root of the JS divergence satisfy the triangle inequality but JS divergence does not.*

6. *The JS divergence can be generalized in order to quantify the difference between more than two probability distributions $p_1, p_2, \ldots, p_m$ with weights $\pi_1, \pi_2, \ldots, \pi_m$ by*

$$JS(p_1, p_2, \ldots, p_m) = \mathbb{H}\left[\sum_{i=1}^{m} \pi_i p_i\right] - \left[\sum_{i=1}^{m} \pi_i \mathbb{H}[p_i]\right], \tag{3.1.8}$$

where $\sum_{i=1}^{m} \pi_i = 1$.

Definition 3.6 ***(Mutual Information)*** *Let X and Y be two random variables with a joint probability distribution $p(x,y)$ and marginal probability distributions $p(x)$ and $p(y)$. The mutual information $\mathbb{MI}(X;Y)$ between X and Y is defined as*

$$\mathbb{MI}(X;Y) = \sum_{x}\sum_{y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \tag{3.1.9}$$

Alternatively, the $\mathbb{MI}(X;Y)$ between X and Y can be written according to $\mathbb{H}(X)$, $\mathbb{H}(Y)$ and $\mathbb{H}(X,Y)$ as

$$\mathbb{MI}(X;Y) = \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y). \tag{3.1.10}$$

The mutual information $\mathbb{MI}(X;Y)$ between two random variables X and Y is the reduction in the uncertainty of Y due to the knowledge of X (or vice versa).

Theorem 3.5 ***(Basic properties of mutual information)***

1. *Non-negativity: $\mathbb{MI}(X;Y) \geq 0$.*

2. *$\mathbb{MI}(X;Y) = 0$ if and only if X and Y are independent.*

3. *$\mathbb{MI}(X;Y)$ is symmetric: $\mathbb{MI}(X;Y) = \mathbb{MI}(Y;X)$.*

4. *$\mathbb{MI}(X;Y) = \mathbb{H}(X) - \mathbb{H}(X|Y)$.*

5. *$\mathbb{MI}(X;Y) \leq \mathbb{H}(X)$.*

3.2 Quantum Information Theory

Before we begin to define quantum information theory, we firstly have to introduce some of the basic principles and notations of quantum mechanics, which are necessary to explain and to understand quantum information theory. The definitions and notations in this section are based on [67–71]. Moreover, we do not consider in this thesis complex vector spaces and only deal with real ones.

3.2.1 Basic definitions of quantum mechanics

Quantum bits

In computer science, the elementary unit of information is represented by a *bit* which is the fundamental concept of classical computation and classical information. A bit describes the unit information of a classical system being in one of two possible states, in either the state 0 or the state 1. Like a classical bit, a *quantum bit* or short *qubit* is the elementary unit of information describing two-state system in the quantum information theory. Two possible states for a qubit are $|0\rangle$ and $|1\rangle$ which are equivalent to the states of a classical bit. The notation, ”$|.\rangle$”, called as *Dirac notation*, is the standard notation of quantum mechanics. Although classical bits and qubits look very similar at first glance, there exists a fundamental difference between them. While a classical bit can be in either the state 0 or 1, a qubit can also be in the state *superposition* which is a linear combination of the states $|0\rangle$ and $|1\rangle$,

$$|\psi\rangle = \alpha\,|0\rangle + \beta\,|1\rangle\,. \tag{3.2.1}$$

Equation 3.2.1 shows that the $|\psi\rangle$ is in a superposition where α and β are coefficients with unit norm

$$|\alpha|^2 + |\beta|^2 = 1.$$

The coefficients α and β are referred to as *quantum probability amplitudes* and their square magnitudes, $|\alpha|^2$ and $|\beta|^2$, indicate the probability of $|\psi\rangle$ for being in states $|0\rangle$ and $|1\rangle$, respectively.

Dirac notation

In quantum mechanics, Dirac notation is a very useful way of expressing unit-length vectors that are used to represent the states of a physical system. Further, the Dirac notation is especially nice to use because it provides a very helpful way of specifying vector and matrix operations of quantum states. The notation consist of *bras* "$\langle .|$" and *kets* "$|.\rangle$" therefore it is also called as "braket ($\langle .|.\rangle$) notation". Basically, a ket $|.\rangle$ indicates in this thesis a column vector of a real column vector space $\mathfrak{A}$ and a bra $\langle .|$ is obtained by transposing a ket $|.\rangle$. In quantum information theory it is usual to consider complex spaces. But for our purposes it suffices to deal with real ones. The multiplication of bras by kets results in the brakets notation "$\langle . | . \rangle$" which is also referred as *standard scalar product* or *inner product* on $\mathfrak{A}$. Suppose we have the vectors $|u\rangle$ and $|v\rangle$. Their inner product, $\langle u|v\rangle$ has the following geometric interpretation. It is equal to $\|u\|\|v\|\cos\alpha$, where α is the angle between the vectors $|u\rangle$ and $|v\rangle$, and $\|u\| = \sqrt{\langle u|u\rangle}$ and $\|v\| = \sqrt{\langle v|v\rangle}$ are the Euclidean norms (lengths) of $|u\rangle$ and $|v\rangle$.

The column vector representation of $|0\rangle$ and $|1\rangle$ is denoted as

$$|0\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \tag{3.2.2}$$

$$|1\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{3.2.3}$$

The bras corresponding to the kets $|0\rangle$ and $|1\rangle$ can be shown as

$$\langle 0| = \begin{pmatrix} 1 & 0 \end{pmatrix}, \tag{3.2.4}$$

$$\langle 1| = \begin{pmatrix} 0 & 1 \end{pmatrix}. \tag{3.2.5}$$

Although a bra does not represent quantum states, it is required for performing calculations like probability amplitudes in the quantum theory. Recall the qubit in the Equation 3.2.1, in order to determine the probability of $|\psi\rangle$ for being in state $|0\rangle$, we will combine the state of the $|\psi\rangle$ with the bra $\langle 0|$ in the following calculation as

$$\begin{aligned}
\langle 0|\psi\rangle &= \langle 0|\left(\alpha|0\rangle + \beta|1\rangle\right) && (3.2.6)\\
&= \alpha\langle 0||0\rangle + \beta\langle 0||1\rangle && (3.2.7)\\
&= \alpha\begin{pmatrix} 1 & 0 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} + \beta\begin{pmatrix} 0 & 1 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \end{pmatrix} && (3.2.8)\\
&= \alpha\cdot 1 + \beta\cdot 0 && (3.2.9)\\
&= \alpha. && (3.2.10)
\end{aligned}$$

Likewise, calculating the quantity $\langle 1|\psi\rangle$ we can determine the probability of $|\psi\rangle$ for being in states $|1\rangle$. Further, we can also calculate the quantities $\langle 0|1\rangle$ and $\langle 1|0\rangle$ as

$$\langle 0|1\rangle = 0, \tag{3.2.11}$$

$$\langle 1|0\rangle = 0, \tag{3.2.12}$$

which show the probability of $|0\rangle$ for being in state $|1\rangle$ and the probability of $|1\rangle$ for being in state $|0\rangle$. Since the vectors $|0\rangle$ and $|1\rangle$ are *orthogonal* to each other and have no overlap, their inner products are equal to zero. In contrast to this, the inner products $\langle 0|0\rangle$ and $\langle 1|1\rangle$

are equal to one,

$$\langle 0|0\rangle = 1, \tag{3.2.13}$$

$$\langle 1|1\rangle = 1. \tag{3.2.14}$$

Outer product

In contrast to the *inner product*, the product of a ket $|u\rangle$ with a bra $\langle v|$ results in a matrix. It is defined as the *outer product* which is used to represent the density operator of a system.

Let $|u\rangle$ and $|v\rangle$ be two qubits given by

$$|u\rangle = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} \text{ and, } |v\rangle = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$$

The outer product is an expression like $|u\rangle\langle v|$ and calculated by

$$|u\rangle\langle v| = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} (b_0, b_1) = \begin{pmatrix} a_0 b_0 & a_0 b_1 \\ a_1 b_0 & a_1 b_1 \end{pmatrix}. \tag{3.2.15}$$

We can perform the bras $\langle .|$ and kets $|.\rangle$ to represent this matrix in Dirac notation as

$$|u\rangle\langle v| = a_0 b_0 |0\rangle\langle 0| + a_0 b_1 |0\rangle\langle 1| + a_1 b_0 |1\rangle\langle 0| + a_1 b_1 |1\rangle\langle 1|. \tag{3.2.16}$$

Recall the qubit in the Equation 3.2.1, combining $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ with the matrix $|u\rangle\langle v|$, we determine the effect of this matrix on $|\psi\rangle$ as

$$(|u\rangle\langle v|)(|\psi\rangle) = |u\rangle\langle v|\psi\rangle = \langle v|\psi\rangle |u\rangle. \tag{3.2.17}$$

This means that $|u\rangle\langle v|$ is a projection of $\mathfrak{A}$ onto the subspace spanned by $|u\rangle$. Every linear operator on $\mathfrak{A}$ can be represented by a linear combination of such products. The transpose A^T of a linear operator $A = \sum_{ij} a_{ij} |u_i\rangle\langle v_j|$ equals $\sum_{ij} a_{ij} |v_j\rangle\langle u_i|$.

Tensor product

The *tensor product* is a fundamental mathematical operation of combining vector spaces to form a new larger vector space. Suppose we have two two-dimensional vectors:

$$|u\rangle = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \text{ and } |v\rangle = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

The tensor product of $|u\rangle \otimes |v\rangle$ is

$$|u\rangle \otimes |v\rangle = |uv\rangle = \binom{a_1}{a_2} \otimes \binom{b_1}{b_2} = \begin{pmatrix} a_1b_1 \\ a_1b_2 \\ a_2b_1 \\ a_2b_2 \end{pmatrix}.$$

We can also calculate the tensor product of qubit states $|0\rangle$ and $|1\rangle$ using their vector representations:

$$|00\rangle = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \ |01\rangle = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad |10\rangle = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \ |11\rangle = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

Likewise, we can derive the tensor product of two matrices that is referred to as the *Kronecker product*. Suppose we have two operators $U = |u_1\rangle\langle u_2|$ and $V = |v_1\rangle\langle v_2|$ and their matrix representations are given by

$$U = |u_1\rangle\langle u_2| = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \quad \text{and} \quad V = |v_1\rangle\langle v_2| = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}.$$

The matrix representation of the tensor product $U \otimes V$ can be written as

$$|u_1\rangle\langle u_2| \otimes |v_1\rangle\langle v_2| = |u_1\rangle\langle v_1| \otimes |u_2\rangle\langle v_2| = U \otimes V,$$

and calculated as

$$U \otimes V = \begin{pmatrix} a_{11}V & a_{12}V \\ a_{21}V & a_{22}V \end{pmatrix} = \begin{pmatrix} a_{11}b_{11} & a_{11}b_{12} & a_{12}b_{11} & a_{12}b_{12} \\ a_{11}b_{21} & a_{11}b_{22} & a_{12}b_{21} & a_{12}b_{22} \\ a_{21}b_{11} & a_{21}b_{12} & a_{22}b_{11} & a_{22}b_{12} \\ a_{21}b_{21} & a_{21}b_{22} & a_{22}b_{21} & a_{22}b_{22} \end{pmatrix}.$$

In general, if $|u_1\rangle, |u_2\rangle \in \mathfrak{A}_1$ and $|v_1\rangle, |v_2\rangle \in \mathfrak{A}_2$ are states, then $|u_1v_1\rangle = |u_1\rangle \otimes |v_1\rangle \in \mathfrak{A}_1 \otimes \mathfrak{A}_2$ and $|u_2v_2\rangle = |u_2\rangle \otimes |v_2\rangle \in \mathfrak{A}_1 \otimes \mathfrak{A}_2$ are tensor products of column vectors. The tensor products of row vectors $\langle u_1v_1| = \langle u_1| \otimes \langle v_1|$ and $\langle u_2v_2| = \langle u_2| \otimes \langle v_2|$ are the transposes of $|u_1v_1\rangle$ and $|u_2v_2\rangle$, where $\langle u_1v_1|u_2v_2\rangle = \langle u_1|v_2\rangle\langle u_1|v_2\rangle$.

The Density operator or matrix

The traditional quantum mechanics distinguishes between *pure states* and *mixed states*: pure state of a quantum system is a vector ($|.\rangle$) with a unit length in a vector space (the quantum states which we have considered until now are also pure states); mixed state is a statistical mixture of two or more states. While the pure states are represented by state vectors, mixed states are described by *density matrices*. A density matrix ρ is an operator representing a quantum state that describes a part of the composite system.

Suppose we have the following two state vectors:

$$\begin{aligned} |u\rangle &= \alpha\,|x\rangle + \beta\,|y\rangle\,, && (3.2.18)\\ |v\rangle &= \gamma|x\rangle + \delta\,|y\rangle\,. && (3.2.19)\end{aligned}$$

The density matrix for each of these states is the outer product of the state vector with itself:

$$\begin{aligned} \rho_u &= |u\rangle\,\langle u|\,, && (3.2.20)\\ \rho_v &= |v\rangle\,\langle v|\,. && (3.2.21)\end{aligned}$$

Now, we can generalize the density matrix for n possible states. Suppose we have n states and a quantum system is in one of these states $|\psi_i\rangle$ where $i = 1, 2, \ldots, n$ with respective probabilities p_i. The notation $\{p_i, |\psi_i\rangle\}$ is called an *ensemble of pure states*. Then the density matrix for the entire system is defined as

$$\rho = \sum_{i=1}^{n} p_i\,|\psi_i\rangle\,\langle\psi_i|\,. \tag{3.2.22}$$

Definition 3.7 ***(Eigenvalues and Eigenstates of the density matrix)*** *Given a density matrix ρ on a vector space, a non-zero state vector $|\psi\rangle$ is defined as* eigenstate *of ρ if the following equation is satisfied*

$$\rho\,|\psi\rangle = \lambda\,|\psi\rangle\,, \tag{3.2.23}$$

where λ is called an eigenvalue *of ρ.*

In order to determine eigenvalues of a density matrix ρ, the characteristic equation $det|\rho - \lambda I| = 0$ is used, where *det* indicates the determinant of the matrix $\rho - \lambda I$ and I denotes the identity matrix. The values of λ are the eigenvalues which represent the solutions of this characteristic equation.

Theorem 3.6 ***(Key properties of a Density operator)*** *An operator ρ is called as a density operator if and only if it satisfies the following properties:*

1. *The density operator ρ is Hermitian*[7]*:*

$$\rho^\dagger = (\sum_i p_i\,|\psi_i\rangle\,\langle\psi_i|)^\dagger = \sum_i p_i\,|\psi_i\rangle\,\langle\psi_i| = \rho.$$

---

[7] A Hermitian matrix is a matrix which is equal to its own transpose.

2. *ρ is positive semi-definite for any state vector $|\psi\rangle$:*

$$\langle\psi|\rho|\psi\rangle \geq 0.$$

3. *ρ has unit trace: $Tr(\rho) = 1$.*

Entangled and Separable States

One of the most striking features of quantum mechanics is the fact that the systems can become *entangled*. Suppose we have two systems A and B and their composite state is given by a density matrix ρ_{AB}. We define these two systems to be entangled if they have interacted sometime in the past and we cannot construct the individual systems again now. According to this interaction, the values of certain properties of system A are correlated with the values that those properties will assume for system B [71]. In contrast, we say that two system are *separable* if they are not entangled indicating that the corresponding composite state ρ_{AB} can be broken into tensor product of these two systems.

More precisely, suppose we have two density operators as

$$\rho_A = \sum_{i,j=1}^{n} \alpha_{ij} |i\rangle \langle j| \text{ and } \rho_B = \sum_{k,l=1}^{n} \beta_{kl} |k\rangle \langle l|.$$

Their tensor product

$$\rho_A \otimes \rho_B = \sum_{i,j,k,l=1}^{n} \alpha_{ij}\beta_{kl} |ik\rangle \langle jl| \tag{3.2.24}$$

is a separable density operator since it is the probability distributions over products of the kind

$$\sum_m p_m \rho_{A,m} \otimes \rho_{B,m} \ (p_m > 0, \ \sum_m p_m = 1). \tag{3.2.25}$$

The reduced density operator: Partial Trace

Suppose we have two systems A and B and their composite state is given by a density operator ρ_{AB} as

$$\rho_{AB} = \sum_{i,j,k,l=1}^{n} \gamma_{ikjl} |ik\rangle \langle jl|. \tag{3.2.26}$$

We define the *reduced density operator* of the system A and B as

$$\rho_A = \quad tr_B(\rho_{AB}) = \quad tr_B(\rho_A \otimes \rho_B), \tag{3.2.27}$$

$$\rho_B = \quad tr_A(\rho_{AB}) = \quad tr_A(\rho_A \otimes \rho_B), \tag{3.2.28}$$

where tr_A and tr_B are the *partial traces* over system A and B, respectively. The partial traces tr_A and tr_B are defined by

$$\mathrm{tr}_A(\rho) = \sum_{k,l=1}^{n} \left(\sum_{i=1}^{20} \gamma_{ikil}\right) |k\rangle \langle l|, \qquad \mathrm{tr}_B(\rho) = \sum_{i.j=1}^{n} \left(\sum_{k=1}^{20} \gamma_{ikjk}\right) |i\rangle \langle j|. \tag{3.2.29}$$

It is important to note that partial traces of density operators are also density operators.

3.2.2 Quantum Information and von Neumann Entropy

In this section, we will consider the basis of quantum information theory. We will begin by defining the von Neumann Entropy of a density operator ρ that is the quantum analogue of the Shannon entropy $\mathbb{H}$. After that, we will go on with quantum definitions of relative entropy, joint entropy, mutual information, and Jensen-Shannon divergence [67–71].

Definition 3.8 ***(Quantum Entropy: von Neumann Entropy)*** *Let ρ be a density operator on a n-dimensional space. The von Neumann entropy of ρ is denoted by $VNE(\rho)$ and defined as*

$$VNE(\rho) = -tr(\rho \log \rho) = -\sum_{i=1}^{n} \lambda_i \log \lambda_i, \tag{3.2.30}$$

where $\lambda_1, \ldots, \lambda_n$ are eigenvalues of ρ and we define $0 \log 0 = 0$.

Definition 3.9 ***(Joint Quantum Entropy)*** *Let A, B be two subsystems and AB be their composite system. Let ρ_{AB}, ρ_A, ρ_B be corresponding density operators of the system and its subsystems. Quantum joint entropy is defined as*

$$VNE(\rho_{AB}) = -tr(\rho_{AB} \log \rho_{AB}). \tag{3.2.31}$$

Definition 3.10 ***(Quantum Mutual Information)*** *The quantum mutual information of a composite system AB defined as*

$$VNE(\rho_A;\rho_B) = VNE(\rho_A) + VNE(\rho_B) - VNE(\rho_{AB}). \tag{3.2.32}$$

Theorem 3.7 ***(Basic properties of quantum entropy)***

1. *Non-negativity: The quantum entropy is non-negative for any density operator ρ_A:*

$$VNE(\rho_A) \geq 0,$$

where the equality is only possible if and only if ρ_A is a pure state.

2. *Upper bound: The maximum value of $VNE(\rho_A)$ for a density operator ρ_A is $\log(n)$, where n is the dimension of the system:*

$$VNE(\rho_A) \leq \log(n).$$

3. *If a composite system ρ_{AB} is in a pure state:*

$$VNE(\rho_A) = VNE(\rho_B).$$

4. *Subadditivity:*

$$VNE(\rho_{AB}) \leq VNE(\rho_A) + VNE(\rho_B).$$

The equality is only possible if the composite system ρ_{AB} equal to the tensor product of two subsystems

$$VNE(\rho_A \otimes \rho_B) = VNE(\rho_A) + VNE(\rho_B),$$

which corresponds to two subsystems with uncorrelated information.

Definition 3.11 ***(Quantum Relative entropy)*** *The quantum relative entropy between two density operators ρ and σ is defined, analogously to the corresponding classical quantity, as*

$$R\,VNE(\rho||\sigma) = tr(\rho \log \rho) - tr(\rho \log \sigma). \tag{3.2.33}$$

Definition 3.12 ***(Quantum Jensen-Shannon divergence)*** *Similar to the classical Jensen-Shannon divergence, the quantum Jensen-Shannon divergence $(QJSD(\rho||\sigma))$ between two density operators ρ and σ is a symmetrized and smoothed version of quantum relative entropy. The $(QJSD(\rho||\sigma))$ is defined as*

$$QJSD(\rho||\sigma) = \frac{1}{2}\left[R\,VNE\left(\rho||\frac{\rho+\sigma}{2}\right) + R\,VNE\left(\sigma||\frac{\rho+\sigma}{2}\right)\right]. \tag{3.2.34}$$

The Equation 3.2.34 can be rewritten in terms of the von Neumann entropy as

$$QJSD(\rho||\sigma) = \mathrm{VNE}\left(\frac{\rho+\sigma}{2}\right) - \frac{1}{2}\mathrm{VNE}(\rho) - \frac{1}{2}\mathrm{VNE}(\sigma). \tag{3.2.35}$$

3.3 Mathematical Foundations

In following we briefly introduce the definition of the β-distribution and its parameter estimation which is necessarily for the development of our significant model. The descriptions in this section are based on [74].

Definition 3.13 ***(β-distribution)*** *A continuous random variable X has a β-distribution over the interval $[0,1]$ with positive shape parameters α and β if its density function is defined as*

$$f_X(x) = \frac{1}{B(\alpha,\beta)} x^{(\alpha-1)(1-\beta)^{(\beta-1)}}, \quad 0 \leq x \leq 1. \tag{3.3.1}$$

In Equation 3.3.1 $B(.,.)$ is the β-function and it is calculated as

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt, \quad \alpha > 0 \text{ and } \beta > 0. \tag{3.3.2}$$

The mean μ and variance σ^2 of the β-distribution are

$$\mu = \frac{\alpha}{\alpha+\beta}, \tag{3.3.3}$$

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{3.3.4}$$

In addtion, the cumulative distribution function of β-distribution is given as

$$F_X(x) = \frac{1}{B(\alpha,\beta)} = \int_0^x t^{\alpha-1}(1-t)^{\beta-1}dt, \quad 0 \leq x \leq 1 \text{ and } \alpha,\ \beta > 0. \tag{3.3.5}$$

For a given sample mean $\hat{\mu}$ and variance $\hat{\sigma}$ of finite sample of size N, the shape parameters α and β of a beta distribution can be estimated as

$$\hat{\alpha} = \hat{\mu}\left(\frac{\hat{\mu}(1-\hat{\mu})}{\hat{\sigma}} - 1\right), \tag{3.3.6}$$

$$\hat{\beta} = (1-\hat{\mu})\left(\frac{\hat{\mu}(1-\hat{\mu})}{\hat{\sigma}} - 1\right), \tag{3.3.7}$$

where $\hat{\mu}$ and $\hat{\sigma}$ can be calculated using following equations as

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{3.3.8}$$

$$\hat{\sigma} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2. \tag{3.3.9}$$

(3.3.10)

4 Applying Classical Information Theory for the Compensatory Mutation Analysis

In this chapter, we will present two entropy-based metrics, namely $\mathbb{U}$-metric and $\mathbb{U}_{D(\alpha)}$-metric, for the detection of compensatory mutations in MSAs. While the former one was developed by Merkl et. al in [22], the latter has been developed during this thesis in order to complete the $\mathbb{U}$-metric. Further, we will present a new MSA-specific statistical model based on multiple testing procedures in order to determine significant $\mathbb{U}$ and $\mathbb{U}_{D(\alpha)}$-values, respectively.

We previously published an important part of this chapter in [4] (see Appendix A 9.1). The descriptions of both the $\mathbb{U}$-metric and the $\mathbb{U}_{D(\alpha)}$-metric as well as the new MSA-specific statistical model approach are based on this publication.

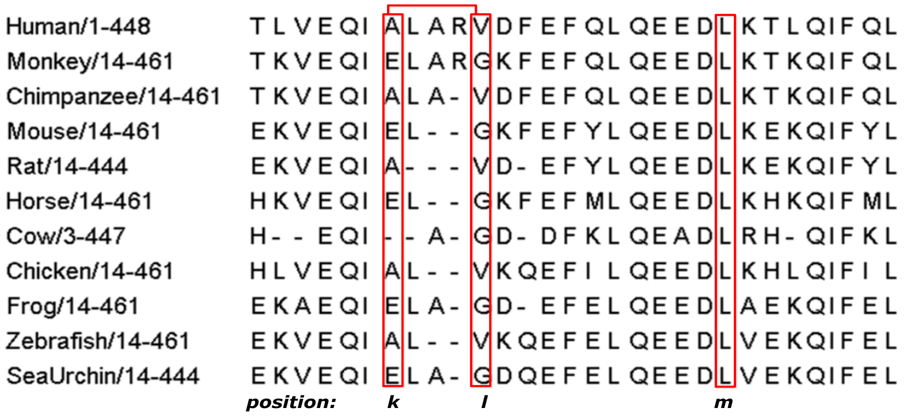
4.1 Detecting compensatory mutations by the $\mathbb{U}$-metric

Let M be an arbitrarily chosen but fixed MSA for the protein under investigation. In order to determine the correlation between columns of M, Merkl et. al. used in [22] a normalized measure of mutual information ranging over the interval $[0,1]$. It is denoted as $\mathbb{U}$-metric and defined as

$$\mathbb{U}(k,l) = 2 \cdot \frac{\mathbb{H}(k) + \mathbb{H}(l) - \mathbb{H}(k,l)}{\mathbb{H}(k) + \mathbb{H}(l)} = 2 \cdot \frac{\mathbb{MI}(k;l)}{\mathbb{H}(k) + \mathbb{H}(l)}, \tag{4.1.1}$$

where $\mathbb{H}(k)$ and $\mathbb{H}(l)$ are the entropy of the empirical amino acid distributions of the columns k and l under study, and $\mathbb{H}(k,l)$ is their joint entropies. The empirical distributions of amino acids are observed based on the occurrence of each amino acid in columns k and l (see Figure 4.1). Likewise, extending the concept of empirical amino acid distributions, we can observe the empirical joint distributions of amino acid pairs in the column pair (k,l).

It is important to note that during the observation of these amino acid distributions we only take into account the standard 20 amino acids and do not consider the gaps ('-') as a real component of the protein. Hence, we exclude them when we determine distributions of amino acids in columns or in column pairs.



Human/1-448	TLVEQI ALARVDFEFQLQEEDLKTLQIFQL
Monkey/14-461	TKVEQI ELARGKFEFQLQEEDLKTKQIFQL
Chimpanzee/14-461	TKVEQI ALA-VDFEFQLQEEDLKTKQIFQL
Mouse/14-461	EKVEQI EL--GKFEFYLQEEDLKEKQIFYL
Rat/14-444	EKVEQI A---VD-EFYLQEEDLKEKQIFYL
Horse/14-461	HKVEQI EL--GKFEFMLQEEDLKHKQIFML
Cow/3-447	H--EQI --A-GD-DFKLQEADLRH-QIFKL
Chicken/14-461	HLVEQI AL--VKQEFILQEEDLKHLQIFIL
Frog/14-461	EKAEQI ELA-GD-EFELQEEDLAEKQIFEL
Zebrafish/14-461	EKVEQI AL--VKQEFELQEEDLVEKQIFEL
SeaUrchin/14-444	EKVEQI ELA-GDQEFELQEEDLVEKQIFEL

Figure 4.1: **A small part of an MSA.** The positions k and l show a strong correlation. At position k the $[A$ to $E]$ mutation cause at position l the $[V$ to $G]$ mutation or vice versa. In addition, the position m demonstrates an example for strictly conserved residues.

As mentioned above, the $\mathbb{U}(k,l)$-value of a column pair (k,l) ranges from 0 to 1. According to the Figure 3.1, one can easily see that if the columns k and l are completely dependent, the individual column entropies are equal to their corresponding joint entropy, i.e.,

$$\mathbb{H}(k) = \mathbb{H}(l) = \mathbb{H}(k,l). \tag{4.1.2}$$

In this case, we have $\mathbb{U}(k,l) = 1$, which is the upper bound of the $\mathbb{U}$-metric. On the other hand, if both columns are completely independent, we can see according to the Figure 3.1 that the sum of individual column entropies corresponds to the their joint entropy, i.e.,

$$\mathbb{H}(k,l) = \mathbb{H}(k) + \mathbb{H}(l), \tag{4.1.3}$$

where $\mathbb{U}(k,l) = 0$, which is the lower bound of the $\mathbb{U}$-metric.

Calculation of the column entropies

Since we utilize the concept of information theory, it is necessary to consider amino acids as a set of a random variable in columns of MSAs. Like in Figure 4.1, let k and l be two columns and n is the number of sequences in the MSA M under study. Further, we assume that X and Y are two random variables with same alphabet $\mathfrak{X}$, where the alphabet size corresponds to 20 amino acids (throughout the remainder of this thesis unless otherwise noted, the alphabet size is always 20). The random variables X and Y characterize the frequencies of amino acids in given columns k and l in M, respectively.

According to this, we can write associated empirical amino acid distributions of each columns as $\hat{p}(x) = Pr\{X = x\},\ \hat{p}(y) = Pr\{Y = y\},\ x,y \in \mathfrak{X}$. Further, the empirical joint amino acid pair distribution is denoted by $\hat{p}(x_i,y_i) = \hat{p}\{X = x, Y = y\}$. Then, the marginal distributions $\hat{p}(x_i)$, $\hat{p}(y_i)$ and joint distribution $\hat{p}(x_i,y_i)$ are calculated as

$$\hat{p}(x_i) = \frac{\#(x_i)}{n}, \tag{4.1.4}$$

$$\hat{p}(y_i) = \frac{\#(y_i)}{n}, \tag{4.1.5}$$

$$\hat{p}(x_i,y_i) = \frac{\#(x_i,y_i)}{n}, \tag{4.1.6}$$

where $\#(x_i)$, $\#(y_i)$, and $\#(x_i,y_i)$ are observed frequencies of amino acids (x_i), (y_i) and pair of amino acids (x_i,y_i) in the given columns k and l and in column pairs (k,l).

Using the Equations 4.1.4 to 4.1.6, we calculate the corresponding column entropies $\mathbb{H}(X)$, $\mathbb{H}(Y)$, and their joint entropy $\mathbb{H}(X,Y)$, respectively.

$$\mathbb{H}(X) = -\sum_{i=1}^{20} \hat{p}(x_i) \log \hat{p}(x_i) \tag{4.1.7}$$

$$\mathbb{H}(Y) = -\sum_{j=1}^{20} \hat{p}(y_j) \log \hat{p}(y_j) \tag{4.1.8}$$

$$\mathbb{H}(X,Y) = -\sum_{i=1}^{20} \sum_{j=1}^{20} \hat{p}(x_i,y_j) \log \hat{p}(x_i,y_j) \tag{4.1.9}$$

Then, we calculate the

$$\mathbb{MI}(X;Y) = -\sum_{i=1}^{20} \hat{p}(x_i) \log \hat{p}(x_i) - \sum_{j=1}^{20} \hat{p}(y_j) \log \hat{p}(y_j) + \sum_{i=1}^{20} \sum_{j=1}^{20} \hat{p}(x_i,y_j) \log \hat{p}(x_i,y_j), \tag{4.1.10}$$

which can be rewritten as:

$$\mathbb{MI}(X;Y) = \sum_{i=1}^{20} \sum_{j=1}^{20} \hat{p}(x_i,y_j) \log \frac{\hat{p}(x_i,y_j)}{\hat{p}(x_i)\hat{p}(y_j)}. \tag{4.1.11}$$

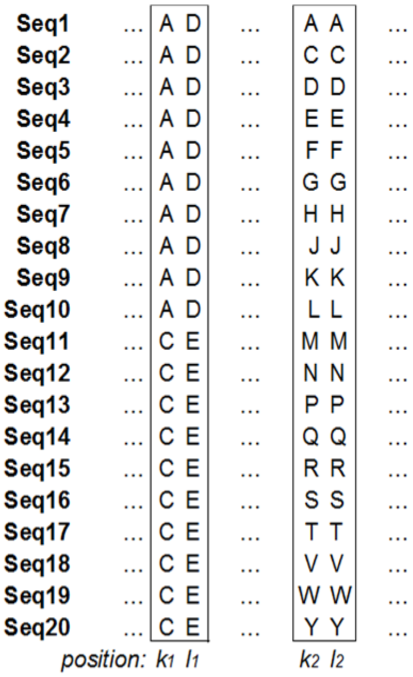
Finally, we determine the correlation between corresponding columns with $\mathbb{U}$-metric as

$$\mathbb{U}(X,Y) = 2 \cdot \frac{\mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y)}{\mathbb{H}(X) + \mathbb{H}(Y)} = 2 \cdot \frac{\mathbb{MI}(X;Y)}{\mathbb{H}(X) + \mathbb{H}(Y)}.$$

The higher the $\mathbb{U}(X,Y)$ value between columns k and l, the stronger the pair-wise co-occurrence of amino acids at these columns.

The alphabet size effect on the mutual information

The mutual information ($\mathbb{MI}$) is a useful method in bioinformatics in order to measure the correlation between two columns in MSAs. However, Martin et al. [8] have argued and showed in detail that the normalizing $\mathbb{MI}$ values with the entropy makes them more suitable than raw $\mathbb{MI}$ values for the correlated/compensatory mutation analysis. This is due to possible differences in the alphabet size of a column.



Seq1	...	A	D	...	A	A	...
Seq2	...	A	D	...	C	C	...
Seq3	...	A	D	...	D	D	...
Seq4	...	A	D	...	E	E	...
Seq5	...	A	D	...	F	F	...
Seq6	...	A	D	...	G	G	...
Seq7	...	A	D	...	H	H	...
Seq8	...	A	D	...	J	J	...
Seq9	...	A	D	...	K	K	...
Seq10	...	A	D	...	L	L	...
Seq11	...	C	E	...	M	M	...
Seq12	...	C	E	...	N	N	...
Seq13	...	C	E	...	P	P	...
Seq14	...	C	E	...	Q	Q	...
Seq15	...	C	E	...	R	R	...
Seq16	...	C	E	...	S	S	...
Seq17	...	C	E	...	T	T	...
Seq18	...	C	E	...	V	V	...
Seq19	...	C	E	...	W	W	...
Seq20	...	C	E	...	Y	Y	...
	position:	k_1	l_1		k_2	l_2	

Figure 4.2: **An artificial MSA.** The positions k_1 and l_1 are slightly conserved and contain only two amino acids. The positions k_2 and l_2 contain all 20 amino acids. Since the computation of $\mathbb{MI}$ values strongly depend on the empirical amino acid distributions of the columns which is also related to the alphabet size, both pairs of columns (k_1, l_1) and (k_2, l_2) have completely different $\mathbb{MI}$ values. The more different amino acids are included in a pair of columns, the higher are their individual column entropy and related joint entropy that often causes a greater $\mathbb{MI}$ value. However, if a column pair is slightly conserved like (k_1, l_1), the alphabet size in these column pair is quite low which results in a less $\mathbb{MI}$ value.

Example: Let M be the MSA under study and let (k_1, l_1) and (k_2, l_2) be two different column pairs in M, where the observed alphabet size of the first column pair is $\mathfrak{X}_{(k_1,l_1)} = 2$ and the observed alphabet size of the second column pair $\mathfrak{X}_{(k_2,l_2)} = 20$ (see Figure 4.2). Further, suppose that the associated empirical amino acid distributions, $\hat{p}(x_{k_i})$, $\hat{p}(y_{l_i})$ and $\hat{p}(x_{k_i}, y_{l_i})$, $i = 1, 2$ are uniformly distributed in both pairs of columns which results in

$$\mathbb{H}(X_{k_1}) = \mathbb{H}(Y_{l_1}) = \mathbb{H}(X_{k_1}, Y_{l_1}), \tag{4.1.12}$$

$$\mathbb{H}(X_{k_2}) = \mathbb{H}(Y_{l_2}) = \mathbb{H}(X_{k_2}, Y_{l_2}). \tag{4.1.13}$$

Now, we calculate the $\mathbb{MI}\left(X_{k_i};Y_{l_i}\right)$ of both pairs of columns as

$$\begin{aligned}
\mathbb{MI}\left(X_{k_1};Y_{l_1}\right) &= -\sum_{1}^{2}\frac{1}{2}\log\frac{1}{2} = -\log\frac{1}{2} = \log 2 = \log|\mathcal{X}_{(k_1,l_1)}| && (4.1.14)\\
\mathbb{U}\left(X_{k_1};Y_{l_1}\right) &= 2\cdot\frac{\mathbb{MI}\left(X_{k_1};Y_{l_1}\right)}{\mathbb{H}\left(X_{k_1}\right)+\mathbb{H}\left(Y_{l_1}\right)} = 1 && (4.1.15)\\
\mathbb{MI}\left(X_{k_2};Y_{l_2}\right) &= -\sum_{1}^{20}\frac{1}{20}\log\frac{1}{20} = -\log\frac{1}{20} = \log 20 = \log|\mathcal{X}_{(k_2,l_2)}| && (4.1.16)\\
\mathbb{U}\left(X_{k_2};Y_{l_2}\right) &= 2\cdot\frac{\mathbb{MI}\left(X_{k_2};Y_{l_2}\right)}{\mathbb{H}\left(X_{k_2}\right)+\mathbb{H}\left(Y_{l_2}\right)} = 1 && (4.1.17)
\end{aligned}$$

The Equations 4.1.14 and 4.1.16 clearly show the alphabet size effect on the $\mathbb{MI}$. Although, both column pairs (k_1,l_1) and (k_2,l_2) are perfectly correlated and have the same $\mathbb{U}$ value, their raw $\mathbb{MI}$-values reflect these correlations completely different because of the alphabet size. Consequently, we can see that the higher is the alphabet size, the greater is the corresponding mutual information. Thus, a normalization method was required in order to reduce the effect of the alphabet size on the $\mathbb{MI}$ values which explains the reason of using the $\mathbb{U}$-metric in this thesis.

4.1.1 Filtering of an MSA

We filter MSAs before their analysis with an approach similar to [22].

First, we delete highly similar and dissimilar sequences in a given MSA to ensure that the sequence identity between any two sequences is at least 20% and no more than 90%.

Second, we remove strictly conserved residue columns (see column m in the Figure 4.1). We say a column is strictly conserved if the percentage of identical residues is greater than 95%. The reason for this is two-fold. First, these strictly conserved columns occur because of the common ancestor of all living organisms thus a compensatory mutation at that positions is often not allowed through the evolution (see Section 2.3). Second, fully conserved columns have an entropy of zero which results in a division through zero (see Section 3.1).

Third, we eliminate the columns which contain more than 25% gaps. As mentioned before, we do not consider the gaps '-' as a real amino acid in a protein as well as in an MSA. Therefore, rows with a gap at position k or l were excluded from the computation of $\mathbb{U}$-metric. If we incorporate them in the calculation, the gaps can provide wrong information on the correlation between columns. As a result of this, the columns with a higher percentage of gaps could be detected by the $\mathbb{U}$-metric as highly correlated although only little information is available in those columns.

Finally, we only consider MSAs that have at least 125 rows remaining after the filtering.

4.2 Determining an MSA-specific lower bound for the significance of $\mathbb{U}$-values

Filtering an MSA ensures that there is always a correlation between each column pairs in this MSA based on the definition of $\mathbb{U}$-metric. However, a challenging problem in bioinformatics is the separation of significant $\mathbb{U}$-values between two or more column pairs from the background noise and unrelated column pairs. Thus, we have developed an MSA-specific statistical model based on multiple testing procedures that quantifies the error made in terms of the false discovery rate described in [27,28]. Below, we will explain our statistical model step by step. The result is an MSA-dependent threshold τ above which $\mathbb{U}$-values are defined as significant.

4.2.1 Step 1: Calculation of p-values

Let M be the MSA for the protein under study. We slightly extend the standard approach of multiple testing procedures introduced in [27, 75–77] with the following assumptions in mind. If regarded as random variables, M's $\mathbb{U}(k,l)$-values follow three different distributions as demonstrated in Figure 4.3: i) a null distribution, F_0 representing background signals; ii) a distribution for those column pairs which are completely unrelated, G_1; iii) a distribution representing the correlation signals we are interested in, G_2.

We assume F_0 to be a β-distribution and M's $\mathbb{U}(k,l)$-values $U_1, U_2, \ldots, U_\mu$ be an independent and identically distributed (i.i.d.) sample, although there are some weak dependencies between them. With respect to the F_0, we first determine the p-values of each $\mathbb{U}$ values.

p-values of $\mathbb{U}$-values

The p-values of $\mathbb{U}$ depend on whether $\mathbb{U}$-values are F_0, G_1 or G_2-distributed. According to Figure 4.4 and the definition of $\mathbb{U}$-metric, we can say that while the G_1-distributed unrelated pair signals have a low $\mathbb{U}$-value, the G_2 distributed correlation signals of column pairs take relatively high $\mathbb{U}$-values, whereas the F_0 distributed background signals have moderate $\mathbb{U}$-values.

Let M be an arbitrarily chosen but fixed MSA and $U_1, U_2, \ldots, U_\mu$ be M's $\mathbb{U}$-values. We calculate the p-value for each $\mathbb{U}$-value as

$$X_\iota = 1 - F_0(U_\iota) = P\{\text{random } F_0 \text{ -distributed value} \geq U_\iota\}, \tag{4.2.1}$$

where X_ι is the p-value of U_ι with respect to F_0. If U_ι is F_0-distributed, then X_ι is uniform over $[0,1]$ (Figure 4.3). If however, U_ι is G_1 distributed, then the corresponding X_ι is relatively high and tends to 1 (Figure 4.3). In contrast, if U_ι is G_2 distributed, then X_ι becomes a low score and tends to 0 (Figure 4.3).

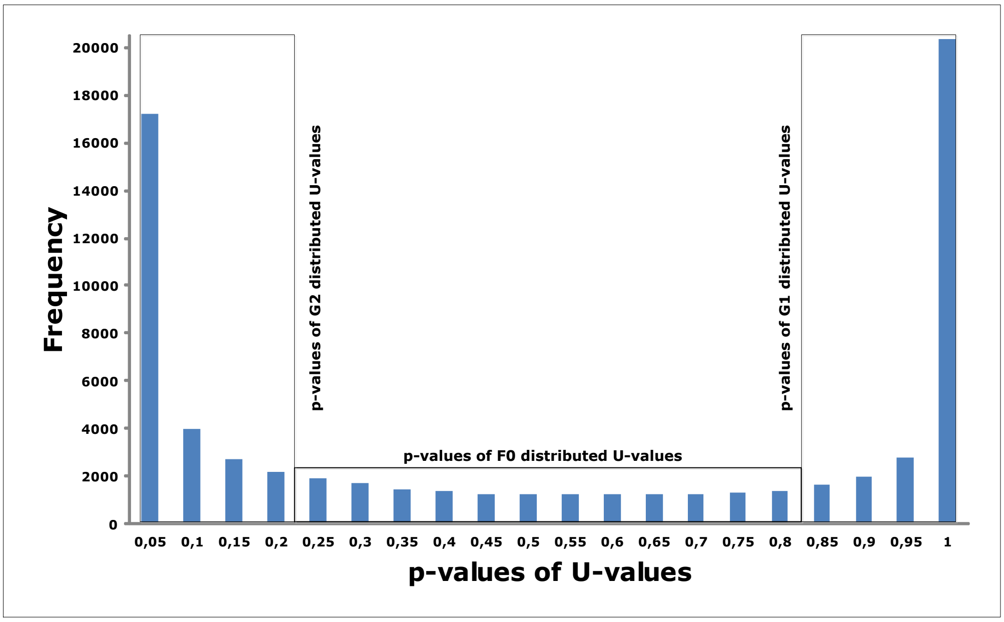


Figure 4.3: **p-value distribution of $\mathbb{U}$-values for human GCK protein (PDB Entry 1V4S).** The p-values close to zero represent the significant pairs by means of which we assess the individual residue position. In contrast, p-values of unrelated pairs tend to one. As one can see, the p-values of F_0 distributed $\mathbb{U}$-values are approximately uniform.

Equation 4.2.1 shows that the p-values of the $\mathbb{U}$-values can be calculated based on either the cumulative distribution function of the β-distribution (see Equation 3.3.5) or by drawing a random sample from the β-distribution with parameters α, $\beta > 0$. However, in both cases we first need to know the α and β. According to equations 3.3.6 and 3.3.7, we can estimate α and β from the expected value and the variance of M's $\mathbb{U}$-values. The expected value is estimated by the sample mean of all $\mathbb{U}$-values of M according to the Equation 3.3.8. However, we do not estimate the variance directly using the Equation 3.3.9, instead we use a more sophisticated approach. If we were to use the Equation 3.3.9 directly, we would get relatively high sample variance because of the G_1 and G_2-distributed $\mathbb{U}$-values. Then, the β- distribution would not represent the F_0- distribution. Hence, the corresponding p-values do not follow the uniform distribution over $[0,1]$ (see Figure 4.5).
follow the real scores (see figure 4.5).

In order to eliminate the effect of G_1 and G_2-distributed $\mathbb{U}$-values on the sample variance we take pattern from [78]. Having drawn an i.i.d. sample $(C_1,C_1'),(C_2,C_2'),\ldots,(C_v,C_v')$ of random column pairs of a sufficient size whose $\mathbb{U}$-values fall in a preassigned subinterval

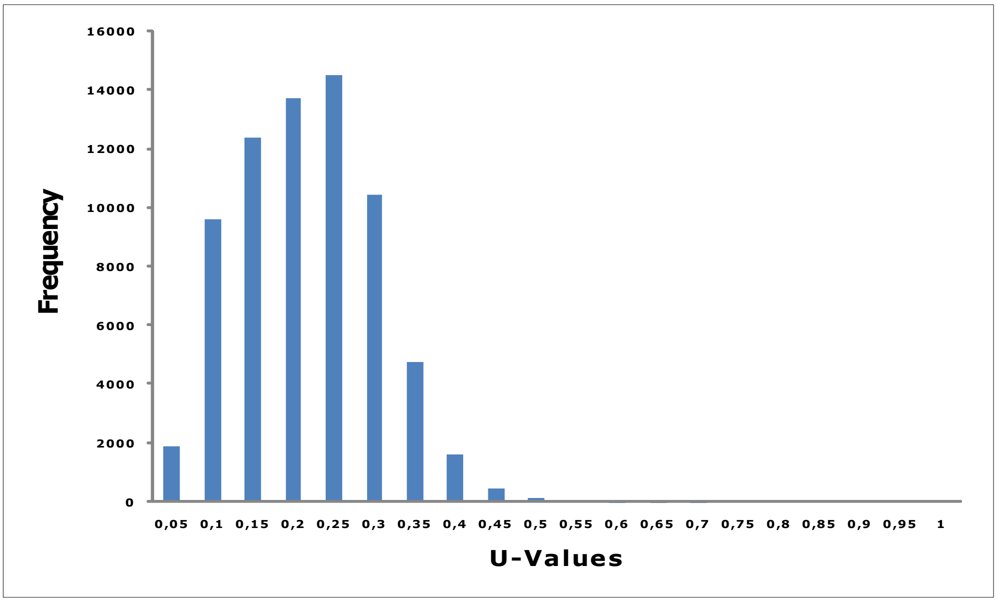


Figure 4.4: **$\mathbb{U}$-value distribution for human GCK protein (PDB Entry 1V4S)**. The higher the $\mathbb{U}$-value between pairs of columns, the stronger the correlation of amino acids at these columns.

of $[0,1]$, we calculate $D_1, D_2, \ldots, D_\nu$ by randomly shuffling the sequence C'_ι, $\iota = 1, 2, \ldots, \nu$. The artificial variance is then estimated according to Equation 3.3.9 as the sample variance of $(C_1, D_1), (C_2, D_2), \ldots, (C_\nu, D_\nu)$. Afterwards, we use the sample mean and the artificial variance to estimate the parameters of the β-distribution.

In order to draw random numbers from the β-distribution, we apply the *BN-algorithm* in [79], which is an acceptance-rejection technique using uniform and normal distributions to generate β-distributed random numbers (see Algorithm 1).

