TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Genomorientierte Bioinformatik

Helix-Helix Interactions in Membrane Proteins: Analysis, Prediction and Applications

Angelika J.M. Fuchs

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. H.-R. Fries
Prüfer der Dissertation:	
	1. Univ.-Prof. Dr. D. Frischmann
	2. Univ.-Prof. Dr. D. Langosch
	3. Prof. D.Sc. N. Ben-Tal, Tel Aviv University, Tel Aviv, Israel

Die Dissertation wurde am 27.10.2009 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 01.02.2010 angenommen.

Abstract

Despite rapidly increasing numbers of available 3D structures, membrane proteins still account for less than 2% of all structures in the Protein Data Bank. In contrast, membrane proteins often constitute key components in a variety of important biological processes and estimations suggest that more than 50% of all prescribed drugs are targeted against this class of proteins. Accordingly, the demand for additional insights into the structural universe of membrane proteins is high motivating the development of reliable structure prediction methods specifically tailored for this class of molecules.

Generally, such methods have to cope with the specific environment of membrane proteins, the lipid bilayer, which restraints both the sequence composition but also the structural diversity of embedded proteins. Still unclear is however, to what extent membrane protein structures are limited in their variety since all structurally characterized membrane proteins so far adopt either a beta-barrel or helix-bundle fold but recent high-resolution structures also indicated a clearly broader structural diversity within these overall fold architectures than initially anticipated.

Several structural features contribute to the distinct characterization of an alpha-helical membrane protein structure. This thesis is focused on one of these characteristics, namely helix interactions formed by helix-helix residue contacts, and aims at a better understanding of membrane protein structural diversity in general but also at the evaluation of new paths in structure prediction and classification of membrane proteins.

First, the diverse nature of helix interactions is illustrated by presenting several newly detected helix interaction motifs that were found to promote high-affine self-association within the genetic screening tool ToxR/POSSYCCAT but importantly could also be recovered from naturally occurring bitopic membrane proteins.

Subsequently, the prediction of helix interactions is addressed. Thereby, the prediction of individual helix-helix contacts was tackled in the first place, on the one hand by conducting the first analysis of co-evolving residues in membrane proteins and on the other hand by developing a novel machine-learning approach trained for the prediction of residue contacts in transmembrane regions. While co-evolving residues were found to carry a strong signal for the detection of interacting transmembrane helices due to their frequent occurrence in close sequence neighborhood to helix-helix contacts, their detection alone was not sufficient to reliably predict helix-helix contacts. However, the neural network based predictor TMHcon incorporating different types of sequence in-

formation significantly improved prediction accuracies up to 26%, therefore constituting the first available method able to predict contacting residues within transmembrane domains with equal accuracy to the best methods available for contact prediction in soluble proteins.

Following the prediction of residue contacts, a detailed analysis is presented addressing the prediction of helix interaction patterns from obtained helix-helix contacts. Using contact predictions derived with TMHcon, interacting helices could be identified with high accuracy (>78%). Interestingly, the sensitivity of obtained predictions can be further improved by incorporating contacts predicted for homologous proteins thereby confirming that helix interactions are likely to be conserved among related proteins.

Finally, a new structural classification approach is introduced identifying proteins with highly similar helix architectures as expressed by their helix interactions. This classification could be shown to closely resemble classification approaches such as SCOP or CATH, which rely on full structure comparisons, thus demonstrating that helix interactions in fact are major structural determinants of membrane proteins. Furthermore, common helix interaction patterns could not only be derived from known structures but also using predicted helix interactions offering the possibility of complementing available sequence-based classification systems of membrane proteins.

As the prediction and classification of helix interactions and accordingly helix architectures constitutes a completely new and valuable field in structural bioinformatics of membrane proteins, it will hopefully gain further interest in the coming years when the number of available membrane protein structures required for the development of such methods is likely to represent the full structure space of membrane proteins even better than is the case at the moment.

Zusammenfassung

Obwohl die Zahl der vorhandenen Membranproteinstrukturen ständig anwächst, ist ihr Anteil an allen Strukturen der Protein Data Bank (PDB) mit nur 2% noch immer verschwindend gering. Dies steht im starken Widerspruch zur biologischen und medizinischen Bedeutung von Membranproteinen, die nicht nur als Schlüsselkomponenten in einer Vielzahl biologischer Prozesse fungieren sondern auch den Angriffspunkt von mehr als 50% aller verschriebenen Medikamente darstellen. Dementsprechend hoch ist das Interesse an Einblicken in die strukturelle Vielfalt von Membranproteinen und damit auch der Bedarf an zuverlässigen Strukturvorhersagemethoden speziell für diese Klasse von Proteinen.

Bei der Entwicklung derartiger Methoden muss insbesondere berücksichtigt werden, dass Sequenzen und Strukturen von Membranproteinen durch Anpassung an ihre spezifische Umgebung - die Lipiddoppelschicht - stark beeinträchtigt sind. Noch ungeklärt ist jedoch, wie weit die strukturelle Vielfalt von Membranproteinen eingeschränkt ist, da zum einen alle Strukturen entweder einer Beta-Barrel oder Helix-Bundle Architektur zugeordnet werden können, zum anderen jedoch immer mehr hochaufgelöste Strukturen auch deutliche Diversität innerhalb dieser übergeordneten Faltungen aufweisen.

Mehrere strukturelle Eigenschaften charakterisieren speziell alpha-helikale Membranproteinstrukturen. Diese Arbeit hat eines dieser Charakteristika zum Thema, nämlich die Interaktionen einzelner Transmembranhelizes. Hauptziele sind dabei zum einen ein besseres Verständnis der strukturellen Vielfalt von Membranproteinen, zum anderen jedoch auch die Entwicklung und Evaluierung neuer Methoden zur Strukturvorhersage und Klassifikation speziell membrangebundener Proteine.

In einer ersten Analyse wird dabei zunächst die Vielfalt beobachteter Helixinteraktionen vorgestellt, indem mehrere neuartige Interaktionsmotife präsentiert werden. Diesen Motifen konnte nicht nur experimentell nachgewiesen werden, dass sie hochaffine Helixinteraktionen ermöglichen, ihre biologische Bedeutung wurde darüber hinaus durch Analyse natürlicher bitopischer Membranproteine bestätigt.

Im Anschluss werden neuartige Methoden zur Vorhersage von Helixinteraktionen vorgestellt. Dabei wird zunächst auf die Vorhersage von Helix-Helix Kontakten zwischen einzelnen Aminosäureresten eingegangen, da diese eine wichtige Information zur Ableitung kompletter Helixinteraktionsmuster darstellen. Zwei Verfahren werden vorgestellt und verglichen, zum einen die Analyse ko-evolvierender Alignmentpositionen, zum

anderen die Vorhersage mittels eines spezifisch für Membranproteine entwickelten neuronalen Netzes. Dabei konnte beobachtet werden, dass ko-evolvierende Reste zwar häufig in direkter Nachbarschaft zu tatsächlichen Kontakten gefunden werden, ihre Genauigkeit jedoch nicht ausreichend ist für eine zuverlässige Vorhersage. Durch die Kombination verschiedener Sequenzeigenschaften in einem neuronalen Netz dagegen wird die Vorhersagegenauigkeit deutlich verbessert. Die entwickelte Kontaktvorhersagemethode namens TMHcon stellt mit einer finalen Genauigkeit von 26% damit die erste verfügbare Methode speziell für Membranproteine dar, die verfügbaren Methoden für lösliche Protein an Genauigkeit gleich kommt.

Nach der Vorhersage einzelner Aminosäurekontakte, wird die Vorhersage ganzer Helixinteraktionsmuster adressiert. Es wird gezeigt, dass interagierende Helizes unter Verwendung vorhergesagter Kontakte mit einer Genauigkeit >78% identifiziert werden können. Die Sensitivität der erhaltenen Vorhersagen kann zusätzlich noch durch Einbeziehung homologer Proteine verbessert werden, wodurch bestätigt wird, dass Helixinteraktionen tendenziell zwischen verwandten Proteinen konserviert sind.

Im letzten Abschnitt der Arbeit wird schließlich gezeigt, wie Helixinteraktionen in einem neuartigen Klassifikationsansatz Verwendung finden, in dem Proteine mit ähnlichen Helixarchitekturen identifiziert werden. Dabei wird gezeigt, dass eine derartige Klassifikation praktisch identisch ist mit strukturellen Klassifikationen der Datenbanken SCOP und CATH, obwohl diese auf kompletten Strukturvergleichen beruhen. Helixinteraktionen stellen daher in der Tat ein wichtiges, wenn nicht gar das am stärksten charakterisierende Merkmal einer Membranproteinfaltung daher. Dies gilt sogar dann, wenn Helixinteraktionen nur vorhergesagt werden, da auch dann Proteine mit ähnlichen Helixarchitekturen mit hoher Genauigkeit erkannt werden konnten.

Mit der Vorhersage und Klassifikation von Helixinteraktionen und daran anschließend ganzen Helixarchitekturen steht der strukturellen Bioinformatik damit ein komplett neues und potentiell wertvolles Gebiet zur Verfügung, das hoffentlich weiter an Bedeutung gewinnen wird, je mehr experimentell gelöste Strukturen von Membranproteinen vorhanden sind.

Acknowledgments

During the past three years of work on this thesis, I was influenced, assisted and motivated by a whole bunch of people. Having the opportunity to thank all of them is a pleasure not to be missed:

Prof. Dr. Dmitrij Frishman :: for being an amazing supervisor offering me scientific freedom and inspiration at the same time; even more important: for demonstrating that science, family, friends, french classes and numerous other interests in fact can be combined into one life.

Prof. Dr. Dieter Langosch :: for initiating and supporting the fruitful collaboration addressing helix interaction motifs in membrane proteins; additionally, for providing me the opportunity of catching a glimpse of membrane protein wetlab work and agreeing to review this thesis.

Prof. D.Sc. Nir Ben-Tal :: for initiating and supporting the project addressing co-evolving residues in membrane proteins; for agreeing to review this thesis even though we have never met in person.

Sindy Neumann and Nadia Latif :: for sharing the office, hundreds of lunch and coffeebreaks with me and still being willing to also spend their freetime with me. Venice 2018 is still on the list!

Dr. Andreas Kirschner :: not only for being an expert in machine learning issues but also for many, many chats making S-Bahn rides and working breaks so much more enjoyable.

Stephanie Unterreitmeier, Jana Herrmann and Johanna Panitz :: for their great work in the lab and a more than pleasant collaboration.

Dr. Thomas Rattei and Dr. Philipp Pagel :: for scientific and technical advice during the past years; special thanks also to Thomas for taking care of our computers virtually 24/7.

All students working with me during my PhD (Jeremias Weihmann, Holger Hartmann, Anita Winkler, Barbara Hummel, Tatjana Goldberg, Gabriele Härtinger, Göksel Kaya, Henrik Failmezger, Michaela Matthes and Boqiao Sun) :: for all your input into my projects and the knowledge I gained from your questions.

All my fellow PhD students and scientists in Weihenstephan (Patrick Tischler, Roland Arnold, Martin Sturm, Qibin Luo, Sheng Zhao, Dr. Erik Granseth, Dr. Antonio Martin-

Galiano, Dr. Pawel Smialowski - to name only some) :: for general and specific support, for discussions (not only related to work) and many fun moments we had together.

Finally:

My family and friends :: for making me laugh when my membrane proteins gave me no reason to; for motivation when I saw no light; for being there when I needed them.

Gholam :: for everything; for more happy moments than I can count; for making my life both easy and fulfilled; for believing in me when I have doubts; for making things possible that others think impossible.

Contents

List of Figures

List of Tables

1 Introduction

Life is incredibly diverse yet simple. Despite hundreds of thousands different species known and catalogued today, all organisms are built up of the same building block: the cell. First described in 1665 by the British scientist Robert Hooke [1], biologists all over the world have put remarkable effort in elucidating function and assembly of this basic unit of life since then, detecting common features and conserved fundamental molecular mechanisms in all analyzed species.

One of the main concepts of a cell is the separation of its content from the surrounding environment. This is achieved by the presence of a phospholipid bilayer, the cell membrane, forming a barrier molecules generally can not bypass without assistance. However, cells need to uptake nutrients in order to sustain life. Waste products of ongoing reactions on the other hand need to get disposed. To this end, proteins are embedded into the cell membrane providing means for transporting molecules in and out of the cell. Additional proteins, inside the membrane or attached to it, allow for the transport of external signals across the membrane giving the cell the possibility to react to environmental stimuli as well as to communicate with other cells. Finally, membranes and their integral proteins play an important role in the energy balance of a cell providing the possibility to build up ion or electron gradients across the membrane which can be used to generate ATP, the energy currency of the cell.

As this work deals with structural properties of a specific class of membrane proteins, namely integral membranes proteins having an alpha-helix bundle architecture, the following introduction aims at summarizing current knowledge regarding this class of proteins. As many properties of membrane proteins can be directly attributed to their surrounding environment, the first section of the introduction provides a short overview about the present view of biological membranes. In the following, special emphasis is put on sequence and structural characteristics of alpha-helical membrane proteins as

well as on known principles guiding the folding process of these proteins. Finally, as one of the major aspects of this work is the development of new methods for the prediction of structural features of membrane proteins, available methodology in this field will be presented.

1.1 Biological membranes

Besides their important role as barrier between the cell content and the environment, biological membranes are also found within eukaryotic cells surrounding intracellular compartments such as the nucleus, mitochondria, chloroplasts, the endoplasmatic reticulum (ER) and the Golgi apparatus. Accordingly, a large variety of different substrates and signals need to be transported across these different membranes influencing strongly the presence of different proteins embedded into the membrane but also composition and organization of the membrane itself. The following paragraphs summarize first how membranes differ among each other with respect to their molecular composition. Secondly, a short overview will be given about present ideas regarding the organization of biological membranes, a field of research facing ongoing evolution from the fluid mosaic model proposed in the seventies [2] to the idea of lipid microdomains becoming increasingly popular over the last fifteen years (for current reviews see [3, 4]).

1.1.1 Membrane composition

Within biological membranes, amphipatic lipids spontaneously arrange into a bilayer structure where hydrophobic lipid tails are held together by non-covalent interactions while hydrophilic head regions are exposed to the aqueous environment which can be either the extracellular or cytosolic space but also the interior of cellular organelles. Three major types of amphiphatic lipids are observed (Figure 1.1A, page 4), namely phospholipids (prominent examples are phosphatiylcholine and phosphatidylserine), glycolipids and steroids with cholesterol being the most common representative of the latter class [5].

The occurrence of lipids and lipid types differs on a number of different scales between individual membranes. Plasma membranes of different organisms vary among each other just as plasma membrane and organelle membranes do of the same eukaryotic organism. While bacterial plasma membranes for example generally lack cholesterol and often contain only a small number of different phospholipids, eukaryotic plasma membranes tend to be enriched in cholesterol and are composed mostly of a larger

variety of phospholipids [5]. Furthermore, different sides of the same lipid bilayer are known to consist of different lipids, an asymmetry specific enzymes are required to maintain [6]. Generally, negatively charged phospholipids dominate the intracellular side of plasma membranes providing a slightly charged environment helpful for binding of membrane-associated or integrated proteins [7]. The enrichment of cholesterol and sphingolipids on the extracellular side on the other hand immobilizes neighboring lipids thereby leading to increased membrane stability but is also important for lipid curvature required for cell structure [8].

Embedded and attached to the membrane bilayer itself are specialized membrane proteins. The ratio protein to lipid is typically 1:1 based on mass proportions which translates to approximately one protein molecule per 50 lipid molecules (assuming a 40 kDa molecular mass for an average protein and 750 Da per lipid molecule) [9]. However, individual membranes may diverge from the 1:1 ratio remarkably such as neuronal plasma membranes where lipids make up roughly 82% of the complete membrane mass or mitochondrial membranes where proteins are dominant contributing 75% of the membrane mass [10]. In both cases, the enrichment of either lipid or protein molecules is tightly coupled to major functions of the respective membranes which is electric isolation in the first case but energy generation in the second case.

1.1.2 Membrane organization: from fluidity to mosaicism

While the asymmetric distribution of lipids across the two individual leaflets of the bilayer is well established, the lateral organization of molecules within the same leaflet is still under extensive experimental research. Following the central 'fluid mosaic model' proposed by Singer and Nicolson in 1972 [2] (Figure 1.1B), lipids are mainly regarded as solvent for membrane embedded proteins and hence the membrane is often referred to as '2D liquid' [2, 5, 11]. Furthermore, the same model suggests free lateral and rotational mobility of membrane molecules leading to unrestricted diffusion and therefore random distribution of lipids and proteins within the membrane [11].

This canonical view has been challenged by experiments including single-particle tracking (SPT) [13], fluorescence recovery after photobleaching (FRAP) [14] and optical laser trapping [15] which found indications that membrane molecules in fact may be hindered in their free lateral diffusion [16]. In accordance with these results, membranes should rather be imaged as environments containing a distinct degree of heterogeneity with so called membrane microdomains (differentiated membrane patches) imposing order on the submicrometer level (Figure 1.1C).

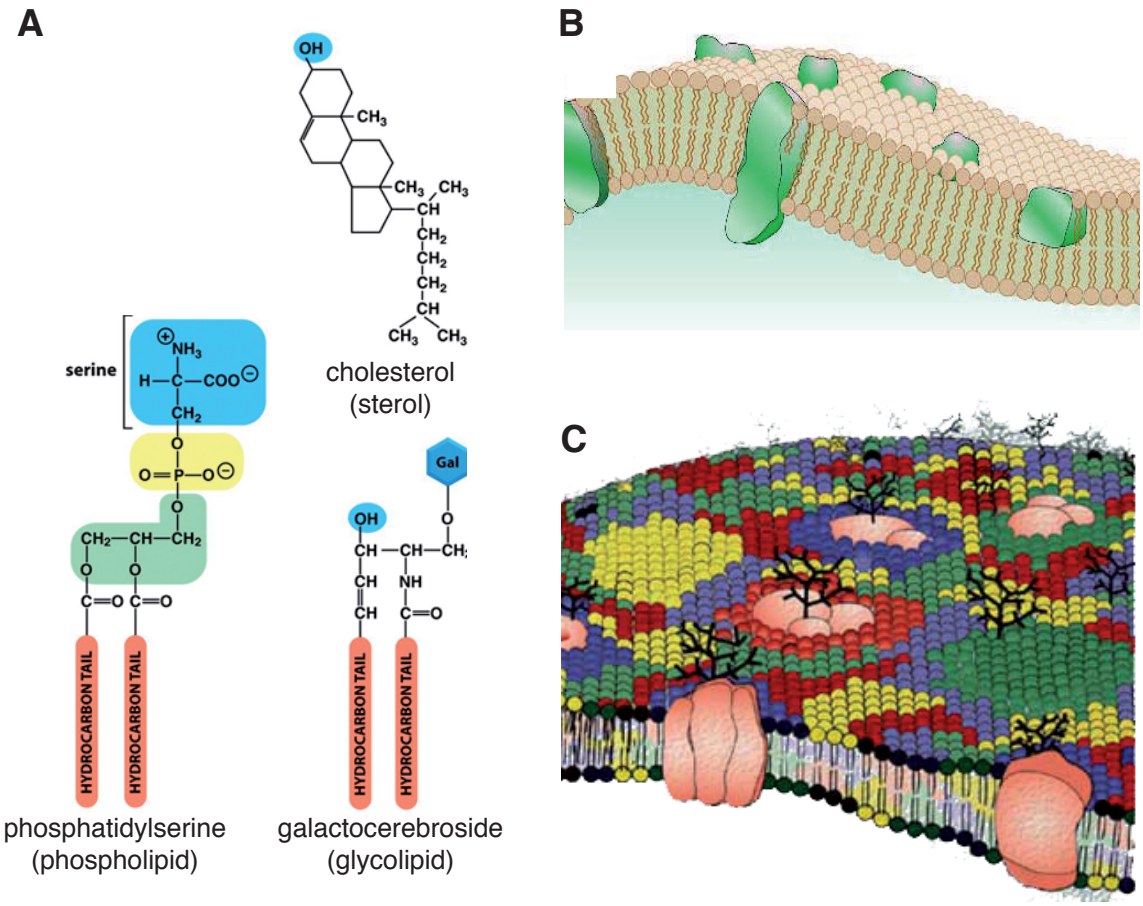

Figure 1.1: Membrane composition and organization. (A) Three major types of amphiphatic lipids observed in biological membranes (figure taken and adapted from [5]). (B) Membrane organization according to the fluid mosaic model proposed by Singer and Nicolson (figure taken from [11]). (C) Membrane heterogeneity caused by microdomains enriched in specific lipids. Membrane proteins may be needed for stabilization of these microdomains (figure taken and adapted from [12]).

While specific structural and functional details often still remain unsolved, two major sources of membrane heterogeneity are presently well accepted [17, 18]. The first type of restriction on protein lateral diffusion arises from barriers formed by the submembrane actin cytoskeleton together with cytoskeleton anchored transmembrane proteins [19]. Lipids and proteins are thought to be confined within corraled areas by this picket fence like structure with long range diffusion between adjacent domains occurring only rarely if the fence fluctuates.

The second line of thought regarding lateral membrane organization emphasizes the role of lipid-lipid interactions and focuses mainly on the analysis of so-called lipid rafts

[20, 3]. This specific class of membrane microdomains corresponds to lateral complexes consisting mainly of sphingomyelin and chosterol whose properties have been extensively analysed using artificial model membranes [21]. Depending on the temperature, such model membranes consisting of a hydrated phospholipid bilayer are observed either in a solid ordered state (termed the S_o phase) or a liquid disordered state (termed the L_d phase). In the presence of cholesterol and sufficient amounts of sphingomyelin and saturated phospholipids, a third L_o phase characterized by lateral mobility in combination with ordered acyl chains can appear besides the L_d phase. Analogous to these observations, lipid rafts are thought to be L_o isles within an L_d phase environment. However, as simplistic model membranes obviously not capture all properties of far more complex plasma membranes, the exact structure and function of lipid rafts is still controversial. A recently proposed model suggests that L_o microdomains are intrinsically unstable and need to be stabilized by membrane proteins [22]. Innovative cell imaging techniques together with *in silico* modelling experiments will be required to gain deeper insights into the organization of biological membranes at microdomain level.

In any case, membrane domains are thought to be transient and small with expected diameters of tens to hundreds of nanometers [17] but highly biologically relevant at the same time as proteins that are supposed to interact are trapped in close neighborhood while other proteins are excluded from a potential interaction. Accordingly, membrane domains have been associated with protein sorting, receptor-mediated signaling and pathogen entry [17, 22].

1.2 Membrane proteins

Alpha-helical membrane proteins constitute between 20 and 30 percent of all ORFs in already sequenced genomes [23]. With a large variety of functions being mediated by membrane proteins and their immense importance for the pharmaceutical industry, remarkable effort has been put over the last years in the experimental as well as computational research on integral membrane proteins. Although this has already produced important insights into occurrence, structure and functions of membrane proteins, available information is still scarce compared to soluble proteins. Especially the number of available 3D structures is still low with less than 2% of all structures in the Protein Data Bank corresponding to membrane proteins. Even though the number of membrane protein structures increases exponentially doubling approximately every third year (White, 2004), structural biology of soluble proteins has an advance of approximately 15 years

which can be mainly attributed to experimental challenges caused by the hydrophobic character of membrane proteins.

Within the following paragraphs, recent findings regarding importance and occurrence of membrane proteins will be summarized. Furthermore, main characteristics of membrane protein topology (describing number and position of transmembrane helices as well as the position of the protein's N-terminus) and 3D structure will be introduced providing the biological background for any computational method addressing membrane protein structures.

1.2.1 Biological and medical importance of membrane proteins

Integral membrane proteins appear in two main architectures: alpha-helix bundle proteins or proteins of beta-barrel type. Together with proteins anchored to the membrane by lipid groups and non-hydrophobic proteins bound in membrane complexes they form the even larger class of membrane-associated proteins.

All-beta integral membrane proteins are found only in the outer membrane of Gram-negative bacteria, mitochondria or chloroplasts. Generally, they form large transmembrane pores, which function mainly as toxins or transporters across the membrane [24]. Alpha-helical membrane proteins on the other hand are often found in oligomeric complexes mediating wide-spread functions such as active transport, ion flow, energy and signal transduction. Important biological processes such as cell-cell-signaling, cell-cell-recognition and the formation of electrical and chemical gradients across membranes are accomplished by this class of proteins [25]. Since alpha-helical membrane proteins are not only much more frequent and much more functionally divers but also the focus of this work, the following paragraphs deal focus only on this class of proteins.

With membrane proteins being key components of a variety of important biological processes, they are also of great interest for the pharmaceutical industry. The large superfamily of G-protein coupled receptors (GPCRs) alone includes receptors for hormones, neurotransmitters, growth factors, light and odor-related ligands [26, 27] and it was estimated that more than 50% of all prescribed drugs are targeted against this protein family [28]. Additionally, several mutations in membrane proteins have been related to the cause of diseases. Nonpolar to polar or charged mutations in the cystic fibrosis conductance regulator (CFTR) for example lead to the clinical pattern of cystic fibrosis [29] and mutations in the insulin receptor may cause diabetes [30]. Hence, membrane proteins are a major factor in the development of new drugs raising the need for further insights into their structural features.

1.2.2 The membrane protein universe: lessons from protein classification

Large-scale classification approaches try to characterize the whole set of genes or proteins for one or several organisms. Membrane proteins have been addressed specifically by a couple of such analyses improving our knowledge of occurrence, diversity and prevalent functions of these proteins.

Membrane protein occurrence Within their ground-breaking first genome-wide analysis of membrane proteins from 16 organisms covering all three kingdoms of life, Wallin and von Heijne could show that membrane proteins cover between 20% and 30% of the ORFs in all analyzed genomes [23]. They further demonstrated that membrane proteins with a small number of transmembrane helices are more frequent than larger membrane proteins although bacterial and archaean genomes have increased numbers of proteins with six and twelve helices corresponding to transporters for small solutes, amino acids or sugars, while eukaryotes have a distinct peak for proteins with seven transmembrane helices representing the important class of G-protein coupled receptors.

In the following, similar analyses have been conducted for individual organisms such as *E.coli* [31] and *S.cerevisiae* [32] using prediction tools in combination with experiments to obtain improved topology models for a large fraction of the membrane proteome of each organism. Among the major findings of both analyses was the prevalent occurrence of membrane proteins with even numbers of transmembrane helices and a $N_{in} - C_{in}$ topology where both N- and C-terminus are found within the cytoplasm suggesting the importance of helix hairpin structures for membrane protein evolution. Furthermore, transporters were found to be the most common functional class of membrane proteins in both *E.coli* and *S.cerevisiae* covering 41% and 32% of all analyzed membrane proteins, respectively, while proteins with six or less transmembrane helices are mostly still lacking a functional annotation.

Protein family classification The classification of proteins into families provides insights into evolutionary relationships among sequences and helps to understand the variety of observed protein sequences. While protein family classification databases such as Pfam [33] generally contain both soluble and membrane proteins, a few protein family classification approaches have been described addressing membrane proteins specifically [34, 35].

In 2002, Liu and colleagues reported in total 637 membrane protein families covering proteins from 26 organisms. In the following, they used the obtained families to determine residues and motifs conserved within related membrane proteins. In contrast, Oberai and colleagues obtained membrane protein families primarily to estimate the number of membrane protein folds in nature and the time required to experimentally solve the structure of a representative set of membrane proteins [35]. They classified roughly 86,000 membrane proteins from 95 genomes into 4075 families with at least 2 members and showed that family size decreases rapidly with few families such as the GPCRs, ABC transporters or the major facilitator family of secondary transporter covering a high number of membrane proteins while most families are found with only a small number of members. From their analysis it seems likely that the space of membrane protein families is already largely saturated given the momentary available sequence data. On a structural level on the other hand, at least ten more years will be required until structural representatives are available for 300 membrane proteins folds which they estimated would cover approximately 80% of all membrane proteins.

Protein fold classification Aiming specifically at the analysis of membrane protein folds, the CAMPS database of membrane proteins [36] classifies membrane proteins directly into clusters likely to represent folds based on sequence similarity and conserved protein topology. From nearly 45,000 membrane proteins from 120 prokaryotic genomes, around 70% could be assigned to 266 fold clusters with at least 15 members, a number very similar to the estimation of Oberai and colleagues [35]. Since at the moment of database construction only 24 of these clusters included a representative structure, structural genomics approaches should aim to choose representatives of the remaining 242 clusters for 3D structure elucidation.

In total, the sequence and structure space of membrane proteins appears to be approachable despite the high percentage of membrane proteins and the experimental difficulties imposed by their high hydrophobicity. With only few new membrane protein families expected to appear in the future [35], computational and experimental biologists seem to have all required sequence data at hand needed for a complete analysis and description of the membrane protein universe. As the number of membrane protein structures increases slowly but exponentially, the next years promise exciting insights into the structural diversity of membrane proteins. Hopefully, this will help to gain a better understanding regarding the fascinating question of how the broad range of membrane proteins functions can be mediated despite strong structural restrictions imposed

by the lipid bilayer.

1.2.3 Membrane proteins in 2D: protein topology

Approaching membrane proteins from a structural perspective, a simplified yet highly informative feature of a membrane protein is its topology describing the number and position of all transmembrane helices and the in/out orientation of the protein with respect to the membrane. Several sequence features have been shown to be deterministic for protein topology with the hydrophobicity of transmembrane segments and an enrichment of positively charged residues in cytoplasmic loops being the most prominent ones. The latter observation has found broad acceptance under the term 'positive-inside rule' especially since it could be shown that this rule holds for both prokaryotic and eukaryotic organisms [37, 38] .

However, as an increasing number of structures and genome-wide studies of membrane proteins become available, not only our understanding of structural diversity of membrane proteins deepens but also our knowledge of membrane protein topology is challenged. While it is still believed that in most cases membrane protein topology is completely defined by the protein's amino acid sequence and hence is conserved within protein families and over the lifetime of a protein, a number of exceptions of this rule have been reported (for an excellent review see [39]).

- Homologous proteins with *opposite topology* were found in large-scale analyses of the *E.coli* and *S.cerevisiae* membrane proteomes [31, 32].

- Even more remarkable, proteins such as the small multidrug resistance proteins EmrE and SugE were discovered that insert into the membrane in both possible orientations with an approximate stoichiometry of 1:1 [40]. Such proteins were termed *dual topology* proteins.

- In addition to proteins with undecided in/out orientation, inefficiently inserting transmembrane helices can give rise to proteins with *multiple topologies* such as the scrapie prion protein which has four reported topologies including a fully cytoplasmic and a fully secreted form [41].

- Finally, cases of *dynamic topology* are discussed where proteins or individual helices change their membrane orientation post-translationally either to adopt their final structure or as part of executing their function as suggested for the protein SecG, a subunit of the SecYEG translocon [42].

So far, topology-predicting programs completely rely on the conventional definition of membrane protein topology assuming a fixed topology for each protein as well as transmembrane helices fully crossing the membrane and oriented largely perpendicular to the membrane. More advanced methods will be required in the future to cope with the diversity of membrane protein topology and structure observed recently.

1.2.4 Membrane proteins in 3D: structural characteristics

Knowing the number and position of transmembrane helices is a first step in understanding the structure of a membrane protein. A full structural description however requires further knowledge especially regarding the packing of individual helices against each other and interactions of helix residues with the lipid environment. Additionally, further structural features such as re-entrant helices or helix kinks need to be considered. The following paragraphs cope with these aspects of membrane protein structure starting with characteristics regarding single transmembrane helices. Afterwards, known types of helix-helix and helix-lipid interactions will be described and recently detected aberrants of regular helix bundle structures are summarized. Especially with the latter being observed frequently in recent membrane protein structure, our view of the diversity of membrane proteins is constantly renewed [43, 44] (Figure 1.2).

Transmembrane helices

Statistical analyses of amino acids in transmembrane helices have consistently observed a distinct distribution of amino acid types along the length of a transmembrane helix. While aliphatic residues and phenylalanine are enriched in the centre of the membrane, tryptophan and tyrosine are frequently found at the border between hydrophobic membrane core and hydrophilic environment [45, 46]. Charged or polar amino acids are only poorly represented in transmembrane segments with a total frequency of lower than 5% [47, 48]. Analyzing the conservation of individual amino acids within transmembrane helices, glycine and proline were found to be significantly enriched in conserved positions, while the opposite was the case for Ile, Val, Met and Thr, which seem to be highly mutable within transmembrane segments [34]. Lately, approaches have been developed to determine the contribution of all twenty amino acids to the free energy of membrane insertion [49, 50]. These experiments provide a biophysical explanation for the observed statistical amino acid distributions as they could demonstrate that for example tryptophan and tyrosine in central helix positions are unfavorable for membrane insertion but

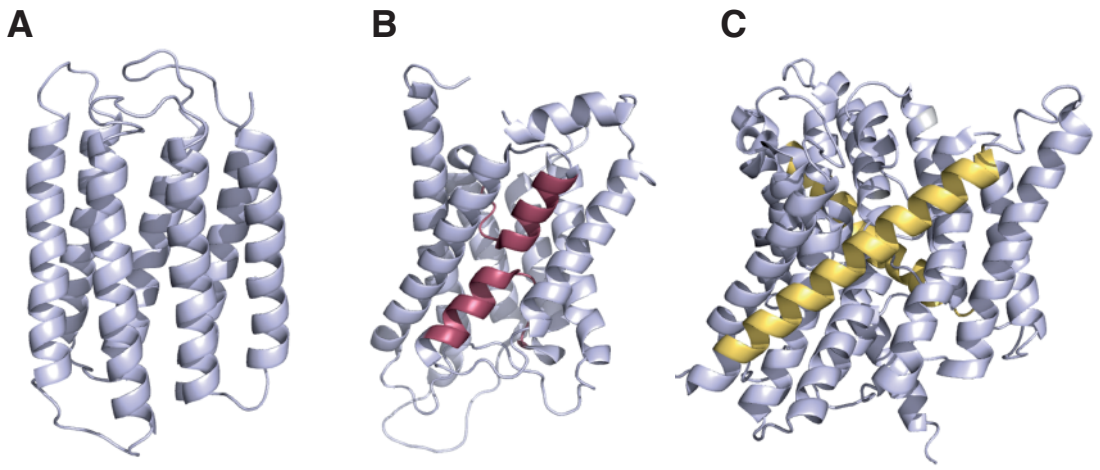

Figure 1.2: Structural diversity of membrane protein structures. (A) Structure of bacteriorhodopsin (PDB 2BRD). Seven transmembrane helices form a regular alpha-helix bundle structure. (B) Structure of aquaporin 1 (PDB 1J4N) containing two re-entrant helices (shown in red). (C) Structure of the H(+)/Cl(-) exchange transporter clcA (PDB 1KPK) consisting of ten transmembrane helices. Helices differ significantly in their length and may be strongly tilted with respect to the membrane (shown in yellow).

less problematic in border positions.

Another important property of transmembrane helices beside amino acid composition is their length. With typically 30 Å thick hydrophobic core membrane regions, transmembrane helices consist in general of 20 to 30 amino acids with 23 amino acids being the average [51]. Helices significantly longer than this average are expected to tilt with respect to the membrane normal in order to adjust to the membrane thickness and prevent a so-called 'hydrophobic mismatch' [52, 53].

Helix-helix packing

With helix-helix interactions being an important stabilizer and determinant of membrane protein structures, major efforts have been put into the understanding of these interactions. Thermodynamic measurements [54, 55, 56] as well as genetic approaches [57, 58] have been developed and applied to determine the strength of individual helix interactions and estimate the effect of mutations on helix assembly (for a recent review see [59]). Furthermore, available membrane protein structures have been rigorously analyzed to obtain insights into the helix-helix packing of membrane proteins and to derive amino acid propensities for the participation in interhelical interactions [60, 61, 62].

While the packing of α-helices in membrane proteins is commonly described by a

'knobs-into-holes' packing introduced originally for soluble coiled coils [63], comparative studies of helix packing have found remarkable differences between soluble and membrane proteins. It has been shown that membrane proteins are on average tighter packed than soluble proteins [48] and that closely packed small residues are the major source for this observation [61]. In contrast, large hydrophobic and aromatic residues have the highest packing values within soluble proteins [61]. Furthermore, membrane proteins were found to have a strong preference for left-handed crossing angles around $+20°$ although angles between -56° and $+67°$ are principally possible as reported by a study of 88 transmembrane interfaces [64]. Within soluble proteins, the preference for a certain range of helix crossing angles seems to be less clear. Additionally, the range of possible interhelical angles is larger [62] and right-handed angles are found more frequently than in membrane proteins [62, 65].

Interestingly, left-handed and right-handed helix interactions within membrane proteins were found to differ in their preferred mode of interaction. Left-handed interactions are often promoted by a heptad motif such as the LxxLxxxLxx motif of leucine zippers, where amino acids at positions a and d are the main contributors to the contact interface [63]. Right-handed interactions on the other hand seem to rely on a regular tetrad pattern with amino acids at positions a and b forming the helix-helix interface [66].

Detailed analyses of amino acids in helix-helix interfaces as well as mutational studies have given further insights into the nature of non-covalent interactions within membrane proteins highlighting the special importance of polar residues as well as small residues for transmembrane helix assembly. Although van der Waals packing is the main determinant of this process and hence residue pairs formed from apolar amino acids (F, L, V, I and A) are the most frequently found ones in membrane helix interfaces [60], several studies observed that the composition of helix interfaces in membrane proteins is even more diverse than in soluble proteins [60, 61]. While soluble proteins have a strong preference for salt-bridge interactions formed by oppositely ionizable amino acids, membrane proteins feature a much broader range of polar-polar interactions covering residue pairs of polar residues such as S, T, Y, N and Q [60]. In average, every transmembrane helix is expected to form at least one hydrogen bond with side chain-backbone hydrogen bonds contributing substantially to this observation as every second hydrogen bond between transmembrane helices seems to be of this type [62]. Experimentally, transmembrane helices with motifs of multiple serine and threonine residues or single glutamine, asparagine, aspartic acid or glutamic acid residues were found to promote strong self interaction [67, 68] further confirming the importance of polar residues for

helix associations. Small residues (G, A and S) on the other hand, were repeatedly found to be among the most overrepresented ones within membrane helix interaction interfaces [61, 62]. The importance of these residues has found even more attention after detecting that motifs consisting of two small residues spaced by three residues ([GAS]xxx[GAS]) are a recurrent theme in helix-helix-interfaces [69, 70, 71]. GxxxG, the most prominent motif of this kind, was originally identified in mutagenesis experiments using the glycophorin A (GlpA) transmembrane helix dimer [72, 73] and was later found not only to be frequently present within transmembrane helices [74] but also to be strongly conserved [34]. Generally, small residues are thought to allow very close contact between transmembrane helices and accordingly extensive van der Waals interactions [75] as well as the formation of Cα-H $\cdots$ O hydrogen bonds across the helical backbone [76]. Lately, measurements of helix interaction energies have indicated however, that GxxxG-containing transmembrane segments may interact with remarkably different strength suggesting that sequence context is equally important for interaction as the GxxxG motif itself [77, 59].

Helix-lipid interactions

While helix-helix interactions are established as major determinants of membrane protein structure for many years now, the influence of surrounding membrane lipids on membrane protein assembly and membrane protein structure in general is just recently emerging. Nevertheless, several well studied examples have demonstrated that interactions between transmembrane helices and membrane lipids modulate different aspects of membrane protein structures including helix-helix interactions and helix tilts (for a recent review see [78]). Accordingly, transmembrane helices are thought to interact not only due to the presence of specific sequence motifs or favorable helix-helix contacts but also as a result of less favorable helix-lipid interaction which would be observed for example in the case of a hydrophobic mismatch between bilayer thickness and transmembrane helix length [79]. As mentioned earlier, such hydrophobic mismatch is thought to be one of the reasons for tilted helices [53]. Additionally, helices are expected to tilt due to interactions between anionic lipid headgroups and positively charged helix anchoring residues [78]. Generally, changes in lipid composition can be expected to significantly alter transmembrane helix assembly which can further propagate changes in extramembranous parts and quarternary structure of the protein. Such changes can be even extensive enough to shape membrane protein function [80].

'Unusual' structural features of membrane proteins

The recent increase in available membrane protein structures - Steve White's list of membrane proteins with solved structure currently contains 193 unique proteins [1]- has gained remarkable insights into the structural variability of alpha-helical membrane proteins. Despite the limited number of overall fold architectures and the general restrictions imposed by the hydrophobic nature of the lipid bilayer, a still increasing number of structural aberrations have been described modulating our view of canonical membrane protein structures (for example structures see Figure 1.2, page 11).

In a typical transmembrane segment the residue at position i forms a hydrogen bond with the residue at position i+4 within the same transmembrane helix. However, several alterations of this canonical helix conformation have been described for membrane proteins such as π-bulges resulting from hydrogen bonds between a residue at position i and a residue at position i+5 [81], helix unwinding [82] and proline-induced kinks [83]. These irregularities lead to local conformational instability and are thought to be primary spots for conformational changes [84]. Additionally, at least one backbone carbonyl group is exposed which can be important for the binding of cofactors [83] or inter-helical hydrogen-bond formation. An analysis of structures in the PDB in 2001 observed, that nearly 50% of all transmembrane segments in alpha-helical membrane proteins contain such elements [85]. Additionally, transmembrane helices may form so-called 'reentrant loops' which cross the membrane only halfway and then return to the side where they entered the membrane [86]. Helices may also be disrupted and much longer and much more tilted than expected from the first available membrane protein structures [62].

Altogether, membrane proteins emerge to be structurally diverse to an extent not anticipated ten years ago. Despite limited variability on amino acid level and restricting influences of the lipid bilayer, evolution has found means of diversification that provide the basis for the wealth of known functions mediated by membrane embedded proteins. With increasing evidence that membrane protein structure is not only dependant on protein sequence but also the specific membrane environment, an additional level of complexity is added to the analysis and understanding of α-helical membrane proteins. Accordingly, the rapid development of experimental techniques for the analysis of membrane protein structures - for reviews see [87, 88] - promises exciting insights over the next years.

[1] http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html, 15th June 2009

1.2.5 Membrane protein folding

While helix association is an important step in the complete membrane protein folding process, equally important is the insertion of transmembrane helices into the lipid bilayer after their synthesis by the ribosome. Accordingly, membrane protein folding has been described as a multi-step process, for example by the two-stage model [89, 90] which suggests that α-helices are formed and inserted into the membrane in a first step and associate then into the final structure during the second step.

The molecular translocation machinery where these processes mostly take place is a multi-subunit complex located in the endoplasmic reticulum membrane of eukaryotes (Sec61 translocon) or the plasma membrane of eubacteria and archaea (SecYEG and SecYEβ, respectively). Due to the X-ray structure of an archaean translocon ([91]) solved in 2004 and advanced experimental systems for analyzing helix insertion efficiency [49, 50], increasing information about the molecular mode of operation of this complex is now at hand (for a recent review see [92]). Proteins meant for secretion or insertion into the membrane are recognized during translation by a signal recognition particle (SRP) and transported to the translocon. There, the nascent protein chain is transported through the so-called hydrophobic collar within the SecY subunit of the translocon with the possibility of releasing transmembrane segments laterally into the surrounding lipid bilayer through a lateral gate formed by two transmembrane helices of SecY. According to recent theories, the decision whether a protein segments leaves the translocon through the lateral gate or not is a direct result of a simple partitioning process [93, 92]. Hydrophobic segments with favorable free energy of insertion due to interactions with surrounding lipids leave the translocon while polar segments prefer to stay in the aqueous environment of the translocon. In case of polytopic proteins containing more than one transmembrane segments, individual helices are believed to be inserted strictly in a N- to C-terminal order although helix interactions between already inserted segments and the helix currently inside the translocon seem to be possible facilitating the membrane insertion of helices likely not hydrophobic enough by themselves [94, 95].

1.3 Structural bioinformatics of membrane proteins

As membrane proteins are challenging to work with experimentally while at the same time their structural diversity is strongly limited by the surrounding membrane, their

analysis using computational methods seems to be both useful and promising. Accordingly, a whole field of structural bioinformatics has opened up developing methods specifically tailored for this class of proteins. Addressed problems range from the sequence based prediction of membrane protein topology over the prediction of individual structural properties such as helix kinks or re-entrant helices to the full 3D structure prediction of membrane proteins (for recent reviews see [43, 44, 96]). The following section summarizes results and open questions in each of these fields.

1.3.1 Topology prediction

Predicting the topology of a membrane protein aims at detecting the correct position of all transmembrane segments as well as the in-out orientation of the protein within the membrane, the latter being equivalent with predicting the position of the N-terminus of the protein. While the prediction of the inside/outside orientation is consistently based on two topogenic signals, namely the enrichment of positively charged amino acids within inside loops [97] and the occurrence of N-terminal cleavage signals indicating an outside position of the N-terminus [98], the detection of transmembrane helices has seen major enhancements over the years. Early hydrophobicity-scannning algorithms such as the Kyte-Doolittle method [99] applied a sliding-window approach, where a hydropathy value was assigned to all amino acids within a window of given size and all segments having a summed hydrophathy value above a certain threshold were predicted to be transmembrane helices. First improvements of these early methods used neural networks or sequence profiles generated from multiple sequence alignments rather than single sequences [100]. A significant increase in prediction accuracy was gained by methods relying on Hidden Markov Models (HMMs) with several hundreds of free parameters required to be optimized using a training set of proteins with known topology. TMHMM [97] as well as HMMTOP 2.0 [101] are prominent examples of this type of methods, which have found widespread application as they were repeatedly ranked among the best methods in comparisons of available topology-predicting tools [102, 103, 104, 105].

Despite the prominence of methods such as TMHMM and HMMTOP, the development of new methods for the prediction of membrane protein topology is still ongoing. Recently proposed methods use dynamic Bayesian networks instead of HMMs [106] or compile and evaluate predictions of several individual predictors into a consensus prediction [107]. Increasing effort is put also into the development of methods predicting simultaneously protein topology and signal peptide sequences as misprediction of signal peptides is known as one major error source for topology prediction [98, 106]. Finally,

as the molecular forces guiding transmembrane helix insertion are better understood, topology predictors try to predict transmembrane helices rather based on experimentally derived physical properties known to be important for membrane insertion instead of statistical analysis of proteins with known topology [108]. As such methods could be shown to perform already with equal accuracy as statistics-based methods [108], they promise further advances in the field of topology prediction as soon as more and more information regarding the folding and insertion of membrane proteins is available.

Generally, state-of-the-art topology prediction methods can be expected to reach full-topology prediction accuracies between 70% and 80% [43, 107], where a correctly predicted topology requires the correct number of transmembrane helices, the correct position of the N-terminus and approximately correct position of all helices. However, with only about 400 membrane proteins having an experimentally confirmed topology [43], the validation and comparison of different prediction methods is still error-prone and varies strongly depending on the used dataset leaving room for further improvements.

1.3.2 3D structure prediction

Given the limited number of folds membrane proteins and alpha-helical membrane proteins specifically can comprise, 3D structure prediction of membrane proteins seems to be clearly easier than for soluble proteins suggesting that currently available computational resources might be sufficient for *ab initio* folding of membrane proteins. However, overall success of 3D structure prediction for membrane proteins is still small especially since available methods for soluble proteins generally can not be directly applied to membrane proteins due to the different environment formed by the membrane. Homology modeling techniques based on known structures would principally be able to produce structural models of membrane proteins with similar accuracy as reported for soluble proteins [109], but the small number of available membrane protein structures strongly limits the practical usage of these methods.

Historically, 3D structure prediction efforts were focused mainly on individual membrane proteins or specific membrane protein families. Especially structure prediction of GPCRs has found significant interest with several methods being developed specifically for this class of proteins [110, 111, 112, 113]. Lately, structural models have been derived for all human GPCR candidates using the threading and refinement protocol TASSER, which could be shown to model bovine rhodopsin with a global Cα RMSD of 4.6 Å.

