TECHNISCHE UNIVERSITÄT MÜNCHEN

Fachgebiet für Bioinformatik

Structure-function relationships in membrane protein families

Sindy Neumann

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. Luksch

Prüfer der Dissertation:

1. Univ.-Prof. Dr. D. Frischmann
2. Univ.-Prof. Dr. D. Langosch

Die Dissertation wurde am 07.02.2012 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 27.04.2012 angenommen.

Abstract

α-helical membrane proteins account for 20-30% of a typical genome and are crucial for such processes as transport, cell division and metabolism. They are associated with various diseases and serve as targets for about 60% of approved drugs. Paradoxically, only about 2% of the protein structures in the Protein Data Bank (PDB) account for membrane proteins. Therefore, structural bioinformatics of membrane proteins is still in its infancy, and the picture of their fold space is only beginning to emerge.

Structure classification is a valuable means for the investigation of structure-function relationships and for studies on the structural diversity of the protein sequence space. This thesis is focused on the structural classification of α-helical membrane proteins.

In the first part, the structural classification of membrane proteins in SCOP and CATH is addressed by a comparative analysis. Using a dataset of 63 α-helical membrane protein structures, a number of differently classified proteins both regarding their domain and fold assignment was observed. The majority of all discrepancies affect single transmembrane helix (TMH), two helix hairpin, and four helix bundle proteins, while domains with more than five TMHs are mostly classified consistently between SCOP and CATH.

In the second part, a hierarchical classification approach specifically tailored to membrane proteins (named CAMPS) is described. In contrast to SCOP and CATH, it is not based on three-dimensional structures, but on sequence similarity and topology conservation (in terms of the number of TMHs and loop lengths) allowing for a large-scale classification of membrane proteins. Using high-order hidden Markov models 1,353 structurally homogeneous clusters (SC-clusters) roughly corresponding to membrane protein folds were found. Only 53 SC-clusters are associated with experimentally determined structures, and for these clusters CAMPS is in reasonable agreement with SCOP and CATH.

In the third and last part, a further development of the classification approach is proposed. In addition to sequence similarity, the number of TMHs and loop length patterns, it also considers helix-helix interactions and allows to identify SC-clusters that are likely to represent the same fold. Using predicted consensus helix architectures a selected set of 431 SC-clusters was joined into 151 superior clusters (termed MCL clusters). By comparing the SC- and MCL clusters to Pfam clans, it could be shown that the sensitivity increased by 30.7%.

Zusammenfassung

α-helikale Membranproteine machen 20-30% eines Genoms aus und sind essentiell für Prozesse wie z.B. Transport, Zellteilung und Metabolismus. Sie sind mit verschiedenen Krankheiten assoziiert und dienen als Targets für circa 60% der zugelassenen Medikamente. Seltsamerweise sind Membranproteine nur zu etwa 2% in der Strukturdatenbank PDB vertreten. Daher ist die Strukturbioinformatik von Membranproteinen noch ganz am Anfang und der Faltungsraum nur ganz wenig erforscht.

Die Strukturklassifikation ist ein nützliches Instrument für die Untersuchung von Struktur-Funktions-Beziehungen und der Analyse der strukturellen Diversität von Proteinsequenzen. Diese Dissertation konzentriert sich auf die strukturelle Klassifikation von α-helikalen Membranproteinen.

Im ersten Teil wird die Strukturklassifikation von Membranproteinen in SCOP und CATH mittels einer vergleichenden Analyse adressiert. Unter Verwendung eines Datensatzes von 63 α-helikalen Membranproteinen wurden mehrere unterschiedlich klassifizierte Proteine entdeckt, sowohl hinsichtlich ihrer Domänenzuordnung, als auch ihrer Faltungszuweisung. Die Mehrheit aller Diskrepanzen betrifft Membranproteine mit ein, zwei oder vier Transmembranhelices (TMH), während Domänen mit mehr als fünf TMHs meistens konsistent in SCOP und CATH klassifiziert sind.

Im zweiten Teil wird ein hierarchischer Klassifikationsansatz (CAMPS) beschrieben, der speziell auf Membranproteine angepasst ist. Im Gegensatz zu SCOP und CATH basiert dieser nicht auf drei-dimensionalen Strukturen, sondern auf Sequenzähnlichkeit und Topologiekonservierung (hinsichtlich der Anzahl der TMHs und der Loop-Längen) und erlaubt somit eine umfangreiche Klassifikation von Membranproteinen. Mit Hilfe von Hidden Markov Modellen höherer Ordnung wurden 1,353 strukturell homogene Cluster (SC-cluster) gefunden, die in etwa Membranproteinfaltungen entsprechen. Nur 53 SC-cluster sind mit einer experimentell bestimmten Struktur assoziiert und für diese stimmt CAMPS im Wesentlichen mit SCOP und CATH überein.

Im dritten und letzten Teil wird eine Weiterentwicklung des Klassifikationsansatzes vorgestellt. Zusätzlich zur Sequenzähnlichkeit, der Anzahl an TMHs und der Loop-Längen, berücksichtigt diese auch Helix-Helix Interaktionen und ermöglicht es SC-cluster zu identifizieren die sehr wahrscheinlich die gleiche Faltung repräsentieren. Unter Verwendung von vorhergesagten Consensus-Helix-Architekturen wurde ein ausgewähltes Set von 431 SC-clustern zu 151 übergeordneten Clustern (genannt MCL cluster) zusam-

mengefasst. Durch den Vergleich der SC- und MCL Cluster mit Pfam Clans konnte gezeigt werden, dass die Sensitivität um 30.7% gesteigert werden konnte.

Acknowledgments

Although only my name appears on the cover of this thesis, a number of people have contributed in various ways to the completion of this work. Accordingly, I would like to thank those people who made it possible.

In the first place, I want to express my sincere gratitude to my supervisor »Prof. Dr. Dmitrij Frishman« for his supervision, advice and guidance.

I am very grateful to »Prof. Dr. Dieter Langosch« for the many inspiring meetings regarding the SNARE project, for giving me the opportunity to learn about membrane proteins in his lab and for reviewing this thesis.

I also thank »Prof. Dr. Harald Luksch« for agreeing to participate in my dissertation committee.

Furthermore, my very special thanks go to my colleagues and co-workers:

»Angelika Fuchs and Nadia Latif« : I will never be able to express my gratitude by words. This work would not have been possible without their help and support. I learned so much from them and I hope that we will retain our friendship for a long time.

»Antonio Martin-Galiano« : for his constant advice on membrane protein bioinformatics and sharing his experience in this field.

»Holger Hartmann« : for the contribution to the CAMPS database through his excellent work on meta-models in the context of his master thesis.

»Roman Sutormin« : for the implementation of some nice additional features for the CAMPS website.

»Thomas Rattei« : for doing a perfect job in maintaining the computer network and his assistance in using the SIMAP database.

»Philipp Pagel« : for the valuable scientific and statistical advice. I very much enjoyed his enthusiasm for science and computers.

»Qibin Luo and Sheng Zhao« : for their warm-heartedness. I will never forget our lovely sushi parties and I truly treasure the time we spent together.

»Jan Kirrbach and Oxana Pester« : for their friendship and the collective laughter. They always found encouraging words for me whenever necessary.

»Frauke, Leonie and Claudia« : for always having an open ear and their help in private matters. Special thanks go to Leonie for proofreading my thesis.

»All my colleagues in Weihenstephan« (in particular Patrick Tischler, Roland Arnold,

Andreas Kirschner, Andre Jehl, Erik Granseth): for creating a perfect working atmosphere and having a good time together.

Last, and most importantly, I wish to thank »my family«. Words alone cannot express what I owe them. Their patient love and support enabled me to complete this work. Thank you!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> "So far, we have only scratched the surface of the world of membrane proteins..."
>
> *(Gunnar von Heijne)*

He is absolutely right! Gunnar von Heijne is one of the leading scientists in membrane protein research. And I wholeheartedly agree with his opinion that we have just reached the tip of the iceberg. The most important cause is that only few membrane protein structures are currently available mainly due to their intrinsic difficulties in experimental determination. Therefore, sequence comparison and structure prediction remain the main tools for investigating membrane protein families.

This thesis copes with the structural classification of membrane proteins in order to explore structure-function relationships. Thus, the following introduction starts with a review on current knowledge about membrane protein research. Similarly, different approaches to classify proteins in general are summarized. Finally, the manifold relations between sequence, structure and function are introduced.

1.1 Membrane proteins

Membrane proteins are crucial to all living organisms because of their key roles in controlling the processes of life. They do not only transport ions, metabolites and proteins across membranes, but also receive chemical signals from outside the cell and propagate electrical impulses. Membrane proteins are also necessary to attach to neighboring cells, to anchor other proteins to specific locations in the cell and to regulate intracellular vesicular transport. And to demonstrate their functional diversity even further,

they also control membrane lipid composition, and organize and maintain the shape of the organelles and the cell itself [1]. Through genome-wide studies it was estimated that membrane proteins constitute 20-30% of proteomes [2–6]. Given their abundance and functional significance, it is not surprising that defects in membrane proteins are associated with many known diseases [1] such as cystic fibrosis [7], color blindness [8], alzheimer [9] and diabetes [10]. Likewise, membrane proteins represent more than 60% of drug targets [11] and are therefore of great pharmaceutical interest as well.

1.1.1 Basic principles

There are two types of membrane proteins differing in the degree to which they span the membrane lipid bilayer. Peripheral membrane proteins are loosely associated with the membrane with no part extending into the hydrophobic bilayer environment. Integral membrane proteins on the other hand are embedded in the phospholipid bilayer [12]. Integral membrane proteins can be further divided into two distinct architectures: α-helix bundle and β-barrel proteins [13]. β-barrel proteins are integrated exclusively in the outer membrane of Gram-negative bacteria, mitochondria and chloroplasts [1]. They constitute 2-3% of the proteome [14] and are also called porins. In contrast, α-helical membrane proteins occur not only in all cellular membranes, but also exhibit a broader functional range and are much more abundant [1]. Therefore, all the following sections and chapters will focus on α-helical membrane proteins.

Membrane proteins differ remarkably from soluble proteins because they do not only reside in an aqueous environment, but are also embedded in a lipid bilayer. The bilayer can be distinguished into three distinct regions [15]. The first part is formed by the hydrophobic core of the bilayer that is $\sim$30 Å thick. The second and third part are the surrounding head group layers on both sides (each of $\sim$15 Å width). With regard to this bilayer thickness, transmembrane helices are usually about 20 to 25 residues in length [16]. The different environments that a membrane protein is faced with result in different amino acid preferences along the bilayer [17]. In the middle of the membrane, transmembrane helices are strongly enriched in hydrophobic residues (such as Ala, Ile, Val and Leu). The aromatic residues Tyr and Trp are abundant in the lipid-water interface near the helix ends. Finally, positively charged residues (Lys and Arg) are four times more abundant in loops located at the cytoplasmic side of membranes than in extracellular loops [18]. This observation was termed the positive-inside rule and was shown to appear in all three kingdoms of life [4, 19].

1.1.2 Membrane structure space

A landmark in the history of membrane proteins was the structure determination of the first high-resolution membrane protein, the photosynthetic reaction center, in 1985 [20]. In fact, it was such a meaningful achievement that Johann Deisenhofer, Robert Hubert and Hartmut Michel were honored with the Nobel prize in chemistry in 1988 [21]. However, it has to be noted that the bacteriorhodopsin structure had already been resolved ten years beforehand [22]. But most probably due to the low resolution (7 Å), the high-resolution (3 Å) structure of the photosynthetic reaction center instead is usually specified as the first resolved membrane protein structure.

In contrast to soluble proteins, membrane proteins reside in a lipid bilayer environment that strongly restricts the range of possible transmembrane protein structures [15]. For a long time, α-helical membrane proteins appeared to adopt a very simple architecture [17, 23]: transmembrane helices are oriented more or less perpendicular to the membrane plane forming an α-helix bundle, in which the helices pack with typical knobs-into-holes packing angles [24, 25]. A very prominent example of this structural simplicity is bacteriorhodopsin (Figure 1.1A).

One major challenge in membrane protein bioinformatics is the paucity of structural data. In spite of their functional and pharmaceutical importance, less than 2% (1,550 out of 79,538[1]) of the structures in the Protein Data Bank (PDB) [27] account for membrane proteins. This is because membrane proteins are extremely difficult to determine using classical techniques and high-resolution structures are hard to obtain [13, 28]. However, significant efforts have been made in the last years to further improve the structure determination technology for membrane proteins [29–32] (for a review see [33]). The number of available unique membrane protein structures increases exponentially and doubles every ∼3 years [17, 34]. While only about 40 unique structures were available in 2000, 310 structures exist today (data taken from the Stephen White laboratory[2]; as of January, 2012). Although this is an encouraging trend, progress in membrane protein structure determination lacks 15 years behind that of soluble proteins and the few structures only give a limited view on membrane protein structure space [17]. This is evident from the structures that appeared recently showing that membrane proteins are much more diverse than initially assumed giving new insights into membrane protein structure and evolution.

[1] Statistics taken from PDBTM [26] as of January, 2012
[2] `http://blanco.biomol.uci.edu/mpstruc/listAll/list`

Irregular structures

Apart from the simple membrane protein structures (Figure 1.1A) with straight transmembrane helices traversing the membrane bilayer, irregularities in the regular structure have recently been found [35, 36] including proline kinks, tilted helices, interface helices, interrupted helices, and reentrant regions.

- **Proline kinks**: Although Proline is known to be a helix-breaker, it is often found in the middle of transmembrane helices of membrane proteins [37, 38] (Figure 1.1B). Instead of breaking the helix, Proline introduces a notable kink in its backbone [38–40]. In several cases, it was shown that some of the transmembrane prolines are fundamental for the structure and function of the respective membrane protein [36, 41–44].

- **Strongly tilted helices**: Some α-helical membrane proteins contain transmembrane helices that are strongly tilted and thus much longer. A good example is the Clc chloride channel [45] (Figure 1.1C) that contains several transmembrane helices that are not parallel to the membrane normal. It seems that helix tilting is a response to hydrophobic mismatch [46, 47] that occurs if the length of the hydrophobic transmembrane helix does not match the hydrophobic thickness of a membrane bilayer [48]. By doing so the length of the transmembrane helix is adjusted to the bilayer thickness. In some cases, such as bacteriorhodopsin [49, 50] and the mechanosensitive channel MscL [48], it could be shown that the helix tilting is fundamental for the functional activity of the protein. It has to be noted that there also exists other compensation mechanisms for hydrophobic mismatch, such as oligomerization or backbone kinking [48, 50].

- **Interface helices**: Helices that reside in the membrane-water interface region and that are situated parallel with the membrane are called interface helices [51] (Figure 1.1D). 30% of the residues in the interface region form interface helices (70% have irregular secondary structure) that are shorter than transmembrane helices. Like transmembrane helices, interface helices consist mainly of hydrophobic residues, but they contain much more polar aromatic residues (Trp, Tyr) [51, 52]. In addition, while in transmembrane helices Trp and Tyr extend their side chains to the membrane bilayer pointing away from the membrane core (known as 'snorkeling' [53]), in interface helices the side chains of Trp and Tyr tend to point towards the membrane core ('anti-snorkeling' [53]). In contrast, polar and charged residues

tend to point away from the membrane center [51, 54]. Although the functional contributions of interface helices are not yet fully understood, it is assumed that they play an ancillary role by, for example, helping to position the transmembrane helices [51, 55].

- **Interrupted helices**: The regular secondary structure of a transmembrane helix can be interrupted such that coils occur in the core of transmembrane helices (Figure 1.1E). By analyzing these coil regions, Kauko *et al.* found that 7% of all residues in the transmembrane helix core region contain coils [56]. Compared to transmembrane helices, these coils contain more polar and charged residues, show an increased preference for Gly and Pro, and are significantly more buried and conserved. Additionally, it could be shown that coil regions are abundant within channels and transporters and are functionally important. For example, the coils in B_{12} ABC transporter seem to participate in the binding of vitamin B_{12} and in protein-protein interactions [56].

- **Reentrant regions**: So far, reentrant regions are most probably the best example for structural peculiarities in α-helical membrane proteins. These are regions that cross the membrane only partly and enter and exit the membrane on the same side [57] (Figure 1.1F). Depending on their secondary structure content, they can be divided into three distinct categories: reentrant regions with a helix-coil-helix motif, with a helix-coil or coil-helix motif, and without regular secondary structure. They were already found in several known membrane protein structures [58–64] and it is estimated that at least 10% of all membrane proteins in a genome contain reentrant regions. Thereby, reentrant regions are most commonly found in channels and transporters [57]. It was recently found that reentrant regions differ remarkably from transmembrane helices with respect to their hydrophobicity [65]. Reentrant regions are not only less hydrophobic than transmembrane helices, but also show a heterogeneous distribution of hydrophobic residues. While hydrophobic residues are equally distributed at all positions in transmembrane helices, reentrant regions are more hydrophobic at positions close to the membrane surfaces and less hydrophobic inside the membrane.