4.2.2 Step 2: Setting an MSA-dependent threshold τ

After the computation of p-values of M's $\mathbb{U}$-values, we need a threshold τ to separate significant p-values of correlation signals from the background and unrelated pair signals. While determining such a significant threshold τ, a certain amount of non-significant $\mathbb{U}$-values is accepted to be significant. The ratio of these falsely accepted values is defined as *False Discovery Rate* (FDR). In order to determine an MSA-dependent threshold τ above which $\mathbb{U}$-

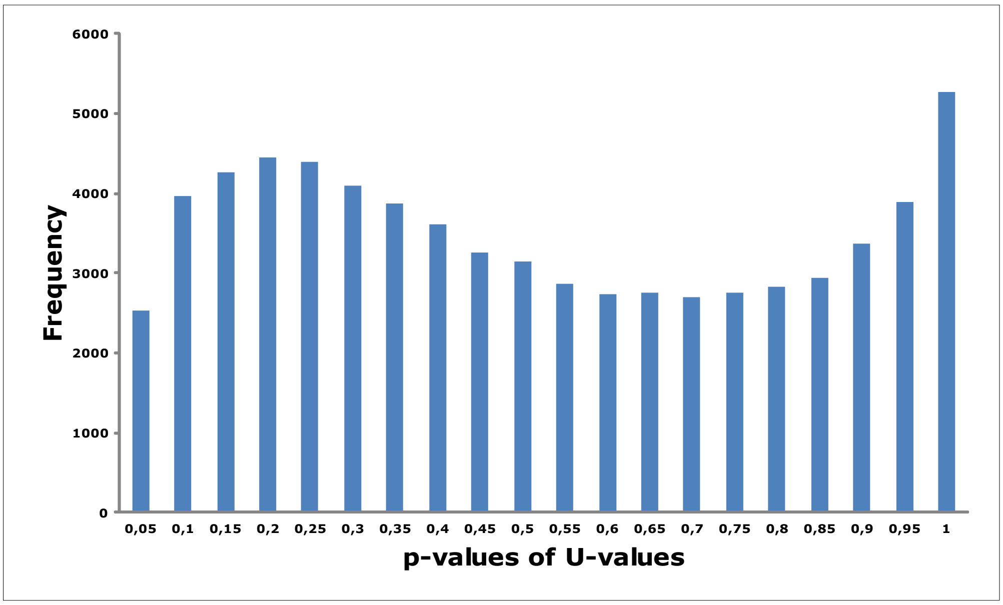


Figure 4.5: **p-value distribution of $\mathbb{U}$-values according to the sample mean and variance for human GCK protein (PDB Entry 1V4S).** The p-values are estimated based on the sample mean and variance which is calculated using all $\mathbb{U}$-values of M. As one can see, if we use directly the sample variance for the estimation of shape parameters α and β of the β-distribution, the p-values of $\mathbb{U}$-values do not follow the uniform distribution over $[0,1]$.

values are defined as significant, we generalize for a preassigned FDR the Storey-Tibshirani procedure devised for multiple testing problems as follows [27, 28].

First, we estimate the fraction γ of the F_0-distributed U_ι's as

$$\hat{\gamma} = \frac{\text{number of } p\text{-values in } [\lambda_1, \lambda_2]}{n(\lambda_2 - \lambda_1)}, \tag{4.2.2}$$

where n is the total number of M's $\mathbb{U}$-values and λ_1, λ_2 are the tuning parameters. λ_1 and λ_2 are chosen such that the fraction of not uniformly distributed p-values that fall into $[\lambda_1, \lambda_2]$ is negligible. The selection of the tuning parameters requires an automated approach during the analysis of M. Algorithm 2 presents our approach for determining λ_1 and λ_2.

Second, knowing the fraction $\hat{\gamma}$ of the F_0-distributed $\mathbb{U}$-values, we say a $\mathbb{U}$-value of sites (i, j) is significant if and only if the p-value $(1 - F_0(\mathbb{U}(i, j)))$ is less than or equal to τ, for

Algorithm 1 The *BN-algorithm:* to sample from β-distribution using only normally and uniformly distributed random numbers [79].

```
1: Input: Shape Parameters $\alpha > 1$ and $\beta > 1$
2: $A \leftarrow \alpha - 1,\ B \leftarrow \beta - 1$
3: $C \leftarrow A + B,\ L \leftarrow C \cdot ln(C)$
4: $\mu \leftarrow \frac{A}{C},\ \sigma \leftarrow \frac{0.5}{\sqrt{C}}$
5: Take a sample $s$ from the standard normal distribution
6: $x \leftarrow s \cdot \sigma + \mu$
if $x < 0$ or $x > 1$ then
    7: Reject it and go to 8
else if Generate a uniform distributed random value then
    8: $u \leftarrow$ Random value $\in [0, 1]$
end if
if $(ln(u) > (A \cdot ln(\frac{x}{A}) + B \cdot ln(\frac{(1-x)}{B}) + L + 0.5 \cdot s^2))$ then
    9: Reject it and go to 8
else if Deliver $x$ as a sample from the $\beta$-distribution with parameters $\alpha$ and $\beta$ then
    return x
end if
```

a threshold $\tau \leq \lambda_1$ that ensures the input FDR, which is estimated by

$$\widehat{\mathrm{FDR}}(\tau) = \frac{\hat{\gamma} n \tau}{\text{number of } p\text{-values} \leq \tau}. \tag{4.2.3}$$

Finally, setting $\tau = 0$ initially, we increase the τ step by step and solve the Equation 4.2.3 iteratively, as long as the result of this computation is less than the preassigned $\widehat{\mathrm{FDR}}$. After that, the last τ upon what the $\widehat{\mathrm{FDR}}$ exceeds its limit is defined as the significance threshold.

Assessment of individual residue sites

A significant $\mathbb{U}$-value simply reflects the correlation between residue sites i and j in an MSA M but does not provide enough information about the site i or j, alone. Because of this, Merkl and Zwick have suggested a method based on concepts of network analysis to characterize individual residue sites [22]. They argued that the consideration of $\mathbb{U}$-values is only a promising method if there is a strong correlation signal between residue sites i and j. However, if all $\mathbb{U}$-values are low in an MSA, a single $\mathbb{U}(i, j)$-value can be easily misclassified due to definition of $\mathbb{U}$ metric. To overcome this problem and to evaluate individual residue positions i and j, we use the connectivity degree technique introduced in [22] and further developed in [4].

Algorithm 2 Determination of the λ_1 and λ_2 according to *M*'s *p*-values.

```
Input: All $p$-values of $M$'s $\mathbb{U}$-values
$\lambda_1 \leftarrow 0,\ \lambda_2 \leftarrow 1$
$\gamma_1 \leftarrow 0,\ \varepsilon \leftarrow 0.01,\ \eta \leftarrow 0.01$
accept $\leftarrow$ false
while (!accept and $(\lambda_1 < \lambda_2)$) do
    Calculate $\gamma_2$ using the Equation 4.2.2
    $\delta = \gamma_2 - \gamma_1$
    if $(\delta < \varepsilon$ and $\delta > -\varepsilon)$ then
        accept $\leftarrow$ true
        $\lambda_1 \leftarrow \lambda_1 - \eta$
        $\lambda_2 \leftarrow \lambda_2 + \eta$
    else
        $\lambda_1 \leftarrow \lambda_1 + \eta$
        $\lambda_2 \leftarrow \lambda_2 - \eta$
        $\gamma_1 \leftarrow \gamma_2$
    end if
end while
return $[\lambda_1, \lambda_2]$
```

Definition 4.1 ***(Connectivity degree)*** *The connectivity degree of a site i with respect to the metric* $\mathbb{U}$ *and the MSA M is defined as number of sites j such that the* $\mathbb{U}(i,j)$*-value of residue sites i and j is significant for M (see Figure 4.6).*

Finally, the site *i* is defined to be $(\mathbb{U},M)$*-significant*, if *i*'s connectivity degree with respect to $\mathbb{U}$ and *M* is greater than or equal to the *n*-th percentile, where we set $n = 90$ in this thesis.

4.3 Enhancing prediction by the $\mathbb{U}_{D(\alpha)}$-metric that models dissimilar compensatory mutations

In this section, we will present our new entropy based metric, called $\mathbb{U}_{D(\alpha)}$-metric. The $\mathbb{U}$ -metric introduced in Section 4.1 is defined as a normalized mutual information that uses only the observed frequency of amino acids in the MSA columns. Thus, it does not consider any physical or biochemical properties of amino acids which are likely to be crucial for the detection of functional or structural important positions. Therefore, we have developed our novel $\mathbb{U}_{D(\alpha)}$-metric which differ from the $\mathbb{U}$-metric by incorporating the significant *BLOSUM62* dissimilar amino acid signals in the prediction of functional or structural important residue sites. Basically, the calculation of the $\mathbb{U}_{D(\alpha)}$-metric is based on transforming the empirical pair distributions of column pairs with a doubly stochastic pair substitution ma-

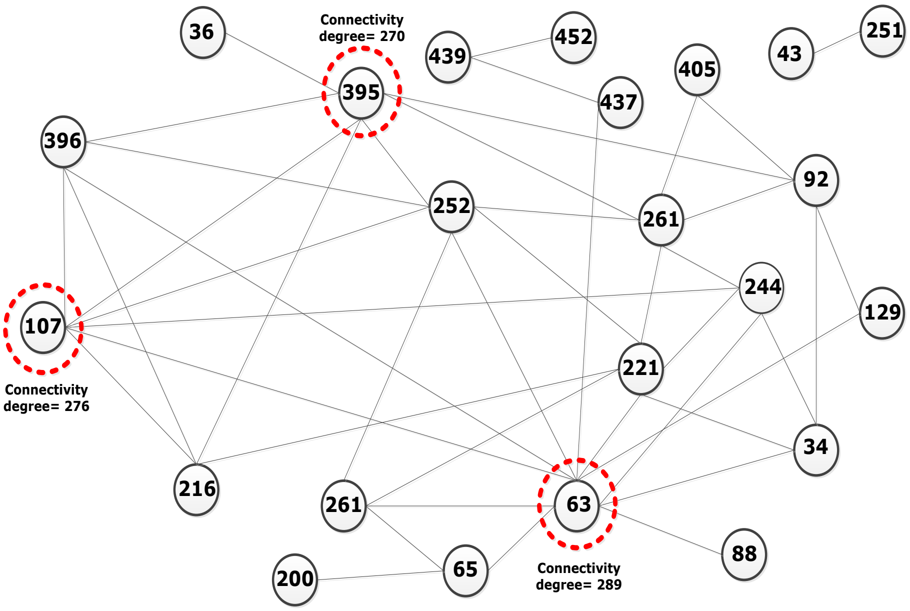


Figure 4.6: **A small part of connectivity degree network of human GCK protein (PDB-Entery 1V4S):** The nodes represent individual residue sites and each vertex indicate a significant correlation ($\mathbb{U}$-value) between residue sites k and l. The vertex count of a node indicates its connectivity degree. The red circles shows residue sites which have three highest connectivity degree in whole network and thus defined as $(\mathbb{U}, M)$-*significant*.

trix. The doubly stochastic matrix reflects the *BLOSUM62* dissimilar amino acid signals which are observed according to the significant column pairs of training MSAs.

4.3.1 Training data set

In order to build our mathematical model we used approximately 1700 protein structures which were randomly chosen from a redundancy free data set with more than 35000 protein structures. The redundancy free data set was prepared by Rainer Merkl's group University of Regensburg. In order to construct the data set, the PISCES server [80] is firstly used with a sequence-similarity cut-off of 25% for the elimination of redundant proteins. Afterwards, we took the corresponding protein structures from the PDB database and the related MSAs were gathered from the HSSP database for our training. Then, we filtered the MSAs in the same way as described in the Section 4.1.1:

- highly similar and dissimilar sequences in the MSAs are removed to ensure that the sequence identity (s_{ij}) between any two sequences is $20\% \leq s_{ij} \leq 90\%$,
- strictly conserved columns in MSAs which contain same residue types more than 95% are removed,
- the columns which contain more than 25% gaps are removed,
- finally, all MSAs with less than 125 sequences are discarded.

More than 17000 MSAs survived the last filtering step. We used approximately 1700 MSAs as training data which we randomly chose from this set. The pdb entries of the corresponding protein structures are listed in Table 9.1.

4.3.2 Preparing substitution matrices

For our approach, we need to prepare a significant substitution matrix and a random substitution matrix.

Let M be an arbitrarily chosen but fixed MSA with m sequences and let k and l be two columns in M. To compute the pair-to-pair substitution scores of amino acid pairs in these columns, we utilize standard concept of computational biology suggested in [81, 82]. For the 20 amino acids, there are 400 different amino acid pairs and each of them can occur in this column pair (k,l). Since we count pair-to-pair substitution of amino acid pairs, there exist 160000 different possible substitution between amino acid pairs for which we use a 400×400 symmetric matrix C. Having that, we calculate pair-to-pair substitution score between pairs of amino acids $[a_{ki}, a_{li}]$ and $[a_{kj}, a_{lj}]$ occurring in columns k and l based on counting the following four individual substitutions:

$$([a_{ki}, a_{li}], [a_{kj}, a_{lj}]), \tag{4.3.1}$$

$$([a_{li}, a_{ki}], [a_{lj}, a_{kj}]), \tag{4.3.2}$$

$$([a_{kj}, a_{lj}], [a_{ki}, a_{li}]), \tag{4.3.3}$$

$$([a_{lj}, a_{kj}], [a_{li}, a_{ki}]), \tag{4.3.4}$$

where i and j indicate the corresponding rows in M. We count the substitutions between pairs of amino acids symmetrically since the correct order of the sequences in MSAs is unknown. Finally, each of these substitution scores are summed up in the matrix C. In Figure 4.7 we give an example in order to show how we count the pair-to-pair substitution in a column pair.

4.3.3 Creating the Doubly Stochastic Matrix to model dissimilar compensatory mutations

A pair $((a_i, a_j), (a_k, a_l))$ of amino acid pairs is defined to be a *formal dissimilar compensatory mutation*, if the *BLOSUM62* score of both (a_i, a_k) and (a_j, a_l) is negative.

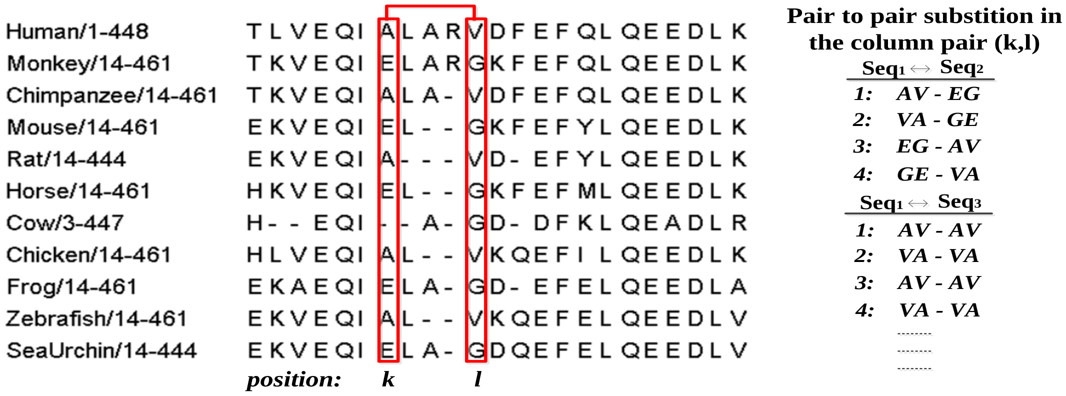

Human/1-448 TLVEQI ALARVDFEFQLQEEDLK
Monkey/14-461 TKVEQI ELARGKFEFQLQEEDLK
Chimpanzee/14-461 TKVEQI ALA-VDFEFQLQEEDLK
Mouse/14-461 EKVEQI EL--GKFEFYLQEEDLK
Rat/14-444 EKVEQI A---VD-EFYLQEEDLK
Horse/14-461 HKVEQI EL--GKFEFMLQEEDLK
Cow/3-447 H--EQI --A-GD-DFKLQEADLR
Chicken/14-461 HLVEQI AL--VKQEFI LQEEDLK
Frog/14-461 EKAEQI ELA-GD-EFELQEEDLA
Zebrafish/14-461 EKVEQI AL--VKQEFELQEEDLV
SeaUrchin/14-444 EKVEQI ELA-GDQEFELQEEDLV
position: k l
Pair to pair substition in the column pair (k,l)
Seq1 ↔ Seq2
1: AV - EG
2: VA - GE
3: EG - AV
4: GE - VA
Seq1 ↔ Seq3
1: AV - AV
2: VA - VA
3: AV - AV
4: VA - VA
....
....
....

Figure 4.7: **Counting pair to pair substitution:** In the column pair (k,l) occurs different pair to pair substitution such as $([A,V] \leftrightarrow [E,G])$, $([A,V] \leftrightarrow [A,V])$ and $([E,G] \leftrightarrow [E,G])$ or vice versa. Since the correct order of the sequences in MSAs is unknown , we count the substitutions symmetrically and store the occurrence values of substitutions in a matrix C.

We use the training data set to estimate a 400×400 doubly stochastic matrix D_{CompMut}. This matrix is our mathematical model of how dissimilar compensatory mutations have affected genomic sequences in the course of evolution. Its construction consists of following five steps.

Step 1:

We calculate a signal and a null set of column pairs. The signal set consists of all $(\mathbb{U},M)$-significant column pairs, where M ranges over all training MSAs. Significant column pairs of each training MSA M are determined individually using the MSA-specific statistical model (see Section 4.2). The null set consists of sufficiently many non-significant column pairs randomly chosen from every training MSA. For both the signal set and the null set we compute a symmetric 400×400 integer-valued matrix of frequencies of pair substitutions C_{alt} and C_{null} as explained in Section 4.3.2.

Step 2:

We compare both C_{alt} and C_{null} matrices in order to find out if a substitution score of signal pairs is smaller than pure chance.

Hence, we define a new matrix C_{sig} by

$$C_{\text{sig}}\big((a_i,a_j),(a_k,a_l)\big) = \begin{cases} C_{\text{alt}}\big((a_i,a_j),(a_k,a_l)\big) & \text{if } \varphi\big((a_i,a_j),(a_k,a_l)\big) = 1; \\ 0 & \text{otherwise;} \end{cases}$$

where $\varphi\big((a_i,a_j),(a_k,a_l)\big) = 1$ if and only if $(a_i,a_j) = (a_k,a_l)$ or

$$\frac{C_{\text{alt}}\big((a_i,a_j),(a_k,a_l)\big)}{\sum_{i',j',k',l'} C_{\text{alt}}\big((a_{i'},a_{j'}),(a_{k'},a_{l'})\big)} > \frac{C_{\text{null}}\big((a_i,a_j),(a_k,a_l)\big)}{\sum_{i',j',k',l'} C_{\text{null}}\big((a_{i'},a_{j'}),(a_{k'},a_{l'})\big)}.$$

Step 3:

We set all entries of the matrix C_{sig} outside the main diagonal that do not represent a formal dissimilar compensatory mutation to zero.

$$C_{\text{CompMut}}\big((a_i,a_j),(a_k,a_l)\big) = \begin{cases} 0, \text{ if BLOSUM62 score of } (a_i,a_k) \text{ or } (a_j,a_l) \geq 0 \\ C_{\text{sig}}\big((a_i,a_j),(a_k,a_l)\big), \text{ otherwise} \end{cases}$$

This results in the compensatory mutation matrix C_{CompMut}. By normalizing C_{CompMut}, we obtain a symmetric matrix P_{CompMut}. For a_i,a_j,a_k,a_l ranging over all amino acids, $P_{\text{CompMut}}\big((a_i,a_j),(a_k,a_l)\big)$ represents an empirical probability distribution on pairs of amino acid pairs.

Step 4:

Afterwards, we calculate the symmetric 400 × 400-matrix

$$S_{\text{CompMut}} = \left(\log \frac{P_{\text{CompMut}}\big((a_i,a_j),(a_k,a_l)\big)}{P^{\text{b}}_{\text{CompMut}}\big(a_i,a_j\big) P^{\text{b}}_{\text{CompMut}}\big(a_k,a_l\big)}\right)_{(a_i,a_j),(a_k,a_l)},$$

where $P^{\text{b}}_{\text{CompMut}}\big(a_i,a_j\big)$ is the marginal distribution of P_{CompMut}.

Step 5:

We set all negative entries of S_{CompMut} to zero. Then we compute the doubly stochastic matrix D_{CompMut} by means of the canonical iterated row-column normalization procedure [83].

4.3.4 Application of doubly stochastic matrix D_{CompMut} for the computation of the new $\mathbb{U}_{D(\alpha)}$-metric

In the previous Section 4.3.3, the matrix D_{CompMut} has been estimated to model significant and *BLOSUM62* dissimilar compensatory mutation signals. Based on the matrix D_{CompMut}, we incorporate compensatory mutation signals in the computation of the $\mathbb{U}$-metric that results in our novel $\mathbb{U}_{D(\alpha)}$-metric. Consequently, for every column pair (i, j) of the input MSA M, we linearly transform the associated empirical pair distribution with the doubly stochastic matrix as

$$D(\alpha) = (1-\alpha)\mathbf{1} + \alpha D_{\text{CompMut}},$$

where $\mathbf{1}$ is the 400×400 unit matrix and $\alpha \in (0,1]$ is a preassigned real number. $\mathbb{U}_{D(\alpha)}(i, j)$ is then defined to be the $\mathbb{U}$-value (see Equation 4.1.1) of this transform.

In the following example, we demonstrate how the empirical pair distribution of a column pair in an MSA is transformed with the doubly stochastic matrix to calculate corresponding $\mathbb{U}_{D(\alpha)}$-value.

Example: Let M be the input MSA and (k,l) be a column pair in M with empirical joint distributions of amino acid pairs $\hat{p}(x_i, y_i) = \hat{p}\{X = x_i, Y = y_i\}$, $x, y \in \mathfrak{X}$, and $i = 1, 2, \ldots, 20$. We write these pair probabilities $\hat{p}(x_i, y_i)$ in a $\hat{P}_{20\times 20}$ matrix.

	y_1	$\cdots$	y_{20}	$\sum_j \hat{p}(x_i, y_j)$
x_1 $\vdots$ x_{20}	$\hat{p}(x_1, y_1)$ $\vdots$ $\hat{p}(x_{20}, y_1)$	$\cdots$ $\ddots$ $\cdots$	$\hat{p}(x_1, y_{20})$ $\vdots$ $\hat{p}(x_{20}, y_{20})$	$\hat{p}_x(x_1)$ $\vdots$ $\hat{p}_x(x_{20})$
$\sum_i \hat{p}(x_i, y_j)$	$\hat{p}_y(y_1)$	$\cdots$	$\hat{p}_y(y_{20})$	1

Table 4.1: $\hat{P}_{20\times 20}$: Observed probabilities of amino acid pairs in column pair (k,l), where $x_i, y_j \in$ 20 amino acids$_{(A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y)}$.

Then, the $\hat{P}_{20\times 20}$ matrix is converted in a vector $\vec{v}$ of length 400 which is necessary to linearly transform the associated pair distribution $\hat{p}(x, y)$ with doubly stochastic matrix D_{CompMut}.

$$\vec{v}^{*}_{400} = \vec{v} \times D_{\text{CompMut}} \tag{4.3.5}$$

After that the vector $\vec{v}^{*}_{400}$ is converted again in a $\hat{P^{*}}_{20\times 20}$ matrix to determine the novel transformed marginal column distributions $\hat{p^{*}}(x)$, $\hat{p^{*}}(y)$ and pair distribution $\hat{p^{*}}(x, y)$.

	y_1	$\cdots$	y_{20}	$\sum\limits_j \hat{p}^*(x_i, y_j)$
x_1	$\hat{p}^*(x_1, y_1)$	$\cdots$	$\hat{p}^*(x_1, y_{20})$	$\hat{p}^*{}_x(x_1)$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_{20}	$\hat{p}^*(x_{20}, y_1)$	$\cdots$	$\hat{p}^*(x_{20}, y_{20})$	$\hat{p}^*{}_x(x_{20})$
$\sum\limits_i \hat{p}^*(x_i, y_j)$	$\hat{p}^*{}_y(y_1)$	$\cdots$	$\hat{p}^*{}_y(y_{20})$	1

Table 4.2: $\hat{P}^*{}_{20\times 20}$: Transformed values of joint amino acid pair distribution

Finally, in order to calculate our new $\mathbb{U}_{D(\alpha)}$-metric, we determine the individual column entropies $\mathbb{H}(k^*)$, $\mathbb{H}(l^*)$ and joint entropy $\mathbb{H}(k^*, l^*)$ based on the transformed distributions $\hat{p}^*(x)$, $\hat{p}^*(y), \hat{p}^*(x,y)$, respectively.

$$\mathbb{U}_{D(\alpha)}(k,l) = 2 \cdot \frac{\mathbb{H}(k^*) + \mathbb{H}(l^*) - \mathbb{H}(k^*, l^*)}{\mathbb{H}(k^*) + \mathbb{H}(l^*)} \tag{4.3.6}$$

Having canonically carried over the definition of a significant site pair and of the connectivity degree of a site to this case, a site i is called $(\mathbb{U}_{D(\alpha)}, M)$-*significant*, if i's connectivity degree with respect to the metric $\mathbb{U}_{D(\alpha)}$ is greater than or equal to the 90-th percentile.

4.4 Discussion

To predict sites of structural or functional importance, we combine the known $\mathbb{U}$-metric of normalized mutual information [22] with our novel metric $\mathbb{U}_{D^{(i)}(1)}$ to enhance the influence of dissimilar compensatory mutations when measuring covariation of two sites. We discuss how we devised $\mathbb{U}_{D(1)}$ introduced in the Section 4.3.

To learn the frequency of compensatory mutations, we took $\mathbb{U}$-significant site pairs as training data. We did that for reasons of computation time regardless of the fact that these data are biased. To deal with this bias, one could carry through the training in an iterative process, with our training being the first iteration. For $i > 0$, in the $(i+1)$-th iteration of this modified training, a doubly stochastic matrix $D^{(i+1)}_{\text{CompMut}}$ is calculated based on $\mathbb{U}_{D^{(i)}(1)}$-significant site pairs. This is done until the training data are stable.

According to Birkhoff's Theorem [84], every doubly stochastic matrix is a convex combination of permutation matrices. Moreover, from the Hardy-Littlewood-Pólya majorization theorem [85] follows that transforming the probability mass function by a doubly stochastic matrix increases entropy. Consequently, by linearly transforming the empirical amino acid pair distribution of a site pair by $D(1)$ before calculating the $\mathbb{U}$-value, we penalized those site pairs whose original distribution does not match the frequency pattern of formal dissimilar compensatory mutations in the training data described in the Section 4.3.1.

The challenge was to separate the signal caused by structural and functional constraints from the background and unrelated pair signal. To address this issue, we studied only metrics μ that satisfy the following condition. The larger the $\mu(k,l)$-value, the larger the probability that the two sites k and l have co-evolved. Our critical assumptions were: i) the $\mu(k,l)$-values follow three different distributions, one for the signal, one for the noise, and one for pairs of completely unrelated sites; ii) there is an MSA-dependent threshold below which the metric μ does not fall with overwhelming probability, when it is applied to the site pairs of functional or structural importance to which μ is sensitive; iii) there is an MSA-dependent threshold significantly smaller then the one in (ii) such that with overwhelming probability there are no $\mu(k,l)$-values of pairs (k,l) of unrelated sites exceeding it.

In order to near-completely eliminate the noise, we filtered both our training and input data. We calculated the significant pairs such that the preassigned false discovery rate was guaranteed by generalizing the Storey-Tibshirani procedure devised for multiple testing problems [27].

Our method to eliminate noise is orthogonal to the technique developed in [9]. Therein, for every pair of sites the so-called average product correction (APC) is calculated as an explicit noise measure, by which the mutual information is then decreased. Furthermore, it generalizes the way Merkl and Zwick [22] as well as Gao et al. [23] cope with noise. According to our judgment, taking only the top 75 high-scoring pairs or the top 25 pairs into account as has been suggested in [22] and [23], respectively, is too conservative.

We based our noise separation technique on rather weak distribution assumptions that are standard practice in multiple hypothesis testing, instead of explicitly model the noise in terms of a metric. We applied the connectivity degree technique due to Merkl and Zwick [22] to significant site pairs with respect to our both metrics.

5 Applying quantum information theory for the detection of functionally or structurally important sites

In this chapter, we will develop our second model for the prediction of important sites in proteins by applying quantum information theory. First, we give an introduction how the notions of this theory are mapped onto those from bioinformatics. After that, we will define two new metrics based on quantum Jensen-Shannon divergence in order identify functionally and/or structurally important residue sites in MSAs.

We publish the context of this chapter in [5] (see Appendix B 9.2). The descriptions and notations are based on this publication.

5.1 Notation mapping from quantum information theory in protein bioinformatics

The 20 amino acids are associated with the standard basis vectors of a 20-dimensional real vector space $\mathfrak{A}$ of column vectors. It is important to note that, we only deal with real vector spaces for our model.

As has been mentioned in the Section 3.2, in Dirac notation a column vector ψ is represented by a ket $|\psi\rangle$. Thus, the 20 amino acids are associated with the basis $|1\rangle, |2\rangle, \dots, |20\rangle$ of $\mathfrak{A}$, where $|i\rangle$ is the column vector whose j-th entry is equal to 1 if $i = j$, and equal to 0 otherwise.

In quantum mechanics unit-length vectors are the states of the system under study. The larger the scalar product of states, the better they are aligned. Thus, we interpret the scalar product of states as similarity measure. If in turn certain not necessarily orthogonal states represent amino acids or amino acid pairs, this is a canonical way to include similarity.

A basis is orthonormal if it consists of states that are pairwise orthogonal. The basis $|1\rangle, |2\rangle, \dots, |20\rangle$ is the standard orthonormal basis of $\mathfrak{A}$. To model amino acid pairs occurring in MSA column pairs, we need two copies $\mathfrak{A}_1$ and $\mathfrak{A}_2$ of $\mathfrak{A}$. Pairs are then represented as tensors $|ij\rangle = |i\rangle \otimes |j\rangle \in \mathfrak{A}_1 \otimes \mathfrak{A}_2 \quad (i, j = 1, 2, \dots, 20)$ which are Kronecker products of the corresponding column vectors of dimension 20. For $i, j = 1, 2, \dots, 20$, the column vector

$|ij\rangle$ has a dimension of 400, where exactly that entry is equal to 1 that corresponds to the pair (i,j). All other coefficients are equal to 0.

The amino acid conservation of an MSA column pair is measured on grounds of its empirical amino acid distribution $\widehat{p} = \left(\widehat{p}_{ij}\right)_{i,j=1,2,\ldots,20}$, where $\widehat{p}_{ij}$ is the relative frequency of the pair of the i-th and the j-th amino acid in that column pair. If we choose a row of the MSA column pair by pure chance, we get the amino acid pair (i,j) with probability $\widehat{p}_{ij}$ $(i,j=1,2,\ldots,20)$. In Dirac notation this is expressed by the density operator

$$\rho(\widehat{p}) = \sum_{i,j=1}^{20} \widehat{p}_{ij} \, |ij\rangle \langle ij| \, . \tag{5.1.1}$$

The logarithm of $\rho(\widehat{p})$ (see Equation 5.1.1) is $\sum_{i,j=1}^{20} \log \widehat{p}_{ij} \, |ij\rangle \langle ij|$, where $\log p$ is set to 0 if $p=0$. This entails $-\rho(\widehat{p})\log\rho(\widehat{p}) = -\sum_{i,j=1}^{20} \widehat{p}_{ij} \log \widehat{p}_{ij} \, |ij\rangle \langle ij|$. Then, the entropy of the distribution $\widehat{p}$ equals the trace of the matrix $-\rho(\widehat{p})\log\rho(\widehat{p})$:

$$\operatorname{tr}\bigl(-\rho(\widehat{p})\log\rho(\widehat{p})\bigr) = -\sum_{i,j=1}^{20} \widehat{p}_{ij} \log \widehat{p}_{ij}. \tag{5.1.2}$$

Positive semi-definite operators of trace 1 are the most general form of density operators. For taking amino acid pair similarity into account when modeling MSA column pairs, the density operator $\rho(\widehat{p})$ of Equation 5.1.1 is swapped for

$$\rho = \sum_{i,j=1}^{20} \widehat{p}_{ij} \, \left|\pi_{ij}\right\rangle \left\langle \pi_{ij}\right| , \tag{5.1.3}$$

where $\left|\pi_{ij}\right\rangle$ is a not necessarily orthonormal basis of states that represents the amino acids pairs (i,j), for $i,j=1,2,\ldots,20$. The corresponding *Gram matrix*

$$\mathcal{A} = \left(\left\langle \pi_{ij} | \pi_{kl} \right\rangle\right)_{i,j,k,l=1,2,\ldots,20} \tag{5.1.4}$$

reflects predefined similarities between the 400 amino acid pairs.

For positive semi-definite operators of trace 1 like ρ there is always an orthonormal basis $\left|\xi_{ij}\right\rangle$ $(i,j=1,2,\ldots,20)$ such that

$$\rho = \sum_{i,j=1}^{20} q_{ij} \, \left|\xi_{ij}\right\rangle \left\langle \xi_{ij}\right| , \tag{5.1.5}$$

where the eigenvalues $q_{ij} \geq 0$ sum up to one. Then, we calculate the *von Neumann entropy* of ρ as

$$\mathrm{VNE}(\rho) = \mathrm{tr}\big(-\rho\log\rho\big) = -\sum_{i,j=1}^{20} q_{ij}\log q_{ij}. \tag{5.1.6}$$

The difference between the Shannon entropy given by Equation 5.1.2 and the von Neumann entropy defined by Equation 5.1.6 is that the more similar the vectors $|\pi_{ij}\rangle$ appearing in Equation 5.1.3 are, the less is the von Neumann entropy compared with the Shannon entropy.

The predefined *Gram matrix* $\mathcal{A}$ (see Equation 5.1.4) does not fully determine the states $|\pi_{ij}\rangle$, for $i, j = 1, 2, \dots, 20$. In order to be able to specify additional conditions for a consistent amino acid pair model, we need measurements as another key notion of quantum mechanics. To this end, let $P_{kl} = |kl\rangle\langle kl|$ be the orthogonal projection from $\mathfrak{A}_1 \otimes \mathfrak{A}_2$ onto the subspace spanned by $|kl\rangle$ so that $P_{kl}|\xi\rangle = \langle kl|\xi\rangle |kl\rangle$, where $k, l = 1, 2, \dots, 20$, and $|\xi\rangle \in \mathfrak{A}_1 \otimes \mathfrak{A}_2$.

By definition, measuring a density matrix ρ given by Equation 5.1.3 with respect to the standard basis $|kl\rangle$ $(k, l = 1, 2, \dots, 20)$ results in the density matrix $\sum_{k,l=1}^{20} P_{kl}\rho P_{kl}$. This measurement determines the relative amino acid pair frequencies of the column pair under study. That is why it is reasonable to require for all $k, l = 1, 2, \dots, 20$

$$P_{kl}\rho P_{kl} = \widehat{p}_{kl}|kl\rangle\langle kl|. \tag{5.1.7}$$

We ensure that a density operator ρ satisfies these conditions having represented it with respect to the standard basis $|kl\rangle$ $(k, l = 1, 2, \dots, 20)$. The relative amino acid pair frequencies $\widehat{p}_{kl}$ have to appear on the main diagonal.

Amino acid pair distributions of MSA column pairs can be marginalized to get the amino acid distributions of the MSA columns concerned. Thus, we use the partial trace which is the analog way in quantum mechanics to observe the amino acid distributions of columns. Let i and j denote the amino acids of the first column, and k and l the amino acids of the second column. If $\rho = \sum_{i,j,k,l=1}^{20} \gamma_{ikjl}|ik\rangle\langle jl|$ is a density operator that describes the amino acid pair distribution of an MSA column pair including pair similarity, then the partial traces of ρ over the first and the second column are:

$$\mathrm{tr}_1\big(\rho\big) = \sum_{k,l=1}^{20}\big(\sum_{i=1}^{20}\gamma_{ikil}\big)|k\rangle\langle l|, \qquad \mathrm{tr}_2\big(\rho\big) = \sum_{i.j=1}^{20}\big(\sum_{k=1}^{20}\gamma_{ikjk}\big)|i\rangle\langle j| \tag{5.1.8}$$

As we have explained in Section 3.2.1, partial traces of density operators are also density operators. Here, they involve amino acid similarity schemes that are "marginals" of the underlying amino acid pair similarity. The Algorithm 3 shows how to implement $\mathrm{tr}_1\big(\rho\big)$ and $\mathrm{tr}_2\big(\rho\big)$.

Algorithm 3 Partial Trace of a density operator ρ.

```
Input: $\rho_{400\times 400}$
$dimV \leftarrow 20$
$dimW \leftarrow 20$
for ($k = 1 \ldots dimV$) do
    for ($l = 1 \ldots dimW$) do
        for ($i = 1 \ldots dimV$) do
            for ($j = 1 \ldots dimW$) do
                $PairIndices[k][l][i][j] \leftarrow psi[k * dimW + l][i * dimW + j]$;
            end for
        end for
    end for
end for
Calculation of the first partial trace, $\mathrm{tr}_1(\rho)$
for ($k = 1 \ldots dimV$) do
    for ($i = 1 \ldots dimV$) do
        $tr_1(\rho)[k][i] \leftarrow 0$
        for ($j = 1 \ldots dimW$) do
            $partialTrace_1(\rho)[k][i] + \leftarrow PairIndices[k][j][i][j]$
        end for
    end for
end for
Calculation of the second partial trace, $\mathrm{tr}_2(\rho)$
for ($l = 1 \ldots dimW$) do
    for ($j = 1 \ldots dimW$) do
        $tr_2(\rho)[l][j] \leftarrow 0$
        for ($i = 1 \ldots dimV$) do
            $partialTrace_2(\rho)[l][j] + \leftarrow PairIndices[i][l][i][j]$
        end for
    end for
end for
```

5.1.1 Calculating the Gram matrices $\mathcal{A}_{\text{ent}}$ and $\mathcal{A}_{\text{sep}}$ to be plugged into Equation 5.1.4

We leverage the matrices C_{alt} and C_{null} (see Section 4.3.3) to calculate of the Gram matrices. The entries of the two matrices are frequencies of pair substitutions: C_{alt} models the correlation signals between pair of amino acid pairs; C_{null} reflects the background signals.

We define two significant pair substitution matrices C_{ent} and C_{sep} from C_{alt} and C_{null} which form the basis of our new metrics $\mathbb{Q}_{\text{ent}}$ and $\mathbb{Q}_{\text{sep}}$ (both metrics will be defined in Section 5.2). The intuition behind C_{ent} is that *BLOSUM62*-based pair similarities are rescaled, whereas C_{sep} is the basis for a new amino acid pair similarity.

$$C_{\text{ent}}\big((a_i,a_j),(a_k,a_l)\big) = \begin{cases} C_{\text{alt}}\big((a_i,a_j),(a_k,a_l)\big) & \text{if } \varphi_{\text{ent}}\big((a_i,a_j),(a_k,a_l)\big) = 1; \\ 0 & \text{otherwise;} \end{cases} \tag{5.1.9}$$

where $\varphi_{\text{ent}}\big((a_i,a_j),(a_k,a_l)\big) = 1$ if and only if either $(a_i,a_j) = (a_k,a_l)$ or the pair $\big(a_i,a_k\big)$ as well as the pair $\big(a_j,a_l\big)$ are *BLOSUM62*-similar and

$$\frac{C_{\text{alt}}\big((a_i,a_j),(a_k,a_l)\big)}{\sum_{i',j',k',l'} C_{\text{alt}}\big((a_{i'},a_{j'}),(a_{k'},a_{l'})\big)} > \frac{C_{\text{null}}\big((a_i,a_j),(a_k,a_l)\big)}{\sum_{i',j',k',l'} C_{\text{null}}\big((a_{i'},a_{j'}),(a_{k'},a_{l'})\big)}. \tag{5.1.10}$$

$$C_{\text{sep}}\big((a_i,a_j),(a_k,a_l)\big) = \begin{cases} C_{\text{alt}}\big((a_i,a_j),(a_k,a_l)\big) & \text{if } \varphi_{\text{sep}}\big((a_i,a_j),(a_k,a_l)\big) = 1; \\ 0 & \text{otherwise;} \end{cases} \tag{5.1.11}$$

where $\varphi_{\text{sep}}\big((a_i,a_j),(a_k,a_l)\big) = 1$ if and only if either $(a_i,a_j) = (a_k,a_l)$ or Equation 5.1.10 is satisfied.

Let C be either C_{ent} or C_{sep}. We define

$$B_{(g,h),(i,j)} = \frac{C^{\alpha}_{(g,h),(i,j)}}{\sqrt{\sum_{\iota,\kappa=1}^{20} C^{2\alpha}_{(\iota,\kappa),(i,j)}}}, \tag{5.1.12}$$

where $\big((g,h),(i,j)\big)$ range over all possible 160000 indices of pairs of amino acid pairs including the main diagonal, and $\alpha \in (0,1)$ was appropriately chosen.

To ensure positive semi-definiteness, we finally set

$$\mathcal{A} = B^T B. \tag{5.1.13}$$

That way we obtain $\mathcal{A}_{\text{ent}}$ as well as $\mathcal{A}_{\text{sep}}$.

5.1.2 Simultaneously ensuring Equations 5.1.3, 5.1.4, and 5.1.7

Having chosen the kets $\left|\pi_{ij}\right\rangle$ that appear in Equation 5.1.3 so that the uniquely determined positive semi-definite square root $\sqrt{\rho}$ of ρ equals $\sum_{i,j=1}^{20} \widehat{q}_{ij} \left|\pi_{ij}\right\rangle \langle ij|$, where $\widehat{q} = \left(\widehat{q}_{ij}\right)_{i,j=1,2,\dots,20}$ represents the amino acid pair distribution we wish to plug into ρ, Equation 5.1.3 is ensured. Moreover, we get

$$\rho = \sum_{g,h,i,j=1}^{20} \sqrt{\widehat{q}_{gh}\mathcal{A}_{(g,h),(i,j)}\sqrt{\widehat{q}_{ij}}} \, |gh\rangle \langle ij| \,. \tag{5.1.14}$$

Therein, $\mathcal{A}$ is either equal to $\mathcal{A}_{\text{ent}}$ or equal to $\mathcal{A}_{\text{sep}}$. (Note, that Equation 5.1.14 corresponds to an approach due to Johansson et al. [86] for MSA columns rather than MSA column pairs.)

Equation 5.1.7 follows now from Equation 5.1.14 in a straightforward manner of swapping $\widehat{p} = \left(\widehat{p}_{ij}\right)_{i,j=1,2,\dots,20}$ for $\widehat{q} = \left(\widehat{q}_{ij}\right)_{i,j=1,2,\dots,20}$. Since the square root ($\sqrt{\rho}$) of ρ is in particular symmetric, we get $\rho = \sum_{g,h,i,j=1}^{20} \sqrt{\widehat{q}_{gh}} \left\langle \pi_{gh} | \pi_{ij} \right\rangle \sqrt{\widehat{q}_{ij}} \, |gh\rangle \langle ij|$. Comparing the latter with Equation 5.1.14, Equation 5.1.4 follows.

5.2 Defining two new metrics $\mathbb{Q}_{\text{ent}}$ and $\mathbb{Q}_{\text{sep}}$

In the Section 4.3.3 a pair $\left((a_i, a_j), (a_k, a_l)\right)$ of amino acid pairs is called a *formal dissimilar compensatory mutation*, if the *BLOSUM62*-score of both (a_i, a_k) and (a_j, a_l) is negative.

To define our first new metric $\mathbb{Q}_{\text{ent}}$ for a certain column pair of a given MSA, let $\widehat{q} = \widehat{p} \cdot D_{\text{CompMut}}$, where D_{CompMut} is the 400×400 doubly stochastic matrix defined in Section 4.3.3, and $\widehat{p}$ is the empirical pair distribution of these two columns. Having determined ρ_{ent} by Equation 5.1.14, where $\mathcal{A}$ is swapped for $\mathcal{A}_{\text{ent}}$, our first metric is

$$\mathbb{Q}_{\text{ent}} = \text{QJSD}\big(\rho_{\text{ent}} \| \rho(\widehat{p})\big). \tag{5.2.1}$$

It is the quantum Jensen-Shannon divergence of an entangled density matrix that incorporates rescaled amino acid pair similarity as well as dissimilar compensatory mutations with a separable density operator representing the empirical amino acid pair frequencies.

Our second new metric $\mathbb{Q}_{\text{sep}}$ replaces the $\mathbb{U}$-metric introduced in the Section 4.1 by a measure that integrates a new amino acid pair similarity score into the framework of quantum Jensen-Shannon divergence. To this end, let ρ_{sep} be defined by Equation 5.1.14, where $\widehat{q}$ is equal to the empirical amino acid pair distribution $\widehat{p}$ of the column pair under study, and $\mathcal{A} = \mathcal{A}_{\text{sep}}$. Then $\mathbb{Q}_{\text{sep}}$-metric is defined as

$$\mathbb{Q}_{\text{sep}} = \text{QJSD}\big(\text{tr}_1\left(\rho_{\text{sep}}\right) \| \, \text{tr}_2\left(\rho_{\text{sep}}\right)\big). \tag{5.2.2}$$

5.3 Discussion

Grosse *et al.* observed in [87] that the Jensen-Shannon divergence (JSD) can be interpreted as mutual information between two (or more) random sources in a special setting particularly appropriate to discriminate between these sources. This is what we need when it comes to predicting important protein sites in an MSA-based approach. It might explain the findings of Capra and Singh [88] on the predictive power of JSD. These two articles encouraged us to utilize quantum Jensen-Shannon divergence (QJSD) in this thesis. As a side effect, a normalization is not necessary, since quantum Jensen-Shannon divergence, like its classical counterpart, ranges over the real interval $[0,1]$.

Several studies have confirmed that detecting coupled MSA-columns is extremely useful in the prediction of important protein sites (see e.g. [4, 8, 9, 15–24]). When using information-theoretic metrics, there is no doubt that it is reasonable to incorporate amino acid pair dissimilarity as well as amino acid similarity in a consistent way such that similarity decreases entropy, whereas dissimilarity increases it. This kind of consistency is important, since entropy is the fundamental building block for most of those metrics. In particular, the Jensen-Shannon divergence between two probability mass functions p and q equals $\mathbb{H}(1/2(p+q)) - 1/2(\mathbb{H}(p) + \mathbb{H}(q))$.

In the Section 4.3 we have presented an amino acid pair dissimilarity model for compensatory mutations. A doubly stochastic matrix transforms the empirical amino acid pair distribution of a column pair. Rescaled pair similarity of *BLOSUM62*-similar pairs is to capture an aspect of coupled MSA column pairs orthogonal to the phenomenon of dissimilar compensatory mutations. It models the amino acid pair transition preferences within those column pairs on the average. As suggested by Caffrey *et al.* [89] as well as Johansson *et al.* [86], it is promising to incorporate them within the framework of quantum information theory. Therein, density matrices replace probability mass functions. The counterpart of the entropy of a probability mass function is the von Neumann entropy (VNE) of a density matrix (see Equation 5.1.6). QJSD corresponds then exactly to JSD.

The challenge was to complement the model presented in the Section 4.3 by additionally incorporating amino acid pair similarity in a way that the two effects interfere but do not interact. We model the 400 amino acid pairs by means of 400 not necessarily orthogonal tensors spanning the tensor product of two copies of a 20-dimensional Hilbert space. This provides us with the opportunity to utilize the notion of entanglement, which in turn is a major resource of quantum information. Moreover, we are in a position to make use of partial traces, which play the role of the marginals in the classical case. Pair similarity is reflected by means of the Gram matrix of these tensors (see Equation 5.1.4). To ensure positive definiteness, which is a key property of density matrices, we used transitivity of similarity (see Equation 5.1.13). Since there is no transitivity of dissimilarity, we kept dissimilarity apart from that Gram matrix. Instead, we carried over our previously defined *BLOSUM62*-dissimilarity model which was used to develop $\mathbb{U}_{D(\alpha)}$-metric (see Section 4.3).

Gram matrix and transformed amino acid pair distribution are joined together by means of Equation 5.1.14 in the final step of our density matrix design. That way we minimize the interaction between the two effects of dissimilarity and similarity.

In order to eliminate the noise and to define an MSA-dependent threshold for significant column pairs, we followed the line of the Section 4.2. The MSA-specific statistical model presented there seems to be of universal applicability. The same is true for the connectivity degree model introduced in the Section 4.2.2. Combining them results in a reliable and robust method to determine significant residue sites.

6 Results

In this chapter, we will analyze important sites of the human epidermal growth factor receptor (EGFR) protein and glucokinase (GCK) protein in order to demonstrate the functionality of our both classical and quantum information theory based methods. We have selected both proteins because their functionally and structurally important sites have been experimentally well studied [29–40].

EGFR is a member of the ErbB (Erythroblastic Leukemia Viral Oncogene Homolog) family receptors. Signaling through this receptor is a highly conserved mechanism from nematode to humans involved in numerous processes, including proliferation, cell fate determination, and tissue specification [90]. Furthermore, several studies have implicated that mutations resulting in misregulation of the activity or action of EGFR led to multiple cancers, including those of the brain, lung, mammary gland, and ovary [29–32].

GCK is a monomeric enzyme catalyzing phosphorylation of glucose to glucose-6-phosphate, which is the first step in the utilization of glucose, at physiological glucose concentration in pancreas and liver. Given the fact that GCK displays low affinity for glucose, it acts as a glucose sensor playing an important role in the regulation of carbohydrate metabolism. Mutations of the GCK gene can lead to maturity onset diabetes of the young (MODY) characterized by an autosomal dominant mode of inheritance and onset early adulthood [37], or familial hyperinsulinemic hypoglycemia type 3 (HHF), which is common cause of persistent hypoglycemia in infancy [91].