Addressing structure prediction of membrane proteins in general, a membrane protein specific version of the Rosetta algorithm for structure prediction has been introduced

in 2006 [114]. There, low-resolution models of membrane proteins are calculated using the Rosetta fragment assembly method in combination with a membrane protein specific energy function modelling residue-residue and residue-lipid interactions within a membrane environment. Subsequently, the method has been further improved by the incorporation of an all-atom physical model able to produce high-resolution structures of membrane proteins [115] and the usage of constraints derived from known helix interactions or mutation experiments which allow for the prediction of membrane proteins up to 300 amino acids [116]. Using these improvements, high-resolution models of small (<150 residues) and large membrane proteins could be obtained with RMSDs <2.5 Å and <4 Å, respectively.

In addition to the work of Barth and colleagues [116], several other approaches have been developed using constraints from experimental or low-resolution structural data to obtain high-resolution models of membrane proteins (reviewed in detail in [117]). Results from FRET experiments or chemical crosslinking for example have been used to distinguish native helix conformations from non-native conformations [118]. Furthermore, cryo-EM structure with a resolution of 5-10 Å have been successfully refined to atomic models by methods assigning and orientating individual transmembrane helices based on observed hydrophobicity, evolutionary conservation and residue co-evolution [119, 120, 121].

1.3.3 Prediction of individual structural features

When alpha-helical membrane proteins were still thought to fold into regular helix bundle structures, structural bioinformatics was dealing mainly with the determination of membrane protein topology followed by the prediction of possible helix arrangements. Recent structures displaying a previously unexpected complexity have opened up a complete new field tackling the prediction of individual structural elements such as for example reentrant helices. Elofsson and van Heijne have termed these kind of prediction tools 2.5D prediction methods [43] since they are positioned between 2D topology predictions and 3D *ab initio* structure predictions.

A commonly addressed task of 2.5D prediction methods is the prediction of lipid exposure for each residue within the membrane based mostly on side chain polarity and sequence conservation since lipid-exposed residues were shown to be generally more hydrophobic and less conserved. Several methods have been published over the last years [122, 123, 124] with reported accuracies up to 88% for correctly predicted individual residues [122]. Furthermore, methods are available for the prediction of proline-induced

kinks [125] as well as the detection of reentrant helices [126, 86]. Such helices were originally observed in the structures of the KcsA potassium channel [127] and the aquaporin-1 water channel [128] and were later shown to contain preferentially small residues and special functional motifs usable for their prediction. Finally, the detection of long and tilted helices is now approachable by a method predicting each residue's distance from the membrane center [129].

1.4 Motivation and overview of this work

With only a very limited number of available 3D protein structures and a high biological and medical importance, membrane proteins are an important research subject for structural bioinformaticians. As the amino acid composition of transmembrane segments deviates remarkably from soluble proteins, the development of structure prediction methods specifically tailored for this class of proteins is required. Given the structural variability observed in recent membrane protein structures, more and more methods are developed that are neither predicting membrane protein topology nor full membrane protein structures, but are addressing specific structural aspects of membrane proteins. This work is placed in this field of structural bioinformatics focusing on the analysis and prediction of residue and full helix interactions within transmembrane domains of alpha-helical membrane proteins. The following paragraphs will give a short overview about all analyses and projects conducted as part of this thesis. Thereby, the term *'helix-helix contact'* is always used to describe pairwise residue interactions between amino acids placed on different transmembrane helices while the term *'helix-helix interaction'* corresponds to two full transmembrane helices connected by at least one helix-helix contact.

Within the following chapter *"Detection of helix interaction motifs"*, results of a experimental and computational study of helix interaction motifs in membrane proteins will be presented. As described in section 1.2.4, several sequence motifs are known that promote strong helix interactions within a membrane environment. Using the ToxR/POSSYCCAT system, a genetic screening tool for high affinity transmembrane helix interactions, further candidates for such sequence motifs were identified by the group of Prof. Dieter Langosch (TU München). Subsequent sequence analysis of naturally occurring membrane proteins proves that the candidate motif FxxGxxxG as well as several identified motifs consisting of a charged amino acid in combination with GxxxG are significantly overrepresented in this dataset thus highlighting their biological rele-

vance.

Within chapter 3, entitled *"Co-evolving residues in membrane proteins"*, the first analysis of concurrently mutating residues within transmembrane domains is presented. Such residues have been frequently analyzed in soluble proteins and have been used for residue contact prediction suggesting their possible application for the prediction of helix-helix contacts in membrane proteins. The executed analysis demonstrates that co-evolving residues alone are not sufficient to reliably predict helix-helix contacts, but that these residues still carry a strong signal for the detection of interacting transmembrane helices due to their frequent occurrence in close sequence neighborhood to helix-helix contacts. A developed consensus predictor combining predictions from several individual prediction algorithms was further able to predict helix-helix contacts with higher accuracy than any single method available.

Following the work on co-evolving residues, the prediction of residue contacts between transmembrane helices was continued by the development of an advanced, neural network based predictor incorporating different types of input information for the identification of helix-helix contacts. As described within the chapter *"Prediction of helix-helix contacts using neural networks"*, the developed method called TMHcon is the first method able to predict contacting residues within transmembrane domains with equal accuracy to the best methods available for contact prediction in soluble proteins. The prediction of helix-helix contacts using a neural network was jointly executed with Dr. Andreas Kirschner (TU München).

Chapter 5 (termed *"Prediction of interacting helices"*) switches the focus from the level of individual amino acids to the level of full transmembrane helices. Here, it is demonstrated, how obtained helix-helix contacts can be used to predict the interaction of transmembrane helices and accordingly the helix architecture of membrane proteins with high accuracy and specificity. While prediction quality can be shown to increase already significantly by using helix-helix contacts predicted with the earlier developed TMHcon method instead of co-evolving residues alone, especially prediction sensitivity can be further improved by incorporating additional contact information predicted within a consensus approach incorporating structurally related proteins obtained from the CAMPS database of membrane proteins ([36], version 2.0 developed by Sindy Neumann, TU München).

Finally, the chapter *"Classification of helix architectures"* introduces a possible field of application where helix interactions (obtained from experimentally determined structures or computationally predicted) can be of great value. After analyzing first the cur-

rent classification of membrane proteins in major structural databases (namely SCOP and CATH), a new classsification system is introduced that specifically addresses membrane protein helix architectures. Using graph representations of transmembrane helices and their interactions, such helix architectures can be visualized and compared among each other. Clustering proteins based on the similarity of these helix interaction graphs is able to closely resemble classification approaches such as SCOP or CATH, which rely on full structure comparisons, thus demonstrating that helix interactions in fact are major structural determinants of membrane proteins. For membrane proteins with no experimentally solved structure available, predicted contacts can be used to identify proteins whose helix architectures have a high likelihood of being similar.

Altogether, the main goal of the work presented in this dissertation is to enrich the field of structural bioinformatics of membrane proteins by a new 2.5D prediction task not addressed in the past by any other research group, but equally valuable to biologist working with membrane proteins as well as bioinformaticians trying to predict full membrane protein structures. For biologists, predicted helix-helix contacts and helix interactions can be helpful to gain insights into the structural organization of membrane proteins when no experimentally solved structure is available. Structural bioinformaticians on the other hand can employ predicted helix-helix contacts to constrain the conformational search space in *ab initio* structure predictions making the detection of the native structure easier approachable. Finally, the large scale application of helix-helix contact and helix architecture predictions complement available sequence clustering approaches permitting additional insights into the structural variability of alpha-helical membrane proteins.

2

Detection and analysis of helix interaction motifs

The importance of individual amino acids for the association of transmembrane helices has been extensively studied in the past [57, 60, 61, 130]. Thereby, recurrent sequence motifs have been discovered that promote strong helix interactions in both artificial and naturally occurring transmembrane sequences (for a review see [65]). So far, the best characterized helix interaction motif is the GxxxG motif, which was first detected by analyzing the dimerisation of human glycophorin A (GpA) [72], but has meanwhile shown to be one of the most frequent sequence motifs in natural transmembrane helices [74]. However, further energetic measurements have indicated that GxxxG alone often may not be sufficient for strong transmembrane helix association [59]. Instead, local sequence context seems to strongly influence dimerisation free energy [77, 131] motivating further analyses of strongly interacting transmembrane domains.

Genetic screening tools such as the TOXCAT [58] and the ToxR/POSSYCAT system [132, 133] can be used to evaluate the effect of sequence context on known interaction motifs but also to detect completely new motifs. Within these systems, self-dimerisation of a bitopic membrane protein (i.e. a membrane protein containing one single transmembrane helix) is linked to expression and accordingly activity of a reporter gene. Recent enhancements of both the TOXCAT and the ToxR/POSSYCAT approach additionally allow for the detection of heteromeric interactions [134, 135].

Within the following chapter, results of two ToxR/POSSYCAT analyses aiming at the detection of high affinity transmembrane helix interactions will be presented. As all experiments were carried out by members of the group of Prof. Dieter Langosch (Chair of Biopolymer Chemistry, TU München), the experimental setup as well as all results gained directly with the ToxR/POSSYCAT system will only be briefly summa-

rized in the first section of this chapter. Following the experimental detection of motif candidates based on combinatorial libraries of transmembrane sequences, computational analyses of naturally occurring bitopic sequences were executed to ensure the biological relevance of all detected motifs. The second section of the chapter summarizes results of these bioinformatic analyses executed by myself confirming the occurrence of individual candidate motifs in a large set of bitopic membrane proteins.

Experimental and computational results presented in this chapter were already published in [136, 137], another publication was recently submitted [138].

2.1 Experimental motif identification using the ToxR/ POSSYCAT system

2.1.1 The ToxR/POSSYCAT system

Using the ToxR/POSSYCAT system, self-interacting transmembrane domains can be selected from combinatorial sequence libraries and the affinity of selected sequences can be characterized, both within an *in vivo* environment. The system is based on the ToxR transcription activator originating from the proteobacterium *Vibrio cholerae* where it is located inside the inner membrane. This transcription factor regulates expression of genes controlled by either the *ctx* or *ompU* promoter and is only active after di- or oligomerisation of its cytoplasmic domain. Naturally, activation of ToxR is triggered by environmental stimuli leading to expression of several target proteins such as cholera toxin, the outer membrane protein OmpU and other virulence factors [139, 140, 141].

For the detection and characterization of high-affinity transmembrane domains with the ToxR/POSSYCAT system (Figure 2.1), chimeric proteins are constructed where the original ToxR transmembrane domain is replaced by the transmembrane domain of interest. Furthermore, the maltose-binding protein (MalE) is attached as periplasmic domain which serves as control for correct membrane insertion as only constructs placing the MalE domain within the periplasma are able to complement the MalE deficiency of *E.coli* PD28 cells. In case the integrated transmembrane domain is able to self-interact, the induced dimerisation of the cytoplasmic ToxR domain activates the expression of a reporter gene controlled by a *ctx* or *ompU* promoter within engineered *E.coli* cells . In *E.coli* EL61 cells, chloramphenicol acetyltransferase is used as reporter gene in combination with the *ompU* promoter indicating self-interacting transmembrane domains by chloramphenicol resistance. Subsequently, interaction affinity can be fur-

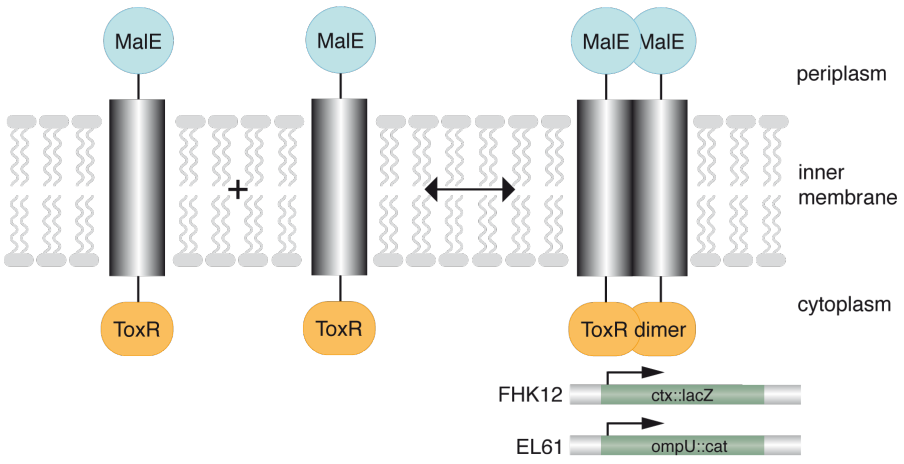

Figure 2.1: The ToxR/POSSYCAT system. After self-interaction of transmembrane domains, cytoplasmic ToxR dimers activate the transcription of reporter genes under the control of a *ctx* or *ompU* promoter. Periplasmic MalE domains allow for the analysis of correct membrane insertion. Figure adapted from [142]

ther quantified by *E.coli* strain FHK12 where the gene *lacZ* encoding β-galactosidase (β-gal) is under control of the *ctx* promoter. By monitoring β-gal activity via the hydrolysis of *o*-nitrophenyl-β-D-galactopyranoside (OPNG), ToxR activity and hence dimerisation affinity of tested transmembrane segments can be compared among different constructs (further details regarding the ToxR/POSSYCAT system are summarized in [73, 132, 136, 137]).

2.1.2 Interaction motifs identified with the ToxR/POSSYCAT system

Within two independent analyses, combinatorial libraries of transmembrane domains were screened for high-affinity self-interactions using the ToxR/POSSYCAT system by Stephanie Unterreitmeier, Jana Herrmann and Johanna Panitz (Chair of Biopolymer Chemistry, Prof. Dieter Langosch, TU München). While both analyses were executed following the same procedure (Figure 2.2A), tested sequence libraries differed substantially from each other to allow for the identification of yet uncharacterized interaction motifs (Figures 2.2B and C).

Briefly, during each analysis a library of randomized helix interface sequences was generated using PCR in combination with a partly degenerated forward primer. From this library, helices with the possibility to interact were selected with the ToxR/POSSYCCAT

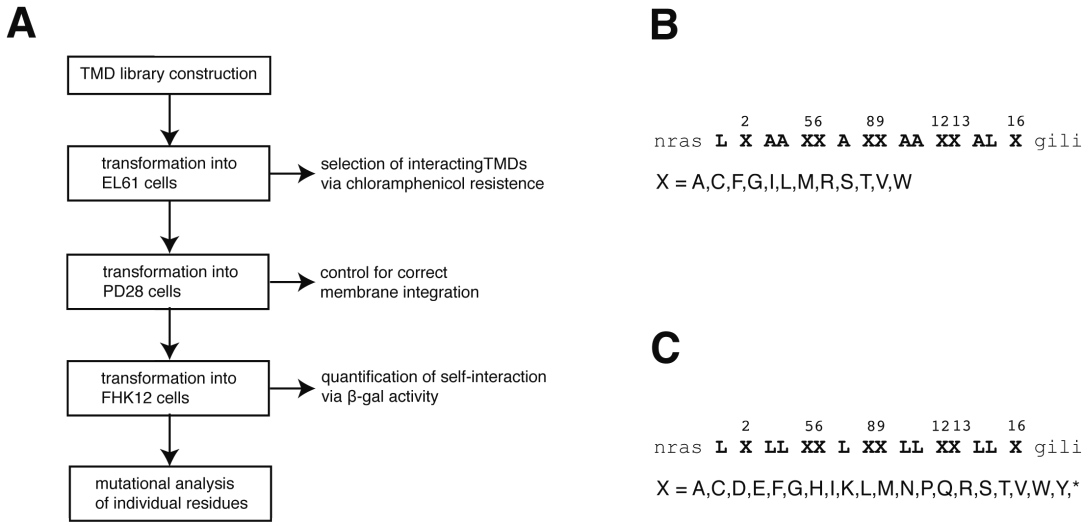

Figure 2.2: Identification of helix interaction motifs using the ToxR/POSSYCAT system. (A) Major experimental steps executed during the identification of helix interaction motifs (TMD is used as abbreviation for transmembrane domain). (B) Interfacial helix positions (X) randomized during the first analysis executed by Stephanie Unterreitmeier. (C) Interfacial helix positions (X) randomized during the second analysis executed by Jana Herrmann and Johanna Panitz.

system by testing the chloramphenicol resistance of EL61 cells transformed with the constructed plasmid library. Identified candidate sequences were then tested for correct membrane insertion using MalE deficient PD28 cells and were further analyzed with respect to their interaction strength by measuring β-gal activity in FHK12 cells. Those transmembrane sequences interacting with high affinity were compiled and sequenced. After detecting manually common patterns within all obtained sequences which might correspond to helix interaction motifs, individual mutational studies were executed to verify the importance of certain amino acids at specific positions within the transmembrane helix interface (Figure 2.2A).

In both analyses, eight positions of a 16-residue heptad repeat motif were randomized during library construction. However, during the first screening process executed by Stephanie Unterreitmeier, mostly hydrophobic amino acids were allowed during randomization and non-interface positions were filled with alanine residues as these residues are known not to be beneficial for helix interaction [143] (Figure 2.2B, the obtained set of sequences is from now on referred to as Library Ala). In the second analysis executed mainly by Jana Herrmann and Johanna Panitz, all naturally occurring amino acids were permitted in interface positions. Furthermore, the remaining positions were occupied

by leucine instead of alanine to facilitate the integration of polar or charged residues in interface positions by increasing the average hydrophobicity of all randomized sequences (Figure 2.2C, this set of obtained sequences is referred to as Library Leu).

In both analyses, high- and low-affinity interactions were identified by comparing measured β-gal activities to a canonical leucine zipper sequence. Within the first analysis, 60 high-affinity clones were detected in total (see Appendix, Table 9.1), while the second analysis resulted in 52 high-affinity sequences (see Appendix, Table 9.2). The sequence composition of high-affinity transmembrane domains was then compared to low-affinity sequences (42 and 22 in the first and second analysis, respectively) in order to detect amino acids and combinations thereof significantly enriched in high-affinity sequences.

Identification of the FxxGxxxG motif

Inspection of the 60 high-affinity sequences detected from Library Ala within the first analysis (Appendix, Table 9.1) revealed several major results (for details see [136]). First, Phe was found to be significantly enriched in high-affinity sequences (3.5-fold enrichment compared to low-affinity sequences, $p<0.001$). Additionally, Gly was slightly more common in high- than low-affinity sequences by a factor of 1.5 resulting again in a significant enrichment ($p<0.001$). Differentiating amino acid frequencies according to positions within the helix interface, Phe was observed most often at position 5 while Gly was preferentially located at positions 8 and 12 leading to frequent formation of the motif FxxGxxxG. In total, this motif was found in nearly 42% of all high-affinity sequences in contrast to only one occurrence in low-affinity sequences.

Site-directed mutagenesis confirmed the importance of Phe at position 5 since β-gal activity decreased remarkably after mutation of this residue to Leu. Furthermore, self-interaction of glycophorin A could be improved by replacing Ile76 with Phe thereby generating the pattern FxxGxxxG together with the naturally present GxxxG motif (further details regarding these experiments are presented in [136]). Based on these results, the sequence motif FxxGxxxG clearly seems to be a potent mediator of helix-helix interactions. However, sequence analysis of bitopic membrane proteins is still required to prove its relevance within naturally occurring transmembrane domains.

Identification of sequence motifs containing histidine

Within the second analysis based on Library Leu covering a broader range amino acids, identified high-affinity sequences varied remarkably from the first analysis (Appendix,

Table 9.2). Here, the most significantly overrepresented amino acid was found to be histidine with a 5-fold enrichment in high-affine sequences ($p<0.001$). Less clearly, but still significantly enriched were Trp and Tyr ($\sim$3-fold enrichment, $p<0.005$ and $p<0.0005$, respectively). Position specific enrichment on the other hand was only detected for His and Gly with His dominating at position 6 and Gly preferentially found at position 13 forming a GxxxG pattern with the first position of the C-terminal vector sequence. These residues form the pattern HxxxxxxGxxxG which was found in ten out of all 52 high-affinity sequences. Furthermore, 18 high-affinity sequences were detected containing histidine at position 6 and Gly, Ser and/or Thr at positions 2, 5, and/or 8 suggesting that histidine can either promote helix interactions via the GxxxG motif or by hydrogen-bond formation with polar side-chains or the backbone of glycine residues.

Again, the importance of His for self-association of transmembrane domains was confirmed by site-directed mutagenesis. Additionally, interaction strength was found to be distinctly improved by either the additional presence of hydrogen bond-forming residues or a C-terminal GxxxG motif as suspected from the analysis of high-affinity transmembrane domains (for further details see [137]).

Identification of sequence motifs consisting of charged residues

In addition to the overrepresentation of histidine, the 52 high-affinity sequences identified from Library Leu showed also a significant overrepresentation of pairs of charged amino acids. Individually, these residues were not detected to be more common in high-affine than in low-affine sequences, positively charged amino acids were even slightly less frequent. However, eleven high-affinity sequences contained at least two charged amino acids with six of these sequences covering even three or four charged amino acids (Appendix, Table 9.2). Notably, in all eleven sequences amino acids of opposite charge were present as well as a GxxxG motif. Given the frequencies of single charged amino acid in the 52 high-affinity sequences, this observation is significant with a p-value of 6.4E-4.

Additional mutational analyses proved that neutral and polar amino acids are not able to replace any of the charged positions without reduced interaction strength although single ionizable side-chains of either charge can also promote helix interaction, yet to a less distinct degree. At the same time, the GxxxG motif seems to be required for high-affine interaction between oppositely charged amino acids while polar residues can further enhance the interaction ([138], manuscript *submitted*).

2.2 Sequence analysis of naturally occurring membrane proteins

To complement the screening for motifs promoting high-affine helix interactions with the ToxR/POSSYCAT system, a database of naturally occurring bitopic transmembrane domains (consisting of one transmembrane helix) was compiled. From this database, membrane protein sequences were identified containing those motifs earlier detected experimentally.

Furthermore, the enrichment of individual motifs within the database was probed using the TMSTAT formalism introduced in the year 2000 by Senes and co-workers for the statistical analysis of amino acid motifs within transmembrane sequences [74]. Applied to a database of 13,606 transmembrane helices originating from both bitopic and polytopic membrane proteins, Senes and colleagues were able to prove that GxxxG is in fact the most strongly overrepresented motif in transmembrane sequences. Here, TMSTAT was applied to the database of solely bitopic proteins in order to match more closely the experimental setup of the ToxR/POSSYCAT system where single transmembrane helices were tested for self-association. As the number of available sequences continues to increase exponentially, enough non-redundant membrane proteins were meanwhile available in public databases to make such an analysis feasible despite the reduction on only a fraction of all membrane proteins.

2.2.1 Materials and methods

Non-redundant database of bitopic membrane proteins

Protein sequences for the analysis of naturally occurring bitopic membrane proteins were obtained from the UniProt Knowledgebase consisting of the intensively annotated Swiss-Prot database and the computer-annotated TrEMBL database [144]. For every independent analysis the latest UniProt release was considered as denoted in Table 2.1.

To select only bitopic membrane proteins, all sequences containing one TRANSMEM annotation in the FT field were extracted from the Swiss-Prot dataset. From the TrEMBL dataset, bitopic proteins were identified using topology prediction programs. For the first analysis addressing the occurrence of FxxGxxxG and related motifs, precalculated TMHMM [97] and SignalP [145] annotations were obtained using the SIMAP database [146]. To exclude mispredicted transmembrane domains, the TMHMM and SignalP predictions were compared for every protein and all transmembrane sequences

overlapping by at least eight amino acids with a predicted signal peptide were eliminated. For the second analysis aiming at the role of histidine for helix association, Phobius predictions [98] were additionally obtained from SIMAP and only proteins containing one predicted transmembrane segment according to TMHMM and Phobius were considered. For the final analysis addressing charged amino acids, only Phobius predictions were used for the identification of bitopic membrane proteins. In any case, all TrEMBL proteins containing one predicted transmembrane segment were combined with the extracted proteins from the Swiss-Prot database resulting in an initial dataset of redundant bitopic proteins (Table 2.1).

Table 2.1: Database of bitopic membrane proteins. The database was updated for each executed analysis to contain the latest content of both Swiss-Prot and TrEMBL.

Analysis	Swiss-Prot release	TrEMBL release	Redundant[a]	Non-redundant[b]
Phe[c]	52.0 (updates until June 12, 2007)	35.0	167,125	19,854
His[d]	55.4	38.0	204,449	20,342
Charge[e]	56.3	39.3	471,336	25,558

[a] Redundant: total number of proteins in the database.
[b] Non-redundant: number of non-redundant proteins in the database.
[c] Phe: analysis of FxxGxxxG and related motifs.
[d] His: analysis of histidine containing motifs.
[e] Charge: analysis of motifs containing one or more charged residues.

To remove sequence redundancy from the initial dataset, the filtering procedure originally introduced by Senes and colleagues for the TMSTAT approach was adapted [74]. First, all transmembrane segments were extended or shortened to a common length of 30 residues. Homologous transmembrane domains were then removed from the dataset by comparing any two sequences in all possible frame shifts using a PAM 100 matrix obtained specifically for transmembrane proteins [47] with a maximal similarity score of 50 or higher being the threshold for identifying homologous transmembrane segments. Proteins obtained from the Swiss-Prot database were kept with higher priority than proteins originating from the TrEMBL dataset. Among Swiss-Prot proteins, transmembrane sequences marked as POTENTIAL, PROBABLE or POSSIBLE were removed in case homology was detected to transmembrane segments either annotated with BY SIMILARITY or without annotation. Additionally, sequences considered to be too hydrophilic (indicated by a hydrophobicity score according to the GES scale [147] < 15) and low-complexity sequences with one amino acid constituting more than half or two amino acids constituting more than two-thirds of the sequence were also excluded from

the analysis, resulting in a final database of non-redundant bitopic transmembrane segments (Table 2.1).

Sequence analysis of bitopic membrane proteins

While the database of bitopic membrane proteins was also used for the identification of naturally occurring transmembrane domains with a certain sequence motif, its main purpose was the statistical evaluation of motif occurrences within a large set of sequences.

If possible, the TMSTAT formalism was applied for this evaluation which explicitly models finite sequence length effects as observed for transmembrane helices and calculates the expected occurrence of a pair or triplet motif by taking into account individual sequence compositions rather than the overall amino acid composition of the database [74]. For every sequence motif under analysis, the algorithm calculates an expectancy distribution describing the probability of finding this motif a certain number of times after randomly permutating each sequence in the database. Briefly, this is done within two steps. First, the probability of observing a pair or triplet motif (XYk or $XYZk_1k_2$) within a *single* sequence a specified number of times is pre-calculated which is dependent on the sequence length l, the sequence distance of motif amino acids k (k_1 and k_2 for triplet motifs) and the occurrence of these amino acids within the analyzed sequence (N_X and N_Y, additionally N_Z for triplets). Within the second step, the full probability distribution $P_{DB}(N_{XYk})$ or $P_{DB}(N_{XYZk_1k_2})$ for this motif and the given database is iteratively calculated from the single sequence probabilities following the recursive equations

$$P_{DB(N)}(N_{XYk}) = \sum_{i=0}^{N_{XYk}} P_{DB(N-1)}(i) P_n(N_{XYk} - 1|l, k, N_{X,n}, N_{Y,n}) \qquad (2.1)$$

or

$$P_{DB(N)}(N_{XYZk_1k_2}) = \sum_{i=0}^{N_{XYZk_1k_2}} P_{DB(N-1)}(i) P_n(N_{XYk_1k_2} - 1|l, k_1, k_2, N_{X,n}, N_{Y,n}, N_{Z,n}) \qquad (2.2)$$

with $P_n(N_{XYk} - 1|l, k, N_{X,n}, N_{Y,n})$ and $P_n(N_{XYk_1k_2} - 1|l, k_1, k_2, N_{X,n}, N_{Y,n}, N_{Z,n})$ corresponding to the precalculated single sequence probabilities for sequence n of the database.

Following the procedure used by Senes et al. [74], the statistical analysis was limited to the most hydrophobic window of all transmembrane domains consisting of either 18 amino acids (analysis of phenylalanine containing motifs) or 23 amino acids (analysis of

histidine motifs and motifs with charged amino acids) which were determined using the GES scale [147]. Slightly longer sequences were used for the latter analyses in order to account for the polar nature of the analyzed residues. For every motif evaluated with the TMSTAT algorithm, the observed occurrence was counted from the non-redundant sequence database and compared to the expected occurrence obtained from the calculated probability distribution:

$$\bar{N}_{XYk} = \sum_{N_{XYk}} N_{XYk} P_{DB}(N_{XYk}) \tag{2.3}$$

The statistical significance of the difference between observed and expected occurrence was assessed using the two-tailed integral of the probability distribution.

In case transmembrane domains consisting of more than 23 amino acids had to be considered or motifs consisting of more than three amino acids or with amino acid spacings of more than five positions were analyzed, the TMSTAT approach was not applicable. In these cases, the expected number of occurrence for a given motif was calculated using a simplified approach relying on average database amino acid frequencies instead of single sequence analysis. To this end, the probability of observing the analyzed motif within a sequence of predefined length was calculated from the individual amino acid frequencies of all contributing motif residues and the number of positions the motif can occur in the given sequence considering the length of both sequence and motif. Multiplying this probability with the number of sequences within the database resulted in the expected occurrence of this motif.

2.2.2 Results

Depending on the previous experiments, different database analyses were executed for identified interaction motifs. In the following, main results of these analyses will be presented starting with the strong candidate motif FxxGxxxG, followed by His-containing motifs and finishing with motifs consisting of one or more charged residues.

Occurrence of FxxGxxxG and related motifs

In order to examine the potential relevance of the FxxGxxxG motif for self-interaction of natural transmembrane domains, its frequency of occurrence was analyzed in comparison to that of related motifs in 19,854 non-redundant bitopic membrane proteins. In total, 2394 sequences from this dataset originated from the intensively annotated Swiss-Prot database while the remaining sequences were obtained from the computer-annotated

TrEMBL database [144]. From this dataset of bitopic protein transmembrane sequences, the abundances of the GxxxG motif and of various F/GxxxG triplets were extracted and compared to their expected occurrences as determined using the TMSTAT formalism [74].

Initially, to test the quality of the obtained dataset and to ascertain the validity of subsequent analyses, the occurrence of the GxxxG motif was calculated which was previously shown to be significantly overrepresented in transmembrane helices of naturally occurring membrane proteins [45, 74]. In total, 12.4% of all sequences contained this motif which is in perfect agreement with the results reported by Senes et al. based on their Swiss-Prot dataset (12.5%) [74]. The overrepresentation of GxxxG in the dataset was highly significant with a p-value of 3.32E-23.

In the second step, frequencies of different triplet motifs containing a Phe residue placed within one helical turn either N- or C-terminal of a GxxxG motif were compared (Table 2.2). Notably, the FxxGxxxG motif was found to be the most overrepresented one of all these motifs, occurring 42% more often than expected (ratio observed/expected = 1.42). This observation is highly significant with a p-value of 1.08E-6. In total, the FxxGxxxG motif was detected 210 times in transmembrane domains of 207 different proteins corresponding to 1% of all sequences in the database. Only the motif GxxxGxF is overrepresented roughly at the same level with a ratio observed/expected of 1.38 (p = 4.38E-6). Some other motifs appeared to be less frequent, yet still significantly overrepresented (GxxFG and GxxxGF). The remaining motifs were found to be either slightly overrepresented or to occur as often as expected.

Additionally, the occurrence of triplet motifs where Phe at the -3 position of GxxxG is replaced by Trp or Tyr was analysed (Table 2.2). In both cases, the observed occurrence was clearly lower than expected (YxxGxxxG, ratio observed/expected = 0.90; WxxGxxxG, ratio observed/expected = 0.80). However, due to the low abundance of these motifs, these underrepresentations are statistically not significant. These results show that the motif FxxGxxxG is significantly overrepresented in naturally occurring transmembrane domains of single-span membrane proteins which suggests its function as interaction motif. The statistical analysis additionally underlines the specific role of Phe compared to other aromatic amino acids.

Based on these results, further experiments were conducted by Stephanie Unterreitmeier. First, one of the 207 bitopic membrane proteins containing a FxxGxxxG motif, the vesicular stomatitis virus G protein (VSV-G), was selected and tested for self-association using the ToxR/POSSYCAT system. After demonstrating its principal

Table 2.2: Statistical analysis of triplet motif frequencies containing GxxxG and one aromatic residue (Phe, Trp, Tyr) at different spacings in a non-redundant database of bitopic membrane proteins. Expected numbers of occurrence and significances were calculated using the TMSTAT approach. From all tested motifs, the motif FxxGxxxG was most strongly overrepresented.

Triplet	Occurrence	Expectation	Odds ratio[a]	Significance (p)
FxxxGxxxG	134	134	1.00	1
FxxGxxxG	210	148	1.42	1.08E-6
FxGxxxG	172	161	1.07	0.411
FGxxxG	221	175	1.26	0.001
GFxxG	211	188	1.12	0.100
GxFxG	204	188	1.09	0.255
GxxFG	243	188	1.29	9.59E-5
GxxxGF	226	175	1.29	1.61E-4
GxxxGxF	222	161	1.38	4.30E-6
GxxxGxxF	185	148	1.25	0.003
GxxxGxxxF	159	134	1.19	0.038
WxxGxxxG	32	40	0.80	0.219
YxxGxxxG	32	36	0.90	0.640

[a] Odds ratio: ratio of observed occurrence divided by the expected occurrence.

potential for significant interaction, further mutational analyses were executed to prove that self-interaction is strongly affected by all residues forming the FxxGxxxG motif. Additionally, the dependence of self-association on the spacing between Phe and the GxxxG motif was examined experimentally. It could be shown that FxxGxxxG is the only motif promoting strong helix association while Phe positioned at a different spacing with respect to GxxxG seems to have no positive effect. Furthermore, the replacement of Phe by another aromatic residues also leads to clearly reduced interaction strength confirming the singularity of the interaction motif FxxGxxxG (for further details refer to [136]).

Occurrence of sequence motifs containing histidine

The biological relevance of His-containing motifs was again analyzed using natural bitopic membrane proteins, however with less clear results as in the case of the FxxGxxxG motif. The observed occurrence of residue pairs consisting of His and either Gly, Ser or Thr that were found to increase self-association experimentally ([G/S/T]xxxH, [G/S/T]H, Hx[G/S/T]) was generally not significantly higher than expected with the TMSTAT approach. Only the motif SxxxH was significantly overrepresented with a

ratio observed/expected of 1.20 (p = 0.004). Similarly, the motif HxxxxxxGxxxG was detected in 12 proteins closely resembling the expected occurrence of 11.5 times. Generally, the TMSTAT analysis of His-containing motifs is strongly limited by the lack of histidine residues within transmembrane helices (<1% in the bitopic dataset) leading to small absolute numbers and hence insignificant results.

However, several membrane proteins could be identified that carry even more complex His-containing motifs consisting of the basic HxxxxxxGxxxG motif in combination with an additional Gly, Ser or Thr residue. Using only the most hydrophobic 23 amino acid stretch of each transmembrane domain, in total five non-redundant sequences with such motifs could be found. Extending the analysis to transmembrane domains elongated to 30 amino acids to account also for flanking regions possibly enriched in polar amino acids, 19 hits were identified. Interestingly, motifs of the type [G/S/T]HxxxxxxGxxxG were most frequent which is in line with the experimental observation that this motif (containing Thr) was found to promote stronger self-interaction than the other tested motifs. While a large fraction of all detected proteins with a His-containing motif were found to be still uncharacterized, several were also functionally annotated such as a number of BNIP3 homologs carrying the motifs THxxxxxxGxxxG and SHxxxxxxGxxxG, the probable ubiquinone biosynthesis protein ubiB (motif HxTxxxxGxxxG) and a N-acetylmuramoyl-L-alanine amidase (motif HxSxxxxGxxxG). Subsequent experimental confirmation of self-association of a 16 residue BNIP3 construct demonstrated that natural bitopic membrane proteins contain variants of His-containing motifs that are likely to self-interact and induce oligomerisation although the enrichment of such motifs in naturally occurring sequences can not be shown significantly.

Occurrence of sequence motifs containing charged residues

Prompted by the enrichment of oppositely charged residues within high-affinity sequences of the second library screen (Library Leu), natural bitopic membrane proteins were analyzed for the occurrence of charged amino acids in combination with GxxxG motifs. Searching for single charged residues placed up to eleven positions up- or downstream of a GxxxG motif resulted in several motifs being significantly enriched in naturally occurring transmembrane domains despite the general low frequency (<2%) of charged amino acids in transmembrane domains (Figure 2.3, Table 2.3). Most strongly overrepresented were the motifs KxxxxxxxxxGxxxG and DxxxxxxxxxGxxxG with ratios observed/expected of 2.46 and 2.54, respectively. In total, 1179 non-redundant transmembrane domains were identified containing GxxxG and one or more charged

residues corresponding to 4.6% of all non-redundant sequences or 37% of all sequences with a GxxxG motif.

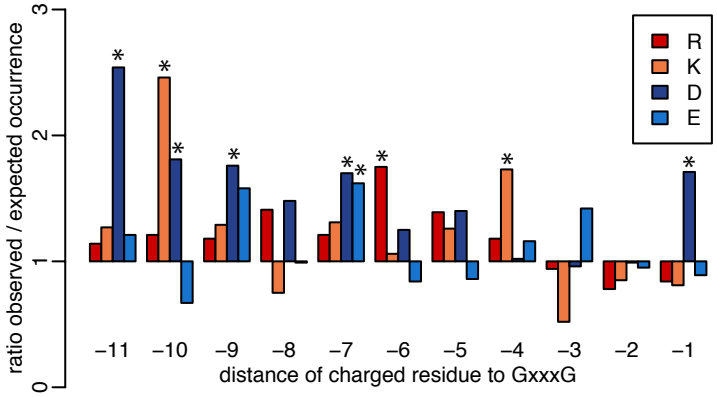

Figure 2.3: Enrichment of motifs consisting of GxxxG and a charged amino acid N-terminal of GxxxG in naturally occurring bitopic membrane proteins. Shown is the relative position of the charged amino acid with respect to the GxxxG motif. From all tested motifs, nine were found to be significantly overrepresented within bitopic membrane proteins (indicated by *).

Table 2.3: Significantly overrepresented motifs consisting of GxxxG and a charged amino acid. In total, 11 motifs significantly enriched in bitopic membrane proteins were detected.

Triplet	Occurrence	Expectation	Odds ratio[a]	Significance (p)
RxxxxxGxxxG	25	14.3	1.75	0.0061
KxxxxxxxxxGxxxG	24	9.8	2.46	0.0001
KxxxGxxxG	25	14.5	1.73	0.0071
GxxxGxK	24	16.4	1.46	0.0451
GxxxGxxxxK	25	13.5	1.85	0.0031
DxxxxxxxxxxGxxxG	15	5.9	2.54	0.0012
DxxxxxxxxxGxxxG	12	6.6	1.81	0.0379
DxxxxxxxxGxxxG	13	7.4	1.76	0.0375
DxxxxxxGxxxG	15	8.9	1.70	0.0360
DGxxxG	22	12.9	1.71	0.0120
ExxxxxxGxxxG	16	9.9	1.62	0.0440

[a] Odds ratio: ratio of observed occurrence divided by the expected occurrence.

All sequences containing GxxxG and a charged amino acid either C- or N-terminal of the GxxxG motif were further analyzed for the presence of polar residues (Cys, Ser, Thr, His, Asn, Gln, Tyr, Gly). To this end, the frequency of polar residues at a specific

distance to the GxxxG motif was counted within all sequences containing a charged amino acid at another position and compared to the expected frequency as obtained from all sequences having a GxxxG motif. In total, 220 different motifs consisting of a charged amino acid, a polar amino acid and the GxxxG motif were tested. Thereof, 92 motifs were enriched within the subset of sequences with charged amino acid, 24 motifs even significantly with $p<0.05$.

Addressing sequence motifs consisting of two charged residues and a GxxxG motif as obtained from the experimental screen, in total 91 transmembrane domains could be detected containing such a motif within the non-redundant dataset of 25,558 sequences. From these domains, 42 had both charged residues placed N-terminal of the GxxxG motif and 24 contained two oppositely charged residues. Apart from several uncharacterized proteins, several functionally annotated proteins were detected among these sequences such as a lipid A biosynthesis lauroyl acyltransferase, a subunit of NADH dehydrogenase and of cytochrome b5. Charged residues were therefore not only found to form several significantly overrepresented motifs with GxxxG, but several natural bitopic membrane proteins even contain motifs with multiple charged residues likely to be important not only for membrane protein function but also for the structure of transmembrane domains as indicated by ToxR/POSSYCAT experiments.

2.2.3 Discussion

Genetic screening tools such as the ToxR/POSSYCAT system are well suited to identify transmembrane sequences with high potential for self-association. Site-specific mutation analyses can further evaluate the contribution of individual amino acids to helix assembly thereby revealing minimal sequence motifs sufficient for successful helix interaction. However, these experiments can not secure that these motifs are indeed biologically relevant and not only an evolutionary possibility. To this end, database analyses of naturally occurring membrane proteins are a necessary step in the analysis of helix interaction motifs.

Here, several sequence motifs were presented that could be shown to promote high-affine helix interaction within the ToxR/POSSYCAT system and whose biological significance was further secured by analysis of bitopic membrane protein sequences. All motifs are variants of the GxxxG motif which is known to promote strong helix interaction [72, 73] although energetic measurements suggested that sequence context may strongly modulate interaction strength [77, 131]. The identification of specific amino acids that stabilize GxxxG-mediated helix interaction is therefore an important step in

the understanding of transmembrane helix association. Interestingly, these amino acids were found to be highly diverse ranging from aromatic phenylalanine to single charged residues or combinations of oppositely charged side-chains, which is in line with previous results suggesting that transmembrane helix interfaces are in fact more diverse than those of soluble proteins [60, 61].

The FxxGxxxG motif is frequently found in bitopic membrane domains

Optimally, sequence analysis and experiments complement each other as observed in the case of the sequence motif FxxGxxxG. This motif, which was experimentally found to promote high-affine helix interaction, was also the most significantly enriched arrangement of Phe and GxxxG in a dataset of non-redundant bitopic membrane proteins. Mutational analysis of one example protein found to contain the FxxGxxxG motif (the viral fusion protein VSV-G) further confirmed that Phe as well as the GxxxG motif are essential for helix interaction in this case. In contrast, similar motifs replacing Phe with other aromatic residues (WxxGxxxG and YxxGxxxG) were found to occur less frequently than expected in natural bitopic transmembrane domains consistent with the experimental observation that both motifs can not promote helix self-interaction. Mechanistically, this suggests that only Phe residues are appropriately sized for close helix association and/or can be properly oriented to enter aromatic π-π interactions or form a hydrogen bond between the Cα-H of Gly and the Phe side-chain.

Naturally, sequence analysis will always reveal motifs significantly overrepresented yet not found to self-interact within the ToxR/POSSYCAT system such as a number of additional motifs containing Phe and GxxxG (Table 2.2). First, overrepresentation may arise not only due to structural but also due to functional reasons. Additionally, within a structural context motifs may also be responsible for heterotypic helix interactions or may require additional residues not included within a specific sequence library. In the case of Phe, previous analyses of helix interactions of polytopic membrane proteins have confirmed the importance of Phe for heterotypic helix interactions [60, 148] suggesting that this might also be the reason causing the enrichment of other F/GxxxG motifs in bitopic transmembrane domains.

His-containing motifs promote helix interaction in artificial and natural transmembrane domains

The imidazole side-chain of histidine has principally both acidic and basic properties making it a prominent participant in intra- and inter-protein hydrogen bonds. Accord-

ing to an analysis of membrane protein structures, more than one-third of all inter-helical hydrogen bonds are formed by either His, Ser or Thr [130]. Using the ToxR/POSSYCAT system, the role of histidine in high-affine helix interactions was now further evaluated. It could be shown that helix interactions mediated by His are strongly dependent on sequence context requiring either the additional presence of polar residues (Ser/Thr/Gly) or a GxxxG motif within the interface. While polar side-chains may serve as potential hydrogen bond partners, the GxxxG motif seems to stabilize helix interaction by properly orienting the histidine side-chains.

Analyzing transmembrane domains of naturally occurring bitopic membrane domains, several examples containing the motif HxxxxxxGxxxG could be identified. However, as histidine is rarely found within transmembrane regions (<1% overall frequency), no statistical significant enrichment of this or related motifs containing His could be observed. Still, histidine is able to promote strong helix interaction within natural transmembrane domains via polar residues and a GxxxG motif as illustrated by the protein BNIP3, one of the examples identified from the database of bitopic proteins. BNIP3 is known to form a homodimer with the motif SHxxAxxxGxxxG forming the interaction interface as discovered using mutagenesis studies and NMR spectroscopy [149, 150]. Thereby, histidine was shown to form multiple hydrogen bonds with the neighboring serine residue. Although other interaction motifs such as the previously presented FxxGxxxG may therefore be easier approachable by evolution leading to more significant overrepresentation within natural membrane proteins, the example of BNIP3 nevertheless clearly demonstrates the biological relevance of the less numerous His-containing motifs such as HxxxxxxGxxxG and variants.

Charge-charge interactions require stabilization via a GxxxG motif

In addition to the detailed analysis of histidine in helix interaction interfaces, the role of charged side-chains and accordingly ionic interactions for helix association was separately addressed with mutation experiments and sequence analysis of bitopic membrane proteins. In contrast to an earlier analysis, which suggested that charged amino acids per se are not beneficial for transmembrane helix interaction [151], it was observed that oppositely charged amino acids are enriched in high-affine transmembrane domains and that heterotypic helix interaction is enhanced by the incorporation of such oppositely charged residues. However, all selected high-affine domains additionally contained a GxxxG motif and subsequent experiments proved that this motif is essential for interaction with pairs of charged amino acids alone not being able to promote strong helix

association. Hence, GxxxG seems to be required for bringing charged residues in appropriate positions and orienting them for proper ionic interactions, similarly as observed in case of the motif FxxGxxxG and also the His-containing motifs.

The importance of the motif GxxxG for ionic interactions was further confirmed by the observation that sequence motifs consisting of a single charged amino acids and GxxxG are commonly enriched in natural bitopic membrane domains (several motifs even significantly). Sequences containing several charged amino acids such as inferred from experimentally selected high-affine transmembrane domains were also found, however less commonly, which is not surprising given the general low frequency of charged amino acids in transmembrane domains (<1% for Arg, Lys, Asp, Glu, respectively). Motifs consisting of a charged residue, a polar residue and GxxxG on the other hand could again be found more commonly as expected, which is in agreement with results gained from the executed mutation experiments indicating that helix interactions promoted by charged residues can be further enhanced by the presence of polar side-chains.

Importantly, significantly overrepresented motifs containing GxxxG together with a charged amino acid were further tested experimentally for their capability to interact in heterotypic fashion and one pair of motifs containing either Asp six position or Lys nine positions N-terminal of GxxxG were in fact found to interact successfully. While several motifs might be still overrepresented due to other reasons (charged amino acids for example might be functionally relevant while the GxxxG motif alone could be required for helix association), ionic interactions between oppositely charged amino acids are clearly one of the possibilities of natural membrane proteins to achieve high-affine helix association if they are stabilized via a GxxxG motif.

3

Co-evolving residues in membrane proteins

In contrast to the preceding chapter which presented results gained from a combination of experimental and computational analyses, all following chapters will concentrate now on the fully computational analysis of helix-helix contacts and helix interactions in membrane proteins.

First, the prediction of residue contacts within transmembrane regions of membrane proteins will be addressed within the current and the subsequent chapter. Available approaches to predict residues participating in helix-helix contacts are generally based on the idea of identifying membrane-exposed and buried residues [122, 152, 153]. However, pairs of contacting residues can not be predicted with these methods. Methods dealing with the pairwise prediction of residue contacts specifically within transmembrane portions of membrane proteins are so far still lacking prompting the evaluation of possible routes in this direction.