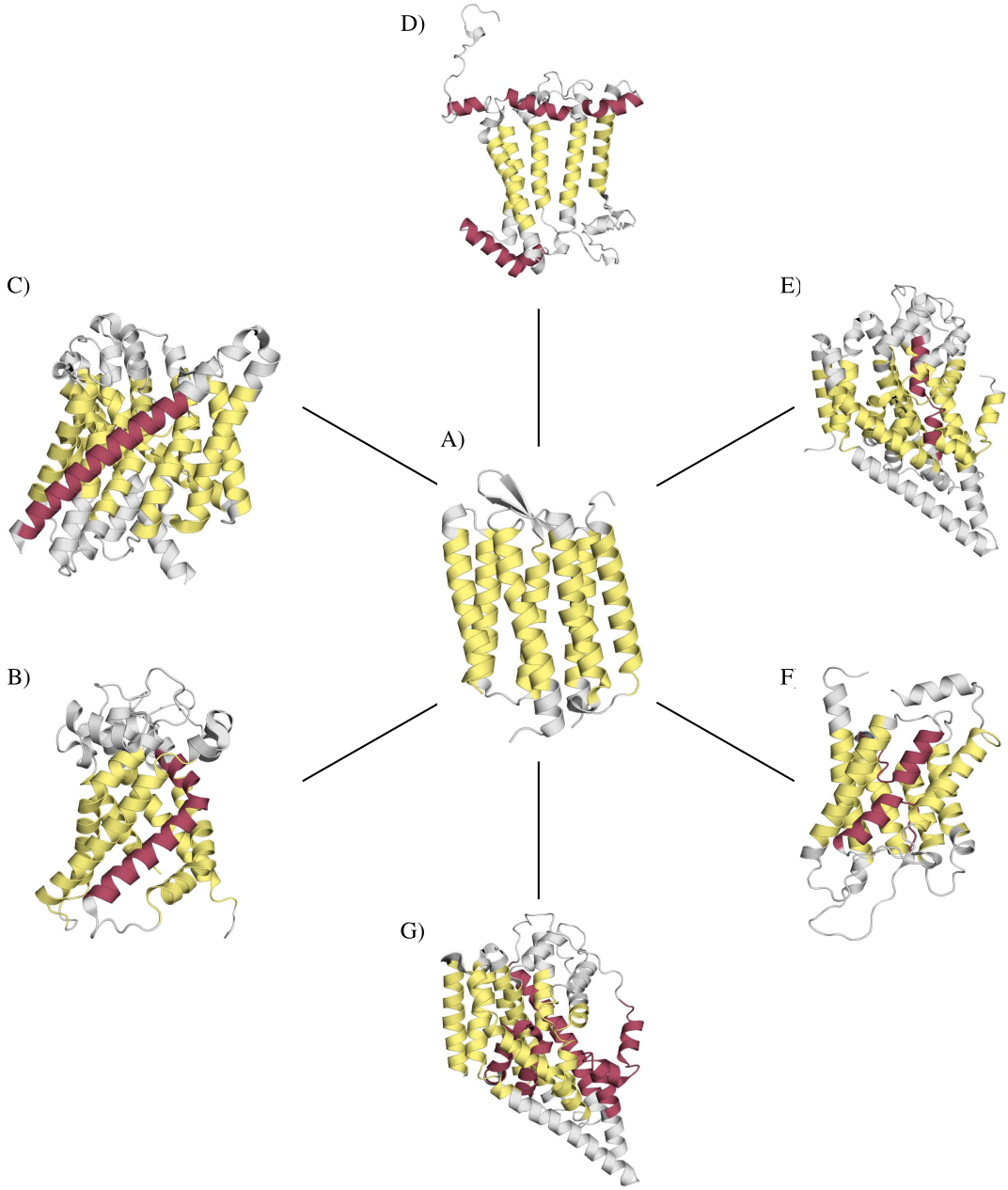

Figure 1.1: Structural diversity of membrane proteins. (A) Simple membrane protein structure of bacteriorhodopsin (PDB code: 1c3w, chain A) with straight transmembrane helices. (B) Structure of the bovine mitochondrial ADP-ATP carrier (PDB code: 2c3e, chain A) with proline kink. (C) Structure of the Clc chloride channel from *S. typhimurium* (PDB code: 1kpl, chain A) with strongly tilted helix. (D) Structure of the photosynthetic reaction center from *R. viridis* (PDB code: 1prc, chain M) with interface helix. (E) Structure of the GltPh (PDB code: 2nwl, chain A) with interrupted helix. (F) Structure of the AQP1 water channel (PDB code: 1j4n, chain A) with reentrant region. (G) Structure of the glutamate transporter homologue from *P. horikoshii* (PDB code: 1xfh, chain A) with strongly tilted and interrupted helices and reentrant regions. Yellow: transmembrane helices, Red: examples of structural irregularities. The figure was drawn using PyMOL [66].

Taken together, many substructures have been found that clearly deviate from the simple membrane protein architectures that we are used to from early membrane protein structures. It is worth mentioning that some membrane proteins even contain several of these irregular substructures. A good example is the eukaryotic glutamate transporter homologue from *Pyrococcus horikoshii* that consists of tilted helices, reentrant regions and interrupted helices [35, 67] (Figure 1.1G). Most importantly, it seems that all these substructures are somehow involved in the function of the corresponding protein. Therefore, it is inevitable to leave behind our view of a simple helix-bundle arrangement as it is overwhelmingly clear that membrane protein structures are much more diverse than initially expected.

3D and 2.5D prediction methods

As only few membrane protein structures are available, computational methods to predict the three-dimensional structure of a membrane protein are in high demand. A couple of prediction methods were already developed [12] and will be briefly summarized here.

One of the main approaches to predict the three-dimensional structure of a protein from its amino acid sequence is homology modeling (also known as comparative modeling). Homology modeling describes the technique of predicting the structure of a given protein sequence (target) based on one or more known three-dimensional structures of homologous proteins (templates) [68]. Given that membrane proteins clearly differ from soluble proteins, it can not be taken for granted that the same methods originally developed for soluble proteins can automatically be applied to membrane proteins as well. However, it could be shown that existing methods perform equally good on membrane proteins [69, 70]. Furthermore, one of the most important steps in homology modeling is the alignment between target and template. Thus, as membrane protein specific substitution matrices [71–75] and alignment methods [76–78] do exist, it is even possible to further optimize existing methods to membrane proteins. That available homology modeling methods can yield better models if specifically tailored to membrane proteins was recently shown by the newly developed MEDELLER algorithm [79].

Similar attempts have also been made in the field of fold recognition methods (also known as threading). Fold recognition, in contrast to homology modeling, does not depend on templates, but rather uses a library of all known folds and tries to find the most compatible one [80]. One fold recognition method developed specifically for membrane proteins was proposed by Taylor *et al.* [81].

Unfortunately, the fundamental problem with these template based methods is the paucity of structural data [12]. Given that only 310 unique membrane protein structures exist up to now, the chance of detecting a suitable template structure is very small and thus limits the applicability.

A possible solution to this problem are the *ab initio* prediction methods that depend solely on the sequence of the target protein [80]. In fact, some membrane protein specific *ab initio* methods already exist [82, 83]. However, these methods are limited too as they have problems in handling large membrane protein structures.

Finally, apart from the methods that try to predict the whole structure of a membrane protein, other methods exist that are specialized in the substructures presented in the previous paragraph. The prediction methods for these substructures were termed 2.5D predictions [35] as the substructures are between the 2D dimension of the topology and the 3D dimension of the tertiary structure. There are tools available for the prediction of kinks [84–86], tilted helices [87, 88], interface helices [55, 57], and reentrant regions [57, 89–92]. And interrupted helices can be predicted by a combination [93] of ZPRED [87, 88] and PsiPred [94].

To summarize, several 3D and 2.5D prediction methods are available. But either they have only a limited applicability or predict only parts of the membrane protein structure. Therefore, the prediction of the three-dimensional structure of membrane proteins is still a very challenging task. More promising in this respect is the prediction of membrane protein topology which will be the focus of the next paragraph.

1.1.3 Membrane protein topology

The topology of a membrane protein describes the number and location of its transmembrane helices together with their orientations relative to the membrane. It is a meaningful feature of membrane proteins as it represents an important intermediate between the primary structure and the fully folded three-dimensional structure [17]. The major determinants of a membrane protein topology are the hydrophobic stretches of the transmembrane helices and the positive-inside rule that helps to define the cytosolic side of the protein [38].

Prediction of topology

Like all methods in bioinformatics, methods for predicting the topology of a membrane protein have evolved from simple approaches to more sophisticated ones. One of the

first methods was developed by Kyte and Doolittle in 1982 [95]. Using a hydropathy scale, a hydrophobicity score is first assigned to each residue in the sequence. Afterwards, a window of specific length slides along the sequence and if the sum of all values in the window exceeds a given threshold, then the hydrophobic stretch is predicted to be a transmembrane helix. Significant improvements have been achieved with the development of machine-learning approaches including neural networks (e.g. PHDhtm [96] and Memsat [97]), hidden Markov models (HMM) [98–102] and more recently, support vector machines [92, 103, 104] and bayesian networks [105]. HMM based methods (such as HMMTOP [99] and TMHMM [98]) were shown to be particularly successful [101, 106, 107] and thus are commonly used. One difficulty that almost all prediction methods struggle with is that signal peptides are hydrophobic as well and are hence often wrongly predicted as transmembrane helices and *vice versa*. Again it was a HMM based method (called Phobius [100]) which could successfully address this problem first by combining signal peptide prediction and topology prediction within one model. Finally, there also a couple of methods available that use the results of multiple prediction programs and combine them into one consensus prediction [108–113].

Although there are already many prediction methods available, there is still room for improvement. As was the case after switching from simple hydropathy scale analysis to advanced machine-learning approaches. With new prediction methods emerging that additionally incorporate new structural features (such as reentrant regions, tilts etc.) it is likely that they will clearly outperform all existing methods. In fact, some methods already exist that go in this direction by either combining different machine-learning approaches and/or predicting reentrant regions as well [90–92, 114, 115]. For example, SPOCTOPUS [91] not only combines HMMs with neural networks, but also predicts transmembrane topology, signal peptides and reentrant regions. However, it is also possible that with new membrane protein structures even more structural deviations will appear that further challenge existing prediction methods. Thus, it seems that membrane proteins will keep us busy for the next few years.

Evolution of topology

There is also another reason why the topology of a membrane protein is given so much attention. In contrast to soluble proteins, the topology represents an additional dimension for membrane protein evolution [17]. The first examples presented in this context show that one membrane protein adopts not necessarily only one topology [17, 36, 38].

- **Opposite topologies**: It could be shown that homologous proteins such as RnfA

and RnfE [116] or YdgE and YdgF [117] have the same number of transmembrane helices and are in agreement with the positive-inside rule, but show opposite topology [17, 38]. These proteins represent examples of divergent evolution of membrane protein topology.

- **Dual topologies**: In contrast to the cases of opposite topology, other proteins do not follow the positive-inside rule and thus adopt a so-called dual topology [17, 36, 118]. That means that these proteins can insert in two opposite orientations [118]. Examples of dual-topology proteins include SugE and CrcB [117].

- **Multiple topologies**: Some proteins can even adopt different topological forms as a consequence of inefficiently inserting transmembrane helices. One example is given by the scrapie prion protein that exists not only as a single-spanning membrane protein in two opposite orientations, but also in a cytoplasmic and fully secreted form [17, 119].

- **Dynamic topologies**: Furthermore, membrane proteins are not necessarily stable entities with respect to their topology. Some of them can reorient their whole topology as part of a reaction cycle. Similarly, also single transmembrane helices can be repositioned during folding and oligomerization [17, 93].

One of the main mechanisms that drive membrane protein evolution is internal gene duplication. The majority of duplications are complete resulting in a duplication of transmembrane helix number and symmetric three-dimensional structures. Particularly membrane proteins with 8, 10 and 12 transmembrane helices seem to have evolved through a complete gene duplication [17, 120, 121]. Thereby, the larger the membrane protein is (i.e. the more transmembrane helices it contains), the higher the frequency to be internally duplicated [122]. The symmetry resulting from such a duplication is apparent in the structures of Lactose permease [123], Clc chloride channels [45], aquaporin-1 [60] and YrbG [124]. But these are only a few examples and Choi *et al.* found that almost half of all known membrane protein structures (that were available at that time) contain internal symmetries [125]. Although the duplication of an odd number of transmembrane helices (called anti-parallel duplication) represents a special case (as the two homologous domains cannot both retain their original orientation), it was found that parallel and anti-parallel duplications are equally common [122]. In context of duplications, Rapp *et al.* suggested a possible scenario for opposite topologies [118, 126]. They hypothesized that proteins with opposite topologies may have emerged from a

duplication of a dual-topology protein followed by divergent topology evolution. And if these proteins are finally fused, the resulting polypeptide adopts an anti-parallel symmetric structure. Therefore, dual-topology proteins form a possible evolutionary path for proteins containing anti-parallel duplications. Another type of internal duplication is the partial duplication, where only part of the primordial protein is duplicated. It was frequently found in proteins with 6 and 7 transmembrane helices likely to have evolved from primordial proteins with 2,3 or 4 and 3,4 or 5 transmembrane helices, respectively [121].

Besides gene duplication, another important force in driving protein evolution is domain recombination. While the first was shown to be common in membrane proteins, the second one is not [127]. Liu *et al.* argue that membrane proteins use an alternative method to gain evolutionary diversity, namely the formation of oligomers that is frequently found among membrane protein structures [15, 127]. However, it has to be mentioned that the uncommon recombination applies to non-homologous membrane domains, but not to the recombination of transmembrane and soluble domains.

1.1.4 Genome-wide analysis

Predicting membrane protein topology on genome-wide scale followed by a comparative analysis of the results for different genomes allows us to address many interesting questions. For example, how abundant are membrane proteins within the studied proteomes and are there differences? Or if there are membrane proteins with a given number of transmembrane helices that are prominent in one organism or kingdom, why not in the others? In the following these questions will be answered through the results of previously published analyses on that subject.

One of the first main results of these studies that is now cited in almost every publication on membrane protein research considers their abundance. The proportion of proteins with one or more predicted transmembrane helices within a proteome reaches 20-30% [2–6]. Whether there is a correlation between the number of transmembrane proteins and the complexity of an organism is ambiguous as the results are controversial [36]. While the results of some studies show that membrane proteins are more abundant in larger genomes [4], others do not affirm such a correlation [5, 128, 129].

A general decreasing trend is observed regarding the transmembrane helix distribution. Accordingly, membrane proteins with few transmembrane helices are more common than proteins with many [3, 4, 130]. But exceptions to this general trend do exist in that certain membrane protein topologies are highly favored. These exceptions are reported

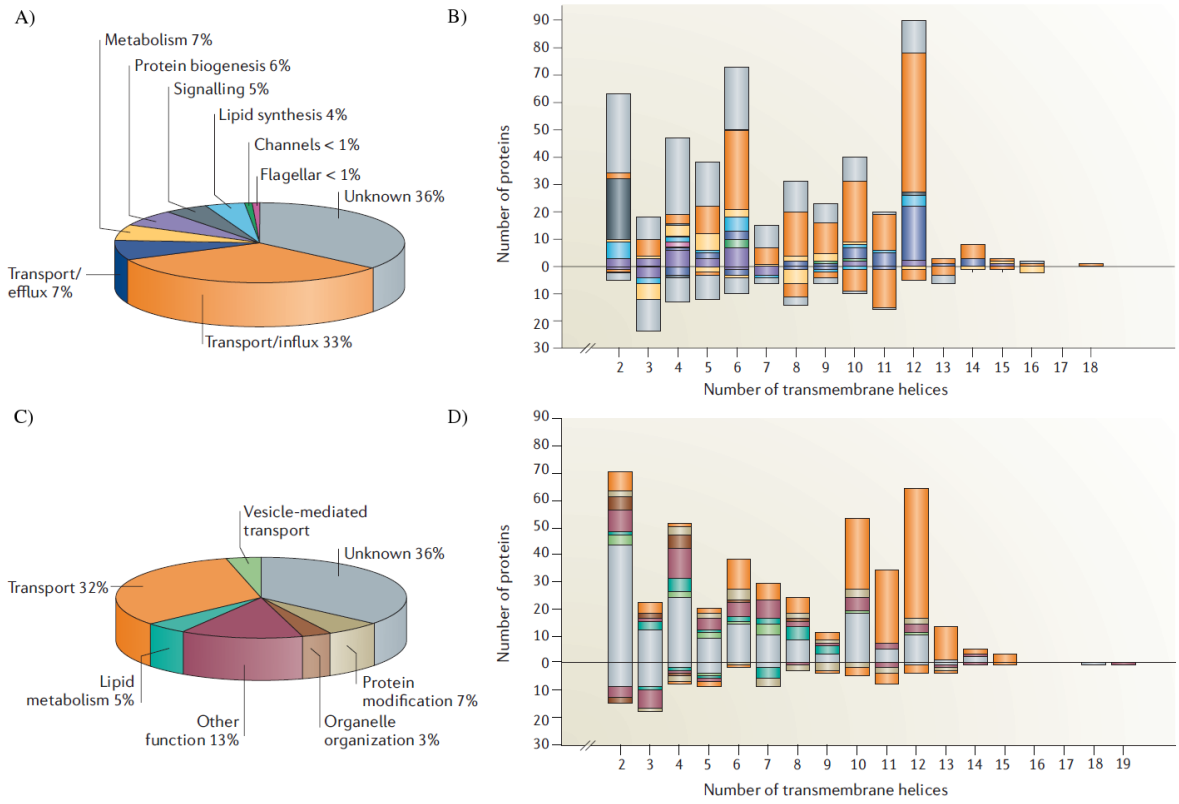

Figure 1.2: Topology distribution in membrane proteomes. (A) Functional categorization of membrane proteins in *E. coli*. (B) Distribution of functional categories with respect to topology in *E. coli*. Proteins with a cytoplasmic C terminus are plotted upwards and those with an extracytoplasmic C terminus are plotted downwards. (C) Same as (A), but for *S. cerevisiae*. (D) Same as (B), but for *S. cerevisiae*. Figure taken from [1].

in many different studies [3, 4, 117, 131–133]. The results of these analyses can be summarized as follows. Proteins with 6 and 12 transmembrane helices are predominant in uni-cellular organisms and constitute small-molecule transporters, sugar transporters and ABC transporters [3, 4]. On the other hand, *C. elegans* and human favor topologies with 7 transmembrane helices [3, 4, 131] which can be explained by the high abundance of G-protein coupled receptors (GPCRs). In human, 5% of the whole proteome and even 15% of the membrane proteome account for GPCRs [134]. Jones called such favored topologies 'transmembrane superfolds' [3] in accordance with the superfolds found for soluble proteins [135]. Furthermore, it was found that the topology with both N- and C-terminus in the cytoplasma is abundant in proteins with an even number of transmembrane helices suggesting that pairs of helices connected by short loops (called helical hairpins) may be a preferred insertion mechanism [38, 117, 132, 133]. Finally, studies

on *S. cerevisiae* [133] and *E. coli* [117] revealed a strong correlation between structure and function (Figure 1.2). While proteins with at least 10 (in case of *S. cerevisiae*) and 12 (in case of *E. coli*) transmembrane helices are most often transporters, most of the proteins with less than 5 transmembrane helices have no functional annotation. These small membrane proteins should be the focus of future functional investigations in order to fill that gap of knowledge.

1.2 Classification of proteins

Large-scale sequencing projects accumulate data at an ever-increasing rate. Therefore, it becomes more and more important to organize similar proteins into groups, a process known as classification, to keep pace with the large amount of data. The most well-known classification approach is probably the biological classification of organisms as for example exemplified by Carl von Linne [136].

At the protein level, there are three important attributes according to which proteins can be classified: sequence, structure and function. These three approaches are described as follows.

1.2.1 Sequence-based classification

One of the most common approaches of protein classification is based on sequence similarity that is measured by using algorithms such as BLAST [137] or FASTA [138]. Sequence-based classification schemes focus on the detection of homologous sequences assuming that proteins with similar sequences are evolutionary related.

The big advantage of establishing protein classifications on sequence information is that it is not limited to structure data. While the UniProt database [139] currently contains more than 19 million sequences (release 2011_12; SwissProt+TrEMBL), only about 78,300 structures are deposited in PDB [27] (as of January, 2012). One disadvantage of these methods is that it is very difficult to define a reliable similarity threshold. On the one hand, if the threshold is too restrictive, only few or no matches are found. On the other hand, if the threshold is too permissive, a lot of false-positives are retrieved.

Many methods exist that group proteins solely based on sequence information [140–144]. Two of the more well-known are Pfam [145, 146] and COG [147, 148].

Pfam

Pfam is a well-established database that classifies protein sequences from the UniProt database into families. The Pfam database consists of two components: Pfam-A and Pfam-B. Pfam-A is a manually curated version of Pfam containing well-characterized protein domain families with high quality alignments. The families of the Pfam-B supplement are automatically generated and are thus of lower quality. However, since Pfam-A does not cover all known proteins, Pfam-B families might be useful when no Pfam-A families are found.