In Section 6.1, we will start with the introducing the structurally or functionally important residue sites of these two human proteins. After that, in Section 6.2 we will go on analyzing both proteins in detail with $\mathbb{U}$-metric and $\mathbb{U}_{D(\alpha)}$-metric presented in Chapter 4. In addition, we will compare the functionality of these metrics with each other to evaluate their performance statistically. In Section 6.3, we will predict the important sites of EGFR and GCK proteins applying our quantum information theory based metrics $\mathbb{Q}_{\text{ent}}$ and $\mathbb{Q}_{\text{sep}}$ presented in the Chapter 5 and analyze the predicted sites in detail. Finally, in Section 6.4, we will make a comparison between the results of both classical and quantum information theory based methods.

We have previously published most of this chapter in [4] and [5] (see Appendix A 9.1 and B 9.2).

6.1 Essential sites of human EGFR and GCK protein

To evaluate the significant residue sites detected by our classical and quantum information theory based methods, we have investigated essential sites of human EGFR (pdb entry 2J6M) and GCK (pdb entry 1V4S) proteins. The essential sites of both proteins have been assigned into three main categories: i) the nsSNP positions and their adjacent sites; ii) residue positions which are located at or near active sites, allosteric sites, or binding sites; iii) residue positions which are nearby strictly conserved sites. Here, we have used "nearby" definition of Nussinov et al. [62] and defined two residues as in contact or adjacent when the distance between their major carbon atoms is less than 6 Å (for the distance calculation see the Section 2.5.4). We have defined positions which are nearby nsSNPs as essential, because several of them are also adjacent to active sites, allosteric sites, binding sites, or strictly conserved sites. Thus, we refer to a significant residue site as "*functionally or structurally important*" if it falls into one of these categories of essential sites.

6.1.1 MSAs and 3-D structures of human EGFR and GCK protein

We gathered the related MSAs of both proteins from the HSSP database (see Section 2.5.2) [54], and their corresponding 3-D structure information were downloaded from PDB database (see Section 2.5.1) [2] which are necessary to determine the distance between residues.

The human EGFR protein has a chain length of 306 residues and its related MSA contains 1551 sequences. Likewise, the human GCK protein has a chain length of 448 residues and its related MSA contains 785 sequences. In Figure 6.1 and 6.2 , we illustrate positions of residues in both proteins with corresponding residue sequence number in PDB files.

6.2 Applying the $\mathbb{U}$-metric and the $\mathbb{U}_{D(\alpha)}$-metric for the human EGFR and GCK protein

We apply our classical information theory based method to predict functionally or structurally important residue positions in two steps. First, we utilize the new MSA-specific statistical method presented in the Section 4.2 for the identification of significant MSA column pairs with respect to either of the two metrics $\mathbb{U}$ and $\mathbb{U}_{D(\alpha)}$. Assuming that M is the MSA under study, these pairs are annotated as $(\mathbb{U},M)$-significant and $(\mathbb{U}_{D(\alpha)},M)$-significant, respectively. Second, we utilized the connectivity degree of a residue site (see in the Section 4.2.2). Recall that the connectivity degree of a residue site indicates the number of its significant coupled mutation pairs. In this case, a site is called (U,M)-significant when the frequency of occurrence of this site in the set of $(\mathbb{U},M)$-significant pairs exceeds the 90-th percentile. Having defined the concept of a $(\mathbb{U}_{D(\alpha)},M)$-significant site analogously, a site

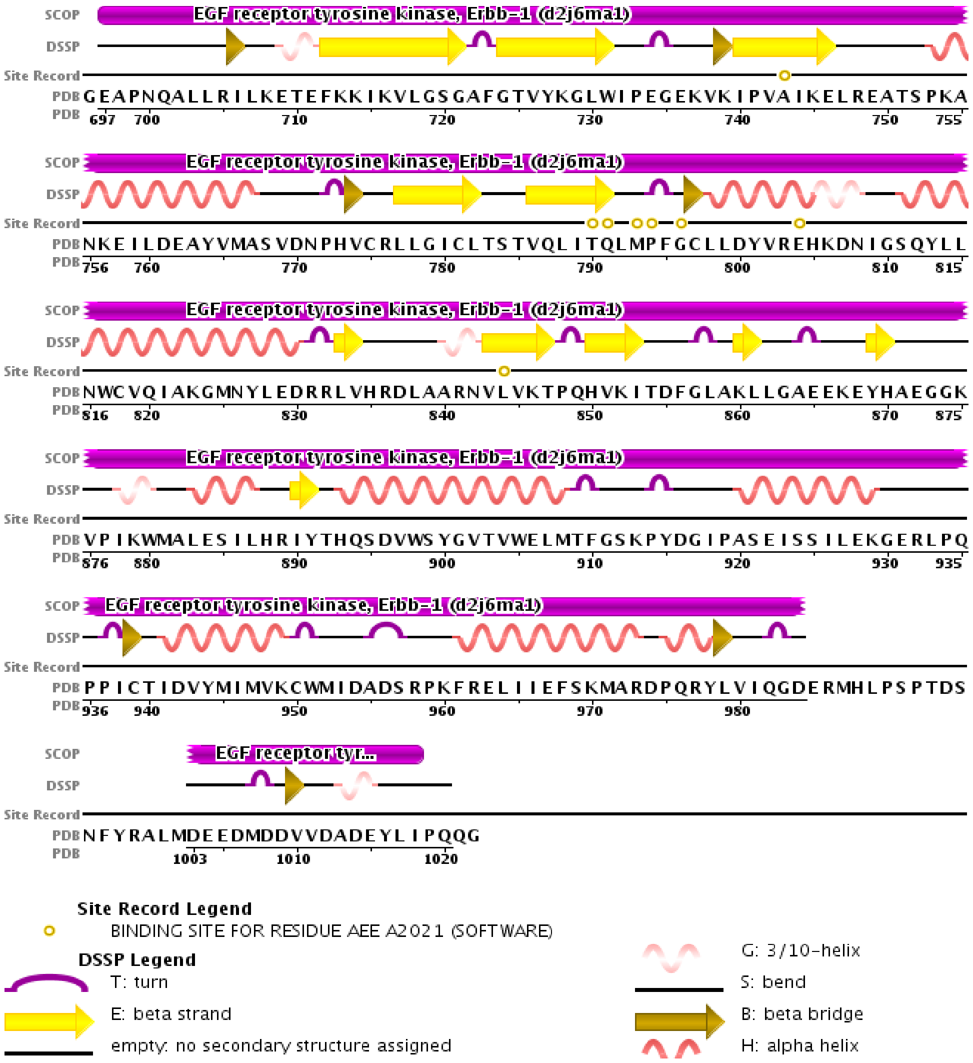

SCOP
DSSP
Site Record
PDB
EGF receptor tyrosine kinase, Erbb-1 (d2j6ma1)
GEAPNQALLRILKETEFKKIKVLGSGAFGTVYKGLWIPEGEKVKIPVAIKELREATSPKA
697 700 710 720 730 740 750 755
NKEILDEAYVMASVDNPHVCRLLGICLTSTVQLITQLMPFGCLLDYVREHKDNIGSQYLL
756 760 770 780 790 800 810 815
NWCVQIAKGMNYLEDRRLVHRDLAARNVLVKTPQHVKITDFGLAKLLGAEEKEYHAEGGK
816 820 830 840 850 860 870 875
VPIKWMALESILHRIYTHQSDVWSYGVTVWELMTFGSKPYDGIPASEISSILEKGERLPQ
876 880 890 900 910 920 930 935
PPICTIDVYMIMVKCWMIDADSRPKFRELIIEFSKMARDPQRYLVIQGDERMHLPSPTDS
936 940 950 960 970 980
EGF receptor tyr...
NFYRALMDEEDMDDVVDADEYLIPQQG
1003 1010 1020
Site Record Legend
BINDING SITE FOR RESIDUE AEE A2021 (SOFTWARE)
DSSP Legend
T: turn
E: beta strand
empty: no secondary structure assigned
G: 3/10-helix
S: bend
B: beta bridge
H: alpha helix

Figure 6.1: **The sequence of the human EGFR protein (PDB-Entry 2J6M)** (image from http://www.rscb.org [2], 05.08.2013).

is defined as CMF-significant[8] with respect to M, when it is either $(\mathbb{U},M)$-significant or $(\mathbb{U}_{D(\alpha)},M)$-significant.

---

[8] We defined the predicted significant sites together as CMF-significant due to our corresponding publication ”Coupled Mutation Finder” [4].

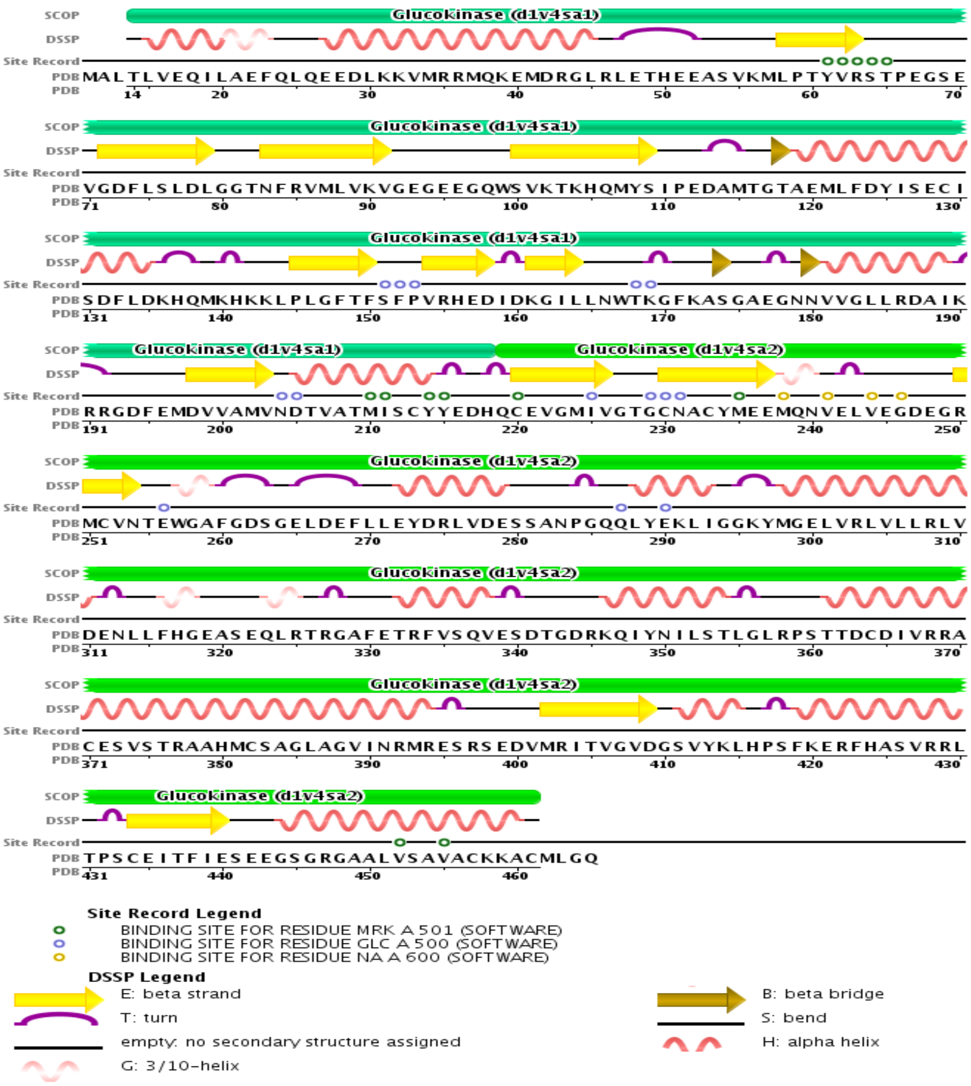

SCOP
DSSP
Site Record
PDB
Glucokinase (d1v4sa1)
MALTLVEQILAEFQLQEEDLKKVMRRMQKEMDRGLRLETHEEASVKMLPTYVRSTPEGSE
14 20 30 40 50 60 70
VGDFLSLDLGGTNFRVMLVKVGEGEEGQWSVKTKHQMYSIPEDAMTGTAEMLFDYISECI
71 80 90 100 110 120 130
SDFLDKHQMKHKKLPLGFTFSFPVRHEDIDKGILLNWTKGFKASGAEGNNVVGLLRDAIK
131 140 150 160 170 180 190
Glucokinase (d1v4sa2)
RRGDFEMDVVAMVNDTVATMISCYYEDHQCEVGMIVGTGCNACYMEEMQNVELVEGDEGR
191 200 210 220 230 240 250
MCVNTEWGAFGDSGELDEFLLEYDRLVDESSANPGQQLYEKLIGGKYMGELVRLVLLRLV
251 260 270 280 290 300 310
DENLLFHGEASEQLRTRGAFETRFVSQVESDTGDRKQIYNILSTLGLRPSTTDCDIVRRA
311 320 330 340 350 360 370
CESVSTRAAHMCSAGLAGVINRMRESRSEDVMRITVGVDGSVYKLHPSFKERFHASVRRL
371 380 390 400 410 420 430
TPSCEITFIESEEGSGRGAALVSAVACKKACMLGQ
431 440 450 460
Site Record Legend
BINDING SITE FOR RESIDUE MRK A 501 (SOFTWARE)
BINDING SITE FOR RESIDUE GLC A 500 (SOFTWARE)
BINDING SITE FOR RESIDUE NA A 600 (SOFTWARE)
DSSP Legend
E: beta strand
T: turn
empty: no secondary structure assigned
G: 3/10-helix
B: beta bridge
S: bend
H: alpha helix

Figure 6.2: **The sequence of the human GCK protein (PDB-Entry 1V4S)** (image from http://www.rscb.org [2], 05.08.2013).

We analyzed both proteins with a false discovery rate (FDR) of 1%. For EGFR, we defined a total of 14339 out of 26079 non-conserved column pairs as significant. 11365 of these significant pairs are detected as $(\mathbb{U}, M)$-significant and 3798 pairs are observed as $(\mathbb{U}_{D(\alpha)}, M)$-significant. Only 824 EGFR pairs are significant with respect to both metrics. On the other hand, for GCK, we identified a total of 32654 out of 69645 non-conserved column pairs as significant where 18106 of them are $\mathbb{U}$-significant and 16241 are $\mathbb{U}_{D(1)}$-significant. Moreover, 1693 pairs are defined as significant for both $\mathbb{U}$ and $\mathbb{U}_{D(1)}$-significant.

Applying the connectivity degree technique, we identified 22 and 36 residue positions as $\mathbb{U}$-significant for human EGFR and GCK proteins, respectively. Additionally, 21 positions of EGFR and 36 positions of GCK were detected as $\mathbb{U}_{D(1)}$-significant. Finally, a total of 43 sites of EGFR and 72 of GCK were found as CMF-significant, and predicted as of structural or functional importance (see Table 6.1 and 6.2). However, there have been no residue sites detected as significant by both metrics.

Residue	Position	Connectivity Degree	Detected by
S	720	170	$\mathbb{U}$-metric
T	725	178	$\mathbb{U}$-metric
Y	727	169	$\mathbb{U}$-metric
E	746	168	$\mathbb{U}$-metric
A	755	177	$\mathbb{U}$-metric
N	756	192	$\mathbb{U}$-metric
E	758	121	$\mathbb{U}_{D(\alpha)}$-metric
I	759	123	$\mathbb{U}_{D(\alpha)}$-metric
Y	764	174	$\mathbb{U}$-metric
M	766	104	$\mathbb{U}_{D(\alpha)}$-metric
A	767	173	$\mathbb{U}$-metric
H	773	125	$\mathbb{U}_{D(\alpha)}$-metric
Q	791	198	$\mathbb{U}_{D(\alpha)}$-metric
D	800	102	$\mathbb{U}_{D(\alpha)}$-metric
N	816	176	$\mathbb{U}$-metric
V	819	177	$\mathbb{U}$-metric
Q	820	147	$\mathbb{U}_{D(\alpha)}$-metric
G	824	99	$\mathbb{U}_{D(\alpha)}$-metric
E	829	101	$\mathbb{U}_{D(\alpha)}$-metric
D	830	95	$\mathbb{U}_{D(\alpha)}$-metric
K	846	183	$\mathbb{U}$-metric
T	847	164	$\mathbb{U}_{D(\alpha)}$-metric
Q	849	173	$\mathbb{U}$-metric
K	860	113	$\mathbb{U}_{D(\alpha)}$-metric
G	863	170	$\mathbb{U}$-metric
E	868	170	$\mathbb{U}$-metric
E	872	170	$\mathbb{U}$-metric
V	876	173	$\mathbb{U}$-metric
K	879	109	$\mathbb{U}_{D(\alpha)}$-metric
Y	891	120	$\mathbb{U}_{D(\alpha)}$-metric
T	892	175	$\mathbb{U}_{D(\alpha)}$-metric
S	899	97	$\mathbb{U}_{D(\alpha)}$-metric
Y	900	119	$\mathbb{U}_{D(\alpha)}$-metric
T	909	158	$\mathbb{U}_{D(\alpha)}$-metric
S	912	174	$\mathbb{U}$-metric
K	913	170	$\mathbb{U}$-metric
D	916	181	$\mathbb{U}$-metric
A	920	172	$\mathbb{U}$-metric
E	922	188	$\mathbb{U}_{D(\alpha)}$-metric
S	924	176	$\mathbb{U}$-metric
E	931	171	$\mathbb{U}$-metric
R	932	95	$\mathbb{U}_{D(\alpha)}$-metric
M	947	106	$\mathbb{U}_{D(\alpha)}$-metric

Table 6.1: 43 CMF**-significant residue sites found by $\mathbb{U}$ and $\mathbb{U}_{D(\alpha)}$-metric in human EGFR protein.**

Residue	Position	Connectivity Degree	Detected by		Residue	Position	Connectivity Degree	Detected by
L	25	229	$\mathbb{U}$-metric		G	264	258	$\mathbb{U}$-metric
M	34	238	$\mathbb{U}$-metric		E	265	250	$\mathbb{U}$-metric
R	36	238	$\mathbb{U}$-metric		L	266	246	$\mathbb{U}$-metric
E	40	302	$\mathbb{U}_{D(\alpha)}$-metric		D	267	246	$\mathbb{U}$-metric
T	60	305	$\mathbb{U}_{D(\alpha)}$-metric		E	268	242	$\mathbb{U}$-metric
R	63	289	$\mathbb{U}$-metric		L	271	286	$\mathbb{U}_{D(\alpha)}$-metric
T	65	213	$\mathbb{U}$-metric		S	281	286	$\mathbb{U}_{D(\alpha)}$-metric
E	67	286	$\mathbb{U}_{D(\alpha)}$-metric		N	283	321	$\mathbb{U}_{D(\alpha)}$-metric
T	82	318	$\mathbb{U}_{D(\alpha)}$-metric		Q	286	238	$\mathbb{U}$-metric
N	83	312	$\mathbb{U}_{D(\alpha)}$-metric		Q	287	332	$\mathbb{U}_{D(\alpha)}$-metric
G	92	224	$\mathbb{U}$-metric		G	294	265	$\mathbb{U}_{D(\alpha)}$-metric
H	105	253	$\mathbb{U}$-metric		E	300	346	$\mathbb{U}_{D(\alpha)}$-metric
M	107	276	$\mathbb{U}$-metric		E	331	241	$\mathbb{U}$-metric
F	123	278	$\mathbb{U}_{D(\alpha)}$-metric		T	332	240	$\mathbb{U}_{D(\alpha)}$-metric
C	129	217	$\mathbb{U}$-metric		R	333	223	$\mathbb{U}$-metric
F	133	230	$\mathbb{U}_{D(\alpha)}$-metric		Q	337	221	$\mathbb{U}$-metric
F	148	234	$\mathbb{U}_{D(\alpha)}$-metric		E	339	264	$\mathbb{U}_{D(\alpha)}$-metric
T	149	277	$\mathbb{U}_{D(\alpha)}$-metric		D	341	299	$\mathbb{U}_{D(\alpha)}$-metric
F	152	258	$\mathbb{U}_{D(\alpha)}$-metric		G	343	235	$\mathbb{U}$-metric
H	156	334	$\mathbb{U}_{D(\alpha)}$-metric		D	363	300	$\mathbb{U}_{D(\alpha)}$-metric
F	171	243	$\mathbb{U}_{D(\alpha)}$-metric		N	391	224	$\mathbb{U}$-metric
N	180	320	$\mathbb{U}_{D(\alpha)}$-metric		E	395	270	$\mathbb{U}$-metric
R	186	257	$\mathbb{U}_{D(\alpha)}$-metric		S	396	265	$\mathbb{U}$-metric
T	206	297	$\mathbb{U}_{D(\alpha)}$-metric		T	405	264	$\mathbb{U}$-metric
T	209	292	$\mathbb{U}_{D(\alpha)}$-metric		S	411	315	$\mathbb{U}_{D(\alpha)}$-metric
C	213	263	$\mathbb{U}$-metric		K	414	318	$\mathbb{U}_{D(\alpha)}$-metric
E	216	227	$\mathbb{U}$-metric		S	418	236	$\mathbb{U}$-metric
D	217	241	$\mathbb{U}_{D(\alpha)}$-metric		F	419	226	$\mathbb{U}_{D(\alpha)}$-metric
E	221	249	$\mathbb{U}$-metric		R	422	252	$\mathbb{U}$-metric
T	228	282	$\mathbb{U}_{D(\alpha)}$-metric		H	424	250	$\mathbb{U}$-metric
E	236	351	$\mathbb{U}_{D(\alpha)}$-metric		T	431	253	$\mathbb{U}$-metric
V	244	240	$\mathbb{U}$-metric		S	433	224	$\mathbb{U}$-metric
R	250	225	$\mathbb{U}$-metric		C	434	227	$\mathbb{U}$-metric
F	260	237	$\mathbb{U}_{D(\alpha)}$-metric		I	439	221	$\mathbb{U}$-metric
G	261	214	$\mathbb{U}$-metric		E	442	319	$\mathbb{U}_{D(\alpha)}$-metric
D	262	295	$\mathbb{U}_{D(\alpha)}$-metric		E	443	315	$\mathbb{U}_{D(\alpha)}$-metric

Table 6.2: 72 CMF**-significant residue sites found by** $\mathbb{U}$ **and** $\mathbb{U}_{D(\alpha)}$**-metric in human GCK protein.**

6.2.1 Position analysis of the Human Epidermal Growth Factor Receptor (EGFR) protein

In order to detect essential mutation positions in the corresponding sequence of human EGFR protein, we determined altogether 43 CMF-significant residue sites (see Table 6.1). 15 of these significant residue sites have been verified as nsSNP sites through the Ensembl database annotation and they are illustrated in Figure 6.3.

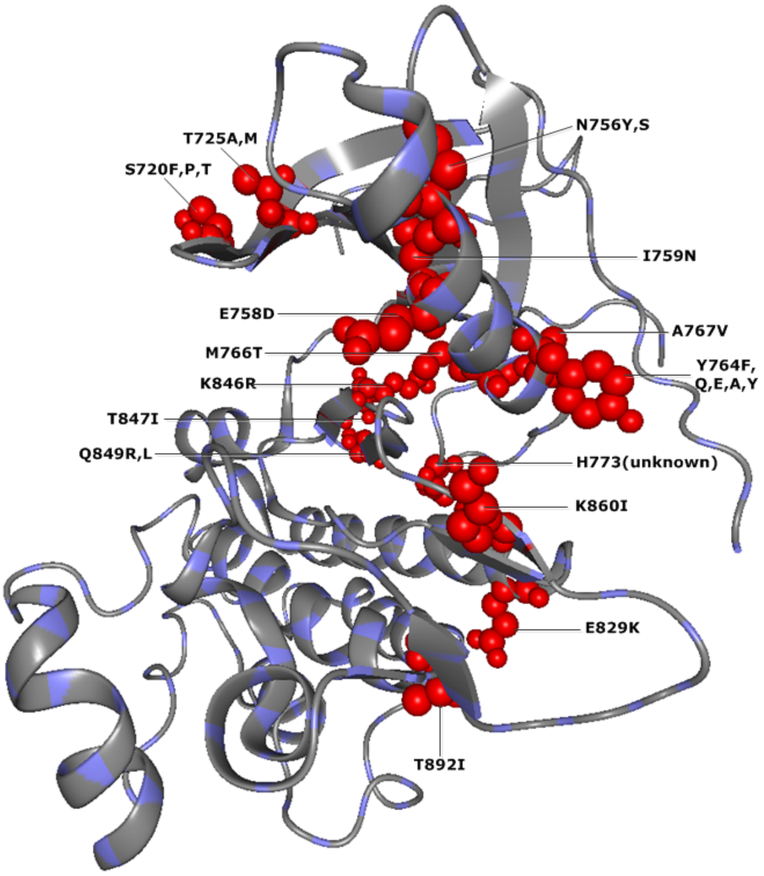


Figure 6.3: CMF**-significant nsSNP positions in human EGFR protein (PDB-Entry 2J6M).** The red spheres correspond to structural localization of 15 different CMF-significant nsSNP positions in the EGFR protein.

Additionally, the CMF-significant sites E746, Q791, and four of the nsSNP positions (I759,Y764,M766 and K846) are also in contact with critical active site regions for the gefitinib binding site in the wild type EGFR kinase [30, 33] (see Figure 6.4).

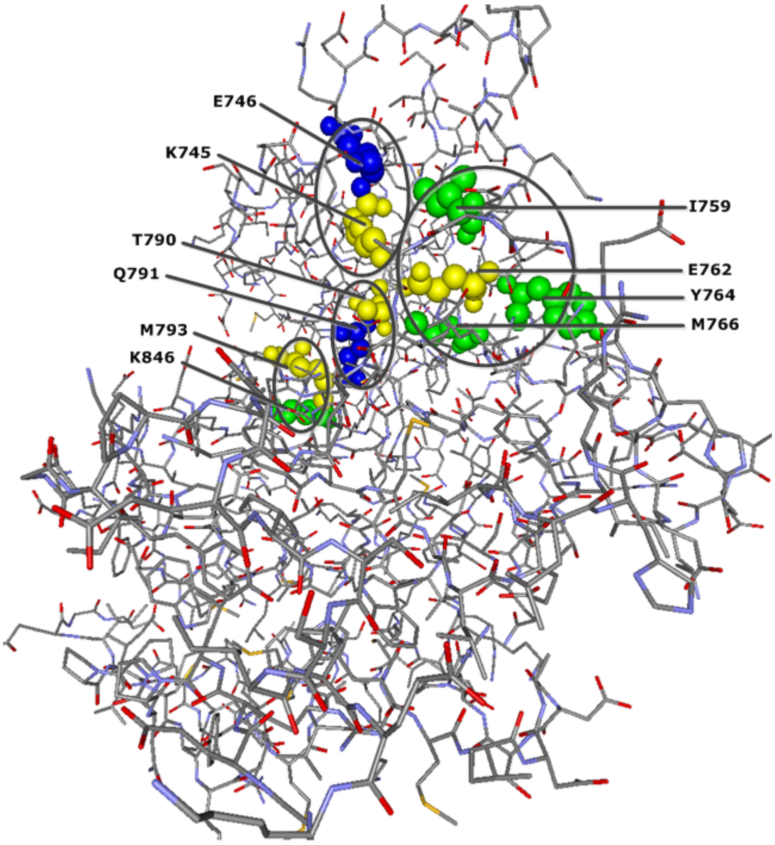


Figure 6.4: CMF-**significant residue positions are in contact with gefitinib binding sites in human EGFR protein (PDB-Entry 2J6M).** Yellow spheres denote positions of the gefitinib binding sites in the wild type kinase. Blue spheres show the localization of CMF-significant adjacent residue positions which are in contact with these binding sites. Additionally, the CMF-significant sites I759, Y764, M766 and K846, shown with green spheres, are already described as nsSNP positions and they are also in contact with gefitinib binding sites E762 and M793, respectively. Circles indicate clusters of gefitinib binding sites and their significant adjacent sites.

Moreover, we observed that 17 further CMF-significant positions are essential sites (see Table 6.3). In total, we have established for EGFR protein the importance of 34 out of 43 CMF-significant residue sites via different resources [30, 33, 39, 40].

CMF-**significant essential sites**	**Nearby nsSNPs or strictly conserved sites**	**Reference**
*Y*727	726^c 743^c	-
		Continued on next page

Table 6.3 – continued from previous page

CMF-**significant essential sites**	**Nearby nsSNPs or strictly conserved sites**	**References**
*H*755	$756^{\mathbf{s}}$, $758^{\mathbf{s}}$	[40]
*D*800	$798^{\mathbf{c}}$	-
*G*824	$773^{\mathbf{s}}$	[40]
*D*830	$829^{\mathbf{s}}$	[40]
*E*868	$892^{\mathbf{s}}$	[40]
*E*872	$873^{\mathbf{s}}$	[39]
*V*876	$877^{\mathbf{c}}$	-
*K*879	$877^{\mathbf{c}}$, $880^{\mathbf{c}}$	-
*Y*891	$892^{\mathbf{s}}$, $895^{\mathbf{c}}$	[40]
*S*899	$880^{\mathbf{c}}$, $896^{\mathbf{c}}$, $898^{\mathbf{c}}$, $901^{\mathbf{c}}$	-
*Y*900	$898^{\mathbf{c}}$, $901^{\mathbf{c}}$	-
*T*909	$906^{\mathbf{c}}$, $936^{\mathbf{c}}$	-
*S*912	$906^{\mathbf{c}}$, 936^{c}	-
*K*913	$914^{\mathbf{c}}$	-
*D*916	$914^{\mathbf{c}}$	-
*M*947	$901^{\mathbf{c}}$, $950^{\mathbf{c}}$	-

Table 6.3: CMF-significant essential sites in human EGFR protein, which are nearby either nsSNPs or strictly conserved sites. $^{\mathbf{s}}$: non-synonymous snp site, $^{\mathbf{c}}$: strictly conserved site.

Although the vast majority of CMF-significant sites are verified to be structurally or functionally important in human EGFR protein, nine CMF-significant sites do not overlap with essential sites. The reason for the high connectivity degree of these unconfirmed significant sites and their role in the EGFR protein is unclear.

6.2.2 Position analysis of the Human Glucokinase (GCK) protein

Applying our both metrics, we determined 72 CMF-significant residue sites to be structurally or functionally important in human GCK protein (see Table 6.2). 16 of these significant residue positions are related to disease associated nsSNP positions [34–36, 39, 40] (see Figure 6.5).

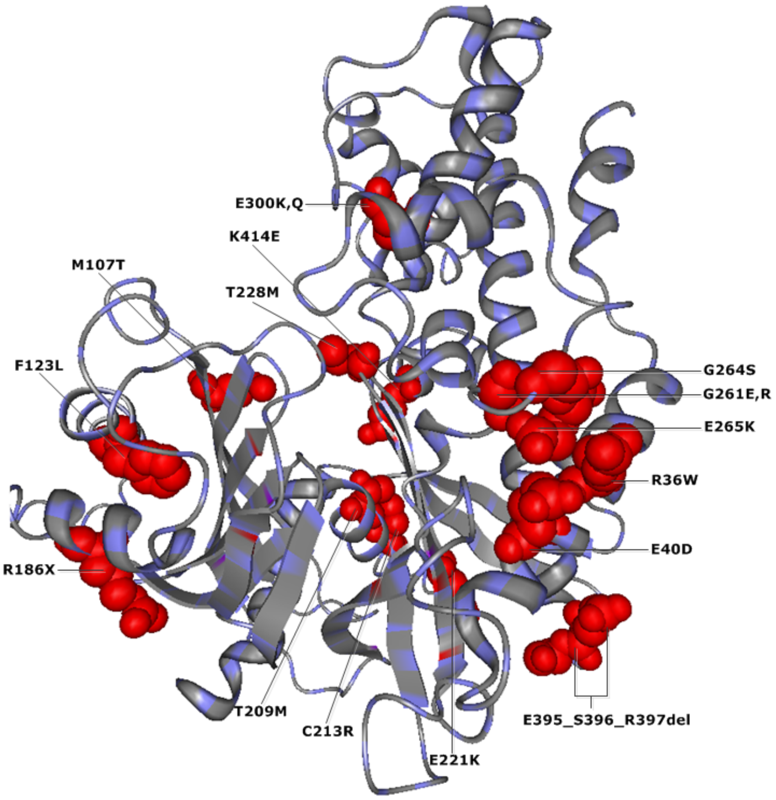


Figure 6.5: CMF-**significant nsSNP positions in human GCK protein (PDB-Entry 1V4S).** Red spheres show the structural localization of 16 different CMF-significant nsSNP positions in the GCK protein.

Furthermore, nine significant sites are found to be in contact with allosteric sites in the GCK protein structure. Among these sites, the R63 is also allosteric site by itself [37] and T209, C213 and E221 overlap with nsSNP regions (see Figure 6.6B). Moreover, the five significant sites T149, F171, T206, Q287, and G294 interact with glucose binding sites K169, D204, N205, and E290 [37] (see Figure 6.6A).

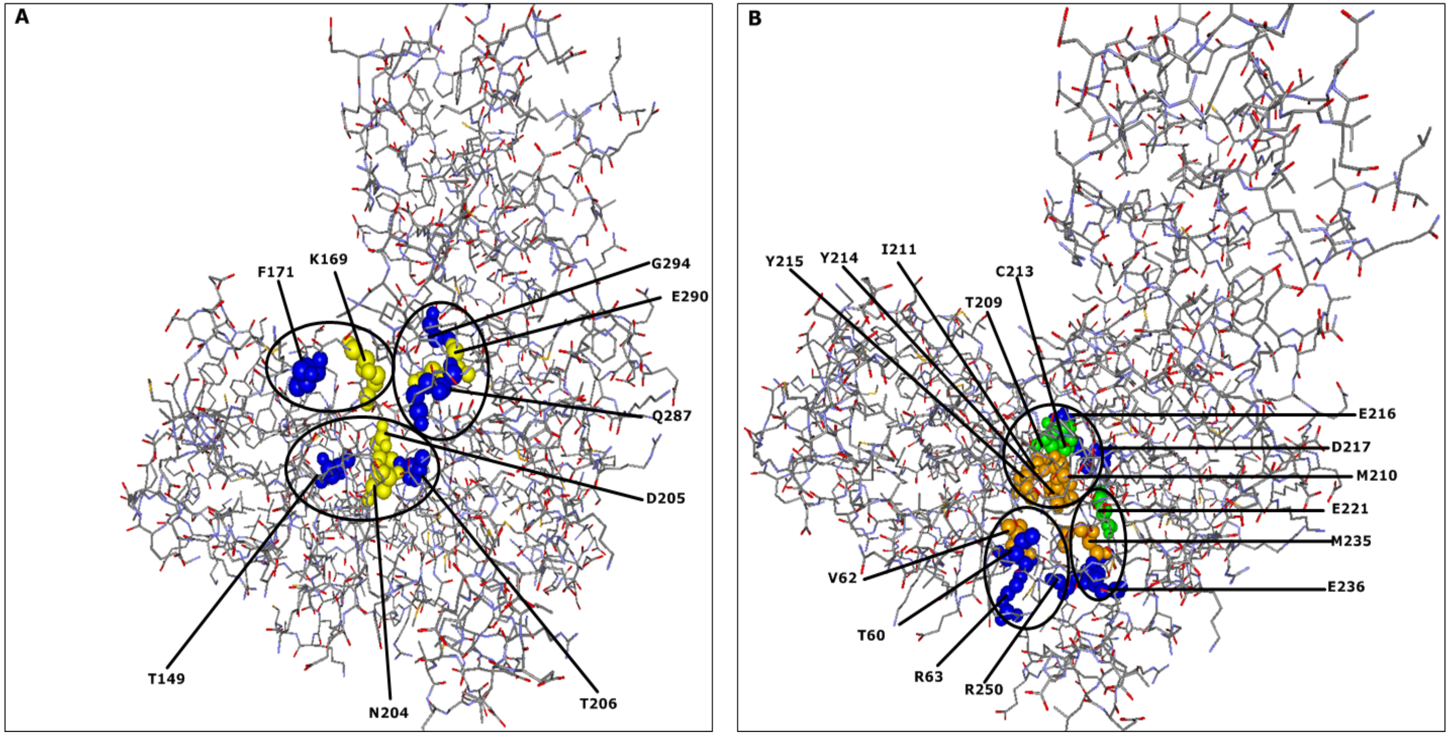


Figure 6.6: CMF-**significant residue positions are in contact with glucose binding site and allosteric site in human GCK protein (PDB-Entry 1V4S).** (A) Yellow spheres show the structural positions of glucose binding sites (active sites). Blue spheres correspond to localization of CMF-significant adjacent residue positions which are in contact with these active sites. (B) Orange spheres denote the allosteric sites. Blue spheres correspond to localization of just significant adjacent residue positions and green spheres indicate the significant residue positions which are already described as nsSNP positions and in contact with these allosteric sites. Additionally, the significant position R63 is allosteric site by itself and it is also in contact with another allosteric site. Circles indicate clusters of glucose binding sites (A), allosteric sites (B), and their significant adjacent sites.

Besides this, there are further 30 CMF-significant essential sites which are nearby nsSNPs or strictly conserved residue positions (see Table 6.4). Altogether, we showed the functionality of 57 positions out of 72 CMF-significant residue sites via different resources [34–40].

CMF-significant essential sites	Nearby nsSNPs or strictly conserved sites	References
$M34$	36^{s}	[39]
$T65$	66^{c}	
$E67$	66^{c}, 68^{c}	-
$T82$	81^{c}	-
$N83$	81^{c}, 108^{s}, 110^{s}	[39]
$H105$	106^{s}	[34]
$C129$	131^{s}, 132^{s}	[34,39]
$F133$	131^{s}, 132^{s}	[34,39]
$F148$	147^{c}, $150^{c,s}$	[39]
$F152$	$150^{c,s}$, 151^{c}	[39]
$H156$	162^{s}	[34]
$N180$	162^{s}, 182^{s}	[34,39]
$F260$	257^{s}, 258^{c}, 259^{s}, 261^{s}	[39]
$D262$	259^{s}, 261^{s}, 264^{s}	[39,92]
$L266$	261^{s}, 264^{s}, 265^{s}	[34,39,92]
$D267$	264^{s}, 265^{s}	[34,92]
$L271$	274^{c}	-
$S281$	278^{c}, 279^{s}	[39]
$Q286$	259^{s}	[39]
$E331$	$299^{c,s}$	[39]
$T332$	295^{c}, $299^{c,s}$	[39]
$R333$	336^{s}	[39]
$Q337$	336^{s}	[39]
$E339$	336^{s}	[39]
$N391$	392^{s}	[34,39]
$S411$	$227^{c,s}$, 410^{c}, 414^{s}	[39]
$S418$	416^{s}	[35]
$F419$	416^{s}	[35]
$E442$	444^{c}	-
$E443$	444^{c}, 445^{c}	-

Table 6.4: CMF-significant essential sites in human GCK protein, which are nearby either nsSNPs or strictly conserved sites. ^{s} : non-synonymous snp site, ^{c} : strictly conserved site.

While we are able to establish the large number of CMF-significant sites as structurally or functionally important in human GCK protein, 15 CMF-significant sites do not overlap with essential sites. Their importance in the GCK protein and the reason of high connectivity degree of these unconfirmed significant sites has not been determined yet.

6.2.3 A comparison between $\mathbb{U}$-metric and $\mathbb{U}_{D(\alpha)}$-metric

Similarities in physical or biochemical properties of amino acids are likely to be crucial for the detection of functionally or structurally important positions of a protein. In contrast to the $\mathbb{U}$-metric, which is a normalized mutual information that uses only the frequencies of occurrences of amino acids in the MSA columns, the novel $\mathbb{U}_{D(\alpha)}$-metric includes dissimilarities according to the *BLOSUM62* matrix when calculating normalized mutual information. As a result, the positions which have undergone dissimilar compensatory mutations are upscaled.

Having applied the $\mathbb{U}$-metric as well as the $\mathbb{U}_{D(\alpha)}$-metric to human EGFR and GCK proteins, the $\mathbb{U}_{D(\alpha)}$-metric has shown better sensitivity and specificity. However, only when we use the both metrics together, the sensitivity is significantly increased, whereas the specificity is only moderately decreased. The details are presented in Table 6.5.

	Sensitivity	Specificity
$\mathbb{U}$-significance	9.7%	91.5%
$\mathbb{U}_{D(\alpha)}$-significance	12.4%	97.2%
CMF-significance	22.1%	88.7%

Table 6.5: **Comparison between $\mathbb{U}$-metric and $\mathbb{U}_{D(\alpha)}$-metric**

It is important to note that the two metrics complement each other, thus we propose to use them together.

6.3 Applying the $\mathbb{Q}_{\text{ent}}$ and the $\mathbb{Q}_{\text{sep}}$-metric for the human EGFR and GCK proteins

In Section 4.2, we pointed out that column pair metric distributions on the interval $[0,1]$ can be modeled based on distributions of background signals F_0, unrelated pair signals G_1, and significant pair signals G_2, where F_0 is a β-distribution. Let $\mathbb{Q}$ be one of the two new metrics introduced in Section 5.2. The p-values $\big(1-F_0(\mathbb{Q}_i)\big)$, $i=1,2,\dots$ are uniformly distributed over $[0,1]$, if $\mathbb{Q}$-values are F_0-distributed. In contrast, p-values tend to zero or one, if $\mathbb{Q}$-values are G_2-distributed or G_1-distributed, respectively (see Figure 6.7). By slightly generalizing the Storey-Tibshirani procedure devised for multiple testing problems [27, 28, 75] to quantify the error made in terms of false discovery rate (FDR), we have developed in our MSA-specific statistical model which successfully separates significant signals of column pairs from background signals and unrelated pair signals. In order to determine significant $\mathbb{Q}$-values of column pairs, we also apply that model in this section. After that, we utilized the connectivity degree technique (see Section 4.2.2) in order to characterize the significance of individual residue sites.

Let M be an MSA, where the protein of interest is represented by M's first row. The definition of a $(\mathbb{Q},M)$-significant site pair (i,j) of the protein is completely analogous to the corresponding notion presented in the previous Section 6.2. A residue site of the protein is defined to be QCMF-significant[9] with respect to the MSA M, if it is $(\mathbb{Q}_{\text{ent}},M)$-significant or $(\mathbb{Q}_{\text{sep}},M)$-significant and its connectivity degree *cut-off* exceeds the 90-th percentile in $(\mathbb{Q}_{\text{ent}},M)$-significant or $(\mathbb{Q}_{\text{sep}},M)$-significant pairs.

In order to compare the performance and functionality of our quantum information theory based method with our classical information theory based method, we tested again the same human proteins using their related MSAs for the evaluation of QCMF-significant residue sites. Likewise, we define here a QCMF-significant residue site as "functionally or structurally important" if it corresponds to one of the essential sites.

6.3.1 Position analysis of the Human EGFR protein

Using the MSA-specific statistical model with a false discovery rate (FDR) of 1% for both $\mathbb{Q}_{\text{ent}}$ and $\mathbb{Q}_{\text{sep}}$-metrics, we first determined altogether 2688 out of 26079 non-conserved column pairs as significant in corresponding MSA of human EGFR protein. 631 of these significant pairs are detected by the $\mathbb{Q}_{\text{ent}}$-metric and 2149 pairs are observed by the $\mathbb{Q}_{\text{sep}}$-metric. Only 92 significant column pairs were detected by both metrics. After that, utilizing the connectivity degree technique, we predicted a total of 33 residue sites in the corresponding sequence of the human EGFR protein as QCMF-significant (see Table 6.6). 12 of them

---

[9] We defined the predicted significant sites together as QCMF-significant due to our corresponding publication "**<u>Q</u>**uantum **<u>C</u>**oupled **<u>M</u>**utation **<u>F</u>**inder" [5].

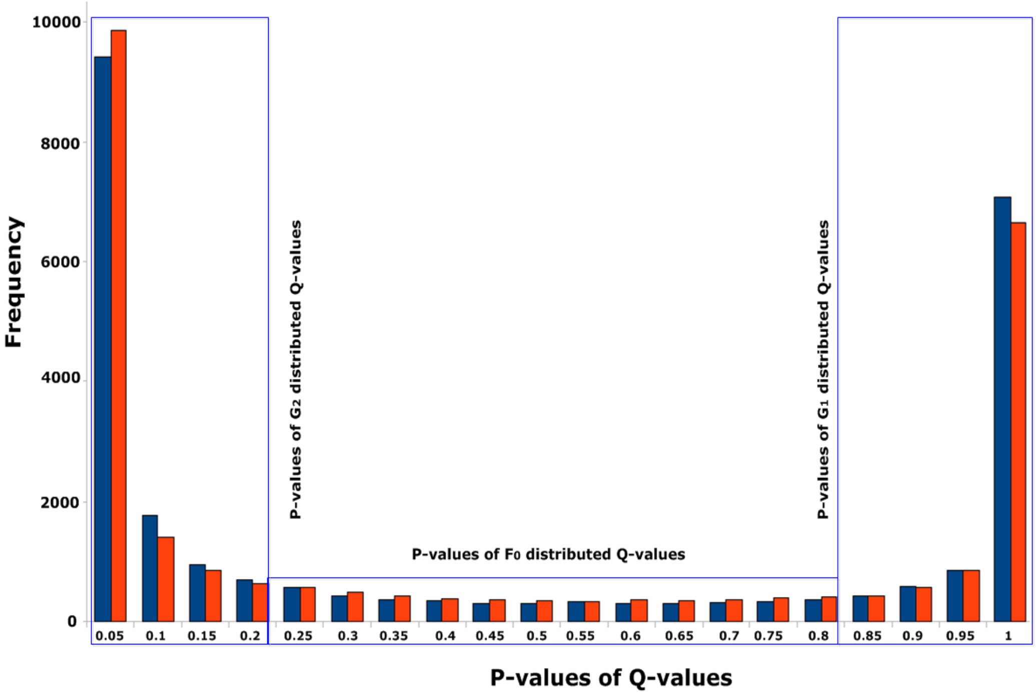


Figure 6.7: **p-value distributions of $\mathbb{Q}_{ent}$ and $\mathbb{Q}_{sep}$-values for human EGFR protein (PDB-Entry 2J6M).** The blue bars illustrate the p-value distribution of the $\mathbb{Q}_{ent}$-values and red bars display the p-value distribution of the $\mathbb{Q}_{sep}$-values.

are only $\mathbb{Q}_{ent}$-significant and 18 residue sites are $\mathbb{Q}_{sep}$-significant, the remaining 3 residue sites (A839,A882 and V902) are determined both $\mathbb{Q}_{ent}$-significant and $\mathbb{Q}_{sep}$-significant.

10 of the QCMF-significant residue sites are in contact with either catalytic residues or critical active site regions for gefitinib binding site in wild type EGFR kinase [30, 33, 58] (see Figure 6.8 and Figure 6.9). Among these sites, the A839 and R841 have been verified as catalytic residue sites through Catalytic Site Atlas [58], the T854 is a gefitinib binding site by itself and the residue sites V845 and A859 are also in contact with nsSNP positions K846, T847 and K860 in human EGFR protein. Moreover, two out of all 33 significant sites are related to disease associated nsSNP positions and their structural localization are illustrated in Figure 6.8.