In order to maintain protein function, mutations which tend to destabilize a particular protein structure may provoke other positions to mutate concurrently in order to compensate for the loss of stability. Amino acid contacts have been suggested to be primary spots of these compensatory processes, making the detection of sequence positions with correlated mutational behavior an important feature for residue contact prediction methods.

While first examples of such compensatory mutational changes were described by analyzing individual families with solved structure [154, 155], several large scale analyses have been executed since then (for example [156, 157, 158, 159]). However, most studies on co-evolving residues so far were conducted on soluble proteins. Membrane proteins were considered only in few individual case studies [160, 161, 121, 120]. Due to the

paucity of three-dimensional structures, a general analysis of co-evolving residues in a large non-redundant dataset of membrane proteins is still missing. With the accumulation of more membrane protein structures in recent years, this analysis has now become feasible.

Within the following chapter, the first large-scale analysis of co-evolving residues in membrane proteins is presented and their potential for the identification of helix-helix contacts is evaluated. First, the introduction will shortly summarize preceding work on co-evolving residues including their current status within contact prediction approaches for soluble proteins. The remaining sections will then introduce the dataset of membrane protein structures used for evaluation purposes, describe execution and main results of all performed analyses and will discuss possible applications for co-evolving residues in membrane proteins based on these results. Furthermore, a new consensus prediction method for correlated mutations in membrane proteins called *HelixCorr* will be introduced which improves the prediction accuracies obtained by individual prediction algorithms.

All main results of this chapter were published in [162].

3.1 Introduction

3.1.1 Detection of co-evolving residues

While the first approach to detect co-evolving residues in a multiple sequence alignment was published already in 1994 [163], a variety of additional detection algorithms have been reported since then. The strategy most commonly used relies on the calculation of a Pearson correlation coefficient to detect alignment positions with similar patterns of amino acid change [163, 164, 165, 156, 166, 167, 168]. Other prediction algorithms try to detect significant co-evolution based on a chi-square goodness-of-fit test comparing the observed co-occurrence of two residues with their expected co-occurrence [169, 170, 171], by using a maximum likelihood approach [172] or through the application of information theory [173, 174, 175]. An alternative approach is constituted by perturbation-based methods such as SCA (Statistical Coupling Analysis) [176, 177, 178] where co-evolution of residues is identified by the analysis of statistical coupling of amino acid distributions. Subalignments having a changed amino acid distribution at certain positions are used to evaluate the effect of this perturbation on the residue compositions at other positions of the alignment.

Common to all prediction algorithms is the generally high number of false positive predictions caused by random noise or misleading phylogenetic signals. Accordingly, several authors have proposed means to reduce false positives by applying special filtering steps [168, 179] or including phylogenetic information about the analyzed sequences in the prediction process [180, 181, 121, 171, 182]. Lately, especially information theory based approaches have been significantly enhanced by filtering procedures addressing background noise [183, 184] giving hope that other currently used methods may also still offer room for further improvement.

3.1.2 Residue contact prediction using co-evolving residues

Originally, most approaches for detecting co-evolving residues were developed with the goal to use these residues for predicting residue contact pairs. Following common practice as performed also in recent CASP experiments [185, 186], obtained contact predictions are generally evaluated and compared by providing contact prediction accuracies corresponding to the fraction of correctly predicted contacts out of all correlated residues found. Several authors calculate also the completeness of the prediction (fraction of correctly predicted contacts out of all real contacts). As both prediction accuracy and completeness depend strongly on the number of predicted contacts, this number is most often selected not dependent on the obtained correlation score but dependent on the length of the protein under analysis to make different predictions better comparable to each other.

Following two independent comparative studies [158, 159], individual prediction methods can be clearly ranked in their ability of predicting residue contacts via the detection of mutationally correlated positions. In both analyses, methods based on Pearson correlation coefficients or using a chi-square goodness-of-fit test clearly outperformed perturbation-based methods or approaches using information theory. However, recent publications have suggested that the reduction of background noise may improve the contact prediction of the latter methods to a level at least equal or even better than the best methods available so far [183, 184]. Independent of the prediction method used, several studies consistently reported that decreasing prediction accuracies can be expected with increasing protein size [163, 156], while alignment size positively correlates with prediction accuracy [175, 187]. Evaluating contact prediction performance on different structural classes (all-β, all-α, α+β, α/β) resulted in high contact prediction accuracies for proteins with mixed secondary structure (α+β, α/β), while the predictive accuracy for all-α proteins was clearly reduced compared to the average accuracy [187, 188].

Despite all progress in the development of new methods, prediction accuracies for structural contacts in globular proteins hardly exceeded 20% with any known method based on co-evolving residues alone (Fodor and Aldrich, 2004), strongly limiting the practical utility of the predicted contacts as structural constraints in *ab initio* structure prediction. While some authors have explained these low contact prediction accuracies with the difficulty of differentiating correlation signal from random noise [172, 165], recent studies indicate that co-evolution of amino acids in fact may originate not only from structural contacts but from a much broader range of biological reasons motivating a couple of other fields of application (see below). Currently best performing contact predictors for soluble proteins are therefore also not based on co-evolving residues alone, but mostly use sequence co-evolution together with other sequence features as input for machine-learning approaches (see section 4.1 as well as [189, 190, 191]).

3.1.3 Other applications for co-evolving residues

In addition to residue contact prediction, co-evolving residues have been employed for a number of other fields of application. When using them as constraints in *ab initio* folding simulations, the global fold of 20 non-homologous proteins with less than 100 amino acids could be successfully predicted within a root-mean-square deviation (RMSD) between 3.0Å and 6.5Å [192, 193]. Furthermore, they have been used in fold recognition experiments [157] and were also found to be helpful in detecting both interacting proteins and interaction regions between two proteins [161, 194, 195]. In 2005, Ranganathan et al. published results which demonstrated that their method called Statistical Coupling Analysis (SCA) was able to detect correlation rules in the WW domain which describe aspects of the fold architecture rather than simple protein contacts. They introduced the concept of a fold correlation backbone which they claimed was nearly sufficient to describe the structural architecture of a protein without additional information [196]. They impressively demonstrated the power of this idea by synthesizing artificial WW domains solely based on the previously derived correlation model of which a substantial percentage was able to fold into functional WW domains *in vitro* [197].

In addition, further contributions have demonstrated that correlated mutations may also occur due to reasons related to protein function. Gloor et al. analyzed 12 mutations of the ATP synthase ϵ subunit and 7 missense mutations of the homeodomain coming to the conclusion that co-evolving residues mutating concurrently with several other residues are more likely to be functional sites than structural contacts [174]. Within a study on the Hsp70-Hop-Hsp90 system, regions previously known to be functionally

important could be identified based on residue co-evolution [198]. Additionally, the authors pointed out that co-evolving amino acids were often found to be in close proximity to functionally important sites. Similar results were obtained in an analysis of correlated mutations within the cytochrome c oxidase subunit I where many co-evolving residues were found adjacent to hypothesized proton pumping channels [199]. In a recent study, Lee et al. provided further evidence for the hypothesis that correlated mutation may be related to functional importance in an analysis of 44 selected protein families [200]. Accordingly, residue co-evolution seems to be a phenomenon employed by evolution to gain variety while concurrently conserving structural hot-spots of a protein but similarly it may also be used to secure a proteins ability to interact with other proteins or conserve its core functionality.

3.2 Materials and methods

3.2.1 Membrane protein datasets

Dataset of high-quality alignments

To obtain a dataset of membrane proteins having a solved structure and carefully curated alignments, protein sequences were taken from the first version of the CAMPS database of membrane proteins [36] covering 120 prokaryotic genomes. For all proteins, CAMPS contains transmembrane segment annotations predicted by TMHMM 2.0 [97]. Furthermore, it provides clusters of related sequences at different granularity levels with precalculated and often manually curated cluster alignments.

For the study of co-evolving residues all SC-clusters were extracted from CAMPS. At this clustering level, the generated groups of proteins roughly correspond to structural folds. As co-evolving residues should be predicted specifically within transmembrane domains, conserved transmembrane regions (TMS cores) were extracted from CAMPS and concatenated to form sequences representing only the transmembrane parts of each protein. From the set of pre-aligned TMS sequences all sequences considered inappropriate for the analysis were discarded. Since highly similar sequences might result in few correlations due to a lack of variability, sequences with a pair-wise identity above a preset threshold were considered redundant and removed. Different thresholds were used in individual predictions ranging from 95% pairwise identity down to 50% identity. The thresholds were chosen dependent on the total number of sequences in the alignment to allow for an optimal tradeoff between a minimal number of required sequences and suf-

ficient sequence diversity for a successful prediction. In addition, sequences with 25% or more gaps within at least one TMS were removed. According to these rules, the number of removed sequences varied between 3 in clusters with well-aligned sequences uniformly covering the cluster sequence space and 346 in the case of a cluster with several tightly connected subclusters. The final number of sequences in individual alignments ranged from 20 to 228.

During an initial prediction step, co-evolving residues were predicted for all SC-clusters of CAMPS. In a subsequent selection step clusters found to be suboptimal due to high sequence diversity were either discarded in case the number of valid sequences in the final multiple alignments was below 15 or replaced by sub-clusters with higher similarity among their members. Inappropriate clusters were identified by either an average pair-wise identity of below 15% or an extremely small number of obtained correlations (less than one per TMS). This procedure was repeated until a cluster was either appropriate for the analysis of correlated mutations or had to be removed due to an insufficient number of sequences. In total, starting from 266 SC-clusters currently available in CAMPS 91 optimal clusters were obtained, 14 of which contained a representative structure (Table 3.1, referred to as dataset `MP_14`).

Table 3.1: High-quality dataset `MP_14` used for the prediction of co-evolving residues.

Protein description	PDB	Chain	TMS_{pred}[a]	TMS_{exp}[b]	L[c]
Na neurotransmitter symporter (snf family)	2A65	A	11	12	249
Probable ammonium transporter	1XQE	A	10	11	253
AcrB bacterial multidrug efflux transporter	1IWG	A	12	12	262
Succinate dehydrogenase cytochrome B-556 subunit	1NEK	C	4	3	77
Succinate dehydrogenase hydrophobic membrane anchor	1NEK	D	3	3	71
Aquaporin Z	1RC2	A	6	6 (8)[d]	135
Nitrate reductase A γ subunit	1Q16	C	5	5	105
Formate dehydrogenase N	1KQF	C	5	4	108
Vitamin B12 transport system permease protein	1L7V	A	7	10	172
Glycerol-3-phosphate transporter	1PW4	A	13	12	262
Mechanosensitive channel protein	1MXM	A	3	3	79
ATP synthase subunit A	1C17	M	5 (6)[e]	4	128
Preprotein translocase secY subunit	1RHZ	A	10	10	228
Fumarate reductase cytochrome B subunit	1QLA	C	5	5	117

[a] TMS_{pred}: number of transmembrane segments predicted with TMHMM.
[b] TMS_{exp}: number of transmembrane segments determined from the PDB structure.
[c] L: length of the alignment consisting only of transmembrane segments.
[d] Protein contains two membrane loops which were not considered.
[e] PDB structure covers only five of all six transmembrane segments.

Full dataset of membrane protein structures

While the high-quality dataset was used to test optimal conditions for the prediction of co-evolving residues in membrane proteins, a second larger dataset was compiled covering all non-redundant alpha-helical membrane protein structures included in the Protein Data Bank of Transmembrane Proteins (PDBTM) [201] and the membrane protein structure dataset provided by the Stephen White laboratory at UC Irvine (`http://blanco.biomol.uci.edu/Membrane_proteins_xtal.html`, further referred to as the White dataset) as of September 17, 2007.

Starting with the non-redundant set of PDB chains containing alpha-helical transmembrane segments obtained from the PDBTM, an initial dataset of those proteins was created whose structure was solved by X-ray with a resolution of less than 3.5Å and which contained at least three transmembrane segments according to the PDBTM annotation. Since this initial set consisting of 50 PDB chains was lacking several prominent membrane proteins with solved structures such as rhodopsin, it was subsequently enriched with sequences from the White dataset. To this end, all chains with less than three transmembrane segments in their PDBTM entry were eliminated from the White dataset. Additionally, all sequences with at least 40% sequence identity to another sequence with better resolution (either within the White dataset or in the initial dataset) or with a resolution worse than 4Å were removed. Both the moderate threshold for sequence identity and the relaxed threshold for structural resolution at this step were concessions needed to be made due to the limited number of available membrane protein structures. The remaining 12 sequences were merged with the sequences from the initial dataset to form the final set of 62 protein chains originating from 52 PDB structures (Appendix Table 9.3). This dataset is from now on also referred to as `MP_62`.

Exact transmembrane segment positions and the in/out topology for each protein were obtained from the recently developed TOPDB [202], which contains comprehensive topology information derived both from literature and public databases for a large number of membrane proteins. For two cases (PDB proteins 2UUH chain A and 1ORQ chain C) no entry could be found in TOPDB, therefore transmembrane positions for these proteins were obtained from PDBTM and the in/out topology from OPM [203]. PDBTM summarizes results obtained with the algorithm TMDET [204], which determines the position of transmembrane regions of membrane proteins from their 3D structure, and is itself one of the databases covered by TOPDB.

The final dataset included proteins with three up to thirteen transmembrane segments with close to 25% of all sequences (15 out of 62) containing ten or more transmembrane

segments. Despite the liberal threshold of 40% sequence identity used for the construction of the dataset, the pairwise sequence identity in the final dataset was low, with less than 2.5% of all possible sequence pairs having a sequence identity above 30% and less than 0.5% having a sequence identity above 35%. Considering only the transmembrane parts of each protein, pairwise sequence identities were slightly higher due to the hydrophobic nature of transmembrane segments. Still, less than 5.5% of all protein pairs had a sequence identity of higher than 35%.

Multiple sequence alignments used for the calculation of correlation scores were derived from initial alignments obtained with PSI-BLAST [205] searches against NCBI's unfiltered NR database [206], with three iterations and the inclusion of related database sequences into the profile with an E-value threshold of $1x10^{-4}$. First, all positions were removed from the full length PSI-BLAST alignment which did not correspond to any transmembrane segment of the PDB sequence resulting in an alignment representing only the transmembrane parts of the reference sequence. Following the procedure developed for the dataset `MP_14` (see above), sequences thought to be inappropriate for the prediction of correlated positions were discarded.

3.2.2 Prediction of co-evolving residues

Co-evolving residues were predicted using seven different prediction algorithms: McBASC [156], OMES [169, 158], CORRMUT [121], CAPS [182], MI [174], SCA [176] and ELSC [178]. For the McBASC algorithm two different substitution matrices (the Miyata matrix [207] and the McLachlan matrix [208]) were evaluated and the OMES algorithm was applied in two different versions, as originally introduced by Kass and Horovitz (OMES-KASS) [169] and in its modified version presented by Fodor and Aldrich (OMES-FODOR) [158].

McBASC

For predictions with the McBASC algorithm, the original method of Gobel et al. [163] with its refinements as introduced by Olmea and Valencia [156] was implemented. To select significantly correlated sequence positions a length-dependent threshold was applied by choosing only the number of highest correlated pairs corresponding to one fifth of the protein length (L/5 criterion, only transmembrane regions were considered for determining the protein length as correlations were only predicted for these parts of the protein).

OMES

First, the original version of the algorithm (OMES-KASS) [169] was implemented where the statistical significance of the difference between observed and expected frequencies is calculated using the chi-square goodness-of-fit test. All covariations with p-values of less than 0.001 were considered to be significantly correlated and the L/5 most significant correlations were selected. Additionally, co-evolving residues were predicted with the modified OMES algorithm as provided by A. Fodor (`www.afodor.net`). Based on the calculated correlation scores the L/5 highest correlated residues were selected.

CORRMUT

For predictions using the CORRMUT algorithm [121] a phylogenetic tree of all sequences in each obtained multiple alignment was calculated and ancestor sequences at internal nodes of the tree were reconstructed with the program FASTML (Pupko, et al., 2000). The Miyata matrix [207] was chosen as substitution matrix. The significance of the derived correlation coefficients was estimated by confidence intervals obtained from a bootstrap procedure using a sample size of 400. Correlation coefficients calculated for each sample were used to derive a mean Pearson correlation coefficient (r) as well as the 95% confidence intervals (rlow,rhigh). To identify significantly correlated residue pairs a minimal threshold of 0.4 for the mean correlation coefficient and a minimal rlow-value of 0.05 was applied. Then, the length/5 highest correlated sequence positions were selected. The thresholds for the mean correlation coefficient and the lower confidence boundary were established in preliminary experiments where they were found to permit the best tradeoff between number of detected correlations and prediction accuracy. Although CORRMUT predictions were only obtained for the high-quality dataset, the number of correlations satisfying these thresholds was clearly lower in comparison to other prediction algorithms such as McBASC or OMES (with the exception of one protein, 1PW4 chain A).

CAPS

Predictions with the CAPS algorithm [182] were executed using the provided program (`http://bioinf.gen.tcd.ie/~faresm/software/caps/`) and recommended standard parameters. Again, a minimal correlation coefficient of 0.4 was applied as threshold and the length/5 highest correlated pairs were selected. Similar to the CORRMUT algorithm, predictions were only obtained for the high-quality dataset as the algorithm

seemed to be highly sensitive to alignment quality. Still, predictions for only five proteins could be obtained (1XQE chain A, 1IWG chain A, 1RC2 chain A, 1L7V chain A, and 1RHZ chain A). In all other cases the number of sequences in the multiple alignment (generally less than 50) seemed to be insufficient for this prediction algorithm.

MI

The MI algorithm was implemented as described earlier in comparative studies on correlated mutations [159]. The L/5 pairs with the highest MI score were selected.

SCA / ELSC

Both algorithms were used in the implementation provided by A. Fodor (`www.afodor.net`). The L/5 correlations with the highest score were again chosen in both cases.

3.2.3 Structural validation

Observed distances between residue pairs were extracted by calculating the minimal distance between side chain or backbone atoms of the two residues. Two residues were considered in contact if their minimal distance was less than 5.5Å. The 5.5Å cutoff was chosen as the maximal distance between a pair of heavy (i.e., non-hydrogen) atoms that is indicative of a direct contact; at larger distances, a third atom may fit in between the atom pair. Other studies on correlated mutations have often used a contact definition based on Cβ-distances and a 8Å cutoff. However, due to the regular backbone conformation of alpha-helical membrane proteins a contact criterion incorporating side chain atoms seems to be better suited for the analysis of helix-helix interactions. Nevertheless, contact prediction accuracies based on Cβ-distances and a contact threshold of 8Å were also calculated showing only minor deviations from the presented results (data not shown).

The prediction accuracy (fraction of correctly predicted contacts out of all correlations found) was calculated from the number of predicted contacts and the number of observed contacts considering only those correlated pairs lying on different transmembrane helices. In order to estimate the significance of the obtained prediction, a p-value was calculated describing the enrichment of contact pairs within all co-evolving residues. This was done based on the hypergeometric distribution and the probability to pick a residue pair in contact by random.

Since correlated positions may contain information beyond mere physical contacts between individual residues, two additional quality measures were used to investigate the prediction outcome. First, the harmonic average Xd as introduced by Pazos et al. [194] was used as a measure of relative proximity rather than direct contact. For the calculation of the Xd value the distribution of Cβ-distances between correlated residues was compared to the distribution of distances for all pairs of positions. Distances from both distributions were grouped into bins of 4Å and the difference between the two distributions was calculated for each bin. The differences were weighed with the inverse of the normalized distance of the corresponding bin and were added. When analyzing the results, a value of Xd = 0 indicates no separation between the two distance distribution, while Xd > 0 indicates a shift of correlated residues towards smaller distances. The larger a positive Xd-value the more successful is the corresponding prediction:

$$Xd = \sum_{i=1}^{n} \frac{P_{ic} - P_{ia}}{d_i n} \tag{3.1}$$

where Pic and Pia are the percentages of correlated and all residue pairs with distance between d_i and d_{i-1}, d_i is the upper limit of each bin (normalized to 60) and n is the number of distance bins (15 for the range from 4 to 60Å).

Additionally, a 'δ-analysis' [209] was used to investigate the position of found correlations with respect to observed helix-helix contacts. Within this analysis the fraction of correlations with residues i and j was calculated which have an observed contact between residues in the interval {i-δ,i+δ} and {j-δ,j+δ}. With δ=4 the fraction of correlations where both participating residues lie within one helix turn of residues forming an interhelical contact was detected. Again a p-value was calculated based on the hypergeometric distribution to estimate the significance of the prediction.

3.2.4 Consensus prediction of co-evolving residues in membrane proteins

A consensus prediction method for co-evolving residues within transmembrane regions was developed combining detected co-evolving residues from different prediction methods while concurrently filtering likely false positive predictions. Within the full consensus approach, all methods were considered except SCA and MI as their individual contact prediction performance was found to be clearly inferior compared to all other methods. Additionally, a reduced version of the consensus method was tested

incorporating only predictions of the four best performing methods (McBASC-Miyata, McBASC-McLachlan, OMES-KASS and CAPS).

Within a first prediction step, the consensus method obtains co-evolving residues using all methods considered for the prediction. These residues are then combined and mapped on transmembrane helix pairs. All correlations found on helices with a total amount of correlations less than a predefined threshold N are removed within a second step as these residues are likely to be false positive predictions.

The developed method is available in both versions (full and reduced) under the name HelixCorr at `http://webclu.bio.wzw.tum.de/helixcorr/`.

3.3 Results and discussion

3.3.1 Selection of optimal sequence alignments

For the analysis of co-evolving residues in membrane proteins a procedure was developed to extract optimal protein clusters and hence optimal sequence alignments from the CAMPS database of membrane proteins [36]. Starting with all 266 clusters corresponding to structural folds (SC-clusters), clusters whose sequences were too diverse to allow for reliable predictions were discarded or replaced by sub-clusters. On the other hand, since highly similar sequences might result in few correlations due to a lack of variability, sequences with a pairwise identity above a pre-set threshold were considered redundant and removed. Based on this selection procedure, 91 optimal protein clusters were selected, of which 14 contained at least one representative protein structure forming the dataset with high-quality alignments (dataset `MP_14`, Table 3.1, page 46).

Multiple alignments for these clusters were obtained by concatenating transmembrane core sequences extracted from CAMPS. Co-evolving residues were extracted using seven different prediction algorithms (McBASC [156], OMES [169, 158], CORRMUT [121], CAPS [182], MI [174], SCA [176], ELSC [178]) which broadly cover the range of prediction approaches known from literature. Additionally, two different substitution matrices (the Miyata matrix [207] and the McLachlan matrix [208]) were evaluated in combination with the McBASC algorithm (McBASC-Miyata, McBASC-McLachlan) and the OMES algorithm was applied in two different versions, as originally introduced by Kass and Horovitz (OMES-KASS) [169] and in its modified version presented by Fodor and Aldrich (OMES-FODOR) [158], resulting in a total of nine different predictions for every multiple alignment. The number of significantly correlated residue pairs was chosen

proportional to the length of the multiple alignment by extracting the top L/5 correlations, with L being the alignment length. However, in the case of the two prediction algorithms CORRMUT and CAPS the number of obtained correlations with a minimal correlation coefficient of 0.4 was less than L/5 in most proteins. Figure 3.1 shows the sequence separation between all co-evolving residue pairs obtained by this procedure. A clearly resolved peak corresponding to a sequence separation of four residues (one turn of an alpha-helix) is observed which confirms that the obtained multiple alignments are indeed well suited for the prediction of co-evolving residues in membrane proteins.

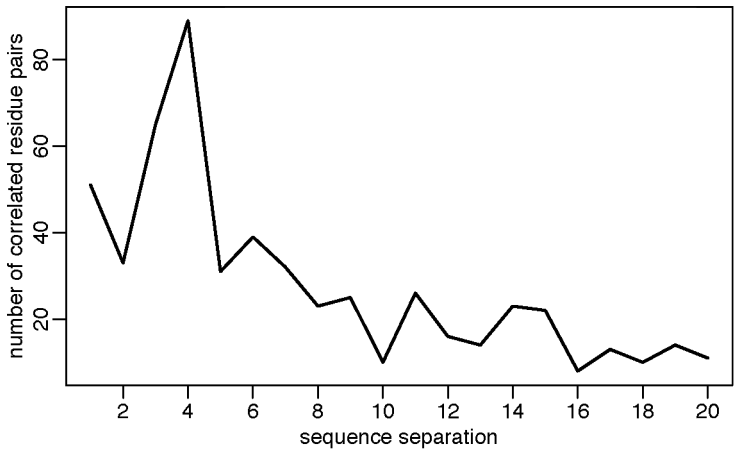

Figure 3.1: Sequence separation of co-evolving residues detected in the dataset `MP_14` with nine different prediction algorithms. The top L/5 correlations were considered for every protein and every prediction method (L being the alignment length). Residue pairs separated by one helix turn are most commonly found to mutate concurrently.

3.3.2 Helix-helix contact predictions obtained with different prediction algorithms

In order to evaluate the ability of individual algorithms to predict structural contacts in membrane proteins, contact prediction accuracies (fraction of correctly predicted contacts out of all correlations found) were calculated for all correlations with residues lying on separate transmembrane segment. For the dataset `MP_14`, between 3% (SCA) and 9% (McBASC-McLachlan) of these correlations were found to be helix-helix contacts (Table 3.2). The McBASC algorithm was slightly better when using it in combination with the McLachlan than the Miyata matrix. While the ELSC algorithm was clearly better than its predecessor SCA, the OMES algorithm in its original version as introduced by Kass

and colleagues was slightly better than its later version published by Fodor and Aldrich. According to a hypergeometric distribution significant predictions with p-values of less than 0.001 were obtained for all prediction methods except SCA.

Table 3.2: Contact prediction accuracies with different prediction algorithms applied to the dataset `MP_14`. CONSENSUS-14 and CONSENSUS-R-5 correspond to two consensus predictions, where predicted correlated mutations obtained with different prediction algorithms are combined. The algorithms McBASC and OMES obtain the highest prediction accuracies of all individual methods, only outperformed by the consensus predictions.

Method	Proteins	Acc [%][a]	P-value	Xd	Acc ($\|\delta\|$=4) [%][b]	P-value
McBASC-Miyata	14	8	2.59E-14	5.6	49	3.91E-30
McBASC-McLachlan	14	9	3.05E-17	5.0	42	1.36E-16
OMES-KASS	14	8	6.37E-13	4.9	43	2.55E-17
OMES-FODOR	14	7	1.38E-11	4.0	38	4.73E-11
CORRMUT	13	7	3.26E-05	4.4	38	9.25E-06
CAPS	5	7	8.06E-04	4.4	42	5.09E-05
MI	14	5	2.08E-04	-1.8	19	0.998
SCA	14	3	0.032	0.48	26	0.152
ELSC	14	7	4.42E-09	3.2	37	2.24E-09
CONSENSUS-14	14	11	1.08E-54	8.5	53	4.1E-100
CONSENSUS-R-5	14	10	4.35E-47	6.7	51	5.18E-82

[a] Acc: prediction accuracy for residues lying on separate transmembrane helices.
[b] Acc ($|\delta|$=4): prediction accuracy for residues lying on separate transmembrane helices with all correlations considered to be correct lying within one helix turn of an observed contact.

Similar results were obtained when the number of selected correlations was not chosen proportional to the length of the multiple alignment used for the prediction but varied over a broad range independent of the protein lengths (Figure 3.2A). Again McBASC used with the McLachlan matrix performed slightly better than the other prediction algorithms. From the two OMES variations the original version (OMES-KASS) was slightly superior to its variation introduced by Fodor and Aldrich except for very small numbers of selected correlations. MI and SCA were found to be the least powerful algorithms in the prediction of helix-helix contacts independent of the number of significantly correlated residue pairs selected. The algorithms CAPS and CORRMUT were excluded from this analysis since the number of significantly correlated residues obtained with these two algorithms was in most proteins clearly smaller than with the other prediction algorithms.

After evaluating the prediction of helix-helix contacts on a small set of membrane protein structures with high-quality multiple alignments, the best performing methods

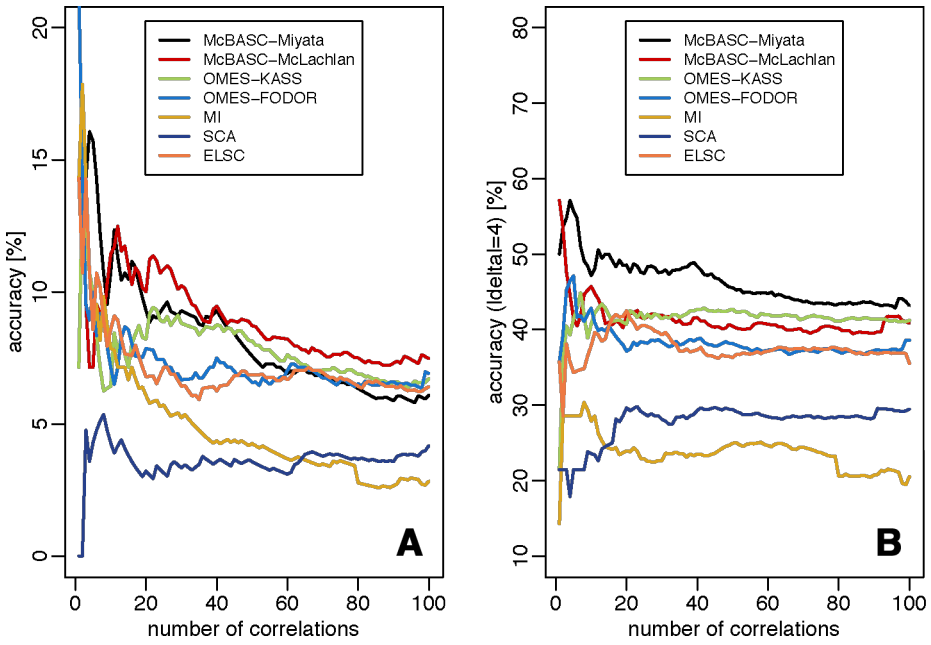

Figure 3.2: Comparative assessment of contact prediction performance of seven methods predicting co-evolving residues on the dataset `MP_14`. (A) Prediction accuracies for helix-helix contacts. McBASC-McLachlan obtains highest prediction accuracies over a wide range of selected correlations. (B) Fraction of correlations lying within one helix turn of a helix-helix contact (accuracy with $|\delta|$=4). Here, the McBASC-Miyata method is superior to all other methods.

were applied to a larger dataset consisting of 62 non-redundant membrane proteins (dataset `MP_62`). Multiple alignments for these proteins were obtained using PSI-BLAST in combination with the same sequence filters used also on the smaller dataset. Again, prediction accuracies were calculated for all prediction algorithms (Table 3.3).

Generally, all tested methods perform with similar accuracy on the large dataset `MP_62` as on the earlier introduced dataset containing only 14 membrane proteins. Again, McBASC-McLachlan shows the highest predictive performance with even slightly increased performance compared to the smaller dataset `MP_14`. In contrast to the smaller dataset however, OMES-FODOR outperforms now the related OMES-KASS method. Generally, this demonstrates, that the small dataset is well chosen resembling closely the larger set of available membrane protein structures. All results gained on the small dataset are therefore likely to hold also for membrane proteins in general. Secondly, the procedure developed for deriving optimal alignments for the analysis of co-evolving residues in membrane proteins seems to be appropriate for a large set of proteins.

Table 3.3: Contact prediction accuracies for 62 membrane proteins (dataset MP_62). CONSENSUS-R-5 corresponds to a consensus prediction, where predicted correlated mutations obtained with different prediction algorithms are combined. Contact prediction accuracies obtained on a reduced set of membrane proteins (MP_14) can be reproduced on an increased number of membrane protein structures.

Method	N(Contacts)	Acc [%][a]	P-value	Acc ($\|\delta\|$=4) [%][b]	P-value
McBASC-Miyata	1589	12	<2.2E-16	49	<2.2E-16
McBASC-McLachlan	1589	14	<2.2E-16	50	<2.2E-16
OMES-KASS	1589	8	<2.2E-16	40	<2.2E-16
OMES-FODOR	1589	12	<2.2E-16	51	<2.2E-16
ELSC	1589	8	<2.2E-16	40	<2.2E-16
CONSENSUS-R-5	3641	14	<2.2E-16	54	<2.2E-16

[a] Acc: prediction accuracy for residues lying on separate transmembrane helices.
[b] Acc ($|\delta|$=4): prediction accuracy for residues lying on separate transmembrane helices with all correlations considered to be correct lying within one helix turn of an observed contact.

3.3.3 Sequence separation between co-evolving residues and helix-helix contacts

Despite the low percentage of correctly predicted contacts, a high fraction of all correlations was detected to be in direct neighborhood of helix-helix-contacts. Starting with a general analysis of residue-residue distances within the dataset MP_14, distances between correlated residues lying on different transmembrane segments were clearly shifted towards smaller values compared to the distance distribution observed for all possible pair of amino acids, as was already described for soluble proteins [157]. When analyzing the results of every prediction method individually (Figure 3.3), this shift towards smaller residues was observed for all algorithms except MI and SCA. In the case of MI, distances between correlations were even shifted towards larger distances compared to the overall distances distribution.

The difference between the two distance distributions can be also quantified using the harmonic average Xd as introduced by Valencia and co-workers [194], where Xd>0 indicates a shift of the population of predicted residue pairs to smaller distances with respect to the population of all pairs. For the dataset MP_14 maximal Xd-values up to 5.6 (McBASC-Miyata) were obtained considering individual prediction methods (Table 3.2, page 54). Intermediate Xd-values between 3.2 and 5.0 were obtained with the methods ELSC, McBASC-McLachlan, OMES-KASS, OMES-FODOR, CAPS and CORRMUT. Apart from the results obtained with MI and SCA (negative Xd-value or Xd close to

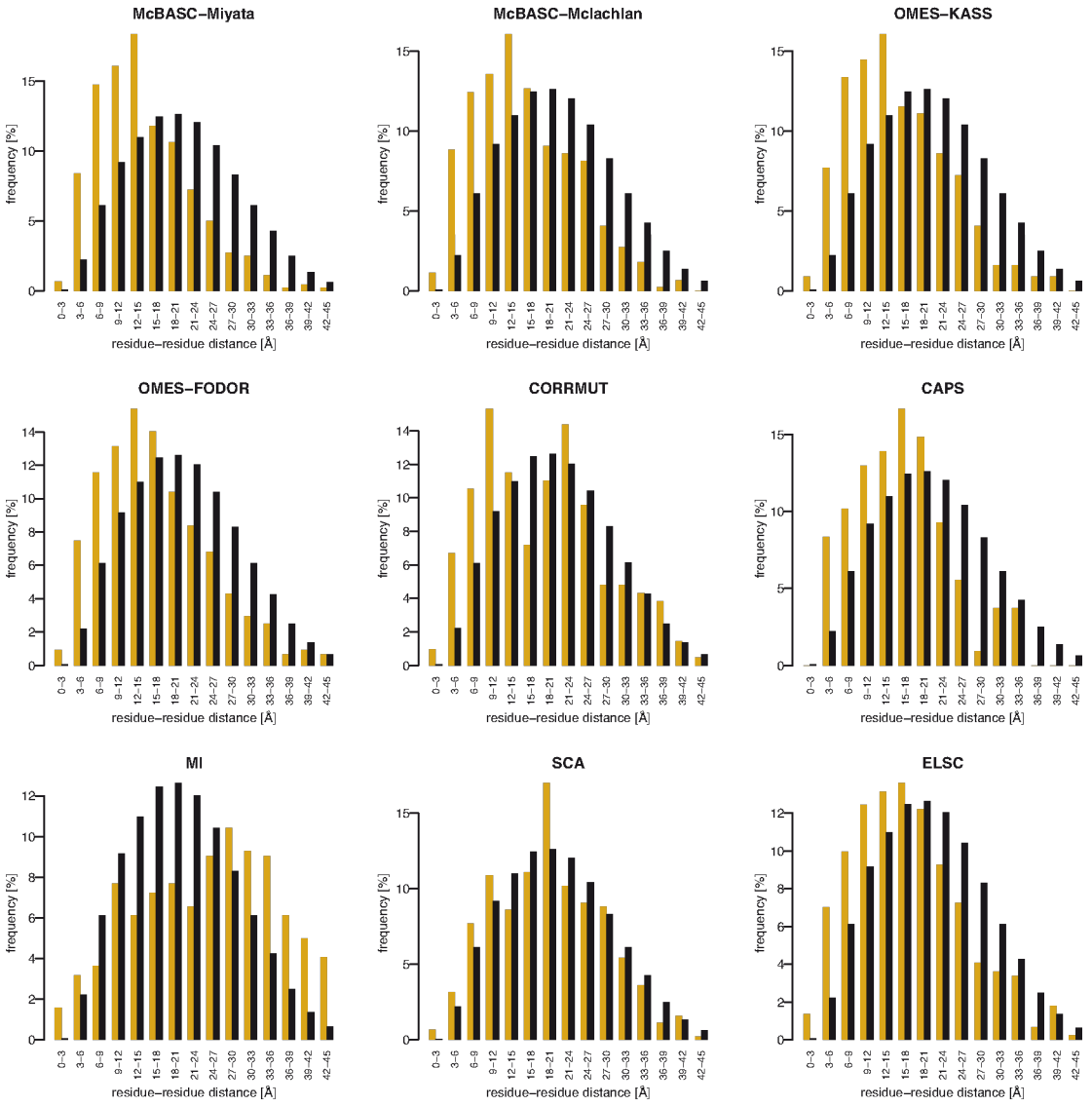

Figure 3.3: Spatial distances of highly co-evolving residues (orange) obtained with nine different prediction methods for the dataset `MP_14` compared to the distribution of all residue distances within 14 membrane proteins (black). For all methods except MI and SCA, distances between co-evolving residues are clearly shifted towards smaller values compared to the full distribution of distances.

zero, respectively), these results are comparable to those obtained for soluble proteins, where a contact prediction accuracy of 9% and a Xd of 4.31 was reported for a dataset of 173 proteins using the McBASC algorithm in combination with the McLachlan matrix [188]. However, it is noteworthy that all-alpha soluble proteins are known to be the most difficult targets for contact prediction using correlated mutations. Using a neural network approach, incorporating also other sequence information such as conservation or

predicted secondary structure, an average prediction accuracy of 7% and Xd-values for individual proteins between -5.0 and 2.7 were reported for this class of proteins [188]. While the contact prediction accuracy is comparable to the results obtained here for alpha-helical membrane proteins, the Xd-values are clearly smaller for all-alpha soluble proteins. The better results obtained for membrane proteins suggest that their typical structural arrangement, with several contacting residues between approximately parallel interacting helices, is more inclined to prompt residues lying in close structural distance to co-evolve than this is the case for soluble all-α proteins.

Using a 'δ-evaluation' [209], where the fraction of correlated positions i and j is calculated with an observed helix-helix contact between residues in the intervals {i-δ,i+δ} and {j-δ,j+δ}, on average up to 49% (McBASC-Miyata) of all detected correlated pairs within the dataset `MP_14` were found to be situated within the same helical turn as an actual contact (accuracy with $|\delta|$=4) (Table 3.2, page 54). The exact fraction differed strongly, depending on the protein and the applied prediction algorithm (data not shown). In individual cases, such as the mechanosensitive channel protein (1MXM), the best prediction was obtained with the MI algorithm, which, on average, performed worse than all other prediction algorithms. Again, these results were fairly consistent between the small and large dataset (Tables 3.2 and 3.3) although a noticeable increase in prediction accuracy ($|\delta|$=4) was observed for the methods McBASC-Mclachlan and OMES-FODOR when testing them on 62 instead of 14 membrane proteins.

As presented earlier for helix-helix contact prediction accuracies, the influence of the selected number of correlated residues on the obtained accuracy with $|\delta|$=4 was also analyzed (Figure 3.2B, page 55). In contrast to the prediction of helix-helix contacts, where the McLachlan matrix performed better than the Miyata matrix, in this analysis best results were obtained using the McBASC algorithm in combination with the Miyata matrix over the full number of analyzed correlations. Results using SCA and MI were again clearly inferior to results from all other prediction algorithms.

In publications on co-evolving residues in soluble proteins, low contact prediction accuracies using correlated mutations have often been attributed to methodological problems in separating real correlated mutational behavior from random noise as well as to co-evolution of distant residues due to long-range interactions [121, 176] or functional reasons [174]. In membrane proteins, pairs or networks of compensatory mutations seem to affect the packing context of transmembrane helices rather than the contacts themselves, as can be concluded from the high fraction of co-evolving residues found in direct neighborhood to helix-helix contacts (illustrated also by Figure 3.5 on page 60). The

surrounding residues of helix-helix-contacts might be generally more amenable to mutational change than the residues in actual contacts, but are still sufficiently important for proper helix interactions to make the compensation of destabilizing amino acid substitutions beneficial for protein stability. Notably, this finding is in line with experimental evidence that helix-helix-interactions mediated both by polar residues and interaction motifs are dependent on the sequence context ([131], see also Chapter 2).

3.3.4 Improvement of prediction accuracies using a consensus approach combining several prediction methods

Based on the observation that results with different prediction algorithms vary remarkably for individual proteins (data not shown), a consensus prediction method was developed combining for every protein co-evolving residues from different prediction methods. Since the two prediction methods SCA and MI were found to perform worse regardless of the prediction quality measure used, these two algorithms were excluded of any consensus approach. Within a first step, the results of the remaining seven predictions were combined to form a initial list of candidate correlations. To further improve the obtained prediction by reducing likely false positives, all correlations lying on a pair of helices with a total number of detected correlations less than a given threshold N were removed resulting in the final consensus prediction (termed CONSENSUS).

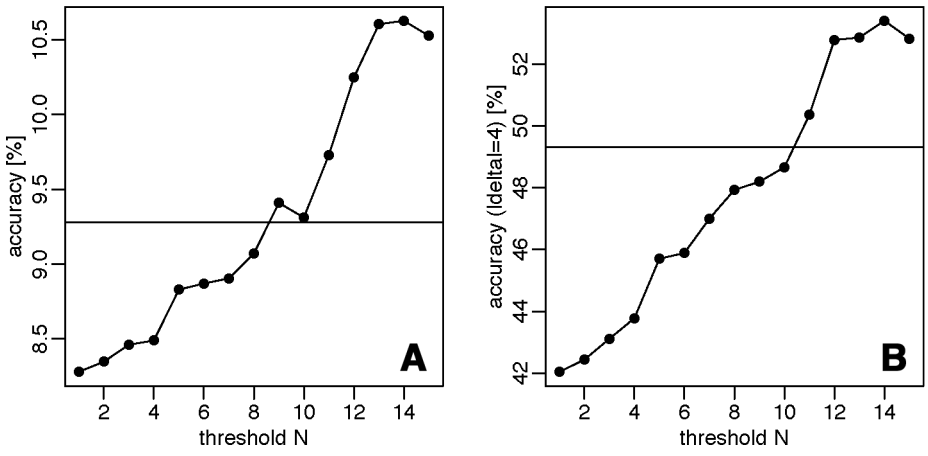

Figure 3.4: Improvement of helix-helix contact prediction accuracy (A) and accuracy ($|\delta|$=4) (B) by applying a consensus approach to the dataset `MP_14`. For comparison, the horizontal line indicates the maximal value obtained with a single prediction algorithm. Prediction accuracies increase the higher the threshold N is chosen.

Figure 3.4 demonstrates that both helix-helix contact prediction accuracy and accuracy ($|\delta|$=4) indeed increase with an increase in this threshold N. With N=14 (CONSENSUS-14) the helix-helix contact prediction accuracy could be improved close to 11% (compared to 9% as best result for a single algorithm) and the accuracy ($|\delta|$=4) could be elevated to 53% (compared to 49% again as best results for an individual algorithm) (Table 3.2, page 54).

For a second consensus approach, correlations detected with the four best performing prediction methods (McBASC-Miyata, McBASC-McLachlan, OMES-KASS and CAPS, selected based on their accuracy with $|\delta|$=4) were combined and again all correlations on helix pairs with less than N correlations in total were removed (reduced consensus or CONSENSUS-R). With N=5 (CONSENSUS-R-5) helix-helix contacts could be predicted with 10% accuracy and the fraction of correlations lying within one helical turn of an actual helix-helix contact was found to be 51% on the small dataset consisting of 14 proteins. The contact map of the AcrB bacterial multidrug efflux transporter (Figure 3.5) illustrates how false positive predictions can be removed by applying this threshold N=5 in comparison to a mere combination of individual prediction methods.

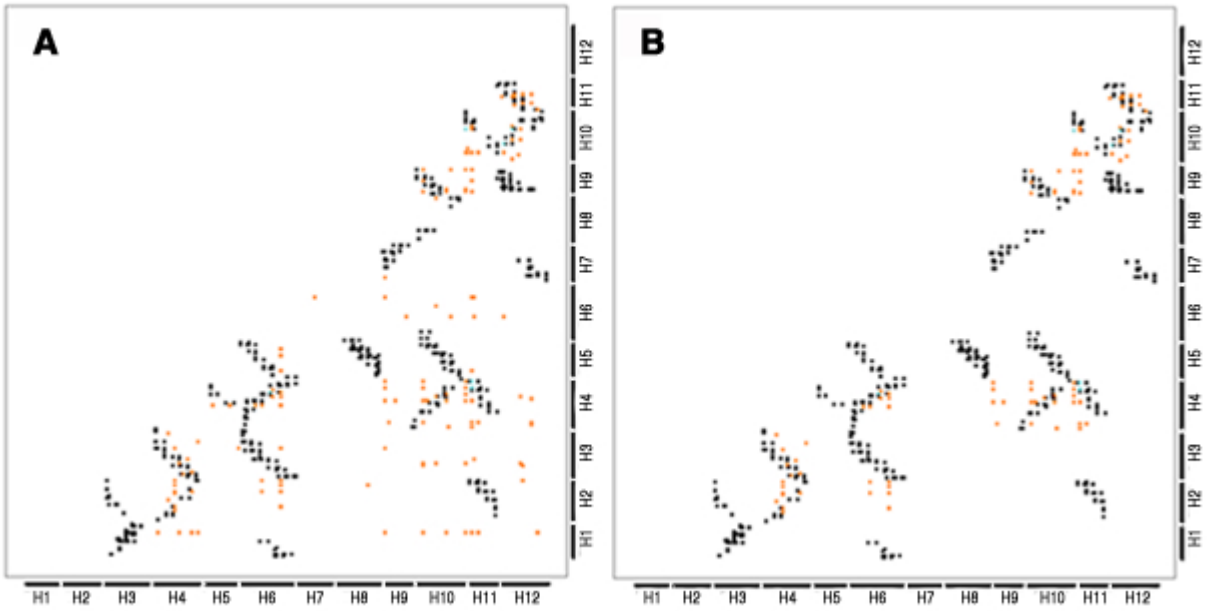

Figure 3.5: Contact maps of the AcrB bacterial multidrug efflux transporter (PDB 1IWG chain A). (A) All correlations detected with the methods McBASC-Miyata, McBASC-McLachlan, OMES-KASS and CAPS. (B) Improved prediction with CONSENSUS-R-5. Observed contacts are shown in black, co-evolving residues in orange, correctly predicted contacts in blue. Individual transmembrane helices are indicated by H1-H12.

3.3.5 Prediction accuracies based on experimentally determined transmembrane segments

All results presented so far for the dataset `MP_14` were based on multiple alignments consisting of transmembrane segments predicted with TMHMM. As shown in Table 3.1 (page 46) these predicted transmembrane helix positions may differ slightly from those determined based on the PDB structure. To evaluate whether these differences have a noticeable effect on the detection of co-evolving residues, multiple alignments consisting of experimentally determined transmembrane segments (as annotated in PDBTM [201]) were obtained. Correlated mutations were predicted with the same procedure used earlier for multiple alignments consisting of predicted transmembrane segments. Helix-helix contact prediction accuracies were found to increase by 1% to 5% while the fraction of correlations within one helix turn of a helix-helix contact increased even by 13% in the case of the MI algorithm (Table 3.4). The analysis of co-evolving residues in membrane proteins, where solved protein structures as well as experimentally determined topologies are generally only available in rare cases, is therefore also significantly dependant on the quality of the predicted transmembrane helix positions.