In order to accomplish the high quality of the Pfam-A alignments, Pfam uses two alignments. The seed alignment is constructed from a representative set of sequences that are described in literature as belonging to the same family. After manual inspection of the alignment quality, an HMM is built that is used to detect additional family members in UniProt. Finally, a full alignment is constructed of all members. Further quality controls are performed once the two alignments have been constructed. Thus, each Pfam family is represented by a multiple alignment and an HMM. In addition, annotation and cross-references to other databases (such as PDB and SCOP) are provided. (species distribution)

Since its first release in 1997, Pfam has not only increased the data, but also developed a lot of new features. One of these features was the grouping of related Pfam families in so-called clans [149]. A clan summarizes two or more families that have arisen from a single evolutionary origin. One of the last developments was the linkage of Pfam families to relevant Wikipedia pages [146].

The whole set of HMMs represents a comprehensive library that can be used to identify and classify domains in protein sequences. The recent version of Pfam (release 26.0) contains 13,672 Pfam-A families covering almost 80% of UniProt and 499 clans [146].

COG

The protein database of Clusters of Orthologous Groups (COGs) classifies proteins from completely sequenced genomes according to orthologous relationships. Each COG includes proteins from at least three distantly related species that are inferred to be orthologs. These are genes from different species that have evolved from a common ancestor through speciation [150].

COGs are derived based on all-against-all sequence comparison of proteins from completely sequenced genomes. First, for each protein its best hit in each of the other

genomes is determined. Then, all triangles formed by best hits are detected (called minimal COGs). Finally, triangles with a common side are merged and each COG is manually refined in order to eliminate false positives.

Since orthologs typically retain the same function in different species [147, 151], it is possible to transfer functional information from one member to an entire COG. The COGs provide a powerful tool for automatic functional annotation of newly sequenced genomes [152–154] and genome-wide evolutionary studies [155, 156].

The current release of the COGs database consists of 4,872 prokaryotic COGs from 66 genomes and 4,852 eukaryotic COGs (termed KOGs) from 7 genomes.

1.2.2 Structure-based classification

Another approach to protein classification is based on the comparison of three-dimensional protein structures. The big advantage of using structural information is that distant evolutionary relationships between proteins can be revealed that are undetectable by sequence analysis [157]. This is because structure is more conserved than sequence [158, 159].

The detection of distant relationships is not the only benefit of structure-based classification databases. These databases also allow the generation of template libraries of unique structures that are needed in structure prediction methods like fold recognition and homology modeling. Furthermore, they simply give an easy accessible overview of the diversity of protein structures and allow to estimate the number of protein folds and families [160, 161].

Among the most commonly used structure-based classification approaches are SCOP [162, 163] (Figure 1.3A) and CATH [164, 165] (Figure 1.3B) that are based on evolutionary relationships. These two databases have become the gold standard in structural bioinformatics.

SCOP

In 1995, the Structural Classification of Proteins (SCOP) database was established by Murzin *et al.* as the first classification approach of determined structures. The basic classification unit in SCOP is the domain. A domain is defined as a compact segment of the polypeptide chain that folds independently [167, 168] and it occurs either in isolation or within a multi-domain protein [169]. That means that proteins are first split into their constituent domains and then each domain is classified separately. SCOP defines the

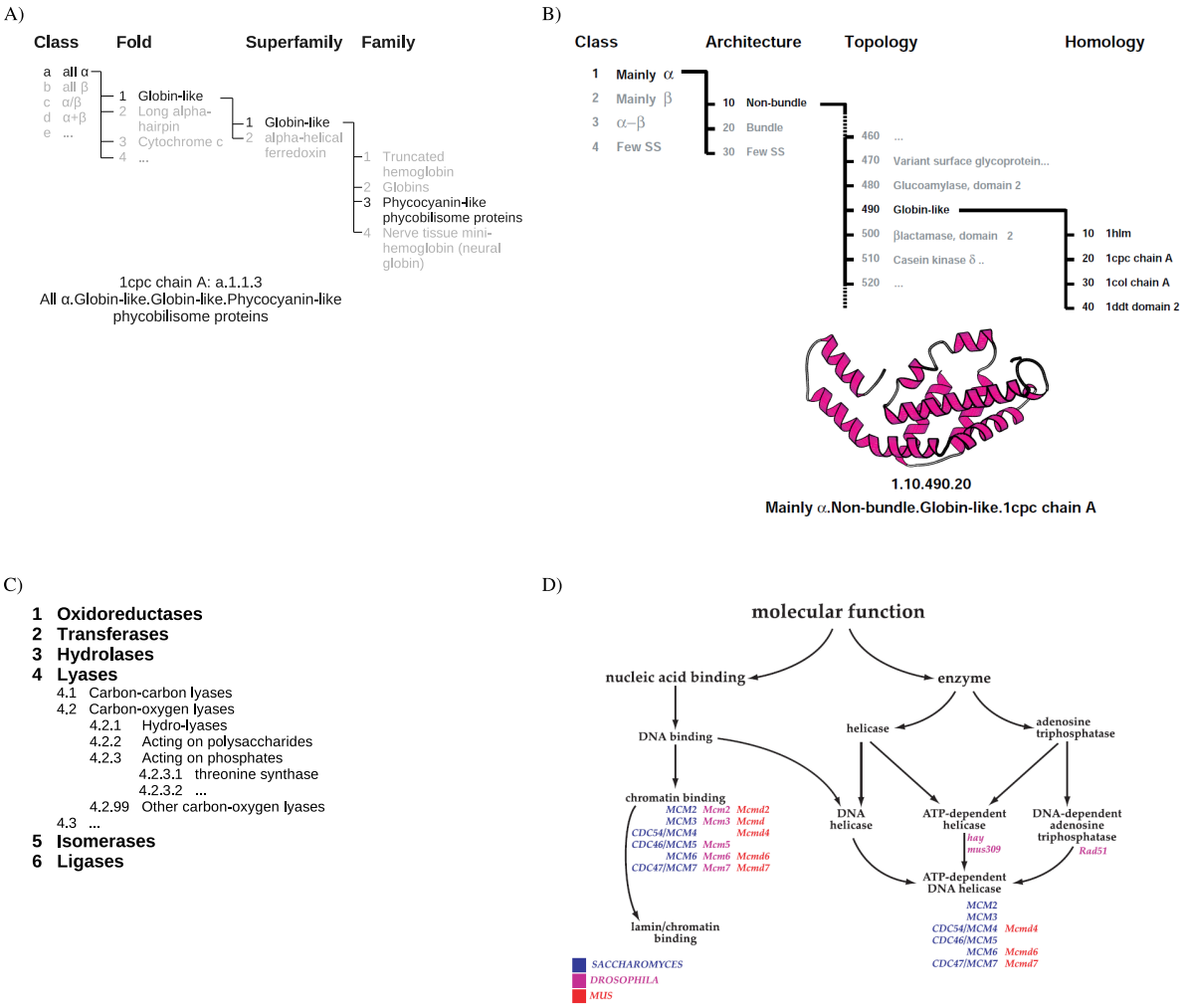

Figure 1.3: Overview of structural and functional classification approaches. Structural classification of SCOP (A) and CATH (B) for C-phycocyanin (PDB code: 1cpc, chain A). Right figure taken from [164]. Examples of the functional classification approaches EC (C) and GO (D). Right figure taken from [166].

boundaries of the domains manually.

Proteins are classified according to a four-level hierarchical scheme [162].

- **Class**: The class is the most general level and considers the content and organization of secondary structure elements. The four main classes are all-alpha, all-beta, alpha+beta and alpha/beta.
- **Fold**: The fold level describes geometrical relationships by considering the overall shape of a protein. Two proteins have the same fold if the major secondary structures are the same and assemble in the same arrangement and with the same

topological connections. Protein domains grouped together in the same fold may not have a common evolutionary origin.

- **Superfamily**: The superfamily level groups together proteins with low sequence identities, but where similarities in structure and function suggest that a common evolutionary origin is probable.

- **Family**: At the family level, proteins are clearly evolutionary related. This is shown either by high sequence identity ($\geq$ 30%) or lower sequence identity but very similar structures and functions. The clustering of proteins according to sequence similarity is the only automatically performed step in SCOP.

With only one exception (family level), all classification steps are manually performed with visual inspection and structure comparison demanding a detailed human expert knowledge. Therefore, SCOP is supposed to be the most accurate classification approach [162]. The problem with such an approach is, however, that the results are not always reproducible.

The latest version of SCOP (release 1.75) contains 1,195 folds, 1,962 superfamilies and 3,902 families.

CATH

While SCOP uses a nearly complete manual approach, CATH (created by Orengo in 1997) classifies proteins in a semi-automatic manner. For example, the definition of domain boundaries is done automatically (in contrast to SCOP).

Common to both approaches is that the classification is on hierarchical levels according to evolutionary and structural relationships. The CATH hierarchy comprises the following levels [164]:

- **Class**: Similar to SCOP, the class level regards the secondary structure content. Three classes are defined: all-alpha, all-beta and alpha/beta.

- **Architecture**: The architecture level considers the overall shape of the structure as determined by the orientation of the secondary structure elements regardless of their connectivity. The assignment of the architecture level is a manual step in CATH.

- **Topology**: At the topology level, the overall shape and connectivity of secondary structure elements are considered. The topology level corresponds to the fold level in SCOP.

- **Homologous superfamily**: Similar to topology but requiring a higher degree of structural similarity coupled with a functional similarity. At this level, proteins may have evolved from a common ancestor.

- **Sequence, Nearly identical, identical**: At these levels, protein domains share a sequence identity of 35%, 95% and 100%, respectively.

At present CATH (release v3.4) contains 40 architectures, 1,282 topologies and 2,549 homologous superfamilies.

With the large increase in the PDB database, manual approaches become more and more difficult furthering the need for an automatic classification. In fact, several methods are already available that completely rely on large-scale structure comparisons and are therefore fully automated [170–178].

1.2.3 Function-based classification

Finally, proteins can also be assigned to groups on the basis of functional similarity. The difficulty with functional classification methods is that there are several ways to define a protein's function. Function can be defined based on enzyme reaction mechanisms, participation in biochemical pathways, functional roles and cellular localization [136]. Furthermore, function can be described at different levels. For example, the term 'lipid transport' describes a protein's function more specific than the term 'transport'.

Among the most widely used function-based classification approaches are the Enzyme Commission [179] (Figure 1.3C) and Gene Ontology [166, 180] (Figure 1.3D).

Enzyme Commission

The Enzyme Commission (EC) provided the first detailed classification of protein function designed as a four-level hierarchy. The first field in the EC number describes the enzymatic class. The second and third field further specify the class. And the fourth field indicates the specific enzymatic activity. For example, the EC number 4.2.3.1 corresponds to a lyase (4.*), more precisely a carbon-oxygen-lyase (4.2.*) that acts on phosphates (4.2.3.*), finally describing a threonine synthase (4.2.3.1).

The ENZYME database [181, 182] contains valuable information for each EC number specified by the Enzyme Commission, such as the list of proteins that are associated with the EC number.

Gene Ontology

The Gene Ontology (GO) Consortium provides a structured, dynamic, controlled vocabulary for describing functions of genes and gene products. As function can be described in several ways, the GO Consortium comprises three independent subschemes called ontologies. An ontology is defined as a vocabulary of well-defined terms with well-defined relationships [166].

- **Biological process**: describes a biological goal to which the gene product contributes (Example term: GO:0016049 'cell growth').
- **Molecular function**: describes the biochemical activity of a gene product (Example term: GO: 0038023 'signaling receptor activity')
- **Cellular component**: describes the location in the cell where a gene product is active (Example term: GO:0005622 'intracellular')

Unlike EC, GO is not limited to enzymes. And in contrast to the strict hierarchical EC numbering, GO terms are organized in a directed acyclic graph (DAG). As such, it is possible that one GO term has multiple parental terms. For example, the term GO:0038023 ('signaling receptor activity') is connected to the parental terms GO:0004871 ('signal transducer activity') and GO:0004872 ('receptor activity'). It is important to note that one gene product can be associated with several GO terms (from the same ontology) reflecting the fact that a protein may be multi-functional.

1.3 From sequence to structure to function

Protein classification methods are a valuable tool for the investigation of sequence-structure and structure-function relationships that give insight into protein evolution. Although it is very common that proteins with similar sequences and similar structures tend to possess similar functions, examples exist showing deviation from this general relationship, as will be shown in the following.

1.3.1 Sequence-structure relationships

Anfinsen's theorem postulates that all information necessary for a protein to fold into its three-dimensional structure residues within its amino acid sequence [183]. Therefore, if two pairs of proteins have similar sequences, it is likely that they also adopt

similar structures [158]. This is the cornerstone of homology modeling a method to predict the three-dimensional structure of a protein using the already known structure of a sequence-similar protein. Several studies have analyzed to which extent sequence similarity ensures structural similarity. It was found that proteins sharing at least 35% sequence identity will almost certainly have similar structures [158, 184, 185]. In the region of 20-35%, which is often called the 'twilight zone', structural similarity is considerably less common [184].

However, several examples were found that show the converse indicating that the relationship between sequence and structure is ambiguous. As such, different sequences can lead to similar structures, and similar sequences to different structures.

Different sequences, similar structures. A very prominent example for the first case are hemoglobin and myoglobin sharing only 25% sequence identity, but having closely related structures (RMSD 1.5 Å) [186]. Such examples highlight the fact that structure is more conserved than sequence [158, 159].

Similar sequences, different structures. Conversely, Kosloff and Kolodny found a significant set of proteins with highly similar sequences (ranging between 50-100%) and substantially different structures [187]. Very interesting examples for this case are pairs of DNA polymerase β proteins and diphtheria toxins (with sequence identities close to 100%) [188]. It is being argued that the structural dissimilarity is caused by functional issues [187, 188]. In the last-mentioned examples, the function involves large conformational changes. Given that some proteins do exist in both open and closed forms, it becomes apparent that sequence similarity does not always imply structure similarity.

1.3.2 Structure-function relationships

Sequence determines structure and structure determines function. Therefore, as sequence can be used to predict structure (e.g. through homology modeling), structure can be used to predict function. This can be done by searching for proteins of known function that are similar in sequence and structure to the target protein. If such proteins could be found, then the function is transferred to the target protein [189–191]. Thus, methods of functional transfer are based on the assumption that structural similarity of proteins implies their functional similarity [192]. 'The great majority of proteins which exhibit significant structural similarity are homologues and perform identical or similar functions' [193] (Figure 1.4).

However, as was already the case for sequence and structure, the relationship between

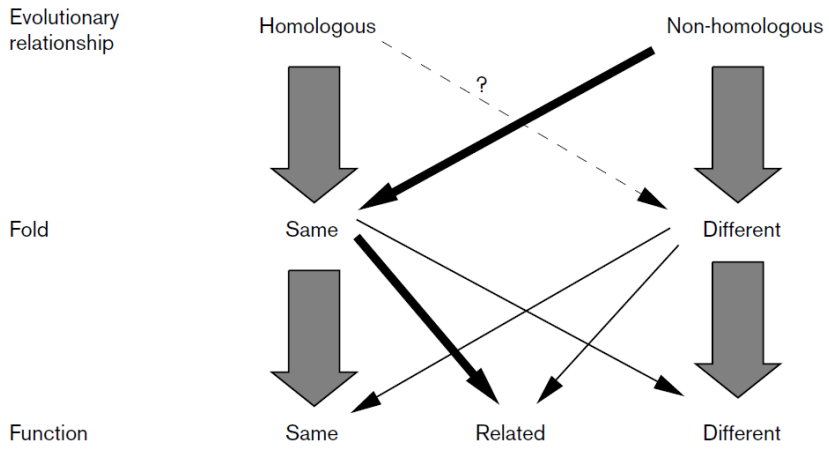

Figure 1.4: Illustration of the relationships between fold and function. The thicker the arrows the higher the relative frequency of the relationships. There is a clear trend that homologous proteins usually have the same fold and function, while the reverse applies to non-homologous proteins. However, deviations from this trend occur, although being less frequent. Figure taken from [193].

structure and function is also not straightforward.

Different structure, similar function. Trypsin and subtilisin proteinases are non-homologous proteins with different structures but the same function [193]. Other examples include carbonic anhydrases, glycosidases and carboxylases [194];

Similar structure, different function. It was found that alpha-beta folds are particularly associated with functional diversity [191, 194]. The TIM-barrel and Rossmann folds are most probably the best known examples in this respect [194, 195]. Similarly, lysozyme and alpha-lactalbumin adopt similar structures, but, although being homologues, perform different functions [193]. All mentioned examples are enzymes. Several researchers argue that the enzyme folds are functionally diverse because their function depends only on a few amino acids and not the whole fold [193, 196].

1.4 Motivation and Outline

Membrane proteins are involved in many essential cellular functions and are of great pharmaceutical interest. However, only few structures are currently available and thus the relationships between structure and function giving valuable insights into protein evolution are only poorly understood. In order to explore their structure-function relationships, it is necessary to have a comprehensive classification system specifically tailored to membrane proteins. In this thesis, existing structure classification methods

are evaluated and new classification methods are proposed.

In chapter 2 ('Classification of membrane proteins based on 3D structure'), the first analysis of the classification of membrane proteins in SCOP and CATH is presented. This analysis addresses the general occurrence of membrane proteins and folds within SCOP and CATH, as well as the differences in their domain and fold assignments. The study was motivated as follows. α-helical membrane proteins share the overall structure of a α-helix bundle with transmembrane helices oriented approximately perpendicular to the membrane. But at the same time, they exhibit a significant variety due to specific structural features such as reentrant regions, strongly tilted helices or interrupted helices (see section 1.1.2, page 3). Therefore, the question arises whether the fold definition initially developed for soluble proteins can be directly applied to membrane proteins as well.

In chapter 3 ('Classification of membrane proteins based on 1D and 2D structure'), a hierarchical classification approach (named CAMPS) is presented that is specifically tailored to membrane proteins and that aims to provide structural membrane protein families with members likely to share the same fold (SC-clusters). While SCOP and CATH are based on three-dimensional structures, CAMPS uses sequence similarity, the number of transmembrane helices and loop length patterns. The CAMPS classification is evaluated against different sequence-based, function-based and structure-based classification methods and the results are discussed in detail. The development of this classification system was motivated by the fact that methods based on three-dimensional structures (such as SCOP and CATH) are not comprehensive since membrane protein structures are rare. Furthermore, methods specifically established for membrane proteins can deal better with their structural characteristics.

In chapter 4 ('Classification of membrane proteins based on helix-helix interactions'), another fold determinant is considered, namely helix-helix interactions. The chapter describes how predicted consensus helix architectures can be used to identify SC-clusters that describe the same fold. A benchmark is performed in order to find the best parameters for the generation of consensus architectures. Then, the method is applied to a selected set of SC-clusters and the joined SC-clusters are further analyzed in terms of structural and functional aspects. The motivation for this work was that CAMPS can not deal with analogous structures resulting from convergent evolution since the approach is mainly based on sequence similarity.

At the end of the thesis, I will present the main conclusions that can be drawn from this work and I will give a short outlook on future work.

Chapter 2

Classification of membrane proteins based on 3D structure

"Science is always wrong. It never solves a problem without creating ten more."