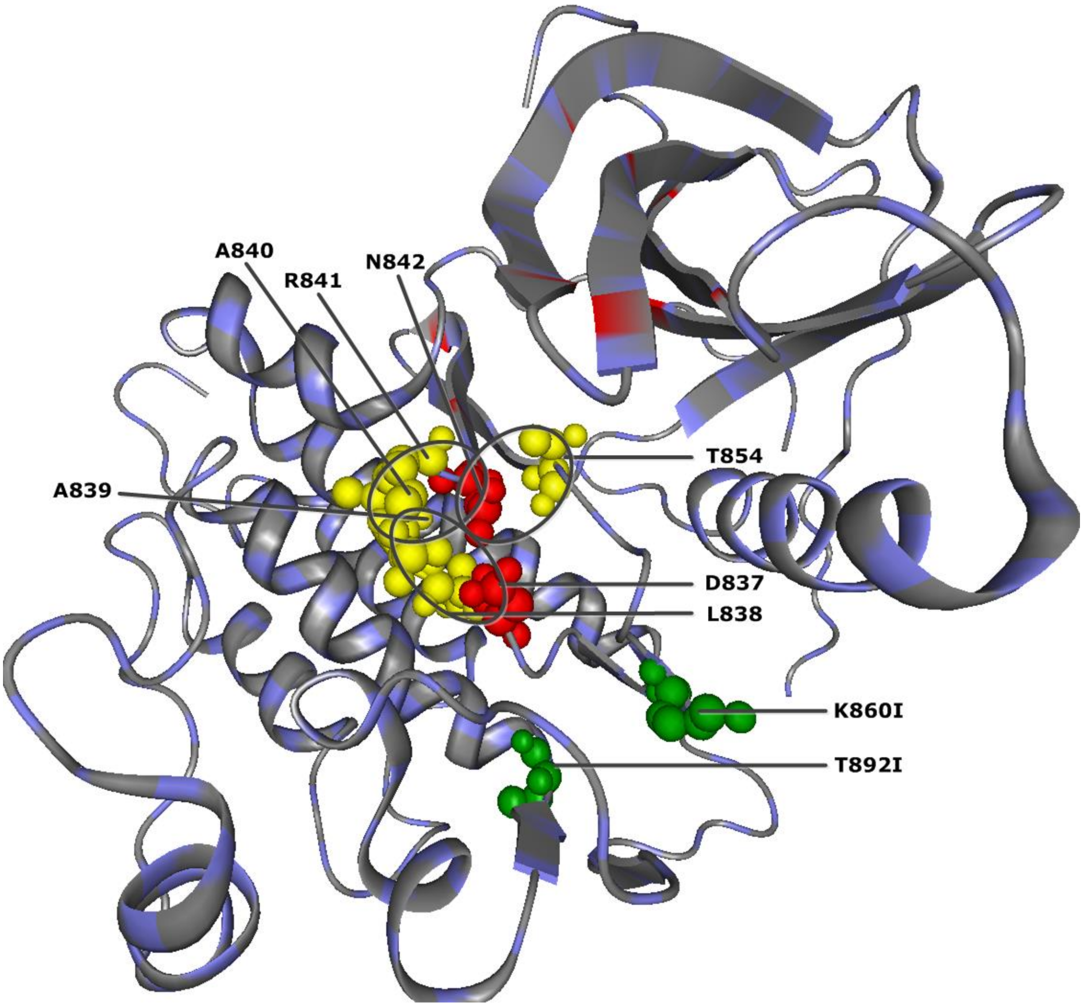


Figure 6.8: QCMF-**significant residue positions are in contact with catalytic residues in human EGFR protein (PDB-Entry 2J6M).** Red spheres denote positions of the catalytic residues. Yellow spheres show the localization of QCMF-significant adjacent residue positions which are in contact with these catalytic residues. Moreover, the QCMF-significant sites A839 and R841 are also catalytic residues by themselves. Green spheres show the structural localization of QCMF-significant nsSNP positions in the EGFR protein. Circles indicate clusters of catalytic residue sites and their significant adjacent sites.

Residue	Position	Connectivity Degree	Detected by		Residue	Position	Connectivity Degree	Detected by
G	729	62	$\mathbb{Q}_{sep}$-metric		L	844	62	$\mathbb{Q}_{sep}$-metric
T	751	18	$\mathbb{Q}_{ent}$-metric		V	845	97	$\mathbb{Q}_{ent}$-metric
N	771	57	$\mathbb{Q}_{sep}$-metric		I	853	21	$\mathbb{Q}_{ent}$-metric
G	779	70	$\mathbb{Q}_{sep}$-metric		T	854	24	$\mathbb{Q}_{ent}$-metric
Q	791	77	$\mathbb{Q}_{sep}$-metric		A	859	24	$\mathbb{Q}_{ent}$-metric
I	792	22	$\mathbb{Q}_{ent}$-metric		K	860	78	$\mathbb{Q}_{sep}$-metric
Q	820	69	$\mathbb{Q}_{sep}$-metric		A	882*	26	$\mathbb{Q}_{ent}$-metric
A	822	35	$\mathbb{Q}_{ent}$-metric		A	882*	54	$\mathbb{Q}_{sep}$-metric
G	824	89	$\mathbb{Q}_{sep}$-metric		Y	891	113	$\mathbb{Q}_{sep}$-metric
M	825	72	$\mathbb{Q}_{sep}$-metric		T	892	45	$\mathbb{Q}_{ent}$-metric
Y	827	119	$\mathbb{Q}_{sep}$-metric		Y	900	132	$\mathbb{Q}_{sep}$-metric
L	828	52	$\mathbb{Q}_{sep}$-metric		V	902*	74	$\mathbb{Q}_{ent}$-metric
V	834	54	$\mathbb{Q}_{ent}$-metric		V	902*	79	$\mathbb{Q}_{sep}$-metric
L	838	52	$\mathbb{Q}_{sep}$-metric		T	909	42	$\mathbb{Q}_{ent}$-metric
A	839*	43	$\mathbb{Q}_{ent}$-metric		G	911	66	$\mathbb{Q}_{sep}$-metric
A	839*	55	$\mathbb{Q}_{sep}$-metric		L	927	24	$\mathbb{Q}_{ent}$-metric
A	840	128	$\mathbb{Q}_{ent}$-metric		G	930	71	$\mathbb{Q}_{sep}$-metric
R	841	68	$\mathbb{Q}_{sep}$-metric		Y	944	51	$\mathbb{Q}_{sep}$-metric

Table 6.6: 33 QCMF**-significant residue sites found by $\mathbb{Q}_{ent}$-metric and $\mathbb{Q}_{sep}$-metric in the human EGFR protein.** * **denotes the residue sites are detected by both metrics $\mathbb{Q}_{ent}$ and $\mathbb{Q}_{sep}$ as significant.**

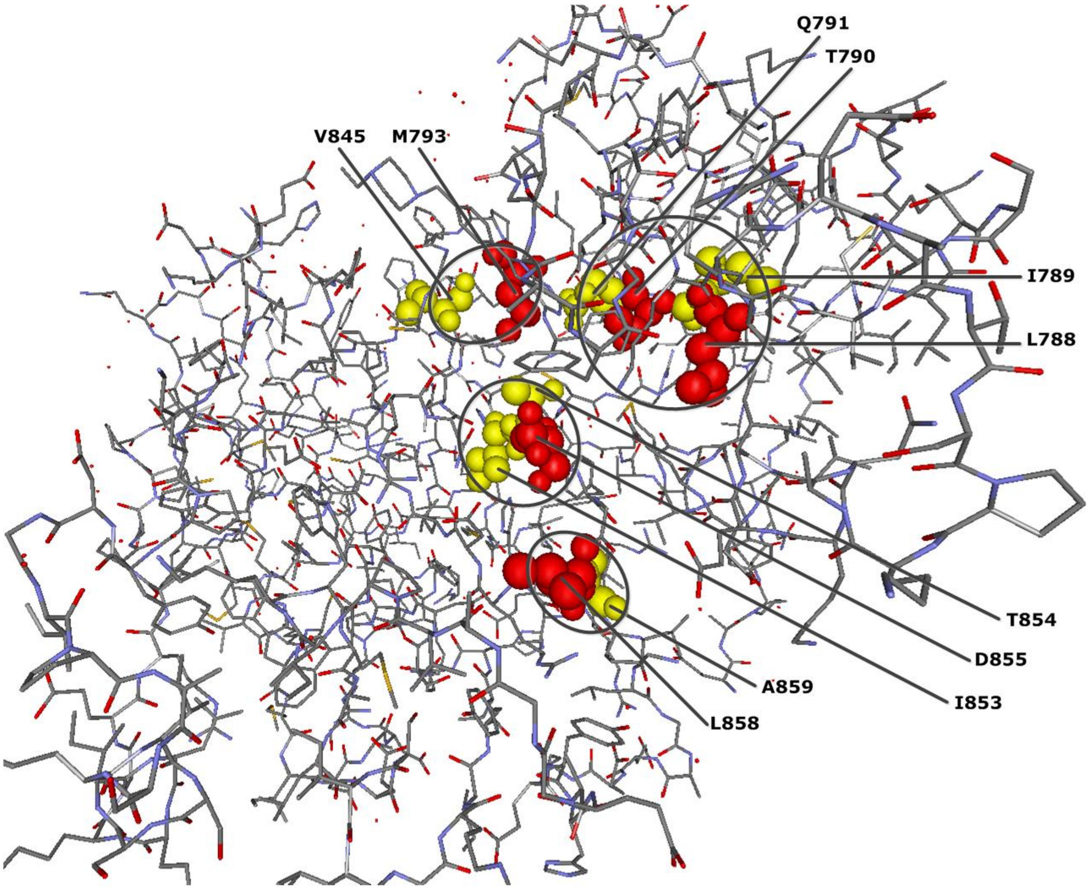


Figure 6.9: QCMF**-significant residue positions are in contact with gefitinib binding sites in human EGFR protein (PDB-Entry 2J6M).** Red spheres show the structural localization of the gefitinib binding sites in the wild type kinase. Yellow spheres show significant adjacent residue positions which are in contact with these binding sites. Moreover, the QCMF-significant site T854 is also binding site by itself and interacts with gefitinib binding sites D855. Circles indicate clusters of gefitinib binding sites and their significant adjacent sites.

Additionally, 13 of out of all QCMF-significant sites are defined as essential sites since they are either nearby strictly conserved residues or nsSNPs (see Table 6.7).

QCMF-**significant essential sites**	**Nearby nsSNPs or strictly conserved sites**	**References**
$N771$	773^{s}	[40]
$G824$	773^{s}	[40]
$Y827$	829^{s}	[40]
$L828$	829^{s}	[40]
$V834$	835^{c}, 836^{s}, 860^{s}	[40,92]
$Y891$	892^{s}, 895^{c}	[40]
$A822$	861^{s}	[29,39,92]
$V844$	796^{c}, 798^{c}, 852^{c}	-
$A882$	884^{c}, 895^{c}, 898^{c}	-
$Y900$	898^{c}, 901^{c}	-
$V902$	880^{c}, 901^{c}	-
$T909$	906^{c}, 936^{c}	-
$G911$	906^{c}	-

Table 6.7: QCMF-significant essential sites in human EGFR protein, which are nearby either nsSNPs or strictly conserved sites. ^{s} : non-synonymous snp site, ^{c} : strictly conserved site.

According to the essential sites of human EGFR protein, we have shown altogether the structural or functional importance of 25 QCMF-significant sites. The remaining 8 significant residue sites (G729, T851, G779, Q820, M825, L927, G930, Y944) do not fall into essential sites and the reason for their significance and their importance in the EGFR protein is currently unclear.

Finally, we compared the new QCMF-significant residue sites for human EGFR protein with our previous CMF-significant residue sites. The new significant residue sites Q791, Q820, G824, K860, Y891, T892, Y900, T909 overlap with CMF-significant residue sites. Interestingly, one of the unconfirmed residue sites (Q820) has been predicted as both QCMF-significant and CMF-significant.

6.3.2 Position analysis of the the Human GCK protein

Like human EGFR protein, applying the MSA-specific statistical model with a FDR of 1% for both $\mathbb{Q}_{\text{ent}}$ and $\mathbb{Q}_{\text{sep}}$-metrics we identified a total of 9853 out of 69645 non-conserved column pairs as significant in human GCK protein. 6070 of them are $(\mathbb{Q}_{\text{ent}}, M)$-significant and 4232 are detected as $(\mathbb{Q}_{\text{sep}}, M)$-significant. Only 449 column pairs are determined as significant with respect to both metrics. Thereupon using the connectivity degree technique, we determined altogether 64 residue sites in human GCK protein as QCMF-significant (see Table 6.8). 30 of them are observed as $\mathbb{Q}_{\text{ent}}$-significant and 30 significant residue sites are detected as $\mathbb{Q}_{\text{sep}}$-significant. Only four residue sites (T82, G223, V253, G407) are defined as significant based on both metrics.

Residue	Position	Connectivity Degree	Detected by		Residue	Position	Connectivity Degree	Detected by
M	37	75	$\mathbb{Q}_{sep}$-metric		G	223*	97	$\mathbb{Q}_{sep}$-metric
T	60	177	$\mathbb{Q}_{ent}$-metric		G	223*	63	$\mathbb{Q}_{ent}$-metric
Y	61	87	$\mathbb{Q}_{sep}$-metric		T	228	310	$\mathbb{Q}_{ent}$-metric
V	62	65	$\mathbb{Q}_{ent}$-metric		A	232	184	$\mathbb{Q}_{ent}$-metric
S	76	297	$\mathbb{Q}_{ent}$-metric		C	233	241	$\mathbb{Q}_{ent}$-metric
L	79	70	$\mathbb{Q}_{sep}$-metric		E	236	153	$\mathbb{Q}_{sep}$-metric
T	82*	95	$\mathbb{Q}_{sep}$-metric		V	253*	89	$\mathbb{Q}_{sep}$-metric
T	82*	318	$\mathbb{Q}_{ent}$-metric		V	253*	64	$\mathbb{Q}_{ent}$-metric
N	83	108	$\mathbb{Q}_{sep}$-metric		F	260	84	$\mathbb{Q}_{sep}$-metric
V	86	144	$\mathbb{Q}_{ent}$-metric		L	271	171	$\mathbb{Q}_{ent}$-metric
V	89	67	$\mathbb{Q}_{ent}$-metric		V	277	107	$\mathbb{Q}_{ent}$-metric
F	123	161	$\mathbb{Q}_{sep}$-metric		S	281	99	$\mathbb{Q}_{sep}$-metric
S	127	296	$\mathbb{Q}_{ent}$-metric		N	283	91	$\mathbb{Q}_{sep}$-metric
F	148	101	$\mathbb{Q}_{sep}$-metric		Q	287	96	$\mathbb{Q}_{sep}$-metric
T	149	318	$\mathbb{Q}_{ent}$-metric		G	294	62	$\mathbb{Q}_{ent}$-metric
F	152	207	$\mathbb{Q}_{sep}$-metric		Y	297	178	$\mathbb{Q}_{sep}$-metric
P	153	95	$\mathbb{Q}_{sep}$-metric		M	298	81	$\mathbb{Q}_{sep}$-metric
H	156	68	$\mathbb{Q}_{sep}$-metric		E	300	82	$\mathbb{Q}_{sep}$-metric
G	162	95	$\mathbb{Q}_{sep}$-metric		T	332	89	$\mathbb{Q}_{ent}$-metric
G	170	82	$\mathbb{Q}_{sep}$-metric		V	374	182	$\mathbb{Q}_{ent}$-metric
F	171	144	$\mathbb{Q}_{sep}$-metric		A	378	302	$\mathbb{Q}_{ent}$-metric
G	175	90	$\mathbb{Q}_{sep}$-metric		A	379	292	$\mathbb{Q}_{ent}$-metric
A	176	63	$\mathbb{Q}_{ent}$-metric		S	383	198	$\mathbb{Q}_{ent}$-metric
G	178	125	$\mathbb{Q}_{sep}$-metric		A	384	55	$\mathbb{Q}_{ent}$-metric
N	180	76	$\mathbb{Q}_{sep}$-metric		A	387	59	$\mathbb{Q}_{ent}$-metric
L	185	96	$\mathbb{Q}_{sep}$-metric		G	388	70	$\mathbb{Q}_{ent}$-metric
A	201	308	$\mathbb{Q}_{ent}$-metric		G	407*	75	$\mathbb{Q}_{sep}$-metric
M	202	105	$\mathbb{Q}_{ent}$-metric		G	407*	143	$\mathbb{Q}_{ent}$-metric
T	206	300	$\mathbb{Q}_{ent}$-metric		V	412	218	$\mathbb{Q}_{ent}$-metric
V	207	303	$\mathbb{Q}_{ent}$-metric		F	419	110	$\mathbb{Q}_{sep}$-metric
A	208	145	$\mathbb{Q}_{sep}$-metric		E	443	110	$\mathbb{Q}_{sep}$-metric
T	209	319	$\mathbb{Q}_{ent}$-metric		G	446	86	$\mathbb{Q}_{sep}$-metric
M	210	85	$\mathbb{Q}_{sep}$-metric		L	451	63	$\mathbb{Q}_{ent}$-metric
Y	215	104	$\mathbb{Q}_{sep}$-metric		S	453	317	$\mathbb{Q}_{ent}$-metric

Table 6.8: 64 QCMF-**significant residue sites found by $\mathbb{Q}_{ent}$-metric and $\mathbb{Q}_{sep}$-metric in human GCK protein.** * **denotes the residue sites are detected by both metrics $\mathbb{Q}_{ent}$ and $\mathbb{Q}_{sep}$ as significant.**

13 out of all QCMF-significant residue sites are in contact with allosteric sites V62, R63, M210, I211, Y214, Y215, M235, V452, V455 and A456 in human GCK protein. Among these significant sites, the *V*62, *M*210, *Y*215 are allosteric sites by themselves [37] and the T209M, G223S and S453del are related to disease associated nsSNP positions. In addition, there are further five QCMF-significant sites F123L, G162D, G175R, T228M and E300K,Q that have been verified as nsSNP positions through annotation databases and previous experimental studies [34–36,38,39,93]. The structural localization of these 18 QCMF-significant residue sites (contact sites and nsSNPs positions) are illustrated in Figure 6.10.

Additionally, eight QCMF-significant sites T149, G170, F171, T206, V207, A208, Q287 and G294 interact with glucose binding sites (active sites) T168, K169, D204, D205 and E290 in human GCK protein [37] (see Figure 6.11) where V207 and A208 are also in contact with allosteric sites M210 and I211.

Moreover, we have observed that 38 QCMF-significant sites are further included in essential sites since they are near nsSNPs or strictly conserved residues in human GCK protein (see Table 6.9).

QCMF-**significant essential sites**	**Nearby nsSNPs or strictly conserved sites**	**References**
*M*37	36^{s}, 39^{s}, 40^{s}	[34,35,39,93]
*S*76	147^{c}	-
*L*79	78^{c}, 80^{c}, 150^{c}	-
*T*82	81^{c}	-
*N*83	81^{c}, 108^{s}, 110^{s}	[39,93]
*V*86	85^{c}, 106^{s}	[34]
*S*127	130^{s}	[36]
*F*148	147^{c}, $150^{c,s}$	[34,35,39,93]
*F*152	$150^{c,s}$, 151^{c}	[35,39,93]
*P*153	154^{s}	[35]
*H*156	154^{s}	[35]
*A*176	119^{s}, 175^{s}	[39]
*G*178	164^{c}	-
*N*180	162^{s}, 182^{s}	[34,35,39],
*L*185	182^{s}, 188^{s}	[35,39,93]
*A*201	147^{c}, 453^{c}	-
*M*202	147^{c}, 203^{s}	[39]
*A*232	223^{s}, 231^{c}	[35,36,93]
*C*233	223^{s}, 234^{c}, 235^{c}	[35,36,93]
*V*253	234^{c}, 254^{c}	-
*F*260	257^{s}, 258^{c}, 259^{s}, 261^{s}	[35,39]
*L*271	274^{c}	-
*V*277	274^{c}, 278^{c}, 279^{s}	[39]
*S*281	278^{c}, 279^{s}	[39]
*Y*297	291^{c}, 295^{c}, 299^{c}, 300^{s}	[39]
*M*298	295^{c}, 299^{c}, 300^{s}	[39]
*T*332	295^{c}, 299^{c}	-

Continued on next page

Table 6.9 – continued from previous page

QCMF-significant essential sites	Nearby nsSNPs or strictly conserved sites	References
*V*374	377^{c}	-
*A*378	377^{c}, 382^{s}	[39]
*A*379	377^{c}, 382^{s}	[39]
*S*383	382^{s}, 385^{s}	[39]
*A*384	382^{s}, 385^{s}	[39]
*A*387	385^{s}	[39]
*S*388	385^{s}, 392^{s}	[34,39]
*V*412	226^{s}, 227^{c}, 410^{c}, 414^{s}, 416^{s}	[36,39]
*F*419	416^{s}	[36]
*E*443	444^{c}, 445^{c}, 447^{s}	[35]
*G*446	444^{c}, 445^{c}, 447^{s}, 448^{c}, 449^{c}	[35]

Table 6.9: QCMF-significant essential sites in human GCK protein, which are nearby either nsSNPs or strictly conserved sites. ^{s} : non-synonymous snp site, ^{c} : strictly conserved site.

In total, we have demonstrated here that according to the essential sites, 62 out of 64 QCMF-significant residue sites are functionally or structurally important for human GCK protein. The remaining two significant residue sites V89, N283 do not overlap with essential sites and the reason for their significance and their role in the GCK protein has not been determined yet.

Lastly, we compared the new QCMF-significant residue sites with the previous CMF-significant sites. The unconfirmed residue site N283 and further 23 significant sites (T60, T82, N83, F123, F148, T149, F152, H156, F171, N180, T206, T209, T228, E236, G260, L271, S281, Q287, G294, E300, T332, F419 and E443) were also determined by the previous method as significant.

6.4 A Comparison between CMF-significant sites and QCMF-significant sites

To further investigate the performance of our methods, we made a statistical comparison between classical information theory based CMF and quantum information theory based QCMF. The CMF method mainly focuses on significant *BLOSUM62*-dissimilar amino acid signals as a model of compensatory mutations and integrates them in the calculation of metrics. As a consequence of only taking into account dissimilar amino acid signals, an important part of its significant sites are verified as disease associated nsSNP positions and just a small part of them were located at or near other essential sites in both proteins. In contrast, considerung simultinously *BLOSUM62*-similar and dissimilar amino acid signals, the result of our new quantum information theory based method is more sensible to catalytic sites, allosteric sites and binding sites.

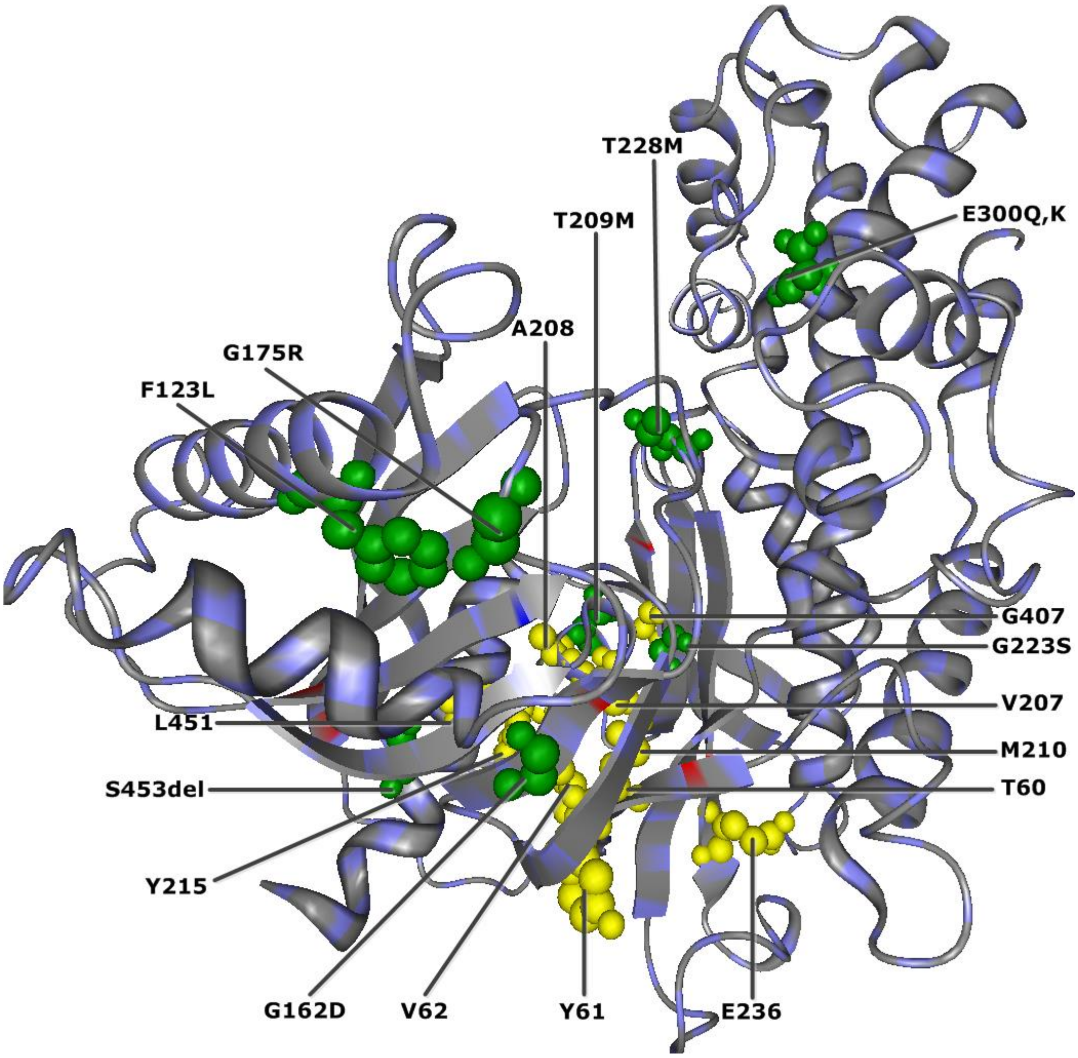


Figure 6.10: QCMF-**significant positions that are either in contact with allosteric sites or related to nsSNPs in human GCK protein (PDB-Entry 1V4S).** Yellow spheres correspond to structural localization of ten QCMF-significant residue sites which are in contact with allosteric sites where the V62, M210, Y215 are denoted as allosteric sites by themselves and they are also in contact with an other allosteric sites. Green spheres indicate eight QCMF-significant nsSNP positions in the GCK protein. Three of them (T209M, G223S and S453del) are further in contact with allosteric sites M210, I211, V452, V455 and A456.

Moreover, when statistically evaluating both methods, we have observed that the new quantum information theory based method significantly outperforms the previous method. The new method reaches an improved performance in identifying essential sites from MSAs of both proteins with a significantly higher Matthews correlation coefficient (MCC) value of 0.215 whereas the previous CMF method reaches only a MCC value of 0.133.

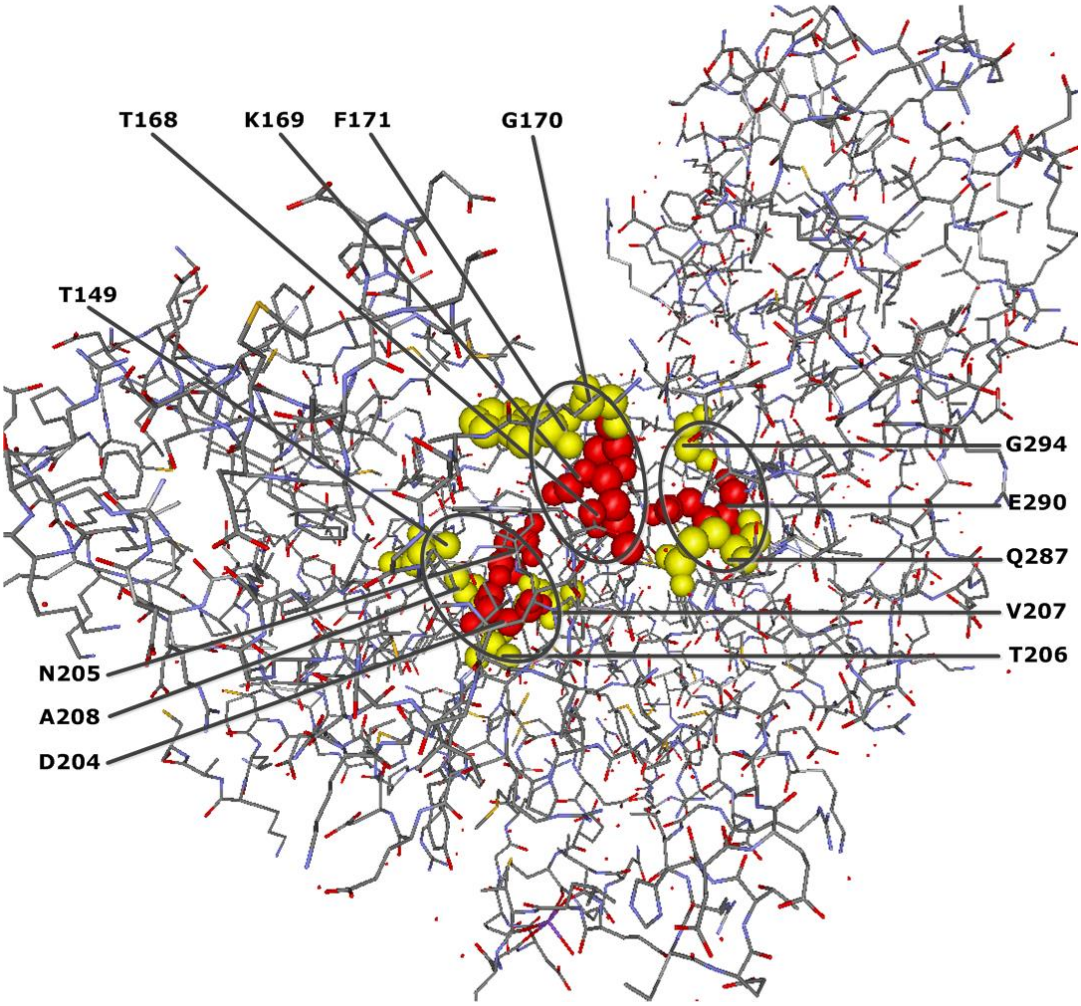


Figure 6.11: QCMF-**significant residue positions are in contact with glucose binding site in human GCK protein (PDB-Entry 1V4S).** (A) Red spheres show the structural positions of the glucose binding sites (active sites) and yellow spheres show the localization of QCMF-significant adjacent residue positions which are in contact with these active sites. The circles indicate clusters of glucose binding sites and their significant adjacent sites.

7 Discussion

In this chapter, we will discuss the reason for setting 90-th percentile *cut-off* for the connectivity degree to evaluate results of our both classical and quantum information theory based methods. Afterwards, we discuss the results presented in the previous chapter in the context of related work.

7.1 Connectivity Degree Cut-off

The connectivity degree technique was introduced by Merkl and Zwick in [22] to characterize an individual residue site. In their study, they used the $\mathbb{U}$-metric, presented in the Section 4.1 and focused on only 75 columns pairs with the highest $\mathbb{U}(k,l)$-values as significant for each MSA. As a result of taking into account only certain number of significant column pair for each MSA, they set a fixed connectivity degree *cut-off* value of 3 for each MSA under investigation. It means if the connectivity degree of a residue site is equal or more than three, the residue is accepted as significant in [22].

However, after applying our MSA-specific significant model, only considering a fixed *cut-off* value for connectivity degree is too conservative and not suitable anymore. Hence, we have further developed the connectivity degree technique by utilizing a confidence interval method to identify a new *cut-off* value based on the Matthews correlation coefficient (MCC).

We have determined for the CMF and QCMF-significant residue sites going through all possible n-th percentiles for $n = 80, 81, \ldots, 99$, the MCC of a joint prediction for human EGFR and GCK protein. The maximal value of MCC is reached for CMF-significant sites if $n = 90$ and likewise, the maximal value of MCC is reached if $n = 88$ for QCMF-significant sites. Although we achieve the best MCC with $n = 88$ for the QCMF-significant sites, we used in the result section 90%-th percentile *cut-off* for the biological evaluation of these significant sites. In this way, we were able to compare both CMF and QCMF-significant residue sites fairly. In Table 7.1 and 7.2 we present the specificity, sensitivity and MCC values of all possible n-th percentiles for result of both methods.

Cut-off	**Specificity**	**Sensitivity**	**MCC**		*Cut-off*	**Specificity**	**Sensitivity**	**MCC**
99%	99.53%	3.16%	0.086		89%	86.38%	23.11%	0.112
98%	97.65%	5.10%	0.065		88%	84.97%	25.30%	0.118
97%	96.24%	6.81%	0.062		87%	82.15%	26.52%	0.096
96%	94.83%	9.00%	0.068		86%	79.34%	27.73%	0.077
95%	93.42%	10.70%	0.067		85%	77.93%	29.68%	0.081
94%	93.42%	13.62%	0.105		84%	76.05%	31.38%	0.077
93%	91.07%	15.08%	0.086		83%	73.23%	32.84%	0.062
92%	90.14%	17.51%	0.101		82%	70.89%	34.06%	0.050
91%	88.73%	19.22%	0.101		81%	69.01%	35.76%	0.047
90%	**88.73%**	**22.14%**	**0.133**		80%	67.13%	37.22%	0.043

Table 7.1: Statistical evaluation of CMF-significant residue sites for both human EGFR and GCK proteins. If connectivity degree *cut-off* is 90-percentile, our classical information theory based method reaches its maximal MCC-value.

Cut-off	**Specificity**	**Sensitivity**	**MCC**		*Cut-off*	**Specificity**	**Sensitivity**	**MCC**
99%	100%	3.16%	0.105		89%	94.83%	22.38%	0.219
98%	100%	5.10 %	0.134		**88%**	**94.36%**	**24.33%**	**0.231**
97%	100%	8.02%	0.170		87%	92.95%	25.30%	0.220
96%	99.06%	9.73%	0.166		86%	91.54%	26.52%	0.212
95%	98.12%	11.43%	0.165		85%	90.14%	27.49%	0.203
94%	97.65%	13.86%	0.182		84%	89.67%	28.22%	0.204
93%	97.18%	15.81%	0.194		83%	89.20%	29.68%	0.212
92%	96.24%	17.51%	0.195		82%	88.73%	31.87%	0.226
91%	96.24%	19.22%	0.211		81%	87.32%	32.84%	0.218
90%	95.30%	21.163%	0.215		80%	86.38%	34.54%	0.222

Table 7.2: Statistical evaluation of QCMF-significant residue sites for both human EGFR and GCK proteins. If connectivity degree *cut-off* is 88-percentile, our quantum information theory based method reaches its maximal MCC-value.

7.2 CMF-significant residue sites

Our results for human EGFR and GCK protein suggest that the large majority of CMF-significant compensatory mutation sites are in agreement with previous experimental studies regarding the functions and stability of these proteins. 15 and 16 CMF-significant sites in human EGRF and GCK proteins, respectively, are verified as disease associated nsSNP positions (see Figures 6.3 and 6.5) where most amino acid substitutions in protein sequences damage structural stability of proteins [41,42,94]. Moreover, we have observed that in both proteins some of CMF-significant nsSNP positions are nearby allosteric sites, binding sites, or catalytic sites each of which are considered to be functionally important [95,96]. Disease associated mutations at these nearby positions are likely to affect protein function [43,97].

Despite the large number of CMF-significant sites demonstrated to be structurally or functionally important for both of the proteins, 9 and 15 significant sites in human EGFR and GCK proteins, respectively, are not included in the essential sites. However, we hypothesize that most of the novel significant sites may play a critical role in both proteins notwithstanding the absence of previous experimental data. Therefore, further progress from the molecular and structural biology end is required not only to assess the importance of these sites, but also for a future perspective on a deeper understanding of the protein structure.

Because we have also used the $\mathbb{U}$-metric, we compared our tool with H2r presented in [22] rather than with those methods developed in [23]. This way, we studied the impact of applying the Storey-Tibshirani procedure in combination with the effect of using the 90-th percentile cut-off for the connectivity degree. We have applied H2r without adding pseudo counts to the human EGFR and GCK protein. For EGFR, the 14 sites T725, A755, N756, A767, Q791, V802, N816, V819, K846, V876, M881, K913, D916, and E931 are identified as significant. Out of these significant sites, ten of these residue sites T725, A755, N756, A767, Q791,K846, V876, M881, K913, and D916 are essential sites. On the other hand, for GCK, H2r identified the 15 residue positions L25, R36, R63, M107, C213, V226, G261, D262, G264, L266, D267, E268, T405, K414, and H416 as significant. Twelve of these sites, namely R36, R63, M107, C213, V226, G261,D262, G264, L266, D267, K414, and H416, are essential sites. However, when using the H2r Web service (`http://www-bioinf.uni-regensburg.de/`) to analyze EGFR and GCK proteins, sensitivity is decreased, while precision is increased. By this service only eight sites for EGFR and nine sites for GCK were found to be significant. Moreover, only five and eight of them are verified as functionally or structurally important for EGFR and GCK proteins, respectively. This difference stems from the fact that the H2r Web service tightens the filtering of the columns. In addition to this, statistically evaluating H2r for EGFR and GCK proteins, we observed a sensitivity of 5.4%, specificity of 96.7%, and a MCC value of 0.047.

The results of this comparison show that a vast majority of functionally or structurally important residue positions cannot be detected without using our novel MSA specific model and both metrics ($\mathbb{U}$ and $\mathbb{U}_{D(1)}$) together.

7.3 QCMF-significant residue sites

The results of our new quantum information theory based method presented in the Section 6.3 show that the vast majority of QCMF-significant residue sites are closely related to functionality and structural stability of both human EGFR and GCK proteins. 10 significant residue sites in EGFR and 19 significant sites in GCK are established as functionally important since they are directly located at or close to catalytic sites, allosteric sites and binding sites which are crucial for maintaining protein functions and for understanding the underlying molecular mechanism (see Figures 6.8, 6.9, 6.10, 6.11). Additionally, 2 significant sites in EGFR and 8 significant sites in GCK (three of them are also in contact with allosteric sites in GCK) are related to disease associated nsSNP regions of both proteins. As has been noted for the result of CMF-significant sites, most disease-causing mutations at these positions in corresponding sequences destroy structural features of proteins, thus affecting protein stability and often results in loss of protein function.

Although the importance of almost all QCMF-significant sites are verified through essential sites of both human proteins, there are still eight and two unconfirmed significant sites in EGFR and GCK proteins, respectively, which do not fall into essential sites. It is interesting to note that some of these unconfirmed sites are also referred as CMF-significant. We therefore believe that most of these unconfirmed sites identified by our present method may have an importance for the function and structural stability of both proteins notwithstanding the absence of previous experimental data. A further comparison reveals that the overlaps between the present QCMF-significant sites and previous CMF-significant sites are quite low, indicating that our both quantum and classical information theory based methods detect considerably different sets of residue sites as functionally and structurally important. The comparison result clearly shows that considering similar and dissimilar amino acid signals simultaneously, our present quantum information theory based method is more sensible to catalytic, allosteric and binding sites, while only focusing on dissimilar signals our previous classical information theory based method deals successfully with nsSNP positions in proteins. A final comparison between the methods is made based on their connectivity degree *cut-off*. As has been mentioned before, we initially set the connectivity degree cut-off value to the 90-th percentile at which the previous method reaches its maximal MCC value. However, going through all possible n-th percentiles for $n = 80, 81, \ldots, 99$, our present method reaches its maximal MCC value of 0.231 if $n = 88$ (see Table 7.2).

Although the QCMF shows a better performance than CMF, the comparison study indicates that the result of the QCMF complements the result of CMF. The simultaneous usage of both methods can significantly improve the identification of important sites of proteins.

8 Conclusion

In the final chapter of this thesis, we summarize our findings and contributions. Furthermore, we give an outlook on future work in this research direction.

8.1 Summary

In this thesis, we reported two new methods based on classical information theory and quantum information theory for the identification of functionally or structurally important non-conserved residue sites by analyzing compensatory mutations in MSAs. In addition, we developed an MSA-specific statistical model by slightly generalizing the Storey-Tibshirani procedure devised for multiple testing problems. This model separates method-based significant compensatory mutation signals from background noise and quantifies further the error made in terms of the false discovery rate.

Our first method (see Chapter 4) is based on classical information theory and includes two metrics. While the first metric (developed by Merkl and Zwick in [22]) does not consider the biochemical constraints of amino acids, the second metric focuses on significant and *BLOSUM62*-dissimilar amino acid signals as a model of compensatory mutations. In addition, these signals are integrated in calculation of its metric using a doubly stochastic matrix to transform the empirical pair distribution of the column pair. Using this approach, we can show how dissimilar compensatory mutations have affected genomic sequences in the course of evolution.

In contrast to the first method, applying principles of quantum information theory, our second method (see Chapter 5) simultaneously models similar and dissimilar amino acid pair signals in the detection of functionally or structurally important sites. This method also includes two metrics: the first metric measures compensatory mutations between pairs of columns; the second metric considers the sequence conservation of columns.

Applying the MSA-specific statistical model for our classical and quantum information theory based methods, we determined for each method as well as each metric the amount of significant compensatory mutation signals in MSAs of two human proteins. To demonstrate the performance and functionality of our methods, we analyzed the structurally or functionally important positions of these proteins. The results show that overlaps between two information theory based methods are quite low, indicating that both methods detect considerably different sets of residue sites as functionally and structurally important. Only focusing on dissimilar signals, the first method successfully deals with nsSNP positions. In contrast, simultaneously considering similar and dissimilar amino acid signals, our second method is more sensible to catalytic, allosteric and binding sites than those found by the first method.

8.2 Outlook

Although we developed two information theory based methods and one statistical model for the identification of functionally and structurally important sites of proteins, there is still need for a method to replace the 90-th percentile *cut-off* by an MSA-dependent threshold for the connectivity degree. The extension of the connectivity degree technique would provide opportunities for future studies, which could also be applied for several biological network analysis.

Another interesting extension of this thesis would be the construction of the matrices, introduced in Section 4.3, taking into account only either functionally or structurally important sites in MSAs. In this thesis, we constructed these matrices simultaneously using functionally and structurally important sites. However, considering them individually for the matrix construction could provide us to measure the effect of biochemical constraints of amino acids in the identification of important residue sites.

Furthermore, in order to get an improved performance in identifying important sites in proteins, our both classical information theory and quantum information theory based methods could be combined using a machine learning model, such as random forests, support vector machine, or other comparable models. The combination of both methods with a suitable machine learning model would result in a significantly better performance for the identification of important sites.

Bibliography

[1] Nelson DL, MCox M: *Lehninger Prenciples of Biochemistry*. New York, NY 10010: W.H Freeman and Company, 41 Madison Avenue, fifth edition edition 2008.

[2] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic Acids Research* 2000, **28**:235–242, [`http://nar.oxfordjournals.org/content/28/1/235.abstract`].

[3] Gloor GB, Martin LC, Wahl LM, Dunn SD: **Mutual Information in Protein Multiple Sequence Alignments Reveals Two Classes of Coevolving Positions**. *Biochemistry* 2005, **44**(19):7156–7165, [`http://pubs.acs.org/doi/abs/10.1021/bi050293e`]. [PMID: 15882054].

[4] Gültas M, Haubrock M, Tüysüz N, Waack S: **Coupled Mutation Finder: A new entropy-based method quantifying phylogenetic noise for the detection of compensatory mutations**. *BMC Bioinformatics* 2012, **13**:225, [`http://www.biomedcentral.com/1471-2105/13/225`].

[5] Gültas M, Düzgün G, Herzog S, Jäger S, Meckbach C, Wingender E, Waack S: **Quantum coupled mutation finder: predicting functionally or structurally important sites in proteins using quantum Jensen-Shannon divergence and CUDA programming**. *BMC Bioinformatics* 2014, **15**:96, [`http://www.biomedcentral.com/1471-2105/15/96`].

[6] Wilson K, Walker J: *Principles and Techniques of Biochemistry and Molecular Biology*. Cambridge University Press, 7th edition 2010.

[7] Altschuh D, Lesk AM, Bloomer AC, Klug A: **Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus**. *Journal of Molecular Biology* 1987, **193**(4):693–707.

[8] Martin LC, Gloor GB, Dunn SD, Wahl LM: **Using information theory to search for co-evolving residues in proteins**. *Bioinformatics* 2005, **21**(22):4116–4124, [`http://bioinformatics.oxfordjournals.org/content/21/22/4116.abstract`].

[9] Dunn SD, Wahl LM, Gloor GB: **Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction**. *Bioinformatics* 2008, **24**(3):333–340.

[10] Chakrabarti S, Panchenko AR: **Structural and Functional Roles of Coevolved Sites in Proteins**. *PLoS ONE* 2010, **5**:e8591, [`http://dx.doi.org/10.1371%2Fjournal.pone.0008591`].

[11] Sandler I, Abu-Qarn M, Aharoni A: **Protein co-evolution: how do we combine bioinformatics and experimental approaches?** *Mol. BioSyst.* 2013, **9**:175–181, [`http://dx.doi.org/10.1039/C2MB25317H`].

[12] DePristo MA, Weinreich DM, Hartl DL: **Missense meanderings in sequence space: a biophysical view of protein evolution**. *Nat Rev Genet Nature Publishing Group* 2005, **6**(9):678–687, [`http://dx.doi.org/10.1038/nrg1672`].

[13] Yeang CH, Haussler D: **Detecting Coevolution in and among Protein Domains**. *PLoS Comput Biol* 2007, **3**(11):e211, [`http://dx.plos.org/10.1371%2Fjournal.pcbi.0030211`].

[14] Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE: **Co-evolution of proteins with their interaction partners**. *Journal of Molecular Biology* 2000, **299**(2):283 – 293, [`http://www.sciencedirect.com/science/article/pii/S002228360093732X`].

[15] Lockless SW, Ranganathan R: **Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families**. *Science* 1999, **286**(5438):295–299, [`http://www.sciencemag.org/content/286/5438/295.abstract`].

[16] Göbel U, Sander C, Schneider R, Valencia A: **Correlated mutations and residue contacts in proteins**. *PROTEINS-STRUCTURE FUNCTION AND GENETICS* 1994, **18**(4):309–317.

[17] Neher E: **How frequent are correlated changes in families of protein sequences?** *Proceedings of the National Academy of Sciences* 1994, **91**:98–102, [`http://www.pnas.org/content/91/1/98.abstract`].

[18] Pollock DD, Taylor WR: **Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution**. *Protein Engineering* 1997, **10**(6):647–657, [`http://peds.oxfordjournals.org/content/10/6/647.abstract`].

[19] Dekker JP, Fodor A, Aldrich RW, Yellen G: **A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments**. *Bioinformatics* 2004, **20**(10):1565–1572.

[20] Atchley, W R and Wollenberg, K R and Fitch, W M and Terhalle, W and Dress, A W: **Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis**. *Mol. Biol. Evol.* 2000, **17**:164.

[21] Tillier ER, Lui TW: **Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments**. *Bioinformatics* 2003, **19**(6):750–755, [`http://bioinformatics.oxfordjournals.org/content/19/6/750.abstract`].

[22] Merkl R, Zwick M: **H2r: Identification of evolutionary important residues by means of an entropy based analysis of multiple sequence alignments**. *BMC Bioinformatics* 2008, **9**:151, [`http://www.biomedcentral.com/1471-2105/9/151`].

[23] Gao H, Dou Y, Yang J, Wang J: **New methods to measure residues coevolution in proteins**. *BMC Bioinformatics* 2011, **12**:206, [`http://www.biomedcentral.com/1471-2105/12/206`].

[24] Codoner FM, Fares M: **Why Should We Care About Molecular Coevolution?** *Evolutionary Bioinformatics* 2008, **4**:29–38.

[25] Noivirt O, Eisenstein M, Horovitz A: **Detection and reduction of evolutionary noise in correlated mutation analysis**. *Protein Engineering Design and Selection* May 2005, **18**(5):247–253, [`http://peds.oxfordjournals.org/content/18/5/247.abstract`].

[26] Wollenberg KR, Atchley WR: **Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap**. *Proceedings of the National Academy of Sciences* 2000, **97**(7):3288–3291, [`http://www.pnas.org/content/97/7/3288.abstract`].

[27] Storey JD, Tibshirani R: **Statistical significance for genomewide studies**. *Proc. Acad. Sci.* 2003, **100**:9440–9445.

[28] Walsh B: **Multiple comparisons: Bonferroni Corrections and False Discovery Rates**. Lecture Notes EEB 581, Department of Ecology and Evolutionary Biology, University of Arizona 2004.

[29] Dixit A, Yi L, Gowthaman R, Torkamani A, Schork NJ, Verkhivker GM: **Sequence and Structure Signatures of Cancer Mutation Hotspots in Protein Kinases**. *PLoS ONE* 2009, **4**(10):e7485, [`http://dx.doi.org/10.1371%2Fjournal.pone.0007485`].

[30] Yun CH, Boggon TJ, Li Y, Woo MS, Greulich H, Meyerson M, Eck MJ: **Structures of Lung Cancer-Derived EGFR Mutants and Inhibitor Complexes: Mechanism of Activation and Insights into Differential Inhibitor Sensitivity**. *Cancer Cell* 2007, **11**(3):217 – 227, [`http://www.sciencedirect.com/science/article/pii/S1535610807000281`].

[31] Zhang H, Berezov A, Wang Q, Zhang G, Drebin J, Murali R, Greene MI: **ErbB receptors: from oncogenes to targeted cancer therapies**. *The Journal of Clinical Investigation* 2007, **117**(8):2051–2058, [`http://www.jci.org/articles/view/32278`].

[32] Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, Harris PL, Haserlat SM, Supko JG, Haluska FG, Louis DN, Christiani DC, Settleman J, Haber DA: **Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non-Small-Cell Lung Cancer to Gefitinib**. *New England Journal of Medicine* 2004, **350**(21):2129–2139, [`http://www.nejm.org/doi/full/10.1056/NEJMoa040938`].

[33] Balius TE, Rizzo RC: **Quantitative Prediction of Fold Resistance for Inhibitors of EGFR**. *Biochemistry* 2009, **48**(35):8435–8448, [`http://pubs.acs.org/doi/abs/10.1021/bi900729a`]. [PMID: 19627157].

[34] Tinto N, Zagari A, Capuano M, De Simone A, Capobianco V, Daniele G, Giugliano M, Spadaro R, Franzese A, Sacchetti L: **Glucokinase Gene Mutations: Structural and Genotype-Phenotype Analyses in MODY Children from South Italy**. *PLoS ONE* 2008, **3**(4):e1870, [`http://dx.plos.org/10.1371%2Fjournal.pone.0001870`].