Table 3.4: Contact prediction accuracies for different prediction algorithms applied to the dataset `MP_14` with experimentally determined transmembrane segments. CONSENSUS-14 and CONSENSUS-R-5 correspond to two consensus predictions. Prediction accuracies improve by 1% to 5% when using experimentally determined instead of predicted transmembrane helices.

Method	Proteins	Acc [%][a]	P-value	Acc ($\|\delta\|$=4) [%][b]	P-value
McBASC-Miyata	14	9	6.22E-14	49	7.82E-28
McBASC-McLachlan	14	10	6.48E-17	46	4.92E-21
OMES-KASS	14	9	1.17E-14	50	2.55E-28
OMES-FODOR	14	8	7.78E-12	41	1.73E-14
CORRMUT	14	7	1.65E-05	33	0.0016
CAPS	5	12	1.29E-06	47	8.31E-07
MI	12	8	3.07E-10	32	0.0014
SCA	14	3	0.0702	27	0.1131
ELSC	14	10	6.48E-17	41	1.73E-14
CONSENSUS-14	14	12	1.41E-53	55	2.48E-95
CONSENSUS-R-5	14	11	3.64E-42	56	6.26E-94

[a] Acc: prediction accuracy for residues lying on separate transmembrane helices.
[b] Acc ($|\delta|$=4): prediction accuracy for residues lying on separate transmembrane helices with all correlations considered to be correct lying within one helix turn of an observed contact.

4

Prediction of helix-helix contacts using neural networks

After conducting the first analysis of co-evolving residues in polytopic membrane proteins (Chapter 3), the results of this study clearly indicated that co-evolving residues alone are not sufficient to predict helix-helix contacts, but that these residues still carry a strong signal for the detection of interacting transmembrane helices due to their frequent occurrence in close sequence neighbourhood to helix-helix contacts. Subsequently, the prediction of helix-helix contacts was further addressed within a second project trying to improve obtained contact predictions by correlated mutations alone. To this end, a neural-network based approach was developed specifically for the prediction of helix-helix contacts in alpha-helical membrane proteins. It integrates sequence profiles, correlated mutations, protein topology, sequence separation and predicted scores for lipid-exposure and hence is the first predictor of residue-residue contacts incorporating membrane protein specific input data.

Neural networks have been used for contact prediction of soluble proteins already for several years (see for example results of the latest CASP competitions, [185, 186]) resulting in important insights regarding usefulness and execution of this approach, which often are also transferable for the application to membrane proteins. The following introduction will therefore summarize the progress within the field of residue contact prediction in general. In the following, the development of the first membrane protein specific contact predictor will be described and its success will be evaluated and compared to those methods originally developed for soluble proteins.

The prediction of helix-helix contacts in membrane proteins using neural networks was jointly executed with Andreas Kirschner (TU München). While Andreas Kirschner was mainly concerned with the realization of the neural network itself, I was responsible

for dataset creation, the selection of possible input features and the testing of external methods. All remaining results were analysed together by Andreas Kirschner and me. Comments in the following sections will further clarify individual contributions.

The results of this chapter were published in [210].

4.1 Introduction

While the first contact predictors available for soluble proteins were mostly based on co-evolving residues (see Chapter 3), recent contact prediction methods can be classified into two categories. First, machine learning methods try to learn the interrelationship between a number of predefined sequence features and the contact state of a given residue pair. In contrast, template-based approaches deduce contacts for the target sequence from template proteins which are assumed to have a similar fold than the target sequence. Here, both approaches are shortly reviewed and the current status of obtained contact prediction accuracies for soluble proteins is presented. Finally, potential applications of predicted residue contacts are discussed.

4.1.1 Sequence-based contact prediction

When contact prediction approaches were still mainly focused on the analysis of co-evolving residues, additional sequence information was already shown to improve the accuracy of obtained predictions significantly [156]. Accordingly, machine learning approaches which are able to incorporate a variety of sequence features have been consistently demonstrated to outperform methods using co-evolving residues alone [211, 188, 187]. Generally, these methods require contact maps of proteins with known structures. During a training phase, the machine learning algorithm tries to deduce association rules between selected sequence features of each protein in the training set and its contact map. These rules are then applied to proteins without known contact map.

Over the years, several different implementations of machine learning approaches have been applied for the residue contact prediction problem with neural networks [211, 188, 189, 191], support vector machines [190] and hidden Markov models [212, 213] being the most commonly used ones. While the first neural network developed for the prediction of residue contacts was based on only a limited number of sequence features (namely sequence profiles, sequence conservation, correlated mutations and predicted secondary structure) [211, 188], additional input features have been incorporated and tested in recent contact predictors. For example, PROFcon, one of the best performing

methods of CASP6 [185, 189], reported increased contact prediction accuracies due to the incorporation of additional sequence features such as predicted solvent accessibility, sequence distance of two residues and global information such as protein length. Similarly, Shackelford and Karplus contributed on of the superior methods to the CASP7 competition with their neural network using a novel statistic for correlated mutational behaviour of two residues in combination with several other sequence features [186, 191]). As indicated by the increased number of contributions in the recently conduced CASP8 experiment [214], the development of new contact prediction methods is still ongoing promising further insights into the correlation between sequence features and residue contact state.

4.1.2 Template-based contact prediction

In contrast to sequence-based contact prediction methods which are completely independent of homologous structures, template-based contact predictions require the availability of related proteins having a solved 3D structure. Within a first step, appropriate template structures which most likely share the same fold as the target protein are identified using threading techniques. After aligning the target sequence to found templates, contacts are then inferred from the template structures [213, 215, 216, 217]. Recent improvements in the field of template-based contact prediction introduced the identification of templates with a meta-server combining several threading programs [216] as well as the usage of machine learning methods for the ranking of contacts obtained from different templates [217]. While contacts obtained from structural templates are frequently used as restraints in 3D structure prediction of soluble proteins [218, 215, 219], membrane proteins are only rarely approachable with these methods due to the lack of available homologous protein structures.

4.1.3 Contact prediction accuracies obtained for soluble proteins

Contact prediction methods have been evaluated independently in several rounds of the CASP experiment (see for example [185, 186, 214]). Results obtained from these evaluations suggest that at least slight increases in prediction performance are observed over the last years as methods performing best within previous CASP experiments are generally outperformed by newly developed methods of recent CASP experiments. However, prediction accuracies are not steadily increasing which seems to be a result of different target difficulties (all-alpha proteins for example were repeatedly shown to be more

difficult to predict than other protein structures [188, 189]).

Within CASP6, three methods performed slightly better than other participants with prediction accuracies between 16% and 23% for long-range contacts separated in sequence by at least 24 positions. While one of these methods (GPCpred) was based on genetic programming, the two remaining methods used neural networks for their prediction. During the CASP7 experiment, again a neural network based predictor was found to predict targets with higher accuracy than the remaining methods although the general level of obtained accuracy was clearly lower than in the preceding CASP round. On average, 13% prediction accuracy was reported for all participating groups (for a sequence separation $\geq$24) and even the best performing methods gained maximal prediction accuracies of roughly 20%. The recently conducted CASP8 experiment resulted again in improved prediction accuracies (mean accuracy of 21.5% over all targets and predictors) with several groups reaching average accuracies of higher than 25%.

Within all CASP evaluations, contacts predicted with specialized contact predictors were also compared to contacts derived from predicted 3D structures. No general trend could be observed with contact specialists performing superior on some targets while 3D structure predictions resulted in better contact predictions for other targets. A recent comparison of sequence-based and template-based contact prediction methods suggests that template-based methods are only superior in case target and template share an sufficient amount of evolutionary and structural similarity [217]. For proteins without appropriate template, predictions of more than 10% higher accuracy were obtained with a sequence-based machine learning approach than using a template-based threading approach.

4.1.4 Applications of contact predictions

Long-range contacts constitute an important information in *ab initio* protein structure predictions as they can be used to constrain the conformational search space. While predicted contacts are principally valuable for this task, the high numbers of false positives obtained with current contact predictors (see above) are still strongly limiting their practical application. According to estimations, one correct contact in every eight positions would be sufficient for guiding protein folding simulations of proteins smaller than 200 amino acids [220]. The number of false positives that can be tolerated on the other hand is less clear. Nevertheless, several efforts have been reported trying to incorporate predicted residue contacts as constraints in *de novo* structure prediction [209, 215, 219].

In addition to the direct incorporation in *ab initio* structure prediction experiments,

predicted residue contacts may be used for the ranking of alternative protein models as well as the refinement of initially course-grained models as recently demonstrated by Latek and Kolinski [221].

4.2 Materials and methods

4.2.1 Dataset

For the development and evaluation of the presented neural network, the non-redundant dataset `MP_62` containing membrane proteins with solved structure was used as introduced already in Chapter 3 (section 3.2.1, page 47).

Briefly, this dataset was obtained from the database PDBTM [201] and the dataset provided by the Stephen White laboratory at UC Irvine (`http://blanco.biomol.uci.edu/Membrane_proteins_xtal.html`). After redundancy removal (for details see section 3.2.1), 62 protein chains remained (Appendix Table 9.3). Transmembrane segment positions and the in/out topology for each protein were obtained from the database TOPDB [202] except for the proteins 2UUH (chain A) and 1ORQ (chain C) which were not included in TOPDB. Topology information for these proteins was obtained from PDBTM, instead.

The dataset was constructed by myself.

4.2.2 Contact definition

In addition to the dataset, the same helix-helix contact definition was used as in the analysis of residue co-evolution within membrane proteins (see section 3.2.3, page 50). According to this criterion, two residues within different transmembrane segments were considered in contact if the minimal distance between side chain or backbone atoms was less than 5.5Å. Thereby, side chain conformations were more appropriately considered than would be the case if contacts were defined based on Cβ-distances as mostly done by contact prediction methods developed for soluble proteins. Accordingly, other studies on helix packing and helix-helix contacts in membrane proteins have also used contact definitions including side chain atoms [60, 62].

Importantly, the difficulty of the contact prediction problem for membrane proteins is not influenced by the choice of contact criterion since the number of observed contacts remains basically the same. Using the contact criterion based on side chain atoms, the observed overall contact density (the number of observed contacts divided by the

number of possible pairs) was 0.021 while the usage of the Cβ contact criterion resulted in a contact density of 0.020 for the dataset of 62 membrane proteins.

4.2.3 Contact density

To estimate the optimal number of contacts to predict per protein, the dependency between the number of transmembrane residues and the amount of helix-helix contacts was estimated. The observed contact density within transmembrane parts of all 62 transmembrane proteins in the dataset was afterwards compared to corresponding values derived for soluble proteins taken from the 25% homology threshold list of the pdb_select database from October 2007 [222]. Two different subsets of pdb_select were used, one comprising all 3652 pdb_select proteins belonging to the SCOP [223] classes all-alpha, all-beta, alpha and beta (a/b), alpha and beta (a+b) and multidomain proteins and one subset consisting of all-alpha proteins only. In any case, contacts were calculated according to the definition given above.

For every dataset linear functions were fitted describing the dependency of the number of observed contacts on the length of the protein (for membrane proteins only the transmembrane parts were considered). The following (rounded) dependencies between the number of considered residues L and the amount of observed contacts C were obtained: For soluble proteins in general C=3.15L-76.5, for all-alpha proteins C=2.5L-75 and for the transmembrane parts of all 62 membrane proteins C=2.25L-100 (see Results, Figure 4.4, page 79). As can be seen from these formulas, contacts between transmembrane segments are generally less frequent than contacts within soluble proteins having the same number of residues (see also Results and discussion).

The contact density analysis was conducted by Andreas Kirschner and myself.

4.2.4 Neural network input features

The prediction of spatial contacts between two amino acid residues is generally based on the analysis of multiple sequence features. These features can be divided into out-of-context features defined for single residues without any contact related information, features targeting properties related to residue pairs in contact, and features that describe global properties of the proteins. Contact prediction is then derived by mapping these features onto the contact state of the residues under observation. Over the last years, machine learning algorithms have become the method of choice to obtain such mapping in an automated fashion (for example see [211, 188, 212, 189, 190, 191, 217]).

The better the chosen features relate to the contact information of two residues, the better the mapping and thus the better the predictive performance of the developed algorithm. Accordingly, for the membrane protein contact prediction problem, prominent features used for globular protein contact prediction were included together with various features that are available for membrane proteins only.

All input features were selected together by Andreas Kirschner and myself, the implementation of the selected features was done by Andreas Kirschner.

Out-of-context features

The used out-of-context features describing individual residues are: windowed PSSM (Position-specific Scoring Matrices) profiles, the position of each residue within the transmembrane helix (cytosolic side of the membrane, hydrophobic core or extracellular side), and the orientation of its side chain, i.e. whether the residue is facing towards the lipophilic membrane or the protein interior.

The PSSM profiles were obtained using PSI-BLAST [205] searches against the NCBI's unfiltered NR database [206], with three iterations and the inclusion of related database sequences into the profile with an E-value threshold of $1\text{x}10^{-4}$. The raw profiles from PSI-BLAST contained scores for all residue positions representing their amino acid preferences. These scores were transformed by the standard logistic function to obtain values in the range [0...1]. In order to include information about adjacent residues as well, a window of five residues to the left and five residues to the right was employed together with the central target residue. An additional feature was included to indicate whether the window was not built properly due to missing data (i.e. at the end of protein sequences).

The position of each residue within the transmembrane helix was encoded by two distinct features. First, a boolean vector of length S was used to represent each transmembrane helix divided into a set of S fragments of equal size. The values of the vector were initialized with 0 and the value at vector index $s = \lfloor \frac{S}{N} \cdot i \rfloor$ was set to 1 with N representing the length of the transmembrane helix, i being the position of the described residue within the transmembrane helix numbered from 1 to N from the N- to C-terminal end and the function $f(x) = \lfloor x \rfloor$ returning the largest integer which is less or equal the real number x. Based on preliminary optimization experiments, the parameter S was fixed at S=7. Second, a boolean vector of size three was used, to encode whether a residue lies close to the extracellular side of the membrane, close to the cytoplasm or within the hydrophobic core of the helix. A region of seven residues was used to define

both the extracellular or cytoplasmic side of the helix.

The side chain orientation of each residue was calculated using LIPS [122], a method for the prediction of transmembrane helix orientation with a reported accuracy of close to 90%. LIPS defines seven helical surfaces called faces which are identified based on the average lipophilicity and the conservation of residues within each face. Large LIPS scores indicate that a particular face is oriented towards the membrane while low scores indicate an orientation towards the hydrophilic membrane protein interior. The helix orientation was encoded in a boolean vector of length seven with the elements in the vector representing the seven helical faces ordered by increasing average lipophilicity. The vector was initialized by zeros. If a residue is member of the helical face with the i-th highest LIPS score, this i-th element was set to 1 in the boolean vector. A single residue can participate in up to three helical faces, as defined by Adamian and Liang [122].

Features of residue pairs

To represent properties pertinent to paired residues, two features were considered: sequence distance between the residues and predicted correlated mutation rates indicating co-evolving residues. The distance between two residues was encoded by a boolean vector of length eight corresponding to sequence separations of less than 25, 50, 75, 100, 150, 200, 300 residues, or more. For a given pair of residues having a sequence separation corresponding to the vector element i, not only this vector element was set to 1 but also all vector elements at positions $\leq$ i.

Residue co-evolution was calculated using three different prediction methods. The algorithm McBASC [156] was applied in two variations, using either the McLachlan [208] or the Miyata [207] substitution matrix, and the OMES algorithm was used in its modified version by Fodor and Aldrich [158]. Multiple sequence alignments used for the calculation of correlation scores were obtained from the PSI-BLAST alignments. First, all positions were removed from the full length PSI-BLAST alignment which did not correspond to any transmembrane segment of the PDB sequence resulting in an alignment representing only the transmembrane parts of the reference sequence. Following the procedure developed during the analysis of residue co-evolution (Chapter 3, section 3.2.1, page 45), sequences thought to be inappropriate for the prediction of correlated positions were discarded. The raw correlation scores were standardized individually for all proteins following the formula $y = \frac{x-min}{max-min}$, where x is the raw correlation score and min and max are the minimal and maximal scores observed for a given protein and

algorithm. Applying this type of standardization conserved relative scores but made results from different proteins comparable. As observed within the analysis of co-evolving residues (Chapter 3, section 3.3.3, page 56), co-evolution in membrane proteins occurs much more often at residue pairs in close vicinity to an actual helix-helix contact than at the contact positions themselves. Therefore, not only correlation scores found for the pair of residues i and j under observation were considered, but also for adjacent residue pairs with a window size of 5 centred around the positions i and j, respectively.

Global features

Two global protein features were considered for the neural network: protein length and the number of transmembrane helices. Both descriptors were again encoded as boolean vectors using the same strategy as described for the sequence distance. The protein length vector had a size of five elements corresponding to protein lengths of less than 100, 200, 400, 800 or more residues. The vector describing the number of transmembrane helices had a length of ten encoding proteins with 3, 4, 5, 6, 7, 8, 9, 10, 11 or 12 and more transmembrane regions.

Combination of features

Each input vector representing a residue pair contains out-of-context features for both participating residues, residue pair features, and global features of the particular protein. To estimate the importance of the various features, input vectors of increasing complexity and thereupon iteratively improved prediction performance were constructed (Figure 4.1). Starting with an input vector consisting of only those features available also for soluble proteins (NN1 and NN2, without and with correlated mutations, respectively), membrane protein specific features were gradually added (NN3: position within transmembrane segment and total number of transmembrane helices; NN4: side chain orientation). The NN4 implementation was additionally evaluated with a dataset that did not include instances with residue pairs from sequentially adjacent helices (termed NN4-distant or NN4-D) in order to estimate the dependence of predictive performance on short range contacts between neighbouring helix pairs. Throughout this work, NN4 and NN4-D, the neural networks based on the full set of input features, are synonymously also referred to as TMHcon. This final version of the contact predictor is available on-line (`http://webclu.bio.wzw.tum.de/tmhcon/`).

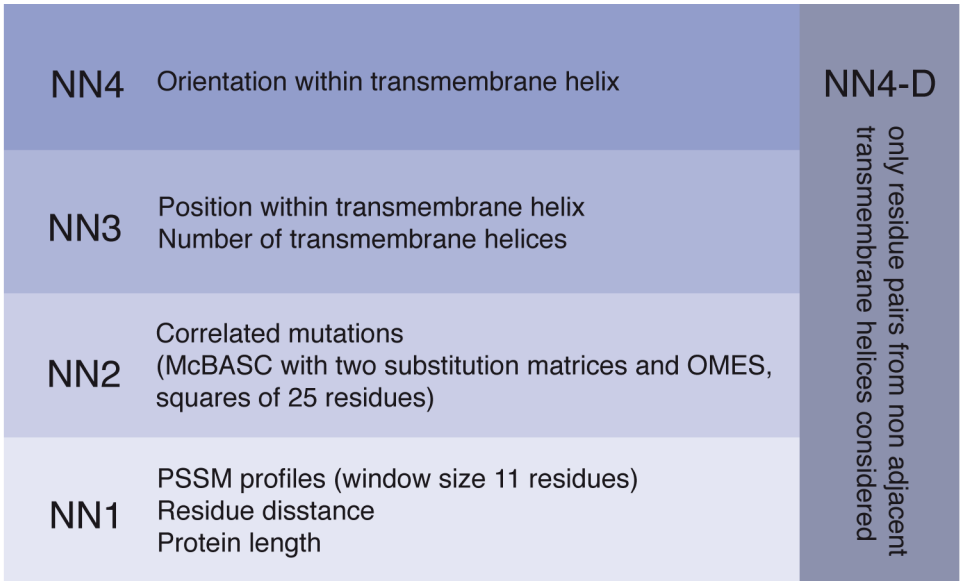

Figure 4.1: Input features used for the prediction of helix-helix contacts of membrane proteins.

4.2.5 Neural network architecture and training

Similar to many contact prediction methods for globular proteins, a feed-forward neural networks specially trained for data with biased class distributions was used. Every network consisted of the same number of input nodes as features available and two output nodes representing the two prediction classes 'contact' (positive class) and 'nonContact' (negative class). The number of hidden nodes was varied in order to optimize prediction performance, and finally an architecture with 90 hidden nodes was chosen.

Generally, during each training iteration of a neural network, called epoch, a set of instances is presented to the neural network, the average error on the given set is estimated and this error is used to calculate the weight update for all node connections. The presentation and weight update process is repeated until a defined stop criterion is reached. The contact prediction network was trained such that for each epoch all positive (contact) instances of the training proteins were chosen as well as a randomly selected equal number of negative instances. The training was iterated over 200 epochs.

The neural network was implemented and trained by Andreas Kirschner.

4.2.6 Measuring contact prediction performance

To assess the prediction performance of the developed neural networks a take-one-out jackknife cross validation was used whereby the method was tested on a single protein

while all other proteins formed the training set. Performance measures were obtained for the test protein and the procedure was repeated for all proteins. The overall prediction performance was calculated by averaging the individually obtained performance results leading to an accurate assessment of method performance.

Following common practice the number of predicted contacts was chosen based on the length of the protein L. Since contacts should be predicted for the transmembrane helices of a protein only, L was calculated as the sum of the lengths of all transmembrane helices of a given protein. Reported contact prediction accuracies are based on the L/5 highest scoring residue pairs, a threshold commonly used in contact prediction assessment [186]. From this number of predicted contacts the prediction accuracy (fraction of correctly predicted contacts out of all predicted contacts) was calculated. Additionally, the coverage (fraction of correctly predicted contacts out of all observed contacts) was calculated. In order to investigate the position of predicted contacts with respect to observed helix-helix contacts, a 'δ-Analysis' [209] was used, calculating the fraction of predicted contacts between residues i and j given an observed contact between residues in the interval {i-δ,i+δ} and {j-δ,j+δ}. To determine the fraction of predicted contacts where both participating residues lie within one helix turn of residues forming an inter-helical contact, δ=4 was used.

Contact prediction performance was evaluated together by Andreas Kirschner and myself.

4.3 Results and discussion

4.3.1 Prediction of helix-helix contacts using neural networks with increasing complexity

Machine learning techniques have been applied for the prediction of amino acid contacts in soluble proteins for more than five years [211, 188, 212, 189, 190, 191, 217]. Here, the first application of neural networks for the specific problem of predicting helix-helix contacts in membrane proteins is presented. Using contact data derived from 62 membrane proteins with solved structure, five neural networks for the prediction of helix-helix contacts were trained. While four of these networks were developed in order to analyze the influence of different input features on the resulting prediction, the neural network NN4-D included the same input features as the network NN4, but was trained only on long-range contacts lying on non-neighbouring transmembrane helices. Such long-range

contacts are particularly important for the discrimination of membrane protein folds resulting from differential helix packing in alpha-helix bundles and therefore, these residue contacts should be predicted with optimal sensitivity and reliability.

Influence of different input features on the prediction of helix-helix contacts

Following the strategy reported for the first contact map predictions using neural networks in globular proteins [211, 188], neural networks of increasing complexity were constructed by incorporation of an increasing number of input features. While the first two neural networks (NN1 and NN2) included only sequence features also available for soluble proteins (e.g. sequence profiles, sequence separation, protein length and correlated mutations), membrane protein specific features were incorporated in neural networks NN3 and NN4 (position of each residue within a transmembrane helix, number of transmembrane helices and orientation of each residue). This step-wise procedure reveals the contribution of individual feature sets, in particular those not available for soluble proteins and therefore missing in earlier studies on contact prediction with neural networks.

In agreement with publications on contact prediction for soluble proteins, the L/5 highest scoring contact pairs were selected for every protein and prediction accuracy (fraction of correctly predicted contacts out of the total number of predictions) and coverage (fraction of correctly predicted contacts out of the total number of observed contacts) were calculated. Additionally, the accuracy ($|\delta|$=4) was determined, a measure describing the fraction of predicted contacts that are found within one helix turn of an observed contact and therefore lie in close sequence neighbourhood to an actual helix-helix contact (Table 4.1).

As seen in Table 4.1, prediction accuracy increases by more than 8% with the addition of more and more input features. While the incorporation of correlated mutations leads to an improvement of 1.6% accuracy, the most significant increase in prediction accuracy of 4.6% is achieved with the first addition of membrane protein specific features in NN3. The incorporation of LIPS scores in NN4 leads to a further improved prediction accuracy of 25.9%. Since the number of analyzed predictions is equal for all neural networks, the coverage increases accordingly. The same trend can be observed for the accuracy ($|\delta|$=4), which increases by more than 13% from NN1 (65.2%) towards NN4 (78.5%). Interestingly, the observed value is basically constant between NN3 and NN4. In both cases around 78% of all predicted contacts are found in close sequence neighbourhood to an observed helix-helix contact. Since the number of predicted contacts located close

Table 4.1: Contact prediction with neural networks of increasing complexity. All values are reported based both on the selection of the L/5 highest scoring residue pairs (L being the length of the concatenated transmembrane segments), and after selecting the expected number of contacts derived using the contact formula for membrane proteins describing the observed number of contacts in dependence on the number of participating residues (see section 4.2.3). Contact prediction accuracies increase with the incorporation of additional sequence features into the prediction process.

Predictor	L/5			Contact density formula		
	Acc [%][a]	Acc ($\|\delta\|$=4) [%][b]	Cov [%][c]	Acc [%][a]	Acc ($\|\delta\|$=4) [%][b]	Cov [%][c]
NN1	17.2	65.2	2.3	10.5	61.2	10.6
NN2	18.9	68.4	2.6	11.4	65.4	11.6
NN3	23.5	78.7	3.2	15.7	70.8	15.8
NN4	25.9	78.5	3.5	15.8	70.7	16.0
NN4-D	14.8	50.2	3.9	10.0	46.0	10.1

[a] Acc: fraction of correctly predicted contacts out of all predicted contacts.
[b] Acc ($|\delta|$=4): fraction of predicted contacts lying within one helix turn of an observed contact.
[c] Cov: fraction of correctly predicted contacts out of all observed contacts.

to an actual contact stays the same while the number of correctly predicted contacts increases from NN3 towards NN4, the addition of LIPS scores seems to be helpful in determining the exact position of helix-helix contacts, which are otherwise only located slightly misplaced from the correct position.

Since the most remarkable increase in prediction accuracy is obtained from NN2 towards NN3 with the inclusion of a feature group defining each residue's position within the transmembrane helix, the predictions of NN3 were investigated in greater detail. As can be seen from the example in Figure 4.2, the given relative position of each residue within the transmembrane helix seems to aid the neural network in detecting the parallel or antiparallel interaction pattern of two transmembrane helices and therefore constrains predicted contacts. Figure 4.2A illustrates observed and the top L/5 predicted contacts for residues on transmembrane helices 1 and 2 from cytochrome B6 (PDB 1VF5 chain A) when using NN2. Here, contacts are predicted for the given two transmembrane helices, but the algorithm is not able to detect in which orientation the two helices are positioned relative to each other, resulting in a significant deviation of the predicted contacts from the known ones. In contrast, NN3 (Figure 4.2B) is able to deduce information on the helix orientation, and thus the predicted contacts lie on the correct diagonal of the contact map. The neural network is constrained by the transmembrane

residue positions: a residue near the extracellular membrane surface cannot contact a residue near the cytoplasmic membrane surface.

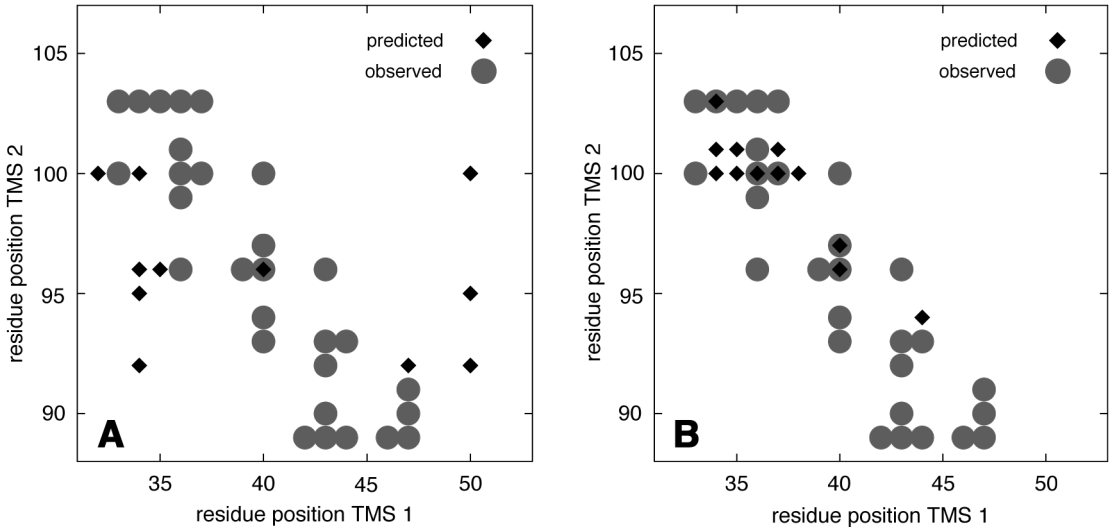

Figure 4.2: Observed and top L/5 predicted contacts between transmembrane helix 1 and transmembrane helix 2 of the protein 1VF5, chain A. (A) Predictions with NN2. (B) Predictions with NN3. NN3 includes information about each residue's position within the transmembrane helix and therefore is aware of the helix orientation. While NN2 has problems to detect the anti-parallel character of the helix interaction and predicts many contacts off diagonal, NN3 exclusively predicts contacts that capture the anti-parallel interaction pattern.

Dependence of the contact prediction performance on the number of transmembrane helices

For the best performing neural network NN4 it was further analyzed how the prediction success depended on the number of transmembrane segments within a protein. All 62 membrane proteins were grouped into subsets of proteins with similar number of transmembrane segments and the prediction accuracy and coverage was calculated for every subset (Table 4.2).

As expected, prediction accuracy decreases for large proteins. For proteins with eight or more transmembrane helices prediction accuracies of close to 20% were obtained, while proteins with less than eight transmembrane segments achieved prediction accuracies of 25% or more. Interestingly, the fraction of predicted contacts in close vicinity to observed contacts (accuracy ($|\delta|$=4)) is largely independent of protein size since in proteins having more than ten transmembrane helices contacts are still detected with an accuracy ($|\delta|$=4)

Table 4.2: Contact prediction using NN4 for subsets of membrane proteins grouped according to their number of transmembrane helices. All values are reported based on the selection of the L/5 highest scoring residue pairs (L being the length of the concatenated transmembrane segments). Proteins with seven transmembrane helices are clearly better predicted than all other proteins.

TMS	N(Proteins)	N(Contacts)	L/5		
			Acc [%][a]	Acc ($\|\delta\|$=4) [%][b]	Cov [%][c]
3-4	19	260	33.1	77.7	7.8
5-6	17	359	25.1	72.4	4.2
7	7	201	40.3	93.5	5.0
8-10	7	242	19.0	71.9	2.6
>10	12	549	20.9	80.1	2.2

[a] Acc: fraction of correctly predicted contacts out of all predicted contacts.
[b] Acc ($|\delta|$=4): fraction of predicted contacts lying within one helix turn of an observed contact.
[c] Cov: fraction of correctly predicted contacts out of all observed contacts.

of more than 80% which is even slightly above the mean value found for all proteins (78.5%, Table 4.1). However, the best contact predictions within the dataset were obtained for proteins with seven transmembrane helices. Five out of the seven proteins in the dataset having this number of transmembrane segments belong to the superfamily of G protein-coupled receptor-like proteins according to the Pfam database [33]. Despite low sequence identity among each other, these proteins typically have a structure largely resembling the canonical alpha-helix bundle structure with only few helix-helix contacts between sequentially distant transmembrane helices [224], facilitating contact prediction for these targets.

Dependency of contact prediction performance on the number of selected contacts

Additionally, the dependency of prediction quality on the number of predicted contacts was evaluated. Figure 4.3 illustrates how the obtained prediction accuracy and the coverage depend on the cutoff for the number of analyzed contacts. While NN2 performs better than NN1 for all tested contact numbers, as do the two neural networks with membrane protein specific input features NN3 and NN4 compared to NN1 and NN2, the improvement of NN4 compared to NN3 is varying with the number of selected contacts. While for large numbers of predicted contacts NN3 and NN4 perform more or

less with equal accuracy and coverage, the highest improvement of prediction accuracy due to addition of LIPS scores as input features in NN4 is obtained for small numbers of predicted contacts (L/3 or less).

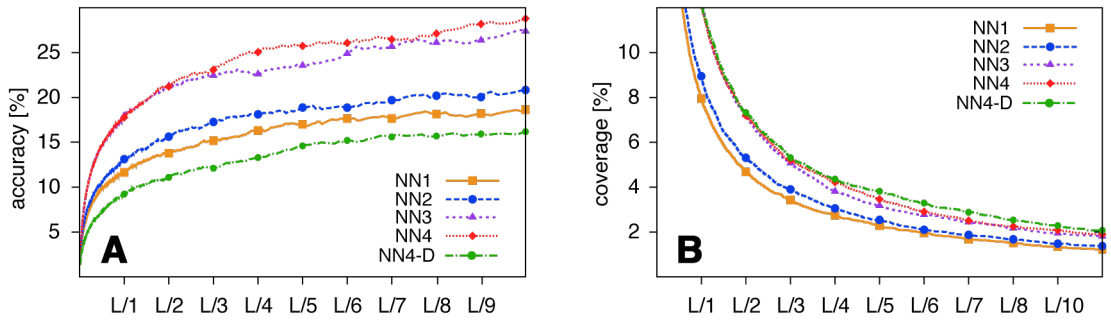

Figure 4.3: Contact prediction accuracy (A) and coverage (B) of different neural networks as a function of the number of predicted contacts (L/X). As expected, the accuracy is increasing while the coverage is decreasing for more stringent criteria. Both performance curves increase steadily for the NN1-NN3 architectures while the improvement of NN4 is less clear cut. L/X scaling for NN4-D is not comparable to the other architectures since the number of possible residue pairs and the number of observed contacts is different.

The same can be observed from Table 4.1 which also summarizes the quality of obtained predictions after selecting a number of predicted contacts determined using a contact formula derived from available membrane protein structures describing the number of expected contacts for a given number of participating residues (Materials and Methods, section 4.2.3). Obviously, a higher number of predicted contacts leads to a decrease in prediction accuracy in favour of an increased coverage. While the increase in prediction accuracy from NN1 towards NN3 is still clearly visible, NN3 and NN4 perform with more or less equal accuracy and coverage in this case.

4.3.2 Contact prediction in membrane proteins compared to soluble proteins

It is well known that the prediction of intra-molecular amino acid contacts gets increasingly difficult with decreasing contact density (fraction of observed contacts among the total number of possible residue pairs) [189]. This is the reason why contact predictions for large proteins are generally less successful than predictions for small proteins [211, 212] and why all-alpha soluble proteins, whose contact density is roughly only half of the contact density found for all-beta proteins [189], were consistently found to pose

special difficulties for the prediction. To evaluate the success of contact predictions obtained with TMHcon (NN4) at least at a very basic level to comparable results obtained for soluble proteins, the contact density was calculated for all membrane proteins in the MP_62 dataset. Figure 4.4 shows the dependency of the number of observed contacts in a protein on the protein length for four different types of proteins: soluble proteins, soluble proteins in the SCOP class all-alpha, the 62 membrane proteins of MP_62 (only transmembrane segments considered), and the 62 membrane proteins of MP_62 (again only transmembrane segments) where residue pairs lying on neighbouring helices were not considered. For all four datasets linear fits were calculated.

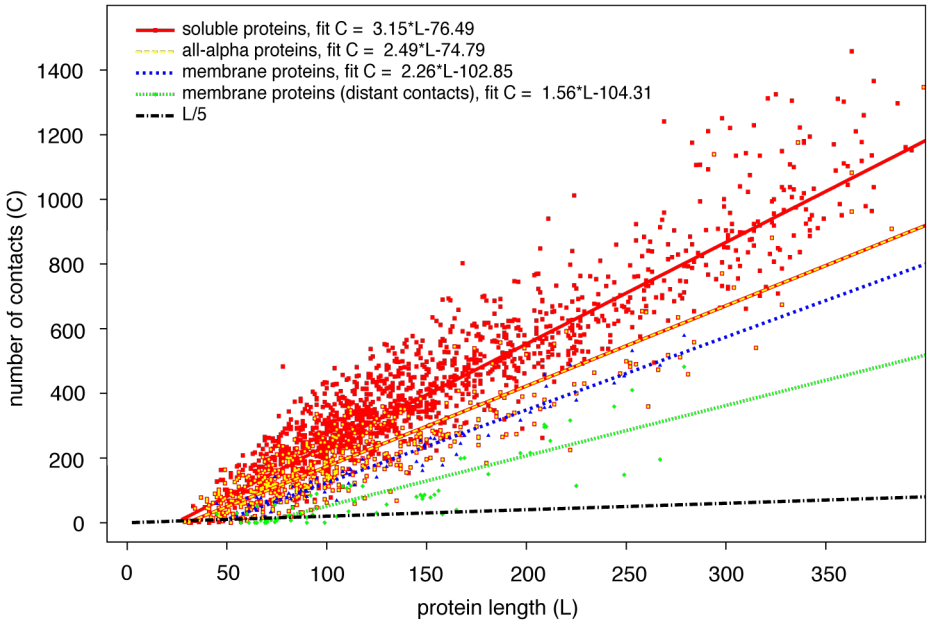

Figure 4.4: Contact density (number of contacts depending on protein length) of membrane proteins compared to soluble proteins. The amount of contacts for any type of protein is linearly proportional to protein length with membrane proteins having generally less contacts than soluble proteins. The fitted curves represent contact functions that can be used for the selection of an appropriate amount of contacts.

While all-alpha soluble proteins were found to possess slightly fewer contacts than soluble proteins in general, as was reported earlier [189], the number of observed contacts within membrane proteins was found to be even more reduced compared to soluble proteins in general and all-alpha soluble proteins in particular. When residue pairs on neighbouring helices were ignored, the number of observed contacts was further decreased significantly. These results indicate that the prediction of helix-helix contacts

in membrane proteins is at least of comparable difficulty to the prediction of intramolecular contacts within all-alpha soluble proteins, if not more difficult. Therefore, it is feasible to compare the performance measures of TMHcon to published values of available contact predictors for soluble proteins. Of special interest is the prediction of contacts within all-alpha soluble proteins due to their structural similarity to membrane proteins and their contact density being more similar to the contact density of membrane proteins than is observed for soluble proteins in general.

Compared to prediction accuracies reported for all-alpha soluble proteins (20% for a L/10 prediction based on 30 proteins at a minimal sequence separation of 8 [225], 24% for a L/2 prediction based on 131 proteins and a minimal sequence separation of 6 [189]), the developed contact predictor TMHcon for membrane proteins performs with equal quality to state-of-the-art methods for soluble proteins. This is also true for the prediction of long-range contacts. Using the neural network NN4-D, which predicts only contacts between non-neighbouring transmembrane helices, a prediction accuracy of 14.8% was obtained (Table 4.1). Reported values for all-alpha soluble proteins with a sequence separation of at least 24 amino acids range between comparable values of 13.5% (L/2 prediction,[189]) and 15.3% (L/10 prediction, [225]).

4.3.3 Comparison to other contact prediction methods

To further assess the benefit of the new contact prediction method specifically developed for membrane proteins, the obtained prediction results were compared to predictions obtained with available state-of-the-art contact prediction methods when applying these methods to the same set of membrane proteins. Despite the fact that these predictors were developed exclusively for soluble proteins, they might still be capable of detecting the contact patterns originating from the alpha-helical bundle structures of membrane proteins. Accordingly, predictions were obtained for all protein in `MP_62` using the contact predictor PROFcon [189], a neural network based predictor ranking among the best performing methods in the CASP6 competition [185], as well as using SVMcon [190], a contact map predictor based on support vector machines, one of the top predictors in the CASP7 experiment [186]. Since both methods returned predicted contacts for the full length sequence of each protein, all contacts lying outside the transmembrane parts or within the same transmembrane helix of a protein were not considered. From the remaining contacts, the top L/5 scoring ones were selected for every protein and every method (Table 4.3).

Using PROFcon, predictions could be obtained for 43 proteins out of the total set

Table 4.3: Contact predictions for 62 membrane proteins using external contact predictors, TMHcon (NN4) and the residue co-evolution consensus predictor HelixCorr. HelixCorr predictions correspond to CONSENSUS-R-5 predictions introduced in section 3.3.4 (page 59). All values are reported based both on the selection of the L/5 highest scoring residue pairs (L being the length of the concatenated transmembrane segments). TMHcon performs clearly better than other contact predictors developed for soluble proteins.

Predictor	N(Proteins)	N(Contacts)	L/5		
			Acc [%]	Acc ($\|\delta\|$=4) [%]	Cov [%]
HelixCorr	62	4822	10.8	51.9	4.4
ProfCon	24	503	4.2	36.8	0.2
SvmCon	62	1600	9.3	55.8	1.3
TMHcon	62	1611	25.9	78.5	3.5

[a] Acc: fraction of correctly predicted contacts out of all predicted contacts.
[b] Acc ($|\delta|$=4): fraction of predicted contacts lying within one helix turn of an observed contact.
[c] Cov: fraction of correctly predicted contacts out of all observed contacts.

of 62 proteins. However, since PROFcon restricts the number of returned contacts to 2L, the number of proteins with predicted contacts within their transmembrane helices was only 24. Based on the L/5 selection criterion, an average contact prediction accuracy of 4.2% was obtained for these 24 proteins. The accuracy ($|\delta|$=4) was found to be 36.8%. Despite these low values, PROFcon was still able to produce comparable results to TMHcon in individual cases with a maximum prediction accuracy of 21% and an accuracy ($|\delta|$=4) of 98% obtained for the ammonia channel AmtB (PDB 2NMR chain A). Using SVMcon, predictions were obtained for all 62 proteins. The average prediction accuracy was 9.3% and the prediction accuracy ($|\delta|$=4) was 55.8%, resulting in total in a clearly superior prediction compared to PROFcon without reaching the prediction accuracies observed with TMHcon. Again the obtained prediction quality was significantly differing among proteins with eight proteins having a prediction accuracy of 20% or more while 23 proteins were found with no correctly predicted contact at all. The best prediction using SVMcon was obtained for the sensory rhodopsin II with a prediction accuracy of 31% and an accuracy ($|\delta|$=4) of 97%. Based on these results it is clear that the development of a membrane protein specific contact predictor is in fact highly valuable since currently available contact predictors are not able to predict contacts within transmembrane helices over a large set of proteins with comparable prediction accuracy as on soluble proteins.

Comparing the neural network based predictions of TMHcon to predictions with the

earlier developed consensus co-evolution method HelixCorr, again TMHcon was found to be clearly superior than the predecessor method. With a prediction accuracy for helix-helix contacts of roughly 10% (Table 4.3), HelixCorr was easily outperformed by even the most basic neural network NN1 reaching a prediction accuracy of 17% (Table 4.1, page 75). The same is true for the prediction accuracy ($|\delta|$=4) where the most basic neural network achieves a more then 10% higher quality score than HelixCorr (64.6% compared to 51.9%). This observation is consistent with reported results for soluble proteins where the prediction of intra-molecular contacts was improved by at least 7% after using a neural network instead of correlated mutations alone [211].

4.3.4 Application of TMHcon to three membrane proteins with recently solved structure

To test TMHcon predictions of helix-helix contacts and interacting helices under 'real-life' conditions, the newly developed method was applied to three membrane proteins whose structure was solved after the construction of the test and trainings data set `MP_62`: the site-2 protease (PDB 3B4R chain B) [226], the sodium-potassium pump (PDB 3B8E chain A) [227] and the plasma membrane proton pump (PDB 3B8C chain A) [228]. None of these proteins had more than 30% sequence identity to any of the proteins in `MP_62`. Transmembrane helix positions determined from the 3D structure were obtained from PDBTM. Additionally, transmembrane helices were predicted using Phobius [98] to simulate the case when no protein structure is available. While Phobius predicted transmembrane helix number and position consistent with the PDBTM annotation in the case of 3B8C, one transmembrane helix was not detected in case of 3B4R, and two were missing in case of 3B8E. Subsequently, helix-helix contacts were predicted with TMHcon (NN4).

While for all three proteins an average prediction accuracy (L/5) for helix-helix contacts close to 20% was obtained for transmembrane helices taken from the PDBTM, this value decreased to only 13% in case transmembrane helices were predicted with Phobius. However, the fraction of predicted contacts within one helix turn of an observed contact was remarkably high both for transmembrane helices taken from PDBTM and predicted by Phobius, resulting in an even higher accuracy ($|\delta|$=4) than in our original data set (87.1% for Phobius, 86.3% for PDBTM). Therefore, the majority of all predicted contacts can be expected to be found on actual interacting helices both for transmembrane helices obtained from 3D structures or predicted by a topology prediction program.

5

Prediction of interacting helices

Membrane proteins show a relatively high structural simplicity compared to soluble proteins due to the severe structural constraints imposed by the lipid bilayer. Thus, in a first approximation, structure prediction of transmembrane domains is basically reduced to the question of how transmembrane segments interact along the membrane.

However, with more and more 3D structures of membrane proteins being available, it is now common understanding that alpha-helical membrane proteins may deviate remarkably from simple helix bundle structures [43]. Already in 1999, a study on helix-packing arrangements proposed a possible number of 1,500,000 different folds for a membrane protein with seven transmembrane helices [229]. Recent studies trying to classify the naturally occurring membrane protein fold space suggested a limited number of 250-500 different membrane protein folds [36, 35]. Nevertheless, the difficulty of membrane protein structure determination has led to the estimation that three more decades will be required to obtain a structural representation of 90% of the current membrane protein sequence space [35]. Therefore, the reliable prediction of helix interaction patterns may be a valuable tool to distinguish membrane proteins of different folds without knowing their structure or to assign a new protein sequence to a known membrane protein fold.

Here, different methods for predicting interacting helices in membrane proteins are introduced and the success of individual methods is compared to each other. First, co-evolving residues alone are used to identify which transmembrane helices of a membrane protein are likely to interact. As these residues were previously found to occur in close sequence neighbourhood to helix-helix contacts (section 3.3.3), they seem well suited for this predictive task. Within the next step, TMHcon contact predictions are applied for the identification of interacting helices. Similar to the prediction of helix-helix contacts, these predictions hopefully will increase again significantly the accuracy of predictions obtained with co-evolving residues alone. Finally, homologous proteins are incorporated

in the prediction process in order to obtain consensus predictions of interacting helices whose sensitivity is again increased compared to helix interactions obtained from single contact predictions.

Parts of this chapter were published in [162] and [210] while the consensus prediction of helix interaction graphs is still unpublished.

5.1 Introduction

5.1.1 Determinants of membrane protein folds

Within structural classification systems such as SCOP and CATH, protein folds are generally defined based on the number, spatial orientation and connectivity of secondary structure elements [230]. For soluble proteins, a large number of different fold architectures have been identified based on this definition, with estimations suggesting a total number of distinct globular folds existing in nature of around 1000 [231]. Applying the same fold definition to alpha-helical membrane proteins would lead to a small number of obtained membrane protein folds as mostly all membrane proteins having the same number of transmembrane helices would be classified to the same alpha-helix bundle fold due to the physical constraints imposed to membrane protein structures by the lipid bilayer.

On the other hand, membrane proteins are structurally divers to an extent not anticipated years ago (see also Introduction, section 1.2.4). According to a recent survey of structural complexity in membrane protein architectures, more than 30% of all analyzed structures contained re-entrant loops and/or incomplete helices within transmembrane domains, which again covered only 25% of all residues suggesting that extramembrane elements are also important determinants of membrane protein structures [232]. An analysis of cytoplasmic and extracellular loops for example suggested that more than 50% of all extramembrane loops are stretched thereby restricting the structural distance between two neighbouring transmembrane helices and supporting their potential to interact [233].