(George Bernard Shaw)

Structural classification of proteins is a valuable tool for the understanding of protein function and evolution, with the power to reveal distant evolutionary relationships that are hidden at the sequence level [157]. Enumeration of protein architectures existing in nature not only gives a good overview of their structural diversity, but also provides invaluable clues into the general principles of protein structure organization [160, 161]. A great practical benefit of structural classification approaches is that they provide a library of unique folds that can be used in structure prediction methods, such as homology modeling and fold recognition. However, the structural classification of soluble proteins is much more advanced than the structural classification of membrane proteins.

For soluble proteins, a clear fold definition exists considering the number, arrangement, and connectivity of secondary structure elements. The total number of distinct protein folds existing in nature is estimated to range from 650 to 8,000, depending on assumptions made and datasets used [135, 161, 197–203]. The two most established resources, SCOP [162, 163] and CATH [164, 165], currently hold about 1,200 protein folds. Finally, the proportion of novel folds among newly determined structures has been steadily decreasing [165, 204–206], from 6% in 1997 to merely 0.4% 10 years later, suggesting that the soluble protein space is nearly exhausted.

In contrast, structure classification of membrane proteins is still in its infancy, pri-

marily because known structures are scarce, but also due to the lack of a clear fold definition. Therefore, the number of membrane protein folds is not known yet. Because of the physical constraints of the membrane bilayer, membrane proteins show a limited range of structural diversity and adopt either an α-helix bundle or β-barrel architecture [13]. However, the large variety of functions mediated by membrane proteins suggests that they attain their structural divergence at a different level than soluble proteins, leading to the question whether current fold definitions initially developed for soluble proteins can be directly applied to membrane proteins as well. Furthermore, considering the idea of a continuous fold space (see section 2.1.2) and the additional structural limitations imposed to membrane proteins by the lipid bilayer, the question arises whether membrane proteins can be reasonably classified into distinct folds at all.

Within this chapter, these issues are addressed by a comparative analysis of the structural classification of membrane proteins in SCOP and CATH. The next two sections briefly introduce comparative studies on structure classification databases and the continuity of the protein fold space. In the results section, three topics are investigated. First, the general occurrence of membrane proteins and folds within SCOP and CATH. Second, the differences in their domain assignments. And third, the differences in their fold assignments. So far, such comparative analyses were focused on the full set of available protein structures. The study presented here is the first comparative analysis of the classification of membrane proteins.

2.1 Introduction

2.1.1 Comparison of SCOP and CATH

SCOP [162, 163] and CATH [164, 165] are both hierarchical classification systems and employ a largely similar fold definition based on the number of secondary structure elements, their spatial orientation and connectivity (see section 1.2.2, page 15). However, they differ in their objectives and the methods used to divide proteins into structural domains, which is the first step in these structural classification approaches. Several studies have already investigated the similarities and differences between the classifications of SCOP and CATH [207–210]. Although they have shown considerable agreement among the classifications, remarkable discrepancies were found as well. Given that the domain partitioning is not a trivial task [211], it is not surprising that some of the differences arise from variations in the applied domain assignment procedure. Csaba *et al.*

found that the domain assignments agree only in about 70% of the cases [210]. Furthermore, several cases were found where SCOP and CATH differ in their fold and homology assignments. For instance, general folds in one database (such as the Rossmann fold) are separated into several more specific folds within the other database (termed the 'fold overlap' problem) [207]. 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. This is possible due to the fact that proteins may be structurally diverse despite a common evolutionary origin [207, 212]. While, some of these disagreements clearly arise as a consequence of specific differences in classification schemes, others may be attributed to the continuity of the protein structure space [212], which will be introduced in the next paragraph.

One of the motivations to perform these analyses was to generate a benchmark set containing protein domains that are consistently classified in several classification systems. Such datasets are highly valuable for the training and evaluation of diverse bioinformatics methods such as automatic structure classification, protein structure prediction [210] and homology detection [213].

2.1.2 Continuity of fold space

SCOP and CATH implicitly assume that the fold families are discrete entities and do not overlap due to local structural similarity [212, 214, 215]. However, this assumption is being questioned and the recent notion of a continuous fold space represents a serious challenge to hierarchical structure classification methods ([208, 212, 214–222] and references therein). It is argued that substructures (corresponding to structural motifs) below the domain level should be used as the basic unit for structural classification [208, 212]. Therefore, if domains are regarded as assemblies of several such substructures leading to local structural similarity of one protein to several other proteins that are not globally related to each other, then the structure space is continuous [212]. In this case, Pascual-Garcia *et al.* suggest to represent the protein structure space as a network and not as a tree in order to allow links between different folds having substructures in common [212]. Such a representation would also bring benefits for functional considerations, since it is suggested that functional relationships between different folds may be missed by grouping structures into non-overlapping folds [214, 219, 220].

Structural overlaps between protein folds were already noticed in 1997 by Orengo *et al.* who used the term 'russian doll effect' [164]. Since then several studies reported on

diverse examples of folds sharing significant structural similarity to other folds [214, 217, 218, 222]. Harrison *et al.* found that most of these folds, which they termed 'gregarious' folds, are α-β proteins and include super-secondary structural motifs, such as α-hairpins [217]. It is unclear to what extent similarities between folds exist. While Harrison *et al.* take the view that the fold space is mostly discrete with only few 'gregarious' folds, Skolnick *et al.* hypothesize that the protein structure space is almost entirely connected [221].

However, it seems more likely that the fold space is not completely continuous with some regions being continuous and others being discrete [212, 217]. It is argued that the different views on fold space are not mutually exclusive and 'both views have their place in practical applications and neither should be neglected or unnecessarily criticized' [220]. Hence, several authors suggest to use the notion of a continuous fold space to complement existing structural classification schemes such as SCOP and CATH [214, 217, 220]. This can be done by adding horizontal links between folds sharing structural substructures [214, 218]. Taken together, although the fold space continuity challenges traditional structural classification approaches, if the classifications are adapted to cope with overlapping folds (e.g. by horizontal connections), they will continue to serve as valuable tools for structural bioinformatics.

2.2 Materials and Methods

2.2.1 Datasets

An initial dataset of α-helical membrane proteins with experimentally determined structure was generated using the Protein Data Bank of Transmembrane Proteins (PDBTM [26]) as of October 1, 2008. From this database, all proteins that had at least one transmembrane helix were extracted according to the PDBTM annotation, yielding a dataset of 2,673 amino acid chains. Using the cd-hit algorithm [223], this dataset was made non-redundant such that no pair of proteins shared more than 95% sequence identity, resulting in a dataset of 381 amino acid chains.

From this initial collection, two datasets were created that contained all protein chains with domain and fold assignment in SCOP v1.73 [162, 163] and CATH v3.2 [164, 165], respectively. SCOP or CATH domains that did not contain at least one transmembrane segment were excluded from consideration. Note that in case of biological units consisting of multiple copies of the asymmetric unit, PDBTM contains multiple instances of

the same chain. These instances have different chain identifiers that are neither listed in PDB nor in SCOP or CATH. Such duplicate chains were ignored. Model structures as well as structures where the fraction of non-alpha carbon atoms is below 70% (for which CATH does not provide classification [224]) were also not considered. In one case, a protein chain (1p49, chain A) with two transmembrane helices and an assignment in SCOP ('alkaline phosphatase-like' fold (SCOP code c.76)) was removed from the SCOP dataset as the domain assignment in SCOP covered not only the transmembrane part but also the globular part of the protein. Similarly, the protein chain 1ehkB was excluded from the CATH dataset, as CATH obviously misclassifies this domain covering only a single transmembrane helix into a fold containing seven transmembrane helix proteins. Finally, the protein 1bzkA was removed from both datasets as the available 3D structure covers only parts of the full protein chain. After these filtering steps, the SCOP dataset included 156 protein chains, whereas the CATH dataset covered 110 protein chains (corresponding to 160 and 119 domains, respectively). These datasets are further referred to as `MP_SCOP` and `MP_CATH`, respectively.

For the comparative analysis of domain assignment and membrane protein fold classification, a third dataset, further referred to as `MP_shared`, was constructed containing proteins with assignments in both classification databases. To this end, all protein chains present in both `MP_SCOP` and `MP_CATH` were extracted yielding 96 chains (corresponding to 99 SCOP and 105 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_shared` dataset contained 63 protein chains corresponding to 64 SCOP and 67 CATH domains that share a sequence identity below 95%.

Using the fold classifications in the two datasets, `MP_SCOP` and `MP_CATH`, all SCOP and CATH folds containing α-helical membrane proteins were identified. The final set of α-helical membrane protein folds contained 34 SCOP (Table 7.1 in the Appendix) and 28 CATH folds (Table 7.2 in the Appendix).

2.2.2 Comparing domain assignments

Using the `MP_SCOP` and `MP_CATH` datasets, the number of single-domain and multi-domain membrane protein chains was calculated separately for both SCOP and CATH. In a second analysis, the domain assignments between SCOP and CATH were directly compared for all proteins in the `MP_shared` dataset. For those proteins where SCOP and CATH agreed in the number of domains, the extent of domain overlap was additionally calculated. This was done by calculating the fraction of residues consistently assigned

by both databases for a pair of corresponding domains. SCOP and CATH were said to agree regarding domain boundaries if this fraction was at least 90% of both individual domains.

2.2.3 Comparing fold assignments

For the analysis of similarities and differences regarding fold assignments of membrane proteins within SCOP and CATH, the `MP_shared` dataset was used. The agreement between fold assignments in SCOP and CATH was considered perfect if the following conditions were satisfied: (i) all proteins that were assigned to the same fold in one database were also assigned to a single fold in the other database, and (ii) no other proteins in the latter database had this fold assignment. If one of the two conditions for fold agreement was not fulfilled, the corresponding SCOP and CATH folds were added to the list of fold disagreements. Whereas fold disagreements of this kind were observed for protein chains classified as single-domain proteins by both databases, disagreements regarding fold assignments naturally resulted also from differences in domain assignments. Another kind of fold disagreement was thus given if PDB chains were classified as single-domain proteins into the same fold in one database, but were classified as two-domain proteins by the other database, with each domain having a separate fold assignment.

To compare the structural similarity of proteins involved in fold disagreements to those with consistent fold assignments, all-against-all protein structure comparisons were made using DaliLite v.3.1 [225]. For each comparison, the structural similarity Z-score and the root mean square deviation (RMSD) were obtained. Fold disagreements caused by domain discrepancies were not considered in this analysis. In case of fold agreements, SCOP domain coordinates were used for structure comparison because of their higher degree of manual curation. For those folds involved in disagreements, SCOP and CATH domain coordinates were used to represent SCOP and CATH folds, respectively. Only in one case (2atk, chain C), CATH domain coordinates were solely used for both the SCOP and CATH fold assignment as the SCOP domain did not cover the whole transmembrane region. As DaliLite did not return any result for several bitopic proteins, especially those with short sequence length, structure comparisons among bitopic proteins were additionally performed using the SSAP algorithm ([226], available via `http://www.cathdb.info/cgi-bin/SsapServer.pl`). The functional consistency of SCOP and CATH folds was furthermore evaluated using GO annotations [166, 180], as provided by the GOA group at the EBI (`http://www.ebi.ac.uk/QuickGO/`).

2.3 Results and Discussion

2.3.1 Membrane protein folds in SCOP and CATH

Membrane proteins with at least one transmembrane helix assigned by PDBTM (`MP_SCOP` and `MP_CATH` datasets, see Materials and Methods) are currently found in 34 SCOP (Table 7.1 in the Appendix) and 28 CATH (Table 7.2 in the Appendix) folds. In SCOP, membrane proteins are primarily classified within the 'membrane and cell surface proteins and peptides' class (f) while CATH does not provide a separate class for membrane proteins. Instead, α-helical membrane proteins are included within the mainly-α class (20 folds) together with α- helical soluble proteins and in the few secondary structures class (eight folds). Within the former class, 4 of the 20 folds containing membrane proteins belong to the orthogonal bundle architecture (CATH code 1.10) and 16 folds to the up-down bundle architecture (1.20). Generally, membrane proteins of the same fold are rarely further subdivided into superfamilies and families in both databases. Only 4 out of 28 CATH membrane protein folds (14.3%) are associated with more than one superfamily (Table 7.2 in the Appendix). In SCOP, three membrane protein folds are further subdivided into more than one superfamily, and only five folds contain more than one family (Table 7.1 in the Appendix), which corresponds to 8.8% and 14.7% of all SCOP membranous folds, respectively. In contrast, 13 and 38% of all globular folds (belonging to SCOP classes a, b, c or d) are associated with more than one superfamily and family, respectively. Not surprisingly, these numbers reflect the substantially higher structural coverage of soluble proteins compared to membrane proteins. Although the number of newly identified folds for soluble proteins is steadily decreasing [227], 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 single transmembrane helix, the two helix hairpin and the four helix bundle.

The number of distinct membrane protein domains that are assigned to one fold ranges from 1 to 30 (SCOP) and 1 to 26 (CATH) domains (Table 7.1 and Table 7.2 in the Appendix). In both SCOP and CATH, folds containing membrane proteins with one to four transmembrane helices represent the largest folds in terms of the number of distinct domains (SCOP: f.17, f.21, f.23; CATH: 1.10.287, 1.20.5, 1.20.85, 1.20.120). Generally, the collection of available membrane protein structures is still too small and biased [228, 229] to allow any conclusions about the most prevalent membrane protein folds in nature. However, it can be noted that the currently most populated folds all have

small numbers of transmembrane helices. This observation is compatible with genome scale analyses of membrane proteins where the fraction of proteins with a given number of transmembrane helices was found to decrease with increasing number of helices [4, 5, 230, 231] (see section 1.1.4, page 11). On the other hand, it might also indicate that proteins with fewer transmembrane helices are harder to classify into different folds because of more limited degree of structural variation (see also later), and that such proteins therefore tend to be assembled into few and larger folds.

Finally, the number of transmembrane helices was found to vary significantly within some folds, according to the annotation taken from PDBTM (Table 7.1 and Table 7.2 in the Appendix). 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 bc_1 complex chain c' fold (1.20.810), whose domains contain either four or eight transmembrane helices. The common classification seems to be caused by local structural similarity between the N-terminal part of cytochrome b and cytochrome b_6 domains [232], although cytochrome b consists of eight transmembrane helices, whereas cytochrome b_6 only has four. In contrast, SCOP splits cytochrome b proteins into two domains and hence, classifies only the N-terminal domain to the same fold as cytochrome b_6 proteins, whereas the C-terminal domain is classified separately.

2.3.2 Comparison of domain assignments

As domains are the basic units of protein structure classification in SCOP and CATH, the agreement in their assignments was analyzed first. SCOP and CATH use different methods to split proteins into domains. While SCOP essentially relies on visual inspection, CATH only employs manual annotation if three different automatic domain assignment methods do not yield a consistent consensus prediction [233]. Accordingly, previous analyses reported significant differences between SCOP and CATH in the resulting domain assignments, considering all proteins in the respective databases [207, 209]. However, as most of the membrane proteins are single-domain proteins [127], one would expect disagreements between domain assignments for membrane proteins to be less frequent than those observed for soluble proteins.

Of the 156 protein chains in the `MP_SCOP` dataset, 152 were classified as single-domain chains and four were split into two domains, while amongst the 110 protein chains from the `MP_CATH` dataset, nine contained two domains (Table 2.1). The observation that CATH identifies more multi-domain proteins than SCOP was already reported in the

Table 2.1: Comparison of α-helical membrane protein fold classification in SCOP and CATH

	SCOP	CATH
General treatment of membrane proteins	Separate class (membrane and cell surface proteins and peptides)	Membrane proteins are classified together with globular proteins
Number of folds	34	28
With more than one superfamily	3	4
With more than one family	5	-
General comparison using independent datasets (`MP_SCOP`, `MP_CATH`)[a,b]		
Number of protein chains with fold assignment	156	110
Domain assignments		
One domain per chain	152	101
Two domains per chain	4	9
Direct comparison using the shared dataset (`MP_shared`)[c]		
Domain assignments (SCOP:CATH)		
1:1	58	
2:1	1	
1:2	4	
Fold agreements	f.3 $\leftrightarrow$ 4.10.220	
	f.13 $\leftrightarrow$ 1.20.1070	
	f.19 $\leftrightarrow$ 1.20.1080	
	f.20 $\leftrightarrow$ 1.10.3080	
	f.24 $\leftrightarrow$ 1.20.210	
	f.29 $\leftrightarrow$ 1.20.1130	
	f.30 $\leftrightarrow$ 1.20.860	
	f.31 $\leftrightarrow$ 1.20.1240	
	f.33 $\leftrightarrow$ 1.20.1110	
Fold disagreement caused by domain disagreement	f.25 $\rightarrow$ 1.10.287 + 1.20.120	
	f.26 $\rightarrow$ 1.20.85 + 1.20.85	
	1.20.810 $\rightarrow$ f.21 + f.32	
Fold disagreement caused by fold overlap	f.14 $\rightarrow$ 1.10.287, 1.20.120	
	f.17 $\rightarrow$ 1.10.287, 1.20.20	
	f.21 $\rightarrow$ 1.20.810, 1.20.950, 1.20.1300	
	f.23 $\rightarrow$ 1.10.8, 1.10.442, 1.20.5, 4.10.49, 4.10.51, 4.10.81, 4.10.91, 4.10.93, 4.10.95, 4.10.540	
	1.10.287 $\rightarrow$ f.14, f.17	
	1.20.5 $\rightarrow$ f.23, j.35, j.37	
	1.20.120 $\rightarrow$ f.14, f.25, f.36	

[a] `MP_SCOP`: Set of PDBTM [26] proteins with an annotation in SCOP.
[b] `MP_CATH`: Set of PDBTM [26] proteins with an annotation in CATH.
[c] `MP_shared`: Set of PDBTM [26] proteins with an annotation in SCOP and CATH.

work of Hadley and Jones [207] and was found to be a direct result of the different domain definitions that are used in the two databases. CATH addresses geometrical aspects while SCOP also incorporates functional considerations.

To further elucidate differences in the domain assignments between SCOP and CATH, the separation into domains was examined using the dataset `MP_shared` containing 63 α–helical membrane proteins with one or more transmembrane helices found both in SCOP and CATH. In 58 cases, the two databases consistently assigned one domain per protein chain. However, this single domain did not always cover the entire protein chain, and in some cases, the sequence positions of domain boundaries differed between SCOP and CATH. Specifically, four cases were observed where SCOP and CATH deviated by more than 10% of their assigned positions, whereas in the remaining 54 cases (85.7% of all proteins in `MP_shared`) the domain assignments were consistent.

Five protein chains were divided into two domains either by SCOP or by CATH. Specifically, SCOP splits cytochrome *b* into two domains (first domain assigned to fold f.21 and second domain assigned to fold f.32, Figure 2.1A) while CATH classifies the full protein as one domain ('cytochrome bc1 complex fold', 1.20.810). Similarly, four cases were only found in CATH, where protein chains were separated into two domains, including two structures of subunit III of the cytochrome *c* oxidase. CATH splits these latter structures into a N-terminal domain with two transmembrane helices and a C-terminal domain with five helices based on the existence of a V-shaped cleft between the two helix bundles (Figure 2.1B) which is known to bind a lipid molecule [234]. The N-terminal domain is assigned to the 'helix hairpin fold' (1.10.287), and the C-terminal one to the 'four helix bundle fold' (1.20.120). In SCOP, the same proteins are classified as single-domain proteins to the 'cytochrome *c* oxidase subunit III like fold' (f.25). The other two two-domain protein chains correspond to the photosynthetic reaction center's L or M chains that have similar structures [235]. Both domains of these structures are assigned to the same CATH fold ('photosynthetic reaction center, subunit m, domain 1' (1.20.85)), whereas one domain spans two and the other three helices. In contrast, SCOP treats them as single-domain chains that belong to the 'bacterial photosystem II reaction centre L and M subunits fold' (f.26).