[35] Capuano M, Garcia-Herrero CM, Tinto N, Carluccio C, Capobianco V, Coto I, Cola A, Iafusco D, Franzese A, Zagari A, Navas MA, Sacchetti L: **Glucokinase (GCK) Mutations and Their Characterization in MODY2 Children of Southern Italy**. *PLoS ONE* 2012, **7**(6):e38906, [`http://dx.doi.org/10.1371%2Fjournal.pone.0038906`].

[36] Garcia-Herrero CM, Rubio-Cabezas O, Azriel S, Gutierrez-Nogues A, Aragones A, Vincent O, Campos-Barros A, Argente J, Navas MA: **Functional Characterization of MODY2 Mutations Highlights the Importance of the Fine-Tuning of Glucokinase and Its Role in Glucose Sensing**. *PLoS ONE* 2012, **7**:e30518, [`http://dx.doi.org/10.1371%2Fjournal.pone.0030518`].

[37] Kamata K, Mitsuya M, Nishimura T, ichi Eiki J, Nagata Y: **Structural Basis for Allosteric Regulation of the Monomeric Allosteric Enzyme Human Glucokinase**. *Structure* 2004, **12**(3):429 – 438, [`http://www.sciencedirect.com/science/article/pii/S0969212604000474`].

[38] Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders**. *Nucleic Acids Research* 2005, **33**(suppl 1):D514–D517, [`http://nar.oxfordjournals.org/content/33/suppl_1/D514.abstract`].

[39] Reichert J, Sühnel J: **The IMB Jena Image Library of Biological Macromolecules: 2002 update**. *Nucleic Acids Research* 2002, **30**:253–254, [`http://nar.oxfordjournals.org/content/30/1/253.abstract`].

[40] Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Gordon L, Hendrix: **Ensembl 2011**. *Nucleic Acids Research* 2011, **39**(suppl 1):D800–D806, [`http://nar.oxfordjournals.org/content/39/suppl_1/D800.abstract`].

[41] Sunyaev S, Ramensky V, Bork P: **Towards a structural basis of human non-synonymous single nucleotide polymorphisms**. *Trends in Genetics* 2000, **16**(5):198 – 200, [`http://www.sciencedirect.com/science/article/pii/S0168952500019880`].

[42] Wang Z, Moult J: **SNPs, protein structure, and disease**. *Human Mutation* 2001, **17**(4):263–270, [`http://dx.doi.org/10.1002/humu.22`].

[43] Burke D, Worth C, Priego EM, Cheng T, Smink L, Todd J, Blundell T: **Genome bioinformatic analysis of nonsynonymous SNPs** . *BMC Bioinformatics* 2007, **8**:301, [`http://www.biomedcentral.com/1471-2105/8/301`].

[44] Xiong J: *Essential Bioinformatics*. New York: Cambridge Universtiy Press 2006.

[45] Lesk AM: *Introduction To Bioinformatics*. Oxford University Press, second edition edition 2005.

[46] Lodish H: *Molecular Cell Biology*. W. H. Freeman, W.H. Freeman 2012, [`http://books.google.de/books?id=hdn7ngEACAAJ`].

[47] Mount DW: *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press, New York, second edition edition 2001.

[48] Hansen B, Jorde LB: *USMLE Step 1 Lecture Notes Biochemistry*. KAPLAN medical 2004.

[49] Barnes MR, Gray IC: *Bioinformatics for Geneticists*. West Sussex PO19 8SQ, England: John Wiley and Sons Ltd, 2003.

[50] Wang S: **On Multiple Sequence Alignment**. *Dissertation*, The University of Texas at Austin 2007.

[51] D H: **Bioinformatics 1, Multiple Sequence Alignment**. Tech. rep., Algorithms in Bioinformatics, University of Tuebingen 2009.

[52] Yun CH, Mengwasser KE, Toms AV, Woo MS, Greulich H, Wong KK, Meyerson M, Eck MJ: **The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP**. *Proceedings of the National Academy of Sciences* 2008, **105**(6):2070–2075, [`http://www.pnas.org/content/105/6/2070.abstract`].

[53] **Protein Data Bank Contents Guide: Atomic Coordinate Entry Format Description**. *wwPDB* 2012, **Version 3.30**.

[54] Schneider R, Sander C: **The HSSP Database of Protein Structure-Sequence Alignments**. *Nucleic Acids Research* 1996, **24**:201–205, [`http://nar.oxfordjournals.org/content/24/1/201.abstract`].

[55] Bairoch A, Boeckmann B: **The SWISS-PROT protein sequence data bank**. *Nucleic Acids Research* 1992, **20**(suppl):2019–2022, [`http://nar.oxfordjournals.org/content/20/supplement/2019.short`].

[56] Kabsch W, Sander C: **Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features**. *Biopolymers* 1983, **22**(12):2577–2637, [`http://dx.doi.org/10.1002/bip.360221211`].

[57] Brown NP, Leroy C, Sander C: **MView: a web-compatible database search or multiple alignment viewer**. *Bioinformatics* 1998, **14**(4):380–381, [`http://bioinformatics.oxfordjournals.org/content/14/4/380.abstract`].

[58] Porter CT, Bartlett GJ, Thornton JM: **The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data**. *Nucleic Acids Research* 2004, **32**(suppl 1):D129–D133, [`http://nar.oxfordjournals.org/content/32/suppl_1/D129.abstract`].

[59] Laskowski RA: **PDBsum: summaries and analyses of PDB structures**. *Nucleic Acids Research* 2001, **29**:221–222, [`http://nar.oxfordjournals.org/content/29/1/221.abstract`].

[60] Laskowski RA, Chistyakov VV, Thornton JM: **PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids**. *Nucleic Acids Research* 2005, **33**(suppl 1):D266–D268, [`http://nar.oxfordjournals.org/content/33/suppl_1/D266.abstract`].

[61] Holland RCG, Down TA, Pocock M, Prlic A, Huen D, James K, Foisy S, Dreager A, Yates A, Heuer M, Schreiber MJ: **BioJava: an open-source framework for bioinformatics**. *Bioinformatics* 2008, **24**(18):2096–2097, [`http://bioinformatics.oxfordjournals.org/content/24/18/2096.abstract`].

[62] Keskin O, Tsai CJ, Wolfson H, Nussinov R: **A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications**. *Protein Science* 2004, **13**(4):1043–1055, [`http://dx.doi.org/10.1110/ps.03484604`].

[63] Deza MM, Deza E: *Encyclopedia of Distances*. Springer Dordrecht Heidelberg London New York 2009.

[64] Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks**. *Proceedings of the National Academy of Sciences* 1992, **89**(22):10915–10919, [`http://www.pnas.org/content/89/22/10915.abstract`].

[65] Pietrokovski S, Henikoff JG, Henikoff S: **The Blocks Database A System for Protein Classification**. *Nucleic Acids Research* 1996, **24**:197–200, [`http://nar.oxfordjournals.org/content/24/1/197.abstract`].

[66] Cover TM, Thomas JA: *Elements of Information Theory*. John Wiley & Sons, Inc, second edition edition 2006.

[67] Desurvire E: *Classical and Quantum Information Theory An Indroduction for the Telecom Scientist*. Cambridge University Press 2009.

[68] Vedral V: *Introduction to Quantum Information Science*. Oxford University Press 2006.

[69] Nielsen MA, Chuang IL: *Quantum Computation and Quantum Information*. Cambridge University Press 2000.

[70] Wilde MM: *Quantum Information Theory*. Cambridge University Press 2013.

[71] McMahon D: *Quantum Computing Explained* . John Wiley and Sons Ltd, 2007.

[72] Lin J: **Divergence measures based on the Shannon entropy**. *Information Theory, IEEE Transactions on* 1991, **37**:145 –151.

[73] Fuglede B, Topsoe F: **Jensen-Shannon divergence and Hilbert space embedding**. In *Information Theory, 2004. ISIT 2004. Proceedings. International Symposium on* 2004:31.

[74] Ewens WJ, Grant GR: *Statistical Methods in Bioinformatics: An Introduction*. Springer-Verlag New York 2001.

[75] Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995, **57**:pp. 289–300, [`http://www.jstor.org/stable/2346101`].

[76] Ferreira JA, Zwinderman AH: **On the Benjamini-Hochberg Method**. *The Annals of Statistics* 2006, **34**(4):pp. 1827–1849, [`http://www.jstor.org/stable/25463486`].

[77] Ferreira JA, Zwinderman AH: **Approximate Power and Sample Size Calculations with the Benjamini-Hochberg Method**. *The International Journal of Biostatistics* 2006, **2**.

[78] Bremm S, Schreck T, Boba P, Held S, Hamacher K: **Computing and visually analyzing mutual information in molecular co-evolution**. *BMC Bioinformatics* 2010, **11**:330, [`http://www.biomedcentral.com/1471-2105/11/330`].

[79] Dieter U, Ahrens JH: **Acceptance-Rejection Techniques For Sampling From The Gamma And Beta Distribution**. Tech. rep., Stanford University 1974.

[80] Wang G, Jr RLD: **PISCES: recent improvements to a PDB sequence culling server**. *Nucleic Acids Research* 2005, **33**(Web-Server-Issue):94–98.

[81] Eyal E, Frenkel-Morgenstern M, Sobolev V, Pietrokovski S: **A pair-to-pair amino acids substitution matrix and its applications for protein structure prediction**. *Proteins: Structure, Function, and Bioinformatics* 2007, **67**:142–153, [`http://dx.doi.org/10.1002/prot.21223`].

[82] Asper RY: **Classifiers for Discrimination of Significant Protein Residues and Protein-Protein Interaction Using Concepts of Information Theory and Machine Learning**. *Dissertation*, Georg-August-University of Goettingen 2011.

[83] Cappellini V, Sommer HJ, Bruzda W, Zyczkowski K: **Random bistochastic matrices**. *J.Phys. A: Math. Theor.* 2009, **42**:23.

[84] Birkhoff G: **Tres observationes sobre et algebra lineal**. *Univ. Nac. Tucaman Rev.* 1946, **A**(5).

[85] Hardy G, Littlewood J, Pólya G: *Inequalities*. Oxford: Oxford University Press, 2nd edition 1952.

[86] Johansson F, Toh H: **Relative von Neumann entropy for evaluating amino acid conservation**. *Journal of Bioinformatics and Computational Biology* 2010, **08**(05):809–823, [`http://www.worldscientific.com/doi/abs/10.1142/S021972001000494X`].

[87] Grosse I, Bernaola-Galván P, Carpena P, Román-Roldán R, Oliver J, Stanley HE: **Analysis of symbolic sequences using the Jensen-Shannon divergence**. *Phys. Rev. E* 2002, **65**:041905, [`http://link.aps.org/doi/10.1103/PhysRevE.65.041905`].

[88] Capra JA, Singh M: **Predicting functionally important residues from sequence conservation**. *Bioinformatics* 2007, **23**(15):1875–1882.

[89] Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES: **Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?** *Protein Science* 2004, **13**:190–2002.

[90] Herbst RS: **Review of epidermal growth factor receptor biology**. *International Journal of Radiation Oncology*Biology*Physics* 2004, **59**(2, Supplement):S21 – S26, [`http://www.sciencedirect.com/science/article/pii/S0360301604003311`].

[91] Thornton PS, Satin-Smith MS, Herold K, Glaser B, Chiu KC, Nestorowicz A, Permutt M, Baker L, Stanley CA: **Familial hyperinsulinism with apparent autosomal dominant inheritance: Clinical and genetic differences from the autosomal recessive variant**. *The Journal of Pediatrics* 1998, **132**:9 – 14, [`http://www.sciencedirect.com/science/article/pii/S0022347698704779`].

[92] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation**. *Nucleic Acids Research* 2001, **29**:308–311, [`http://nar.oxfordjournals.org/content/29/1/308.abstract`].

[93] Valentinova L, Beer NL, Stanik J, Tribble ND, van de Bunt M, Huckova M, Barrett A, Klimes I, Gasperikova D, Gloyn AL: **Identification and Functional Characterisation of Novel Glucokinase Mutations Causing Maturity-Onset Diabetes of the Young in Slovakia**. *PLoS ONE* 2012, **7**(4):e34541, [`http://dx.doi.org/10.1371%2Fjournal.pone.0034541`].

[94] Cheng TMK, Lu YE, Vendruscolo M, Lio' P, Blundell TL: **Prediction by Graph Theoretic Measures of Structural Effects in Proteins Arising from Non-Synonymous Single Nucleotide Polymorphisms**. *PLoS Comput Biol* 2008, **4**(7):e1000135, [`http://dx.doi.org/10.1371%2Fjournal.pcbi.1000135`].

[95] Bao L, Cui Y: **Functional impacts of non-synonymous single nucleotide polymorphisms: Selective constraint and structural environments**. *FEBS Letters* 2006, **580**(5):1231 – 1234, [`http://www.sciencedirect.com/science/article/pii/S0014579306000755`].

[96] Ng PC, Henikoff S: **SIFT: predicting amino acid changes that affect protein function**. *Nucleic Acids Research* 2003, **31**(13):3812–3814, [`http://nar.oxfordjournals.org/content/31/13/3812.abstract`].

[97] Ramensky V, Bork P, Sunyaev S: **Human non-synonymous SNPs: server and survey**. *Nucleic Acids Research* 2002, **30**(17):3894–3900, [`http://nar.oxfordjournals.org/content/30/17/3894.abstract`].

9 Appendix

9.1 Appendix A: Coupled Mutation Finder

Gültas *et al. BMC Bioinformatics* 2012, **13**:225
http://www.biomedcentral.com/1471-2105/13/225

BMC
Bioinformatics

METHODOLOGY ARTICLE **Open Access**

Coupled mutation finder: A new entropy-based method quantifying phylogenetic noise for the detection of compensatory mutations

Mehmet Gültas[1*], Martin Haubrock[2], Nesrin Tüysüz[3] and Stephan Waack[1*]

Abstract

Background: The detection of significant compensatory mutation signals in multiple sequence alignments (MSAs) is often complicated by noise. A challenging problem in bioinformatics is remains the separation of significant signals between two or more non-conserved residue sites from the phylogenetic noise and unrelated pair signals. Determination of these non-conserved residue sites is as important as the recognition of strictly conserved positions for understanding of the structural basis of protein functions and identification of functionally important residue regions. In this study, we developed a new method, the Coupled Mutation Finder (*CMF*) quantifying the phylogenetic noise for the detection of compensatory mutations.

Results: To demonstrate the effectiveness of this method, we analyzed essential sites of two human proteins: epidermal growth factor receptor (EGFR) and glucokinase (GCK). Our results suggest that the *CMF* is able to separate significant compensatory mutation signals from the phylogenetic noise and unrelated pair signals. The vast majority of compensatory mutation sites found by the *CMF* are related to essential sites of both proteins and they are likely to affect protein stability or functionality.

Conclusions: The *CMF* is a new method, which includes an MSA-specific statistical model based on multiple testing procedures that quantify the error made in terms of the false discovery rate and a novel entropy-based metric to upscale BLOSUM62 dissimilar compensatory mutations. Therefore, it is a helpful tool to predict and investigate compensatory mutation sites of structural or functional importance in proteins. We suggest that the *CMF* could be used as a novel automated function prediction tool that is required for a better understanding of the structural basis of proteins. The *CMF* server is freely accessible at http://cmf.bioinf.med.uni-goettingen.de.

Background

A multiple sequence alignment (MSA) of proteins contains a set of aligned amino acid sequences in which homologous residues of different sequences are placed in same columns. Therefore, functionally or structurally important amino acids and their positions both of which are often strictly conserved are easily detectable with MSAs [1-3]. On the other hand, detection of important non-conserved residue positions related to several essential conserved residues requires a more sophisticated approach. The usage of methods such as correlation analysis allow the identification of important non-conserved residue positions based on their correlated mutation manners [4,5] due to functional coupling of mutation positions. This coupling might stem from one mutation in a certain site affecting a compensating mutation at another site, even if both related residue sites are distantly positioned in the protein structure. Moreover, these coupled mutations can result from spatial, physical, or chemical restrictions or signaling of allostery [4,5]. Thus, determination of these positions is as crucial as the recognition of strictly conserved positions for the understanding of the structural basis of protein functions, and for the identification of functionally important residue regions which might be disease associated, responsible for

---

*Correspondence: gueltas@cs.uni-goettingen.de; waack@cs.uni-goettingen.de
[1] Institute of Computer Science, University of Göttingen, Goldschmidtstr. 7, 37077, Göttingen, Germany
Full list of author information is available at the end of the article

BioMed Central

© 2012 Gültas et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

the maintenance of internal protein volume, or possibly form key sites for interactions within or between proteins [6-9].

Until now, a variety of studies have employed Pearson's correlation coefficient methods [10-12], perturbation based methods [9,13] and mutual information (MI) based methods [6,14-17] because of their simplicity and efficiency for the detection of coupled mutations in MSAs. However, due to background noise, all of these methods interfere with the identification of compensatory mutation signals [14,18,19]. Hence, the significant compensatory mutation signals must be separated from the background noise that might occur as a result of: i) false signals arising from insufficient data; ii) sites with low or high conservation biasing the signal; iii) phylogenetic noise. While the first two types of noise can be easily overcome by appropriately filtering the data [16], phylogenetic noise can only be eliminated to some extent by excluding highly similar sequences from the MSA [19].

Recently, several methods such as bootstrapping, simulation or randomization methods have been utilized in order to minimize the influence of phylogenetic linkage and stochastic noise [15,20,21]. Dunn et al. [19] have introduced the *average product correction* (APC), to adjust MI for background effects. Merkl and Zwick, in their study, [16] have used a normalized MI (see Equation 1) and focused on only 75 residue pairs with the highest normalized MI values as significant for each MSA. Gao et al. [17] have pursued a similar approach, where they have replaced the metric used in [16] with the amino acid background distribution (MIB). While the reduction of background noise in the model of Dunn et al. is not quantified, the approaches of Gao et al. and Merkl and Zwick appear to be over-conservative, yet specific.

Despite the presence of a variety of different methods as mentioned above, to date there is still need for a method to deal with the noise as well as for powerful metrics to improve the sensitivity. In this study, we have developed such a method called Coupled Mutation Finder (CMF). The CMF includes an MSA-specific statistical model based on multiple testing procedures described in [22,23] and a novel entropy-based metric to upscale dissimilar compensatory mutations and also to complement the normalized MI metric used in [16]. Unlike previous normalized MI based studies [16,17], we have separated metric-based significant compensatory mutation signals from background noise with respect to our MSA specific statistical model that quantifies the error made in terms of the false discovery rate.

To demonstrate the performance and functionality of the CMF, we analyzed the structurally or functionally important positions of two human proteins, namely epidermal growth factor receptor (EGFR) and glucokinase (GCK). The proteins have been chosen because their functionally and structurally important positions have been experimentally well studied previously [24-35]. As a result, the *CMF* detects in these two proteins disease associated amino acid mutations (non-synonymous single nucleotide polymorphisms (nsSNPs)), not strictly conserved catalytic or binding sites, and residues that are nearby one of these sites or in the close neighborhood of a strictly conserved positions, which are likely to affect protein stability or functionality [36-38].

Results

Our method to predict functionally or structurally important residue positions is composed of two steps. First, we have devised a new MSA-specific statistical method for the identification of significant MSA column pairs with respect to either of the two metrics $\mathbb{U}$ and $\mathbb{U}_{D(\alpha)}$. Assume that M is the MSA under study, these pairs are annotated as $(\mathbb{U}, M)$-significant and $(\mathbb{U}_{D(\alpha)}, M)$-significant, respectively. Second, we utilized the connectivity degree of a residue site with respect to a metric introduced in [16]. The connectivity degree of a residue site indicates the number of its significant coupled mutation pairs. In this case, a site is called (U,M)-significant when the frequency of occurrence of this site in the set of $(\mathbb{U}, M)$-significant pairs exceeds the 90-th percentile. Having defined the concept of a $(\mathbb{U}_{D(\alpha)}, M)$-significant site analogously, a site is defined as *CMF*-significant with respect to M, when it is either $(\mathbb{U}, M)$-significant or $(\mathbb{U}_{D(\alpha)}, M)$-significant.

In this study, we analyzed human EGFR (pdb entry 2J6M) and GCK (pdb entry 1V4S) proteins with a false discovery rate (*FDR*) of 1%. For the preceding one, we defined a total of 14339 out of 26079 non-conserved column pairs as significant. 11365 of these significant pairs are detected as $(\mathbb{U}, M)$-significant and 3798 pairs are observed as $(\mathbb{U}_{D(\alpha)}, M)$-significant. Only 824 EGFR pairs are significant with respect to both metrics. On the other hand, for GCK, we identified a total of 32654 out of 69645 non-conserved column pairs as significant where 18106 of them are $\mathbb{U}$-significant and 16241 are $\mathbb{U}_{D(1)}$-significant. Moreover, 1693 pairs are defined as significant for both $\mathbb{U}$ and $\mathbb{U}_{D(1)}$-significant.

Applying the connectivity degree technique, we identified 22 and 36 residue positions as $\mathbb{U}$-significant for human EGFR and GCK proteins, respectively. Additionally, 21 positions of EGFR and 36 positions of GCK were detected as $\mathbb{U}_{D(1)}$-significant. Finally, a total of 43 sites of EGFR and 72 of GCK were found as *CMF*-significant, and predicted as of structural or functional importance. However, there have been no residue sites defined as significant with respect to either metric.

Essential sites of human EGFR and GCK proteins

To evaluate the *CMF*-significant residue sites, we have investigated essential sites of human EGFR (pdb entry

2J6M) and GCK (pdb entry 1V4S) proteins. The essential sites of both proteins have been assigned into three main categories: i) the nsSNP positions and their adjacent sites; ii) residue positions which are located at or near active sites, allosteric sites, or binding sites; iii) residue positions which are nearby strictly conserved sites. Here, we have used "nearby" definition of Nussinov et al. [39] and defined two residues as in contact or adjacent when the distance between their major carbon atoms is less than 6 Å. We have defined positions which are nearby nsSNPs as essential, because several of them are also adjacent to active sites, allosteric sites, binding sites, or strictly conserved sites. Thus, we refer to a *CMF*-significant residue site as "functionally or structurally important" if it falls into one of these categories of essential sites.

Position analysis of the Human Epidermal Growth Factor Receptor (EGFR) protein

The epidermal growth factor receptor (EGFR) is a member of the ErbB (Erythroblastic Leukemia Viral Oncogene Homolog) family receptors. Signaling through this receptor is a highly conserved mechanism from nematode to humans involved in numerous processes, including proliferation, cell fate determination, and tissue specification [40]. Furthermore, several studies have implicated that mutations resulting in misregulation of the activity or action of EGFR led to multiple cancers, including those of the brain, lung, mammary gland, and ovary [24-27]. Here, in order to detect essential mutation positions in corresponding sequence of human EGFR protein, we determined altogether 43 *CMF*-significant residue sites (see Additional file 1). 15 of these significant residue sites have been verified as nsSNP sites through the Ensembl database annotation and they are illustrated in Figure 1.

Additionally, the significant sites E746, Q791, and four of the nsSNP positions (I759,Y764,M766 and K846) are also in contact with critical active site regions for gefitinib binding site in the wild type EGFR kinase [25,28] (see Figure 2).

Moreover, we observed that 17 further *CMF*-significant positions are essential sites (see Table 1). In total, we

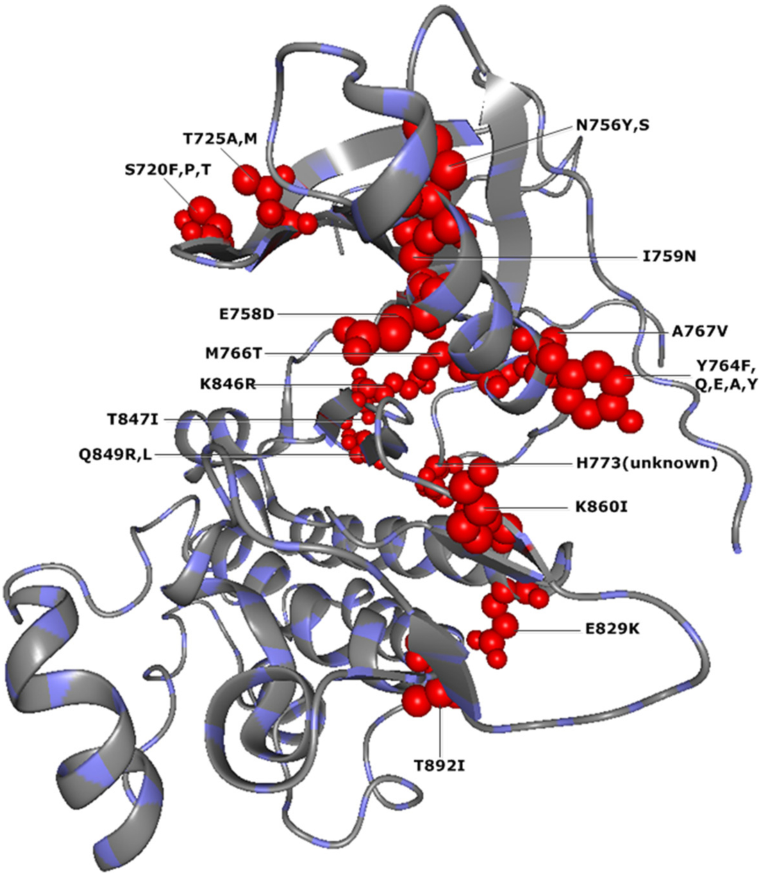


Figure 1 ***CMF*-significant nsSNP positions in human EGFR protein (PDB-Entry 2J6M).** The red spheres correspond to structural localization of 15 different nsSNP positions found by *CMF* as significant in the EGFR protein.

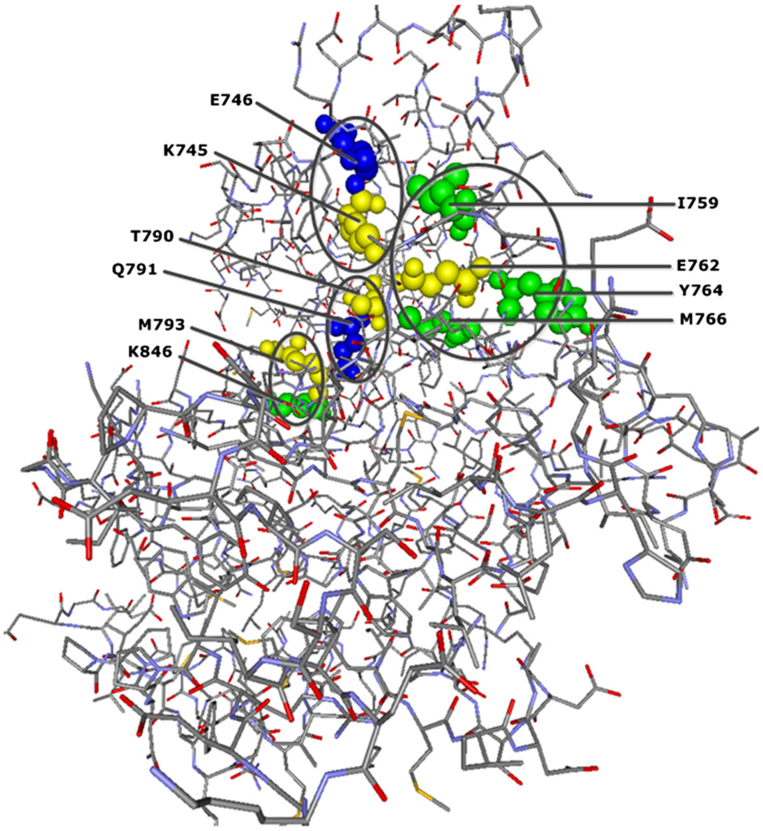


Figure 2 ***CMF*-significant residue positions are in contact with gefitinib binding sites in human EGFR protein (PDB-Entry 2J6M).** Yellow spheres denote positions of the gefitinib binding sites in the wild type kinase. Blue spheres show the localization of significant adjacent residue positions found by *CMF* which are in contact with these binding sites. Moreover, the *CMF*-significant sites I759, Y764, M766 and K846, shown with green spheres, are already described as nsSNP positions and they are also in contact with gefitinib binding sites E762 and M793, respectively. The circles indicate clusters of gefitinib binding sites and their significant adjacent sites.

have established here for EGFR protein the importance of 34 out of 43 *CMF*-significant residue sites via different resources [25,28,35].

Although the vast majority of *CMF*-significant sites are verified to be structurally or functionally important in human EGFR protein, nine *CMF*-significant sites do not overlap with essential sites. The reason for the high connectivity degree of these unconfirmed significant sites and their role in the EGFR protein is unclear.

Position analysis of the Human Glucokinase (GCK) protein

Glucokinase (GCK) is a monomeric enzyme catalyzing phosphorylation of glucose to glucose-6-phosphate, which is the first step in the utilization of glucose, at physiological glucose concentration in pancreas and liver. Given the fact that GCK displays low affinity for glucose, it acts as a glucose sensor playing an important role in the regulation of carbohydrate metabolism. Mutations of the GCK gene can lead to maturity onset diabetes of the young (MODY) characterized by an autosomal dominant mode of inheritance and onset early adulthood [32], or familial hyperinsulinemic hypoglycemia type 3 (HHF), common cause of persistent hypoglycemia in infancy [41].

Applying our method, we found 72 CMF-significant sites to be structurally or functionally important in human GCK protein (see Additional file 2). 16 of these significant residue positions are related to disease associated nsSNP positions [29-31,34,35] (see Figure 3).

Furthermore, nine significant sites are found to be in contact with allosteric sites in the GCK protein structure. Among these sites, the R63 is also allosteric site by itself [32] and T209, C213 and E221 overlap with nsSNP regions (see Figure 4B). Moreover, the five significant sites T149, F171, T206, Q287, and G294 interact with glucose binding sites K169, D204, N205, and E290 [32] (see Figure 4A).

Besides this, there are further 30 *CMF*-significant essential sites which are nearby nsSNPs or strictly conserved residue positions (see Table 2). Altogether, we showed the functionality of 57 positions out of 72 CMF-significant residue sites via different resources [29-35].

Table 1 ***CMF*-significant essential sites in human EGFR protein, which are nearby either nsSNPs or strictly conserved sites**

CMF-significant essential sites	Nearby nsSNPs, or strictly conserved sites	Reference
*Y*727	726^c 743^c	-
*H*755	756^s, 758^s	[35]
*D*800	798^c	-
*G*824	773^s	[35]
*D*830	829^s	[35]
*E*868	892^s	[35]
*E*872	873^s	[34]
*V*876	877^c	-
*K*879	877^c, 880^c	-
*Y*891	892^s, 895^c	[35]
*S*899	880^c, 896^c, 898^c, 901^c	-
*Y*900	898^c, 901^c	-
*T*909	906^c, 936^c	-
*S*912	906^c, 936^c	-
*K*913	914^c	-
*D*916	914^c	-
*M*947	901^c, 950^c	-

^{s}: non-synonymous snp site, ^{c}: strictly conserved site.

While we are able to establish the large number of *CMF*-significant sites as structurally or functionally important in human GCK protein, 15 *CMF*-significant sites do not overlap with essential sites. Their importance in the GCK protein and the reason of high connectivity degree of these unconfirmed significant sites has not been explicitly determined yet.

A comparison between $\mathbb{U}$-metric and $\mathbb{U}_{D(\alpha)}$-metric

Similarities in physical or biochemical properties of amino acids are likely to be crucial for the detection of functionally or structurally important positions of a protein. In contrast to the $\mathbb{U}$-metric, which is a normalized mutual information that uses only the frequencies of occurrences of amino acids in the MSA columns, the novel $\mathbb{U}_{D(\alpha)}$-metric includes dissimilarities according to the BLOSUM62 matrix when calculating normalized mutual information. As a result the positions which have undergone dissimilar compensatory mutations are upscaled.

Having applied the $\mathbb{U}$-metric as well as the $\mathbb{U}_{D(\alpha)}$-metric to human EGFR and GCK proteins, the $\mathbb{U}_{D(\alpha)}$-metric has shown better sensitivity and specificity. However, only when we use the both metrics together, the sensitivity is significantly increased, whereas the specificity is only moderately decreased. The details are presented in Table 3.

It is important to note that the two metrics complement each other. Thus, we propose to use them together.

CMF as a Web service

We have implemented a *CMF* Web service (http://cmf.bioinf.med.uni-goettingen.de) that takes an MSA in multiple FASTA format and a real number from $(0, 1)$ interpreted as false discovery rate as input. It reports the results via email.

Discussion

To predict sites of structural or functional importance, we combine the known $\mathbb{U}$-metric of normalized mutual information [16] with a novel metric $\mathbb{U}_{D^{(i)}(1)}$ to enhance the influence of dissimilar compensatory mutations when measuring covariation of two sites. We discuss how we devised $\mathbb{U}_{D(1)}$ in Methods section.

To learn the frequency of compensatory mutations, we took $\mathbb{U}$-significant site pairs as training data. We did that for reasons of computation time regardless of the fact that these data are biased. To deal with this bias, one could carry through the training in an iterative process, with our training being the first iteration. For $i > 0$, in the $(i + 1)$-th iteration of this modified training, a doubly stochastic matrix $D^{(i+1)}_{\text{CompMut}}$ is calculated based on $\mathbb{U}_{D^{(i)}(1)}$-significant site pairs. This is done until the training data are stable.

According to Birkhoff's Theorem [43], every doubly stochastic matrix is a convex combination of permutation matrices. Moreover, from the Hardy-Littlewood-Pólya majorization theorem [44] follows that transforming the probability mass function by a doubly stochastic matrix increases entropy. Consequently, by linearly transforming the empirical amino acid pair distribution of a site pair by $D(1)$ before calculating the $\mathbb{U}$-value, we penalized those site pairs whose original distribution does not match the frequency pattern of formal dissimilar compensatory mutations in the training data described in the Methods section.

The challenge was to separate the signal caused by structural and functional constraints from the background. To address this issue, we studied only metrics μ that satisfy the following condition. The larger the $\mu(k, l)$-value, the larger the probability that the two sites k and l have co-evolved. Our critical assumptions were: i) the $\mu(k, l)$-values follow three different distributions, one for the signal, one for the noise, and one for pairs of completely unrelated sites; ii) there is an MSA-dependent threshold below which the metric μ does not fall with overwhelming probability, when it is applied to the site pairs of functional or structural importance to which μ is sensitive; iii) there is an MSA-dependent threshold significantly smaller then the one in (ii) such that with overwhelming probability

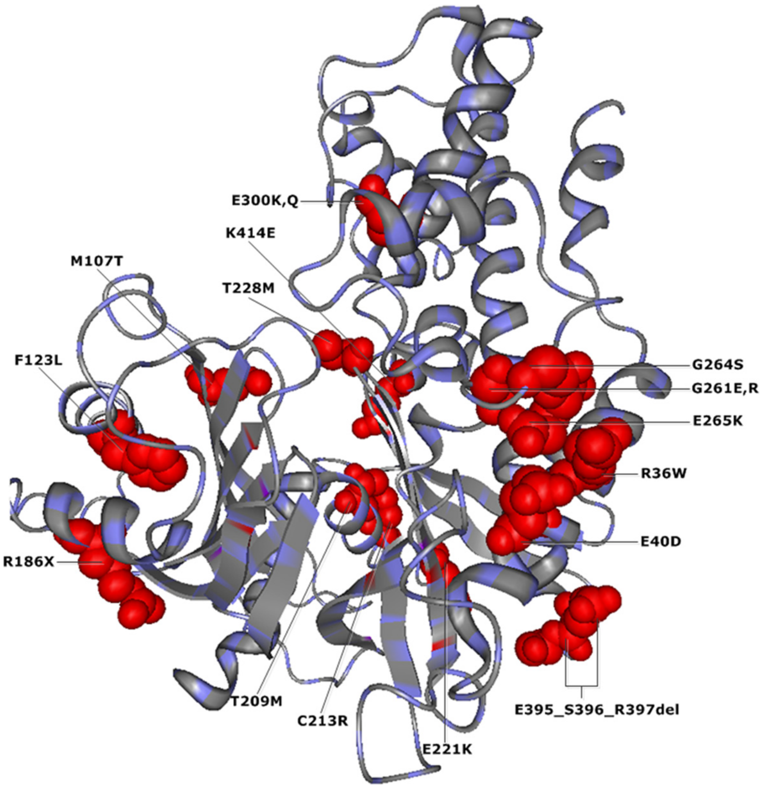


Figure 3 CMF-significant nsSNP positions in human GCK protein (PDB-Entry 1V4S). Red spheres show the structural localization of 16 different nsSNP positions found by *CMF* as significant in the GCK protein.

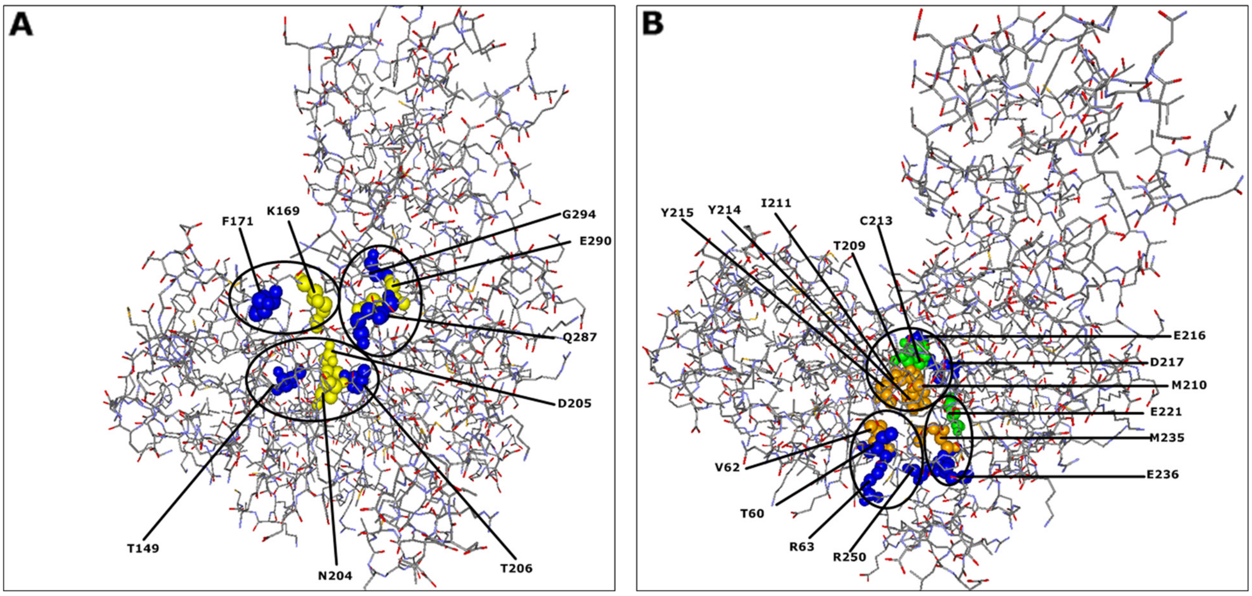


Figure 4 CMF-significant residue positions are in contact with glucose binding site and allosteric site in human GCK protein (PDB-Entry 1V4S). (**A**) Yellow spheres show the structural positions of the glucose binding sites (active sites). Blue spheres correspond to localization of significant adjacent residue positions found by *CMF* which are in contact with these active sites. (**B**) Orange spheres denote the allosteric sites. Blue spheres correspond to localization of just significant adjacent residue positions and green spheres indicate the significant residue positions which are already described as nsSNP position and in contact with these allosteric site. Additionally, the significant position R63 is allosteric site by itself and it is also in contact with an other allosteric site. The circles indicate clusters of glucose binding sites (**A**), allosteric sites (**B**), and their significant adjacent sites.

Table 2 *CMF*-significant essential sites in human GCK protein, which are nearby either nsSNPs or strictly conserved sites

CMF-significant essential sites	Nearby nsSNPs or strictly conserved sites	Reference
*M*34	36^{s}	[34]
*T*65	66^{c}	
*E*67	66^{c}, 68^{c}	-
*T*82	81^{c}	-
*N*83	81^{c}, 108^{s}, 110^{s}	[34]
*H*105	106^{s}	[29]
*C*129	131^{s}, 132^{s}	[29,34]
*F*133	131^{s}, 132^{s}	[29,34]
*F*148	147^{c}, $150^{c,s}$	[34]
*F*152	$150^{c,s}$, 151^{c}	[34]
*H*156	162^{s}	[29]
*N*180	162^{s}, 182^{s}	[29,34]
*F*260	257^{s}, 258^{c}, 259^{s}, 261^{s}	[34]
*D*262	259^{s}, 261^{s}, 264^{s}	[34,42]
*L*266	261^{s}, 264^{s}, 265^{s}	[29,34,42]
*D*267	264^{s}, 265^{s}	[29,42]
*L*271	274^{c}	-
*S*281	278^{c}, 279^{s}	[34]
*Q*286	259^{s}	[34]
*E*331	$299^{c,s}$	[34]
*T*332	295^{c}, $299^{c,s}$	[34]
*R*333	336^{s}	[34]
*Q*337	336^{s}	[34]
*E*339	336^{s}	[34]
*N*391	392^{s}	[29,34]
*S*411	$227^{c,s}$, 410^{c}, 414^{s}	[34]
*S*418	416^{s}	[30]
*F*419	416^{s}	[30]
*E*442	444^{c}	-
*E*443	444^{c}, 445^{c}	-

^{s}: non-synonymous snp site, ^{c}: strictly conserved site.

there are no $\mu(k,l)$-values of pairs (k,l) of unrelated sites exceeding it.

In order to near-completely eliminate the noise, we filtered both our training and input data. We calculated the significant pairs such that the preassigned false discovery rate was guaranteed by generalizing the Storey-Tibshirani procedure devised for multiple testing problems [22].

Our method to eliminate noise is orthogonal to the technique developed in [19]. Therein, for every pair of sites the so-called average product correction (APC) is calculated as an explicit noise measure, by which the mutual information is then decreased. Furthermore, it generalizes the way Merkl and Zwick [16] as well as Gao et al. [17] cope with noise. According to our judgment, taking only the top 75 high-scoring pairs or the top 25 pairs into account as done in [16,17], respectively, is too conservative.

We based our noise separation technique on rather weak distribution assumptions that are standard practice in multiple hypothesis testing, instead of explicitly model the noise in terms of a metric. We applied the connectivity degree technique due to Merkl and Zwick [16] to significant site pairs with respect to our metrics. The cut-off for the connectivity degree was set to the 90-th percentile. That way we defined significant sites. Finally, a site was defined to be *CMF*-significant, if it was μ-significant, where μ is either $\mathbb{U}$ or $\mathbb{U}_{D(1)}$.

Why did we set the cut-off value for the connectivity degree to the 90-th percentile? Going through all possible n-th percentiles for $n = 80, 81, \ldots, 99$, the Matthews correlation coefficient (MCC) of a joint prediction for human EGFR and GCK proteins is maximal if $n = 90$.

It is plausible that the number of functionally or structurally important sites does not only depend on the length of the protein. Therefore, the 90-th percentile cut-off should be replaced by an MSA-dependent threshold in future studies.

Our results for human EGFR and GCK proteins suggest that the large majority of significant compensatory mutation sites found by *CMF* are in agreement with previous experimental studies regarding the functions and stability of these proteins. 15 and 16 *CMF*-significant sites in human EGRF and GCK proteins, respectively, are verified as disease associated nsSNP positions (see Figures 1 and 2) where most amino acid substitutions in protein sequences damage structural stability of proteins [36,37,45]. Moreover, we have observed that in both proteins some of *CMF*-significant nsSNP positions are nearby allosteric sites, binding sites or catalytic sites each of which are considered to be functionally important [46,47] (see Figures 2 and 4). Disease associated mutations at these nearby positions are likely to affect protein function [38,48].

Despite the large number of *CMF*-significant sites demonstrated to be structurally or functionally important for both of the proteins, 9 and 15 significant sites in human EGFR and GCK proteins, respectively, are not included in essential sites. However, we hypothesize that most of the novel significant sites may play a critical role in both proteins notwithstanding the absence of previous

Table 3 Comparison between $\mathbb{U}$-metric and $\mathbb{U}_{D(\alpha)}$-metric

	Sensitivity	Specificity
$\mathbb{U}$-significance	9.7%	91.5%
$\mathbb{U}_{D(\alpha)}$-significance	12.4%	97.2%
CMF-significance	22.1%	88.7%

experimental data. Therefore, further progress from the molecular and structural biology end is required not only to assess the importance of these sites, but also for a future perspective on a deeper understanding of protein structure.

Because we have also used the $\mathbb{U}$-metric, we compared our tool with H2r presented in [16] rather than with those methods developed in [17]. This way, we studied the impact of applying the Storey-Tibshirani procedure in combination with the effect of using the 90-th percentile cut-off for the connectivity degree. We have applied H2r without adding pseudo counts to the human EGFR and GCK protein. For EGFR, the 14 sites T725, A755, N756, A767, Q791, V802, N816, V819, K846, V876, M881, K913, D916, and E931 are identified as significant. Out of these significant sites, ten of these residue sites T725, A755, N756, A767, Q791,K846, V876, M881, K913, and D916 are essential sites. On the other hand, for GCK, H2r identified the 15 residue positions L25, R36, R63, M107, C213, V226, G261, D262, G264, L266, D267, E268, T405, K414, and H416 as significant. Twelve of these sites, namely R36, R63, M107, C213, V226, G261,D262, G264, L266, D267, K414, and H416, are essential sites. However, when using the H2r Web service (http://www-bioinf.uni-regensburg.de/) to analyze EGFR and GCK proteins, sensitivity is decreased, while precision is increased. By this service only eight sites for EGFR and nine sites for GCK were found to be significant. Moreover, only five and eight of them are verified as functionally or structurally important for EGFR and GCK proteins, respectively. This difference stems from the fact that the H2r Web service tightens the filtering of the columns. In addition to this, statistically evaluating H2r for EGFR and GCK proteins, we observed a sensitivity of 5.4%, specificity of 96.7%, precision of 75.9%, and a Matthews correlation coefficient value of 0.047. On the other hand, the CMF reaches precision of 79.1%, and a Matthews correlation coefficiant value of 0.133. For sensitivity and specificity of the *CMF* refer to the last row of Table 3.

The results of this comparison show that a vast majority of functionally or structurally important residue positions cannot be detected without using our novel MSA specific model and both metrics ($\mathbb{U}$ and $\mathbb{U}_{D(1)}$) together.

Conclusions

The *CMF* is a new method which includes an MSA-specific statistical model based on multiple testing procedures that quantifies the error made in terms of the false discovery rate and a novel entropy-based metric to upscale BLOSUM62 dissimilar compensatory mutations. Hence, it shows how dissimilar compensatory mutations have affected genomic sequences in the course of evolution. The method is able to predict significant compensatory mutation positions in protein sequences. We suggest that CMF could be used as a novel automated function prediction tool that is required for a better understanding of the structural basis of proteins.

Methods

In this section we describe the training data used and the methods applied and partly developed. Our descriptions follows the structure of Figure 5, i.e. we start with the data

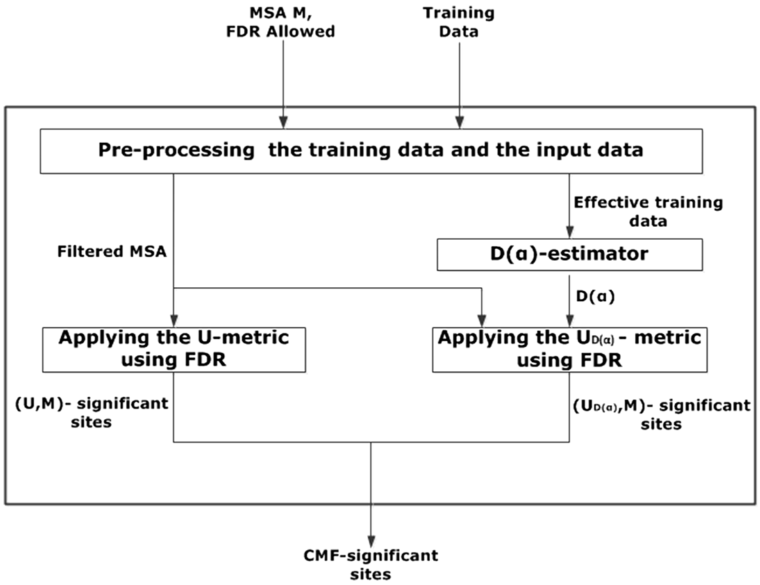

MSA M,
FDR Allowed
Training
Data
Pre-processing the training data and the input data
Effective training data
Filtered MSA
D(α)-estimator
D(α)
Applying the U-metric using FDR
Applying the U_D(α) - metric using FDR
(U,M)- significant sites
(U_D(α),M)- significant sites
CMF-significant sites

Figure 5 Flowchart of the CMF-analysis.

and the preprocessing and systemically work towards the *CMF*-significant site prediction.