Nevertheless, helix-helix interactions are still believed to be the main characteristic of membrane protein structures as they also drive the folding process according to the two-stage model [90]. While helix-helix interactions between sequentially adjacent helices contribute to the canonical helix bundle structure, long-range interactions between non-neighbouring helices define an additional level of structural complexity between proteins

having the same number of transmembrane helices. Such long-range interactions have been shown to occur frequently and have been suggested to be essential for full assembly of the protein [233]. A complete overview of helix architectures arising from varying helix interactions however is still missing.

5.1.2 Graph visualization of helix architectures

For the analysis of transmembrane helix architectures a novel way of visualization is suggested here and used throughout the remaining parts of this thesis. Thereby, the transmembrane domain of a protein is represented by a graph where the transmembrane helices constitute the graph nodes while the interactions between pairs of helices correspond to the edges of this graph. Edges are weighted with the number of individual residue contacts between two helices (for two examples see Figure 5.1).

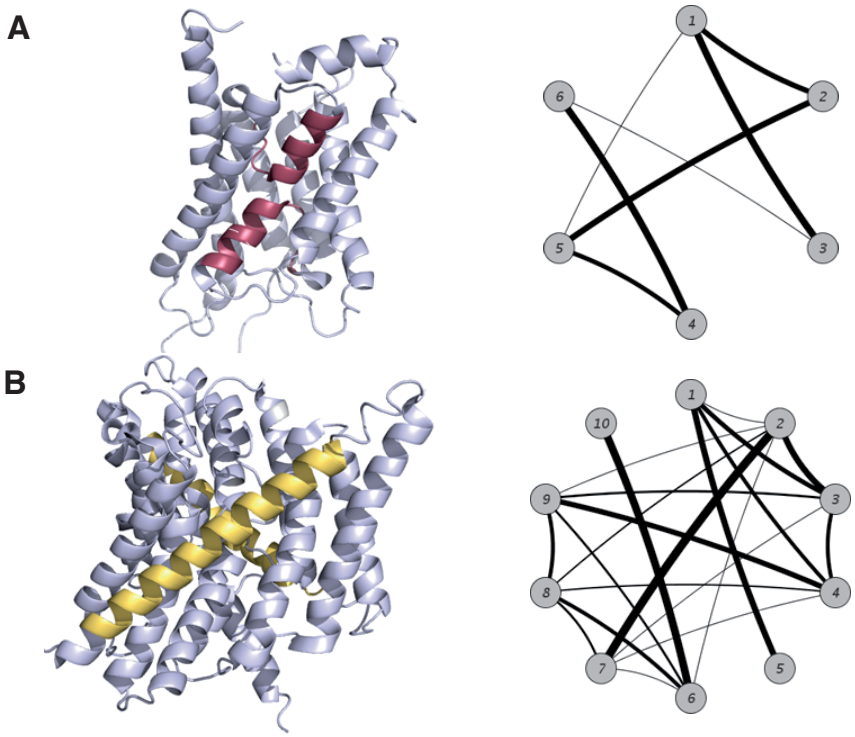

Figure 5.1: Example graph visualizations of helix architectures. (A) Structure of aquaporin 1 (PDB 1J4N) containing two re-entrant helices (shown in red) and six transmembrane helices. (B) Structure of the H(+)/Cl(-) exchange transporter clcA (PDB 1KPK) consisting of ten transmembrane helices. Strongly tilted helices with respect to the membrane are shown in yellow.

Such helix interaction graphs are well suited to visualize and compare the diversity of helix bundles consisting of the same number of transmembrane helices (see also Chapter 6). Furthermore, internal symmetries of membrane protein structures as observed in the case of aquaporin 1 (Figure 5.1A) are immediately approachable and protein substructures or domains can be identified as strongly connected subgraphs having few connections with remaining helices.

5.2 Materials and methods

5.2.1 Prediction of helix-helix-interactions using co-evolving residues

For the prediction of interacting helices based on co-evolving residues only, a dataset of helix-helix pairs was extracted from the membrane protein dataset `MP_14`. In total, 370 possible helix pairs were obtained considering predicted transmembrane helices from all proteins with at least four transmembrane segments. Predicted transmembrane segments were compared against transmembrane helix positions determined using structural information as obtained from the Protein Data Bank of Transmembrane Proteins (PDBTM) [201]. The comparison revealed a total number of 3 missing helices, 3 additionally predicted helices and 3 cases where segments were either joined or split (Table 3.1, page 46). These transmembrane segments were not included in the further analysis resulting in 325 helix pairs, of which 166 were considered to be in contact since they contained at least one residue pair having a minimal distance of less than 5.5Å.

Interacting helices were predicted using several consensus helix pair prediction methods where both the required number of correlated mutations for a predicted helix pair was varied and combinations of different prediction methods for correlated mutations were tested. In each case the sensitivity (TP/T) and specificity (TN/N) of the observed prediction was calculated and compared to a random prediction obtained by calculating the expected number of correctly predicted helix pairs based on the probability for a contacting helix pair. The significance of each prediction was evaluated using a chi-square test. As best predictions were obtained with the developed consensus predictor HelixCorr ('reduced' version, for details regarding HelixCorr see section 3.2.4 on page 51), prediction accuracies are only reported for this predictor.

5.2.2 Prediction of interacting helices with TMHcon

The prediction of interacting helices using TMHcon was evaluated using the dataset `MP_62`. For this dataset, a total number of 1486 helix pairs was obtained. Thereby, helix positions as determined from the 3D structure were directly obtained from TOPDB [202] or alternatively PDBTM [201] if a protein was not available in TOPDB. From this set of helix pairs, 714 helix pairs (48.0%) were considered to be in contact since they contained at least one residue pair corresponding to a helix-helix contact (spatial distance of maximal 5.5Å). For comparison, a second set of helix pairs was obtained where helix positions were predicted with the topology prediction program Phobius [98] instead of using helix positions determined from 3D structures. To this end, topology predictions were obtained for all proteins in `MP_62` and those predicted transmembrane helices were used for evaluation that overlapped by at least 50% of all positions with a transmembrane helix as listed in TOPDB/PDBTM. In total, the number of helix pairs was reduced due to mispredicted helices to 1212 helix pairs with 554 pairs (45.7%) connected by at least one helix-helix contact and hence classified as interacting helix pair.

Interacting transmembrane helices were predicted based on the number of residue contacts predicted for every helix pair with TMHcon using either one or both neural networks NN4 and NN4-D. To this end, an initial list of predicted contacts was compiled based on two different strategies, either using the protein length based L/5 criterion or employing the contact density formulas describing the number of observed contacts for a given number of residues obtained earlier (section 4.2.3, page 68). Subsequently, every helix pair was predicted as interacting for which a number of predicted contacts was found on this initial list exceeding a predefined threshold. Several thresholds for this required number of predicted contacts were evaluated by calculating the sensitivity and specificity of each obtained prediction. The significance of each prediction was calculated based on a chi-square test. In case both networks NN4 and NN4-D were used, two initial contact lists were obtained and different contact thresholds were tested for both networks resulting in an optimal trade-off between prediction sensitivity and specificity.

5.2.3 Consensus prediction of helix interactions

The prediction of helix interactions using additional contact information predicted for homologous sequences was executed based on the CAMPS database (version 2.0). Similar to the first CAMPS version used during the analysis of co-evolving residues in

membrane proteins (see section 3.2.1 on page 45), membrane proteins are classified in this database according to sequence and topology similarity into clusters thought to represent membrane protein folds. However, in addition to prokaryotic genomes covered in CAMPS1.0, the new CAMPS2.0 database also considers eukaryotic genomes. In total, 286,476 different proteins from 535 genomes are included which are classified into 1384 structurally correlated (SC-)clusters. As all proteins belonging to the same SC-cluster are expected to share the same fold and hence mostly the same helix interactions, these clusters seem to be an appropriate starting point for the consensus prediction of helix interactions.

For testing, all SC-clusters were selected containing a protein with at least 95% sequence identity (at 95% sequence coverage) to a PDB structure without considering theoretical models and structures with >4Å resolution. In case more than one representative structure was obtained for a given SC-cluster, their number of transmembrane helices was obtained from TOPDB (alternatively PDBTM if a protein was not present in TOPDB) and from all proteins with the same number of transmembrane helices the one with the best resolution was selected. The remaining list of representative structures was further filtered based on 40% sequence identity resulting in the final dataset consisting of 34 PDB proteins from 32 SC-clusters (Appendix, Table 9.4). Proteins 2R6G Chain F/2R6G Chain G and 3BEH Chain D/1ORQ Chain C originated from the same SC-cluster but differed in their number of transmembrane helices and hence were all included in the subsequent evaluation process as their helix interaction patterns are likely to be different. In the following, the final dataset will be referred to as `MP_CAMPS`.

The prediction of helix interaction consensus graphs from a number of structurally related proteins was executed and evaluated within several steps. First, TMHcon predictions were obtained for all proteins in `MP_CAMPS` based both on transmembrane helices as obtained from TOPDB/PDBTM and predicted with Phobius. Helix interactions were calculated from the TMHcon contact predictions as described above (section 5.2.2) and used later on as reference to evaluate the subsequent consensus predictions. For the actual consensus graph generation, related sequences were selected for each protein from the same SC-cluster which were required to have an identical number of transmembrane helices predicted with Phobius as the representative protein structure. The number of selected sequences was set to 40 during initial optimization experiments (data not shown) as a larger set of sequences was not found to improve prediction accuracy but significantly reduced calculation speed. In case SC-clusters were smaller than the required number of sequences, all appropriate sequences were collected. From this set of

structurally related proteins, a consensus helix interaction graph was obtained by calculating helix-helix contacts for each protein with TMHcon and predicting all helix pairs of the consensus graph as interacting that were also positively predicted in a predefined fraction of all analyzed proteins. To obtain optimal results, two main parameters of the full consensus prediction process were varied and adjusted to each other, first the number of helix-helix contacts required for the prediction of an interacting helix pair within a single protein and second the required fraction of positive predictions over all proteins. The parameter setting resulting in optimal prediction sensitivity at a given specificity was selected for the final consensus predictions. Finally, consensus predictions and predictions obtained for single PDB proteins were combined to evaluate the extent of overlap between these two prediction strategies and test whether individual predictions can be further improved by the addition of contact information from homologous proteins

5.3 Results and discussion

5.3.1 Prediction of interacting helices using co-evolving residues

Going from the residue contact level to the helix interaction level, the applicability of co-evolving residues for the prediction of interacting helix pairs in membrane proteins was first evaluated since co-evolving residues were found to frequently appear in close neighbourhood to helix-helix contacts. To minimize the number of incorrectly predicted interactions without loosing too much valuable information, the CONSENSUS-R approach was used for this analysis, which combines results from four prediction algorithms which were earlier found to perform best on membrane proteins (McBASC-Miyata, McBASC-McLachlan, OMES-KASS and CAPS, see Table 3.2).

Out of 325 helix pairs obtained from 14 membrane proteins (dataset `MP_14`, Table 3.1), the 166 actual interacting pairs were predicted with varying specificity and sensitivity, depending on the number N of correlated mutations required for a positive prediction (Table 5.1). For example, with N=5, i.e., where a helix pair is predicted as interacting in case at least 5 correlations were found for this helix pair, interacting helices could be predicted with a sensitivity of 42% and a specificity of 83%. A prediction accuracy of 71.9% could be achieved in this case. According to a chi-square test, this prediction is significant with a p-value of 2.19E-06. By raising the threshold N to higher values, the specificity of the prediction rises at the expense of a smaller number of predicted

interactions. For example, with N=7 about one fourth (27.1%) of all interacting helices in the dataset can be predicted with a specificity of around 92.5% (p-value 7.14E-06). The prediction accuracy increases to 78.9%.

Table 5.1: Prediction of interacting helices for the dataset `MP_14` based on co-evolving residues using a consensus approach. Sensitivity and specificity can be adjusted by varying the number of required correlations for the positive prediction of an interacting helix pair.

Required correlations	Accuracy [%]	Sensitivity [%]	Specificity [%]	P-value
1	56.8	78.3	37.7	2.30E-03
2	61.5	66.2	56.6	5.54E-05
3	64.1	56.0	67.3	3.85E-05
4	65.0	45.8	74.2	2.74E-04
5	71.9	41.6	83.0	2.19E-06
6	72.5	30.1	88.0	1.09E-04
7	78.9	27.1	92.5	7.14E-06
8	78.3	21.7	93.7	1.33E-04
9	84.8	16.9	96.9	9.20E-05
10	88.5	13.9	98.1	1.62E-04

Applying the same procedure to the dataset `MP_62` (Appendix Table 9.3) resulted in predictions with similar sensitivity and specificity (Table 5.2, page 93). Here, 42.9% of all interacting helices could be predicted with a specificity of 79.8% while 30.4% of all interacting helices could be predicted with a specificity of 90.3% (see Table 5.2 on page 93). For these predictions, accuracies of 66.2% and 74.2%, respectively, were achieved.

These results demonstrate that co-evolving residues in fact can be used to identify interacting helices with good accuracy although most of them are not residue contacts themselves. However, especially the sensitivity of the prediction leaves further room for improvements.

5.3.2 Improved prediction of interacting helices with TMHcon

After demonstrating the capability of TMHcon to predict helix-helix contacts in membrane proteins with equal accuracy to state-of-the-art methods for soluble proteins (Chapter 4), the potential application of these predicted contacts for the identification of interacting helices should be evaluated and compared to the results obtained with co-evolving residues alone (section 5.3.1).

Based on the dataset of 62 proteins (`MP_62`) used also for helix-helix contact prediction, interacting helices were predicted from helix-helix contacts using two different strategies.

On the one hand, the L/5 highest scoring contact pairs were selected (with L being defined as the sum of the transmembrane segments' lengths) and every helix pair was predicted as interacting with at least one predicted contact. However, as can be seen from Figure 4.4 (page 79), the solely sequence length dependent threshold L/5 is much too restrictive to obtain a number of contacts typical for an alpha-helical membrane protein. Additionally, it can be observed that the number of contacts per helix pair predicted by the neural network NN4 tends to increase with the number of observed contacts per helix pair (Figure 5.2). After selecting predicted contacts based on the contact density formula for alpha-helical membrane proteins introduced in section 4.2.3 (page 68), helix pairs with more than 5 actual helix-helix contacts were found to have on average 15 predicted contacts (median: 8) while helix pairs with only a small number of helix-helix contacts between one and five had nine predicted contacts on average (median: 3).

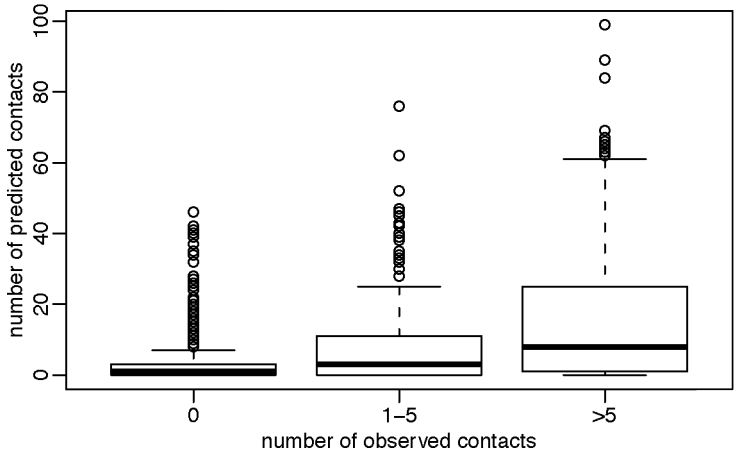

Figure 5.2: Dependency of the number of contacts predicted with TMHcon (neural network NN4) on the number of observed contacts. Helices with a given number of observed helix-helix contacts (0, 1-5, or more than 5) were grouped. The number or predicted contacts increases in average with the number of observed contacts.

Based on these observations a second prediction strategy was developed for interacting helices where the initial number of predicted contacts for every protein was derived from the contact density formula. Afterwards, a threshold of required contacts for an interacting helix pair was applied to remove wrongly predicted interacting helices. Similar to the approach introduced in section 5.3.1 where co-evolving residues alone were used for the identification of interacting helices, this contact threshold can be used to

achieve predictions of increasing specificity at the cost of decreasing sensitivity. Using these two strategies (termed length-based prediction and contact-based prediction) interacting helix predictions were obtained for all four neural networks developed earlier (see section 4.2.4, page 68) and specificity, sensitivity, accuracy and significance (based on a chi-square test) were calculated for each prediction (Table 5.2). Since contact-based predictions made at different thresholds of required contacts are hard to compare, results are always provided for those predictions having the specificity closest to 80% and 90%. As can be seen from Table 5.2, the contact-based selection resulted in a more significant prediction for all of four neural networks than the length-based selection.

Prediction performance of neural networks with increasing complexity

A comparison of the performance of different neural networks produced similar results to those obtained in the analysis of predicted helix-helix contacts (Table 5.2). Predictions based on the same selection strategy showed a clear increase in accuracy and sensitivity (accompanied by decreasing p-value) at the same specificity level with increasing complexity of the used neural network. For example, using length-based (L/5) selection, all four neural networks resulted in a prediction of interacting helices with a specificity between 87% and 89%, while the prediction accuracy increased from 72% towards 78%. At the same time, the sensitivity increased by 9%, and the p-value decreased from 1.76E-25 to 3.72E-46. The same can be observed using the contact-based selection strategy. When predictions with the same specificity (for example 90%) were compared, again accuracy and sensitivity increased (3% and 8%, respectively, in the case of 90% specificity) while the p-value decreased (from 4.43E-34 towards 8.48E-51, again for predictions with 90% specificity).

Prediction of interacting helices distant in sequence

For every prediction, the fraction of predicted interacting helices that are neighbouring in sequence were calculated. Despite all deviations from the canonical alpha-bundle structure found in membrane proteins, neighbouring helices still have a clearly higher probability for interaction with each other compared to non-neighbouring helix pairs (80.5% compared to 37.9% for non-neighbouring helix pairs in the dataset `MP_62`). Therefore, a primitive way of predicting interacting helices in membrane proteins would be to predict all neighbouring helices as interacting and non-neighbouring helices as not interacting. While this prediction method would lead to a high prediction accuracy of 80.5% for `MP_62`, its subsequent application for the discrimination of different membrane protein

Table 5.2: Prediction of interacting transmembrane helices using helix-helix contacts predicted by neural networks of increasing complexity. All results are based on the dataset MP_62. For comparison, results obtained with HelixCorr (the consensus predictor based on correlated mutations) are reported as well. Predicted contacts used for the identification of interacting helices were selected with two different procedures. L/5 corresponds to the length based selection of predicted contacts while CX describes the number X required contacts for an interacting helix pair after compiling an initial list of contact predictions using the contact density formula for membrane proteins described in section 4.2.3.

Method	Threshold	N(predicted)[a]	Neighbour[b] [%]	Accuracy [%]	Sensitivity [%]	Specificity [%]	P-value
HelixCorr	C7	462	38.1	66.2	42.9	79.8	7.36E-21
	C11	292	42.8	74.3	30.4	90.3	2.35E-23
NN1	L/5	359	60.4	72.1	36.3	87.0	1.76E-25
	C4	494	57.9	71.3	49.3	81.6	2.71E-36
	C9	336	73.2	77.4	36.4	90.2	4.43E-34
NN2	L/5	327	65.1	76.1	34.9	89.9	2.28E-30
	C3	531	57.1	71.8	53.4	80.6	5.06E-42
	C7	366	72.4	79.5	40.8	90.3	2.05E-43
NN3	L/5	380	69.7	76.1	40.5	88.2	1.97E-36
	C4	565	59.3	72.9	57.7	80.2	1.04E-50
	C10	373	79.6	79.6	41.6	90.2	8.39E-45
NN4	L/5	413	65.4	78.0	45.1	88.2	3.72E-46
	C4	587	57.1	72.2	59.4	78.9	5.10E-51
	C9	397	75.6	80.4	44.7	89.9	8.48E-51
NN4-D	C7	324	-	58.0	43.8	80.7	1.76E-18
	C10	212	-	66.5	32.9	89.9	3.49E-21
NN4/ NN4-D	C9/C10	552	54.3	74.8	57.8	82.0	2.09E-56
	C9/C15	485	61.9	78.1	53.1	86.3	2.24E-58

[a] N(predicted): number of predicted interacting helices.
[b] Neighbour: percentage of neighbouring helix pairs out of the total number of predicted interacting helices.

folds would be impossible, since no differences in the helix packing of proteins with the same number of transmembrane helices could be determined. Optimally, one would therefore wish to obtain predictions with a small fraction of neighbouring helices (possibly close to the naturally occurring fraction of 39.9% as observed for the tested dataset MP_62), to get a maximum of information about the specific fold of a protein.

A comparison of NN1 and NN2 (Table 5.2) reveals that the incorporation of correlated mutations as input feature results in predictions of higher sensitivity and accuracy at equal specificity with a slightly smaller fraction of neighbouring helices in the set of pre-

dicted helices using the contact-based selection (with 90% specificity 73.2% neighbouring helices with NN1 and 72.4% with NN2). The additionally detected interacting helices are therefore primarily long distance helix pairs, implying that co-evolving residues are generally independent of sequence separation (see also the discussion of HelixCorr results later).

In contrast, the first incorporation of membrane protein specific features (residue position within the transmembrane helix as well as the total number of transmembrane helices) within NN3 resulted in a strong increase of the number of neighbouring helices in the prediction (at 90% specificity 79.6% with NN3 compared to 72.4% with NN2). This demonstrates a general tendency of the neural network to learn about the helix-bundle structure of membrane proteins from basic membrane protein specific input features. The addition of LIPS scores within NN4 reduces the fraction of neighbouring helices again to a final value of 75.6% for the prediction with 90% specificity. Since the fraction of falsely predicted non-interacting neighbouring helices decreases at the same time (from 17.2% with NN3 towards 15.3% with NN4), the inclusion of LIPS scores (the predicted orientation of each residue towards the membrane or the protein interior) seems to prevent the incorrect prediction of those amino acid residues as being in the contact state which would originally be well positioned on neighbouring helices to form a contact in a perfect helix bundle structure.

In order to increase the fraction of non-neighbouring helices in the final prediction a neural network was trained especially on long-range contacts by omitting all helix-helix contacts from neighbouring helices from the training set (network NN4-D). Using contacts predicted by this neural network and selected according to the contact formula derived for non-neighbouring helices (see section 4.2.3) a prediction of distant interacting helices was obtained. Due to the increased difficulty of predicting contacts on non-neighbouring helix pairs resulting from the smaller contact density, the sensitivity and accuracy of this prediction was clearly lower than those obtained for the full dataset (Table 5.2). However, at 80% specificity still 43.8% of all distant interacting helices could be correctly predicted. More than 32% of these interacting helices were predicted with close to 90% specificity. To enhance the original NN4 prediction with long distant interactions, helix pairs predicted as interacting by NN4-D were combined with the initial NN4 prediction. After adding all helix pairs with at least 10 predicted contacts (corresponding to the 90% specificity prediction of NN4-D), the significance of the prediction increased to 2.1E-56 (Table 5.2). While still 57.8% of all interacting helices were predicted with a specificity of 82%, the fraction of neighbouring helices

decreased to only 54.3%. This prediction was further improved by raising the threshold of required contacts for NN4-D, corresponding to the increased difficulty of long-range contact prediction. With 15 required contacts a final prediction with a significance of 2.2E-58, a sensitivity of 53.1% and a specificity of 86.3% was obtained. The fraction of neighbouring helices was only 61.9%, a clear improvement compared to the original NN4 prediction.

Comparison to predictions based on co-evolving residues

Comparing the results obtained with the combination of the neural networks NN4 and NN4-D to the results obtained with co-evolving residues, the increase in prediction quality from HelixCorr towards TMHcon was quite remarkable (Table 5.2).

For predictions with both 80% and 90% specificity, TMHcon predictions with basically equal specificity to comparable HelixCorr predictions resulted in a clearly higher sensitivity and accuracy. An increase in accuracy of up to 6% (HelixCorr with 7 required correlations (C7) compared to TMHcon with 4 required contacts (C4)) and an increase in sensitivity of up to 16% (again HelixCorr/C7 compared to TMHcon/C4) was observed. The significance of the prediction increased from 7.4E-21 to 5.1E-51. However, it must be noted, that the fraction of neighbouring helix pairs is significantly lower in the case of HelixCorr compared to any prediction obtained by a neural network (42.8% with HelixCorr/C11 compared to maximal 79.6% with NN3/C10). While neural networks tend to learn that neighbouring transmembrane helices have a higher probability for interacting with each other, co-evolving residues are much more independent of this fact. Since the obtained predictions favour specificity over sensitivity, resulting in a limited number of predicted interacting helix pairs, this leads to an enrichment of neighbouring helices in the prediction of the neural networks. In contrast, the prediction from HelixCorr with a fraction of close to 40% neighbouring helices resembles nearly perfectly the naturally occurring fraction of 39.9% neighbouring helices in the total set of interacting helices (285 out of 714).

Prediction of interacting helices based on predicted transmembrane helices

Experimentally derived membrane protein topologies are only available in rare cases similar to membrane protein structures. Hence, in most cases transmembrane helices need to be predicted using state-of-the-art topology prediction programs. Such programs can be expected to predict the correct topology of a protein with an accuracy of roughly 70% in case no experimental constraints are available [43]. A noticeable fraction of

all proteins will therefore be predicted with an incorrect number or at least slightly misplaced transmembrane helices. To test how this affects the prediction of interacting helices, a second dataset of helix pairs was obtained from the dataset MP_62 where transmembrane helices were predicted using the program Phobius [98]. For evaluation, all predicted helices were used that overlapped by at least 50% of all positions with an observed transmembrane helix thereby allowing a distinct degree of misplacement. Nevertheless, the final dataset of helix pairs was reduced from 1486 helix pairs to 1212 helix pairs due to incorrectly predicted transmembrane helices.

Interacting helices were predicted once relying on the length-based approach (using the best L/5 helix-helix contacts) and once relying on the best performing contact-based approaches where predictions from networks NN4 and NN4-D were combined. Using the length-based approach, prediction quality was noticeably inferior when using predicted instead of structurally determined transmembrane helices (Table 5.3). At equal specificity, accuracy decreased by more than 2.5% and sensitivity decreased by more than 4%. Using the contact-based approaches C9/C10 and C9/C15 the decrease in prediction quality was less eminent as predictions using the same prediction parameters were clearly more sensitive but less specific in case of predicted transmembrane helices.

Table 5.3: Prediction of interacting helices with TMHcon based on predicted transmembrane helices. All results are based on the dataset MP_62. Predicted contacts used for the identification of interacting helices were selected either based on protein length (L/5) or using the contact density formulas for membrane proteins described in section 4.2.3 (CX, with the number X describing the number of required contacts for an interacting helix pair). While L/5 based predictions are less sensitive but equally specific using predicted transmembrane helices, contact density formula based predictions are shifted towards higher sensitivity at reduced specificity.

Method	Threshold	TMS[a]	Accuracy [%]	Sensitivity [%]	Specificity [%]
NN4	L/5	TOPDB	78.0	45.1	88.2
		Phobius	75.3	40.8	88.8
NN4/NN4-D	C9/C10	TOPDB	74.8	57.8	82.0
		Phobius	63.5	63.5	77.5
NN4/NN4-D	C9/C15	TOPDB	78.1	53.1	86.3
		Phobius	73.5	56.0	83.0

[a] TMS: transmembrane helix positions were obtained either from experimentally determined structures (TOPDB) or using the topology prediction program Phobius.

Overall, the prediction of interacting helices seems to be more robust than the predic-

tion of helix-helix contacts when using predicted instead of experimentally determined transmembrane helices (see also sections 4.3.4 and 5.3.2). Missing or additionally predicted helices may prompt the neural network to deduce wrong residue contacts since on the one hand neighbouring helices may appear as non-neighbouring and vice versa. Furthermore, misplaced transmembrane helices may obscure the correct position of a residue with respect to the membrane (a residue close to the cytoplasmic side for example may appear as one of the central residues) complicating the detection of correct helix-helix contact positions. Interestingly, this seems to affect the prediction of interacting helices only when the number of used residue contacts is small (L/5 predictions). For contact-based predictions where larger numbers of helix contacts are used in the first place, losses in specificity are largely balanced by gains in sensitivity indicating that even incorrectly predicted residue contacts are still preferentially lying on interacting helices. The shift towards more sensitivity and less specificity seems to be mainly an artefact caused by predicted transmembrane helices being generally slightly longer than corresponding helices determined from the 3D structures. Therefore, more predicted helix-helix contacts are selected in the first place increasing prediction sensitivity while reducing prediction specificity at the same time.

Prediction of interacting helices for three membrane proteins with recently solved structure

To exemplary evaluate the prediction of interacting helices under 'real-life' conditions, the same three membrane proteins were used as introduced in section 4.3.4 (page 82) when testing contact predictions obtained with TMHcon. Again, transmembrane helix positions were obtained once from PDBTM and once from predictions with the topology program Phobius [98] to simulate the case when no protein structure is available. Helix-helix contacts were predicted both with NN4 and NN4-D (for results see section 4.3.4) and then used for the prediction of helix-helix interaction patterns for all three proteins. The same prediction parameters were used as in the most significant earlier prediction, thus requiring at least 9 predicted contacts by NN4 or 15 predicted contacts by NN4-D to predict a helix pair as interacting.

As can be seen from Figure 5.3, predicted helix interaction graphs closely resemble the expected interaction patterns although several edges may be missing. This is even true for proteins such as 3B4R (chain B) and 3B8E (chain A) where Phobius was not able to predict the correct number of transmembrane helices.

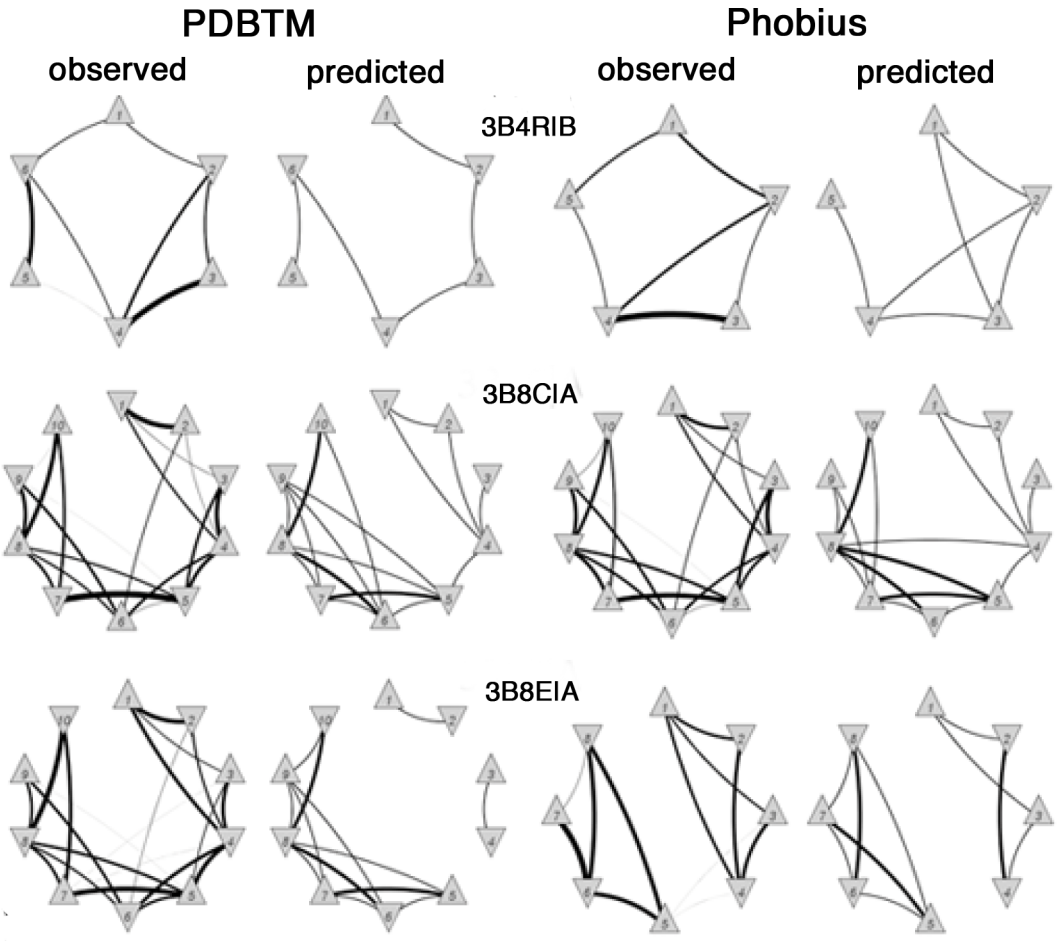

Figure 5.3: Prediction of interacting helices for three example membrane proteins (3B4R chain B, 3B8C chain A, 3B8E chain A). Helices were predicted as interacting with at least 9 or 15 helix-helix contacts predicted by NN4 or NN4-D, respectively. Bold edges in the case of predicted helix interactions indicate positive predictions by NN4 and NN4-D while in the case of observed helix interactions they correlate with the number of observed residue contacts. For all proteins, predicted helix interaction graphs closely resemble observed helix interaction graphs.

5.3.3 Prediction of consensus helix interaction graphs

Information from homologous proteins has been successfully incorporated into a number of structural prediction tasks in the past (see for example [234, 235]). Here, a similar approach is presented trying to recover helix interactions of a given protein using predicted contact information from a number of related proteins. Furthermore, the possibility is evaluated whether such homologous proteins can also be helpful for improving the prediction quality of helix interaction graphs by combining individual predictions into a

consensus graph of increased sensitivity.

Ideally, proteins used for the construction of consensus helix interaction graphs should be sequentially diverse (to add new information to the prediction process) while still folding into the same helix architecture with highly similar helix interactions. The CAMPS database (version 2.0) was used to identify such proteins as membrane proteins are classified in this database into clusters corresponding to membrane protein folds (termed SC-clusters) based on sequence and topology similarity (manuscript *in preparation*). In total, 34 proteins could be identified that are associated with a known 3D structure and are classified within CAMPS either into different SC-clusters or the same SC-cluster but vary in their number of transmembrane segments (dataset `MP_CAMPS`, Appendix Table 9.4). For each protein, 40 related proteins from the same SC-cluster were selected whose predicted number of transmembrane segments was consistent with the representative protein structure. These proteins were used then to calculate consensus interaction graphs representing the corresponding protein structure (for details see Materials and methods, section 5.2.3 on page 87).

Reproduction of helix interaction patterns from structurally related sequences

After obtaining consensus helix interaction graphs for all proteins in `MP_CAMPS`, the accuracy of these graphs was calculated and compared to predictions obtained for the original PDB sequences of `MP_CAMPS`. Additionally, the individual helix graphs used for constructing the consensus graph were evaluated against the observed helix graphs and the average prediction accuracy, sensitivity and specificity of these graphs was determined and again compared to the final consensus graphs (Table 5.4).

Independent of the strategy applied for predicting helix interactions from helix-helix contacts - length-based (L/5) vs. contact-based (CX/CX), for further explanations see section 5.3.2 - consensus graphs were capturing observed helix interactions at least equally good as predictions obtained from the PDB sequences directly and clearly better than the average helix prediction based on proteins used for the construction of the consensus graphs. Using for example the best L/5 predicted residue contacts and predicting all helices as interacting with at least one helix-helix contact, the sensitivity of consensus graphs was even nearly 9% higher (at only slightly reduced specificity) than both the PDB sequence based predictions and than the average of all individual consensus sequence based predictions. For contact-based predictions with 90% specificity, consensus predictions are still 1.5% more sensitive than the single PDB protein predictions and 5% more sensitive than the average single consensus sequence prediction.

Table 5.4: Consensus prediction of interacting helices using structurally related sequences. All results are based on the dataset MP_CAMPS. Helix interactions were predicted either based on individual proteins having a PDB structure or within a consensus approach using 40 proteins structurally related to the representative PDB structure (termed consensus sequences). In all cases, transmembrane helices were predicted with Phobius. Helix interactions predicted with the consensus graph are at least equally good as predictions obtained from the PDB sequences directly and clearly superior to predictions obtained for single consensus sequences.

Graph type	Contact threshold[a]	Consensus threshold[b]	Accuracy [%]	Sensitivity [%]	Specificity [%]
PDB[c]	L/5	-	76.1	40.7	89.3
	C6/C15		71.4	59.0	80.3
	C20/C17		78.1	44.7	89.5
Consensus	L/5	0.3	77.3	49.5	87.6
	C9/C18	0.3	72.4	60.4	80.3
	C17/C23	0.4	80.2	46.2	90.2
Average[d]	L/5	-	74.7	40.5	88.2
	C9/C18		73.0	52.8	83.3
	C17/C23		77.9	41.3	90.0

[a] Contact threshold: strategy to predict helix interactions for a single protein (for details see Table 5.2). For contact-based predictions optimal results at 80% and 90% specificity are shown.

[b] Consensus threshold: fraction of single helix graphs required to contain a specific edge to transfer it to the consensus graph.

[c] PDB: evaluation of all helix predictions obtained directly from the corresponding PDB sequences with transmembrane helices being predicted with Phobius.

[d] Average: evaluation of all individual helix predictions obtained for homologous proteins used during consensus graph generation. All helix predictions were evaluated against the corresponding PDB structure and the average accuracy, sensitivity and specificity of these predictions was calculated.

This is highly encouraging as proteins used during consensus graph construction generally shared only minor sequence identity with the PDB structure whose helix interactions should be predicted. In nearly 80% of all cases, the average sequence identity between PDB structure and all consensus sequences was below 40%. First, this demonstrates that CAMPS SC-clusters in fact contain proteins whose helix architectures are likely to be highly similar as this prediction success would hardly be possible otherwise. Furthermore, helix interaction predictions of individual proteins can be combined into a common prediction thereby filtering out wrongly predicted helix interactions and/or gaining additional correct helix predictions as can be seen from the increased sensitivity of consensus predictions (more than 5% gain for all presented contact thresholds) compared to the average predictions of individual consensus sequences.

Adding consensus information to PDB helix interaction graphs

Although consensus graphs were found to reproduce membrane protein helix architectures at least equally good as predictions obtained from the PDB sequences themselves, their prediction accuracy especially for contact-based helix predictions was also not significantly better leaving room for further improvement. To this end, helix interactions predicted from PDB sequences directly were combined with helix interactions obtained with the consensus approach in order to test the overlap between these predictions and evaluate the potential of further prediction improvements. Prediction parameters of individual helix predictions (number of initially selected helix-helix contacts and number of required contacts for predicting a helix interactions) were varied conjointly with the consensus threshold (fraction of single helix graphs required to contain a specific edge to transfer it to the consensus graph) until optimal prediction conditions were found (Table 5.5).

Table 5.5: Combining PDB helix interaction graphs with consensus information. All results are based on the dataset `MP_CAMPS`. Helix interactions predicted for individual proteins having a PDB structure were enriched with additional helix interactions predicted with a consensus approach using 40 proteins structurally related to the representative PDB structure. In all cases, transmembrane helices were predicted with Phobius. By adding consensus information to PDB predictions, prediction sensitivity can be improved by 4 - 13%.

Graph type	Contact threshold[a]	Consensus threshold[b]	Accuracy [%]	Sensitivity [%]	Specificity [%]
PDB[c]	L/5	-	76.1	40.7	89.3
	C6/C15		71.4	59.0	80.3
	C20/C17		78.1	44.7	89.5
PDB+[d]	L/5	0.4	76.3	53.3	86.5
	C21/C13	0.6	72.8	63.2	80.3
	C21/C20	0.6	79.5	49.9	89.3

[a] Contact threshold: strategy to predict helix interactions for a single protein (for details see Table 5.2). For contact-based predictions optimal results at 80% and 90% specificity are shown.
[b] Consensus threshold: fraction of single helix graphs required to contain a specific edge to transfer it to the consensus graph.
[c] PDB: prediction of helix interactions based on the corresponding PDB sequences alone.
[d] PDB+: prediction of helix interactions by combining consensus information with predictions obtained for the corresponding PDB sequences.

In fact, the prediction of helix interactions can be further improved when adding consensus information. Using the best L/5 helix-helix contacts for all individual helix predictions, the sensitivity increases by nearly 13% after the addition of consensus pre-

dictions while the specificity decreases only by 3%. Compared to consensus predictions alone (Table 5.4), the sensitivity still increases by 4% at equal specificity. Similarly, using contact-based helix predictions the sensitivity of predictions with 80% and 90% specificity rises by 4% and 5%, respectively. This demonstrates, that consensus helix interaction graphs on the one hand are largely consistent with interaction graphs predicted for single PDB proteins since the increase in sensitivity is rather moderate. However, at least a small number of helix interactions is obtained from consensus predictions that can not be derived from the PDB sequences themselves. When appropriately combined into a common prediction, these additional interactions can contribute to increased prediction sensitivity while not reducing prediction specificity. Figure 5.4 demonstrates this increase in sensitivity for the example of bovine rhodopsin (PDB 1U19, chain A), the first G-protein coupled receptor with experimentally solved structure [224]. In Figure 5.4A the helix interactions as obtained from the structure are shown with the edge weights corresponding to the number of observed helix-helix contacts. Using only the PDB sequence (Figure 5.4B), all interactions between sequentially neighbouring helices are correctly predicted but only one distant interaction (between helices 3 and 7) is obtained. Adding further consensus information (Figure 5.4C), one interaction between neighbouring helices ist lost (helices 5 and 6) since more restrictive contact thresholds are applied for individual predictions but three additional distant interactions are gained which significantly contribute to an improved reproduction of the original helix architecture.

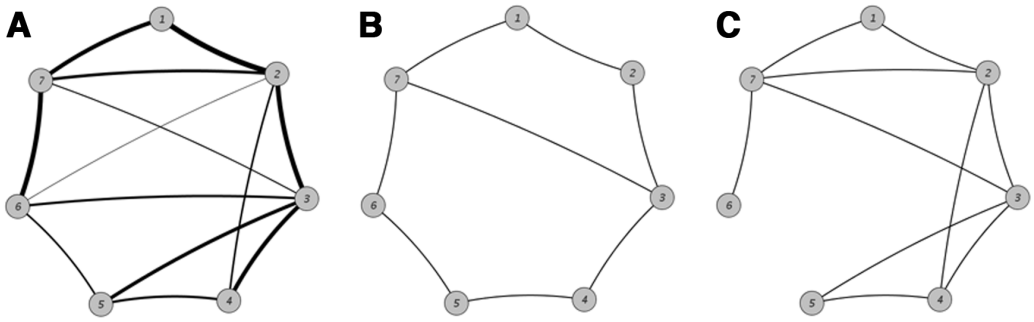

Figure 5.4: Consensus prediction of helix interactions for bovine rhodopsin (PDB 1U19, chain A). All predictions are obtained with the contact-based approach using optimal prediction parameters determined for a minimal prediction specificity of 80%. (A) Helix interactions as observed from the 3D structure. (B) Helix interactions predicted from the PDB sequence alone. (C) Helix interactions predicted with additional consensus information. Several interactions between non-neighbouring helices are predicted only by the consensus approach.

6

Classification of helix architectures

For nearly two decades, structural classification approaches try to organize the jungle of available protein structures. For soluble proteins, between 80-90% of all protein domains originating from completely sequenced genomes can already be classified to a known structural family indicating that the fold space of soluble proteins is already fairly well covered in available structure-based classification databases and new folds are less likely to be identified [236, 237]. Structural classification of membrane proteins on the other hand is still it its infancies due to the overall small number of available experimentally solved structures. However, due to the technological increase in structure determination of membrane proteins [238, 239], the number of available unique membrane protein structures has been significantly growing over the past years allowing now for a first glimpse into the structural universe of membrane proteins based on experimentally determined structures.

Still unclear is, to what extent structural classification procedures derived originally mostly for soluble proteins can be directly applied to membrane proteins. Difficulties result especially from the lack of a uniform fold definition for membrane proteins as all integral membrane proteins on the one hand adopt either an overall alpha-helix bundle or beta-barrel architecture but vary within several structural features such as helix interactions, loop lengths or the presence of structural irregularities like tilted helices and reentrant loops. Furthermore, the continuity of protein fold space has been heavily discussed over the past years (see for example [240, 241, 242]). For membrane proteins whose structures are additionally restricted by the lipid bilayer, the question arises whether membrane proteins can be reasonably classified into distinct folds at all.

Here, these problems are addressed by proposing a completely membrane protein specific structural classification system that identifies common helix architectures based on similarities between helix interaction graphs. Within a short introduction, available

structural classification approaches are first briefly reviewed and the hotly debated idea of a potentially continuous fold space is introduced together with the resulting implications for the classification of protein structures. Then, a short analysis of occurrence and classification of membrane proteins in the structural databases SCOP [243, 223] and CATH [244, 237] is presented summarizing the current status of membrane protein structural classification. Finally, the major part of the chapter describes the new classification system which clusters membrane proteins based on similarity scores obtained from comparing their helix interaction graphs. After testing how well obtained helix architectures agree with protein folds as obtained from SCOP and CATH, several questions regarding the diversity of alpha-helical membrane proteins are addressed such as 1) how diverse can membrane protein sequences be to still fold into a common helix architecture or 2) is the protein folding space of membrane proteins rather discrete or continuous according to the similarity among all helix architectures containing the same number of transmembrane helices. Finally, it is evaluated, whether the accuracy and sensitivity of predicted helix interaction graphs is already sufficient for classifying proteins without experimentally solved structure available.

The analysis of membrane proteins in SCOP and CATH was conducted conjointly with Sindy Neumann (TU München) and was recently submitted for publication [245]. Comments in the section Material and methods will clarify individual contributions. The helix interaction based classification of membrane proteins is still unpublished.

6.1 Introduction

6.1.1 Structural classification of proteins

Structure-based classification of proteins provides a helpful resource to reveal their evolutionary relationships and to obtain an easily accessible overview of the existing protein fold space and the number of naturally observed folds. In addition, classification approaches have found widespread application in many areas of structural bioinformatics, including homology modelling, fold recognition, and structural genomics.

Several resources for structural classification of proteins exist, with SCOP [243, 223] and CATH [244, 237] arguably being the most comprehensive ones. Both databases use a hierarchic classification system and rely on a largely similar definition of a protein fold which takes into account the number of secondary structure elements, their spatial orientation, and connectivity [243, 244]. While SCOP and CATH incorporate different levels

of manual supervision in their classification procedure, other classification approaches have been proposed that completely rely on large-scale structure comparisons and are therefore fully automated [246, 240, 247, 248, 249].

Given the varying classification procedures it is not surprising that several studies have found remarkable differences between individual fold classifications [250, 240, 251]. First, these differences may arise from variations in the applied domain assignment procedure, which generally is the first step within each classification approach since structural domains are used as classification entities. Furthermore, classification databases may disagree in their fold and homology assignments. Large folds in one database might be divided into several more specific folds within another classification system, leading to proteins belonging to the same fold in the first case but to different folds in the latter case. Even more drastic discrepancies can be observed where one database classifies two proteins into an evolutionary related family while another classification approach places the same pair of proteins into completely different folds due to the fact that proteins may be structurally diverse despite a common evolutionary origin [250, 252].

However, all comparative analyses of structural classification systems have been executed on datasets consisting of mostly soluble proteins. A comparable analysis specifically focusing on membrane proteins is still missing.

6.1.2 The protein fold space: discrete or continuous?

While many disagreements between different structural classification approaches can be directly attributed to differences in their classification methodologies (see above), the idea of structural classification itself is challenged by the recently discussed notion of a continuous protein structure space, which would naturally complicate the classification of proteins into discrete fold categories (see for example [240, 241, 242, 253]).