Finally, one amino acid chain (AcrB protein; 1iwg, chain A) is split into two domains by both SCOP and CATH. Although this protein is not yet officially classified in CATH and hence is not included in the `MP_shared` dataset, its CATH domain assignments are already available, defining six globular and two transmembrane domains identical to the domain assignment in SCOP.

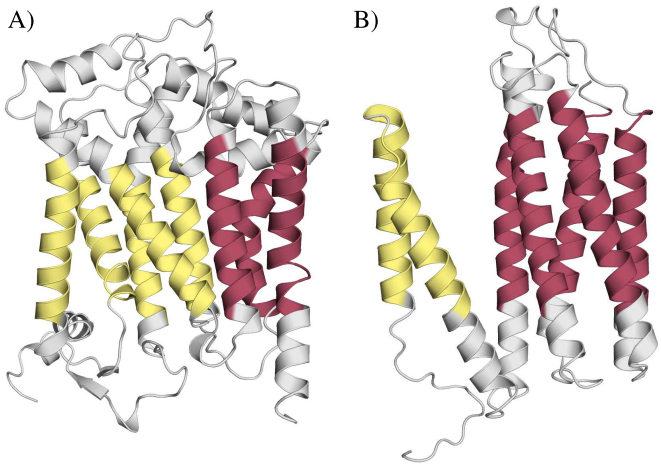

Figure 2.1: Domain assignment discrepancies between SCOP and CATH. (A) Mitochondrial cytochrome *b* subunit of the cytochrome bc_1 complex (PDB code: 1be3, chain C) is classified as a single-domain protein in CATH, but is divided into two domains in SCOP. (B) Mitochondrial cytochrome *c* oxidase, subunit III (PDB code: 1oco, chain C) constitutes a single domain in SCOP, but is separated into two domains in CATH. Two-domain assignments are indicated, with the transmembrane helices of each domain in a different color. Single-domain assignments encompass both colorings (yellow and red). Transmembrane helix coordinates were extracted from PDBTM [26]. The figure was drawn using PyMOL [66].

Summarizing, the fraction of membrane protein domains with consistent domain assignment (84.4 and 80.6% of all SCOP and CATH domains in `MP_shared`) is currently indeed slightly higher than the fraction of consistently assigned globular domains (69.3 and 67.9% for SCOP and CATH, respectively [209]). Such higher consistency for membrane proteins may be due to the fact that most of the membrane proteins with known structures are single-domain proteins. From all multi-domain membrane proteins found in SCOP or CATH (two and six proteins, respectively), only the AcrB protein was consistently assigned with two domains in SCOP and CATH indicating that the assignment of multi-domains *per se* is not easier for membrane protein than for soluble proteins.

2.3.3 Comparison of fold assignments - Fold agreements

To determine the agreement between SCOP and CATH with respect to their fold classification, the fold assignments of all proteins in the `MP_shared` dataset were compared. As `MP_shared` is a subset of `MP_SCOP` and `MP_CATH`, it does not cover all known α-helical membrane protein folds (Table 7.1 and Table 7.2 in the Appendix). Although the total number of membrane protein folds in SCOP is much higher than in CATH (34 folds

compared to 28, respectively), CATH classifies the proteins within the `MP_shared` set into more folds than SCOP (19 folds in SCOP and 26 in CATH, Table 7.1 and Table 7.2 in the Appendix).

Nine folds were found to contain exactly the same domains in SCOP and CATH (Table 2.1). In total, 21 chains, or 33.3% of the full `MP_shared` dataset, were assigned to these folds. All 21 proteins were classified as single-domain proteins by both databases with good positional agreement (only in one case the domains overlapped by less than 90%). SCOP and CATH even agree to a large extent regarding the names of these folds.

Considering only proteins from `MP_shared`, the nine folds identical between SCOP and CATH contained between one and six distinct domains. Six out of these nine cases affected proteins with six or more transmembrane helices. 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. In the remaining three cases (folds f.3/4.10.220, f.30/1.20.860 and f.31/1.20.1240 each consisting of a single protein chain), the number of transmembrane helices was one, two and three, respectively.

By comparing the structures of each fold using DaliLite, a high degree of structural similarity among all proteins of the same fold was observed. Average Z-scores for comparisons of the same fold (Table 2.2) varied between 23.9 (fold agreement f.13/1.20.1070) and 44.3 (fold agreement f.24/1.20.210). On the other hand, trying to align two domains covering both 10 transmembrane helices but classified consistently into two different folds (PDB code 2exw, chain A from fold f.20/1.10.3080 and PDB code 1iwo, chain A from fold f.33/1.20.1110) resulted in a maximal Z-score of 1.0. For comparison, a Z-score of 2.0 and higher was suggested by the authors of DaliLite to indicate a common fold [225], whereas a Z-score above 20 means that two structures are true homologues (see the DaliLite help file at `http://www.ebi.ac.uk/Tools/dalilite/`).

Table 2.2: All-against-all structure comparisons between membrane proteins with agreeing and disagreeing fold assignments in SCOP and CATH using DaliLite [225].

Folds	Number of proteins	Number of comparisons	Maximal Z-score	Minimal Z-score	Average Z-score
Fold agreements[a]					
f.3 ↔ 4.10.220	1	0	-	-	-
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.20 ↔ 1.10.3080	1	0	-	-	-
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
f.30 ↔ 1.20.860	1	0	-	-	-
f.31 ↔ 1.20.1240	1	0	-	-	-
f.33 ↔ 1.20.1110	1	0	-	-	-
Fold disagreements[b]					
f.14	2	1	3.4	3.4	3.4
f.17	4	6	9.6	2.2	4.8
f.21	7	18[c]	19.6	2.8	6.5
f.23	18	58[d]	3.4	0.1	2.2
f.25	3	3	31.1	20.6	24.3
f.36	4	6	20.9	17.7	19.4
1.10.287	6	13[e]	8.8	1.8	4.0
1.20.5	10	36[f]	3.4	0.1	2.0
1.20.120	8	28	27.6	3.8	10.4
1.20.810	2	1	17.0	17.0	17.0
1.20.950	2	1	6.6	6.6	6.6
1.20.1300	3	3	9.4	4.1	6.2
4.10.81	2	1	2.9	2.9	2.9

[a] Structure comparisons were executed using SCOP domain coordinates.
[b] Structure comparisons were executed using domain coordinates from the respective database (CATH coordinates for CATH folds and SCOP coordinates for SCOP folds).
[c] For three comparisons, DaliLite did not yield a result.
[d] For 95 comparisons, DaliLite did not yield a result.
[e] For two comparisons, DaliLite did not yield a result.
[f] For nine comparisons, DaliLite did not yield a result.

2.3.4 Comparison of fold assignments - Fold disagreements

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 and

Jones [207], and for SCOP and CATH, it arises from differences between the manual fold assignment within SCOP and the largely automatic approach based on automatic structure comparisons within CATH. While the discrepancies in domain assignments occurred three times (Table 2.1) and were already discussed earlier, seven cases of fold overlaps within the `MP_shared` dataset were observed. Remarkably, all seven fold overlaps involve only domains with one to five transmembrane helices, with single transmembrane helix, two helix hairpin, and four helix bundle domains being particularly strongly represented.

Case 1: 1.20.120 → f.14, f.25, f.36

The first case of fold overlap encompasses six distinct protein chains that correspond to one potassium channel (1ors, chain C), four different chains of acetylcholine receptor (1oed, chains B-E), and one ubiquinol oxidase subunit III (1fft, chain C). All proteins are assigned to the same CATH fold, the 'four helix bundle fold' (1.20.120), but to three different folds in SCOP. The potassium channel is assigned to the 'voltage-gated potassium channel fold' (f.14), the acetylcholine receptors to the 'neurotransmitter-gated ion channel transmembrane pore fold' (f.36), and ubiquinol oxidase subunit III to the 'cytochrome *c* oxidase subunit III-like fold' (f.25). Each of the four SCOP folds corresponds to a different CATH superfamily within the 'four helix bundle fold' (1.20.120). Interestingly, despite their common four helix bundle architecture, structural similarity among some of the six protein structures is remarkably low. The average Z-score of the six protein chains was only 10.4 (Table 2.2) corresponding to a clearly decreased average Z-score compared to those folds where SCOP and CATH agree in their classification. Protein chains such as 1oedB (acetylcholine receptor, beta chain) and 1fftC (ubiquinol oxidase) are both assigned to the four helix bundle fold (1.20.120), although the root mean square deviation (RMSD) is as high as 7.63 Å and their structure comparison score calculated by the SSAP algorithm [236] is below 70. The latter threshold is motivated by the CATH approach where two structures are assigned to the same fold if their SSAP score is greater than 70 [164]. As CATH uses the single-linkage clustering procedure [237] assigning a new protein to a given fold as soon as there is one member of the fold to which it is structurally similar, structures without significant similarity can end up in the same fold (this effect is also known as chaining). In fact, as just recently described by Pascual-Garcia and colleagues [212], the single-linkage clustering approach of CATH is one of the major sources resulting in differences to the SCOP classification system which instead applies the average-linkage procedure. As long as folds are structurally clearly distinct from each other, the impact of these clustering differences is expected

to be minimal. However, for more similar structures differences between the clustering mechanisms are likely to have a more prominent effect on the classification results, as seems to be the case for the membrane four helix bundle proteins attributed to CATH fold 1.20.120.

Case 2: f.14 → 1.10.287, 1.20.120

The second case of fold overlap involves two protein chains representing potassium channel (1ors, chain C and 2atk, chain C). In fact, these PDB chains represent different conformations of the same protein (1ors, chain C: opened conformation; 2atk, chain C: closed conformation) [238, 239] where the conformational change is known to cause significant structural differences [240] as can be seen in Figure 2.2A and Figure 2.2B. Both structures are assigned to the 'voltage-gated potassium channel fold' in SCOP (f.14). In contrast, CATH classifies each protein chain not only to different folds, but even to different architectures (2atk, chain C: orthogonal bundle architecture (1.10), helix hairpin fold (1.10.287); 1ors, chain C: up-down bundle architecture (1.20), four helix bundle fold (1.20.120)). Although representing the same protein, the 1ors (chain C) and 2atk (chain C) structures show only small similarity with a Z-score of 3.4 (Table 2.2). In this example, SCOP seems to strongly consider functional aspects over structural similarity to assign proteins to the same fold.

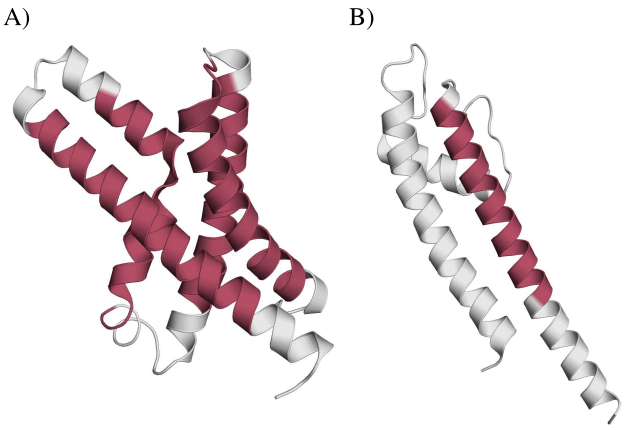

Figure 2.2: Potassium channels assigned to the voltage-gated potassium channel fold in SCOP (f.14) despite high structural diversity. (A) KvAP potassium channel voltage sensor (PDB code: 1ors, chain C). (B) KcsA potassium channel (PDB code: 2atk, chain C). The coordinates for the transmembrane helices of each domain (show in red) were extracted from PDBTM [26]. The figure was drawn using PyMOL [66].

Case 3: f.21 → 1.20.810, 1.20.950, 1.20.1300

The third case covers six distinct chains including one cytochrome b_6 (2d2c, chain A), three fumarate reductases (2bs2, chain C; 2b76, chains C and D), one succinate dehydrogenase (2acz, chain C) and one formate dehydrogenase (1kqf, chain C). As in the previous case, these proteins all belong to the same SCOP fold ('heme-binding four helical bundle' (f.21)), but to different CATH folds ('fumarate reductase cytochrome b subunit' (1.20.950); 'cytochrome bc1 complex, chain C' (1.20.810); 'three helical TM bundles of succinate and fumarate reductases' (1.20.1300)). Thereby, each of the four families of the SCOP fold f.21 corresponds to a different CATH superfamily. Again, some of the structures assigned to the four helical bundle fold in SCOP are very different with respect to the number of transmembrane helices, the loop lengths and helix tilts, with an average Z-score between them of only 6.5 (Table 2.2) and hence even smaller than the average Z-score of the CATH 'four helix bundle fold' (1.20.120) discussed above. In the particular case of the fold f.21, the affected proteins are still assigned to the same SCOP fold, since they all bind heme(s) and their overall structural similarity implies a common evolutionary origin, as suggested by Andreeva and colleagues [169]. On the other hand, CATH disregards functional aspects and identifies enough structural differences to assign these proteins to different folds.

Case 4: f.17 → 1.10.287, 1.20.20

While the previous cases of fold disagreements involved proteins with three to five transmembrane helices, two more disagreements were found for folds containing proteins with two transmembrane helices. Both cases involve one ubiquinol oxidase (1fft, chain B) and two cytochrome c oxidase subunits 2 (1ar1, chain B and 2dyr, chain B) that are all assigned to the 'transmembrane helix hairpin fold' in SCOP (f.17) and to the 'helix hairpin fold' in CATH (1.10.287). The two fold disagreements are caused by a fourth protein chain that is either assigned to another SCOP or another CATH fold. The CATH classification of the F_1F_O ATPase subunit c protein (1c0v, chain A) to the 'F_1F_O ATP synthase fold' (1.20.20) gives rise to the first fold overlap. Thus, compared to the other three protein chains, the fourth protein chain is not only assigned to a different CATH fold, but even to a different CATH architecture. Thereby, the two CATH folds correspond to different SCOP superfamilies of f.17. In line with all other cases of fold disagreement mentioned above, the structural similarity of the four protein chains is again rather low with an average Z-score of 4.8 (Table 2.2).

Case 5: 1.10.287 → f.14, f.17

The second disagreement between SCOP and CATH fold assignments caused by two helix hairpin proteins is similar to the previous one, but differs in that all four chains are assigned to the same CATH fold ('helix hairpin' (1.10.287)), but to different SCOP folds, whereas each SCOP fold corresponds to a different CATH superfamily of 1.10.287. The fourth protein chain (2atk, chain C) causing the discrepancy in the fold assignment is a voltage-gated potassium channel that is classified to the 'voltage-gated potassium channel fold' in SCOP (f.14). The structural similarity among the four structures is again remarkably low as indicated by the average Z-score of 4.0 (Table 2.2).

Case 6: f.23 → 1.10.8, 1.10.442, 1.20.5, 4.10.49, 4.10.51, 4.10.81, 4.10.91, 4.10.93, 4.10.95, 4.10.540

Finally, for proteins containing only a single transmembrane helix two cases of fold overlap were found as well. The first case covers 18 protein chains corresponding to eight cytochrome *c* oxidases (1m56, chain D; 2dyr, chains D, G and I-M), five chains of the cytochrome bc_1 complex (1be3, chains E and K; 1ezv, chain D; 2ibz, chains G and I), two subunits of the photosystem I (1jb0, chains F and J), one photosynthetic reaction center subunit H (1aig, chain H), one formate dehydrogenase subunit (1kqf, chain B) and one cytochrome *f* (1vf5, chain C). All protein chains are assigned to the 'single transmembrane helix fold' in SCOP (f.23). In contrast, CATH separates the same chains not only into ten different folds (1.10.8, 1.10.442, 1.20.5, 4.10.49, 4.10.51, 4.10.81, 4.10.91, 4.10.93, 4.10.95 and 4.10.540), but even into different architectures (orthogonal bundle, up-down bundle and irregular) and classes (mainly alpha and few secondary structures class). Except in one case (CATH classification 4.10.81.10 corresponds to SCOP superfamilies f.23.7 and f.23.18), all CATH superfamilies involved in this case of fold overlap coincide with distinct SCOP superfamilies of fold f.23. The main cause for this differential fold classification seem to be extramembraneous structural elements that are mostly ignored by SCOP while CATH produces a structurally more meaningful classification of bitopic proteins with single transmembrane helix domains being split into several folds depending on their globular portions (Figure 2.3).

Case 7: 1.20.5 → f.23, j.35, j.37

The second case of fold overlap involving single transmembrane helix proteins includes ten protein chains: five chains of the cytochrome bc_1 complex (1be3, chains E and K;

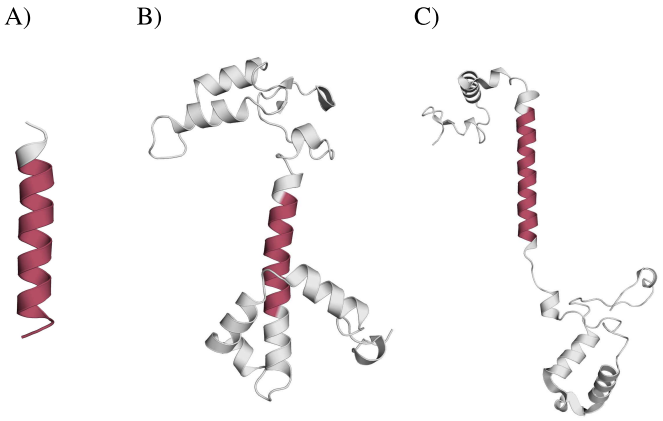

Figure 2.3: Common classification of bitopic membrane protein domains into SCOP fold f.23 despite different structural elements in their globular portions. In contrast, all three proteins are classified into different folds within CATH. (A) Subunit of cytochrome bc_1 (PDB code: 1be3, chain K). (B) Photosystem 1 reaction centre subunit 3 (PDB code: 1jb0, chain F). (C) Cytochrome *c* oxidase subunit 4 (PDB code: 2dyr, chain D). The coordinates for the transmembrane helices of each domain (shown in red) were extracted from PDBTM [26]. The figure was drawn using PyMOL [66].

1ezv, chain D; 2ibz, chains G and I), one cytochrome *c* oxidase (1m56, chain D), one cytochrome *f* (1vf5, chain C), one formate dehydrogenase subunit (1kqf, chain B), one phospholamban (1n7l, chain A) and one glycophorin A (1afo, chain A). Other than in the previous case, all these chains are classified to the same CATH fold ('single alpha-helices involved in coiled-coils or other helix-helix interface' (1.20.5)), but to three different SCOP folds (f.23, j.35, j.37) from two different SCOP classes. In total, eight of the ten protein chains are members of the 'single transmembrane helix fold' (f.23) in SCOP, and only two are classified to the few secondary structures class (1n7l, chain A and 1afo, chain A). Thereby, different CATH superfamilies of 1.20.5 correspond to different SCOP superfamilies. Generally, CATH classifies all domains having either no globular parts or only globular stretches without secondary structure within fold 1.20.5. The reason for the separate classification of proteins 1n7l (chain A) and 1afo (chain A) within SCOP is therefore unclear.