Training data set and pre-processing

We used a redundancy free set of more than 35000 protein structures computed in Rainer Merkl's Lab at the University of Regensburg in the following way. The protein structures were taken from the protein data base (http://www.pdb.org/). The PISCES services [49] was applied to assess proteins on sequence similarity and equality of 3D-data. The related MSAs were gathered from the HSSP data base (http://swift.cmbi.ru.nl/gv/hssp/).

Taking pattern from [16], we filtered every MSA obtained as follows. First, highly similar and dissimilar sequences were deleted to ensure that the sequence identity between any two sequences is at least 20% and no more than 90%. Second, we removed strictly conserved residue columns, where the percentage of identical residues is greater than 95%. Third, we eliminated the residue columns which contain more than 25% gaps. Finally, we discarded all MSAs with less than 125 sequences. More than 17000 MSAs survived the last filtering step. We used approximately 1700 MSAs as training data which we randomly chose from this set. The pdb entries of the corresponding protein structures are listed in Additional file 3.

Detecting compensatory mutations by the $\mathbb{U}$-metric

In [16] a normalized measure of mutual information ranging over the interval $[0,1]$ is successfully used to detect important residues. It is defined as

$$\mathbb{U}\left(i,j\right) := 2 \cdot \frac{\mathbb{H}\left(i\right) + \mathbb{H}\left(j\right) - \mathbb{H}\left(i,j\right)}{\mathbb{H}\left(i\right) + \mathbb{H}\left(j\right)}, \tag{1}$$

where $\mathbb{H}\left(i\right)$ and $\mathbb{H}\left(j\right)$ are the entropy of the empirical amino acid distributions of the columns i and j, and $\mathbb{H}\left(i,j\right)$ is their joint entropy.

We determine an MSA-dependent threshold τ above which $\mathbb{U}$-values are defined as significant. Let M be the MSA for the protein under investigation. We extend a standard approach of multiple testing theory [22,50,51] with the following assumptions in mind. M's $\mathbb{U}(k,l)$-values follow three different distributions. The null distribution F_0 represents background signals. The distributions G_1 and G_2 model the unrelated pairs and the signal pairs, respectively.

We assume F_0 to be a β-distribution, and M's $\mathbb{U}(k,l)$-values $U_1, U_2, \ldots, U_\mu$ to be an independent and identically distributed (iid) sample.

Let $X_\iota := 1 - F_0(U_\iota)$ be the p-value of U_ι with respect to F_0. If U_ι is F_0-distributed, then X_ι is uniform over $[0,1]$. However, if U_ι is G_1-distributed or G_2-distributed, then X_ι is skewed to 1 or to 0 (see Figure 6). According to [22,23], the fraction γ of the U_ι's that are F_0-distributed is estimated by

$$\hat{\gamma} := \frac{\text{number of } p\text{-values in } [\lambda_1, \lambda_2]}{\mu(\lambda_2 - \lambda_1)}.$$

The tuning parameters λ_1 and λ_2 are chosen such that the fraction of not uniformly distributed p-values that fall into $[\lambda_1, \lambda_2]$ is negligible.

We call a pair of sites (i,j) of the protein under study *$(\mathbb{U}, M)$-significant* if and only if the p-value $1 - F_0(\mathbb{U}\left(i,j\right))$ is less than or equal to τ, for a threshold $\tau \leq \lambda_1$ that ensures the input false discovery rate *FDR*, which in turn can be estimated by

$$\widehat{FDR}(\tau) = \frac{\hat{\gamma}\mu\tau}{\text{number of } p\text{-values} \leq \tau}.$$

In order to determine the parameters of the β-distribution F_0, it is sufficient to estimate the expected value and the variance. The expected value is estimated by the sample mean of all $\mathbb{U}$-values of M. As for the variance, we take pattern from [52]. Having drawn an iid sample $(C_1, C_1'), (C_2, C_2'), \ldots, (C_\nu, C_\nu')$ of random column pairs of a sufficient size whose $\mathbb{U}$-values fall in a preassigned subinterval of $[0,1]$, we calculate $D_1, D_2, \ldots, D_\nu$ by randomly shuffling C_ι' for every $\iota = 1, 2, \ldots, \nu$. The variance is then estimated as the sample variance of $(C_1, D_1), (C_2, D_2), \ldots, (C_\nu, D_\nu)$.

The *connectivity degree* of a site i with respect to the metric $\mathbb{U}$ and the MSA M is defined as number of sites j such that (i,j) is $(\mathbb{U}, M)$-significant [16]. Site i is defined to be *$(\mathbb{U}, M)$-significant*, if i's connectivity degree with respect to $\mathbb{U}$ and M is greater than or equal to the 90-th percentile. The $(\mathbb{U}, M)$-significant sites of a protein do not coincide with those predicted by H2r [16]. The connectivity degrees attained and the threshold used substantially differ. In particular, the latter one is data-dependent rather than constant.

Enhancing prediction by the $\mathbb{U}_{D(\alpha)}$-metric that models dissimilar compensatory mutations

A pair $\left((a_i, a_j), (a_k, a_l)\right)$ of amino acid pairs is defined to be a *formal dissimilar compensatory mutation*, if the BLOSUM62 score both of (a_i, a_k) and (a_j, a_l) is negative.

We use the training data set of approximately 1700 MSAs described above to estimate a 400×400 doubly stochastic matrix D_{CompMut}. This matrix is our mathematical model of how dissimilar compensatory mutations have affected genomic sequences in the course of evolution. Its training consists of five phases.

Phase 1. We calculate a signal and a null set of column pairs. The signal set consists of all $(\mathbb{U}, M)$-significant column pairs, where M ranges over all training MSA. The null set consists of sufficiently many column pairs

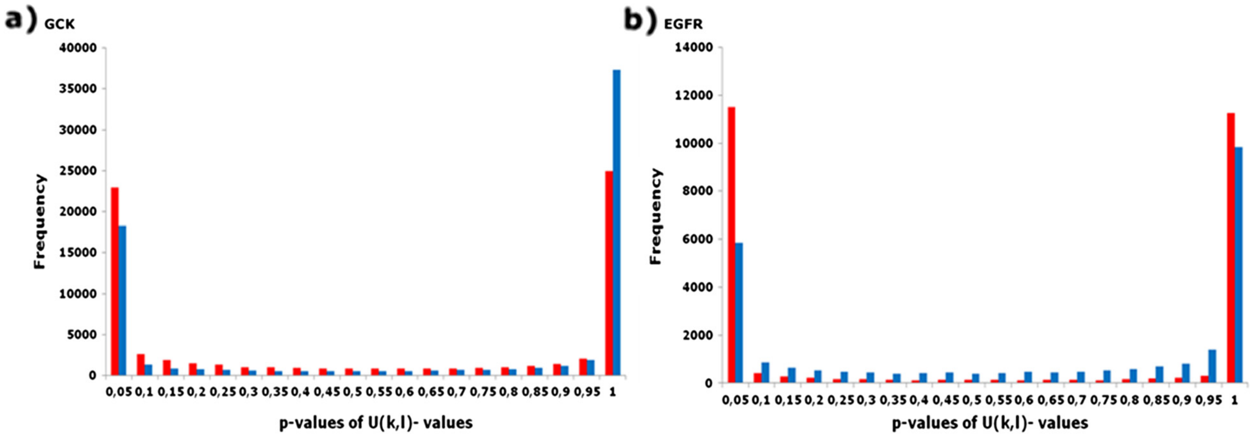


Figure 6 Four p-value distributions of (transformed) normalized mutual information values for human GCK and EFGR proteins having PDB-ID 1V4S and 2J6M, respectively. The bar charts illustrate the two steps of our model: i) blue bars show the *p*-value distribution of the $\mathbb{U}\,(i,j)$-scores; ii) red bars display the *p*-value distribution of the $\mathbb{U}_{D(1)}\,(i,j)$-values. The p-values close to zero represent the significant pairs by means of which we assess the individual residue position. As one can see, within [0.25, 0.70] these four distributions are approximately uniform.

randomly chosen from every training MSA. For both the signal set and the null set we compute a symmetric 400×400 integer-valued matrix of frequencies of pair substitutions C_{alt} and C_{null}. To this end, the method used to compute BLOSUM62 matrices [53] is applied to count residue pair substitutions in MSA column pairs rather than residue substitution in columns.

Phase 2. Using C_{alt} and C_{null}, we define the matrix C_{sig} by

$$C_{\mathrm{sig}}\left((a_i,a_j),(a_k,a_l)\right) := \begin{cases} C_{\mathrm{alt}}\left((a_i,a_j),(a_k,a_l)\right) & \text{if } \varphi\left((a_i,a_j),(a_k,a_l)\right)=1; \\ 0 & \text{otherwise;} \end{cases}$$

where $\varphi\left((a_i,a_j),(a_k,a_l)\right) = 1$ if and only if $(a_i,a_j) = (a_k,a_l)$ or

$$\frac{C_{\mathrm{alt}}\left((a_i,a_j),(a_k,a_l)\right)}{\sum_{i',j',k',l'} C_{\mathrm{alt}}\left((a_{i'},a_{j'}),(a_{k'},a_{l'})\right)} > \frac{C_{\mathrm{null}}\left((a_i,a_j),(a_k,a_l)\right)}{\sum_{i',j',k',l'} C_{\mathrm{null}}\left((a_{i'},a_{j'}),(a_{k'},a_{l'})\right)}.$$

Phase 3. We set all entries of the matrix C_{sig} outside the main diagonal that do not represent a formal dissimilar compensatory mutation to zero. This results in the matrix C_{CompMut}. By normalizing C_{CompMut}, we obtain a symmetric matrix P_{CompMut}. For a_i, a_j, a_k, a_l ranging over all amino acids, $P_{\mathrm{CompMut}}\left((a_i,a_j),(a_k,a_l)\right)$ represents an empirical probability distribution on pairs of amino acid pairs.

Phase 4. We calculate the symmetric 400×400-matrix

$$S_{\mathrm{CompMut}} := \left(\log \frac{P_{\mathrm{CompMut}}\left((a_i,a_j),(a_k,a_l)\right)}{P^{\mathrm{b}}_{\mathrm{CompMut}}(a_i,a_j)\,P^{\mathrm{b}}_{\mathrm{CompMut}}(a_k,a_l)}\right)_{(a_i,a_j),(a_k,a_l)},$$

where $P^{\mathrm{b}}_{\mathrm{CompMut}}\left(a_i,a_j\right)$ is the marginal distribution of P_{CompMut}.

Phase 5. We set all negative entries of S_{CompMut} to zero. Then we compute the doubly stochastic matrix D_{CompMut} by means of the canonical iterated row-column normalization procedure [54].

Now we define our new $\mathbb{U}_{D(\alpha)}$-metric based on D_{CompMut}. For every column pair (i,j) of the input MSA M, we linearly transform the associated empirical pair distribution with the doubly stochastic matrix

$$D(\alpha) := (1-\alpha)\mathbf{1} + \alpha D_{\mathrm{CompMut}}$$

where $\mathbf{1}$ is the 400×400 unit matrix, D_{CompMut} is the result of training phase 5, and $\alpha \in (0,1]$ is a preassigned real number. $\mathbb{U}_{D(\alpha)}\left(i,j\right)$ is then defined to be the $\mathbb{U}$-value (see Equation 1) of this transform.

Having canonically carried over the definition of a significant site pair and of the connectivity degree of a site to this case, a site i is called $(\mathbb{U}_{D(\alpha)}, M)$*-significant*, if i's connectivity degree with respect to the metric $\mathbb{U}_{D(\alpha)}$ is greater than or equal to the 90-th percentile.

Finally, a site is defined to be *CMF-significant* with respect to the MSA M, if it is $(\mathbb{U}, M)$-significant or $(\mathbb{U}_{D(\alpha)}, M)$-significant. The *CMF*-significant sites are predicted as functionally or structurally important ones.

Principally, the controlling parameter $\alpha \in (0, 1]$ can be adjusted by the user. We set α to 1 to allow the two sets of $(\mathbb{U}, M)$-significant and $(\mathbb{U}_{D(\alpha)}, M)$-significant positions to complement each other.

Note, that the matrix S_{CompMut} could be replaced with another scoring matrix meaningful in this context.

Additional files

Additional file 1: EGFR significant sites. *CMF*-significant residue sites of the human epidermal growth factor receptor (EGFR) protein.

Additional file 2: GCK significant sites. *CMF*-significant residue sites of the human glucokinase (GCK) protein.

Additional file 3: Pdb entries of training MSAs. Pdb entries of redundancy free data set.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SW developed the model underlying CMF. MG developed the model together with SW, designed and implemented the tool, and interpreted the results together with NT and MH. All authors read and approved the manuscript.

Acknowledgements

We thank Rainer Merkl for placing a redundancy free set of more than 35000 protein structures at our disposal. Special thank goes to Torsten Schöps for setting up the CMF Web server. Mehmet Gltas acknowledges financial support by the Deutsche Forschungsgemeinschaft (DFG: WA 766/7-1).

Author details

[1]Institute of Computer Science, University of Göttingen, Goldschmidtstr. 7, 37077, Göttingen, Germany. [2]Department of Bioinformatics, University of Göttingen, Goldschmidtstr. 1, 37077 Göttingen, Germany. [3]Erasmus MC Stem Cell Institute, Department of Cell Biology, Erasmus Medical Center, Rotterdam, The Netherlands.

Received: 18 May 2012 Accepted: 23 August 2012
Published: 11 September 2012

References

1. Jeon J, Yang JS, Kim S: **Integration of Evolutionary Features for the Identification of Functionally Important Residues in Major Facilitator Superfamily Transporters.** *PLoS Comput Biol* 2009, **5**(10):e1000522. [http://dx.doi.org/10.13712Fjournal.pcbi.1000522]
2. Sadovsky E, Yifrach O: **Principles underlying energetic coupling along an allosteric communication trajectory of a voltage-activated K+ channel.** *Proc Nat Acad Sci* 2007, **104**(50):19813–19818. [http://www.pnas.org/content/104/50/19813.abstract]
3. Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N: **Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues.** *Bioinformatics* 2002, **18**(suppl 1):S71—S77. [http://bioinformatics.oxfordjournals.org/content/18/suppl_1/S71.abstract]
4. Wilson K, Walker J: *Principles and Techniques of Biochemistry and Molecular Biology*. 7th edition. New York: Cambridge University Press; 2010.
5. Altschuh D, Lesk AM, Bloomer AC, Klug A: **Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus.** *J Mol Biol* 1987, **193**(4):693–707.
6. Martin LC, Gloor GB, Dunn SD, Wahl LM: **Using information theory to search for co-evolving residues in proteins.** *Bioinformatics* 2005, **21**(22):4116–4124.
7. Yeang CH, Haussler D: **Detecting Coevolution in and among Protein Domains.** *PLoS Comput Biol* 2007, **3**(11):e211. [http://dx.plos.org/10.13712Fjournal.pcbi.0030211]
8. Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE: **Co-evolution of proteins with their interaction partners.** *J Mol Biol* 2000, **299**(2):283–293. [http://www.sciencedirect.com/science/article/pii/S002228360093732X]
9. Lockless SW, Ranganathan R: **Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families.** *Science* 1999, **286**(5438):295–299. [http://www.sciencemag.org/content/286/5438/295.abstract]
10. Göbel U, Sander C, Schneider R, Valencia A: **Correlated mutations and residue contacts in proteins.** *Proteins-Struct Funct Genet* 1994, **18**(4):309–317.
11. Neher E: **How frequent are correlated changes in families of protein sequences?** *Proc Nat AcadSci* 1994, **91:**98–102. [http://www.pnas.org/content/91/1/98.abstract]
12. Pollock DD, Taylor WR: **Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution.** *Protein Eng* 1997, **10**(6):647–657. [http://peds.oxfordjournals.org/content/10/6/647.abstract]
13. Dekker JP, Fodor A, Aldrich RW, Yellen G: **A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments.** *Bioinformatics* 2004, **20**(10):1565–1572.
14. Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW: **Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis.** *Mol Biol Evol* 2000, **17:**164.
15. Tillier ER, Lui TW: **Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments.** *Bioinformatics* 2003, **19**(6):750–755. [http://bioinformatics.oxfordjournals.org/content/19/6/750.abstract]
16. Merkl R, Zwick M: **H2r: Identification of evolutionary important residues by means of an entropy based analysis of multiple sequence alignments.** *BMC Bioinformatics* 2008, **9:**151. [http://www.biomedcentral.com/1471-2105/9/151]
17. Gao H, Dou Y, Yang J, Wang J: **New methods to measure residues coevolution in proteins.** *BMC Bioinformatics* 2011, **12:**206. [http://www.biomedcentral.com/1471-2105/12/206]
18. Codoner FM, Fares M: **Why Should We Care About Molecular Coevolution?** *Evolutionary c* 2008, **4:**29–38.
19. Dunn SD, Wahl LM, Gloor GB: **Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction.** *Bioinformatics* 2008, **24**(3):333–340.
20. Noivirt O, Eisenstein M, Horovitz A: **Detection and reduction of evolutionary noise in correlated mutation analysis.** *Protein Eng Design and Sel* 2005, **18**(5):247–253. [http://peds.oxfordjournals.org/content/18/5/247.abstract]
21. Wollenberg KR, Atchley WR: **Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap.** *Proc Nat Acad Sci* 2000, **97**(7):3288–3291. [http://www.pnas.org/content/97/7/3288.abstract]
22. Storey JD, Tibshirani R: **Statistical significance for genomewide studies.** *Proc Acad Sci* 2003, **100:**9440–9445.
23. Walsh B: **Multiple comparisons: Bonferroni Corrections and False Discovery Rates.** Lecture Notes EEB 581, Department of Ecology and Evolutionary Biology, University of Arizona 2004.
24. Dixit A, Yi L, Gowthaman R, Torkamani A, Schork NJ, Verkhivker GM: **Sequence and Structure Signatures of Cancer Mutation Hotspots in Protein Kinases.** *PLoS ONE* 2009, **4**(10):e7485. [http://dx.doi.org/10.13712Fjournal.pone.0007485]
25. Yun CH, Boggon TJ, Li Y, Woo MS, Greulich H, Meyerson M, Eck MJ: **Structures of Lung Cancer-Derived EGFR Mutants and Inhibitor Complexes: Mechanism of Activation and Insights into Differential Inhibitor Sensitivity.** *Cancer Cell* 2007, **11**(3):217–227. [http://www.sciencedirect.com/science/article/pii/S1535610807000281]
26. Zhang H, Berezov A, Wang Q, Zhang G, Drebin J, Murali R, Greene MI: **ErbB receptors: from oncogenes to targeted cancer therapies.** *J Clin Invest* 2007, **117**(8):2051–2058. [http://www.jci.org/articles/view/32278]
27. Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, Harris PL, Haserlat SM, Supko JG, Haluska FG, Louis DN, Christiani DC, Settleman J, Haber DA: **Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non-Small-Cell Lung Cancer to Gefitinib.** *New England J Med* 2004, **350**(21):2129–2139. [http://www.nejm.org/doi/full/10.1056/NEJMoa040938]

28. Balius TE, Rizzo RC: **Quantitative Prediction of Fold Resistance for Inhibitors of EGFR.** *Biochemistry* 2009, **48**(35):8435–8448. [PMID: 19627157]. [http://pubs.acs.org/doi/abs/10.1021/bi900729a]
29. Tinto N, Zagari A, Capuano M, De Simone A, Capobianco V, Daniele G, Giugliano M, Spadaro R, Franzese A, Sacchetti L: **Glucokinase Gene Mutations: Structural and Genotype-Phenotype Analyses in MODY Children from South Italy.** *PLoS ONE* 2008, **3**(4):e1870. [http://dx.plos.org/10.13712Fjournal.pone.0001870]
30. Capuano M, Garcia-Herrero CM, Tinto N, Carluccio C, Capobianco V, Coto I, Cola A, Iafusco D, Franzese A, Zagari A, Navas MA, Sacchetti L: **Glucokinase (GCK) Mutations and Their Characterization in MODY2 Children of Southern Italy.** *PLoS ONE* 2012, **7**(6):e38906. [http://dx.doi.org/10.13712Fjournal.pone.0038906]
31. Garcia-Herrero CM, Rubio-Cabezas O, Azriel S, Gutierrez-Nogues A, Aragones A, Vincent O, Campos-Barros A, Argente J, Navas MA: **Functional Characterization of MODY2 Mutations Highlights the Importance of the Fine-Tuning of Glucokinase and Its Role in Glucose Sensing.** *PLoS ONE* 2012, **7:**e30518. [http://dx.doi.org/10.13712Fjournal.pone.0030518]
32. Kamata K, Mitsuya M, Nishimura T, ichi Eiki J, Nagata Y: **Structural Basis for Allosteric Regulation of the Monomeric Allosteric Enzyme Human Glucokinase.** *Structure* 2004, **12**(3):429–438. [http://www.sciencedirect.com/science/article/pii/S0969212604000474]
33. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic Acids Res* 2005, **33**(suppl 1):D514—D517. [http://nar.oxfordjournals.org/content/33/suppl_1/D514.abstract]
34. Reichert J, Sühnel J: **The IMB Jena Image Library of Biological Macromolecules: 2002 update.** *Nucleic Acids Res* 2002, **30:**253–254. [http://nar.oxfordjournals.org/content/30/1/253.abstract]
35. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Gordon L, Hendrix M, Hourlier T, Johnson N, Kähäri A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Larsson P, Longden I, McLaren W, Overduin B, Pritchard B, Riat HS, Rios D, Ritchie GRS, Ruffier M, *et al*: **Ensembl 2011.** *Nucleic Acids Res* 2011, **39**(suppl 1):D800—D806. [http://nar.oxfordjournals.org/content/39/suppl_1/D800.abstract]
36. Sunyaev S, Ramensky V, Bork P: **Towards a structural basis of human non-synonymous single nucleotide polymorphisms.** *Trends in Genet* 2000, **16**(5):198–200. [http://www.sciencedirect.com/science/article/pii/S0168952500019880]
37. Wang Z, Moult J: **SNPs, protein structure, and disease.** *Human Mutation* 2001, **17**(4):263–270. [http://dx.doi.org/10.1002/humu.22]
38. Burke D, Worth C, Priego EM, Cheng T, Smink L, Todd J, Blundell T: **Genome bioinformatic analysis of nonsynonymous SNPs.** *BMC Bioinformatics* 2007, **8:**301. [http://www.biomedcentral.com/1471-2105/8/301]
39. Keskin O, Tsai CJ, Wolfson H, Nussinov R: **A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications.** *Protein Science* 2004, **13**(4):1043–1055. [http://dx.doi.org/10.1110/ps.03484604]
40. Herbst RS: **Review of epidermal growth factor receptor biology.** *Int J Radiat Oncol Biol Phys* 2004, **59**(Supplement 2):S21—S26. [http://www.sciencedirect.com/science/article/pii/S0360301604003311]
41. Thornton PS, Satin-Smith MS, Herold K, Glaser B, Chiu KC, Nestorowicz A, Permutt M, Baker L, Stanley CA: **Familial hyperinsulinism with apparent autosomal dominant inheritance: Clinical and genetic differences from the autosomal recessive variant.** *J Pediatrics* 1998, **132:**9–14. [http://www.sciencedirect.com/science/article/pii/S0022347698704779]
42. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res* 2001, **29:**308–311. [http://nar.oxfordjournals.org/content/29/1/308.abstract]
43. Birkhoff G: **Tres observationes sobre et algebra lineal.** *Univ Nac Tucaman Rev* 1946, **A**(5):147–151.
44. Hardy G, Littlewood J, Pólya G: *Inequalities*. 2nd edition. Oxford: Oxford University Press; 1952.
45. Cheng TMK, Lu YE, Vendruscolo M, Lio' P, Blundell TL: **Prediction by Graph Theoretic Measures of Structural Effects in Proteins Arising from Non-Synonymous Single Nucleotide Polymorphisms.** *PLoS Comput Biol* 2008, **4**(7):e1000135. [http://dx.doi.org/10.13712Fjournal.pcbi.1000135]
46. Bao L, Cui Y: **Functional impacts of non-synonymous single nucleotide polymorphisms: Selective constraint and structural environments.** *FEBS Letters* 2006, **580**(5):1231–1234. [http://www.sciencedirect.com/science/article/pii/S0014579306000755]
47. Ng PC, Henikoff S: **SIFT: predicting amino acid changes that affect protein function.** *Nucleic Acids Res* 2003, **31**(13):3812–3814. [http://nar.oxfordjournals.org/content/31/13/3812.abstract]
48. Ramensky V, Bork P, Sunyaev S: **Human non-synonymous SNPs: server and survey.** *Nucleic Acids Res* 2002, **30**(17):3894–3900. [http://nar.oxfordjournals.org/content/30/17/3894.abstract]
49. Wang G, Dunbrack Jr RLD: **PISCES: recent improvements to a PDB sequence culling server.** *Nucleic Acids Res* 2005, **33**(Web-Server-Issue): 94–98.
50. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.** *J R Stat Soc Ser B (Methodological)* 1995, **57:**289–300. [http://www.jstor.org/stable/2346101]
51. Ferreira JA, Zwinderman AH: **On the Benjamini-Hochberg Method.** *Ann Stat* 2006, **34**(4):1827–1849. [http://www.jstor.org/stable/25463486]
52. Bremm S, Schreck T, Boba P, Held S, Hamacher K: **Computing and visually analyzing mutual information in molecular co-evolution.** *BMC Bioinformatics* 2010, **11:**330. [http://www.biomedcentral.com/1471-2105/11/330]
53. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks.** *Proc Nat Acad Sci* 1992, **89**(22):10915–10919. [http://www.pnas.org/content/89/22/10915.abstract]
54. Cappellini V, Sommer HJ, Bruzda W, Zyczkowski K: **Random bistochastic matrices.** *J Phys A: Math Theor* 2009, **42:**23.

doi:10.1186/1471-2105-13-225
Cite this article as: Gültas *et al.*: Coupled mutation finder: A new entropy-based method quantifying phylogenetic noise for the detection of compensatory mutations. *BMC Bioinformatics* 2012 **13**:225.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

9.2 Appendix B: Quantum Coupled Mutation Finder

Gültas *et al. BMC Bioinformatics* 2014, **15**:96
http://www.biomedcentral.com/1471-2105/15/96

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Quantum coupled mutation finder: predicting functionally or structurally important sites in proteins using quantum Jensen-Shannon divergence and CUDA programming

Mehmet Gültas[1,2*], Güncel Düzgün[1], Sebastian Herzog[1], Sven Joachim Jäger[1], Cornelia Meckbach[1], Edgar Wingender[2] and Stephan Waack[1*]

Abstract

Background: The identification of functionally or structurally important non-conserved residue sites in protein MSAs is an important challenge for understanding the structural basis and molecular mechanism of protein functions. Despite the rich literature on compensatory mutations as well as sequence conservation analysis for the detection of those important residues, previous methods often rely on classical information-theoretic measures. However, these measures usually do not take into account dis/similarities of amino acids which are likely to be crucial for those residues. In this study, we present a new method, the Quantum Coupled Mutation Finder (QCMF) that incorporates significant dis/similar amino acid pair signals in the prediction of functionally or structurally important sites.

Results: The result of this study is twofold. First, using the essential sites of two human proteins, namely epidermal growth factor receptor (EGFR) and glucokinase (GCK), we tested the QCMF-method. The QCMF includes two metrics based on quantum Jensen-Shannon divergence to measure both sequence conservation and compensatory mutations. We found that the QCMF reaches an improved performance in identifying essential sites from MSAs of both proteins with a significantly higher Matthews correlation coefficient (MCC) value in comparison to previous methods. Second, using a data set of 153 proteins, we made a pairwise comparison between QCMF and three conventional methods. This comparison study strongly suggests that QCMF complements the conventional methods for the identification of correlated mutations in MSAs.

Conclusions: QCMF utilizes the notion of entanglement, which is a major resource of quantum information, to model significant dissimilar and similar amino acid pair signals in the detection of functionally or structurally important sites. Our results suggest that on the one hand QCMF significantly outperforms the previous method, which mainly focuses on dissimilar amino acid signals, to detect essential sites in proteins. On the other hand, it is complementary to the existing methods for the identification of correlated mutations. The method of QCMF is computationally intensive. To ensure a feasible computation time of the QCMF's algorithm, we leveraged Compute Unified Device Architecture (CUDA).
The QCMF server is freely accessible at http://qcmf.informatik.uni-goettingen.de/.

---

*Correspondence: gueltas@cs.uni-goettingen.de, waack@cs.uni-goettingen.de;
[1] Institute of Computer Science, University of Göttingen, Goldschmidtstr. 7, 37077 Göttingen, Germany
[2] Institute of Bioinformatics, University of Göttingen, Goldschmidtstr. 1, 37077 Göttingen, Germany

BioMed Central

© 2014 Gültas et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

Background

Multiple sequence alignments (MSAs) of homologous protein sequences give us information about two major features of the proteins of interest. The first one consists of easily detectable highly conserved residue sites that are obviously important for the structure and/or the function of the protein; while the second one corresponds to compensatory (coupled) mutations between two or more residue sites that also contain crucial information on the structural and functional basis of proteins [1]. These compensatory mutations occur according to the functional coupling of mutation positions which might be explained as one mutation in a certain site affecting a compensating mutation at another site, even if both related residue sites are distantly positioned in the protein structure [2-5]. In particular, such mutations at essential residue sites are likely to destroy protein structure which often results in loss of the protein function [6,7]. Thus, recognition of these residue sites is as important as the strictly conserved positions for the understanding of the structural basis of protein functions and for the identification of functionally important residue positions [5,8,9].

Although the strictly conserved residue sites are easily detectable and interpretable in MSAs, the detection of important non-conserved compensatory mutation sites needs more complex approaches. Today, due to the simplicity and efficiency, the mutual-information-based metrics (MI-metrics) are often used to measure the co-evolutionary relationship between residue sites in MSAs [4-6,10-13]. However, the MI-metrics strongly depend on the amino acid distributions observed in the MSA columns rather than on physical or biochemical constraints of amino acids that are likely to be crucial for the detection of functionally or structurally important compensatory mutations in a protein sequence. Further, according to the phylogenetic relationship of protein sequences and background noise, there is always a MI-value between each column pair in an MSA. Therefore, the challenging problems in bioinformatics for the detection of significant compensatory mutation signals are: i) the minimization of the influence of phylogenetic relationships of protein sequences by incorporating physical or biochemical properties of amino acids in the calculation; ii) the separation of significant signals from the background noise or unrelated pair signals.

In order to eliminate the influence of phylogeny and noise effects of MI, Dunn et al. [6] have introduced the average product correction (APC). Subtracting APC from MI, they obtained their MIp metric. However, in their model the reduction of background noise is not quantified. On the other hand, Gao et al. [13] have integrated amino acid background distribution (MIB) in the calculation of their MI-metric and focused on only 25 column pairs of each MSA with the highest normalized MI values as significant to reduce noisy effect which seems to be over-conservative, yet specific.

Large efforts have been made in the last few years to improve local-correlation-measure-based approaches to residue co-evolution when it comes to modeling effects that rely on spatial proximity (see [14] for an overview). In this case, it is necessary to disentangle direct and indirect correlations. Classical mutual information, for example, is high not only if the two sites under study are close in 3D space. Quite the contrary, any local measure of correlation, not just mutual information, is limited by the transitivity effect.

To overcome this problem, global statistical models of protein families are employed. The direct-coupling analysis (DCA) works as follows. Maximizing the entropy subject to preserving the single and pair residue frequencies observed, a joint probability distribution on all possible members of the protein family is derived. Utilizing this distribution, considerable progress in predicting residue-residue contacts in 3-dimensional protein structures was made [15-17]. Protein Sparse Inverse Covariance (PSICOV) [18] achieves disentanglement of direct and indirect correlations by inverting a residue-residue covariance matrix. In [19] further progress was made by integrating structural context and sequence co-evolution information.

There is merely a small number of methods that incorporate amino acid similarity in the prediction of functionally or structurally important sites. In this context, it is natural to partition the amino acids into chemically similar groups before applying an information-theoretic measure like the Shannon entropy [20,21]. It was reported that many other methods fail to outperform this simple partition approach [22]. However, quantum information theory supplies a well-studied and powerful framework to integrate such similarity, where the classical Shannon entropy is swapped for the von Neumann entropy (VNE). Caffrey et al. [23] and Johansson et al. [24] have firstly introduced VNE to multiple sequence alignment analysis although they did not treat amino acid pair similarity.

Recently, a new method called Coupled Mutation Finder (CMF) has been introduced by Gültas et al. [5] to deal with phylogenetic noise as well as background signals and to quantify the error made in terms of the false discovery rate. The CMF method only focuses on BLOSUM62-dissimilar amino acid pairs as a model of compensatory mutations and integrated them in the calculation of normalized MI-metrics using a doubly stochastic matrix to transform the empirical pair distribution of the column pair. However, the CMF disregards amino acid pair similarity which can be also crucial for the detection of functionally or structurally important sites in MSAs.

In this study, we present a new method called Quantum Coupled Mutation Finder (QCMF) which extends the CMF algorithm [5] by additionally incorporating amino acid pair similarity. To this end, the QCMF invokes principles from quantum information theory, in particular for the first time in the context of MSA analysis quantum entanglement as a major resource of quantum information. Amino acid pair distributions are replaced by entangled density matrices from quantum mechanics which encompass in our case both empirical pair distributions, possibly transformed by the doubly stochastic matrix used in [5], and pair similarity. Following Capra and Singh [22] who pointed out that it is hard to improve upon metrics based on Jensen-Shannon divergences, we quantify the effect of both amino acid pair similarity and amino acid pair dissimilarity by the quantum Jensen-Shannon divergence between an entangled density matrix and the one that simply represents the amino acid pair frequencies.

The QCMF algorithm is strongly based on the matrix operations that are computationally intensive. When analyzing a single MSA, the computational time of these matrix operations rise very quickly due to the huge number of column pairs. In order to speed up the running time of the QCMF, we implemented its algorithm using Compute Unified Device Architecture (CUDA). CUDA is an efficient parallel computing architecture developed by NVIDIA that utilizes graphic processing units (GPUs) for general-purpose scientific and engineering applications [25]. Nowadays, GPUs are often used for computationally challenging problems in bioinformatics [26-29] and several other scientific fields [30-32].

Results

Our main focus in this study was to investigate whether quantum information theory based measures could contribute beyond conventional measures to the identification of important residue sites. The Results section of this work twofold. First, to test the functionality of QCMF-significant individual residue sites we analysed the essential sites of two human proteins: epidermal growth factor receptor (EGFR) (pdb entry 2J6M) and glucokinase (GCK) (pdb entry 1V4S). The functionally and structurally important sites of both proteins have been experimentally investigated in several studies previously [33-44] and their positions were summarized in [5] as essential sites. The essential sites of these proteins consist of several non-conserved residue sites which are directly located at or near disease associated amino acid mutation (non-synonymous single nucleotide polymorphisms (nsSNPs)) sites, catalytic sites, protein binding sites and so on, each of which are likely to affect protein stability or functionality (see [5] and references therein). In addition, residue sites are defined to be in contact according to the "nearby" definition of Nussinov et al. [45] if their carbon major atoms have a distance of less than or equal to 6 Å. Consequently, we defined an individual QCMF-significant residue site as "functionally or structurally important" if it corresponds to one of these essential sites.

Second, to further investigate the performance of QCMF and to make a comparison with the previous methods (CMF [5], MIp [6], and PSICOV [18]), we selected a non-redundant set of proteins prepared by Janda et al. [46]. Although the dataset contains 216 proteins, we eliminated a few proteins due to inconsistency between corresponding MSAs and PDB files, so that we finally ended up with a dataset of 153 proteins (see Additional file 1).

The MSAs for each protein, which contain after filtering at least 125 independent sequences, were derived from the HSSP-database [47] that merges primary structure information and tertiary structure information of proteins.

Finally, we define QCMF-significant sites as follows. Let M be an MSA, with the protein of interest being the first row of M. A site pair as well as an individual site of the protein are said to be QCMF-significant with respect to the MSA M, if they are $(\mathbb{Q}_{\text{ent}}, M)$-significant or $(\mathbb{Q}_{\text{sep}}, M)$-significant. The latter two notions and the underlying two co-evolutionary column pair metrics $\mathbb{Q}_{\text{ent}}$ and $\mathbb{Q}_{\text{sep}}$ are defined in the Methods section. If the MSA M is fixed, we speak of $\mathbb{Q}_{\text{ent}}$-significance and $\mathbb{Q}_{\text{sep}}$-significance, rather than of $(\mathbb{Q}_{\text{ent}}, M)$-significance and $(\mathbb{Q}_{\text{sep}}, M)$-significance, respectively.

QCMF-significant residue sites in the Human Epidermal Growth Factor Receptor (EGFR) protein

Using the MSA-specific statistical model with a false discovery rate (FDR) of 1% for both QCMF-metrics, we first determined altogether 2688 out of 26079 non-conserved column pairs as significant in corresponding MSA of human EGFR protein. 631 of these significant pairs were detected by $\mathbb{Q}_{\text{ent}}$-metric, and 2149 pairs were detected by $\mathbb{Q}_{\text{sep}}$-metric. Only 92 significant column pairs were detected by both metrics. After that, utilizing the connectivity degree technique, we predicted in total 33 residue sites in corresponding sequence of human EGFR protein as QCMF-significant (see Additional file 2). 12 of them are only $\mathbb{Q}_{\text{ent}}$-significant and 18 residue sites are $\mathbb{Q}_{\text{sep}}$-significant, the remaining 3 residue sites (A839, A882 and V902) are both $\mathbb{Q}_{\text{ent}}$-significant and $\mathbb{Q}_{\text{sep}}$-significant.

10 of the QCMF-significant residue sites are in contact with either catalytic residues or critical active site regions for gefitinib binding site in wild type EGFR kinase [34,37,48] (see Figure 1 and Figure 2). Among these sites, the A839 and R841 have been verified as catalytic residue sites through the Catalytic Site Atlas [48]. The T854 is a gefitinib binding site by itself and the residue sites V845 and A859 are also in contact with nsSNP positions K846,

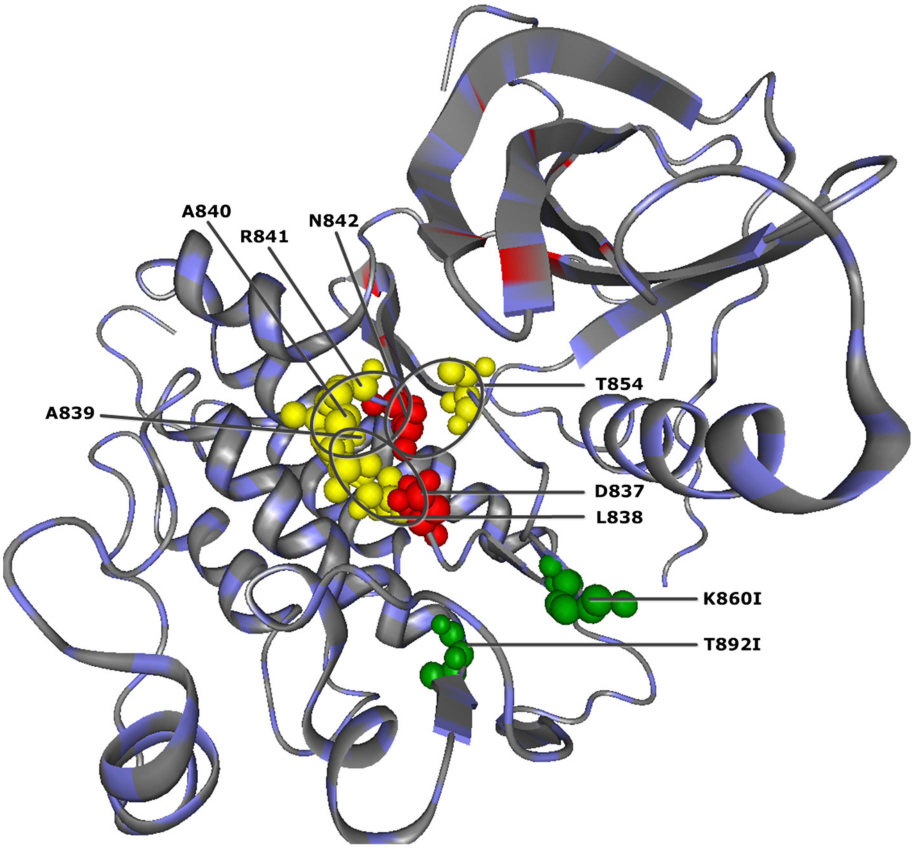


Figure 1 QCMF-significant residue positions are in contact with catalytic residues in human EGFR protein (PDB-Entry 2J6M). Red spheres denote positions of the catalytic residues. Yellow spheres show the localization of significant adjacent residue positions found by QCMF which are in contact with these catalytic residues. Moreover, the QCMF-significant sites A839 and R841 are also catalytic residues by themselves. Green spheres show the structural localization of nsSNP positions found by QCMF as significant in the EGFR protein. The circles indicate clusters of catalytic residue sites and their significant adjacent sites.

T847 and K860 in human EGFR protein. Moreover, two out of all 33 significant sites are related to disease associated nsSNP positions and their structural localization are illustrated in Figure 1.

Additionally, 13 out of all QCMF-significant sites are referred to as essential sites, each of them are either nearby strictly conserved residues or nsSNPs (see Table 1).

According to the essential sites of human EGFR protein, published in [5], we have shown altogether the structural or functional importance of 25 QCMF-significant sites. The remaining 8 significant residue sites (G729, T851, G779, Q820, M825, L927, G930, Y944) do not fall into essential sites and the reason for their significance and their importance in the EGFR protein is currently unclear.

QCMF-significant residue sites in the Human Glucokinase (GCK) protein

Like human EGFR protein, applying the MSA-specific statistical model with a FDR of 1% for both QCMF-metrics we identified a total of 9853 out of 69645 non-conserved column pairs as significant in the human GCK protein (pdb entry 1V4S). 6070 of them were $(\mathbb{Q}_{\text{ent}}, M)$-significant and 4232 were detected as $(\mathbb{Q}_{\text{sep}}, M)$-significant. Only 449 column pairs were detected as significant with respect to both metrics. Thereupon using the connectivity degree technique, we determined altogether 64 residue sites in the human GCK protein as QCMF-significant (see Additional file 3). 30 of them are determined as $\mathbb{Q}_{\text{ent}}$-significant and further 30 significant residue sites are determined as $\mathbb{Q}_{\text{sep}}$-significant. Only four residue sites (T82, G223, V253, and G407) are significant based on both metrics.

13 of QCMF-significant sites are in contact with allosteric sites V62, R63, M210, I211, Y214, Y215, M235, V452, V455 and A456 in the human GCK protein. Among these significant sites, the $V62$, $M210$, $Y215$ are allosteric sites by themselves [41] and the T209M, G223S and S453del are related to disease associated nsSNP positions. In addition, there are further five QCMF-significant sites (F123L, G162D, G175R,

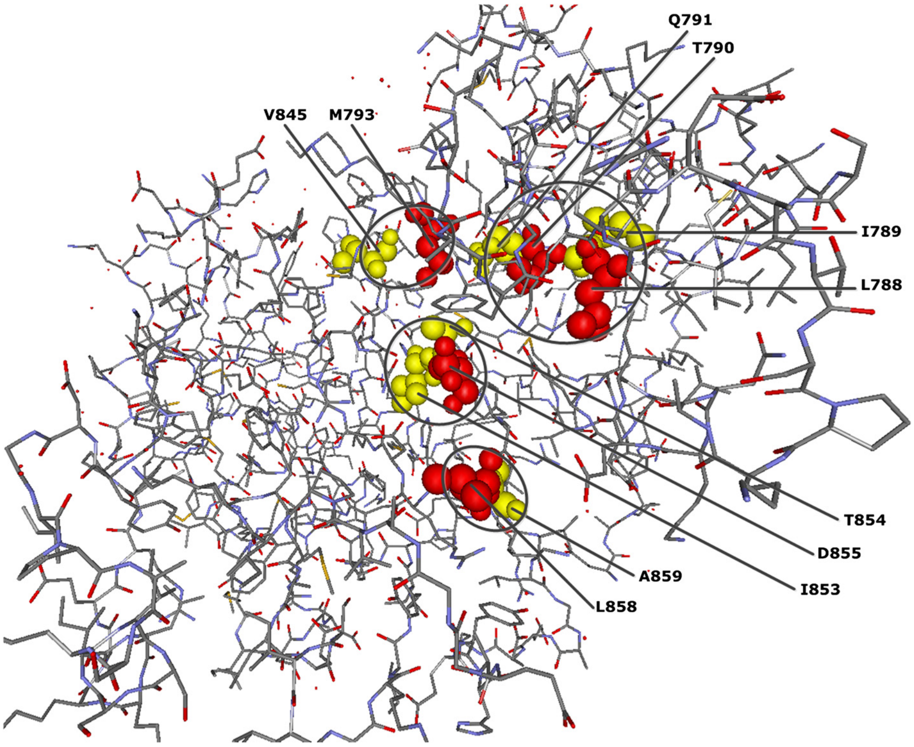


Figure 2 QCMF-significant residue positions are in contact with gefitinib binding sites in human EGFR protein (PDB-Entry 2J6M). Red spheres show the structural localization of the gefitinib binding sites in the wild type kinase. Yellow spheres show QCMF-significant adjacent residue positions which are in contact with these binding sites. Moreover, the QCMF-significant site T854 is also a binding site by itself and interacts with gefitinib binding site D855. The circles indicate clusters of gefitinib binding sites and their significant adjacent sites.

Table 1 QCMF-significant essential sites in the human EGFR protein, which are nearby either nsSNPs or strictly conserved sites

QCMF-significant essential sites	Nearby nsSNPs, or strictly conserved sites	Reference
N771	773^{s}	[44]
G824	773^{s}	[44]
Y827	829^{s}	[44]
L828	829^{s}	[44]
V834	835^{c}, 836^{s}, 860^{s}	[44,49]
Y891	892^{s}, 895^{c}	[44]
A822	861^{s}	[43,49,50]
V844	796^{c}, 798^{c}, 852^{c}	-
A882	884^{c}, 895^{c}, 898^{c}	-
Y900	898^{c}, 901^{c}	-
V902	880^{c}, 901^{c}	-
T909	906^{c}, 936^{c}	-
G911	906^{c}	-

^{s}: non-synonymous snp site, ^{c}: strictly conserved site.

T228M, and E300K,Q) that have been verified as nsSNP positions through annotation databases and previous experimental studies [38-40,42,43,51]. The structural localization of these 18 QCMF-significant sites (contact sites and nsSNPs positions) are illustrated in Figure 3.

Additionally, eight significant sites T149, G170, F171, T206, V207, A208, Q287 and G294 in contact with glucose binding sites (active sites) T168, K169, D204, D205 and E290 in human GCK protein [41] (see Figure 4) where V207 and A208 are also in contact with the allosteric sites M210 and I211.

Moreover, we have also observed that 38 QCMF-significant sites are further included in essential sites since they are nearby nsSNPs or strictly conserved residues in human GCK protein (see Table 2).

In total, we have demonstrated here that according to the essential sites of GCK, 62 out of 64 QCMF- significant sites are functionally or structurally important for human GCK protein. The remaining two significant residue sites V89 and N283 do not overlap with essential sites and the reason for their significance and their role in the GCK protein is still unclear.

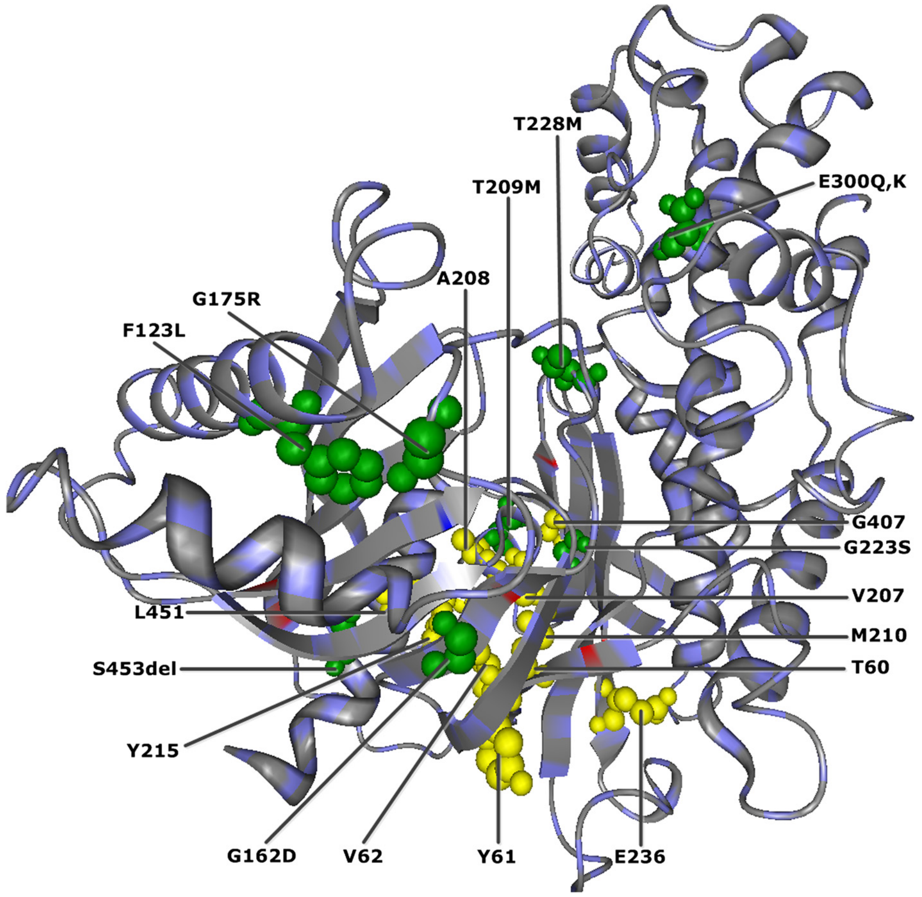


Figure 3 QCMF-significant positions that are either in contact with allosteric sites or related to nsSNPs in human GCK protein (PDB-Entry 1V4S). Yellow spheres correspond to structural localization of ten significant residue sites which are in contact with allosteric sites where V62, M210, and Y215 are denoted as allosteric sites by themselves and they are also in contact with an other allosteric sites. Green spheres indicate eight significant nsSNP positions in the GCK protein. Three of them (T209M, G223S and S453del) are further in contact with allosteric sites M210, I211, V452, V455 and A456.