Rooted in the idea that short polypeptides (corresponding to structural motifs) form basic evolutionary units [254, 255], compact domains are suggested to be constructed from several such substructures leading to local structural similarity of one protein to several other proteins that are not globally related to each other. Accordingly, the protein structure space can be visualized not as composition of clearly distinct fold entities but rather appears as network with different folds being connected by commonly shared structural fragments [256]. Different views exist regarding the degree of continuity of this network. While some studies suggest a mostly discrete fold space where only a small subset of all folds (termed "gregarious" folds) are linked to several other folds thereby serving as network hubs [257, 258], others proposed a highly continuous fold

space using less restrictive similarity requirements [259]

However, while discussions regarding the nature of the protein fold space are still ongoing, classification approaches such as SCOP and CATH are well established in the scientific community and will continue to serve as valuable tools for structural biologists and bioinformaticians. Nevertheless, adjustments in the classification procedures might be necessary to cope with the observed overlap caused by common supersecondary structure motifs [252, 258].

6.2 Materials and methods

6.2.1 Membrane proteins in SCOP and CATH

For the analysis of membrane proteins in SCOP and CATH, all proteins were identified that were classified in SCOP v1.73 [223] and/or CATH v3.2 [237] and contained at least two transmembrane segments according to the annotation in PDBTM [201]. After filtering redundancy at 95% sequence identity, the SCOP dataset contained 88 protein chains and the CATH dataset contained 71 protein chains (corresponding to 92 and 80 unique classified domains, respectively). These domains were spread over 27 SCOP folds and 17 CATH folds summarized in Tables 9.5 and 9.6 of the Appendix.

Comparison of domain and fold assignments

For the comparative analysis of domain assignments and membrane protein fold classifications a common dataset was constructed (further referred to as `MP_SCOP_CATH`) containing proteins with assignments in both classification databases. To this end, all protein chains classified in SCOP and CATH were extracted yielding 58 chains (corresponding to 60 SCOP and 63 CATH domains). Redundancy at the domain level was removed from this set using the SCOP unique identifier (sunid) describing distinct domains. The final non-redundant `MP_SCOP_CATH` dataset contained 42 protein chains corresponding to 43 SCOP and 46 CATH domains all sharing a sequence identity below 95% (Appendix Table 9.7).

The occurrence of multi-domain assignments within alpha-helical membrane proteins classified by SCOP and CATH was calculating for each database separately using all membrane protein chains derived for this database. Furthermore, domain assignments were directly compared between SCOP and CATH for all proteins in the dataset `MP_SCOP_CATH`. For those proteins with an equal number of domains in SCOP and CATH,

the extent of domain position overlap was additionally analyzed. To this end, the fraction of residues consistently assigned by both databases was calculated and two domains were said to agree regarding their domain boundaries if this fraction exceeded 90% of the length of both individual domains.

Similarities and differences in the fold assignment of membrane proteins within SCOP and CATH were again analyzed using the `MP_SCOP_CATH` dataset. To this end, SCOP and CATH folds containing exactly the same protein domains were identified and classified as 'fold agreements'. Folds sharing at least one overlapping protein domain without constituting a fold agreement were added to the list of 'fold disagreements'. Such fold disagreements were further subcategorized depending on whether they directly arose from differences in domain assignments or whether they affected identical domains that were classified differently leading to 1:N or N:M fold relationships also called 'fold overlaps'. In order to compare the structural similarity of proteins involved in fold overlaps to those of fold agreements, all-against-all protein structure comparisons were executed using DaliLite v.3.1 [260]. SCOP domain coordinates were used for structure comparisons due to the higher degree of manual inspection. Only in one case (PDB 2atkC), CATH domain coordinates were used as the SCOP domain did not cover the whole transmembrane region.

The comparative analysis of membrane proteins in SCOP and CATH was initiated and conducted in an initial version by myself. This initial analysis was extended and turned into its final version presented here including all structural comparisons by Sindy Neumann.

Analysis of four helix bundle proteins

In order to compare membrane proteins to alpha-helical soluble structures in terms of diversity and classification, a detailed analysis was conducted based on the class of four helix bundle proteins. Initial datasets of both soluble and membrane four helix bundles were constructed by manually selecting all folds and corresponding protein domains from the all-alpha and membrane protein classes in SCOP with a fold description containing the terms '4 helices' and 'bundle'. In total, 38 soluble and three transmembrane four helix bundle folds were identified containing 2601 and 78 protein domain entries, respectively. After removing all domains with redundant SCOP domain sunid and insufficient positional agreement (<90%) with a corresponding CATH domain, the final dataset of four helix bundle domains was obtained consisting of 188 soluble and 11 membrane four-helix bundle domains. It should be noted that the protein 1c17 Chain

M (ATP synthase subunit A) was included within the dataset of membrane four-helix bundle proteins although its structure was not solved experimentally since both SCOP and CATH provide a classification for this protein despite their general exclusion of model structures. Again, all-against-all protein structure comparisons were executed for all protein domains of the final datasets using DaliLite v3.1 [260]. Obtained similarity scores (Z-scores) as well as the fraction of aligned residues with respect to the smaller of the two compared structures (coverage) were used to compare proteins classified into the same fold in both SCOP and CATH to proteins classified either together within only one database or separately in both databases.

The analysis of soluble and membrane four helix bundle proteins was executed by myself.

6.2.2 Classification of helix architectures

Dataset of membrane protein structures

To derive a structural classification protocol based specifically on transmembrane helix interactions, all protein chains having a solved 3D structure and at least four annotated transmembrane helices were obtained from PDBTM [201]. Proteins with less transmembrane helices were not considered as their helix architectures (defined by the observed helix interactions) lack the combinatorial diversity required for such a classification system. After removing sequence redundancy at 95% sequence identity, 182 protein chains remained of which 33 chains were classified also in both SCOP and CATH and 31 more chains had either a SCOP or CATH annotation. Helix interaction graphs were obtained for all proteins in this dataset using the corresponding PDB structure and the transmembrane helix annotations obtained from PDBTM.

Within the dataset, the number of observed transmembrane helices differed between four and 13 with seven transmembrane helices being the most prevalent number (59 protein chains). For all further analyses, all protein chains were treated as single domain proteins which is in line with results reported by Liu and colleagues stating that multi-domain membrane proteins are generally scarce [261] and with own results from the analysis of membrane proteins in SCOP and CATH where no protein could be found that was consistently annotated with two domains in both databases (see section 6.3.1 on page 113).

Similarity score HISS

A necessary requirement for the new classification system was the development of a scoring system that appropriately captures the similarity of two helix architectures as represented by their helix interaction graphs. Thereby, helix interactions should be weighted by the number of residue contacts observed between two helices and non-neighbouring helices should have a higher impact than neighbouring helices as especially long-distant interactions define the specific helix architecture of a protein. The new similarity score HISS (**H**elix **I**nteraction **S**imilarity **S**core) satisfies both requirements.

Given the two helix interaction graphs A and B, both containing the same number of transmembrane helices, the one-sided HISS score ($HISS_{A \rightarrow B}$) formulates how well helix architecture A is recovered from helix architecture B by calculating the fraction of edges from structure A that are also present in structure B:

$$HISS_{A \rightarrow B} = \frac{\sum_{edges(A \wedge B)} w_{dist} \cdot w_{con}}{\sum_{edges(A)} w_{dist} \cdot w_{con}} \tag{6.1}$$

Thereby, all edges can be weighted differently according to the number of helix-helix contacts they are based on (w_{con}) and/or dependent on whether they connect sequentially adjacent helices or not (w_{dist}). Here, the following weighting schemes were chosen. First, sequentially neighbouring helix interactions were down weighted by a factor $w_{dist} = 1 - a$ (a<1) and distant helix interactions were simultaneously up weighted by a factor $w_{dist} = 1 + b$ with $b = \frac{a \cdot N(\text{neighbouring edges})}{N(\text{distant edges})}$. Using the latter formula, the total weight of all distant helix interactions is balanced with the total weight of all neighbouring helix interactions. Accordingly, distant helix interactions are considered to be proportionally even more important in case only few such interactions are present in relation to the number of neighbouring interactions than in case of many distant helix interactions. The number of individual helix-helix contacts of each helix interaction was encoded by categorizing all edges into "weak", "intermediate" and "strong" interactions according to their number of helix-helix contacts. "Weak" interactions (<5 helix-helix contacts) were down weighted by $w_{con} = w$ (w<1), "intermediate" interactions (between 5 and 15 helix-helix interactions) were weighted with $w_{con} = 1$ and "strong" interactions (>15 helix-helix interactions) were up weighted with $w_{con} = s$ (s>1). These categories were used instead of deriving w_{con} directly from the number of helix-helix contacts to be less dependent on small variations in the number of observed helix-helix contacts which often may differ simply dependent on the used contact criterion. All parameters (a, b, w and s) were fixed during a subsequent parameter optimization experiment (see below).

As a perfect $HISS_{A\to B} = 1$ does not mean that two helix architecture are identical, but rather implies that architecture A is a subset of architecture B, the final similarity of two structures was calculated as the average HISS score of the two one-sided HISS scores:

$$HISS(A, B) = avg(HISS_{A\to B}, HISS_{B\to A}) \tag{6.2}$$

Parameter optimization

Optimal parameters for the comparison of two helix architectures using HISS scores were determined by repeatedly calculating pairwise similarities for a subset of all PDB chains selected earlier and comparing how well these similarity scores recover other structural classification approaches. SCOP and CATH were chosen as gold standards and hence all PDB chains from the full list of 182 proteins obtained earlier were selected having a classification in at least one of these databases resulting in 64 protein chains. All possible pairs of proteins with the same number of transmembrane helices were formed and classified into two groups containing either protein pairs classified to the same fold in SCOP and/or CATH or to different folds. Protein pairs classified differently in SCOP and CATH were not considered within this analysis. In total, the first group (same fold) contained 102 protein pairs while the second group (proteins from different folds) included 143 protein pairs. Considering only proteins with more than four helices, the number of protein pairs in both groups decreased to 95 and 116 pairs, respectively. In the following analysis, the first group of pairs was used as set of true positives while the second group formed the set of true negative instances.

HISS scores were calculated for all protein pairs with varying score parameters a, b, w and s (see above). For each parameter setting, it was evaluated how well these HISS scores were suited to classify all protein pairs into protein pairs belonging to the same or to different folds. A receiver operator characteristic (ROC) curve was calculated, which plots the achieved true positive rate against the false positive rate, with any point above the diagonal corresponding to better predictions than random. The quality of different classifications was compared using the Area Under the Curve (AUC) measure, with AUC values above 0.5 indicating classifications reproducing the gold standard classification better than random. Finally, the classification with the best AUC value was selected and the corresponding parameter setting was selected for the following classification of all PDB chains.

Clustering of helix architectures

To obtain the final classification of helix architectures, HISS scores were obtained for all protein pairs from the full dataset of 182 protein chains varying maximal by one transmembrane helix. The difference of one single transmembrane helix was permitted as the addition of one helix C- or N-terminal is evolutionary likely but might not alter significantly the observed helix interactions of already present helices. Practically, proteins with different helix numbers were compared by removing separately either the first or the last helix of the larger protein, comparing each substructure to the smaller protein and taking the maximum of the two calculated HISS scores.

All calculated HISS scores above 0.85 (0.9 for proteins with different transmembrane helix numbers) were subjected to MCL clustering [262] with an implementation of the algorithm obtained from `http://www.micans.org/mcl/`. Scores below 0.85 were neglected as this threshold resulted in maximal classification sensitivity and specificity during the previous parameter optimization experiments with respect to the corresponding SCOP/CATH classifications. Ultimately, the final MCL clusters were considered to constitute unique helix architectures forming the basic unit of the new classification approach.

Classification of predicted helix architectures

To test whether structural similarities of membrane proteins can also be derived from predicted helix architectures, all analyses executed with helix architectures obtained from PDB structures were also done after predicting helix interactions using helix-helix contacts obtained with TMHcon. As membrane proteins with four transmembrane helices were found to cause problems for structural classification approaches in preceding analyses (see Results and Discussion, section 6.3.1), these proteins were excluded. Therefore, helix interactions were predicted for 152 remaining proteins of which 54 were annotated in SCOP and/or CATH.

To determine the best method for distinguishing proteins with similar helix architecture from those with different helix architecture even when only a fraction of all observed helix interactions are predicted while other interactions are wrongly predicted, HISS scores were calculated for all protein pairs having the same number of transmembrane helices and a consistent annotation ('same fold' vs 'different fold') in SCOP and CATH. Thereby, different methods for predicting helix interactions (length-based prediction and contact-based predictions) were combined with different variations of HISS

scores (with or without edge weights) to select the optimal strategy to discriminate proteins classified to the same fold in SCOP and CATH from those classified to different folds. Finally, all proteins were clustered with the MCL algorithm [262] using similarity scores obtained with this optimal strategy for all proteins having the same number of transmembrane helices.

6.3 Results and discussion

6.3.1 Classification of membrane proteins in SCOP and CATH

So far, comparative analyses of structure classification databases have generally been carried out on the full set of available PDB proteins [250, 251]. Membrane proteins, which account for only 2% of all PDB entries, were therefore never in the focus of any previous work. Here, results of the first comparative analysis of occurrence and classification of alpha-helical membrane proteins within the two most commonly used structure classification databases, SCOP [243, 223] and CATH [244, 237], are presented. Special attention is focused on the question how these two databases cope with the fact that alpha-helical membrane proteins share the overall structure of a largely parallel alpha-helix bundle, while at the same time comprising a significant variety due to specific structural features such as helix interaction patterns or helix tilts. Furthermore, observed classification similarities and discrepancies as well as quantitative structure comparisons will be used to evaluate how continuous the currently known structure space of membrane proteins is in order to access the feasibility of structural classification approaches for membrane proteins now and in the future.

Membrane protein folds in SCOP and CATH

Membrane proteins with at least two transmembrane helices assigned by PDBTM are currently found within 27 SCOP (Appendix Table 9.5) and 17 CATH folds (Appendix Table 9.6). In SCOP, membrane proteins are classified within the class 'Membrane and cell surface proteins and peptides' while CATH does not provide a separate class for membrane proteins. Instead, alpha-helical membrane proteins are included within the mainly-alpha class together with alpha-helical soluble proteins. Within this class, two of the 17 folds containing membrane proteins belong to the orthogonal bundle architecture (CATH code 1.10) and 15 folds to the up-down bundle architecture (CATH code 1.20).

Generally, membrane proteins of the same fold are rarely further subdivided into

superfamilies and families in both databases. Only three out of 17 CATH membrane protein folds (17.6%) are associated with more than one superfamily. In case of SCOP, only two membrane protein folds are further subdivided into more than one superfamily and only four folds contain more than one family, which corresponds to 7% and 15% of all SCOP membranous folds, respectively. For comparison, 13% and 38% of all globular folds (belonging to SCOP classes 'a', 'b', 'c' or 'd') are associated with more than one superfamily or family, respectively. The number of distinct membrane protein domains assigned to one fold varies only slightly ranging from 1 to 9 (SCOP) and 1 to 14 (CATH) domains (Appendix Tables 9.5 and 9.6). Not surprisingly, these numbers reflect the substantially higher structural coverage of soluble proteins compared to membrane proteins. While the number of newly identified folds for soluble proteins is steadily decreasing [237], structure determination of membrane proteins is far from saturation, limiting the number of folds with several unrelated representatives to a small number of well studied folds, such as the two-helix hairpin and the four-helix bundle fold.

Unexpectedly, the number of transmembrane helices can vary significantly within the same fold according to the annotation taken from PDBTM (Appendix Tables 9.5 and 9.6). For example, protein domains assigned to the 'heme-binding four-helical bundle' fold in SCOP (f.21) were found to contain between three and five transmembrane helices. Within the CATH database, the biggest variance was found for the 'cytochrome bc1 complex; chain c' fold (1.20.810), whose domains contain either four or eight transmembrane helices corresponding to the cytochrome b6 of the b6f complex and the cytochrome b of the bc1 complex, respectively. Local similarity between the N-terminal heme-binding part of cytochrome b and cytochrome b6 [263] seems to cause the common classification of these proteins.

Similarities and discrepancies between SCOP and CATH domain assignments

Since domains are the basic units of protein structure classification in SCOP and CATH, the agreement of their assignments was analyzed first. As SCOP and CATH use different methods to decompose proteins into domains (visual inspection compared to largely automatic domain assignments), previous analyses reported significant differences between these two databases [250, 251]. However, since most membrane proteins are single-domain proteins [261], one would expect disagreements between domain assignments for membrane proteins to be less frequent than those observed for soluble proteins.

In total, four protein chains (4.5% of all chains) were classified as multi-domain within SCOP, while amongst the 71 protein chains from CATH nine (12.7%) contained two

domains. The observation that CATH identifies more multi-domain proteins than SCOP was already reported in the work of Hadley and Jones [250] and was found to be a direct result of the different domain definitions that are used in the two databases with CATH addressing rather geometrical aspects while SCOP incorporates also functional considerations.

Addressing differences in the domain assignments between SCOP and CATH more specifically for individual proteins, the separation into domains was analyzed for all 42 alpha-helical membrane proteins classified by both databases (dataset `MP_SCOP_CATH`, Appendix Table 9.7). In 37 cases, the two databases consistently assigned one domain per protein chain. However, several cases were observed where this single domain was not covering the entire protein chain resulting in three cases where SCOP and CATH deviated by more than 10% of their assigned domain positions, while in the remaining 34 cases (81% of all proteins in `MP_SCOP_CATH`) domain position assignments were consistent between SCOP and CATH. In total five protein chains were divided into two domains either by SCOP or by CATH, with the majority (four chains) being assigned with a single domain in SCOP but two domains in CATH.

Interestingly, the obtained results vary not much from similar results reported for soluble proteins. There, 82% of all chains were found to agree in the number of assigned domains [250] compared to 88% reported here for membrane protein chains. However, a larger number of membrane protein structures will be required to confirm this trend in the future.

Similarities and discrepancies between SCOP and CATH fold classifications

The agreement between SCOP and CATH with respect to their fold classification was again compared using the dataset `MP_SCOP_CATH`. All domains of this dataset were consistently assigned to 15 folds in both SCOP and CATH, although the composition of individual folds was found to vary in several cases. Eight folds were found to contain exactly the same domains in SCOP and CATH (Table 6.1). In total, 20 chains (47.6% of `MP_SCOP_CATH`) containing each exactly one domain were assigned to these folds. With only one exception where proteins with 12 and 13 transmembrane helices were found within the same fold (SCOP fold f.24 / CATH fold 1.20.210), the number of transmembrane helices was completely conserved within each of these folds.

Disagreements between SCOP and CATH fold assignments can be caused either by discrepancies in domain assignments or by intrinsic differences in the classification process. This latter type of disagreement was termed the fold overlap problem by Hadley

Table 6.1: Comparison of membrane protein fold assignments in SCOP and CATH. Agreements and differences in the fold classification were obtained for all proteins in `MP_SCOP_CATH`. Discrepancies were further classified into two categories dependent on what caused the difference (differing domain assignments or split/merging of folds). Roughly 50% of all membrane proteins are classified differently with helix hairpins and four helix bundles accounting for nearly all observed fold overlaps.

Relationship	SCOP fold	CATH fold
Fold agreements	f.13	1.20.1070
	f.19	1.20.1080
	f.20	1.10.3080
	f.24	1.20.210
	f.29	1.20.1130
	f.30	1.20.860
	f.31	1.20.1240
	f.33	1.20.1110
Fold disagreement caused by domain disagreement	f.26	1.20.85 + 1.20.85
	f.21 + f.32	1.20.810
Fold disagreement caused by fold overlap[a]	**f.14**, f.25, **f.36**	**1.20.120**
	f.14, f.17	**1.10.287**
	f.21	**1.20.810, 1.20.950**, 1.20.1300
	f.14	**1.10.287, 1.20.120**
	f.17	**1.10.287, 1.20.20**

[a] Folds marked in bold correspond to folds containing two or four-helix bundle proteins.

and Jones [250] and for SCOP and CATH it arises from differences between the manual fold assignment within SCOP and the largely automatic approach based on structure comparisons within CATH. While the first type of discrepancy occurred three times (Table 6.1) and involved five proteins as discussed above, additional five cases of fold overlaps were observed within the `MP_SCOP_CATH` dataset. Remarkably, all five fold overlaps involve domains with two or four transmembrane helices.

All-against-all protein structure comparisons executed with DaliLite v.3.1 [260] were used to compare the structural similarity of proteins involved in fold overlaps to those from fold agreements (Table 6.2). For fold agreements, average Z-scores varied between 23.9 (fold f.13/1.20.1070) and 44.3 (fold f.24/1.20.210). Proteins from fold overlaps on the other hand were found to be structurally much more divers with average Z-scores ranging between 3.4 and 11.3. Importantly, this observation is not necessarily caused by fold overlaps generally consisting of less transmembrane segments than fold agreements as for example two domains covering both ten transmembrane helices but classified into two different folds from the list of fold agreements (f.20/1.10.3080 and f.33/1.20.1110)

resulted in a Z-score of 1.0.

Table 6.2: All-against-all structure comparisons between membrane proteins classified in SCOP and CATH. Structure comparisons were executed using DaliLite [260] for all folds containing at least two domains. Average similarity scores obtained for folds identical in SCOP and CATH (fold agreements) are clearly higher than for fold discrepancies.

Folds	N(prot)[a]	N(comp)[b]	Max(Z-score)[c]	Min(Z-score)[c]	Avg(Z-score)[c]
Fold agreements					
f.13 / 1.20.1070	6	15	35.6	8.5	23.9
f.19 / 1.20.1080	4	6	30.7	17.8	24.1
f.24 / 1.20.210	4	6	57.5	34.1	44.3
f.29 / 1.20.1130	2	1	34.1	34.1	34.1
Fold disagreements					
1.20.120 / f.14,f.25,f.36	6	15	20.9	3.8	11.3
f.14 / 1.10.287,1.20.120	2	1	3.4	3.4	3.4
f.21 / 1.20.810,1.20.950,1.20.1300	6	14[d]	9.5	3.2	5.9
f.17 / 1.10.287,1.20.20	4	6	9.6	2.2	4.8
1.10.287 / f.14,f.17	4	5[d]	9.6	2.0	5.7

[a] N(prot): number of proteins within analyzed fold(s).
[b] N(comp): number of pairwise comparisons executed with DaliLite.
[c] Z-score: similarity score as obtained from DaliLite.
[d] For one comparison, DaliLite did not yield a result.

Associated with the reduced structural similarity, two main reasons can be identified causing the presence of observed fold overlaps. First, the single-linkage clustering approach of CATH results in differences to the SCOP classification system which instead applies an average linkage procedure [252]. As long as folds are structurally clearly distinct from each other, the impact of these clustering differences is likely to be minimal. However, the more continuous the fold space the more prominent is the effect of different clustering methods on classification results, as seems to be the case for membrane four helix bundle proteins (see also below). Additionally, functional reasons may prompt SCOP to classify proteins within the same fold despite low structural similarity that are separated into several folds within CATH. For example, fold f.21 of SCOP contains four helix bundle proteins that all bind heme(s). CATH disregards this functional aspect and identifies enough structural differences to assign these proteins to different folds.

In summary, the comparative analysis of membrane proteins in SCOP and CATH shows that available membrane protein structures with six and more transmembrane helices are either very similar to each other (and thus are classified consistently to the same fold) or sufficiently diverse that SCOP and CATH both assign them to different

folds. Accordingly, the current structure space adopted by these proteins seems to be mostly discrete and a structural classification similar to soluble proteins is possible. Apparently more difficult is the classification of membrane proteins with two to five transmembrane helices as can be seen from the fact that all cases of fold overlap in the field of alpha-helical membrane proteins involve proteins with five or less transmembrane segments. Obviously, these proteins are on the one hand diverse enough that both CATH and SCOP separate them into several individual folds, but on the other hand display differences too subtle to be captured using the classic definition of a fold, leading to largely deviating classifications within SCOP and CATH.

Structural diversity of four helix bundle proteins

As four helix bundle proteins were found to pose problems for structural classification of membrane proteins, it was further analyzed whether this was due to intrinsic properties of this particular architecture, or whether the structural restrictions imposed by the lipid bilayer additionally impede the classification. According to early studies soluble four helix bundles comprise significant variety in their pattern of interhelical angles despite the low number of helices [264, 265]. In order to test whether this diversity is specific for soluble proteins and facilitates their classification, a non-redundant dataset of 188 soluble four-helix bundle domains from 28 SCOP folds was analyzed and compared to an analogous dataset consisting of 11 distinct membrane four helix bundle domains from 3 SCOP folds.

First, the consistency between SCOP and CATH with respect to their fold classification of four helix bundles was analyzed. To this end, the frequency of all domain pairs classified into the same fold in one database that are also assigned to the same fold in the other database was calculated. For soluble proteins, these percentages were remarkably high with 88% of all co-classified SCOP domains appearing also in the same fold in CATH and as many as 94% of all CATH co-classified domain pairs being in the same SCOP fold. As discussed above, the consistency between SCOP and CATH is much lower for membrane four helix bundles than for other membrane folds. For these proteins, from all domain pairs within the same SCOP four helix bundle fold, only 48% were also found in the same CATH fold while 71% of all CATH co-classified domain pairs were also in the same SCOP fold. No case of complete fold agreement between SCOP and CATH could be found for membrane four helix bundles while 11 such fold agreements were detected for soluble proteins. It thus appears that membrane four helix bundle proteins in fact are more difficult to classify than their soluble counterparts,

although in principle such low degree of agreement between SCOP and CATH for this particular architecture could be a statistical artefact caused by the paucity of available structural data.

In order to further explain the observed difficulties in the classification of membrane four helix bundle proteins, all-against-all structure comparisons were executed using DaliLite and the observed structural variety of membrane and soluble four helix bundle domains was compared (Table 6.3). For soluble proteins, comparisons between proteins consistently classified to different folds in SCOP and CATH (category 3 comparisons) either retrieved no detectable similarity (45.4% of all comparisons) or the obtained fractions of aligned residues (coverage) and Z-scores were clearly smaller on average than for proteins classified either in one or in both databases to the same fold. Furthermore, analyzing specifically those folds containing exactly the same domains in SCOP and CATH it was observed that these folds in fact represent distinct regions of the structure space of four helix bundle domains as structure comparisons of proteins within each fold returned in all cases at least twice as high average Z-scores than comparisons of fold members with proteins not belonging to the respective fold (data not shown).

Table 6.3: All-against-all structure comparisons of membrane and soluble four helix bundle domains. Structure comparisons were executed using DaliLite [260] for 188 soluble and 11 membrane four helix bundle domains. The structural diversity of membrane proteins is generally less distinct than observed for soluble proteins.

		Category 1[a]	Category 2[b]	Category 3[c]
soluble	N(comparisons)	1321	242	16015
	NA[d]	3.5%	4.1%	45.4%
	Avg(coverage)[e]	74.3%	70.9%	57.1%
	Avg(Z-score)	6.7	5.3	2.1
membrane	N(comparisons)	10	15	30
	NA[d]	-	6.7%	-
	Avg(coverage)[e]	87.5%	74.3%	69.9%
	Avg(Z-score)	14.4	6.0	4.5

[a] Category 1: protein pairs classified to the same fold in SCOP and CATH.
[b] Category 2: protein pairs classified to the same fold either in SCOP or CATH but to a different fold in the respective other database.
[c] Category 3: protein pairs classified to separate folds in SCOP and CATH.
[d] NA: percentage of comparisons where DaliLite did not return any result.
[e] Avg(coverage): average fraction of aligned residues with respect to the smaller of the two compared structures.

The structural space for the respective membrane proteins on the other hand seems

to be much more continuous since structural differences among all proteins, no matter whether they are classified within SCOP and/or CATH to the same or to a different fold, are less pronounced. Average coverage and average Z-scores in all three analyzed categories of comparisons (Table 6.3) were found to be higher than for soluble proteins. Even proteins classified to different folds in both databases still had an average Z-score of 4.5 and an average coverage of 69.9% which is comparable to those soluble proteins classified to the same fold in one database but to separate folds in the other database. Additionally, DaliLite was able to detect at least a minimal similarity (Z-score > 2.0) among all proteins that were assigned to different folds, which was clearly not the case for their soluble counterparts where 45% of all comparisons did not retrieve any result and additional 30% resulted in a Z-score of less than 2.0. As structural variations are more fine-grained for membrane four helix bundle proteins, their classification naturally represents a harder problem as long as the same rules are applied as for soluble proteins. This problem is likely to persist also with more solved structures becoming available unless a more specific fold definition for membrane proteins is at hand.

6.3.2 Classification of helix architectures obtained from PDB structures

Structural classification approaches such as SCOP and CATH base their classification on structural similarity between full domains. For membrane proteins, this includes transmembrane regions just as extramembranous parts although the contribution of individual regions to the final classification may vary. Here, a new structural classification system for membrane proteins is proposed that specifically addresses the similarity of transmembrane helix bundles as expressed in their helix interactions. To this end, a new structural similarity score (termed HISS) was developed that quantifies to which extent two helix interaction graphs resemble each other and this score was used to cluster the full set of available membrane protein structures.

Consistency with SCOP and CATH

HISS, the similarity score derived for comparing helix interaction graphs, specifies the average fraction of all edges of a given graph found also in the second graph and vice versa. Thereby, interactions between neighbouring helices and those based on a small number of residue contacts can be down weighted while distant and strong helix interaction can be up weighted. Within a first analysis, optimal weights were determined

using a subset of membrane protein structures that were consistently classified in both SCOP and CATH or were present in only one of these databases. For each set of tested parameters, a ROC curve was calculated describing how well calculated HISS scores can be used to distinguish protein pairs classified to the same fold in SCOP and/or CATH from those proteins separated into different folds. The AUC ("area under the ROC curve") statistic was used to compare the influence of different parameters and estimate how well SCOP and CATH can be reproduced in general by a similarity score considering only helix interactions (Table 6.4).

Table 6.4: Classification of proteins from SCOP/CATH using the helix interaction similarity score HISS. Only proteins classified consistently in SCOP and CATH either to the same fold or different folds were used (64 protein chains in total). Considering only proteins with more than four transmembrane helices, SCOP and CATH can be reproduced nearly perfectly with an AUC of close to 1.

Parameters			AUC_{total} [a]	AUC_{4+} [b]
a [c]	w [d]	s [e]		
0.0	1.0	1.0	0.955	0.992
0.1	1.0	1.0	0.956	0.991
0.5	1.0	1.0	0.955	0.978
0.0	0.9	1.1	0.957	0.993
0.0	0.7	1.3	0.964	0.997
0.0	0.5	1.5	0.966	0.998
0.1	0.9	1.1	0.961	0.993
0.1	0.7	1.3	0.967	0.996
0.1	0.5	1.5	0.969	0.997

[a] AUC_{total}: area under the curve statistic using all proteins classified consistently in SCOP and CATH.
[b] AUC_{4+}: area under the curve statistic using only proteins with more than four transmembrane helices classified consistently in SCOP and CATH.
[c] Parameter a: factor used to reduce the weight from neighbouring helix interactions and increase the weight from distant interactions.
[d] Parameter w: edge weight used for helix interactions with less than five helix-helix contacts.
[e] Parameter s: edge weight used for helix interactions with more than fifteen helix-helix contacts.

As can be seen from Table 6.4, a HISS score based classification agrees nearly perfectly with the consensus classification of SCOP and CATH used as reference. Using all protein chains as a test set, AUC values were consistently above 0.95, restricting the dataset to proteins with more than four helices, AUC values even increased to values higher than 0.99 (for comparison, a perfect classification would result in an AUC value of 1.0). Down weighting neighbouring helix interactions while concurrently up weighting distant helices has no positive effect by itself as can be seen when changing parameter 'a'

alone but can be helpful in combination with the remaining two parameters 'w' and 's'. Weighting edges according to the number of observed helix-helix contacts on the other hand results in a slightly improved classification in all tested parameter settings. Overall the best classification for all proteins (AUC=0.969) was obtained when down weighting neighbouring and weak interactions by 0.1 and 0.5, respectively and up weighting strong interactions by factor 1.5. For proteins with more than four helices, the best classification was achieved with weights for weak and strong interactions of 0.5 and 1.5. Using a HISS score of 0.85 as requirement for classifying two proteins with more than four helices to the same "helix architecture fold", 98.9% of all proteins classified to the same fold in SCOP/CATH would also be added to the same helix architecture and 97.4% of all protein pairs not belonging to the same SCOP/CATH fold would also be separated to different helix architectures. Only one protein pair would be wrongly classified to separate folds while only three protein pairs would be incorrectly classified to the same fold.

Importantly, these results demonstrate that helix interactions alone are basically sufficient to reproduce the structural classification of membrane proteins as proposed by SCOP and CATH. Accordingly, helix interactions are not only one of the characteristic determinants of a membrane structure but seem to clearly outweigh other structural features such as properties of extramembranous elements, helix tilts or helix kinks, at least for proteins with more than four transmembrane helices where the combinatorial freedom of possible helix interaction patterns is sufficient to distinctly separate different helix architectures from each other. Consistent with previous results presented for the comparison of SCOP and CATH, membrane proteins with four transmembrane segments again stand out due to the limited number of possible helix interaction patterns leading to an increased number of false positives (proteins with a highly similar helix interaction graph but different SCOP and/or CATH folds). As suggested already earlier, the classification of membrane four helix bundles seems to require special rules addressing a most likely more continuous fold space or might even not be possible at all. Therefore, these proteins were excluded from all further analyses.

Clustering of helix architectures

To obtain the final classification of helix architectures, all membrane proteins with solved structure and more than four transmembrane segments were subjected to MCL clustering [262] based on HISS scores. In contrast to the previous analysis, HISS scores were also obtained for proteins differing in their number of transmembrane helices to detect

those cases where one transmembrane helix was added at the protein's C- or N-terminus without altering significantly the helix interaction pattern of the remaining helices. In total, 152 proteins were clustered resulting in 20 helix architectures with at least two members and 14 protein chains not classified at all. Figure 6.1 displays the members of one helix architecture (HA13, see Table 6.5) demonstrating the amount of diversity observed within one cluster with respect to weak helix interactions (formed by few helix-helix contacts) and the conservation of strong helix interactions.

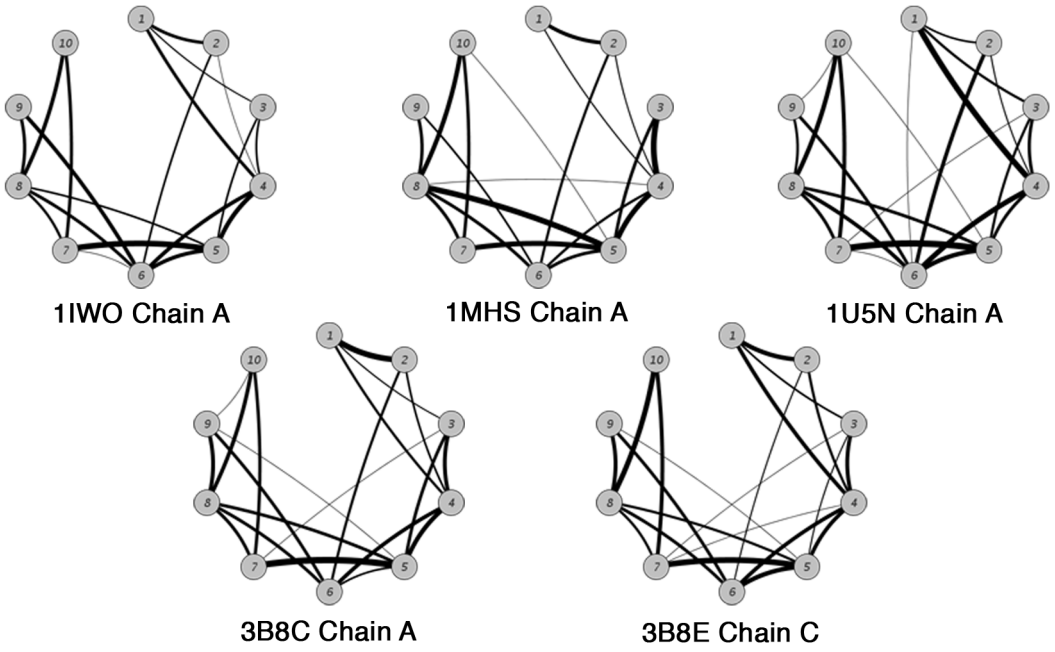

Figure 6.1: Example MCL cluster HA13 containing five membrane protein with highly similar helix architecture. All cluster proteins correspond to transporters although with different transporter specificity. A common pattern of strong helix interactions is clearly visible.

As can be seen from Table 6.5 and Figure 6.3, found helix architectures having at least two members cover proteins with five up to twelve transmembrane helices. The highest number of distinct architectures was obtained for proteins with six transmembrane segments. The cluster with the most member proteins on the other hand is helix architecture (HA) 9 with 54 proteins having seven transmembrane segments. This cluster includes structures of bacteriorhodopsin, archaean and eukaryotic rhodoposins as well as other 7TM receptors which all were found to share the same overall helix architecture despite a rather low average sequence identity of 28%. This observation is consistent with results from a recently conducted large-scale modelling experiment of all human

GPCRs which suggested that the structural core of these receptors is highly conserved even though variation in helix packing, helix kinks and loop conformations is possible [266]. Similar to the cluster of 7TM receptors, sequence identity among proteins of the same helix architecture is generally rather low with thirteen of twenty clusters having an average pairwise sequence identity below 40% indicating that proteins with surprisingly diverse sequences may fold into structures having the same pattern of helix interactions.

Table 6.5: Detected helix architectures using HISS scores and MCL clustering. In total, 152 protein chains were clustered into 20 helix architectures covering at least two members, 14 proteins remained unclassified.

Helix architecture	Members	TMS[a]	SCOP[b]	CATH[c]	Avg(ident) [%][d]
HA1	10	5	f.26	-	38.8
HA2	3	5	f.21/f.25	1.20.120	38.5
HA3	13	6	f.19	1.20.1080	37.0
HA4	6	6 (5)[e]	f.21	1.20.950	35.8
HA5	4	6	-	-	50.6
HA6	2	6	-	-	29.1
HA7	2	6	f.42	-	22.8
HA8	2	6	f.51	-	40.7
HA9	54	7 (6)[f]	f.13/f.37	1.20.1070	28.0
HA10	4	7	f.25	-	57.0
HA11	2	7	-	1.20.1450	60.9
HA12	5	8	-	1.20.810	54.9
HA13	5	10	f.33	1.20.1110	30.4
HA14	3	10	f.41	-	30.6
HA15	2	10	f.20	1.10.3080	80.9
HA16	2	10	f.22	-	32.3
HA17	5	11	f.29	1.20.1130	56.2
HA18	3	11	f.44	-	26.5
HA19	8	12	f.38	-	27.4
HA20	6	12 (13)[g]	f.24	1.20.950	39.6

[a] TMS: number of transmembrane segments characteristic for this helix architecture.
[b] SCOP: SCOP classification(s) found for members of this helix architecture.
[c] CATH: CATH classification(s) found for members of this helix architecture.
[d] Avg(ident): average pair wise sequence identity between all members of this cluster.
[e] Two proteins of this cluster had only five transmembrane segments in contrast to the majority of proteins with six transmembrane helices.
[f] One proteins of this cluster had only six transmembrane segments in contrast to the majority of proteins with seven transmembrane helices.
[g] One proteins of this cluster had thirteen transmembrane segments in contrast to the majority of proteins with twelve transmembrane helices.

Generally, all clusters are highly conserved with respect to the number of transmembrane segments, only three cases were observed where one or two proteins had one

transmembrane segment more or less than the majority of all proteins (HA4, HA9 and HA20, Figure 6.2). Further analysis of these cases showed that proteins of HA4 (Figure 6.2A) and HA20 (Figure 6.2B) in fact seem to share highly similar helix architectures due to evolutionary relationship since structures of these clusters can be superimposed with RMSD values ≤ 2.5Å even though they have different numbers of transmembrane segments. For HA20 the common evolutionary origin is also recognized by SCOP where all proteins are classified not only to the same fold but also to the same superfamily and family (f.24.1.1). The classification of protein 2HYD (chain A, six transmembrane helices) to helix architecture 9 on the other hand seems to be an artefact as no significant structural similarity could be detected between 2HYD and 7TM receptors (Dali Z-score < 3.0). Accordingly, the helix interaction graph similarity seems to be caused by the common presence of a strongly connected four helix bundle (helices 3-6) and a rather loosely connected helix hairpin (helices 1 and 2) which seems to be a repeated pattern in membrane protein structures (see below) not necessarily implying evolutionary relationship.

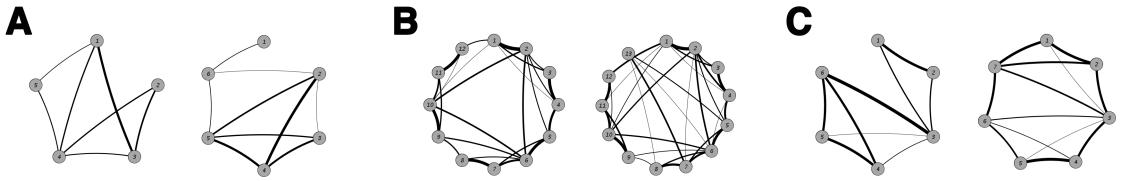

Figure 6.2: Membrane proteins with differing number of transmembrane helices classified to the same helix architecture. (A) Helix architecture 4 with proteins 3DHW (chain A, five transmembrane helices) and 3D31 (chain C, six transmembrane helices). (B) Helix architecture 20 with proteins 2DYR (chain A, twelve transmembrane helices) and 1EHK (chain A, thirteen transmembrane helices). (C) Helix architecture 9 with proteins 2HYD (chain A, six transmembrane helices) and 1E12 (chain A, seven transmembrane helices). The common classification may arise either due to evolutionary relationship (HA4 and HA20) or as result of a clustering artefact (HA9).

Analysing the coverage of found helix architectures with SCOP and/or CATH annotations, in total eight helix architectures contained proteins with annotations in both databases and ten more helix architectures covered members with an annotation in at least one structural database (Table 6.5). However, often only a subset of all members of these clusters were found in SCOP and/or CATH. In case several proteins of the same helix architecture in fact were annotated in either SCOP or CATH, these annotations were highly consistent as expected from the previous detailed comparison of helix ar-

chitecture based classification with SCOP and CATH (see page 119). Only two helix architectures (HA2 and HA9) united proteins from two different SCOP folds. In the first case (HA2), one of the cluster proteins was differently classified by SCOP to the 'heme-binding four-helical bundle' fold (f.21) due to the presence of a heme group although the protein covers five helices whose interaction pattern is highly similar to two other five helix bundle proteins. The second discrepancy arises from incorrectly classifying the six transmembrane helix protein 2HYD (chain A) to the same helix architecture as known 7TM receptors discussed already above.

Overall, the proposed classification of helix architectures constitutes a comprehensive classification approach of all known membrane protein structures while SCOP and CATH both include only a subset of these structures. Therefore, it combines on the one hand information present separately in SCOP and CATH but also generates new information by identifying completely new helix architectures not present in SCOP and CATH or adding proteins based on their helix interaction patterns to already known folds.

Diversity of found helix architectures

Inspecting representative helix interaction graphs for all helix architectures with at least two members (Figure 6.3), several observations can be made. First, even for proteins with a limited number of five to seven transmembrane helices, clearly distinct helix architectures can be differentiated illustrating impressively the structural variety open to alpha-helical membrane proteins. This impression is even enforced when considering also singleton proteins which were not classified with other proteins to one of the extracted helix architecture clusters and therefore can be expected to represent to a large degree distinct helix patterns by themselves (Figure 6.4, some singleton proteins also share at least visual similarity with other proteins of the dataset which might cause a common classification once more membrane protein structures are available, see for example proteins 3G5U and 2GIF and helix architecture HA 19).

Especially for proteins with ten or more transmembrane helices it is furthermore apparent that most helix architectures are densely packed with many helix interactions among distant helices. Only in few cases (for example HA11 or singleton protein 1LVI) the network of interactions is sparse with a majority of all interactions taking place between sequentially close helices. Again, this demonstrates the surprisingly large space of structural variation open to membrane proteins suggesting that many new helix architectures will be observed with additional structures becoming available.

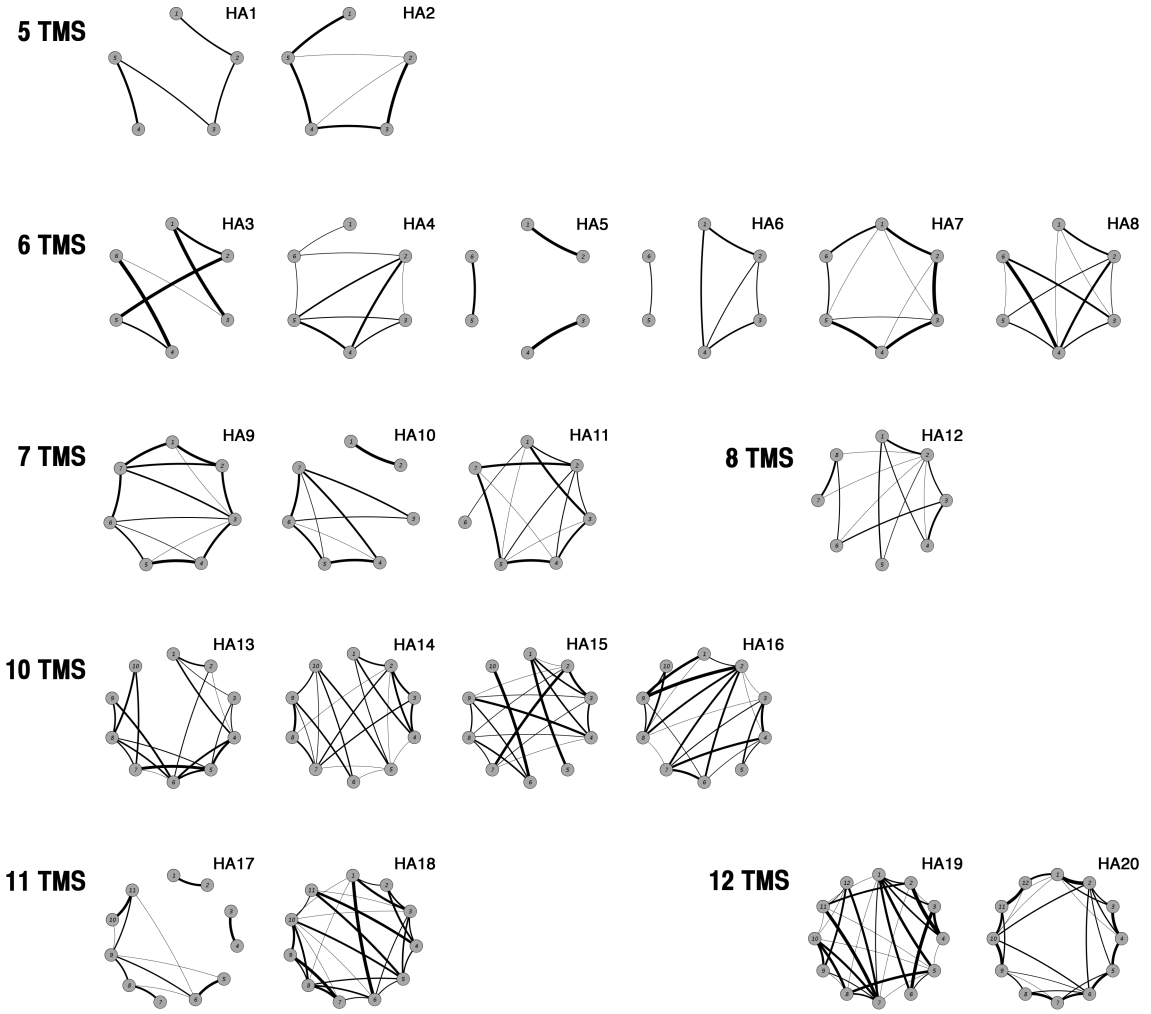

Figure 6.3: Representative helix interaction graphs for all helix architectures with at least two member proteins. Twenty architectures containing proteins with five to 12 transmembrane segments (TMS) were obtained each represented by a distinct helix interaction pattern.