Reasons for fold overlaps

Further investigating the reasons for these fold overlaps, it must first be noted that no database is in general more structurally consistent than the other when it comes to the classification of membrane proteins with more than one transmembrane helix, as can

be seen from the average structural similarity of proteins classified to the same SCOP or CATH fold (Table 2.2). For example, proteins from CATH folds covering several SCOP folds (1.10.287 and 1.20.120) are on average as structurally similar to each other as proteins from SCOP folds covering several CATH folds (f.14, f.17 and f.21), with average Z-scores ranging between 3.4 and 10.4. Similarly, the classification of four helix bundle proteins, while approached completely differently in both databases, results in folds with average Z-scores between 6.6 (fold 1.20.950) and 17.0 (fold 1.20.810) in CATH and 6.5 (fold f.21) and 24.3 (fold f.25) in SCOP, demonstrating that no classification system groups together structurally more similar folds than the other.

This finding is surprising as the CATH classification system is exclusively based on structural considerations while SCOP also considers functional and evolutionary aspects [207]. Inspecting the observed cases of fold overlaps more closely (see above), it becomes apparent that functional considerations in fact explain some of the observed membrane protein fold overlaps but may lead to both structurally more diverse as well as more consistent folds in SCOP. For example, the two potassium channels classified together to SCOP fold f.14 (see above; Figure 2.2) share a rather low structural similarity (Z-score 3.4, Table 2.2). The opposite effect is observed in the case of CATH fold 1.20.120 (see above) covering proteins with only a weak structural similarity that are in fact not fulfilling the requirements of a common CATH classification (required SSAP score $>$ 70). Although these proteins end up in the same CATH fold due to the single-linkage clustering approach used during classification [237], SCOP places them in different folds (f.25 and f.36), which have homogenous functional GO assignments [166, 180] and high average Z-scores (24.3 for fold f.25 and 19.4 for fold f.36).

In contrast, functional considerations seem to be irrelevant for the classification of bitopic membrane protein domains (having only one transmembrane helix). DaliLite structure comparisons generally resulted in either very low Z-scores or no result at all for most pairwise comparisons of bitopic proteins because of their short sequence length (Table 2.2, folds 1.20.5, 4.10.81, and f.23). Therefore, additional comparisons using the SSAP algorithm [226] were conducted as well as manual structure inspection to analyze the structural consistency of bitopic membrane protein folds in SCOP and CATH. As was already described above, SCOP collects most single helix domains of transmembrane proteins within fold f.23 irrespective of the presence of any additional extramembraneous domains or their functional annotation (see Figure 2.3). While CATH classifies bitopic membrane domains into different folds (including 1.20.5) depending on their globular parts. Accordingly, SSAP structure comparisons of proteins from CATH fold 1.20.5

result in high SSAP scores (>80) and/or low RMSD values (< 3Å), while SCOP fold f.23 combines proteins with only intermediate SSAP scores (≤60) and high RMSD values (>15Å) (e.g., see Figure 2.3).

In general, SCOP comprises functionally more consistent folds as can be observed from available GO annotations for all affected proteins without being necessarily more or less structurally consistent. Although all CATH folds covering several SCOP folds contain proteins with completely different GO annotations, several SCOP folds combining proteins from different CATH folds, such as folds f.14 and f.21, share common GO annotations. Of course, functionally consistency is also not exhaustive in SCOP. Especially folds with only few transmembrane helices and a very general description (such as folds f.17 ('transmembrane helix hairpin') and f.23 ('single transmembrane helix')) aggregate proteins with inconsistent GO annotations.

Summary

From the comparative analysis of membrane proteins in SCOP and CATH it can be concluded that the currently available membrane protein structures with six and more helices are either very similar to each other with average DaliLite Z-scores ranging between 23.9 and 44.3 (and thus are classified to the same fold) or sufficiently diverse for SCOP and CATH to be able to consistently assign them to different folds. However, it must be noted that most of the folds within the `MP_shared` dataset differ in the number of transmembrane helices, facilitating the classification of the corresponding proteins. More than one fold is observed only in proteins with 10, 4, 2 and 1 helices. In the first case, structures of the Clc chloride channel (1ots, chain A) and the calcium ATPase (transmembrane domain M; 1xp5, chain A) are so dissimilar (DaliLite Z-score of 1.0) that both SCOP and CATH concordantly separate them into two different folds. In contrast, the classification of membrane proteins with one to five transmembrane helices seems to be more difficult as highlighted by the fact that almost no cases of fold agreement can be detected for these proteins at the moment. Only 3 out of 21 proteins for which SCOP and CATH agree in their fold classification have less than six transmembrane helices. Instead, all cases of fold overlap affecting α-helical membrane proteins involve proteins with one to five transmembrane segments.

This observation might have several reasons. One possibility is that larger, multi-helix proteins might contain more specific traits facilitating their separation into different folds. Recent publications have highlighted the presence of previously unseen structural features such as reentrant or tilted helices within membrane proteins (for a review, see

[35, 241]). According to the recent estimates, these features may occur quite frequently. For example, 10% of all polytopic membrane proteins are expected to contain reentrant regions [57]. Naturally, the possibility for structural modification and variation significantly increases with the number of available transmembrane helices. Accordingly, although proteins with less than six helices are still diverse enough for both CATH and SCOP to place them into several individual folds, their differences seem to be too subtle to be captured using the classic definition of a fold (primarily based on the number, connectivity and orientation of secondary structure elements), leading to largely deviating classifications within SCOP and CATH.

Another possible explanation for the discrepancy between the fold attribution in the two databases may lie in the potential adaptability of membrane-embedded proteins with only few transmembrane helices, which is possibly related to the evolutionary origin of primordial membrane proteins. Under the standard evolutionary model, with RNA and proteins preceding the emergence of cellular membranes [242], the problem of the first membrane proteins arises as a typical 'chicken and egg' paradox: while a lipid membrane would be useless without membrane transporting systems, the respective membrane proteins would need membranes to evolve. A plausible resolution of this paradox has recently been offered based on a combination of structural and phylogenetic analyses [243]. The suggested solution implies that the evolution of membrane proteins started from simple amphiphilic, α-helical hairpins capable of being incorporated into the membrane as oligomeric pores. This membrane architecture is retained by the membrane oligomer of c-subunits in the F_1F_O- type ATP synthase [244], where each hairpin is stabilized by interactions with its neighbors. Accordingly, the structure of simpler, two and four helix membrane proteins might be essentially dependent on the interaction with neighboring α-helices and, hence, depend on the partners in case of oligomeric structures. This variation could, at least partly, account for the discrepancy in fold attribution between databases.

2.4 Summary

- SCOP provides a separate class for membrane proteins; in CATH membrane proteins are classified together with soluble proteins
- SCOP contains 34 and CATH 28 membrane protein folds
- SCOP also considers functional aspects in fold classification → clearly different structures can be assigned to the same fold
- CATH uses single-linkage clustering → chaining effect can cause different structures to be assigned to the same fold
- Folds involved in fold agreements show much higher internal structural similarity than those folds involved in fold disagreements
- Reasonable agreement for domains with ≥ 6 TMHs
  - structure space seems to be discrete
  - classification similar to soluble proteins possible
- Many discrepancies for domains with 1-5 TMHs
  - structure space seems to be continuous
  - redefinition of fold necessary (more fine-grained structural features such as reentrant regions and helix-helix interactions)

2.5 Clarification of contribution

The comparative analysis of membrane protein classification in SCOP and CATH (as published in [245]) was carried out jointly with Angelika Fuchs. Angelika Fuchs accomplished an initial version of this comparison (for proteins with at least 3 transmembrane helices). In the final version of the comparative analysis she also performed studies on four-helix bundle proteins (results not shown here). I redid the analysis using an updated dataset that also contains membrane proteins with less than 3 transmembrane helices. The investigation of agreements and disagreements in the classifications was done by myself (such as the structural comparisons, literature research etc.).

Chapter 3

Classification of membrane proteins based on 1D and 2D structure

> "Commenting your code is like cleaning your bathroom - you never want to do it, but it really does create a more pleasant experience for you and your guests."
>
> *(Ryan Campbell)*

The comparative analysis of the structural classification of α-helical membrane proteins in SCOP and CATH (see Chapter 2, page 23) has shown that the globular fold definition is applicable only to a limited degree and thus should be adapted for membrane proteins by incorporating more fine-grained structural features. However, even if a new fold definition would be applied on membrane proteins in the context of structure-based classification, we still have to face the problem of the paucity of structural data. Therefore, such an approach only gives a limited view to the structural diversity of membrane proteins.

Given the current pace of membrane protein structure determination, the only option available today to explore the structural variety of the vast majority of membrane proteins is by sequence comparison and structure prediction. Existing classification systems for membrane proteins (such as GPCRDB [246] or TCDB [247]) operate at the sequence [248–251] and/or function level, but ignore topology. One method that also considers structural information was presented by Oberai *et al.* [250]. To organize the membrane sequence space into families, they developed an algorithm optimized for membrane proteins using sequence similarity and information about predicted TMHs. Using a dataset covering 95 genomes, they found 4,075 membrane protein families con-

taining sequences with at least two transmembrane helices (TMHs). Predicted TMHs served to define domain boundaries at the post-processing stage of the clustering approach. Another classification explicitly considering predicted structural organization is the CAMPS (Computational Analysis of the Membrane Protein Space) database that was introduced in 2006 [252]. The first version of the database, CAMPS 1.0, was based on sequence clustering and TMH prediction to identify structurally homogeneous clusters (SC-clusters) whose members are likely to share the same fold. Using a dataset of membrane proteins with at least three TMHs from 120 prokaryotic genomes, 266 SC-clusters were found, representing the estimate of the number of prokaryotic membrane protein folds at that time. The analysis was restricted to proteins with at least three TMHs, since most of the proteins with less helices are typically lipid-anchored proteins and integral membrane proteins were in the focus of that work. Furthermore, TMH prediction methods often falsely predict signal peptides as TMHs. Thus, using the cut-off of three TMHs minimizes the risk to include non-membrane proteins into the analysis considerably. In contrast to the approach of Oberai *et al.* structural information (in terms of predicted TMHs) directly affects the clustering in CAMPS such that clusters are required to show a certain degree of structural homogeneity.

In this chapter, a new version of the CAMPS database (CAMPS 2.0) is presented that features the following novelties. First, membrane proteins of eukaryotic and viral origin were incorporated into the classification process. Second, the membrane protein fold definition was revised by considering information about loop lengths in addition to sequence similarity and the number of TMHs. Finally, the empirically derived rules to derive structurally homogeneous clusters were replaced by a more sophisticated approach based on meta-models. The usage of meta-models (that are derived from so-called hidden Markov models) and the incorporation of loop lengths are two major modifications in CAMPS 2.0. Therefore, a short introduction is given at the beginning of this chapter. Subsequently, the new release is described and the results of comprehensive comparisons against sophisticated classification approaches are presented.

3.1 Introduction

3.1.1 Hidden Markov models

Hidden Markov models (HMMs) are statistical models that have been extensively applied in the field of speech recognition and computational biology for many different

problems including the prediction of genes [253], coiled coils [254], signal peptides [255] and transmembrane helices [98, 99] (see also section 1.1.3, page 8), as well as for sequence alignment and homology detection [256]. A short introduction will be given here according to [257] and [258]. For a more detailed description see [259].

Each HMM is a finite model composed of a set of states and a set of symbols. Thereby, each state emits symbols depending on its own emission probabilities and the states are connected by transition probabilities. In the example shown in Figure 3.1A, the HMM consists of two states and the symbols that each state can emit correspond to the nucleobases A, C, G and T. A HMM can be considered as a model that generates sequences. Therefore, a sequence of states (Figure 3.1B) is generated by moving from state to state according to the transition probabilities. As each state emits a symbol according to the emission probabilities, an observable sequence of symbols is generated as well (Figure 3.1C). In our example, the symbol sequence corresponds to a DNA sequence. The sequence of states is a so-called Markov chain. This is also a probabilistic model generating a sequence, in which the probability of a symbol only depends on the preceding symbol. As the state sequence is not observed and is hidden, it is called a hidden Markov chain, and therefore the whole model a hidden Markov model.

The usual scenario is that we are given a symbol sequence and we want to infer the hidden state sequence. For example, in the special case of transmembrane helix prediction, the symbol sequence corresponds to the amino acid sequence and the state sequence contains the information which residue belongs to a transmembrane helix or to a interhelical loop. It is possible that many state sequences can generate the given symbol sequence. In order to find the most probable one, algorithms such as the Viterbi algorithm can be used.

A very important step in generating HMMs is their training that defines the parameters of the model (i.e. the transition and emission probabilities etc.) that are unknown at the beginning. The standard training algorithms are the Baum-Welch and the Forward-Backward algorithms.

3.1.2 Importance of interhelical loop regions

Studies on α-helical membrane proteins mainly focus on their transmembrane domains. However, given that less than 30% of residues are within transmembrane domains [260], it is more than reasonable to also include interhelical regions in the structural and functional investigations of α-helical membrane proteins. In fact, the theory on how membrane protein folding occurs is switching from the 'two-stage' model [261] to the

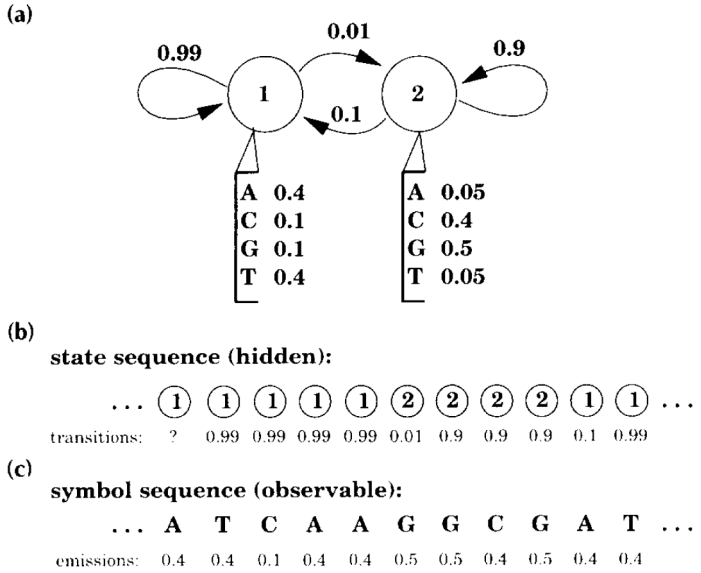

Figure 3.1: Example of a simple two-state hidden Markov model (HMM) describing DNA sequence. (A) State 1 and State 2 produce AT-rich and CG-rich sequences, respectively. Arrows indicate state transitions and their probabilities. Each state has its own emission probability distribution for generating a symbol (either A, C, G, or T; shown below the states). By moving from state to state, the HMM generates two sequences. A hidden state sequence (B) and a symbol sequence that is observable (C). Figure taken from [257].

'three-stage' model [262]. According to the first model, membrane protein folding is separated into two distinct stages. First, transmembrane helices are formed and inserted into the lipid bilayer. Second, the embedded transmembrane helices interact with each other to form a helix bundle. Based on experimental evidence, the second model suggests an additional third stage that also considers the binding of ligands and the folding of extramembraneous loops.

There is also evidence that some interhelical loops contribute to the function of α-helical membrane proteins [263–265]. This is also confirmed by two approaches demonstrating that loop length patterns can be used for functional classification [266, 267]. Otaki and Firestein applied their approach to the large superfamily of G-protein coupled receptors (GPCRs). It is known that all GPCRs share a seven transmembrane helix topology, but differ remarkably at the sequence level and are therefore further divided into families [134]. They have shown that loop length patterns are conserved among family members suggesting a functional significance in the GPCR superfamily [267]. Similar results were obtained by the study of Wistrand *et al.* [268].

Apart from the functional importance, there are some indications that interhelical

loops are also important for the structure of α-helical membrane proteins [269–271]. While the studies of Kim *et al.* and Landin *et al.* revealed that certain loops are essential for the stability of bacteriorhodopsin [269] and rhodopsin [270], Tastan *et al.* showed that interhelical loops affect the assembly of transmembrane helices as stretched loops constrain the distance between two adjacent helices.

Taken together, interhelical loops seems to play a meaningful role in the function and structure of α-helical membrane proteins. Hence, it seems promising to also consider loops in the structural classification of membrane proteins as an additional fold determinant. In fact, for soluble proteins it could already been shown that fold classification based on secondary structure can be further improved by adding loop information [272].

3.2 Materials and Methods

3.2.1 Dataset

A dataset of α-helical membrane protein sequences was created using the SIMAP [273] database that contains pre-calculated similarity scores between all publicly available amino acid sequences as well as annotation features. Only sequences from completely sequenced genomes with at least three predicted transmembrane helices (TMHs; according to Phobius [100]) were considered. Since SIMAP covers sequences from all genome sequencing projects, some genomes were represented several times. In such cases, the RefSeq [274] genome was chosen in order to avoid redundancy at the genome level. If a RefSeq genome was not available one genome instance was selected randomly. The final dataset of α-helical membrane proteins contained 494,679 sequences from 1,702 genomes, including 57 archaeal, 792 bacterial, 134 eukaryotic and 719 viral genomes. This dataset was used to set up the CAMPS database. It has to be noted that through the subsequent steps (initial clustering and SC-cluster derivation; see below), the number of sequences and genomes was further reduced (see Results and Discussion).

3.2.2 Analysis of domain content

For the analysis of domain content of α-helical membrane proteins a non-redundant dataset of 373,800 sequences was derived sharing no more than 90% sequence identity using the cd-hit algorithm [223]. For each sequence, Pfam-A [145, 146] domain annotations were obtained from SIMAP. Furthermore, each Pfam-A domain was classified as

either soluble (if it contained no predicted TMHs), transmembrane (if it contained predicted TMHs and soluble regions not longer than 120 residues), or hybrid (if it contained predicted TMHs and soluble regions longer than 120 residues).

3.2.3 Initial clustering

All sequences from the membrane protein dataset were initially clustered using the MCL [275] algorithm (version mcl-08-157) with an inflation value of 1.1 (default is 2.0) and a scheme of 7 (corresponds to default). The inflation value controls the granularity of the clustering (with higher values increasing the granularity). The smallest possible inflation value was used, since the purpose of the initial clustering was to obtain a very coarse grained clustering. The clustering was conducted based on pairwise FASTA [138] similarity scores between sequences that were extracted from the SIMAP database. Edges corresponding to alignments above a variable E-value threshold or not covering two or more TMHs and at least 40% of all predicted TMHs [252] were not considered during the clustering. Furthermore, the clustering was performed hierarchically at various E-value thresholds, starting at E-value threshold 1E-5 up to E-value 1E-100. The clusters obtained with the previous, less stringent E-value threshold were clustered independently of each other at the successive, more stringent threshold. Depending on the E-value threshold, the number of initial clusters ranged from 6,278 (1E-5) to 35,868 (1E-100) (for more information see Table 7.3 in the Appendix).