Individual residue site comparison between QCMF-significant sites and previous CMF-significant sites

We compared QCMF-significant residue sites for both human EGFR and GCK proteins with the significant residue sites given in [5] of the previous CMF-method. The CMF-method detected for both human proteins, 43 sites in EGFR and 72 sites in GCK as significant.

For the EGFR protein we found that the QCMF-significant residue sites Q791, Q820, G824, K860, Y891, T892, Y900, T909 overlap with results of the CMF-method. Interestingly, one of the unconfirmed residue sites, the Q820, has been predicted by both QCMF-method and CMF-method as significant.

For GCK protein, we observed that in total 24 QCMF-significant sites (T60, T82, N83, F123, F148, T149, F152, H156, F171, N180, T206, T209, T228, E236, G260, L271, S281, N283, Q287, G294, E300, T332, F419 and E443) were also determined by the CMF-method as significant. Although both methods detected residue site N283 as significant, it corresponds to one of the unconfirmed residue sites for GCK, currently.

The CMF has been developed using normalized mutual information (MI) measures in order to detect important residue positions in MSAs. The method mainly focuses on significant BLOSUM62-dissimilar amino acid signals as a model of compensatory mutations and integrates them in the calculation of normalized MI-metrics. As a consequence of mainly taking into account dissimilar amino acid signals, an important part of CMF-significant sites were verified as disease associated nsSNP positions and just a small part of them were located at or near the catalytic sites, allosteric sites and binding sites in both proteins.

Moreover, when statistically evaluating both methods, we have observed that the QCMF significantly outperforms the QCMF-method. The QCMF reaches an improved performance in identifying essential sites from MSAs of both proteins with a significantly higher

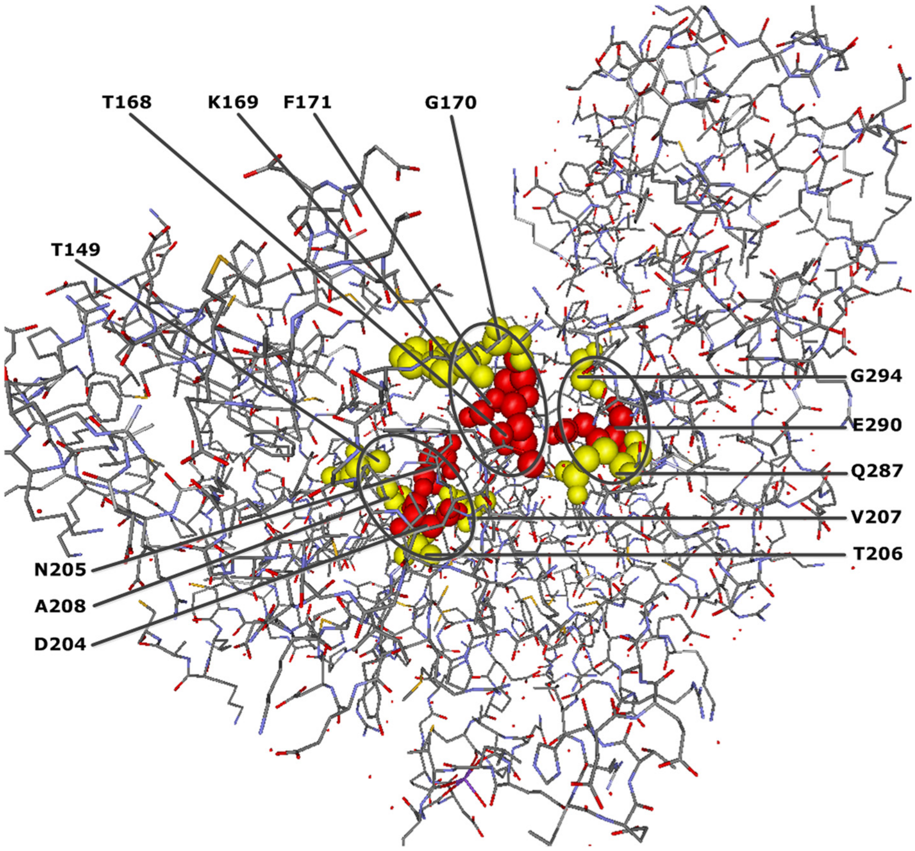


Figure 4 QCMF-significant residue positions are in contact with glucose binding site in human GCK protein (PDB-Entry 1V4S). (A) Red spheres show the structural positions of the glucose binding sites (active sites) and yellow spheres show the localization of significant adjacent residue positions found by QCMF which are in contact with these active sites. The circles indicate clusters of glucose binding sites and their significant adjacent sites.

Matthews correlation coefficient (MCC) value of 0.215 whereas the CMF reaches only a MCC value of 0.133.

Significant residue pair comparison

To analyze whether the quantum-information-theory-based measures proposed in this study complements the coventional methods for the detection of correlated (co-evolutionary) mutations, we made pairwise comparisons between our new QCMF, MIp [6], PSICOV [18], and CMF [5].

All four methods take as input an MSA satisfying certain admissibility criteria. The problem is that QCMF and CMF output the set of QCMF-significant sites and CMF-significant sites of M's reference protein, respectively, whereas PSICOV and MIp result in sets of important residue pairs. To make these outputs comparable, we extend them in all cases.

Let $\mathcal{V}_{\text{QCMF}}$ denote the output of QCMF on any admissible MSA M. We extend this set to what we call the QCMF-significant residue network $\mathcal{N}_{\text{QCMF}} := (\mathcal{V}_{\text{QCMF}}, \mathcal{E}_{\text{QCMF}})$ of M as follows. Any two elements of $\mathcal{V}_{\text{QCMF}}$ are connected by an undirected edge belonging to $\mathcal{E}_{\text{QCMF}}$ if and only if the corresponding column pair is QCMF-significant.

The CMF-significant residue network $\mathcal{N}_{CMF}$ is analogously defined.

In order to get a sufficiently large number MIp-significant and PSICOV-significant residue pairs, for every input MSA we simply took the top-ranking 10% as MIp-significant and PSICOV-significant, respectively.

We then utilized the connectivity degree technique in the same way as we did for CMF and QCMF to calculate the set of MIp-significant sites $\mathcal{V}_{\text{MIp}}$ and the set of PSICOV-significant sites $\mathcal{V}_{\text{PSICOV}}$.

For all four methods we used the 90th, the 95th and the 99th percentile as *cut-off* values.

Finally, the edge sets $\mathcal{E}_{\text{MIp}}$ and $\mathcal{E}_{\text{PSICOV}}$ were determined by full analogy with the calculation of $\mathcal{E}_{\text{QCMF}}$ and $\mathcal{E}_{\text{CMF}}$. Thus we obtained the MIp-significant residue network $\mathcal{N}_{\text{MIp}}$ and the PSICOV-significant residue network $\mathcal{N}_{\text{PSICOV}}$.

Table 2 QCMF-significant essential sites in the human GCK protein, which are nearby either nsSNPs or strictly conserved sites

QCMF-significant essential sites	Nearby nsSNPs or strictly conserved sites	Reference
*M*37	$36^{s}, 39^{s}, 40^{s}$	[38,39,43,51]
*S*76	147^{c}	
*L*79	$78^{c}, 80^{c}, 150^{c}$	-
*T*82	81^{c}	-
*N*83	$81^{c}, 108^{s}, 110^{s}$	[43,51]
*V*86	$85^{c}, 106^{s}$	[38]
*S*127	130^{s}	[40]
*F*148	$147^{c}, 150^{c,s}$	[38,39,43,51]
*F*152	$150^{c,s}, 151^{c}$	[39,43,51]
*P*153	154^{s}	[39]
*H*156	154^{s}	[39]
*A*176	$119^{s}, 175^{s}$	[43]
*G*178	164^{c}	
*N*180	$162^{s}, 182^{s}$	[38,39,43],
*L*185	$182^{s}, 188^{s}$	[39,43,51]
*A*201	$147^{c}, 453^{c}$	
*M*202	$147^{c}, 203^{s}$	[43]
*A*232	$223^{s}, 231^{c}$	[39,40,51]
*C*233	$223^{s}, 234^{c}, 235^{c}$	[39,40,51]
*V*253	$234^{c}, 254^{c}$	
*F*260	$257^{s}, 258^{c}, 259^{s}, 261^{s}$	[39,43]
*L*271	274^{c}	
*V*277	$274^{c}, 278^{c}, 279^{s}$	[43]
*S*281	$278^{c}, 279^{s}$	[43]
*Y*297	$291^{c}, 295^{c}, 299^{c}, 300^{s}$	[43]
*M*298	$295^{c}, 299^{c}, 300^{s}$	[43]
*T*332	$295^{c}, 299^{c}$	
*V*374	377^{c}	
*A*378	$377^{c}, 382^{s}$	[43]
*A*379	$377^{c}, 382^{s}$	[43]
*S*383	$382^{s}, 385^{s}$	[43]
*A*384	$382^{s}, 385^{s}$	[43]
*A*387	385^{s}	[43]
*S*388	$385^{s}, 392^{s}$	[38,43]
*V*412	$226^{s}, 227^{c}, 410^{c}, 414^{s}, 416^{s}$	[40,43]
*F*419	416^{s}	[40]
*E*443	$444^{c}, 445^{c}, 447^{s}$	[39]
*G*446	$444^{c}, 445^{c}, 447^{s}, 448^{c}, 449^{c}$	[39]

^{s} : non-synonymous snp site, ^{c} : strictly conserved site.

We performed the method comparison edge-oriented, with the number of overlapping edges as measure. We applied all four methods to the 153 MSAs (see Additional files 1) described at the very beginning of this section and calculated the numbers $\left|\mathcal{E}^{(i)}_{\mathrm{QCMF}}\right|$, $\left|\mathcal{E}^{(i)}_{\mathrm{CMF}}\right|$, $\left|\mathcal{E}^{(i)}_{\mathrm{PSICOV}}\right|$, $\left|\mathcal{E}^{(i)}_{\mathrm{MIp}}\right|$, $\left|\mathcal{E}^{(i)}_{\mathrm{QCMF}} \cap \mathcal{E}^{(i)}_{\mathrm{MIp}}\right|$, $\left|\mathcal{E}^{(i)}_{\mathrm{QCMF}} \cap \mathcal{E}^{(i)}_{\mathrm{PSICOV}}\right|$, $\left|\mathcal{E}^{(i)}_{\mathrm{QCMF}} \cap \mathcal{E}^{(i)}_{\mathrm{CMF}}\right|$, $\left|\mathcal{E}^{(i)}_{\mathrm{MIp}} \cap \mathcal{E}^{(i)}_{\mathrm{PSICOV}}\right|$, $\left|\mathcal{E}^{(i)}_{\mathrm{MIp}} \cap \mathcal{E}^{(i)}_{\mathrm{CMF}}\right|$ and $\left|\mathcal{E}^{(i)}_{\mathrm{PSICOV}} \cap \mathcal{E}^{(i)}_{\mathrm{CMF}}\right|$ on each of them, where the connectivity cut-off ranges over the 90th, the 95th and the 99th percentile, and $i = 1, 2, \ldots, 153$. Summing up the 153 numbers in each of these groups results in the numbers $\sum_{i=1}^{153} \left|\mathcal{E}^{(i)}_{\mathrm{QCMF}}\right|$, $\sum_{i=1}^{153} \left|\mathcal{E}^{(i)}_{\mathrm{CMF}}\right|$, $\sum_{i=1}^{153} \left|\mathcal{E}^{(i)}_{\mathrm{PSICOV}}\right|$, $\sum_{i=1}^{153} \left|\mathcal{E}^{(i)}_{\mathrm{MIp}}\right|$, $\sum_{i=1}^{153} \left|\mathcal{E}^{(i)}_{\mathrm{QCMF}} \cap \mathcal{E}^{(i)}_{\mathrm{MIp}}\right|$, $\sum_{i=1}^{153} \left|\mathcal{E}^{(i)}_{\mathrm{QCMF}} \cap \mathcal{E}^{(i)}_{\mathrm{PSICOV}}\right|$, $\sum_{i=1}^{153} \left|\mathcal{E}^{(i)}_{\mathrm{QCMF}} \cap \mathcal{E}^{(i)}_{\mathrm{CMF}}\right|$, $\sum_{i=1}^{153} \left|\mathcal{E}^{(i)}_{\mathrm{MIp}} \cap \mathcal{E}^{(i)}_{\mathrm{PSICOV}}\right|$, $\sum_{i=1}^{153} \left|\mathcal{E}^{(i)}_{\mathrm{MIp}} \cap \mathcal{E}^{(i)}_{\mathrm{CMF}}\right|$ and $\sum_{i=1}^{153} \left|\mathcal{E}^{(i)}_{\mathrm{PSICOV}} \cap \mathcal{E}^{(i)}_{\mathrm{CMF}}\right|$, which are displayed in Tables 3 and 4.

Table 3 shows that all methods detect with the same connectivity degree cut-off a comparable number of edges in the corresponding significant residue network.

Table 4 highly suggests that all four methods carry distinct information. The overlap between any two of them is less than or equal to 10%. This indicates that, under the assumption that each of them models important aspects of co-evolution, they complement each other perfectly. In particular, this is true for QCMF as a quantum-information-science-based service compared with the other three established tools that are based on conventional methods.

Implementation of QCMF: Parallel computing using CUDA

The computation of both QCMF metrics (Equations 7 and 8) is strongly based on matrix operations. Therefore, we implement QCMF algorithm using CUDA [25] which is very suitable to perform large number of vector and matrix operations in real time. This results in a dramatic reduction of computational time of QCMF.

In this study, we use the CUDA 4.0 architecture (Toolkit) with several linear algebra libraries such as MAGMA [52], LAPACK [53], BLAS [54], GotoBLAS [55], CUBLAS [25] together (see Figure 5) to speed up the running time of the QCMF algorithm. Since our program requires a cooperative multi threading to not fall in any asynchronicity or locks we extended the magma library with dynamic scheduling features according to [56]. Further, in order to be able to compare the performance, we also implemented the QCMF algorithm onto CPU architecture alone. Both implementations were performed on an Intel Core™ i7-3770K Processor operating at 3.9GHz, with 16 GB of DDR3 RAM and a GeForce GTX 680

Table 3 Total number of edges in method-dependent significant residue networks with respect to various connectivity degree cut-offs

	Total number of edges in significant residue networks		
Connectivy degree cut-off	**90%th percentile**	**95%th percentile**	**99%th percentile**
$\sum_{i=1}^{153} \left\lvert \boldsymbol{\mathcal{E}}_{\text{QCMF}}^{(i)} \right\rvert$	82561	20411	435
$\sum_{i=1}^{153} \left\lvert \boldsymbol{\mathcal{E}}_{\text{MIp}}^{(i)} \right\rvert$	90636	24094	1454
$\sum_{i=1}^{153} \left\lvert \boldsymbol{\mathcal{E}}_{\text{PSICOV}}^{(i)} \right\rvert$	80489	21596	1088
$\sum_{i=1}^{153} \left\lvert \boldsymbol{\mathcal{E}}_{\text{CMF}}^{(i)} \right\rvert$	87208	23893	936

graphics card using the Ubuntu 13.04 operating system (64-bit version).

Applying the QCMF algorithm for human EGFR protein with CPU alone and with CUDA acceleration, the average computational time of a column pair was 0.7117 seconds and 0.0301 seconds, respectively. Similarly, for human GCK protein, the average computational time of a column pair was 0.6977 seconds with CPU alone and 0.0299 seconds with CUDA acceleration. Consequently, the algorithm took $\sim$ 310 minutes for human EGFR protein and $\sim$ 811 minutes for GCK protein with CPU alone. On the other hand, applying the CUDA acceleration it took only $\sim$ 13 minutes for EGFR and $\sim$ 39 minutes for GCK protein. The comparison between the average times indicates that the required computational time of QCMFalgorithm with the CUDA acceleration was significantly faster than with CPU alone (approximately more than 23 times faster).

Methods

We predict important sites of a protein by detecting co-evolving residues. Our measures of co-evolution are quantum-Jensen-Shannon-divergence-based metrics of column pairs of a multiple sequence alignment, with the protein under study being the reference row. The quantum Jensen-Shannon divergence in turn has the von Neumann entropy as main building block.

The von Neumann entropy was originally defined in the framework of quantum mechanics. We elucidate it in the subsequent section as far as it is necessary to understand our methods. Researchers interested in learning more are referred to the excellent textbook due to Vedral [57]. A comprehensive reference book was published by Nielsen and Chuang [58].

This section is organized as follows. In the first four subsections we recapitulate techniques developed in [5] which we leverage in this study. This concerns the definition of significant site pairs and of significant individual sites, the preparation of the training data set used, and the computation of a doubly stochastic matrix D as our model of compensatory mutations on grounds of two counting matrices C_{alt} and C_{null}. These two matrices also form the basis of the two amino acid pair similarity matrices $\mathcal{A}_{\text{ent}}$ and $\mathcal{A}_{\text{sep}}$, which in turn give rise to our new quantum-information-science-based metrics $\mathbb{Q}_{\text{ent}}$ and $\mathbb{Q}_{\text{sep}}$. The last four subsection are dedicated to their definitions.

Significant column pairs and significant position with respect to a certain metric

Let M be an MSA, where the protein of interest is represented by M's first row, and let $\mathbb{E}$ be a metric which assigns to every MSA column pair (γ_1, γ_2) a real number $\mathbb{E}(\gamma_1, \gamma_2) \in [0, 1]$. We call $\mathbb{E}$ a *co-evolutionary column pair metric* if it models a biologically meaningful co-evolutionary signal: The larger the metric value on (γ_1, γ_2), the more likely co-evolution between position γ_1 and position γ_2 has occurred.

Let $\widehat{p}_{(i,j)}$ be the empirical relative amino acid pair frequency of the i-th and the j-th amino acid in column pair (γ_1, γ_2), where $i, j = 1, 2, \ldots, 20$. (When choosing a row of

Table 4 Total number of edges in two networks of different type with respect to various connectivity degree cut-offs

	Total number of common edges in two networks of different type		
Connectivy degree cut-off	**90%th percentile**	**95%th percentile**	**99%th percentile**
$\sum_{i=1}^{153} \left\lvert \boldsymbol{\mathcal{E}}_{\text{QCMF}}^{(i)} \cap \boldsymbol{\mathcal{E}}_{\text{MIp}}^{(i)} \right\rvert$	898	77	0
$\sum_{i=1}^{153} \left\lvert \boldsymbol{\mathcal{E}}_{\text{QCMF}}^{(i)} \cap \boldsymbol{\mathcal{E}}_{\text{PSICOV}}^{(i)} \right\rvert$	735	64	0
$\sum_{i=1}^{153} \left\lvert \boldsymbol{\mathcal{E}}_{\text{QCMF}}^{(i)} \cap \boldsymbol{\mathcal{E}}_{\text{CMF}}^{(i)} \right\rvert$	4036	474	1
$\sum_{i=1}^{153} \left\lvert \boldsymbol{\mathcal{E}}_{\text{MIp}}^{(i)} \cap \boldsymbol{\mathcal{E}}_{\text{PSICOV}}^{(i)} \right\rvert$	9094	1488	11
$\sum_{i=1}^{153} \left\lvert \boldsymbol{\mathcal{E}}_{\text{MIp}}^{(i)} \cap \boldsymbol{\mathcal{E}}_{\text{CMF}}^{(i)} \right\rvert$	3343	474	6
$\sum_{i=1}^{153} \left\lvert \boldsymbol{\mathcal{E}}_{\text{PSICOV}}^{(i)} \cap \boldsymbol{\mathcal{E}}_{\text{CMF}}^{(i)} \right\rvert$	2618	368	2

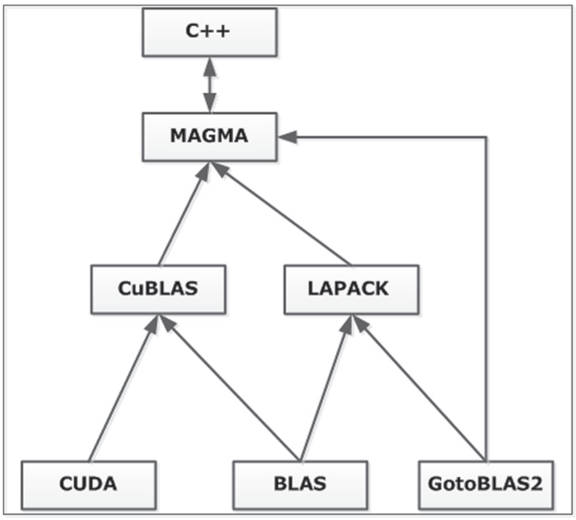


Figure 5 **Linking of the CUDA environment using C++.**

this column pair by pure chance, acid pair (i,j) is drawn with probability $\widehat{p}_{(i,j)}$.) In the subsequent subsection we recapitulate the way developed in [5] to identify significant columns and significant column pairs with respect to $\mathbb{E}$.

A well-studied example (see [5,12]) of a co-evolutionary column pair metric is the normalized mutual information

$$\mathbb{U}(\gamma_1,\gamma_2) := 2 \cdot \frac{\mathbb{H}(\gamma_1)+\mathbb{H}(\gamma_2)-\mathbb{H}(\gamma_1,\gamma_2)}{\mathbb{H}(\gamma_2+\mathbb{H}\gamma_2)}, \tag{1}$$

where $\mathbb{H}(\gamma_1,\gamma_2)$, $\mathbb{H}(\gamma_1)$, and $\mathbb{H}(\gamma_2)$ denote the Shannon entropy of the empirical pair distribution $\left(\widehat{p}_{(i,j)}\right)_{i,j=1,2,\ldots,20}$ of the column pair (γ_1,γ_2) and its two marginals.

In order to identify significant column pairs of the MSA under study with respect to the metric $\mathbb{E}$, in [5] we have pointed out, that the distribution of $\mathbb{E}$ can be regarded as a mixture of a background β-distribution F_0, an unrelated pair distribution G_1, and a distribution G_2 of presumably co-evolving pairs.

The p-values $1 - F_0(\mathbb{E})$ are then uniformly distributed over $[0,1]$ given the underlying $\mathbb{E}$-values are F_0-distributed. In contrast, p-values tend to zero or one, if $\mathbb{E}$-values are G_2-distributed or G_1-distributed, respectively.

If, moreover, there is a sub-interval of $[0,1]$ which contains only data from the background distribution, on grounds of a result due to Storey and Tibshirani [59,60] we determined in [5] an MSA-dependent threshold for $\mathbb{E}$-values. A column pair is said to be $(\mathbb{E},M)$-significant, if its $\mathbb{E}$-value is above the threshold, where the false discovery rate is bounded by a predefined constant.

Figure 6 is a typical pictorial representation of metric distributions which can be treated that way to detect significant pairs.

We applied that model in this study.

We utilized the connectivity degree technique, introduced in [12] and developed further in [5], in order to define the $(\mathbb{E},M)$-significance of individual residue sites. The connectivity degree of a position γ_1 is the number of positions γ_2 so that the site pair (γ_1,γ_2) is $(\mathbb{E},M)$-significant. A site of the protein of interest is then called $(\mathbb{E},M)$-significant, if its connectivity degree *cut-off* exceeds the 90-th percentile.

Training data set and pre-processing

Following [5], a redundancy free set of more than 35000 protein structures is our starting point. This collection was compiled in Rainer Merkl's Lab at the University of Regensburg. The protein structures were taken from the protein data base (http://www.pdb.org/). The PISCES services [61] was applied to assess proteins on sequence similarity and equality of 3D-data. The related MSAs were gathered from the HSSP data base (http://swift.cmbi.ru.nl/gv/hssp/).

Taking pattern from [12], we filtered every MSA obtained as follows. First, highly similar and dissimilar sequences were deleted to ensure that the sequence identity between any two sequences is at least 20% and no more than 90%. Second, we removed strictly conserved residue columns, where the percentage of identical residues is greater than 95%. Third, we eliminated the residue columns which contain more than 25% gaps. Finally, we discarded all MSAs with less than 125 sequences. More than 17000 MSAs survived the last filtering step. We used approximately 1700 MSAs published in [5] as our *training data set* which we randomly chose from this set.

Setting up the counting matrices C_{alt} and C_{null}

The entries of the two matrices are frequencies of pair substitutions calculated from our training data set described in the foregoing subsection. Informally spoken, matrix C_{alt} models the signal, whereas C_{null} reflects the background.

In line with [5], we calculated a signal and a null set of column pairs. The signal set consists of all $(\mathcal{U},M)$-significant column pairs, where M ranges over all training MSA. The null set consists of sufficiently many column pairs randomly chosen from every training MSA. For both the signal set and the null set we computed a symmetric 400 × 400 integer-valued matrix of frequencies of pair substitutions C_{alt} and C_{null}. To this end, the method used to compute BLOSUM62 matrices [62] is applied to count residue pair substitutions in MSA column pairs rather than residue substitution in columns.

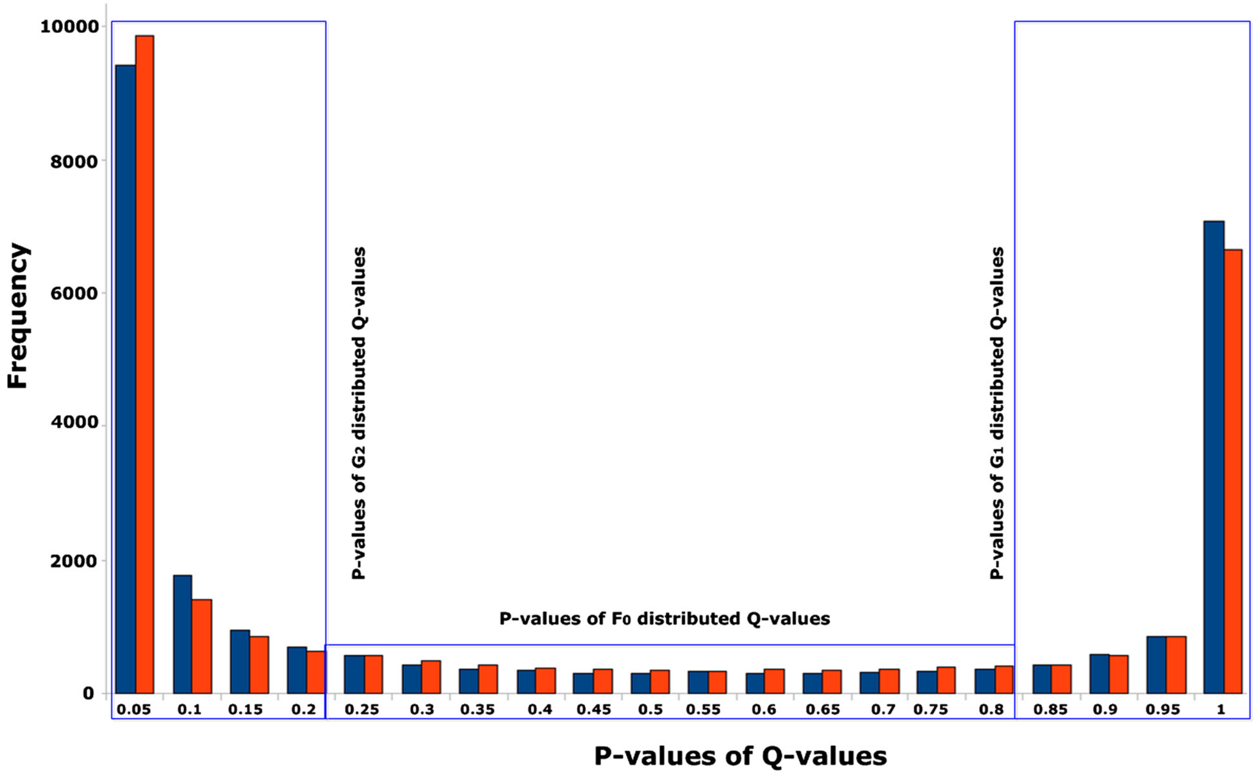


Figure 6 p-value distributions of $\mathbb{Q}_{\text{ent}}$ and $\mathbb{Q}_{\text{sep}}$-values for human EGFR protein (PDB-Entry 2J6M). The blue bars illustrate the $\boldsymbol{p}$-value distribution of the $\mathbb{Q}_{\text{ent}}$-values and red bars display the p-value distribution of the $\mathbb{Q}_{\text{sep}}$-values.

Computing a doubly stochastic matrix *D*

According to [5], a pair $\left((a_i,a_j),(a_k,a_l)\right)$ of amino acid pairs is said to be a *formal dissimilar compensatory mutation*, if the BLOSUM62 score both of (a_i,a_k) and (a_j,a_l) is negative.

Using C_{alt} and C_{null}, we define the matrix C_{CompMut} by

$$C_{\text{CompMut}}\left((a_i,a_j),(a_k,a_l)\right) := \begin{cases} C_{\text{alt}}\left((a_i,a_j),(a_k,a_l)\right) & \text{if } \varphi_{\text{CompMut}}\left((a_i,a_j),(a_k,a_l)\right)=1; \\ 0 & \text{otherwise;} \end{cases}$$

where $\varphi_{\text{CompMut}}\left((a_i,a_j),(a_k,a_l)\right)=1$ if and only if either $(a_i,a_j)=(a_k,a_l)$ or $\left((a_i,a_j),(a_k,a_l)\right)$ is a formal dissimilar compensatory mutation and

$$\frac{C_{\text{alt}}\left((a_i,a_j),(a_k,a_l)\right)}{\sum_{i',j',k',l'} C_{\text{alt}}\left((a_{i'},a_{j'}),(a_{k'},a_{l'})\right)} > \frac{C_{\text{null}}\left((a_i,a_j),(a_k,a_l)\right)}{\sum_{i',j',k',l'} C_{\text{null}}\left((a_{i'},a_{j'}),(a_{k'},a_{l'})\right)}.$$

By normalizing C_{CompMut}, we obtain a symmetric matrix P_{CompMut}. For a_i,a_j,a_k,a_l ranging over all amino acids, $P_{\text{CompMut}}\left((a_i,a_j),(a_k,a_l)\right)$ represents an empirical probability distribution on pairs of amino acid pairs.

We then calculated the symmetric 400×400-matrix

$$S_{\text{CompMut}} := \left(\log \frac{P_{\text{CompMut}}\left((a_i,a_j),(a_k,a_l)\right)}{P^{\text{b}}_{\text{CompMut}}\left(a_i,a_j\right) P^{\text{b}}_{\text{CompMut}}\left(a_k,a_l\right)}\right)_{(a_i,a_j),(a_k,a_l)},$$

where $P^{\text{b}}_{\text{CompMut}}\left(a_i,a_j\right)$ is the marginal distribution of P_{CompMut}.

Having set all negative entries of S_{CompMut} to zero, the doubly stochastic matrix D is computed by means of the canonical iterated row-column normalization procedure [63].

The doubly stochastic D is used to linearly transform empirical amino acid pair distributions of column pairs. If the pair distribution is regarded as a 400-dimensional row vector, matrix D is multiplied from the right. If then, for example, the resulting distribution is plugged into Equation 1, column pairs containing formal dissimilar compensatory mutations the D-transition probability of which is relatively large tend to be up-scaled.

The idea of the subsequent subsections is to design a model of MSA column pairs that takes formal dissimilar compensatory mutations regarded as pair dissimilarities as well as pair similarities into account. The challenge is to implement this in a way such that these two effects interfere but do not interact. This is necessary since a similarity relation is transitive, whereas a dissimilarity relation is not.

Setting up the two counting matrices C_{ent} and C_{sep}

We set up two significant pair substitution matrices C_{ent} and C_{sep} from C_{alt} and C_{null} which form the basis of our new metrics $\mathbb{Q}_{\mathrm{ent}}$ and $\mathbb{Q}_{\mathrm{sep}}$. The intuition behind C_{ent} is that the component-wise BLOSUM62-based pair similarity is rescaled, whereas C_{sep} leads to a new amino acid pair similarity.

$$C_{\mathrm{ent}}\left((a_i,a_j),(a_k,a_l)\right) := \begin{cases} C_{\mathrm{alt}}\left((a_i,a_j),(a_k,a_l)\right) & \text{if } \varphi_{\mathrm{ent}}\left((a_i,a_j),(a_k,a_l)\right)=1;\\ 0 & \text{otherwise;}\end{cases}$$

where $\varphi_{\mathrm{ent}}\left((a_i,a_j),(a_k,a_l)\right) = 1$ if and only if either $(a_i,a_j) = (a_k,a_l)$ or the following two conditions are satisfied. First, the amino acids a_i and a_k as well as the amino acids a_j and a_l are BLOSUM62-similar. Second,

$$\frac{C_{\mathrm{alt}}\left((a_i,a_j),(a_k,a_l)\right)}{\sum_{i',j',k',l'} C_{\mathrm{alt}}\left((a_{i'},a_{j'}),(a_{k'},a_{l'})\right)} > \frac{C_{\mathrm{null}}\left((a_i,a_j),(a_k,a_l)\right)}{\sum_{i',j',k',l'} C_{\mathrm{null}}\left((a_{i'},a_{j'}),(a_{k'},a_{l'})\right)}. \tag{2}$$

$$C_{\mathrm{sep}}\left((a_i,a_j),(a_k,a_l)\right) := \begin{cases} C_{\mathrm{alt}}\left((a_i,a_j),(a_k,a_l)\right) & \text{if } \varphi_{\mathrm{sep}}\left((a_i,a_j),(a_k,a_l)\right)=1;\\ 0 & \text{otherwise;}\end{cases}$$

where $\varphi_{\mathrm{sep}}\left((a_i,a_j),(a_k,a_l)\right) = 1$ if and only if either $(a_i,a_j) = (a_k,a_l)$ or Equation 2 is satisfied.

Calculating the two amino acid pair similarity matrices $\mathcal{A}_{\mathrm{ent}}$ and $\mathcal{A}_{\mathrm{sep}}$

Recall that a matrix A is positive definite (positive semidefinite), if there is an orthogonal matrix U (defining property $U^{-1} = U^T$) such that $U\,AU^T$ is a diagonal matrix, where the coefficients in the main diagonal are strictly positive (non-negative).

Let us call a 400×400-matrix $\mathcal{A}$ a *amino acid pair similarity matrix*, if $\mathcal{A}$ is positive definite and the entries in the main diagonal are equal to 1, whereas the off-diagonal elements $\mathcal{A}_{(g,h),(i,j)}$ ($(g,h) \neq (i,j)$) are greater than or equal to 0, but less than 1.

The entries of an amino acid pair similarity matrix $\mathcal{A}$ are interpreted as follows. The closer $\mathcal{A}_{(g,h),(i,j)}$ to 1, the more similar are the amino acid pairs (g,h) and (i,j).

Let C be either C_{ent} or C_{sep}. We define

$$B_{(g,h),(i,j)} := \frac{C^{\alpha}_{(g,h),(i,j)}}{\sqrt{\sum_{\iota,\kappa=1}^{20} C^{2\alpha}_{(\iota,\kappa),(i,j)}}},$$

where $\left((g,h),(i,j)\right)$ ranges over all possible 160000 indices of pairs of amino acid pairs including the main diagonal, and $\alpha \in (0,1)$ was set to 0.1 in order to enhance the effect of similarity.

Because of the fact, that matrix B is not in any case positive definite, we finally set

$$\mathcal{A} := B^T B, \tag{3}$$

which is justified by the transitivity of similarity. That way the amino acid similarity matrices $\mathcal{A}_{\mathrm{ent}}$ and $\mathcal{A}_{\mathrm{sep}}$ are obtained from the counting matrices C_{ent} and C_{sep}, respectively.

Amino acid pair similarity matrices generalize amino acid similarity matrices used by Johansson et al. [24] for evaluating amino acid conservation.

Modeling MSA column pairs and single columns by means of density matrices

Let (γ_1,γ_2) be a column pair of a multiple sequence alignment, let $\left(\widehat{p}_{(i,j)}\right)_{i,j=1,2,\ldots,20}$ be the empirical amino acid pair distribution in these columns, let $\left(\widehat{q}_{(i,j)}\right)_{i,j=1,2,\ldots,20}$ be the linear transform of $\left(\widehat{p}_{(i,j)}\right)_{i,j=1,2,\ldots,20}$ by the doubly stochastic matrix D, and let $\mathcal{A}$ be an amino acid pair similarity matrix.

Recall, that the trace of a matrix is the sum of its coefficients in the main diagonal.

Taking pattern from quantum mechanics, we model column pair (γ_1,γ_2) by a positive semi-definite 400×400-matrix the trace of which is equal to 1, a so-called *density matrix*. Regarding the two distributions $\left(\widehat{p}_{(i,j)}\right)_{i,j=1,2,\ldots,20}$ and $\left(\widehat{q}_{(i,j)}\right)_{i,j=1,2,\ldots,20}$ as 400×400-diagonal matrices the main diagonal of which are formed by the probabilities $\widehat{p}_{(i,j)}$ and $\widehat{q}_{(i,j)}$, respectively, we integrate the classical model into the quantum-mechanics-based one.

Generalizing the approach for amino acid used in [24] to amino acid pairs, our density matrices are of the shape

$$\rho\left(\widehat{r},\mathcal{A}\right) := \left(\sqrt{\widehat{r}_{(g,h)}}\,\mathcal{A}_{(g,h),(i,j)}\sqrt{\widehat{r}_{(i,j)}}\right)_{g,h,i,j=1,2,\ldots,20}, \tag{4}$$

where $\widehat{r}_{(i,j)}$ is either $\widehat{p}_{(i,j)}$ or $\widehat{q}_{(i,j)}$ $(i,j = 1,2,\ldots,20)$. Using this denotation, the diagonal density matrices considered in the preceding paragraph are equal to some $\rho\left(\widehat{r},\mathbb{1}\right)$, where $\mathbb{1}$ is the 400×400-identity matrix.

In this study, we regard individual MSA columns only as components of column pairs. In the classical case, where MSA-column pair (γ_1,γ_2) is modeled by an MSA-dependent amino acid pair distribution $\widehat{r}$ (either $\left(\widehat{p}_{(i,j)}\right)_{i,j=1,2,\ldots,20}$ or some derivative), the columns γ_1 and γ_2 are represented by the corresponding marginals $\widehat{r}_1$ and $\widehat{r}_2$ of $\widehat{r}$.

In quantum information science, the counter part of the marginals $\widehat{r}_1$ and $\widehat{r}_2$ of $\widehat{r}$ are the partial traces $\mathrm{tr}_2(\rho)$ and $\mathrm{tr}_1(\rho)$ of ρ. They are 20×20 density matrices defined by

$$(\mathrm{tr}_1(\rho))_{ij} := \sum_{k=1}^{20} \rho_{kkij} \qquad (\mathrm{tr}_2(\rho))_{ij} := \sum_{k=1}^{20} \rho_{ijkk},$$

where $i, j = 1, 2 \ldots, 20$. As opposed to the indices of the marginals, matrix $\mathrm{tr}_1(\rho)$ models column γ_2, whereas matrix $\mathrm{tr}_2(\rho)$ represents column γ_1.

Defining our two new metrics $\mathbb{Q}_{\text{ent}}$ and $\mathbb{Q}_{\text{sep}}$

To begin with, we define the von Neumann entropy $\mathrm{VNE}(\rho)$ of a diagonal density matrix ρ as the Shannon entropy of its main diagonal coefficients regarded as a probability distribution.

The crucial property of a density matrix ρ is that there exists an orthogonal matrix U such that $U\rho U^T$ is a diagonal density matrix, where the diagonal elements are uniquely determined up to their order. Thus we are justified to finally define

$$\mathrm{VNE}(\rho) := \mathrm{VNE}\left(U\rho U^T\right), \tag{5}$$

where U is an orthogonal matrix diagonalizing ρ in a way just mentioned.

In principle, the following holds true. The larger the off-diagonal coefficients of the similarity matrix $\mathcal{A}$, the smaller the von Neumann entropy of the density matrix according to Equation 4 compared with the Shannon entropy of the probability distribution $\widehat{r}_{(i,j)}$ $(i, j = 1, 2, \ldots, 20)$.

In order to compare two density matrices ρ and σ of the same dimension, we make use of the quantum Jensen-Shannon divergence:

$$\mathrm{QJSD}(\rho\|\sigma) := \mathrm{VNE}\left((\rho + \sigma)/2\right) - \left(\mathrm{VNE}(\rho) + \mathrm{VNE}(\sigma)\right)/2. \tag{6}$$

It can be shown that $0 \leq \mathrm{QJSD}(\rho\|\sigma) \leq 1$, where 0 is attained if and only if the two density matrices ρ and σ are equal. As oppose to the case of Equation 1, we have thus avoided a normalization.

We are now in a position to define our new two metrics for a certain column pair of a given MSA. As before, the amino acid pair distribution $\widehat{q}$ is given by $\widehat{p} \cdot D$, where D is the 400×400 doubly stochastic matrix described above, $\widehat{p}$ is the empirical pair distribution of these two columns, and $\mathbb{1}$ is the 400×400-identity matrix.

Then our first metric $\mathbb{Q}_{\text{ent}}$ is defined by

$$\mathbb{Q}_{\text{ent}} := \mathrm{QJSD}\left(\rho\left(\widehat{q}, \mathcal{A}_{\text{ent}}\right) \| \rho\left(\widehat{p}, \mathbb{1}\right)\right) \tag{7}$$

(see Equation 4). This metric measures the difference between a density matrix combining rescaled amino acid pair similarity with dissimilar compensatory mutations and the empirical amino acid pair distribution. The index "*ent*" indicates that here we make use of quantum entanglement, which in turn is a major resource of quantum information science. (Entangled 400×400-density matrices are those that cannot be represented as a convex combination of Kronecker products of 20×20-density matrices. Note, that the Kronecker product of density matrices is the analog of the classical product of probability distributions).

Our second new metric $\mathbb{Q}_{\text{sep}}$ is given by

$$\mathbb{Q}_{\text{sep}} := \mathrm{QJSD}\left(\mathrm{tr}_1\left(\rho\left(\widehat{p}, \mathcal{A}_{\text{sep}}\right)\right) \| \mathrm{tr}_2\left(\rho\left(\widehat{p}, \mathcal{A}_{\text{sep}}\right)\right)\right). \tag{8}$$

The density operator $\rho\left(\widehat{p}, \mathcal{A}_{\text{sep}}\right)$ is entangled. However, before finally calculating the metric, we separate the columns of the pair by applying the two partial trace operators.

Using the example of the human EGFR protein (PDB-Entry 2J6M), Figure 6 illustrates that the method we developed in [5] to determine significant column pairs is well-applicable for both $\mathbb{Q}_{\text{ent}}$ and $\mathbb{Q}_{\text{sep}}$. The results presented in this work prove that $\mathbb{Q}_{\text{ent}}$ as well as $\mathbb{Q}_{\text{sep}}$ are powerful co-evolutionary column pair metrics.

Discussion

Grosse *et al.* observed in [64] that the Jensen-Shannon divergence (JSD) can be interpreted as mutual information between two (or more) random sources in a special setting particularly appropriate to discriminate between these sources. This is what we need when it comes to predicting important protein sites in an MSA-based approach. It might explain the findings of Capra and Singh [22] on the predictive power of JSD. These two articles encouraged us to utilize quantum Jensen-Shannon divergence (QJSD) in this study. As a side effect, a normalization is not necessary, since quantum Jensen-Shannon divergence, like its classical counterpart, ranges over the real interval $[0, 1]$.

Several studies have confirmed the fact that detecting coupled MSA-columns is extremely useful in the prediction of important protein sites (see e.g. [4-6,10-13,65-70]). When using information-theoretic metrics, there is no doubt that it is reasonable to incorporate amino acid pair dissimilarity as well as amino acid similarity in a consistent way such that similarity decreases entropy, whereas dissimilarity increases it. This kind of consistency is important, since entropy is the fundamental building block for most of those metrics. In particular, the Jensen-Shannon divergence between two probability mass functions (pmfs) p and q equals $\mathbb{H}(1/2(p + q)) - 1/2(\mathbb{H}(p) + \mathbb{H}(q))$.

In [5] an amino acid pair dissimilarity model for compensatory mutations is presented. A doubly stochastic matrix transforms the empirical amino acid pair distribution of a column pair.

Rescaled pair similarity of BLOSUM62-similar pairs is to capture an aspect of coupled MSA column pairs orthogonal to the phenomenon of dissimilar compensatory mutations. It models the amino acid pair transition preferences within those column pairs on the average. As suggested by Caffrey *et al.* [23] as well as Johansson *et al.* [24], it is promising to incorporate them within the framework of quantum information theory. Therein, density matrices replace pmfs. The counterpart of the entropy of a pmf is the von Neumann entropy (VNE) of a density matrix (see Equation 5). QJSD corresponds then exactly to JSD (see Equation 6).

The challenge was to complement the model presented in [5] by additionally incorporating amino acid pair similarity in a way that the two effects interfere but do not interact. We model an MSA column pair by means of a 400×400-density matrix, rather than amino acid pair distributions. This provides us with the opportunity to utilize the notion of entanglement, which in turn is a major resource of quantum information. In our model, partial traces play the role of the marginals in the classical case. Pair similarity is reflected by means of positive definite pair similarity matrices (see Equation 3), where positive definiteness, which is a key property of density matrices, can only be ensured by using transitivity of similarity. Since there is no transitivity of dissimilarity, we kept dissimilarity apart from that similarity matrix. Instead, we carried over the CMF dissimilarity model of [5]. Similarity matrix and transformed amino acid pair distribution are joined together by means of Equation 4 in the final step of our density matrix design. That way we minimize the interaction between the two effects of dissimilarity and similarity.

In order to eliminate the noise and to define an MSA-dependent threshold for significant column pairs, we followed the line of [5]. The model presented there seems to be of universal applicability. The same is true for the connectivity degree model introduced in [12] and further developed in [5]. Combining them results in a reliable and robust method to determine significant residues.

The results we present in this study show that the vast majority of QCMF-significant residue sites are closely related to functionality and structural stability of both human EGFR and GCK proteins. 10 significant residue sites in EGFR and 19 significant sites in GCK are established as functionally important since they are directly located at or close to catalytic sites, allosteric sites and binding sites which are crucial for maintaining protein functions and for understanding the underlying molecular mechanism (see Figures 1,2,3,4). Additionally, 2 significant sites in EGFR and 8 significant sites in GCK (three of them are also in contact with allosteric sites in GCK) are related to disease associated nsSNP regions of both proteins. As has been noted in [5], most disease-causing mutations at these positions in corresponding sequences destroy structural features of proteins, thus affecting protein stability and often results in loss of protein function.

Although the importance of almost all QCMF-significant sites are verified through essential sites of both human proteins, there are still eight and two unconfirmed significant sites in EGFR and GCK proteins, respectively, which do not fall into essential sites. It is interesting to note that some of these unconfirmed sites are also referred as significant by CMF [5]. We therefore believe that most of these unconfirmed sites identified by our present method may have an importance for the function and structural stability of both proteins notwithstanding the absence of previous experimental data. A further comparison reveals that the overlaps between the results of the QCMF method and the CMF method are quite low, indicating that both methods detect considerably different sets of residue sites as functionally and structurally important. The comparison results clearly show that considering similar and dissimilar amino acid signals simultaneously, our present method is more sensible to catalytic, allosteric and binding sites, while only focusing on dissimilar signals the previous method deals successfully with nsSNP positions in proteins.