Nevertheless, several structural patterns are also observed repeatedly in multiple helix graphs. Four helix bundles where all (or nearly all) helices interact with each other distinctly are noticeable especially in structures with five to seven transmembrane helices (for example, HA4, HA6, HA9 and HA10). Similarly, several structures contain clearly disconnected helix hairpin structures consisting of two mutually connected helices with no (or only weak) interactions to other helices (for example HA5, HA6, HA10 and HA17). Interestingly, this observation is consistent with a recent theory describing the evolutionary origin of membrane proteins which proposed that the evolution of

membrane proteins started from simple amphiphilic, alpha-helical hairpins [267] which repeatedly duplicated to form multi-helix bundle proteins. Indeed, helix interaction patterns of present membrane protein suggest that helix hairpins and subsequently bundles of four helices form important building blocks which are combined in various ways to form more complex membrane protein structures.

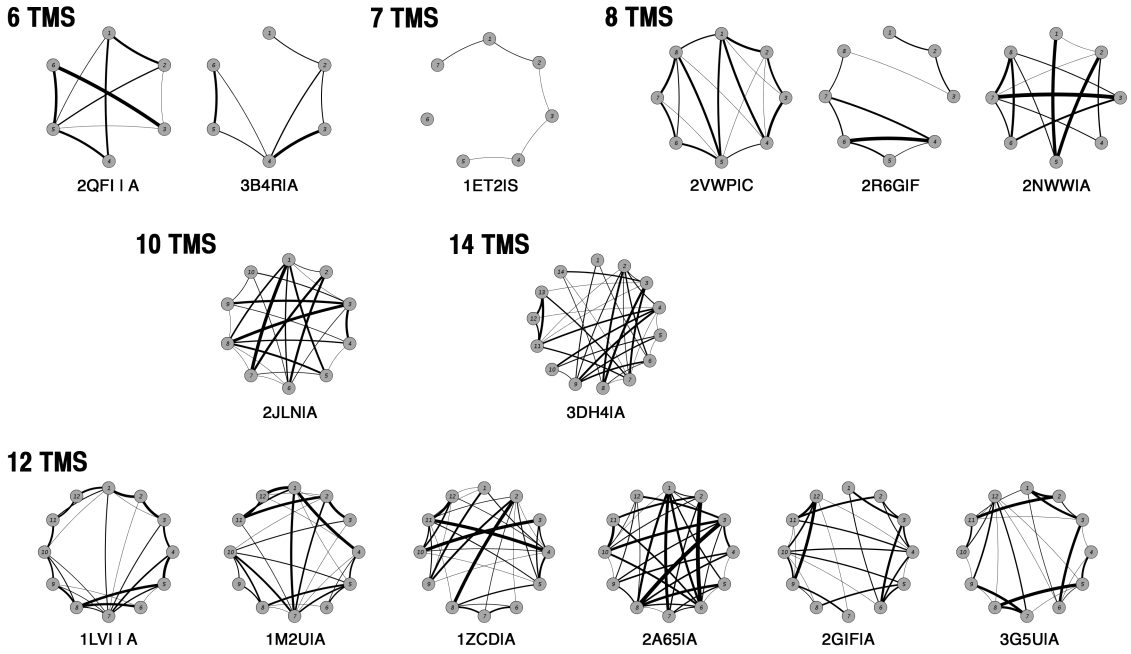

Figure 6.4: Helix interaction graphs for all singleton proteins not classified to one of the derived twenty helix architectures shown in Figure 6.3. In total 14 proteins remained unclustered with the majority containing 12 transmembrane segments (TMS).

Finally, the analysis of all found helix architectures confirms the frequent occurrence of internal symmetry in membrane protein structures noticed previously already from individual structures such as the ones of lactose permease [268], chloride ion channel [269] and the protein EmrD [270]. For proteins with six, eight, ten and twelve helices, always at least one helix architecture is observed which is split into two subgraphs which are mostly identical but are also highly connected to each other and are therefore forming rather one helix bundle than two distinct structural domains. Again, this agrees with a recently proposed mechanism of membrane protein evolution where so-called dual topology proteins are duplicated and subsequently fused to give rise to proteins with a doubled number of transmembrane helices [40].

6.3.3 Classification of predicted helix architectures

After demonstrating that membrane proteins can be structurally classified based on the similarity of their helix interaction patterns, a final analysis was conducted evaluating whether this is also possible for proteins where no 3D structure is already available. In this case, helix interactions need to be predicted from sequence as introduced in Chapter 5. Generally, these predictions are highly accurate but suffer from limited sensitivity (see section 5.3.2) which might make the identification of similar structures difficult if not impossible.

Therefore, the principal similarity of predicted helix interaction graphs and the possibility of discriminating proteins with similar and different structures was first evaluated using all protein pairs classified consistently in SCOP and CATH either to the same fold or to different folds (Table 6.6). As four helix bundle proteins were earlier shown to pose a problem to structural classification in general and helix interaction graph based classification specifically (sections 6.3.1 and 6.3.2), only proteins with at least five transmembrane helices were considered resulting in a test set of 211 protein pairs of which 95 had the same fold assignment in SCOP/CATH while the remaining protein pairs had the same number of transmembrane helices but different fold assignments. Helix interactions were predicted for all proteins based on helix-helix contacts obtained with TMHcon with different strategies (using either only the L/5 best predicted residue contacts or the two step contact-based filtering procedure where a large set of residue contacts is selected in the first step but only those helix pairs are predicted as interacting with a minimal number of these residue contacts). Additionally, HISS similarity scores were calculated once treating all predicted helix interactions equally and once where helix interactions with few residue contacts were down weighted while interactions with many predicted contacts were up weighted.

As can be seen from Table 6.6, proteins from the same fold in SCOP and CATH have constantly higher HISS scores than proteins from different folds independent of the prediction strategy used for obtaining interacting helices and from the HISS calculation method applied. Accordingly, the classification of proteins into "same" or "different" fold based on HISS scores results in a classification well above random as can be concluded from the reported AUC ("area under the ROC curve") values which were found to be as high as 0.86 (a random prediction would result in an AUC value of 0.5). Here, contact-based helix predictions (C9/C10 and C9/C15) were superior to length-based L/5 predictions which is not surprising as these predictions were also already observed to predict individual helix interactions with higher accuracy (section 5.3.2, Table 5.2).

Table 6.6: Classification of proteins in SCOP and CATH using predicted helix interactions. Helix interactions were predicted using different strategies - length-based (L5) and contact-based (C9/C10, C9/C15) - and HISS scores were calculated with and without weighting edges differently. Proteins with similar structures can be recognized with good accuracy (up to 65% sensitivity at >90% specificity).

	HISS[a]	Avg($HISS_{same}$)[b]	Avg($HISS_{diff}$)[c]	AUC[d]	Score[e]	Sensitivity [%]	Specificity [%]
L/5	unweighted	0.806	0.627	0.775	0.80	56.8	82.8
					0.85	40.0	91.4
	weighted	0.843	0.649	0.815	0.80	66.3	81.0
					0.88	41.1	91.4
C9/C10[f]	unweighted	0.864	0.723	0.849	0.83	71.6	81.9
					0.86	58.9	92.2
	weighted	0.821	0.633	0.865	0.76	74.7	80.1
					0.81	58.9	90.5
C9/C15[f]	unweighted	0.890	0.749	0.846	0.88	72.6	78.4
					0.90	65.3	92.2
	weighted	0.864	0.665	0.848	0.82	76.8	81.0
					0.88	53.7	91.4

[a] HISS scores were calculated once without weighting helix interactions differently (unweighted) and once down weighting helix interactions with <5 predicted residue contacts by a factor 0.5 and up weighting interactions with >15 residue contacts by a factor 1.5 (weighted).
[b] Avg($HISS_{same}$): average HISS score for proteins classified to the same fold in SCOP and CATH.
[c] Avg($HISS_{diff}$): average HISS score for proteins classified to different folds in SCOP and CATH.
[d] AUC: area under the curve describing how well proteins of the same fold can be differentiated from proteins from different folds (AUC=0.5 would correspond to a random prediction).
[e] Score: HISS score threshold used to identify proteins with the same helix architecture.
[f] Helix interactions were predicted from TMHcon residue contacts with every helix pair considered interacting having nine predicted contacts from NN4 or 10 (C9/C10) respectively 15 (C9/C15) contacts from NN4-D.

Generally, weighted HISS scores result in slightly better predictions than unweighted HISS scores confirming again that helix interactions with many residue contacts tend to cumulate also higher numbers of predicted contacts (see also Figure 5.2 on page 91).

Using specific HISS score thresholds to predict proteins as belonging to the same fold, the sensitivity and specificity of such a prediction can be calculated by testing how many of all proteins that actually belong to the same fold satisfy this threshold and how many proteins that are classified to separate folds in SCOP and CATH have HISS scores below this threshold. The lower this score threshold is chosen, the more sensitive is such a prediction at the cost of reduced specificity. Aiming at a specificity of 80%, the best prediction is obtained with C9/C15 helix interaction predictions and weighted

HISS scores with a sensitivity of nearly 77%. Similarly, the best prediction with 90% specificity (again C9/C15 helix predictions but unweighted HISS scores) resulted in a sensitivity of 65%.

In summary, these results are highly encouraging with respect to the structural classification of membrane proteins. Of course, similar structures can not be identified with equal quality as based on known structures (for comparison see Table 6.4), but still a large fraction of all proteins having the same helix architecture can be recognized with high specificity. Importantly, this similarity can also be determined using predicted helix interactions in case the sequence similarity of two analyzed proteins is to low to confidently assign a common fold. For example, bovine rhodopsin (PDB 1GZM, chain A) was found to have HISS scores ≥ 0.9 with several other rhodopsins such as halorhodopsin (PDB 1E12, chain A) or sensory rhodopsin (PDB 1XIO, chain A) although the sequence similarity among these proteins is too low to obtain a proper sequence alignment. While existing classification approaches specifically addressing membrane protein always use sequence similarity as major criterion for a common classification [36, 35], the combination of predicted helix interactions and HISS scores offers the completely new possibility of deriving structural similarity originating for example from convergent evolution that is not approachable by these other classification systems.

Finally, these promising results can also be repeated on the full set of PDB proteins. Using HISS scores incorporating edge weights in combination with the MCL clustering algorithm, the original helix architecture based clustering obtained based on known 3D structures could be reproduced with predicted helix interactions with a sensitivity of 68.2% and a specificity of 82%. Thereby, both values were again calculated on the basis of protein pairs, i.e. sensitivity for example describes the fraction of protein pairs classified to the same fold in the prediction based classification out of all protein pairs classified together in the original structure based classification. On the cluster level, eleven of the original twenty helix architecture clusters were also found in the classification using predicted helix interactions albeit usually with reduced size. Notably, classification errors (missing helix architectures as well as wrongly co-classified proteins) appeared especially for proteins with six or less transmembrane helices where wrongly predicted helix interactions have a stronger impact simply because of the reduced number of possible interactions. For proteins with six transmembrane helices for example, only helix architectures HA3 and HA5 were correctly identified resulting in four missing helix architectures, which equals the number of missing helix architecture for all proteins with seven or more transmembrane helices combined.

Nevertheless, these results again confirm that structural similarities and even more so common helix architectures can be deduced also from predicted helix interaction patterns opening new perspectives for the large-scale analysis and classification of alpha-helical membrane proteins.

7 Conclusions

Overall, this thesis aims at a better understanding of helix interactions occurring in alpha-helical membrane proteins. It intends to provide new algorithms specifically developed for membrane proteins that can be used to predict contacts and interactions both on a residue level and on the full helix level. Furthermore, the field of structural classification is presented as one possible area of application where patterns of helix interactions (either obtained from known structures or predicted with the introduced methods) can be successfully used for identifying proteins with common helix architectures and accordingly highly similar folds.

Based on the obtained results presented in preceding chapters, several major conclusions can be drawn.

7.1 Helix interactions im membrane proteins are promoted by a diverse range of amino acids and interaction motifs

The amino acid composition of transmembrane protein domains is strongly biased towards hydrophobic residues in order to adjust to the lipophilic environment of the membrane. Nevertheless, a large number of experimental studies and sequence analyses of membrane protein structures have detected a surprising variety regarding the amino acids promoting strong helix interactions that were suggested to be even more diverse than found in soluble proteins [60, 61]. Even the currently best analyzed recurrent sequence pattern GxxxG, frequently reported to be a potent helix interaction motif, has lately been shown to be dependent on local sequence context for strong helix interaction [59, 77, 131].

Here, results of experimental and computational analyses of bitopic membrane proteins further demonstrated that again a diverse range of amino acids can lead in combination with the GxxxG motif to high-affine helix interactions. Several potent helix interaction motifs were identified consisting of a GxxxG portion and one or more other residues including the aromatic phenylalanine, the polar histidine and combinations of oppositely charged amino acids. As histidine and charged amino acids alone were not found to promote strong helix interactions, the GxxxG motif seems to be especially important for positioning these additional residues appropriately which are then able to contribute to high-affine helix interactions by forming aromatic π-π interactions, hydrogen bonds or ionic interactions.

Although all experimentally found motifs could be detected also in natural bitopic membrane proteins, the frequency with which they occurred varied significantly with the FxxGxxxG motif being found significantly more often than expected in several hundred proteins while motifs containing histidine or charged amino acids were found only in a limited number of sequences. Considering the hydrophobicity of phenylalanine, the formation of strong helix interactions by combining this residue with two glycines naturally seems to be evolutionary much easier approachable than the inclusion of amino acids occurring rarely in transmembrane domains such as histidine or charged residues. However, as the latter residues often may be functionally relevant as well, their additional structural importance shown here further highlights that membrane proteins are diverse to an extent often not fully appreciated yet.

7.2 Residue co-evolution affects the sequence neighbourhood of helix-helix contacts

Within Chapter 3, the first analysis of residue co-evolution in alpha-helical membrane proteins was presented. In agreement with studies conducted on soluble proteins, it could be observed that only a small fraction of predicted correlations actually involved pairs of residues in physical contact. However, up to 50% of all strongly correlated residue pairs with individual prediction methods were found be in close vicinity to interhelical contacts. Combining the outcome of several prediction methods into a consensus prediction this fraction could even be further increased to more than 55%. While recent publications analyzing co-evolving residues have already highlighted that residue co-evolution may have also other than structural reasons [174, 198, 200], these results additionally indicate that also on a structural level co-evolution not only occurs to maintain specific

amino acids required for a structural contact but also influences the correct formation of a helix-helix contact by affecting the sequence context of this contact. Accordingly, residue co-evolution appears to be a comprehensive manifestation of the complex task evolution has to cope with when balancing the test of new sequence variants with the need of maintaining the functionality of an organism's proteome.

With respect to the successful and valuable prediction of helix-helix contact, the analysis of co-evolving residues in membrane proteins demonstrated both limits and potentials. On the one hand, obtained prediction accuracies were clearly too low to make co-evolving residues alone a useful prediction method for helix-helix contacts. On the other hand, their frequent occurrence in close sequence neighbourhood to real helix-helix contacts suggested that they might be an important source to derive helix pairs that are likely to be in direct contact, possibly along with the approximate region of interaction. Furthermore, given that prediction accuracies were largely consistent with those reported for soluble proteins, the combination of residue co-evolution with other sequence features promised further gain in prediction accuracy as already demonstrated also for soluble proteins. Both aspects were further evaluated with subsequent analyses (see below).

7.3 Helix-helix contacts in membrane proteins can be predicted with equal accuracy than soluble residue contacts

While a large number of algorithms were already available for the prediction of membrane protein topology or the prediction of lipid-exposed surfaces (for reviews see [43, 96]), here the first method for the prediction of helix-helix contacts using neural networks was introduced. This method is specific for alpha-helical membrane proteins due to two reasons. First, the neural network was trained on a data set of 62 membrane proteins with solved structure. Secondly, sequence features were included that can only be derived for membrane proteins with alpha-helix bundle fold. With a final prediction accuracy of close to 26%, the newly developed method called TMHcon not only performs with equal accuracy as reported for current contact predictors on soluble proteins, but also easily outperforms these methods when using them on membrane proteins. Interestingly, from all proteins in the dataset consistently best results were obtained for the important class of proteins with seven transmembrane helices indicating that the helix architecture of

these proteins is specifically approachable to the implemented neural network.

While the experimental determination of membrane protein structures remains to be a difficult and time-consuming process, computational methods for the prediction of structural features of membrane proteins are required to close the gap between available sequence and structure data of membrane proteins. The contact predictor TMHcon will hopefully contribute to this task by providing on the one hand potential constraints for *ab initio* structure prediction experiments of membrane proteins and on the other hand enough structural information for distinguishing different helix architectures (see below).

7.4 Helix interaction patterns can be obtained with high reliability from predicted helix-helix contacts

Membrane protein structures can differ from each other in a number of structural features, including their number of transmembrane helices in the first place as well as length and folding of extramembranous loops or helix abnormalities such as kinks. Additionally, proteins with the same number of transmembrane helices can be further characterized by their specific patterns of helix interactions defining the helix bundle architecture of each protein. Within this thesis, a novel graph visualization for these helix architectures is introduced depicting individual helices as graph nodes that are connected in case of observed helix-helix contacts. These so-called helix interaction graphs promise easy access to the detection of structural similarities and differences as they constitute a high-level representation of membrane protein structures.

However, as membrane protein structures are rarely available, it was furthermore necessary to test, how well such helix interaction graphs can be predicted from sequence. While other prediction approaches are imaginable (a machine-learning algorithm for example might deduce possible interactions from helix properties directly), here a two-step approach was evaluated where helix interactions are derived from prior predicted helix-helix contacts (either using co-evolving residues or with the more complex neural network approach TMHcon). Importantly, such a prediction could be shown to reach accuracies of close to 80% even though the accuracies of used residue contact predictions are not exceeding 26%. Thereby, the prediction of interacting helices benefits from the fact that also wrongly predicted contacts still have a strong tendency to be in close sequence neighbourhood to observed helix-helix contacts and therefore cluster on actually interacting helices. While this characteristic has so far not found any major

application in the analysis of soluble proteins, the helix bundle structures of membrane proteins is offering the perfect environment for actually exploiting this information.

While co-evolving residues alone were already found to predict interacting helices with good specificity yet limited sensitivity, two main ways of further improving the prediction could be demonstrated. First, an increase in prediction accuracy during the initial prediction of helix-helix contacts resulted also in better helix interaction predictions as could be seen from using contacts predicted with TMHcon instead of co-evolving residues alone. Secondly, helix-helix contacts predicted not for the test sequence itself but for structurally related proteins can further contribute to an increased prediction sensitivity without reducing specificity significantly. Combining all tested improvement strategies, helix interactions could be predicted with a sensitivity of 63% at a specificity of 80%. Whether these values will be further improved in the future or constitute already the theoretically reachable optimum is hard to tell in general, although the idea of using predicted residue contacts seems to be largely exhausted given that contact prediction methods in the field of soluble proteins were not found to significantly increase in their prediction performance over the last years. Nevertheless, the prediction of helix interactions and accordingly helix architectures constitutes a completely new and valuable field in structural bioinformatics of membrane proteins, hopefully motivating other researchers to contribute to this problem.

7.5 Structural classification of membrane proteins is possible - with limitations

Comparing the structural classification of membrane proteins within SCOP and CATH, both databases were found to agree to a large extent when it comes to the classification of domains with five or more transmembrane helices. Discrepancies previously described for soluble proteins (differing domain assignments and fold overlap problems) were detected mostly for proteins with two or four transmembrane helices. A comparison to soluble four helix bundle proteins revealed that this observation is not automatically tied to a possibly limited structural variability of four helix bundles per se, but rather is specific for membrane proteins. Since structure comparisons additionally indicate that four helix bundle membrane domains in fact display a generally higher similarity among each other than comparable soluble helix bundles, their structure space seems to be highly continuous making their classification intrinsically more difficult.

Obviously, membrane proteins with more transmembrane helices are also structurally

restricted and hence likely to be more similar between each other than soluble helix bundles. However, with an increasing number of transmembrane segments the spectrum of possible structural variations (especially the number of complex helix interaction patterns) grows as well. Given the current status of membrane proteins in structural databases such as SCOP and CATH but also considering the helix architecture based classification presented in this work, the identification of distinct folds for membrane proteins having at least five transmembrane helices seems to possible and accordingly a classification similar to that for soluble proteins can be executed for these proteins.

7.6 Membrane proteins can be classified according to recurrent helix interaction patterns

Combining the analysis and prediction of helix interactions with the structural classification of membrane proteins, a new classification approach was proposed trying to cluster proteins based on similar helix interaction graphs. Thereby, nearly all known membrane protein structures with five or more transmembrane helices could be assigned to twenty recurrent helix architectures confirming the principal possibility of obtaining distinct membrane protein folds even though membrane proteins are structurally more restricted than soluble proteins. Furthermore, the obtained classification of helix architectures was largely consistent with general structural classification approaches such as SCOP and CATH demonstrating that helix interactions constitute maybe the most distinctive characteristic of membrane protein structures.

Importantly, common helix interaction patterns can not only be derived from known structures but also using predicted helix interactions. While this was shown here only using a small subset of all available membrane protein sequences (namely those having also an experimentally determined structure), the development of such a classification system incorporating membrane proteins from all currently sequenced organisms promises exciting insights into the structural diversity but also the evolution of membrane proteins in general.

8
Bibliography

[1] R. Hooke. *Micrographia: Some Physiological Descriptions of Minute Bodies Made by Magnifying Glasses with Observations and Inquiries Thereupon.* Royal Society London, 1665.

[2] S. J. Singer and G. L. Nicolson. The fluid mosaic model of the structure of cell membranes. *Science*, 175(23):720–731, Feb 1972.

[3] M. Edidin. The state of lipid rafts: from model membranes to cells. *Annu Rev Biophys Biomol Struct*, 32:257–283, 2003.

[4] K. Jacobson, O. G. Mouritsen, and R. G. W. Anderson. Lipid rafts: at a crossroad between cell biology and physics. *Nat Cell Biol*, 9(1):7–14, Jan 2007.

[5] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular biology of the cell.* Garland Science New York, 2002.

[6] D. L. Daleke. Phospholipid flippases. *J Biol Chem*, 282(2):821–825, Jan 2007.

[7] T. J. McIntosh and S. A. Simon. Roles of bilayer material properties in function and distribution of membrane proteins. *Annu Rev Biophys Biomol Struct*, 35:177–198, 2006.

[8] F. R. Maxfield and I. Tabas. Role of cholesterol and lipid organization in disease. *Nature*, 438(7068):612–621, Dec 2005.

[9] D. J. Müller, N. Wu, and K. Palczewski. Vertebrate membrane proteins: structure, function, and insights from biophysical approaches. *Pharmacol Rev*, 60(1):43–78, Mar 2008.

[10] G. Guidotti. The composition of biological membranes. *Arch Intern Med*, 129(2):194–201, Feb 1972.

[11] D. M. Engelman. Membranes are more mosaic than fluid. *Nature*, 438(7068):578–580, Dec 2005.

[12] M. Edidin. Lipids on the frontier: a century of cell-membrane bilayers. *Nat Rev Mol Cell Biol*, 4(5):414–418, May 2003.

[13] M. J. Saxton and K. Jacobson. Single-particle tracking: applications to membrane dynamics. *Annu Rev Biophys Biomol Struct*, 26:373–399, 1997.

[14] E. Yechiel and M. Edidin. Micrometer-scale domains in fibroblast plasma membranes. *J Cell Biol*, 105(2):755–760, Aug 1987.

[15] M. Edidin, S. C. Kuo, and M. P. Sheetz. Lateral movements of membrane glycoproteins restricted by dynamic cytoplasmic barriers. *Science*, 254(5036):1379–1382, Nov 1991.

[16] K. Jacobson, E. D. Sheets, and R. Simson. Revisiting the fluid mosaic model of membranes. *Science*, 268(5216):1441–1442, Jun 1995.

[17] M. Edidin. Shrinking patches and slippery rafts: scales of domains in the plasma membrane. *Trends Cell Biol*, 11(12):492–496, Dec 2001.

[18] C. A. Day and A. K. Kenworthy. Tracking microdomain dynamics in cell membranes. *Biochim Biophys Acta*, 1788(1):245–253, Jan 2009.

[19] M. Edidin. Patches, posts and fences: proteins and plasma membrane domains. *Trends Cell Biol*, 2(12):376–380, Dec 1992.

[20] K. Simons and E. Ikonen. Functional rafts in cell membranes. *Nature*, 387(6633):569–572, Jun 1997.

[21] S. Mukherjee and F. R. Maxfield. Membrane domains. *Annu Rev Cell Dev Biol*, 20:839–866, 2004.

[22] J. F. Hancock. Lipid rafts: contentious only from simplistic standpoints. *Nat Rev Mol Cell Biol*, 7(6):456–462, Jun 2006.

[23] E. Wallin and G. von Heijne. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci*, 7(4):1029–1038, Apr 1998.

[24] S. Galdiero, M. Galdiero, and C. Pedone. beta-barrel membrane bacterial proteins: structure, function, assembly and interaction with lipids. *Curr Protein Pept Sci*, 8(1):63–82, Feb 2007.

[25] G. von Heijne. The membrane protein universe: what's out there and why bother? *J Intern Med*, 261(6):543–557, Jun 2007.

[26] M. C. Lagerstrom and H. B. Schioth. Structural diversity of g protein-coupled receptors and significance for drug discovery. *Nat Rev Drug Discov*, 7(4):339–357, Apr 2008.

[27] A. Marchese, S. R. George, L. F. Kolakowski, K. R. Lynch, and B. F. O'Dowd. Novel gpcrs and their endogenous ligands: expanding the boundaries of physiology and pharmacology. *Trends Pharmacol Sci*, 20(9):370–375, Sep 1999.

[28] K. Lundstrom. Structural genomics of gpcrs. *Trends Biotechnol*, 23(2):103–108, Feb 2005.

[29] A. G. Therien, F. E. Grant, and C. M. Deber. Interhelical hydrogen bonds in the cftr membrane domain. *Nat Struct Biol*, 8(7):597–601, Jul 2001.

[30] M. T. Malecki. Genetics of type 2 diabetes mellitus. *Diabetes Res Clin Pract*, 68 Suppl1:S10–S21, Jun 2005.

[31] D. O. Daley, M. Rapp, E. Granseth, K. Melen, D. Drew, and G. von Heijne. Global topology analysis of the escherichia coli inner membrane proteome. *Science*, 308(5726):1321–1323, May 2005.

[32] H. Kim, K. Melen, M. Osterberg, and G. von Heijne. A global topology map of the saccharomyces cerevisiae membrane proteome. *Proc Natl Acad Sci U S A*, 103(30):11142–11147, Jul 2006.

[33] R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H.-R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. L. Sonnhammer, and A. Bateman. The pfam protein families database. *Nucleic Acids Res*, 36(Database issue):D281–D288, Jan 2008.

[34] Y. Liu, D. M. Engelman, and M. Gerstein. Genomic analysis of membrane protein families: abundance and conserved motifs. *Genome Biol*, 3(10):research0054.1–research0054.12, Sep 2002.

[35] A. Oberai, Y. Ihm, S. Kim, and J. U. Bowie. A limited universe of membrane protein families and folds. *Protein Sci*, 15(7):1723–1734, Jul 2006.

[36] A. J. Martin-Galiano and D. Frishman. Defining the fold space of membrane proteins: the camps database. *Proteins*, 64(4):906–922, Sep 2006.

[37] G. von Heijne. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. *EMBO J*, 5(11):3021–3027, Nov 1986.

[38] J. Nilsson, B. Persson, and G. von Heijne. Comparative analysis of amino acid distributions in integral membrane proteins from 107 genomes. *Proteins*, 60(4):606–616, Sep 2005.

[39] G. von Heijne. Membrane-protein topology. *Nat Rev Mol Cell Biol*, 7(12):909–918, Dec 2006.

[40] M. Rapp, E. Granseth, S. Seppälä, and G. von Heijne. Identification and evolution of dual-topology membrane proteins. *Nat Struct Mol Biol*, 13(2):112–116, Feb 2006.

[41] S. J. Kim, R. Rahbar, and R. S. Hegde. Combinatorial control of prion protein biogenesis by the signal sequence and transmembrane domain. *J Biol Chem*, 276(28):26132–26140, Jul 2001.

[42] S. Nagamori, K. ichi Nishiyama, and H. Tokuda. Membrane topology inversion of secg detected by labeling with a membrane-impermeable sulfhydryl reagent that causes a close association of secg with seca. *J Biochem*, 132(4):629–634, Oct 2002.

[43] A. Elofsson and G. von Heijne. Membrane protein structure: prediction versus reality. *Annu Rev Biochem*, 76:125–140, 2007.

[44] S. J. Fleishman and N. Ben-Tal. Progress in structure prediction of alpha-helical membrane proteins. *Curr Opin Struct Biol*, 16(4):496–504, Aug 2006.

[45] I. T. Arkin and A. T. Brunger. Statistical analysis of predicted transmembrane alpha-helices. *Biochim Biophys Acta*, 1429(1):113–128, Dec 1998.

[46] M. B. Ulmschneider and M. S. Sansom. Amino acid distributions in integral membrane protein structures. *Biochim Biophys Acta*, 1512(1):1–14, May 2001.

[47] D. T. Jones, W. R. Taylor, and J. M. Thornton. A mutation data matrix for transmembrane proteins. *FEBS Lett*, 339(3):269–275, Feb 1994.

[48] M. Eilers, S. C. Shekar, T. Shieh, S. O. Smith, and P. J. Fleming. Internal packing of helical membrane proteins. *Proc Natl Acad Sci U S A*, 97(11):5796–5801, May 2000.

[49] T. Hessa, H. Kim, K. Bihlmaier, C. Lundin, J. Boekel, H. Andersson, I. Nilsson, S. H. White, and G. von Heijne. Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature*, 433(7024):377–381, Jan 2005.

[50] T. Hessa, S. H. White, and G. von Heijne. Membrane insertion of a potassium-channel voltage sensor. *Science*, 307(5714):1427, Mar 2005.

[51] T. A. Eyre, L. Partridge, and J. M. Thornton. Computational analysis of alpha-helical membrane protein structure: implications for the prediction of 3d structural models. *Protein Eng Des Sel*, 17(8):613–624, Aug 2004.

[52] O. G. Mouritsen and M. Bloom. Mattress model of lipid-protein interactions in membranes. *Biophys J*, 46(2):141–153, Aug 1984.

[53] S. H. Park and S. J. Opella. Tilt angle of a trans-membrane helix is determined by hydrophobic mismatch. *J Mol Biol*, 350(2):310–318, Jul 2005.

[54] K. G. Fleming and D. M. Engelman. Specificity in transmembrane helix-helix interactions can define a hierarchy of stability for sequence variants. *Proc Natl Acad Sci U S A*, 98(25):14340–14344, Dec 2001.

[55] B. D. Adair and D. M. Engelman. Glycophorin a helical transmembrane domains dimerize in phospholipid bilayers: a resonance energy transfer study. *Biochemistry*, 33(18):5539–5544, May 1994.

[56] L. Cristian, J. D. Lear, and W. F. DeGrado. Use of thiol-disulfide equilibria to measure the energetics of assembly of transmembrane helices in phospholipid bilayers. *Proc Natl Acad Sci U S A*, 100(25):14772–14777, Dec 2003.

[57] D. Langosch, B. Brosig, H. Kolmar, and H. J. Fritz. Dimerisation of the glycophorin a transmembrane segment in membranes probed with the toxr transcription activator. *J Mol Biol*, 263(4):525–530, Nov 1996.

[58] W. P. Russ and D. M. Engelman. Toxcat: a measure of transmembrane helix association in a biological membrane. *Proc Natl Acad Sci U S A*, 96(3):863–868, Feb 1999.

[59] K. R. MacKenzie and K. G. Fleming. Association energetics of membrane spanning alpha-helices. *Curr Opin Struct Biol*, 18(4):412–419, Aug 2008.

[60] L. Adamian and J. Liang. Helix-helix packing and interfacial pairwise interactions of residues in membrane proteins. *J Mol Biol*, 311(4):891–907, Aug 2001.

[61] M. Eilers, A. B. Patel, W. Liu, and S. O. Smith. Comparison of helix interactions in membrane and soluble alpha-bundle proteins. *Biophys J*, 82(5):2720–2736, May 2002.

[62] M. Gimpelev, L. R. Forrest, D. Murray, and B. Honig. Helical packing patterns in membrane and soluble proteins. *Biophys J*, 87(6):4075–4086, Dec 2004.

[63] D. Langosch and J. Heringa. Interaction of transmembrane helices by a knobs-into-holes packing characteristic of soluble coiled coils. *Proteins*, 31(2):150–159, May 1998.

[64] J. U. Bowie. Helix packing angle preferences. *Nat Struct Biol*, 4(11):915–917, Nov 1997.

[65] A. R. Curran and D. M. Engelman. Sequence motifs, polar interactions and conformational changes in helical membrane proteins. *Curr Opin Struct Biol*, 13(4):412–417, Aug 2003.

[66] D. Langosch, E. Lindner, and R. Gurezka. In vitro selection of self-interacting transmembrane segments–membrane proteins approached from a different perspective. *IUBMB Life*, 54(3):109–113, Sep 2002.

[67] J. P. Dawson, J. S. Weinger, and D. M. Engelman. Motifs of serine and threonine can drive association of transmembrane helices. *J Mol Biol*, 316(3):799–805, Feb 2002.

[68] F. X. Zhou, M. J. Cocco, W. P. Russ, A. T. Brunger, and D. M. Engelman. Interhelical hydrogen bonding drives strong interactions in membrane proteins. *Nat Struct Biol*, 7(2):154–160, Feb 2000.

[69] J. M. Mendrola, M. B. Berger, M. C. King, and M. A. Lemmon. The single transmembrane domains of erbb receptors self-associate in cell membranes. *J Biol Chem*, 277(7):4704–4712, Feb 2002.

[70] M. C. Overton, S. L. Chinault, and K. J. Blumer. Oligomerization, biogenesis, and signaling is promoted by a glycophorin a-like dimerization motif in transmembrane domain 1 of a yeast g protein-coupled receptor. *J Biol Chem*, 278(49):49369–49377, Dec 2003.

[71] S.-F. Lee, S. Shah, C. Yu, W. C. Wigley, H. Li, M. Lim, K. Pedersen, W. Han, P. Thomas, J. Lundkvist, Y.-H. Hao, and G. Yu. A conserved gxxxg motif in aph-1 is critical for assembly and activity of the gamma-secretase complex. *J Biol Chem*, 279(6):4144–4152, Feb 2004.

[72] M. A. Lemmon, J. M. Flanagan, H. R. Treutlein, J. Zhang, and D. M. Engelman. Sequence specificity in the dimerization of transmembrane alpha-helices. *Biochemistry*, 31(51):12719–12725, Dec 1992.

[73] B. Brosig and D. Langosch. The dimerization motif of the glycophorin a transmembrane segment in membranes: importance of glycine residues. *Protein Sci*, 7(4):1052–1056, Apr 1998.

[74] A. Senes, M. Gerstein, and D. M. Engelman. Statistical analysis of amino acid patterns in transmembrane helices: the gxxxg motif occurs frequently and in association with beta-branched residues at neighboring positions. *J Mol Biol*, 296(3):921–936, Feb 2000.

[75] M. M. Javadpour, M. Eilers, M. Groesbeek, and S. O. Smith. Helix packing in polytopic membrane proteins: role of glycine in transmembrane helix association. *Biophys J*, 77(3):1609–1618, Sep 1999.

[76] A. Senes, I. Ubarretxena-Belandia, and D. M. Engelman. The calpha —h...o hydrogen bond: a determinant of stability and specificity in transmembrane helix interactions. *Proc Natl Acad Sci U S A*, 98(16):9056–9061, Jul 2001.

[77] A. K. Doura, F. J. Kobus, L. Dubrovsky, E. Hibbard, and K. G. Fleming. Sequence context modulates the stability of a gxxxg-mediated transmembrane helix-helix dimer. *J Mol Biol*, 341(4):991–998, Aug 2004.

[78] T. K. M. Nyholm, S. Ozdirekcan, and J. A. Killian. How protein transmembrane segments sense the lipid environment. *Biochemistry*, 46(6):1457–1465, Feb 2007.

[79] J. Ren, S. Lew, J. Wang, and E. London. Control of the transmembrane orientation and interhelical interactions within membranes by hydrophobic helix length. *Biochemistry*, 38(18):5905–5912, May 1999.

[80] R. Phillips, T. Ursell, P. Wiggins, and P. Sens. Emerging roles for lipids in shaping membrane-protein function. *Nature*, 459(7245):379–385, May 2009.

[81] M. Kolbe, H. Besir, L. O. Essen, and D. Oesterhelt. Structure of the light-driven chloride pump halorhodopsin at 1.8 a resolution. *Science*, 288(5470):1390–1396, May 2000.

[82] C. Toyoshima, M. Nakasako, H. Nomura, and H. Ogawa. Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 a resolution. *Nature*, 405(6787):647–655, Jun 2000.

[83] H. Luecke, B. Schobert, H. T. Richter, J. P. Cartailler, and J. K. Lanyi. Structure of bacteriorhodopsin at 1.55 a resolution. *J Mol Biol*, 291(4):899–911, Aug 1999.

[84] H. Luecke. Atomic resolution structures of bacteriorhodopsin photocycle intermediates: the role of discrete water molecules in the function of this light-driven ion pump. *Biochim Biophys Acta*, 1460(1):133–156, Aug 2000.

[85] R. P. Riek, I. Rigoutsos, J. Novotny, and R. M. Graham. Non-alpha-helical elements modulate polytopic membrane protein architecture. *J Mol Biol*, 306(2):349–362, Feb 2001.

[86] H. Viklund, E. Granseth, and A. Elofsson. Structural classification and prediction of reentrant regions in alpha-helical transmembrane proteins: application to complete genomes. *J Mol Biol*, 361(3):591–603, Aug 2006.

[87] A. Engel and H. E. Gaub. Structure and mechanics of membrane proteins. *Annu Rev Biochem*, 77:127–148, 2008.

[88] E. P. Carpenter, K. Beis, A. D. Cameron, and S. Iwata. Overcoming the challenges of membrane protein crystallography. *Curr Opin Struct Biol*, 18(5):581–586, Oct 2008.

[89] J. L. Popot and D. M. Engelman. Helical membrane protein folding, stability, and evolution. *Annu Rev Biochem*, 69:881–922, 2000.

[90] J. L. Popot and D. M. Engelman. Membrane protein folding and oligomerization: the two-stage model. *Biochemistry*, 29(17):4031–4037, May 1990.

[91] B. V. den Berg, W. M. Clemons, I. Collinson, Y. Modis, E. Hartmann, S. C. Harrison, and T. A. Rapoport. X-ray structure of a protein-conducting channel. *Nature*, 427(6969):36–44, Jan 2004.

[92] S. H. White and G. von Heijne. How translocons select transmembrane helices. *Annu Rev Biophys*, 37:23–42, 2008.

[93] S. U. Heinrich, W. Mothes, J. Brunner, and T. A. Rapoport. The sec61p complex mediates the integration of a membrane protein by allowing lipid partitioning of the transmembrane domain. *Cell*, 102(2):233–244, Jul 2000.

[94] H. Sadlish, D. Pitonzo, A. E. Johnson, and W. R. Skach. Sequential triage of transmembrane segments by sec61alpha during biogenesis of a native multispanning membrane protein. *Nat Struct Mol Biol*, 12(10):870–878, Oct 2005.

[95] N. M. Meindl-Beinker, C. Lundin, I. Nilsson, S. H. White, and G. von Heijne. Asn- and asp-mediated interactions between transmembrane helices during translocon-mediated membrane protein assembly. *EMBO Rep*, 7(11):1111–1116, Nov 2006.

[96] M. Punta, L. R. Forrest, H. Bigelow, A. Kernytsky, J. Liu, and B. Rost. Membrane protein prediction methods. *Methods*, 41(4):460–474, Apr 2007.

[97] A. Krogh, B. Larsson, G. von Heijne, and E. L. Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J Mol Biol*, 305(3):567–580, Jan 2001.

[98] L. Kall, A. Krogh, and E. L. L. Sonnhammer. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol*, 338(5):1027–1036, May 2004.

[99] J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, 157(1):105–132, May 1982.

[100] B. Rost, R. Casadio, P. Fariselli, and C. Sander. Transmembrane helices predicted at 95 *Protein Sci*, 4(3):521–533, Mar 1995.

[101] G. E. Tusnady and I. Simon. The hmmtop transmembrane topology prediction server. *Bioinformatics*, 17(9):849–850, Sep 2001.

[102] C. P. Chen and B. Rost. State-of-the-art in membrane protein prediction. *Appl Bioinformatics*, 1(1):21–35, 2002.

[103] L. Kall and E. L. L. Sonnhammer. Reliability of transmembrane predictions in whole-genome data. *FEBS Lett*, 532(3):415–418, Dec 2002.

[104] S. Moller, M. D. Croning, and R. Apweiler. Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, 17(7):646–653, Jul 2001.

[105] H. Viklund and A. Elofsson. Best alpha-helical transmembrane protein topology predictions are achieved using hidden markov models and evolutionary information. *Protein Sci*, 13(7):1908–1917, Jul 2004.

[106] S. M. Reynolds, L. Kall, M. E. Riffle, J. A. Bilmes, and W. S. Noble. Transmembrane topology and signal peptide prediction using dynamic bayesian networks. *PLoS Comput Biol*, 4(11):e1000213, Nov 2008.

[107] A. Bernsel, H. Viklund, A. Hennerdal, and A. Elofsson. Topcons: consensus prediction of membrane protein topology. *Nucleic Acids Res*, 37(Web Server issue):W465–W468, Jul 2009.

[108] A. Bernsel, H. Viklund, J. Falk, E. Lindahl, G. von Heijne, and A. Elofsson. Prediction of membrane-protein topology from first principles. *Proc Natl Acad Sci U S A*, 105(20):7177–7181, May 2008.

[109] L. R. Forrest, C. L. Tang, and B. Honig. On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. *Biophys J*, 91(2):508–517, Jul 2006.

[110] J. M. Baldwin. The probable arrangement of the helices in g protein-coupled receptors. *EMBO J*, 12(4):1693–1703, Apr 1993.

[111] J. M. Baldwin, G. F. Schertler, and V. M. Unger. An alpha-carbon template for the transmembrane helices in the rhodopsin family of g-protein-coupled receptors. *J Mol Biol*, 272(1):144–164, Sep 1997.

[112] V. M. Unger, P. A. Hargrave, J. M. Baldwin, and G. F. Schertler. Arrangement of rhodopsin transmembrane alpha-helices. *Nature*, 389(6647):203–206, Sep 1997.

[113] R. J. Trabanino, S. E. Hall, N. Vaidehi, W. B. Floriano, V. W. T. Kam, and W. A. Goddard. First principles predictions of the structure and function of g-protein-coupled receptors: validation for bovine rhodopsin. *Biophys J*, 86(4):1904–1921, Apr 2004.

[114] V. Yarov-Yarovoy, J. Schonbrun, and D. Baker. Multipass membrane protein structure prediction using rosetta. *Proteins*, 62(4):1010–1025, Mar 2006.

[115] P. Barth, J. Schonbrun, and D. Baker. Toward high-resolution prediction and design of transmembrane helical protein structures. *Proc Natl Acad Sci U S A*, 104(40):15682–15687, Oct 2007.

[116] P. Barth, B. Wallner, and D. Baker. Prediction of membrane protein structures with complex topologies using limited constraints. *Proc Natl Acad Sci U S A*, 106(5):1409–1414, Feb 2009.

[117] S. J. Fleishman, V. M. Unger, and N. Ben-Tal. Transmembrane protein structures without x-rays. *Trends Biochem Sci*, 31(2):106–113, Feb 2006.

[118] K. Sale, J.-L. Faulon, G. A. Gray, J. S. Schoeniger, and M. M. Young. Optimal bundling of transmembrane helices using sparse distance constraints. *Protein Sci*, 13(10):2613–2627, Oct 2004.

[119] T. Beuming and H. Weinstein. Modeling membrane proteins based on low-resolution electron microscopy maps: a template for the tm domains of the oxalate transporter oxlt. *Protein Eng Des Sel*, 18(3):119–125, Mar 2005.

[120] S. J. Fleishman, S. Harrington, R. A. Friesner, B. Honig, and N. Ben-Tal. An automatic method for predicting transmembrane protein structures using cryo-em and evolutionary data. *Biophys J*, 87(5):3448–3459, Nov 2004.

[121] S. J. Fleishman, O. Yifrach, and N. Ben-Tal. An evolutionarily conserved network of amino acids mediates gating in voltage-dependent potassium channels. *J Mol Biol*, 340(2):307–318, Jul 2004.

[122] L. Adamian and J. Liang. Prediction of transmembrane helix orientation in polytopic membrane proteins. *BMC Struct Biol*, 6:13, 2006.

[123] T. Beuming and H. Weinstein. A knowledge-based scale for the analysis and prediction of buried and exposed faces of transmembrane domain proteins. *Bioinformatics*, 20(12):1822–1835, Aug 2004.

[124] Y. Pilpel, N. Ben-Tal, and D. Lancet. kprot: a knowledge-based scale for the propensity of residue orientation in transmembrane segments. application to membrane protein structure prediction. *J Mol Biol*, 294(4):921–935, Dec 1999.

[125] S. Yohannan, S. Faham, D. Yang, J. P. Whitelegge, and J. U. Bowie. The evolution of transmembrane helix kinks and the structural diversity of g protein-coupled receptors. *Proc Natl Acad Sci U S A*, 101(4):959–963, Jan 2004.

[126] G. Lasso, J. F. Antoniw, and J. G. L. Mullins. A combinatorial pattern discovery approach for the prediction of membrane dipping (re-entrant) loops. *Bioinformatics*, 22(14):e290–e297, Jul 2006.

[127] D. A. Doyle, J. M. Cabral, R. A. Pfuetzner, A. Kuo, J. M. Gulbis, S. L. Cohen, B. T. Chait, and R. MacKinnon. The structure of the potassium channel: molecular basis of k+ conduction and selectivity. *Science*, 280(5360):69–77, Apr 1998.

[128] T. Walz, T. Hirai, K. Murata, J. B. Heymann, K. Mitsuoka, Y. Fujiyoshi, B. L. Smith, P. Agre, and A. Engel. The three-dimensional structure of aquaporin-1. *Nature*, 387(6633):624–627, Jun 1997.

[129] E. Granseth, H. Viklund, and A. Elofsson. Zpred: predicting the distance to the membrane center for residues in alpha-helical membrane proteins. *Bioinformatics*, 22(14):e191–e196, Jul 2006.

[130] L. Adamian and J. Liang. Interhelical hydrogen bonds and spatial motifs in membrane proteins: polar clamps and serine zippers. *Proteins*, 47(2):209–218, May 2002.

[131] D. Schneider and D. M. Engelman. Motifs of two small residues can assist but are not sufficient to mediate transmembrane helix interactions. *J Mol Biol*, 343(4):799–804, Oct 2004.

[132] R. Gurezka and D. Langosch. In vitro selection of membrane-spanning leucine zipper protein-protein interaction motifs using possyccat. *J Biol Chem*, 276(49):45580–45587, Dec 2001.

[133] E. Lindner, S. Unterreitmeier, A. N. J. A. Ridder, and D. Langosch. An extended toxr possyccat system for positive and negative selection of self-interacting transmembrane domains. *J Microbiol Methods*, 69(2):298–305, May 2007.

[134] E. Lindner and D. Langosch. A toxr-based dominant-negative system to investigate heterotypic transmembrane domain interactions. *Proteins*, 65(4):803–807, Dec 2006.

[135] H. Yin, J. S. Slusky, B. W. Berger, R. S. Walters, G. Vilaire, R. I. Litvinov, J. D. Lear, G. A. Caputo, J. S. Bennett, and W. F. DeGrado. Computational design of peptides that target transmembrane helices. *Science*, 315(5820):1817–1822, Mar 2007.