3.2.4 Determination of TMH core regions

For each initial cluster, a permissible range of TMH numbers was defined as described in Martin-Galiano and Frishman, 2006 [252]. Briefly, the most common number of TMHs among the cluster members was determined and some minor deviations were allowed (e.g. if almost all cluster members had four TMHs, the permissible TMH range was 3-5; see Martin-Galiano and Frishman, 2006 [252] for details). For each cluster, all those members were selected sharing less than 90% sequence identity (based on cd-hit [223] calculations) and whose TMH number was within the cluster's TMH range. If this procedure yielded at least 15 cluster members, they were multiply aligned using ClustalW [276] (which yielded better results than MUSCLE [277], data not shown). If the number of retained cluster members exceeded 400, only the 400 most divergent members were retained for alignment. Using these alignments, TMH cores corresponding to consecutive regions in the multiple alignment where 35% of the aligned sequences were

predicted to be within a TMH [252] were determined. If fewer than three TMH cores were found or the number of TMH cores was not within the TMH range, no TMH cores were assigned to the cluster. Regions outside of the TMH cores were considered to represent loop regions. Thus, each initial cluster was finally represented by a set of TMH cores and loops that was used for the generation of meta-models (see below).

3.2.5 Derivation of SC-clusters

Out of the set of initial membrane protein clusters those clusters were selected whose members are structurally homogeneous. This was done by incorporating the information on the number of TMHs and the length of extramembraneous loops into meta-models, as described below. Such clusters were called SC-clusters (structurally correlated) because the structural properties of their constituent proteins are similar.

Definition of meta-models

The intention is to combine topology, loop lengths and sequence similarity for the classification process since it is presumed that all these features determine a membrane protein fold. For this purpose meta-models were utilized that consist of a set of hidden Markov models (HMMs) each representing either a single TMH or a single loop region within a membrane protein. The individual submodels are connected to form one high-level model (Figure 3.2). Given a specific membrane protein family, a collection of HMMs representing loop and helix regions was shown to be able to detect remote members of this family. While the architecture of the meta-model developed by Wistrand *et al.* [268] was specifically adjusted to the family of G protein-coupled receptors, a family independent definition of a meta-model was developed capturing both sequence and topology information of any set of training sequences.

Submodel architecture

Different types of HMM architectures were utilized to model TMHs and loops. Using the set of TMH cores and loops inferred from the cluster alignments, HMMs were constructed as follows. A dedicated HMM was created for each TMH core, with the number of states corresponding to the length of the aligned core region. Depending on the number of gaps, the core region of an individual cluster member can also be shorter than the consensus length for all cluster members. Therefore, the helix model must be able to represent helices of different length simultaneously in order to account for structural variation in

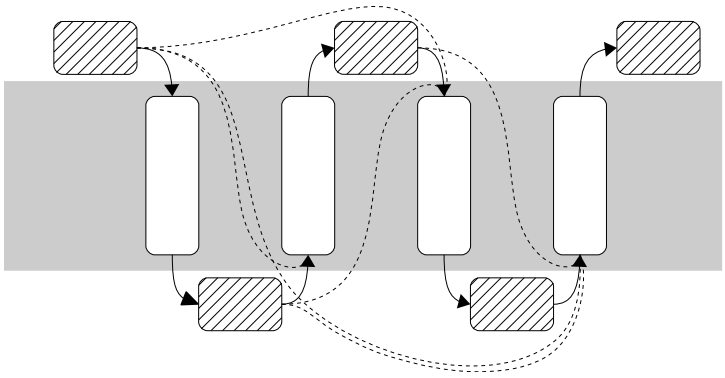

Figure 3.2: Meta-model architecture for a membrane protein with four TMHs. The boxes indicate the submodels of the meta-model, with hatched boxes representing loop regions and white boxes representing TMHs. The gray colored plane indicates the membrane bilayer. Solid arrows correspond to transitions between individual submodels and dotted arrows indicate transitions resulting from missing helices. Figure taken and adapted from [278].

the cluster. To this end transitions between a given state and all successive states were allowed by introducing silent states that emit no symbols, thus creating the possibility of skipping one or more states (Figure 3.3A).

Additionally, two different architectures were used to model the loop regions. The first model deals with loops longer than twelve residues and contains an adjustable number of states (depending on loop length conservation) and an additional globular state in the middle (Figure 3.3B). To determine the optimal number of states for this model, the mean loop length μ and its standard deviation σ were calculated. If the loop length was detected to be conserved ($\sigma < 2$) the number of states corresponding to the mean loop length was selected. If not, the number of states was set to twelve since the first and the last six amino acids of each loop were found to carry the most information, with additional states not resulting in a higher classification accuracy (data not shown). In order to allow longer loops to fit to the model as well, the globular state is able to perform self-transitions. On the other hand, the same model can also match shorter loops by using the transitions between normal and silent states. The second loop model was used for short loops with up to twelve residues (Figure 3.3C). In this model the number of states equals the maximally observed loop length within a given cluster. Again silent states allow for skipping individual states.

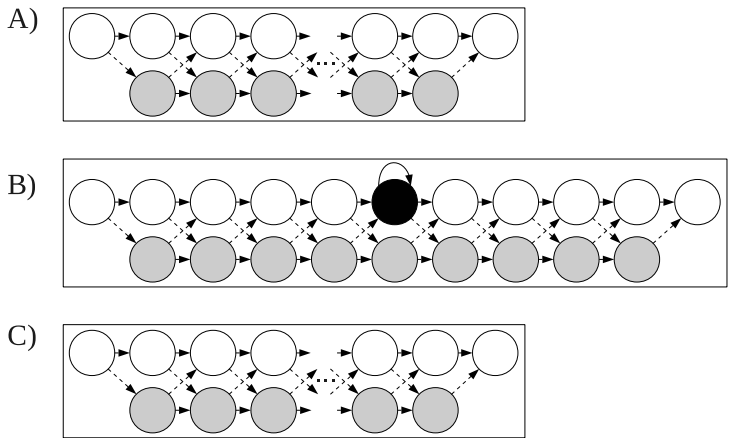

Figure 3.3: Helix and loop model architectures. (A) Helix model with normal states (white) and silent states (gray) that emit no symbols. Silent states allow transitions from one state to all successive states and are used to model helices of different lengths. Arrows denote possible transitions between two states. (B) Model for loops having a length greater than twelve amino acids. If the loop length is conserved within the cluster (standard deviation σ of mean loop length < 2), the number of states is set to the mean loop length, otherwise the model contains 12 states. In both cases, the model contains an additional globular state (black). (C) Model for loops having a length shorter than twelve amino acids. The number of states equals to the maximal loop length within the cluster. No globular state exists in this model. Figure taken and adapted from [278].

Model training

Before the model parameters were estimated by training they were initially set uniformly to 0.05 for the emission parameters (corresponding to 20 possible amino acids) and to $1/S$ for all transitions starting in node i, with S being the total number of successive states for node i. In addition to emission and transition probabilities, the probability of a sequence to start in state i was initialized by setting this parameter to 1 for the first state and to 0 for any other state. The actual parameter estimation for each HMM was executed using the iterative Baum-Welch algorithm from the Jahmm library v0.6.1 (`http://code.google.com/p/jahmm/`) with the default value of nine iterations as the stopping criterion.

After the separate training of each helix and loop model, the models were connected to each other to obtain one meta-model for the respective set of training sequences. To this end, transition probabilities between the individual submodels were introduced. Transition probabilities from the last state of a helix model to the first state in the following loop model were always set to 1. For the reverse case (loop-helix transition), the probability was set to the observed percentage of sequences containing the subsequent

helix in the training set. The remaining fraction of sequences missing this helix was used to model a transition skipping this helix.

Model optimization

After the training was finished, the models were best adapted to the training sequences and could not generalize to unseen data. To overcome this overfitting of meta-models, emission and transition probabilities were further processed after the training.

Transition probabilities needed to be adapted since the training may cause some transitions to be modeled with the probability of zero. To avoid this, predefined pseudocounts were added to allow transitions not seen in the training sequences with at least minimal probabilities by using the following equation:

$$\hat{a}_{ij} = \frac{a_{ij} + \eta}{1 + \eta\,|a_i|}$$

where $\hat{a}_{ij}$ and a_{ij} are the final and initial probabilities for the transition from state i to state j, η is a predefined pseudocount, and a_i is the number of transitions from state i. Based on preliminary optimization experiments, η was set to 0.003 (data not shown). These pseudocounts were not only used for transition probabilities within an HMM, but also for transition probabilities connecting two HMMs. Emission probabilities were adapted using Dirichlet mixtures which represent prior information about amino acid distributions observed in different sequence contexts such as positions with polar residues, small residues, and highly conserved residues. The combination of this information with the observed amino acid frequencies is expected to result in a greater generalization capability of the obtained model allowing the detection of remotely related sequences. The 9-component mixture estimated on the BLOCKS database [279] was used, which also includes a context for hydrophobic amino acids, and the adjusted emission probabilities were calculated as suggested by Sjölander *et al.* [280].

Generation of meta-models

In a second step an iterative procedure was developed using the generalized meta-model definition to detect from the set of initial sequence clusters those whose members are likely to share the same membrane protein fold (SC-clusters).

The derivation of SC-clusters is based on the idea that such clusters are supposed to contain proteins with sufficient sequence and structure similarity to be classified to

their originating cluster with high sensitivity and selectivity within a cross validation experiment. While this would also be true for protein families (as opposed to folds), the SC-cluster detection procedure was started with clusters generated at highly permissive E-value thresholds and then E-value thresholds were gradually made more and more stringent, accepting any cluster as soon as the pre-defined criteria for successful cross validation were met. For several protein families sharing the same fold this should be the case at an E-value where the tested sequence cluster still covers all of these families. The procedure is thus based on the notion that a structural fold is a broader concept than a functional family. The exact steps of the iterative SC-cluster detection procedure are as follows (Figure 3.4):

Step 1: Start with all initial clusters at the least restrictive E-value threshold (1E-5) as the training set.
Step 2: Build a candidate meta-model for each cluster in the training set.
Step 3: Classify all members of the clusters in the training set using 10-fold cross validation against all meta-models in the training set ('test step').
Step 4: Remove clusters with less than 95% correctly classified members and replace them with their child clusters at the next more restrictive E-value threshold. The training set now includes these newly added clusters as well as all the clusters accepted at earlier cross validation rounds.
Step 5: Go to Step 2 and iterate until no cluster is left to be split or until the most restrictive E-value threshold has been reached (1E-100).

The clusters resulting from this iterative procedure are assumed to represent structural membrane protein families whose members are likely to share the same fold and are termed SC-clusters. A short description was derived for each SC-cluster using the Pfam-A [145, 146] and SUPERFAMILY [281] assignments (as extracted from SIMAP [273]) of their members. The assignment that could be found for at least half of all members was inherited as the SC-cluster description. If no such assignment was available, the SC-cluster was termed 'Uncharacterized SC-cluster'.

Classification of individual sequences

In the third step of the SC-cluster detection procedure, all cluster sequences were classified against a set of meta-models. A scoring mechanism was developed to identify the best fitting meta-model. Every sequence was compared to each meta-model by scoring all possible paths using the forward algorithm implemented in the Jahmm library.

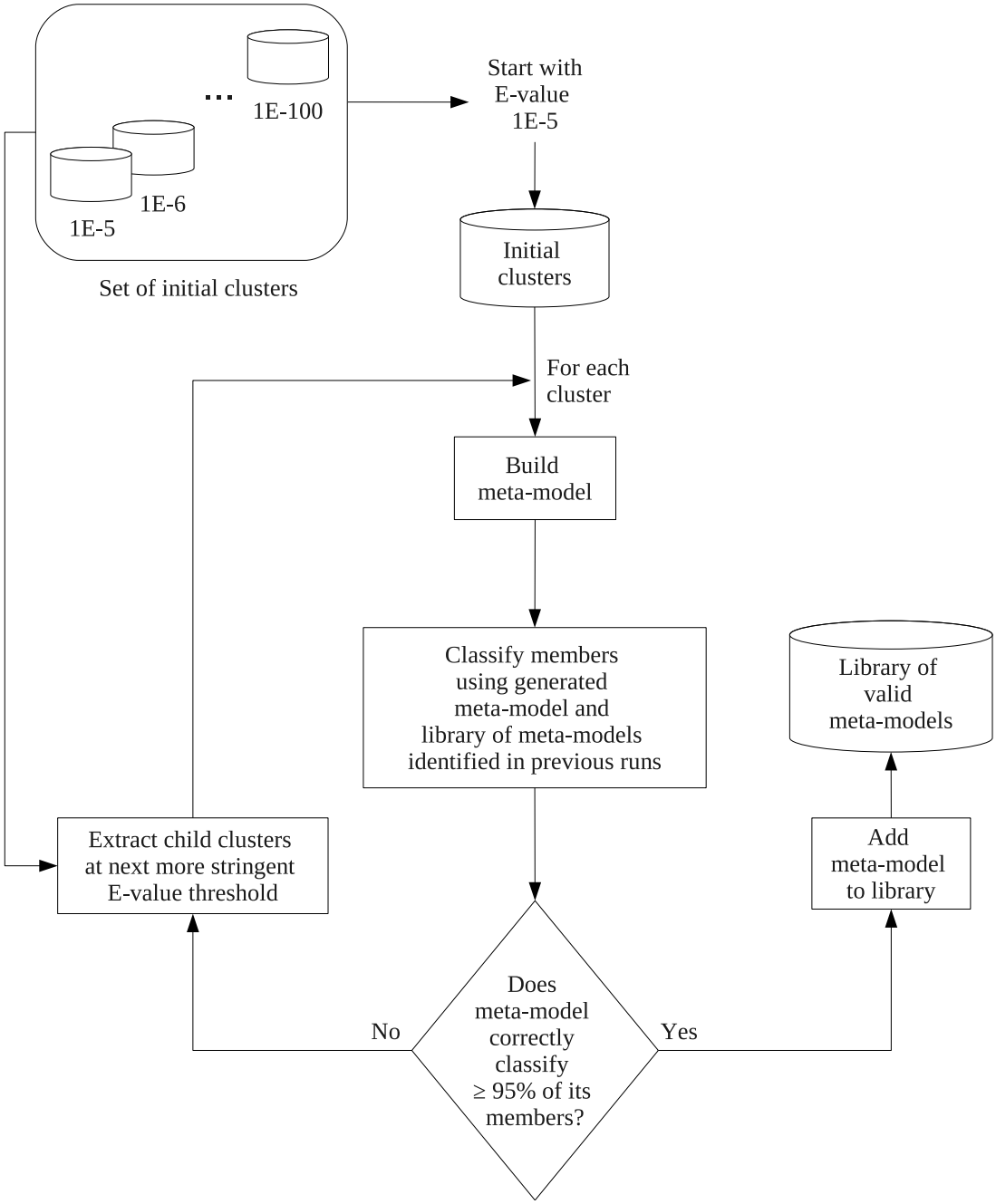

Figure 3.4: Flowchart of the procedure to identify SC-clusters using meta-models.

Given a meta-model and a sequence, the forward algorithm calculates a value between 0 and 1 indicating the probability that the meta-model generated the sequence. The final score was calculated as the natural logarithm of that probability. For those cases where the best score was below the minimal score threshold, implying that even the

best meta-model gave no score reliable enough for classification, the sequence was classified as 'unknown'. Otherwise, the sequence was assigned to the highest scoring model. The optimal value for the minimal score threshold was determined in preliminary optimization experiments with the goal to obtain the best trade-off between sensitivity (fraction of classified cluster members out of all sequences in the cluster) and accuracy (fraction of correctly classified cluster members within the set of all classified members of the cluster) (data not shown). The classification method is used in almost the same manner on the project website to classify an unknown query sequence submitted by the user. The only difference is that before the query sequence is checked against the meta-models, TMHs are predicted using Phobius [100]. Then, the sequence will only be compared to those meta-models, where the number of predicted TMHs is within the TMH range of the underlying SC-cluster. This step was implemented to speed up the classification of unknown query sequences. Please note that the forward algorithm used in this work calculates the probability that the entire query sequence was generated by a given meta-model (which is comparable to a global alignment), leading to poor scores when multi-domain membrane proteins are compared with a meta-model built from single-domain proteins. In a future CAMPS release, it is intended to incorporate an improved algorithm that will be able to perform both global and local alignments with respect to the sequence.

In order to filter unreliable matches, a score cut-off was determined for each meta-model such that any match scoring above this cut-off is very likely to be a true positive. The scores for each SC-cluster sequence against all meta-models (all-against-all comparison) were calculated yielding a list of scores of all members and a list of scores of all non-members for each meta-model. Since each meta-model represents a SC-cluster, members of a meta-model were defined as those SC-cluster sequences that were assigned to this SC-cluster (see above) and non-members as those sequences that were assigned to one of the other SC-clusters. For each meta-model, different score cut-offs were tested and the sensitivity (fraction of all members that score above the cut- off) and specificity (fraction of all non-members that score below the cut-off) was calculated. Then, the score cut-off achieving at least 90% specificity was chosen. The mean sensitivity over all defined meta-model specific score cut-offs was 96.7% and the mean specificity was 95.7%. These cut-offs are used to classify an unknown sequence such that only those matches are returned that score above the respective cut-offs. Since these cut-offs do not always fully exclude false positives, the uncertainty of each match is quantified by reporting a P-value and a Z-score. The P-value is a probability and is calculated as

the fraction of non-members with a score at least as good as the one obtained for the match. The Z-score is defined as the difference of the score obtained for the match and the mean score of all non-members divided by the standard deviation of the scores of all non-members.

3.2.6 Derivation of FH- and MD-clusters

Each SC-cluster was further subdivided into two types of subclusters: functionally homogeneous (FH-clusters) and modeling distance clusters (MD-clusters). For the identification of FH-clusters, each SC-cluster was split based on the Pfam [145, 146] domain architecture (i.e. the sequential order of domains) of its cluster members. SIMAP already contains an integrated clustering that is based on sequence similarity and protein domain architecture [282]. Briefly, sequences were clustered using the MCL algorithm [275] and a measure for assessing the similarity of two domain architectures as described in Lin *et al.* [283]. This measure combines three indices, the Jaccard index (which measures the number of shared domains), the Goodman-Kruskal γ function (which measures the similarity of shared domain arrangement) and the domain duplicate index (which measures the similarity of domain duplications). The domain architecture cluster assignment was extracted for all members of the respective SC-cluster from SIMAP and members with the same assignment were merged, yielding FH-clusters whose members have the same or similar domain architecture and are thus likely to have the same function [284]. Finally, a representative domain architecture was determined for each FH-cluster by selecting the most common architecture among its members.

To derive the second type of subclusters (MD-clusters) at least 70% of all members were required to share a sequence identity of at least 30%. This condition was first evaluated for the entire SC-cluster and if it was met then all members of the SC-cluster were assigned to one MD-cluster. Otherwise the subclusters at the next more stringent E-value cut-off (determined by the initial clustering step; see above) were considered and the conditions were checked again. This procedure was repeated until the requirements were fulfilled or no subclusters were left. A short description was defined for each FH- and MD-cluster in the same way as it was done for SC-clusters (see above).