The final comparison between QCMF and CMF on EGFR and GCK proteins is made by inspecting several connectivity degree cut-offs. We initially set it to the 90-th percentile at which CMF reaches its maximal MCC value. Going through all possible n-th percentiles for $n = 80, 81, \ldots, 99$, QCMF reaches its maximal MCC value of 0.231 if $n = 88$. What we got can be summarized as follows. On the one hand QCMF shows a better performance than CMF in identifying important residue sites. On the other hand QCMF complements CMF. This is because of the fact that the method of QCMF is more information rich than that of CMF. QCMF simultaneously uses similar and dissimilar amino acid pair signals, whereas CMF's method focuses only on amino acid pair dissimilarity.

To confirm the educated guess that QCMF complements conventional methods both from information theory and statistics, we applied QCMF, CMF [5], MIp [6] and PSICOV [18] to the 153 MSAs described at the beginning of the Results section. In sum, each of these methods detects different residue pairs as important, where the pairwise overlap is bounded from above by 10%. The reason for that is that the four methods model different aspects of amino acid pair co-evolution. Consequently, they carry distinct information.

To further improve the specificity of QCMF it is promising to combine its quantum-information-theory-based framework with the direct pair distribution derived in DCA (see e.g. [15] or [16]).

Conclusions

In this work, we report a new method, QCMF, applying principles of quantum information theory. In contrast to the previous method CMF which focused on dissimilar amino acid signals, QCMF simultaneously models similar and dissimilar amino acid pair signals in the detection of functionally or structurally important sites. QCMF includes two metrics based on quantum Jensen-Shannon divergence. While the first metric measures compensatory mutations between pairs of columns, the second metric considers the sequence conservation of columns. Results show that QCMF reaches an improved performance in identifying important sites from MSAs and it predicts a quite different set of residue sites as functionally and structurally important (in comparison to the previous method). Further, results indicate that the residue sites found by QCMF are more sensible to catalytic sites, allosteric sites and binding sites than those found by the previous method. On the top of that, a pairwise comparison with existing methods shows that QCMF is complementary to them when it comes to predicting co-evolving residue site pairs.

Additional files

Additional file 1: Pdb entries. Pdb entries of 153 test proteins.

Additional file 2: EGFR significant sites. QCMF-significant residue sites of the human epidermal growth factor receptor (EGFR) protein.

Additional file 3: GCK significant sites. QCMF-significant residue sites of the human glucokinase (GCK) protein.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SW developed the model underlying QCMF. EW adjusted the model together with SW and interpreted the results. MG developed the model together with SW, designed and implemented the tool together with SH, SJJ and GD, and interpreted the results together with EW and CM. SH designed and carried out the CUDA programming. GD and SJJ supported SW in developing the model and MG in designing and implementing the tool. CM interpreted the results together with EW and MG. All authors read and approved the manuscript.

Acknowledgements

We thank Rainer Merkl from Regensburg for his helpful advice and insights at early stages of this project and for comments on the methods.
Mehmet Gültas acknowledges financial support by the Deutsche Forschungsgemeinschaft (DFG: WA 766/7-1).

Received: 10 June 2013 Accepted: 26 March 2014
Published: 3 April 2014

References

1. Gloor GB, Martin LC, Wahl LM, Dunn SD: **Mutual information in protein multiple sequence alignments reveals two classes of Coevolving positions.** *Biochemistry* 2005, **44**(19):7156–7165. [http://pubs.acs.org/doi/abs/10.1021/bi050293e]. [PMID: 15882054]
2. Wilson K, Walker J: *Principles and Techniques of Biochemistry and Molecular Biology*. 7th edition. New York: Cambridge University Press; 2010.
3. Altschuh D, Lesk AM, Bloomer AC, Klug A: **Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus.** *J Mol Biol* 1987, **193**(4):693–707.
4. Martin LC, Gloor GB, Dunn SD, Wahl LM: **Using information theory to search for co-evolving residues in proteins.** *Bioinformatics* 2005, **21**(22):4116–4124.
5. Gültas M, Haubrock M, Tüysüz N, Waack S: **Coupled mutation finder: a new entropy-based method quantifying phylogenetic noise for the detection of compensatory mutations.** *BMC Bioinformatics* 2012, **13:**225. [http://www.biomedcentral.com/1471-2105/13/225]
6. Dunn SD, Wahl LM, Gloor GB: **Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction.** *Bioinformatics* 2008, **24**(3):333–340.
7. Chakrabarti S, Panchenko AR: **Structural and functional roles of Coevolved sites in proteins.** *PLoS ONE* 2010, **5:**e8591. [http://dx.doi.org/10.1371%2Fjournal.pone.0008591].
8. Sandler I, Abu-Qarn M, Aharoni A: **Protein co-evolution: how do we combine bioinformatics and experimental approaches?** *Mol BioSyst* 2013, **9:**175–181. [http://dx.doi.org/10.1039/C2MB25317H]
9. DePristo MA, Weinreich DM, Hartl DL: **Missense meanderings in sequence space: a biophysical view of protein evolution.** *Nat Rev Genet Nat Publishing Group* 2005, **6**(9):678–687. [http://dx.doi.org/10.1038/nrg1672]
10. Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW: **Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis.** *Mol Biol Evol* 2000, **17:**164.
11. Tillier ER, Lui TW: **Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments.** *Bioinformatics* 2003, **19**(6):750–755. [http://bioinformatics.oxfordjournals.org/content/19/6/750.abstract]
12. Merkl R, Zwick M: **H2r: Identification of evolutionary important residues by means of an entropy based analysis of multiple sequence alignments.** *BMC Bioinformatics* 2008, **9:**151. [http://www.biomedcentral.com/1471-2105/9/151]
13. Gao H, Dou Y, Yang J, Wang J: **New methods to measure residues coevolution in proteins.** *BMC Bioinformatics* 2011, **12:**206. [http://www.biomedcentral.com/1471-2105/12/206].
14. de Juan D, Pazos F, Valencia A: **Emerging methods in protein co-evolution.** *Nat Rev Genet* 2013, **14:**249–261.
15. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M: **Direct-coupling analysis of residue coevolution captures native contacts across many protein families.** *Proc Nat Acad Sci* 2011, **108**(49):E1293–E1301. [http://www.pnas.org/content/108/49/E1293.abstract]
16. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C: **Protein 3D structure computed from evolutionary sequence variation.** *PLoS ONE* 2011, **6**(12):e28766.
17. Cheng RR, Morcos F, Levine H, Onuchic JN: **Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information.** *Proc Nat Acad Sci* 2014. [http://www.pnas.org/content/early/2014/01/17/1323734111.abstract]
18. Jones DT, Buchan DWA, Cozzetto D, Pontil M: **PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments.** *Bioinformatics* 2012, **28**(2):184–190.
19. Kamisetty H, Ovchinnikov S, Baker D: **Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era.** *Proc Nat Acad Sci* 2013, **110**(39):15674–15679. [http://www.pnas.org/content/110/39/15674.abstract]
20. Williamson R: **Information theory analysis of the relationship between primary sequence structure and ligand recognition among a class of facilitated transporters.** *J Theor Biol* 1995, **174:**179–188.
21. Mirny J, Shakhnovich E: **Universally conserved position in protein folds: reading evolutionary signals about stability, folding, kinetics, and function.** *J Mol Biol* 1999, **291:**10930–10935.
22. Capra JA, Singh M: **Predicting functionally important residues from sequence conservation.** *Bioinformatics* 2007, **23**(15):1875–1882.
23. Caffrey DR, Somaroo S, Hughes JD, Mintseris J: **Huang ES: Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?** *Protein Sci* 2004, **13:**190–2002.
24. Johansson F, Toh H: **Relative von Neumann entropy for evaluating amino acid conservation.** *J Bioinformatics Comput Biol* 2010,

08(05):809–823. [http://www.worldscientific.com/doi/abs/10.1142/S021972001000494X]

25. **NVIDIA CUDA Zone.** [http://www.nvidia.com/object/cuda_home_new.html]
26. Liu Y, Wirawan A, Schmidt B: **CUDASW++ 3.0: accelerating Smith-Waterman protein database search by coupling CPU and GPU SIMD instructions.** *BMC Bioinformatics* 2013, **14:**117. [http://www.biomedcentral.com/1471-2105/14/117]
27. Manavski S, Valle G: **CUDA compatible GPU cards as efficient hardware accelerators for Smith-Waterman sequence alignment.** *BMC Bioinformatics* 2008, **9**(Suppl 2):S10.
28. Lui Y, Maskell D, Schmidt B: **CUDASW++: optimizing Smith-Waterman sequence database searches for CUDA-enabled graphics processing units.** *BMC Res Notes* 2009, **2:**73.
29. Wirawan A, Kwoh C, Hieu N, Schmidt B: **CBESW: Sequence alignment on the Playstation 3.** *BMC Bioinformatics* 2008, **9:**377.
30. Ufimtsev I, Martinez T: **Graphical processing units for quantum chemistry.** *Comput Sci Eng* 2008, **10**(6):26–34.
31. Stone J, Hardy D, Ufimtsev I, Schulten K: **GPU-accelerated molecular modeling coming of age.** *J Mol Graph Model* 2010, **29**(2):116–125.
32. Michalakes J, Vachharajani M: **GPU acceleration of numerical weather prediction.** *Parallel Process Lett* 2008, **18**(4):531–548.
33. Dixit A, Yi L, Gowthaman R, Torkamani A, Schork NJ, Verkhivker GM: **Sequence and structure signatures of cancer mutation Hotspots in protein Kinases.** *PLoS ONE* 2009, **4**(10):e7485. [http://dx.doi.org/10.1371%2Fjournal.pone.0007485]
34. Yun CH, Boggon TJ, Li Y, Woo MS, Greulich H, Meyerson M, Eck MJ: **Structures of lung cancer-derived EGFR mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity.** *Cancer Cell* 2007, **11**(3):217–227. [http://www.sciencedirect.com/science/article/pii/S1535610807000281]
35. Zhang H, Berezov A, Wang Q, Zhang G, Drebin J, Murali R, Greene MI: **ErbB receptors: from oncogenes to targeted cancer therapies.** *J Clin Invest* 2007, **117**(8):2051–2058. [http://www.jci.org/articles/view/32278]
36. Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, Harris PL, Haserlat SM, Supko JG, Haluska FG, Louis DN, Christiani DC, Settleman J, Haber DA: **Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to Gefitinib.** *N Engl J Med* 2004, **350**(21):2129–2139. [http://www.nejm.org/doi/full/10.1056/NEJMoa040938]
37. Balius TE, Rizzo RC: **Quantitative prediction of fold resistance for inhibitors of EGFR.** *Biochemistry* 2009, **48**(35):8435–8448. [http://pubs.acs.org/doi/abs/10.1021/bi900729a]. [PMID: 19627157].
38. Tinto N, Zagari A, Capuano M, De Simone A, Capobianco V, Daniele G, Giugliano M, Spadaro R, Franzese A, Sacchetti L: **Glucokinase gene mutations: structural and genotype-phenotype analyses in MODY children from South Italy.** *PLoS ONE* 2008, **3**(4):e1870. [http://dx.plos.org/10.1371%2Fjournal.pone.0001870]
39. Capuano M, Garcia-Herrero CM, Tinto N, Carluccio C, Capobianco V, Coto I, Cola A, Iafusco D, Franzese A, Zagari A, Navas MA, Sacchetti L: **Glucokinase (GCK) mutations and their characterization in MODY2 children of Southern Italy.** *PLoS ONE* 2012, **7**(6):e38906. [http://dx.doi.org/10.1371%2Fjournal.pone.0007485]
40. Garcia-Herrero CM, Rubio-Cabezas O, Azriel S, Gutierrez-Nogues A, Aragones A, Vincent O, Campos-Barros A, Argente J, Navas MA: **Functional characterization of MODY2 mutations highlights the importance of the fine-tuning of glucokinase and its role in glucose sensing.** *PLoS ONE* 2012, **7:**e30518. [http://dx.doi.org/10.1371%2Fjournal.pone.0038906]
41. Kamata K, Mitsuya M, Nishimura T, ichi Eiki J, Nagata Y: **Structural basis for allosteric regulation of the monomeric allosteric enzyme human glucokinase.** *Structure* 2004, **12**(3):429–438. [http://www.sciencedirect.com/science/article/pii/S0969212604000474]
42. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic Acids Res* 2005, **33**(suppl 1):D514–D517. [http://nar.oxfordjournals.org/content/33/suppl_1/D514.abstract]
43. Reichert J, Sühnel J: **The IMB Jena image library of biological macromolecules: 2002 update.** *Nucleic Acids Res* 2002, **30:**253–254. [http://nar.oxfordjournals.org/content/30/1/253.abstract]
44. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Gordon L, Hendrix M, Hourlier T, Johnson N, Kähäri A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Larsson P, Longden I, McLaren W, Overduin B, Pritchard B, Riat HS, Rios D, Ritchie GRS, Ruffier M, Schuster M, et al.: **Ensembl 2011.** *Nucleic Acids Res* 2011, **39**(suppl 1):D800–D806. [http://nar.oxfordjournals.org/content/39/suppl_1/D800.abstract]
45. Keskin O, Tsai CJ, Wolfson H, Nussinov R: **A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications.** *Protein Sci* 2004, **13**(4):1043–1055. [http://dx.doi.org/10.1110/ps.03484604]
46. Janda JO, Busch M, Kuck F, Porfenenko M, Merkl R: **CLIPS-1D: analysis of multiple sequence alignments to deduce for residue-positions a role in catalysis, ligand-binding, or protein structure.** *BMC Bioinformatics* 2012, **13:**55. [http://www.biomedcentral.com/1471-2105/13/55]
47. Sander C, Schneider R: **Database of homology derived protein structures and the structural meaning of sequence alignment.** *Proteins* 1991, **9**(1):56–69.
48. Porter CT, Bartlett GJ, Thornton JM: **The Catalytic site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data.** *Nucleic Acids Res* 2004, **32**(suppl 1):D129–D133. [http://nar.oxfordjournals.org/content/32/suppl_1/D129.abstract]
49. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res* 2001, **29:**308–311. [http://nar.oxfordjournals.org/content/29/1/308.abstract]
50. Dixit A, Yi L, Gowthaman R, Torkamani A, Schork NJ, Verkhivker GM: **Sequence and structure signatures of cancer mutation hotspots in protein kinases.** *PLoS ONE* 2009, **4**(10):e7485. [http://dx.doi.org/10.1371%2Fjournal.pone.0030518].
51. Valentinova L, Beer NL, Stanik J, Tribble ND, van de Bunt M, Huckova M, Barrett A, Klimes I, Gasperikova D, Gloyn AL: **Identification and functional Characterisation of novel glucokinase mutations causing maturity-onset diabetes of the young in Slovakia.** *PLoS ONE* 2012, **7**(4):e34541. [http://dx.doi.org/10.1371%2Fjournal.pone.0007485]
52. Bosma W, Cannon J, Playoust C: **The Magma algebra system. I. The user language.** *J Symbolic Comput* 1997, **24**(3–4):235–265. [http://dx.doi.org/10.1006/jsco.1996.0125]. [Computational algebra and number theory (London, 1993)]
53. Anderson E, Bai Z, Bischof C, Blackford S, Demmel J, Dongarra J, Du Croz J, Greenbaum A, Hammarling S, McKenney A, Sorensen D: *LAPACK Users' Guide*. 3rd edition. Philadelphia, PA: Society for Industrial and Applied Mathematics; 1999.
54. **An updated set of basic linear algebra subprograms (BLAS).** *ACM Trans Math Softw* 2002, **28**(2):135–151. [http://doi.acm.org/10.1145/567806.567807]
55. Goto K, Geijn RAvd: **Anatomy of high-performance matrix multiplication.** *ACM Trans Math Softw* 2008, **34**(3):12:1–12:25. [http://doi.acm.org/10.1145/1356052.1356053]
56. Lifflander J, Evans GC, Arya A, Kale L: **Dynamic Scheduling for Work Agglomeration on Heterogeneous Clusters.** In *Parallel and Distributed Processing Symposium Workshops PhD Forum (IPDPSW), 2012 IEEE 26th International.* 2012:2404–2413. doi:10.1109/IPDPSW.2012.297.
57. Vedral V: *Introduction to Quantum Information Science (Oxford Graduate Texts)*. New York: Oxford University Press Inc.; 2006.
58. Nielsen MA, Chuang IL: *Quantum Computation and Quantum Information*: Cambridge University Press; 2000.
59. Storey JD, Tibshirani R: **Statistical significance for genomewide studies.** *Proc Acad Sci* 2003, **100:**9440–9445.
60. Walsh B: *Multiple comparisons: Bonferroni corrections and false discovery rates.* Lecture Notes EEB 581, Department of Ecology and Evolutionary Biology, University of Arizona, 2004.
61. Wang G, Jr RLD: **PISCES: recent improvements to a PDB sequence culling server.** *Nucleic Acids Res* 2005, **33**(Web-Server-Issue):94–98.
62. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks.** *Proc Nat Acad Sci* 1992, **89**(22):10915–10919. [http://www.pnas.org/content/89/22/10915.abstract]
63. Cappellini V, Sommer HJ, Bruzda W, Zyczkowski K: **Random bistochastic matrices.** *J Phys A: Math Theor* 2009, **42:**23.
64. Grosse I, Bernaola-Galván P, Carpena P, Román-Roldán R, Oliver J, Stanley HE: **Analysis of symbolic sequences using the Jensen-Shannon**

divergence. *Phys Rev E* 2002, **65:**041905. [http://link.aps.org/doi/10.1103/PhysRevE.65.041905]

65. Göbel U, Sander C, Schneider R, Valencia A: **Correlated mutations and residue contacts in proteins.** *Proteins-Struct Funct Genet* 1994, **18**(4):309–317.
66. Neher E: **How frequent are correlated changes in families of protein sequences?** *Proc Nat Acad Sci* 1994, **91:**98–102. [http://www.pnas.org/content/91/1/98.abstract]
67. Pollock DD, Taylor WR: **Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution.** *Protein Eng* 1997, **10**(6):647–657. [http://peds.oxfordjournals.org/content/10/6/647.abstract]
68. Lockless SW, Ranganathan R: **Evolutionarily conserved pathways of energetic connectivity in protein families.** *Science* 1999, **286**(5438):295–299. [http://www.sciencemag.org/content/286/5438/295.abstract]
69. Dekker JP, Fodor A, Aldrich RW, Yellen G: **A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments.** *Bioinformatics* 2004, **20**(10):1565–1572.
70. Codoner FM, Fares M: **Why should we care about molecular coevolution?** *Evol Bioinform* 2008, **4:**29–38.

doi:10.1186/1471-2105-15-96
Cite this article as: Gültas *et al.*: Quantum coupled mutation finder: predicting functionally or structurally important sites in proteins using quantum Jensen-Shannon divergence and CUDA programming. *BMC Bioinformatics* 2014 **15**:96.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

9.3 Appendix C: PDB enteries of training protein MSAs

PDB Enteries of Training MSAs

Table 9.1: Pdb enteries of training MSAs.

2V4EA	1CKTA	2EOFA	1TO9A	3BQ7A	3EWSA	2VMQA	1CO8A
1ZYEA	1DUNA	2HW5A	2GBRA	4GTUA	1GGOA	1KY9A	1YY9A
2ET1A	2BHJA	1CL7H	1HV9A	1NMLA	1PZ1A	3C05A	2NQMA
2R0OA	2WFQA	3CYFA	1UN3A	2OTUB	1INDL	3ESYA	1BMTA
2Q3OA	2DJQA	1MCIA	1PTQA	1Y65A	2J0IA	1OV9A	2JJYA
1YKKB	2VXLA	1CQGA	1UDCA	2A0WA	1FYTB	1BL3A	1R36A
2P79A	2PT6A	1AQKL	1IRBA	3DGZA	2IFYA	1KENL	1I88A
1RUZH	3D5TA	1BUHB	3B84A	1WHIA	2ISSA	2CUBA	2QQNH
2DWBA	2EAOA	1S9PA	2EMXA	2FTSA	2FEMA	1W1B1	1NPKA
1QKWA	1XQCA	2B1UA	3F9PA	1ZUTA	2YRHA	2GOYA	2FFYA
1TYBE	1R9NA	3BT8A	2V8QE	1SJJA	2IY0B	2ZJBA	1UMNA
1PTZA	1IOLA	2R4RH	2CPUA	2EOHA	1T5BA	5BJ3A	1D6MA
1VHPA	1IFVA	1QORA	3FF1A	2D5YA	1MXSA	2QAPA	2GIXA
5REQA	2FGQX	2IPMA	1O4RA	3BEAA	1YIUA	2RHEA	1T5XA
3CSWA	2Z2YA	1WYOA	2E9UA	1P3OB	1C0FA	1CYWA	2HQKA
1PRGA	2OWGA	1AW2A	1Z1DB	2V3XA	5NLLA	2F3LA	1CUKA
2ETRA	2GIVA	1VLZB	2VFHA	2BA0E	1PKUA	1OUIA	1AW8E
2HANA	1SSMA	1NEKC	1PPAA	1AYM4	3HXPA	2K5SA	3F0UX
2RFJA	1UGAA	1OMWG	1P4SA	3D01A	2ADAA	1H25C	2VSUC
1T2QL	1A5EA	2Z83A	1P4WA	2ITWA	1QTWA	1U8SA	1DJEA
3GL1A	2AYMA	3D8XA	1FHJA	3GKEA	2VPJA	1R8ME	2BIOA
2JB9A	8CA2A	2CHGA	2P71A	1PMTA	1J2EA	1VP7A	3CZMA
1IAGA	2OGHA	2CH4W	2VMKA	1TABI	1B5ZA	2R69H	1CS8A
1B8IB	1B50A	1JDIA	2K1ZA	1GVLA	2VL9A	2UW1B	1I92A
1Y62A	1F2ZA	1HVFA	1CCUA	2GZOA	1AG9A	2BBLA	1NFIE
1MEXL	1GFGA	1TZPA	3BW3A	1BBIA	1ZS6A	1A5ZA	2QEJC
1UHEB	1BAFH	1QQJA	3B9JB	2ZM1A	1L3WA	1GF6A	2Q4PA
2VFBA	1EJIA	3FLCO	1WSZA	2F91A	2FSVC	1IQ3A	1MJ8L
1OJXA	2JFPA	1OVAA	1FAIL	1A2XA	1PKRA	3CLUD	3EDYA
1M93C	1QKFA	2DNMA	1A6TA	1YMGA	1T3IA	1ZL9A	3BQYA
1HBRB	1K8XB	2ED3A	1B8GA	1YAGA	1G9XA	2RG3A	1YACA
2P77A	2RIVB	2PBIB	1HVUA	1QF5A	2V36A	1F4DA	2ODBB
2PAEA	2OTGC	2IPCA	1Q3IA	1HW7A	1BPVA	3G3DA	1QOGA
2A9HA	1Y7TA	1LBBA	1NG9A	3D5FA	3CWBA	2EW8A	1DTGA
1CL7L	1E20A	1GN6A	2P6BB	1USMA	2MASA	1ECFA	2QL9B
2QQUA	3D21A	3HLUA	3HYGA	2QOBA	1F54A	1CTAA	2H79A

Continued on next page

Table 9.1 Pdb enteries of training MSAs – continued from previous page

2W2NE	1NQBA	1ZH9A	2E2RA	1HWUA	2HZIA	1FKOA	2BW4A
1BIZA	1RVFL	2HY0A	1M40A	1MF8B	1Z95A	2ZD2A	2IUFA
2GRRA	3CTZA	1O3SA	1FSKB	2JO6A	1LDLA	3DOKA	1MP5A
2O36A	1MY5A	1J19A	1M7NA	1C5CL	1VH7A	1JI8A	1J8YF
1GQRA	2J58A	1XHGA	2H88A	1RT3B	1KZ9A	2HZLA	2UZCA
1AN2A	1CGPA	1ND9A	2VS4A	1IRDA	2V55B	2F7XE	3MATA
1ATYA	1X7ZA	1BCZA	2W69A	2ER8A	1U6ZA	2W72A	2NT8A
1O2FB	3HMMA	1QL9A	1FSAA	1TMF3	2QV1C	2W1VA	1ATNA
3CBJA	1QJOA	2P91A	1XF7A	1GYPA	1GROA	1AD1A	2P4VA
2ON5A	2ZPBA	1QGKB	1BWWA	2AQMA	2BM1A	2JNPA	1W0ZU
2QPUA	1X7AL	3PBHA	2E09A	1Y59T	1RHSA	1IH0A	1YJRA
1FI3A	1YFDA	2B3HA	1W0TA	3FYSA	1PKLA	3F6UL	1YDBA
3HHHA	3EPF3	1IQQA	2CCGA	1V2XA	1N0EA	2VH9A	2K79A
2VGLS	3BV6A	1ISNA	2QRZA	3BZXB	2KBOA	1AW8A	2GOLA
2YQCA	1Q6OA	2JLJA	2HLPA	2BDQA	3C5QA	2G38A	1P4CA
3BDMD	2APSA	1MZBA	2GJ3A	3E22A	1YI8B	1XYKA	2FVCA
2QQNL	2OPOA	3GKAA	1EJRC	1TWYA	1A6VH	1WCDJ	3B6BA
1UHFA	2EOLA	1JPTL	1Q5XA	3E7CA	2QIEA	2ESNA	2ETJA
1YBZA	1OSPO	3I8VA	1UZ6E	1ZBRA	3GFTA	2H1VA	1E3IA
2Z26A	2Q3YA	1P7KA	2NYBA	2O5KA	3AIGA	1I7HA	2UW2A
1Z8YJ	1HL4A	1SI2A	1IN0A	1YX5B	2VV6A	1W74A	1NW2A
1N0XH	1TMOA	3D3PA	2UUMX	1UGKA	1OXSC	3I4SA	3FWYA
1DDKA	2GH9A	2QSQA	2JUZA	1Z3SA	3HNNA	3HJBA	1Z7Z1
2GI4A	1MHWA	2AXYA	2FHTA	2SNWA	1ES0A	2GQQA	2HDCA
1K9UA	2Z39A	2G6AA	2BH8A	1JZ8A	1ATGA	2RD1A	1TA8A
3DDKA	1W9JA	1SNOA	2QV1A	1P5JA	1TZAA	3DAJA	1GWTA
1SAUA	1WG2A	2GV0A	3C3TA	1J95A	2GMLA	1LH0A	2DJ6A
2IUEA	1KTBA	1S1TA	3GLBA	1GQ2A	1RDAA	1S3IA	1L8BA
1T13A	1YCLA	1VJAU	2ASCA	2GQRA	2OMZA	2IV1A	2HFGL
2ZURX	1YABA	3G3NA	3DLPX	1I5EA	1SVXB	2FL5A	1KZMA
1TX2A	2IO0B	3HHFA	3B3DA	2VEDA	2P2BA	1C1EH	2DP3A
1NTYA	2BZRA	2BIGA	3DEXA	3ERHA	1W8BL	1SYIA	1C2OA
3HP7A	1UFVA	1D3AA	1SW1A	1MODA	3BNDA	2F9OA	1A6FA
1FRBA	1DEJA	3DZYD	2RGDA	1BVKB	1ZD9A	3BC7A	3D9AL
1Z3IX	1E1DA	1G7JA	1E9EA	1DB3A	1SVIA	1R64A	1Z07A
1EQ2A	5CA2A	1J6YA	1JNRB	2JOTA	2OAXA	2HEDA	2CU7A
1CGJI	1PCAA	3GL6A	1UN2A	2EIGA	2ZHGA	1Y791	1ZG4A
1Z7BA	1PE0A	1FPSA	1NBMD	2CBYA	1SJ1A	1BRMA	2NRLA
4PBGA	1YKTB	1Q52A	1AG8A	1T11A	1M8VA	2QW7A	2NZDC
1QACA	2PCJA	2EP0A	3DQGA	2JEXA	2EPVA	3CX5J	2C5LA
1DGFA	1C6RA	2QCUA	2RGXA	2I24N	1NEEA	1FNDA	1Y4RB
2B8TA	2VIIA	2Z6QA	2QMXA	2IU8A	1F99B	1B0TA	2Z5II
1EAEA	1KTRL	2TECI	3FS2A	2ID4A	3IY7A	1H8BA	2PIAA
1M5AB	2BMOA	1QDEA	3FKWA	2ZA7A	1A0JA	1YEZA	1UEKA
1RVXB	1Q7JA	1VS3A	1K1CA	1JO8A	2AUVA	1JBBA	1O9PA
2ZOKA	3HOAA	1MM3A	3DRSA	2ENCA	1KDCA	3EC5A	1CRKA
1IS8A	3HILA	2OF2A	1N1JA	3CX5C	1PY5A	2NWLA	1XD8A

Continued on next page

Table 9.1 Pdb enteries of training MSAs – continued from previous page

1KAWA	1ULQA	1KEAA	2QO4A	3GKNA	2Q66A	2IF2A	1QHHA
1E1CA	2MHAA	1CMZA	1BPTA	2VD2A	1NCAH	3IL8A	1TSZA
2DQ3A	2BQHA	1TM1I	1FPOA	2F0IA	1BIYA	3C1YA	2PTQA
1HVEA	2DFBA	1A40A	2P6KA	1KTRH	1H8SA	1IY9A	1CMIA
2QBUA	1D4M2	1MPSM	2HLEA	1GP6A	1GP7A	3EDTB	2QDYA
1SROA	1PSKL	2V6OA	3BQVA	2AWGA	2EOOA	2HHEB	1BJGA
2AA2A	2ECNA	1OUBA	3EANA	2EXYA	3DCTA	1TOZA	1BZZA
1J3DA	2JIDA	2P2CA	2ECUA	3H02A	3DJNB	1AJJA	1FLJA
2FY9A	2O3FA	2YSSB	2D3EA	2OZPA	3ERZA	1CICA	2AF5A
1QTMA	1NEKD	1UN8A	1BHYA	2DFEA	1CNKA	1UVQA	1BFTA
2JPBA	1AQAA	1CHVS	2BURB	3HMUA	2JXBA	3EG7A	1FTSA
2NQAA	1DOTA	2J41A	1CJCA	1U07A	1ZGKA	1W4HA	1K1AA
1B4SA	1BQ0A	1JYMA	2IT6A	2IBAA	2J8MA	2P9YA	3COVA
2EK7A	1QN2A	2PWZA	1H2VZ	1UUEA	1YIVA	2BEMA	3G9YA
1OSEA	1KNQA	1YGSA	2ARJB	1VZYA	1HW3A	3GNLA	2QJXA
2WIOA	1QXOA	1BW9A	2K6WA	2JXMB	1L4ZA	1LO0L	3EO0A
2ZREA	2F1EA	1VIOA	1EIAA	1CS3A	1ATPE	2DVWB	2GTNA
2AZHA	1JNLL	2Z8WC	1H1HA	2ESBA	1QZ8A	1JWIB	1PPJD
1ZT7A	1GTIA	2D0SA	1MEYC	1I3DA	1X3SA	1YNLH	1OGWA
1KHUD	1Q6UA	2H8FA	1FDDA	1DVAL	1HI6A	1AHPA	1OAKL
1D5BB	1WUIS	1BC8C	3B8KA	1PJZA	2A7VA	1QFKL	1MY7A
2H3LA	1Y69K	2PILA	2BJYA	1HK0X	1JO0A	1NBUA	2P5MA
2IFFY	1RDBA	1NDUA	1CPJA	1GX3A	3HMFA	2PPTA	1IV0A
1R3RA	3E8JA	1DI2A	3DGVA	2YWLA	1Y6WA	1ZC3A	1QGHA
2DGTA	1DRBA	1WA5A	1FSEA	2RA3A	2HHOB	2SPOA	1UD7A
1W54A	1FFTB	1YZGA	1AE1A	1BXWA	1K82A	3BJFA	3BRPA
1AD9B	1BLUA	1QCNA	1MQOA	1G79A	1IYUA	3G7LA	2J65A
3C90X	2G9MA	1OPCA	1NGGA	1FY7A	3D5KA	1K5VA	1G4UR
2CH7A	2A3VA	2DJ7A	1D7AA	2RIPA	2NU8A	1MU2A	2AABL
2DDYA	1ZV8A	1ZOVA	1DXQA	9PAIA	2BS3A	1D9ZA	1DR8A
1CGHA	1OPHA	1U9LA	2EY2A	2BV5A	3CB0A	3EYMA	2NLLB
1C6SA	2QWTA	1FORL	1ZBA1	2DLUA	1DKKA	1OI7A	3GPDR
2DBYA	1BLLE	2FQ3A	2V4JC	2PPYA	2OXBA	1AVHA	1EG5A
2V7HA	1RUE1	2OXFA	2V4OA	2CKFB	1UFYA	1D4ZA	1E4XI
2JFNA	1T46A	2Q81A	1DO1A	1DE0A	1FXLA	1K25A	3H4JA
1QDUA	2DLHA	2O5XH	2ENPA	1JBAA	1EVTC	1AB0A	1MQIA
1YPHE	1YNWA	3BWTA	1H1ZA	1RVJL	3BQDA	2EKXA	2YV4A
1BWVA	1HNAA	1B1BA	1GQ8A	1HKOA	1CH9A	1NMEB	1TFBA
2A83A	2YQDA	1RCIA	2EMLA	1KQFA	1RD5A	3A0IX	3H43A
1HX0A	2VR6A	1JY0A	1Z7WA	2NNYA	1FIGL	1GGWA	1VYDA
1ZM6A	1A0NB	1DVIA	1UI8A	1LKYB	2GMIA	3EXBA	1ZPEA
2NMVB	2R8SL	2P2XA	1JG8A	1EQ4A	1PWKA	2OP2A	1YS4A
1GWOA	5CHYA	3BWUD	2ZV3A	3BO9A	2R31A	2F2EA	1ROEA
2JXWA	2NTFA	2KCXA	1HCHA	1HYRB	1FBIH	3E0OA	2C7LB
1CEEA	2UYOA	2FOLA	1E21A	2IW9A	1J36A	1J7MA	3BZ6A
1AABA	1C30B	1C1JA	2DJBA	1YMAA	1PIZA	1NSUA	1EBDC
1KZLA	1EUGA	1O6LA	1SLMA	2EPZA	1A14L	2K9NA	1EW2A

Continued on next page

Table 9.1 Pdb enteries of training MSAs – continued from previous page

1I3OB	2CJ3A	1CRWG	1N2SA	2CZ8A	1I8KB	1TCWA	2DWWA
2MEBA	2Q20A	1SHMB	1BMGA	1YIBA	3DO6A	1YFSA	1KI1A
1AUUA	1R8QA	2F1VA	1TW0A	1WOHA	1V67A	1ZAAC	3D4DA
3CRRA	3DX6A	1MJXA	1NG6A	3BSZL	1K4QA	2MCG1	1VZ0A
2VWSA	3BCQA	1NN7A	2D3AA	2YV1A	1WESA	2A77L	2VD8A
1PS8A	1BX0A	2CW2A	2VXPA	2QMCB	1Q41A	1B3RA	3ENSA
1F9MA	2C7WA	2YWWA	2GHJA	1EAH4	1QOJA	1H5QA	2E3HA
1IJYA	2IM5A	2OSSA	3CG0A	1Z9DA	1UJ0A	1FD9A	3BA7A
1QPDA	1FDQA	2ELIA	1YVFA	3BK8A	1GYCA	1HLKA	1HAXB
1TAHA	1TE0A	1JGLL	2ELZA	1R4PA	3E70C	2YUJA	1DOZA
2VRPB	2Q6GA	1OIDA	1AR61	3BCOA	1BTYA	2J0FA	1AB6A
1I7QA	1DXHA	2V09A	2QHNA	2ESWA	2E74B	9PAPA	1OAUL
1KNOA	3B8EA	1KV0A	1TJRA	2B1KA	1VRPA	1YUCA	2Q5WD
1UWAA	2OS0A	1I1ZA	2CU8A	1SB6A	2JMMA	1L0HA	1R5PA
1I7GA	1AIHA	2C2VC	7FD1A	1E93A	2PHCB	3BIGA	1DXUB
2W1GA	1ES1A	2D31A	3BA4A	3GA5A	2R751	1ZMUA	1HFSA
2P5CA	2GTDA	1H1TA	2WGJA	1DZRA	2HI9A	2V7AA	1CIAA
1MJSA	3GF4A	1HROA	1JXSA	1S17A	2QOFA	1O6AA	3EMNX
2DI2A	1O1JB	2HHLA	3IY3B	2OS1A	1BJJA	1L7TH	1TF3A
2DKRA	3I5TA	3BOSA	4MDHA	1QS0A	2GGMA	3CZJA	1EEPA
966CA	1B89A	1F11A	2BPQA	1JHEA	3I6VA	3GL3A	1J1LA
2ZRUA	2H8FB	3GHZA	1MQ1A	1XGQB	2RIWA	1X09A	1RRPA
1IOQA	1RJ5A	1FCSA	1LW0B	2HF1A	2OWDA	1FCYA	1DWKA
3BH0A	3HJWA	2POIA	1XFRA	1YFKA	1MX2A	1EV5A	2EEHA
3EWFA	3E02A	1BA0A	3HFML	1DPJA	2C0RA	2DUWA	1E6UA
2JBMA	1EVFA	2FWLA	2HGMA	2IG2H	1C83A	2GIMA	1F62A
1NSKR	3F4LA	1CC5A	1VK2A	2GOUA	1ROVA	1PYOA	1DF1A
1Q0XH	2FS2A	1IBEA	2C1NA	1CT2I	2Z10A	1MHLA	2I3HA
1D3WA	3CTQA	1JSTB	2JQHA	1H9SB	1F6NL	1YS9A	2JNVA
2A6HC	1HFPA	2ZOMA	3C5JB	8ATCB	2Q3BA	1VPCA	1HJ5A
1C1BB	1RW7A	3EPYA	1CFPA	1ZCNA	1IX8A	1E4EA	1COV4
1XM5A	2B1XB	3EZPA	2ZUQC	1YZLA	1IX9A	2DMLA	1ATLA
2OQ3A	2DZTA	2POVA	3BF0A	1XTTA	3BSXA	3BLMA	1U4JA
1OAQL	1QM0A	2BUWB	1T47A	2A0LC	2OB4A	1SUPA	2VGGA
6CCPA	3DHXA	1QG4A	1KTQA	2J5PA	2W9DH	1Y4AI	3BQUD
1F00I	1PB1A	2D1XA	1ABOA	3BKWA	3D09A	1K2FA	2IC9A
1EWLA	2GP6A	1MHYD	3FIFA	1P52A	2C4MA	1GP2G	2R9SA
2EMHA	2VJ3A	1CMYB	2W2GA	3B2NA	1RDO1	1CVAA	1P3JA
1QSFA	2VH1A	2DPNA	1RSRA	2ATXA	1KJ3H	1AUWA	2A02A
1MHSA	1AS0A	2RGPA	1IR9A	1V11A	3GL5A	3CHVA	1PUZA
1MBEA	3DHQA	1ZL3A	1MUTA	1NI5A	1GFJA	3HH1A	1HBAB
1FOD4	1UIEA	3FL0A	1CB9A	2F7LA	2A7SA	1CSYA	2NSHA
1ZUXA	1F4HA	1F6LL	2DMOA	1I9EA	2VG5A	1XECA	2BJDA
1A9EA	3C00B	6FABL	2R4FA	1GXSB	1PMRA	1LOFD	2A66A
1BF3A	1X0MA	2DKFA	2D4VA	1SQNA	1NDHA	1QV6A	3E64A
1US5A	1KKTA	1KOLA	1DSXA	2PTJA	1Q2XA	1NCQB	1EK6A
1EZOA	1S2JA	1RTTA	2BO1A	1YC7A	2RN2A	3CYYA	3DRMA

Continued on next page

Table 9.1 Pdb enteries of training MSAs – continued from previous page

3BZTA	1FRFS	2GFYA	2H1OA	1GZ3A	3B5JA	1EX2A	1O6DA
2NXIA	2PCPA	1LPQA	1A9ZA	3B3FA	3DFGA	3EPD3	2DJ1A
1SU0B	1L7IL	1ND4A	1HVDA	3CZCA	2G9HA	2H0PA	2YXWA
3IY5A	1OXNA	2O3KA	1X5SA	1UEXC	1XHFA	1DTSA	1YNJD
3RSPA	2Q1ZA	2GDNA	1IP7A	2QO9A	2DKSA	1HJRA	2J4WD
3D54B	3CW0A	1VJWA	1JLAA	1MIQA	1M04A	3BDMB	1TXAA
3BOWA	1JWQA	1UTXA	2H75A	1BKMA	2AX9A	3FR4A	2BIMA
1VIIA	2RNJA	1AKNA	1LAJA	1F05A	2BLFA	1BIIA	2DLYA
1WMVA	1JJEA	2CFYA	2IS8A	1DOSA	1HVAA	2QO7A	1IAOB
1NI2A	1VHUA	1U06A	1UMRC	1L1DA	1CTLA	1MZAA	3ES3A
1JYSA	1OSFA	1L9QA	2ITVA	1EGJH	1XY0A	2E54A	1EUJA
2EY5A	2NV6A	1UDOA	1JGYM	1Z91A	1ZHCA	1UGBA	2QXCA
1R57A	2IOVA	1S96A	2QE0A	1HJ8A	2RJTA	1UULA	1M57C
2D2QA	1NTGA	3EMOC	2PF0A	2C6NA	1T1TA	1LI9A	1AYZA
2DA2A	1JCAA	2KJKA	1IBDA	1VZUA	2K91B	3DK7A	1F91A
2EOIA	2B1ZA	1DI0A	1Z8RA	2I4QA	3C5WA	1DD4C	2DNPA
1IGJA	1YP1A	2RHZA	1F8EA	3CF6R	1FTCA	2PSYA	2O2CA
1Y0SA	1ULNA	1EKSA	2NLJC	1BKLA	4TMKA	2UZ9A	2TGIA
3EKOA	1KA9H	1EGWA	2BOYA	1F3OA	4UBPB	2AQQA	2OX4A
2I7GA	2OGIA	2ENIA	2ORBH	1GNGA	1EA5A	2Q6NA	2A7RA
1WH3A	1ZYSA	2NL8A	1DPEA	2E2AA	2ZQQA	3ETPA	2NS1B
1FVCA	2Q3RA	1POUA	3BU2A	2TBSA	2P3UB	2AFPA	1QBZA
1D2AA	1QQDB	3G5LA	2VULA	2CT2A	1MZRA	1X3XA	1HILB
2VSLA	1T2MA	3CLNA	1AAMA	3D4MA	1RVTH	1F1AA	1SNNA
3DD4A	3DPRE	1NYOA	1G2SA	2KCPA	2CX4A	1KN1A	3EUB2
1A5MC	1M2IA	2QWOB	3E3BX	3FGHA	1YNYA	1W5CF	2HKYA
2GFWA	3C9WA	2BJMH	1D0QA	1N3BA	2VCQA	1ZVVW	1HIXA
1I2AA	3DGEA	1QHHD	2FMKA	1I5SA	3D57A	1KJQA	1LKJA
1P8DA	3D66A	1E28A	1MA3A	1GH4A	1UHCA	2WGMA	2C0DA
2P4BA	3CPIG	2DXSA	2B0LA	2FTNA	3HCSA	1H85A	2HFPA
2RMLA	1MREL	2CCSA	1AV8A	2DAOA	1UEGA	1RL2A	2GNCA
1F66D	1Q8IA	2JPCA	2Z43A	3E2YA	1G16A	1FL5A	1G5ZA
2AWTA	1JR9A	1OD6A	1O7JA	1I5DA	2A1WH	1YYHA	2HC8A
1WEJH	2QJRA	1EGAA	2QGUA	1CVEA	1GEJA	2QE7D	1H6KZ
1ASGA	1JUDA	1YNIA	1Q98A	1XGUB	1BO0A	2JEJA	1O78A
1M3KA	2FMJA	1UW9A	2KL5A	1D8YA	3DUUA	8PRKA	2F2BA
1P8PA	1POHA	1Z3NA	1QFWL	5CPAA	2HW0A	2YXRA	1CE2A
1PYTC	1QWDA	1L8QA	3BW4A	2WA8A	3EGVB	3BA1A	2P8WS
2RGNC	3C4CA	2WGXA	1P2HA	1NUCA	1JBOB	1LJYA	2DK2A
3FTTA	1PKDA	1ZNCA	1PYCA	2JK6A	2G0FA	3HEIA	1DN0A
1OY5A	3HOFA	1QS0B	1H9QA	3B9YA	1X0AA	1FWGC	1G8OA
1PVC1	1PR3A	1O12A	2OTGA	2EDYA	3PMGA	1FUEA	2D4AB
2GRXC	1CDOA	2DT1A	1Z8PA	3G4DA	1JTOA	3BEWA	3IY3A
1LQWA	3IY2B	1P4LB	2K48A	1T2KA	1BUPA	1A4IA	

9.4 Appendix D: Curriculum Vitae

Curriculum Vitae

PERSÖNLICHE DATEN

Name:	Mehmet Gültas
Geburtstag:	17.05.1984
Geburtsort:	Kirikkale/Türkei
Familienstand:	Ledig
Staatsangehörigkeit	Türkisch
Abschluss:	Promotion

AUSBILDUNG

09/1998–06/2002	Kirikkale Anadolu Lisesi (Gymnasium), Türkei, Abschluss Abitur.
09/2002–06/ 2006	Studium der Statistik und Computer Science an der Karadeniz Technische Universität, mit Abschluss B.Sc.
09/2006–09/2007	Erfolgreicher Abschluss der Deutschen Sprachprüfung.
10/2007–10/2009	Studium der Angewandten Informatik mit dem Nebenfach Bioinformatik an der Georg-August Universität Göttingen, mit Abschluss M.Sc.
10/2009–06/2013	Doktorand und wissenschaftlicher Mitarbeiter im Institut für Informatik an der Georg-August Universität Göttingen, mit Abschluss Promotion.
Seit 01.07.2013	Wissenschaftlicher Mitarbeiter im Institut für Bioinformatik an der Universitätsmedizin, Georg-August Universität Göttingen.

BERUFSERFAHRUNG

03/2008–11/2008	Studentische Hilfskraft bei der Fakultät für Geowissenschaften und Geographie, Georg-August Universität Göttingen.
02/2009–06/2009	Studentische Hilfskraft bei der Fakultät für Geo- und Umweltwissenschaften, Ludwig-Maximilians Universität München.
10/2008–11/2009	Studentische Hilfskraft beim Deutschen Forschungszentrum für Luft- und Raumfahrt.
Seit 01.11.2009	Wissenschaftliche Betreung sowohl unterschiedliche Vorlesungen als auch Bachelor, Master oder Forschungsorientierte Projektarbeiten.

BESONDERE KENNTNISSE

Sprachkenntnisse:	Türkisch (Muttersprache), Deutsch - fliessend in Wort und Schrift, Gute Englisch Kenntnisse.
PC-Kenntnisse:	Fundierte Programmierungs-Kenntnisse, insbesondere Java, PHP, Sql, Html, JavaScript, Css und Flash.
Sonstiges:	Langjährige Tätigkeit in der Selbstverwaltung des Studentenwohnheims sowie Tätigkeit als Vorsitzender.

PUBLIKATIONEN

[1] Erpenbeck D, Voigt O, Gültas M, Wörheide G: *The sponge genetree server-providing a phylogenetic backbone for poriferan evolutionary studies*. Zootaxa 2008,1939:58-60.

[2] Körner MC, Schöbel A, Ashauer K, Schulz K, Gieshold M, Rex R, Zhang H, Gültas M: *Book Chapter: Projekt Promotoranalyse in Gene Graphen, Organismen; Modellierungs- und Anaysemethoden in der Systembiologie* Shaker Verlag 2010.

[3] Gültas M, Haubrock M, Tüysüz N, Waack S: *Coupled Mutation Finder: A new entropy-based method quantifying phylogenetic noise for the detection of compensatory mutations*. BMC Bioinformatics 2012, 13:225.

[4] Gültas M, Düzgün G, Herzog S, Jäger SJ, Meckbach C, Wingender E, Waack S: *Quantum Coupled Mutation Finder: Predicting functionally or structurally important sites in proteins using quantum Jensen-Shannon divergence and CUDA programming*. BMC Bioinformatics 2014, 15:96.

[5] Dong Z, Wang K, Gültas M, Welter M, Dang L, Wierschin T, Stanke M, Waack S: *CRF-based models of protein surfaces improve protein-protein interaction site predictions*. BMC Bioinformatics 2014, under review.

[6] Gültas M, Haubrock M, Wlochowitz D, Zeidler S, Waack S, Wingender E: *TF-Spiker: A novel information theory-based method for the identification of functionally important transcription factors in co-expressed genes*. Bioinformatics 2014, under review.

Conferences, Workshops, Meetings

- Statistical and dynamical models in biology and medicine (October, 2011, Göttingen): Poster präsentation
- Workshop über Algorithmen und Komplexität, 63. Theorietag (Januar, 2012, Brandennburg): Vortag über "A new entropy-based method quantifying phylogenetic noise for the detection of compensatory mutations"
- German Conference of Bioinformatics (September, 2012, Jena): Poster präsentation
- Meeting Gene Regulation and Information Theory (April, 2013, Halle): Poster präsentation
- German Conference of Bioinformatics (September, 2013, Göttingen): Poster präsentation