[136] S. Unterreitmeier, A. Fuchs, T. Schäffler, R. G. Heym, D. Frishman, and D. Langosch. Phenylalanine promotes interaction of transmembrane domains via gxxxg motifs. *J Mol Biol*, 374(3):705–718, Nov 2007.

[137] J. R. Herrmann, J. C. Panitz, S. Unterreitmeier, A. Fuchs, D. Frishman, and D. Langosch. Complex patterns of histidine, hydroxylated amino acids and the gxxxg motif mediate high-affinity transmembrane domain interactions. *J Mol Biol*, 385(3):912–923, Jan 2009.

[138] J. R. Herrmann, A. Fuchs, J. C. Panitz, T. Eckert, S. Unterreitmeier, D. Frishman, and D. Langosch. Ionic interactions promote transmembrane helix-helix association depending on sequence context. *J Mol Biol*, submitted, 2009.

[139] V. L. Miller, R. K. Taylor, and J. J. Mekalanos. Cholera toxin transcriptional activator toxr is a transmembrane dna binding protein. *Cell*, 48(2):271–279, Jan 1987.

[140] K. M. Ottemann and J. J. Mekalanos. The toxr protein of vibrio cholerae forms homodimers and heterodimers. *J Bacteriol*, 178(1):156–162, Jan 1996.

[141] K. Skorupski and R. K. Taylor. Control of the toxr virulence regulon in vibrio cholerae by environmental stimuli. *Mol Microbiol*, 25(6):1003–1009, Sep 1997.

[142] S. Unterreitmeier. *Selektion und Charakterisierung hochaffiner Transmembrandomänen aus kombinatorischen Plasmidbanken*. PhD thesis, Fakultät Wissenschaftszentrum Weihenstephan, Technische Universität München, 2008.

[143] R. Gurezka, R. Laage, B. Brosig, and D. Langosch. A heptad motif of leucine residues found in membrane proteins can drive self-assembly of artificial transmembrane segments. *J Biol Chem*, 274(14):9265–9270, Apr 1999.

[144] U. Consortium. The universal protein resource (uniprot) 2009. *Nucleic Acids Res*, 37(Database issue):D169–D174, Jan 2009.

[145] J. D. Bendtsen, H. Nielsen, G. von Heijne, and S. Brunak. Improved prediction of signal peptides: Signalp 3.0. *J Mol Biol*, 340(4):783–795, Jul 2004.

[146] T. Rattei, P. Tischler, R. Arnold, F. Hamberger, J. Krebs, J. Krumsiek, B. Wachinger, V. Stümpflen, and W. Mewes. Simap–structuring the network of protein similarities. *Nucleic Acids Res*, 36(Database issue):D289–D292, Jan 2008.

[147] D. M. Engelman, T. A. Steitz, and A. Goldman. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu Rev Biophys Biophys Chem*, 15:321–353, 1986.

[148] L. Adamian, R. Jackups, T. A. Binkowski, and J. Liang. Higher-order interhelical spatial interactions in membrane proteins. *J Mol Biol*, 327(1):251–272, Mar 2003.

[149] E. S. Sulistijo and K. R. MacKenzie. Sequence dependence of bnip3 transmembrane domain dimerization implicates side-chain hydrogen bonding and a tandem gxxxg motif in specific helix-helix interactions. *J Mol Biol*, 364(5):974–990, Dec 2006.

[150] E. S. Sulistijo and K. R. Mackenzie. Structural basis for dimerization of the bnip3 transmembrane domain. *Biochemistry*, 48(23):5106–5120, Jun 2009.

[151] D. Shigematsu, M. Matsutani, T. Furuya, T. Kiyota, S. Lee, G. Sugihara, and S. Yamashita. Roles of peptide-peptide charge interaction and lipid phase separation in helix-helix association in lipid bilayer. *Biochim Biophys Acta*, 1564(1):271–280, Aug 2002.

[152] Z. Chen and Y. Xu. Structure prediction of helical transmembrane proteins at two length scales. *J Bioinform Comput Biol*, 4(2):317–333, Apr 2006.

[153] P. W. Hildebrand, S. Lorenzen, A. Goede, and R. Preissner. Analysis and prediction of helix-helix interactions in membrane channels and transporters. *Proteins*, 64(1):253–262, Jul 2006.

[154] D. Altschuh, A. M. Lesk, A. C. Bloomer, and A. Klug. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J Mol Biol*, 193(4):693–707, Feb 1987.

[155] D. Altschuh, T. Vernet, P. Berti, D. Moras, and K. Nagai. Coordinated amino acid changes in homologous protein families. *Protein Eng*, 2(3):193–199, Sep 1988.

[156] O. Olmea and A. Valencia. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold Des*, 2(3):S25–S32, 1997.

[157] O. Olmea, B. Rost, and A. Valencia. Effective use of sequence correlation and conservation in fold recognition. *J Mol Biol*, 293(5):1221–1239, Nov 1999.

[158] A. A. Fodor and R. W. Aldrich. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins*, 56(2):211–221, Aug 2004.

[159] I. Halperin, H. Wolfson, and R. Nussinov. Correlated mutations: advances and limitations. a study on fusion proteins and on the cohesin-dockerin families. *Proteins*, 63(4):832–845, Jun 2006.

[160] L. Oliveira, P. B. Paiva, A. C. M. Paiva, and G. Vriend. Sequence analysis reveals how g protein-coupled receptors transduce the signal to the g protein. *Proteins*, 52(4):553–560, Sep 2003.

[161] M. Filizola, O. Olmea, and H. Weinstein. Prediction of heterodimerization interfaces of g-protein coupled receptors with a new subtractive correlated mutation method. *Protein Eng*, 15(11):881–885, Nov 2002.

[162] A. Fuchs, A. J. Martin-Galiano, M. Kalman, S. Fleishman, N. Ben-Tal, and D. Frishman. Co-evolving residues in membrane proteins. *Bioinformatics*, 23(24):3312–3319, Dec 2007.

[163] U. Gobel, C. Sander, R. Schneider, and A. Valencia. Correlated mutations and residue contacts in proteins. *Proteins*, 18(4):309–317, Apr 1994.

[164] E. Neher. How frequent are correlated changes in families of protein sequences? *Proc Natl Acad Sci U S A*, 91(1):98–102, Jan 1994.

[165] W. R. Taylor and K. Hatrick. Compensating changes in protein multiple sequence alignments. *Protein Eng*, 7(3):341–348, Mar 1994.

[166] D. D. Pollock and W. R. Taylor. Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. *Protein Eng*, 10(6):647–657, Jun 1997.

[167] S. Vicatos, B. V. B. Reddy, and Y. Kaznessis. Prediction of distant residue contacts with the use of evolutionary information. *Proteins*, 58(4):935–949, Mar 2005.

[168] P. J. Kundrotas and E. G. Alexov. Predicting residue contacts using pragmatic correlated mutations method: reducing the false positives. *BMC Bioinformatics*, 7:503, 2006.

[169] I. Kass and A. Horovitz. Mapping pathways of allosteric communication in groel by analysis of correlated mutations. *Proteins*, 48(4):611–617, Sep 2002.

[170] S. M. Larson, A. A. D. Nardo, and A. R. Davidson. Analysis of covariation in an sh3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. *J Mol Biol*, 303(3):433–446, Oct 2000.

[171] O. Noivirt, M. Eisenstein, and A. Horovitz. Detection and reduction of evolutionary noise in correlated mutation analysis. *Protein Eng Des Sel*, 18(5):247–253, May 2005.

[172] D. D. Pollock, W. R. Taylor, and N. Goldman. Coevolving protein residues: maximum likelihood identification and relationship to structure. *J Mol Biol*, 287(1):187–198, Mar 1999.

[173] N. D. Clarke. Covariation of residues in the homeodomain sequence family. *Protein Sci*, 4(11):2269–2278, Nov 1995.

[174] G. B. Gloor, L. C. Martin, L. M. Wahl, and S. D. Dunn. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry*, 44(19):7156–7165, May 2005.

[175] L. C. Martin, G. B. Gloor, S. D. Dunn, and L. M. Wahl. Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, 21(22):4116–4124, Nov 2005.

[176] S. W. Lockless and R. Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438):295–299, Oct 1999.

[177] G. M. Süel, S. W. Lockless, M. A. Wall, and R. Ranganathan. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Biol*, 10(1):59–69, Jan 2003.

[178] J. P. Dekker, A. Fodor, R. W. Aldrich, and G. Yellen. A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. *Bioinformatics*, 20(10):1565–1572, Jul 2004.

[179] F. M. Codoner, S. O'Dea, and M. A. Fares. Reducing the false positive rate in the non-parametric analysis of molecular coevolution. *BMC Evol Biol*, 8:106, 2008.

[180] I. N. Shindyalov, N. A. Kolchanov, and C. Sander. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng*, 7(3):349–358, Mar 1994.

[181] G. Chelvanayagam, A. Eggenschwiler, L. Knecht, G. H. Gonnet, and S. A. Benner. An analysis of simultaneous variation in protein structures. *Protein Eng*, 10(4):307–316, Apr 1997.

[182] M. A. Fares and S. A. A. Travers. A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses. *Genetics*, 173(1):9–23, May 2006.

[183] S. D. Dunn, L. M. Wahl, and G. B. Gloor. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24(3):333–340, Feb 2008.

[184] B.-C. Lee and D. Kim. A new method for revealing correlated mutations under the structural and functional constraints in proteins. *Bioinformatics*, Epub ahead of print, Jul 2009.

[185] O. Grana, D. Baker, R. M. MacCallum, J. Meiler, M. Punta, B. Rost, M. L. Tress, and A. Valencia. Casp6 assessment of contact prediction. *Proteins*, 61 Suppl 7:214–224, 2005.

[186] J. M. G. Izarzugaza, O. Grana, M. L. Tress, A. Valencia, and N. D. Clarke. Assessment of intramolecular contact predictions for casp7. *Proteins*, 69 Suppl 8:152–158, 2007.

[187] N. Hamilton, K. Burrage, M. A. Ragan, and T. Huber. Protein contact prediction using patterns of correlation. *Proteins*, 56(4):679–684, Sep 2004.

[188] P. Fariselli, O. Olmea, A. Valencia, and R. Casadio. Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. *Proteins*, Suppl 5:157–162, 2001.

[189] M. Punta and B. Rost. Profcon: novel prediction of long-range contacts. *Bioinformatics*, 21(13):2960–2968, Jul 2005.

[190] J. Cheng and P. Baldi. Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics*, 8:113, 2007.

[191] G. Shackelford and K. Karplus. Contact prediction using mutual information and neural nets. *Proteins*, 69 Suppl 8:159–164, 2007.

[192] A. R. Ortiz, A. Kolinski, and J. Skolnick. Fold assembly of small proteins using monte carlo simulations driven by restraints derived from multiple sequence alignments. *J Mol Biol*, 277(2):419–448, Mar 1998.

[193] A. R. Ortiz, A. Kolinski, and J. Skolnick. Nativelike topology assembly of small proteins using predicted restraints in monte carlo folding simulations. *Proc Natl Acad Sci U S A*, 95(3):1020–1025, Feb 1998.

[194] F. Pazos, M. Helmer-Citterich, G. Ausiello, and A. Valencia. Correlated mutations contain information about protein-protein interaction. *J Mol Biol*, 271(4):511–523, Aug 1997.

[195] F. Pazos and A. Valencia. In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins*, 47(2):219–227, May 2002.

[196] M. Socolich, S. W. Lockless, W. P. Russ, H. Lee, K. H. Gardner, and R. Ranganathan. Evolutionary information for specifying a protein fold. *Nature*, 437(7058):512–518, Sep 2005.

[197] W. P. Russ, D. M. Lowery, P. Mishra, M. B. Yaffe, and R. Ranganathan. Natural-like function in artificial ww domains. *Nature*, 437(7058):579–583, Sep 2005.

[198] S. A. A. Travers and M. A. Fares. Functional coevolutionary networks of the hsp70-hop-hsp90 system revealed through computational analyses. *Mol Biol Evol*, 24(4):1032–1044, Apr 2007.

[199] Z. O. Wang and D. D. Pollock. Coevolutionary patterns in cytochrome c oxidase subunit i depend on structural and functional context. *J Mol Evol*, 65(5):485–495, Nov 2007.

[200] B.-C. Lee, K. Park, and D. Kim. Analysis of the residue-residue coevolution network and the functionally important residues in proteins. *Proteins*, 72(3):863–872, Aug 2008.

[201] G. E. Tusnady, Z. Dosztanyi, and I. Simon. Pdb_tm: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res*, 33(Database issue):D275–D278, Jan 2005.

[202] G. E. Tusnady, L. Kalmar, and I. Simon. Topdb: topology data bank of transmembrane proteins. *Nucleic Acids Res*, 36(Database issue):D234–D239, Jan 2008.

[203] M. A. Lomize, A. L. Lomize, I. D. Pogozheva, and H. I. Mosberg. Opm: orientations of proteins in membranes database. *Bioinformatics*, 22(5):623–625, Mar 2006.

[204] G. E. Tusnady, Z. Dosztanyi, and I. Simons. Tmdet: web server for detecting transmembrane regions of proteins by using their 3d coordinates. *Bioinformatics*, 21(7):1276–1277, Apr 2005.

[205] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, Sep 1997.

[206] D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. Dicuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, O. Khovayko, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, J. Ostell, K. D. Pruitt, G. D. Schuler, M. Shumway, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R. L. Tatusov, T. A. Tatusova, L. Wagner, and E. Yaschenko. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 36(Database issue):D13–D21, Jan 2008.

[207] T. Miyata, S. Miyazawa, and T. Yasunaga. Two types of amino acid substitutions in protein evolution. *J Mol Evol*, 12(3):219–236, Mar 1979.

[208] A. D. McLachlan. Tests for comparing related amino-acid sequences. cytochrome c and cytochrome c 551 . *J Mol Biol*, 61(2):409–424, Oct 1971.

[209] A. R. Ortiz, A. Kolinski, P. Rotkiewicz, B. Ilkowski, and J. Skolnick. Ab initio folding of proteins using restraints derived from evolutionary information. *Proteins*, Suppl 3:177–185, 1999.

[210] A. Fuchs, A. Kirschner, and D. Frishman. Prediction of helix-helix contacts and interacting helices in polytopic membrane proteins using neural networks. *Proteins*, 74(4):857–871, Mar 2009.

[211] P. Fariselli, O. Olmea, A. Valencia, and R. Casadio. Prediction of contact maps with neural networks and correlated mutations. *Protein Eng*, 14(11):835–843, Nov 2001.

[212] G. Pollastri and P. Baldi. Prediction of contact maps by giohmms and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics*, 18 Suppl 1:S62–S70, 2002.

[213] Y. Shao and C. Bystroff. Predicting interresidue contacts using templates and pathways. *Proteins*, 53 Suppl 6:497–502, 2003.

[214] I. Ezkurdia, O. Grana, J. M. G. Izarzugaza, and M. L. Tress. Assessment of domain boundary predictions and the prediction of intramolecular contacts in casp8. *Proteins*, Epub ahead of print, Jul 2009.

[215] Y. Zhang, A. Kolinski, and J. Skolnick. Touchstone ii: a new approach to ab initio protein structure prediction. *Biophys J*, 85(2):1145–1164, Aug 2003.

[216] S. Wu and Y. Zhang. Lomets: a local meta-threading-server for protein structure prediction. *Nucleic Acids Res*, 35(10):3375–3382, 2007.

[217] S. Wu and Y. Zhang. A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics*, 24(7):924–931, Apr 2008.

[218] A. Sali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, 234(3):779–815, Dec 1993.

[219] Y. Zhang and J. Skolnick. Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc Natl Acad Sci U S A*, 101(20):7594–7599, May 2004.

[220] W. Li, Y. Zhang, and J. Skolnick. Application of sparse nmr restraints to large-scale protein structure prediction. *Biophys J*, 87(2):1241–1248, Aug 2004.

[221] D. Latek and A. Kolinski. Contact prediction in protein modeling: scoring, folding and refinement of coarse-grained models. *BMC Struct Biol*, 8:36, 2008.

[222] U. Hobohm and C. Sander. Enlarged representative set of protein structures. *Protein Sci*, 3(3):522–524, Mar 1994.

[223] A. Andreeva, D. Howorth, J.-M. Chandonia, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin. Data growth and its impact on the scop database: new developments. *Nucleic Acids Res*, 36(Database issue):D419–D425, Jan 2008.

[224] K. Palczewski, T. Kumasaka, T. Hori, C. A. Behnke, H. Motoshima, B. A. Fox, I. L. Trong, D. C. Teller, T. Okada, R. E. Stenkamp, M. Yamamoto, and M. Miyano. Crystal structure of rhodopsin: A g protein-coupled receptor. *Science*, 289(5480):739–745, Aug 2000.

[225] R. M. MacCallum. Striped sheets and protein contact prediction. *Bioinformatics*, 20 Suppl 1:i224–i231, Aug 2004.

[226] L. Feng, H. Yan, Z. Wu, N. Yan, Z. Wang, P. D. Jeffrey, and Y. Shi. Structure of a site-2 protease family intramembrane metalloprotease. *Science*, 318(5856):1608–1612, Dec 2007.

[227] J. P. Morth, B. P. Pedersen, M. S. Toustrup-Jensen, T. L.-M. Sorensen, J. Petersen, J. P. Andersen, B. Vilsen, and P. Nissen. Crystal structure of the sodium-potassium pump. *Nature*, 450(7172):1043–1049, Dec 2007.

[228] B. P. Pedersen, M. J. Buch-Pedersen, J. P. Morth, M. G. Palmgren, and P. Nissen. Crystal structure of the plasma membrane proton pump. *Nature*, 450(7172):1111–1114, Dec 2007.

[229] J. U. Bowie. Helix-bundle membrane protein fold templates. *Protein Sci*, 8(12):2711–2719, Dec 1999.

[230] M. B. Swindells, C. A. Orengo, D. T. Jones, E. G. Hutchinson, and J. M. Thornton. Contemporary approaches to protein structure classification. *Bioessays*, 20(11):884–891, Nov 1998.

[231] Y. I. Wolf, N. V. Grishin, and E. V. Koonin. Estimating the number of protein folds and families from complete genome data. *J Mol Biol*, 299(4):897–905, Jun 2000.

[232] J. Arce, J. N. Sturgis, and J.-P. Duneau. Dissecting membrane protein architecture: An annotation of structural complexity. *Biopolymers*, 91(10):815–829, Oct 2009.

[233] O. Tastan, J. Klein-Seetharaman, and H. Meirovitch. The effect of loops on the structural organization of alpha-helical membrane proteins. *Biophys J*, 96(6):2299–2312, Mar 2009.

[234] B. Rost and C. Sander. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, 19(1):55–72, May 1994.

[235] L. Kall, A. Krogh, and E. L. L. Sonnhammer. An hmm posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics*, 21 Suppl 1:i251–i257, Jun 2005.

[236] C. Yeats, J. Lees, A. Reid, P. Kellam, N. Martin, X. Liu, and C. Orengo. Gene3d: comprehensive structural and functional annotation of genomes. *Nucleic Acids Res*, 36(Database issue):D414–D418, Jan 2008.

[237] A. L. Cuff, I. Sillitoe, T. Lewis, O. C. Redfern, R. Garratt, J. Thornton, and C. A. Orengo. The cath classification revisited–architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res*, 37(Database issue):D310–D314, Jan 2009.

[238] S. J. Opella and F. M. Marassi. Structure determination of membrane proteins by nmr spectroscopy. *Chem Rev*, 104(8):3587–3606, Aug 2004.

[239] M. Caffrey. Crystallizing membrane proteins for structure determination: use of lipidic mesophases. *Annu Rev Biophys*, 38:29–51, 2009.

[240] I. N. Shindyalov and P. E. Bourne. An alternative view of protein fold space. *Proteins*, 38(3):247–260, Feb 2000.

[241] R. Kolodny, D. Petrey, and B. Honig. Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction. *Curr Opin Struct Biol*, 16(3):393–398, Jun 2006.

[242] D. Petrey and B. Honig. Is protein classification necessary? toward alternative approaches to function annotation. *Curr Opin Struct Biol*, 19(3):363–368, Jun 2009.

[243] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–540, Apr 1995.

[244] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. Cath–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, Aug 1997.

[245] S. Neumann, A. Fuchs, A. Mulkidjanian, and D. Frishman. Current status of membrane protein structure classification. *Proteins*, submitted, 2009.

[246] L. Holm and C. Sander. Touring protein fold space with dali/fssp. *Nucleic Acids Res*, 26(1):316–319, Jan 1998.

[247] S. Dietmann, J. Park, C. Notredame, A. Heger, M. Lappe, and L. Holm. A fully automatic evolutionary classification of protein folds: Dali domain dictionary version 3. *Nucleic Acids Res*, 29(1):55–57, Jan 2001.

[248] P. Rogen and B. Fain. Automatic classification of protein structure by using gauss integrals. *Proc Natl Acad Sci U S A*, 100(1):119–124, Jan 2003.

[249] V. Sam, C.-H. Tai, J. Garnier, J.-F. Gibrat, B. Lee, and P. J. Munson. Towards an automatic classification of protein structural domains based on structural similarity. *BMC Bioinformatics*, 9:74, 2008.

[250] C. Hadley and D. T. Jones. A systematic comparison of protein structure classifications: Scop, cath and fssp. *Structure*, 7(9):1099–1112, Sep 1999.

[251] R. Day, D. A. C. Beck, R. S. Armen, and V. Daggett. A consensus view of fold space: combining scop, cath, and the dali domain dictionary. *Protein Sci*, 12(10):2150–2160, Oct 2003.

[252] A. Pascual-Garcia, D. Abia, A. R. Ortiz, and U. Bastolla. Cross-over between discrete and continuous protein structure space: insights into automatic classification and networks of protein structures. *PLoS Comput Biol*, 5(3):e1000331, Mar 2009.

[253] A. S. Yang and B. Honig. An integrated approach to the analysis and modeling of protein sequences and structures. i. protein structural alignment and a quantitative measure for protein structural distance. *J Mol Biol*, 301(3):665–678, Aug 2000.

[254] N. V. Grishin. Fold change in evolution of protein structures. *J Struct Biol*, 134(2-3):167–185, 2001.

[255] A. N. Lupas, C. P. Ponting, and R. B. Russell. On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J Struct Biol*, 134(2-3):191–203, 2001.

[256] I. Friedberg and A. Godzik. Connecting the protein structure universe by using sparse recurring fragments. *Structure*, 13(8):1213–1224, Aug 2005.

[257] A. Harrison, F. Pearl, R. Mott, J. Thornton, and C. Orengo. Quantifying the similarities within fold space. *J Mol Biol*, 323(5):909–926, Nov 2002.

[258] A. Cuff, O. C. Redfern, L. Greene, I. Sillitoe, T. Lewis, M. Dibley, A. Reid, F. Pearl, T. Dallman, A. Todd, R. Garratt, J. Thornton, and C. Orengo. The cath hierarchy revisited-structural divergence in domain superfamilies and the continuity of fold space. *Structure*, 17(8):1051–1062, Aug 2009.

[259] D. Kihara and J. Skolnick. The pdb is a covering set of small protein structures. *J Mol Biol*, 334(4):793–802, Dec 2003.

[260] L. Holm, S. Kääriäinen, P. Rosenström, and A. Schenkel. Searching protein structure databases with dalilite v.3. *Bioinformatics*, 24(23):2780–2781, Dec 2008.

[261] Y. Liu, M. Gerstein, and D. M. Engelman. Transmembrane protein domains rarely use covalent domain recombination as an evolutionary mechanism. *Proc Natl Acad Sci U S A*, 101(10):3495–3497, Mar 2004.

[262] A. J. Enright, S. V. Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*, 30(7):1575–1584, Apr 2002.

[263] G. M. Soriano, M. V. Ponamarev, C. J. Carrell, D. Xia, J. L. Smith, and W. A. Cramer. Comparison of the cytochrome bc1 complex with the anticipated structure of the cytochrome b6f complex: Le plus ça change le plus c'est la même chose. *J Bioenerg Biomembr*, 31(3):201–213, Jun 1999.

[264] S. R. Presnell and F. E. Cohen. Topological distribution of four-alpha-helix bundles. *Proc Natl Acad Sci U S A*, 86(17):6592–6596, Sep 1989.

[265] N. L. Harris, S. R. Presnell, and F. E. Cohen. Four helix bundle diversity in globular proteins. *J Mol Biol*, 236(5):1356–1368, Mar 1994.

[266] Y. Zhang, M. E. Devries, and J. Skolnick. Structure modeling of all identified g protein-coupled receptors in the human genome. *PLoS Comput Biol*, 2(2):e13, Feb 2006.

[267] A. Y. Mulkidjanian, M. Y. Galperin, and E. V. Koonin. Co-evolution of primordial membranes and membrane proteins. *Trends Biochem Sci*, 34(4):206–215, Apr 2009.

[268] J. Abramson, I. Smirnova, V. Kasho, G. Verner, H. R. Kaback, and S. Iwata. Structure and mechanism of the lactose permease of escherichia coli. *Science*, 301(5633):610–615, Aug 2003.

[269] R. Dutzler, E. B. Campbell, and R. MacKinnon. Gating the selectivity filter in clc chloride channels. *Science*, 300(5616):108–112, Apr 2003.

[270] Y. Yin, X. He, P. Szewczyk, T. Nguyen, and G. Chang. Structure of the multidrug transporter emrd from escherichia coli. *Science*, 312(5774):741–744, May 2006.

9
Appendix

Table 9.1: High-affinity transmembrane domains identified with the ToxR/POSSYCAT system from a combinatorial sequence library of transmembrane domains containing alanine residues at all non-interface positions. Reported is the relative β-gal activity of each sequence with respect to a canonical leucine zipper. All sequences with a relative β-gal activity >1.5 were classified as high-affine.

Sequence	Signal	Sequence	Signal
.L..GI.VG..GT..V	1.6	.A..AI.GL..GI..G	2.5
.V..CA.CG..GW..T	1.6	.V..FF.GI..SC..T	2.5
.C..WC.CG..FM..G	1.6	.V..FI.GM..GG..V	2.5
.C..CG.VT..WF..A	1.7	.W..FA.GW..GI..A	2.6
.C..FC.GW..GS..M	1.7	.A..FV.GV..GC..I	2.6
.I..AA.GG..FG..I	1.7	.V..FS.GF..AS..F	2.7
.G..CF.GW..GM..S	1.8	.A..FM.GF..GS..W	2.7
.C..CS.VG..WM..C	1.8	.G..FA.GL..GM..A	2.7
.V..FS.MF..AG..T	1.8	.F..GT.FG..TV..L	2.7
.M..TW.SG..WG..V	1.8	.C..VA.LS..VG..T	2.7
.S..WF.FG..TF..A	1.8	.A..FI.GC..GF..S	2.7
.I..CM.GA..GA..S	1.9	.T..IV.SF..GM..G	2.9
.I..FC.GA..AG..W	1.9	.A..FW.GF..GA..T	2.9
.I..GI.CG..IS..I	1.9	.V..CM.AS..VS..M	3.0
.V..VS.TA..IF..T	2.0	.I..FV.GV..GV..G	3.0
.W..AA.MF..GF..R	2.0	.V..FV.GV..GM..T	3.0
.W..GF.CG..WS..S	2.0	.M..CV.MS..VS..T	3.1
.V..LA.VF..GV..G	2.1	.I..FC.GF..GT..F	3.1
.A..FV.GC..GG..F	2.1	.G..FG.VF..GV..G	3.1
.F..SL.GC..GC..T	2.1	.V..FA.GL..GF..C	3.1
.A..CF.GG..CG..F	2.1	.I..FF.GM..GV..G	3.1
.S..AS.VG..FG..M	2.2	.C..FS.GF..GM..M	3.2
.I..CI.VG..GG..S	2.2	.G..FL.GA..GA..F	3.3
.F..GI.CG..MG..T	2.2	.W..VV.VS..TS..T	3.3
.A..AF.LG..IT..W	2.3	.W..MG.WS..IS..T	3.4
.F..GC.CG..MC..A	2.3	.I..CF.VG..GG..S	3.4
.L..WA.GT..GG..I	2.3	.V..WW.AT..SS..C	3.4
.L..MG.WA..WG..G	2.4	.V..WW.ST..TS..V	3.5
.G..FL.GC..GC..A	2.4	.I..CI.VS..TS..T	3.7
.M..IW.SG..WG..V	2.4	.V..WW.AT..TS..C	3.8

Table 9.2: High-affinity transmembrane domains identified with the ToxR/POSSYCAT system from a combinatorial sequence library of transmembrane domains containing leucine residues at all non-interface positions. Reported is the relative β-gal activity of each sequence with respect to a canonical leucine zipper. All sequences with a relative β-gal activity >1.0 were classified as high-affine.

Sequence	Signal	Sequence	Signal
.I..GY.LY..VA..A	1.0	.A..QY.VV..ET..L	2.5
.T..FH.VL..TW..G	1.2	.G..DR.DM..LG..L	2.5
.T..AH.SL..WA..A	1.3	.I..NS.TS..TG..L	2.5
.I..GW.AY..NA..W	1.3	.R..ER.TI..TG..G	2.5
.S..GN.YT..TG..I	1.6	.R..VN.TM..VG..L	2.6
.T..GH.VA..TH..V	1.6	.G..ST.AS..KA..V	2.7
.V..AH.TY..CW..W	1.6	.C..GH.SS..AG..L	2.8
.G..GH.IL..IH..V	1.7	.R..TR.EA..GG..I	2.8
.T..GH.AI..EF..I	1.7	.D..DK.DW..AG..L	2.9
.A..AG.AG..VG..S	1.8	.L..TH.LV..SG..C	3.0
.S..GH.SF..WG..T	1.8	.G..TH.SA..IG..T	3.0
.V..ER.AW..NG..M	2.0	.W..CY.VG..SG..T	3.3
.W..CH.TG..LG..A	2.1	.K..YF.TG..AG..S	3.3
.G..AN.GT..TG..L	2.1	.G..SV.SG..GA..M	3.5
.C..KD.ML..GG..I	2.1	.F..VT.AD..AN..S	3.6
.F..NH.SG..FG..L	2.1	.G..SH.SS..GG..L	3.6
.V..TN.GC..FG..I	2.1	.R..HT.DG..LG..I	3.6
.V..LH.AL..CN..T	2.2	.R..DR.YD..LG..I	3.6
.T..GF.GG..GE..G	2.3	.V..NF.AG..GG..G	3.7
.V..LR.AL..CY..S	2.3	.G..DR.CY..VG..G	3.7
.T..LH.CY..IM..I	2.3	.S..IY.GG..CG..L	3.7
.R..TH.VA..GG..S	2.3	.I..PC.GS..GG..Q	3.8
.I..GH.AI..LN..T	2.3	.T..TH.SC..GG..T	3.8
.Q..GH.VS..AG..W	2.3	.Y..AT.SL..NC..M	4.0
.V..SG.GS..GN..P	2.4	.N..LF.SG..TG..G	4.1
.R..ED.EI..AG..A	2.5	.A..SR.EG..HG..L	4.2

Table 9.3: PDBTM non-redundant dataset of membrane protein structures (MP_62) used for the prediction of co-evolving residues and the development of TMHcon.

PDB	Chains	Description	TMS[a]	Res [Å][b]	Species
1AIG	L	Photosynthetic reaction center	5	2.6	Rhodobacter sphaeroides
1BCC	C	Cytochrome bc1 complex	8	3.2	Gallus gallus
1EYS	M	Photosynthetic reaction center	5	2.2	Thermochromatium tepidum
1FFT	A; C	Ubiquinol oxidase	12; 5	3.5	Escherichia coli
1FX8	A	Glycerol facilitator (Glpf)	6	2.2	Escherichia coli
1JB0	A; L	Photosystem I	11; 3	2.5	Synechococcus elongatus
1KQF	C	Formate dehydrogenase	4	1.6	Escherichia coli
1L7V	A	BtuCD vitamin B12 transporter	10	3.2	Escherichia coli
1M0K	A	Bacteriorhodopsin	7	1.4	Halobacterium salinarium
1NEK	C;D	Succinate dehydrogenase	3; 3	2.6	Escherichia coli
1ORQ	C	Potassium channel	4	3.2	Mus muculus
1PW4	A	Glycerol-3-phosphate transporter	12	3.3	Escherichia coli
1Q16	C	Nitrate reductase A (NarGHI)	5	1.9	Escherichia coli
1QLE	C	Cytochrome c oxidase	7	3.0	Paracoccus denitrificans
1RH5	A	Protein conducting channel	10	3.2	Methanococcus jannaschii
1U19	A	Rhodopsin	7	2.2	Bos taurus
1VF5	A;B	Cytochrome b6f complex	4; 3	3.0	Mastigocladus laminosus
1XIO	A	Sensory rhodopsin	7	2.0	Anabaena sp.
1XME	A	Cytochrome ba3 oxidase	13	2.3	Thermus thermophilus
1YEW	B,C	Methane monooxygenase	7; 4	2.8	Methylococcus capsulatus
1ZCD	A	Na(+)/H(+) antiporter NhaA	12	3.5	Escherichia coli
2A65	A	Na(+):neurotransmitter symporter	12	1.7	Aquifex aeolicus vf5
2A79	B	Shaker Kv1.2 potassium channel	4	2.9	Rattus norvegicus
2AGV	A	Calcium ATPase 1	10	2.4	Oryctolagus cuniculus
2AXT	A;B;C;D	Photosystem II	5; 6; 6; 5	3.0	Thermosynechococcus elongatus
2B2F	A	Ammonium transporter Amt-1	11	1.7	Archaeoglobus fulgidus
2B76	C;D	Quinol fumarate reductase FrdA	3; 3	3.3	Escherichia coli
2BG9	A	Nicotinic Acetylcholine Receptor	4	4.0	Torpedo marmorata
2BHW	A	Light-harvesting complex II	3	2.5	Pisum sativum
2BL2	A	V-type ATPase	4	2.2	Enterococcus hirae
2BS2	C	Quinol-fumarate reductase	5	1.8	Wolinella succinogenes
2C3E	A	Mitochondrial ADP-ATP carrier	6	2.8	Bos taurus
2CFP	A	Lactose permease	12	3.3	Escherichia coli
2EVU	A	Aquaporin aqpM	6	2.3	Methanobacterium thermoautotrophicum
2EXW	A	H(+)/Cl(-) exchange transporter	10	3.2	Escherichia coli
2F93	A	Sensory rhodopsin II	7	2.0	Natronomonas pharaonis
2FBW	C; D	Succinate dehydrogenase	3; 3	2.1	Gallus gallus
2FYN	A	Cytochrome bc1 complex	8	3.2	Rhodobacter sphaeroides
2GFP	A	Multidrug transporter EmrD	12	3.5	Escherichia coli
2GIF	A	Acriflavine resistance protein B	12	2.9	Escherichia coli
2GSM	A	Cytochrome c oxidase	12	2.0	Rhodobacter sphaeroides
2HI7	B	DsbB-DsbA-ubiquinone complex	4	3.7	Escherichia coli
2HYD	A	Multidrug ABC transporter SAV1866	6	3.0	Staphylococcus aureus
2IC8	A	GlpG	6	2.1	Escherichia coli
2JAF	A	Halorhodopsin	7	1.7	Halobacterium salinarium
2NMR	A	Ammonia channel	11	2.1	Escherichia coli
2NR9	A	Rhomboid peptidase GlpG	6	2.2	Haemophilus influenzae
2NWL	A	Aspartate transporter GltPh	8	3.0	Pyrococcus horikoshii
2O9D	A	Aquaporin Z	6	2.3	Escherichia coli
2OAU	A	Mechanosensitive channel MscS	3	3.7	Escherichia coli
2ONK	C	ABC transporter ModBC	6	3.1	Archaeoglobus fulgidus
2UUH	A	Leukotriene C4 Synthase	4	2.2	Homo sapiens

[a] TMS: number of transmembrane segments experimentally determined from the PDB structure.
[b] Res: resolution of 3D structure.

Table 9.4: CAMPS non-redundant dataset of membrane protein structures (MP_CAMPS) used for the prediction of consensus helix interaction graphs.

PDB	Chain	Description	Res [Å][a]	TMS_{pred}[b]	TMS_{exp}[c]	HomSeq[d]
1FFT	A	Ubiquinol oxidase	3.5	14	12	40
1FFT	C	Ubiquinol oxidase	3.5	5	5	40
1J4N	A	Aquaporin 1	2.2	6	6	40
1JB0	A	Photosystem I P700 Apoprotein A1	2.5	11	11	15
1M0K	A	Bacteriorhodopsin	1.4	7	7	39
1NEN	C	Succinate dehydrogenase cytochrome b-556 subunit	2.9	3	3	40
1OKC	A	ADP/ATP carrier protein	2.2	3	6	1
1ORQ	C	Potassium channel	3.2	4	4	20
1PW4	A	Glycerol-3-phosphate transporter	3.3	12	12	40
1Q16	C	Nitrate reductase A gamma chain	1.9	5	5	40
1SQX	C	Cytochrome b	2.6	9	8	3
1U19	A	Rhodopsin	2.2	7	7	40
1VF5	A	Cytochrome b6	3.0	4	4	19
1ZCD	A	Na(+)/H(+) antiporter 1	3.5	11	12	16
2A65	A	Na(+):neurotransmitter symporter	1.7	12	12	40
2AKI	C	Preprotein translocase secE subunit	4.0	3	3	44
2AXT	A	Photosystem Q(B) protein	3.0	7	5	14
2AXT	C	Photosystem II CP43 protein	3.0	7	6	11
2EXW	A	H(+)/Cl(-) exchange transporter clcA	3.2	10	10	40
2GFP	A	Multidrug transporter EmrD	3.5	11	12	40
2GIF	A	Acriflavine resistance protein B	2.9	12	12	40
2HI7	B	Disulfide bond formation protein B	3.7	4	4	40
2IC8	A	GlpG	2.1	6	6	40
2NQ2	B	ABC transporter permease protein HI1471	2.4	8	10	40
2NWL	A	Aspartate transporter GltPh	3.0	8	8	40
2Q7R	C	Arachidonate 5-lipoxygenase-activating protein	4.0	3	4	2
2QFI	A	Ferrous-iron efflux pump fieF	3.8	6	6	40
2R6G	F	Maltose transport system permease protein malF	2.8	8	8	40
2R6G	G	Maltose transport system permease protein malG	2.8	6	6	40
2YVX	B	Mg2+ transporter MgtE	3.5	5	5	40
2ZBG	A	Sarcoplasmic/endoplasmic reticulum calcium ATPase 1	2.6	8	10	40
3B8E	C	Sodium/potassium-transporting ATPase subunit alpha-1	3.5	8	10	40
3B9W	A	Ammonium transporter	1.3	11	11	40
3BEH	D	Mll3241 protein	3.1	6	6	40

[a] Res: resolution of 3D structure.
[b] TMS_{pred}: number of transmembrane segments predicted with Phobius.
[c] TMS_{exp}: number of transmembrane segments determined from the PDB structure.
[d] HomSeq: number of homologous sequences used for the construction of consensus helix interaction graphs.

Table 9.5: SCOP folds containing membrane proteins with at least two transmembrane helices.

Fold	Description	Domains[a]	Superfamilies	Families	Min (TMS)[b]	Max (TMS)[b]
f.13	Family A G protein-coupled receptor-like	6	1	2	2	7
f.14	Voltage-gated potassium channels	5	1	1	1	4
f.16	Gated mechanosensitive channel	1	1	1	2	2
f.17	Transmembrane helix hairpin	5	3	3	2	2
f.19	Aquaporin-like	4	1	1	6	6
f.20	Clc chloride channel	1	1	1	10	10
f.21	Heme-binding four-helical bundle	9	3	5	3	5
f.22	ABC transporter involved in vitamin B12 uptake	1	1	1	10	10
f.24	Cytochrome c oxidase subunit I-like	4	1	1	12	13
f.25	Cytochrome c oxidase subunit III-like	3	1	1	5	7
f.26	Bacterial photosystem II reaction centre L and M subunits	2	1	1	5	5
f.29	Photosystem I subunits PsaA/PsaB	2	1	1	11	11
f.30	Photosystem I reaction center subunit X, PsaK	1	1	1	2	2
f.31	Photosystem I reaction center subunit XI, PsaL	1	1	1	3	3
f.32	Domain/subunit of cytochrome bc1 complex (Ubiquinol-cytochrome c reductase)	2	1	1	3	3
f.33	Calcium ATPase, transmembrane domain M	1	1	1	10	10
f.34	Mechanosensitive channel protein MscS (YggB) transmembrane region	1	1	1	3	3
f.35	Multidrug efflux transporter AcrB transmembrane domain	1	1	1	6	6
f.36	Neurotransmitter-gated ion-channel transmembrane pore	4	1	1	4	4
f.37	ABC transporter transmembrane region	1	1	1	6	6
f.38	MFS general substrate transporter	2	1	2	12	12
f.41	Preprotein translocase SecY subunit	1	1	1	10	10
f.42	Mitochondrial carrier	1	1	1	6	6
f.43	Chlorophyll a-b binding protein	1	1	1	3	3
f.44	Ammonium transporter	1	1	1	11	11
f.49	Proton glutamate symport protein	1	1	1	8	8
f.51	Rhomboid-like	2	1	1	6	6

[a] Domains: number of distinct domains according to SCOP unique identifiers (sunid) for protein domains.
[b] TMS: number of transmembrane segments as defined by PDBTM.

Table 9.6: CATH folds containing membrane proteins with at least two transmembrane helices.

Fold	Description	Domains[a]	Superfamilies	Min (TMS)[b]	Max (TMS)[b]
1.10.287	Helix hairpins	14	6	2	3
1.10.3080	Clc chloride channel	2	1	10	10
1.20.20	F1F0 ATP synthase	3	1	2	2
1.20.85	Photosynthetic reaction center, subunit M; domain 1	13	1	2	3
1.20.120	Four helix bundle (hemerythrin (Met), subunit A)	9	3	4	5
1.20.210	Cytochrome C oxidase; chain A	5	1	12	13
1.20.810	Cytochrome Bc1 complex; chain C	7	1	4	8
1.20.860	Alpha-t-alpha	1	1	2	2
1.20.950	Fumarate reductase cytochrome B subunit	2	2	4	5
1.20.1050	Glutathione S-transferase Yfyf (class pi); chain A, domain 2	1	1	4	4
1.20.1070	Rhodopsin 7-helix transmembrane proteins	9	1	7	7
1.20.1080	Glycerol uptake facilitator protein	8	1	6	6
1.20.1110	Calcium-transporting ATPase transmembrane domain	1	1	9	10
1.20.1130	Photosystem I p700 chlorophyll A apoprotein A1	2	1	11	11
1.20.1240	Photosystem 1 reaction centre subunit Xi; chain L	1	1	3	3
1.20.1300	Three helical TM bundles of succinate and fumarate reductases	3	1	3	3
1.20.1450	Particulate methane monooxygenase; chain B	1	1	7	7

[a] Domains: number of distinct domains according to a representative set of CATH domains at 95% sequence identity.
[b] TMS: number of transmembrane segments as defined by PDBTM.

Table 9.7: Non-redundant dataset MP_SCOP_CATH of membrane protein structures present in both SCOP and CATH.

PDB	Chain	CATH	SCOP
1AIG	L	1.20.85.10, 1.20.85.10	f.26.1.1
1AR1	B	1.10.287.90	f.17.2.1
1C0V	A	1.20.20.10	f.17.1.1
1E12	A	1.20.1070.10	f.13.1.1
1EHK	A	1.20.210.10	f.24.1.1
1EYS	M	1.20.85.10, 1.20.85.10	f.26.1.1
1EZV	C	1.20.810.10	f.21.1.2, f.32.1.1
1FFT	A	1.20.210.10	f.24.1.1
1FFT	B	1.10.287.90	f.17.2.1
1FFT	C	1.20.120.80	f.25.1.1
1FX8	A	1.20.1080.10	f.19.1.1
1GZM	A	1.20.1070.10	f.13.1.2
1IWO	A	1.20.1110.10	f.33.1.1
1J4N	A	1.20.1080.10	f.19.1.1
1JB0	A	1.20.1130.10	f.29.1.1
1JB0	B	1.20.1130.10	f.29.1.1
1JB0	K	1.20.860.20	f.30.1.1
1JB0	L	1.20.1240.10	f.31.1.1
1KQF	C	1.20.950.20	f.21.1.1
1M0K	A	1.20.1070.10	f.13.1.1
1M56	A	1.20.210.10	f.24.1.1
1OED	A	1.20.120.370	f.36.1.1
1OED	B	1.20.120.370	f.36.1.1
1OED	C	1.20.120.370	f.36.1.1
1OED	E	1.20.120.370	f.36.1.1
1ORS	C	1.20.120.350	f.14.1.1
1QLE	C	1.10.287.70, 1.20.120.80	f.25.1.1
1UAZ	A	1.20.1070.10	f.13.1.1
1XIO	A	1.20.1070.10	f.13.1.1
2ABM	A	1.20.1080.10	f.19.1.1
2ACZ	C	1.20.1300.10	f.21.2.2
2ATK	C	1.10.287.70	f.14.1.1
2B6O	A	1.20.1080.10	f.19.1.1
2B76	C	1.20.1300.10	f.21.2.2
2B76	D	1.20.1300.10	f.21.2.2
2BS2	C	1.20.950.10	f.21.2.1
2D2C	A	1.20.810.10	f.21.1.2
2DYR	A	1.20.210.10	f.24.1.1
2DYR	B	1.10.287.90	f.17.2.1
2DYR	C	1.10.287.70, 1.20.120.80	f.25.1.1
2EXW	A	1.10.3080.10	f.20.1.1
2F93	A	1.20.1070.10	f.13.1.1

List of publications

Publications included in this thesis:

- S. Unterreitmeier, **A. Fuchs**, T. Schäffler, R.G. Heym, D. Frishman and D. Langosch.
  Phenylalanine promotes interaction of transmembrane domains via GxxxG motifs.
  J Mol Biol. 2007, **374**:705-18.

- **A. Fuchs**, A.J. Martin-Galiano, M. Kalman, S. Fleishman, N. Ben-Tal and D. Frishman.
  Co-evolving residues in membrane proteins.
  Bioinformatics 2007, **23**:3312-9.

- **A. Fuchs***, A. Kirschner* and D. Frishman
  Prediction of helix-helix contacts and interacting helices in polytopic membrane proteins using neural networks.
  Proteins 2009, **74**:857-71.

- J.R. Herrman, J.C. Panitz, S. Unterreitmeier, **A. Fuchs**, D. Frishman and D. Langosch.
  Complex patterns of histidin, hydroxylated amino acids and the GxxxG motif mediate high-affinity transmembrane domain interactions.
  J Mol Biol. 2009, **385**:912-23

- S. Neumann*, **A. Fuchs***, A. Mulkidjanian and D. Frishman.
  Current status of membrane protein structure classification.
  Proteins 2009, *submitted.*

- J.R. Herrmann, **A. Fuchs**, J.C. Panitz, T. Eckert, S. Unterreitmeier, D. Frishman and D. Langosch
  Ionic interactions promote transmembrane helix-helix association depending on sequence context.
  J Mol Biol. 2009, *submitted.*

Publications not included in this thesis:

- B. Eyüboglu, K. Pfister, G. Haberer, D. Chevalier, **A. Fuchs**, K.F. Mayer and K. Schneitz.
  Molecular characterisation of the STRUBBELIG-RECEPTOR FAMILY of genes encoding putative leucine-rich repeat receptor-like kinases in Arabidopsis thaliana.
  BMC Plant Biol 2007, **7**:16.

- A. Kowarsch, **A. Fuchs**, D. Frishman and P. Pagel.
  Correlated mutations: a hallmark of phenotypic amino acid substitutions
  PLoS Comput Biol. 2009, *submitted.*

- M. Gersoni*, **A. Fuchs***, Y. Fridman, M. Corral-Debrinski, A. Aharoni, D. Frishman and D. Mishmar.
  Subunit co-evolution predicts direct interactions within the membrane-bound oxidative phosphorylation complex I.
  In preparation.

* These authors contributed equally.