3.2.7 Comparing SC-clusters with other databases

To analyze the relationships between CAMPS SC-clusters and Pfam [145, 146] (release 24.0) families and clans, the underlying sequence database of Pfam was used, called

pfamseq containing 9,421,015 sequences from UniProt [139] (release 15.6). Sequences having both a Pfam and a SC-cluster assignment were identified using md5 checksums, yielding 421,422 common sequences. For each sequence of this dataset the domain organization was extracted from the swisspfam file which is released with Pfam (and contains Pfam domain annotations for SwissProt entries). Sequences having more than one annotated Pfam domain and those where a Pfam domain covered less than 90% of the complete sequence were excluded from consideration. If available, Pfam clan assignments for each domain family were extracted from the Pfam-C file. The final dataset included 94,337 sequences having both CAMPS SC-cluster and Pfam family assignments and 56,756 sequences having both CAMPS SC-cluster and Pfam clan assignments. The two databases were then compared in both directions by analyzing Pfam family and clan assignments for sequences having the same CAMPS SC-cluster assignment (forward comparison) as well as SC-cluster assignments for sequences having the same family and clan assignment (reverse comparison).

Similarly, the relationship between the TCDB [247] and CAMPS databases was investigated. The FASTA sequence file from TCDB (downloaded from `http://www.tcdb.org` as of September 13, 2010) was used including 5,835 sequences. The sequence mapping between TCDB and CAMPS was performed using md5 checksums. 2,085 and 647 sequences with SC-cluster assignments were also found to possess TCDB family and superfamily assignments, respectively. TCDB and CAMPS were then compared in the same way as Pfam and CAMPS (see above).

For the comparison of CAMPS with SCOP [162, 163] (v1.75) and CATH [164, 165] (v3.3) all 3,837 redundant α-helical transmembrane proteins from the PDBTM [26] database (v2.3; September 13, 2010) were used. For the subset of 921 non-identical proteins, the mapping between CAMPS and PDBTM was performed using BLAST [137] (version 2.2.19) and a 90% sequence identity threshold yielding a common set of 146 sequences. There are two reasons for such limited overlap. First, only membrane proteins with at least 3 TMHs are included in CAMPS while the set of 3,837 PDBTM proteins also contains proteins with fewer TMHs. Second, CAMPS covers only completely sequenced genomes while PDBTM also contains structures from partially sequenced genomes. After excluding sequences with more than one SCOP or CATH transmembrane domain, datasets of 54 CAMPS sequences also having SCOP and CATH annotations, respectively, were collected. Again, the databases were compared in both directions.

For all four comparisons, the median sequence identity of agreements and disagree-

ments was calculated using the respective sequences from the reference group (i.e. if the Pfam family distribution within each SC-cluster was analyzed, the reference group was the particular SC-cluster and if the distribution of SC-cluster assignments was investigated within each Pfam family, the reference group was the particular Pfam family). The pairwise sequence identities were obtained from SIMAP and the median over all pairwise values over all cases of agreement and disagreements, respectively, was determined. For those cases where CAMPS did not agree with CATH or SCOP, SIMAP did not report sequence identity values (indicating that the values are very low and thus are not contained in SIMAP). For these sequence pairs, the sequence identity was calculated using the FASTA [138] algorithm. Similarly, the median standard deviation of the number of TMHs was calculated based on Phobius [100] predictions. In case of CATH and SCOP, the structural similarity of proteins involved in agreements and disagreements was also calculated using DaliLite [225]. The Z-score for each pairwise comparison was obtained and the median value was defined in the same way as for sequence identity.

3.2.8 Comparing FH-clusters with ENZYME

To compare FH-clusters with the ENZYME nomenclature database [181, 182] the EC (Enzyme Commission [179]) annotations were extracted for the proteins listed in the enzyme.dat file (downloaded from `http://www.expasy.org`; release of July 13, 2010) that could also be found in the CAMPS FH-clusters using the SwissProt entry name as the mapping criterion. All proteins having more than one EC annotation were not considered. The final dataset comprised 4,509 proteins covering 106 EC codes and 127 FH-clusters. EC codes and FH-clusters were compared in both directions as before.

3.2.9 Mapping to external databases

In order to exhibit relationships to other external databases membrane protein sequences from the SC-, FH- and MD-clusters were searched using BLAST [137] (version 2.2.19) against the following databases: DrugBank [285] (version 2.5), GPCRDB [246] (version 9.9.1), MPtopo [286], OMIM [287], OPM [288], PDBTM [26] (version 2.3), TargetDB [289], TCDB [247], TOPDB [290] and VFDB [291]. Only those matches with a sequence identity of at least 40% and an E-value better or equal to 1E-3 were considered. Furthermore, it was required that at least two TMHs and at least 40% of all TMHs of the CAMPS sequence are covered by the alignment.

3.2.10 Availability

CAMPS 2.0 is available at `http://webclu.bio.wzw.tum.de/CAMPS2.0/`.

3.3 Results and Discussion

3.3.1 Modifications and improvements in CAMPS 2.0

The new release of CAMPS (CAMPS 2.0) contains substantial changes compared to the previous database version (CAMPS 1.0), summarized in Table 3.1. An overview of the whole method is given in Figure 3.5. One of the major changes is the inclusion of membrane proteins from eukaryotic and viral genomes. While the first release was built on membrane proteins from 120 genomes of prokaryotic origin [252], CAMPS 2.0 covers 1,253 genomes in total (849 prokaryotic, 134 eukaryotic, 270 viral genomes). This 10-fold increase in the number of genomes comes is accompanied with an almost 16-fold increase in the number of unique membrane protein sequences assigned to SC-clusters (26,360 versus 413,714 sequences). Several methods used in the classification pipeline were also replaced by more sophisticated ones. In CAMPS 1.0, the TMH predictions from TMHMM [98] were used. One drawback of this prediction method is that signal peptides are often falsely predicted as TMHs. To circumvent this problem the new release uses predictions from Phobius [100], a tool combining TMH and signal peptide predictions. Furthermore, single linkage clustering was replaced by Markov clustering [275] for generating the initial set of membrane protein clusters. The latter approach is much better suited for dealing with large sequence collections and with multi-domain proteins.

Besides the incorporation of eukaryotic membrane proteins, additional structural features of membrane proteins were included into the process of detecting SC-clusters whose members are likely to have the same fold. In CAMPS 2.0 not only sequence similarity and the number of TMHs are used as fold determinants, but also information about loop lengths. They contain valuable biological information for classifying proteins [266, 267, 272, 292] and are important for structural assembly of the TMHs [271]. Together with the enhancement of fold determinants the process of determining SC-clusters was also changed. In the original version SC-clusters were found based on empirically derived quality criteria applied to sequence similarity and the conservation of TMH number within the cluster. In the new release these rules were replaced by so-called meta-models that are particularly useful for the detection of remote homologues.

Table 3.1: Differences between the first (CAMPS 1.0) and the second (CAMPS 2.0) CAMPS release

	First release	Second release
Dataset[a]		
Sequences	26,360	413,714
Genomes	120	1,253
Archaea	15	57
Bacteria	105	792
Eukarya	-	134
Viruses	-	270
Transmembrane helix (TMH) prediction	TMHMM	Phobius
Initial clustering	Single-linkage clustering	Markov clustering
Fold determinants	Sequence similarity Number of TMHs Loop lengths	
SC-cluster generation	Empirical rules	Meta-models
Subclustering of SC-clusters	OH-clusters: 95% of all members have same COG assignment	FH-clusters: all members have same domain architecture cluster assignment
	MD-clusters: 70% of all members share at least 30% sequence identity	
Linked databases	COG	Pfam, PDB, TCDB eggNOG, GenBank, UniProt, ENZYME, GO, SUPERFAMILY, DrugBank, GPCRDB, MPtopo, OMIM, OPM, TargetDB, TOPDB, VFDB

[a] The numbers given correspond to those sequences and genomes that are covered by the SC-, FH-, and MD-clusters.

Meta-models are higher-order Markov models that are composed of multiple HMMs that represent either a TMH or a loop region. For example, if the membrane proteins of the underlying cluster contain four TMHs, the respective meta-model contains nine HMMs in total (four for the TMHs and five for the loops) that are connected in the correct order (see Materials and Methods). The advantage of using meta-models is that sequence information can be easily combined with topology information (i.e. the number of TMHs and loops). The concept of meta-models was previously employed by Möller

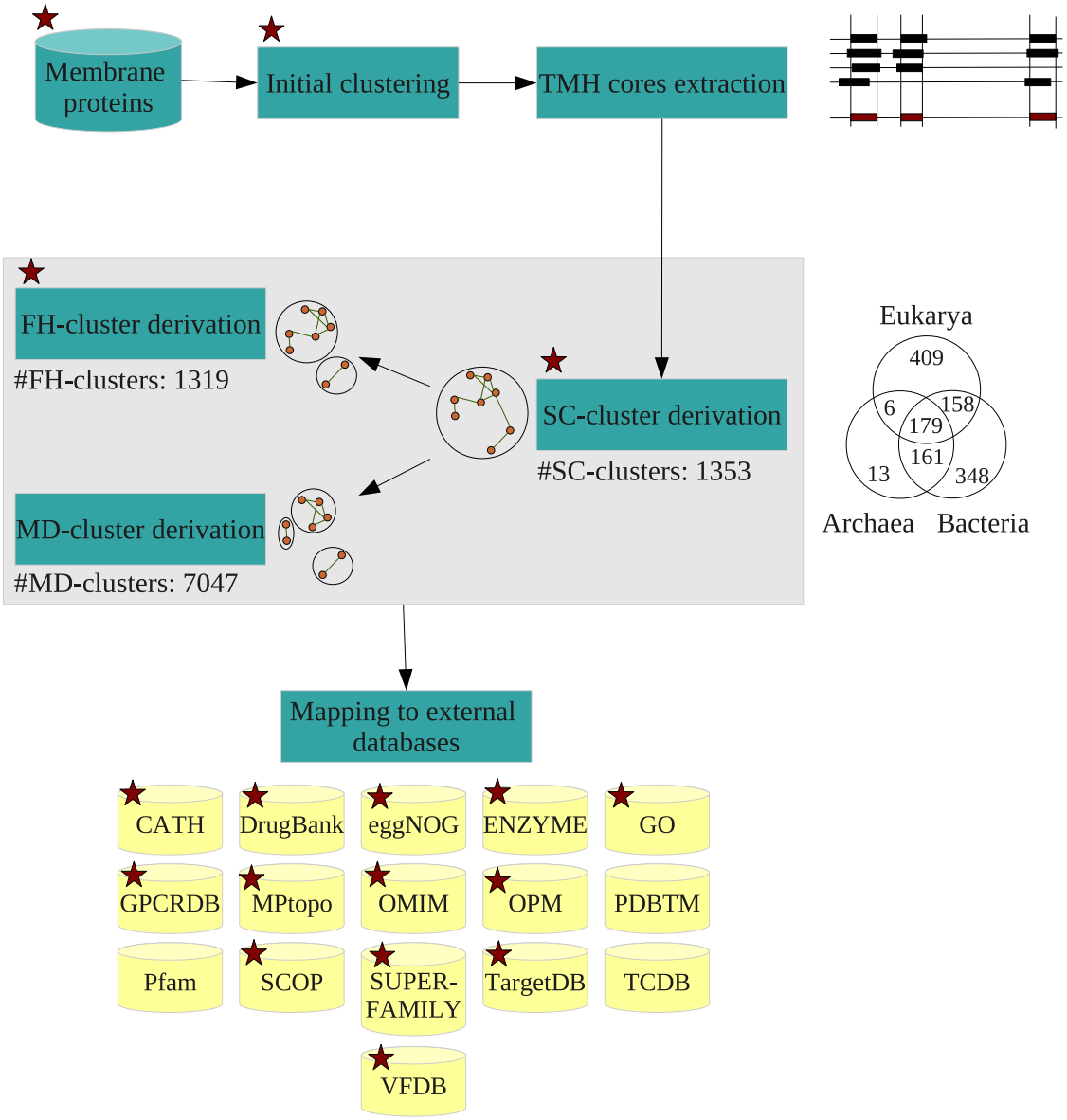

Figure 3.5: Pipeline of the CAMPS 2.0 database release. Elements of the pipeline added or changed in the new release are marked with stars.

and colleagues in their tool 7TMHMM that predicts the specificity of G-protein coupled receptors (GPCRs) [293]. Since 7TMHMM can not identify new GPCRs another tool named GPCRHMM was proposed that is specifically trained for the recognition of GPCRs lacking sequence similarities to known sequences [268]. In this work the idea of meta-models was adapted to detect clusters of membrane proteins that are structurally homogeneous and are likely to share the same fold (SC-clusters).

Just like in CAMPS 1.0, SC-clusters were further divided into two types of subclusters.

The first type is called MD-clusters (modeling distance), grouping membrane proteins having at least 30% sequence identity. The second type, designed to approach the level of protein function, was called OH-clusters (ortholog homogeneous) in CAMPS 1.0 and contained proteins with the same COG [147, 148] assignment. Since orthologous group assignments are not available for many eukaryotic proteins (e.g. the eggNOG [294] database currently covers 575 prokaryotic, but only 55 eukaryotic genomes), a different technique to derive functionally homogeneous subclusters was used in CAMPS 2.0. This technique is based on the idea that proteins sharing the same domain architecture tend to have the same function [284]. The DODO tool uses this idea to detect orthologous proteins based on domain architectures [295]. Using sequence clusters based on the similarity of domain architectures available from the SIMAP [273] database a subcluster was considered as an FH-cluster if all its members had the same domain architecture cluster assignment.

Finally, the links to external databases were further extended. Initially, CAMPS was linked to four databases: COG [147, 148], Pfam [145, 146], PDB [27], and TCDB [247]. CAMPS 2.0 is additionally linked to sequence databases (GenBank [296], UniProt [139]), function databases (ENZYME [181, 182], GO [166, 180]), family databases (SUPERFAMILY [281]), membrane protein resources (GPCRDB [246], MPtopo [286], OPM [288], TOPDB [290]), structural genomics repositories (TargetDB [289]) and biomedical databases (DrugBank [285], OMIM [287], VFDB [291]).

3.3.2 Empirical rules versus meta-models

In CAMPS 1.0 [252] empirically derived rules were used to identify SC-clusters by imposing constraints on their structural (in terms of TMHs) and sequence variability. In this work, these rules were replaced by meta-models and an advanced fold definition was used that also incorporates loop lengths. In order to investigate whether these changes improved fold classification of membrane proteins the new classification procedure was applied to the CAMPS 1.0 dataset. For a reasonable comparison, the restriction to the CAMPS 1.0 dataset was necessary as the empirical rules were initially developed specifically for prokaryotic membrane proteins and a revision of these rules would have been necessary if they had to be applied to eukaryotic proteins as well. Since meta-models are trained with at least 15 sequences having TMH core annotations (see Materials and Methods), only those SC-clusters derived by meta-models were compared to the original empirically derived SC-clusters also having at least 15 members with core annotations.

The first observed difference was the number of recognized SC-clusters - 233 SC-

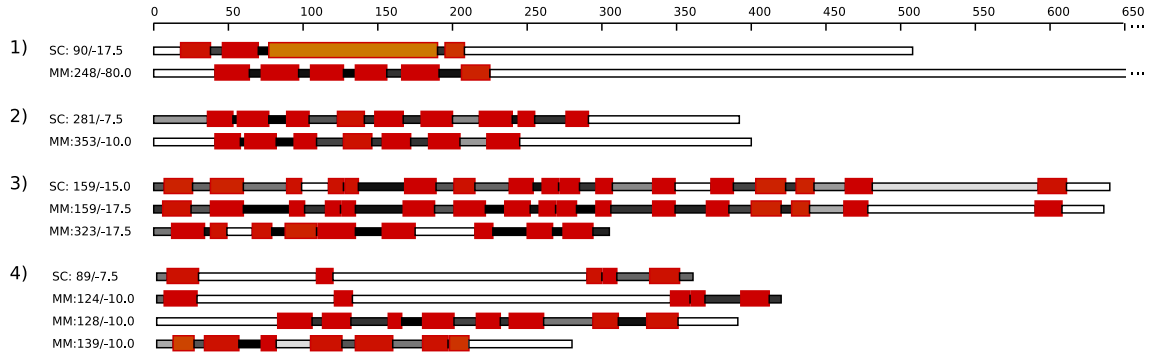

Figure 3.6: Topology diagram of rule-based and meta-model based SC-clusters. Red colored boxes correspond to transmembrane helices, gray colored boxes to loops. The darker the color, the more conserved is the region. Empirical rule-based SC-clusters (upper lines) are marked with the prefix 'SC', meta-model based SC-clusters (lower lines) with the prefix 'MM'. 1)+2) One rule-based SC-cluster conforms to one meta-model based SC-cluster differing in their E-value thresholds and the number of transmembrane helices. 3)+4) One rule-based SC-cluster is associated with more than one meta-model based SC-cluster. Figure taken and adapted from [278].

clusters based on empirically derived rules versus 249 SC-clusters based on meta-models. Secondly, the number of proteins covered by meta-model based SC-clusters (15,016) was comparable to that covered by rule based SC-clusters (15,497). Thus, while the number of covered proteins remained practically unchanged, the new approach delivered slightly more SC-clusters. This can be explained by the incorporation of the loop lengths in the meta-model approach presented here. Using these additional criteria, members of a SC-cluster must not only possess similar sequences and a similar number of TMHs, but also display homogeneity in terms of their loop length patterns, resulting in clusters of smaller size. Furthermore, membrane proteins can have very similar architectures, but vary widely in their loop lengths as was already shown for the GPCRs [267, 292]. In this case, it is necessary to split the underlying cluster in order to meet all requirements of the advanced fold definition. Figure 3.6 displays the topology of representative cases where membrane proteins of rule based SC-clusters were classified differently using meta-models. In ten cases, one rule based SC-cluster corresponded to one meta-model based SC-cluster (see examples 1 and 2 in Figure 3.6) differing in their E-value thresholds, and in seven cases they also differed in terms of the number of TMHs. The third type of correspondence was the partitioning of a rule-based SC-cluster into two or more meta-model based SC-clusters (see examples 3 and 4 in Figure 3.6), which occurred six times. In all six cases it was found that the meta-model based SC-clusters matching one rule-based SC-cluster had completely different topologies indicating different folds.

Altogether, in almost every case where there was no one-to-one correspondence between rule based SC-clusters and meta-model based SC-clusters, loop lengths and the number of TMHs were significantly better conserved within the meta-model based SC-clusters. The conservation of loop lengths is secured by the usage of the meta-models and since they are modeled explicitly this conservation should improve automatically. The standard deviation of loop lengths between two TMHs went down from 3.88 for rule based SC-clusters to 3.55 for meta-model based SC-clusters. Overall, these results demonstrate better performance of the meta-model based procedure (caused by the incorporation of loop lengths that affect the assembly of the TMHs [271]) in finding clusters that are assumed to represent membrane protein folds.

3.3.3 Domain content of membrane proteins

Expansion of the CAMPS database and, in particular, incorporation of eukaryotic genomes raised the question to which extent the well-known difficulties in clustering globular proteins caused by multi-domain proteins (promiscuous domains etc.) also apply to membrane proteins. In fact, very little is known about the domain structure of membrane proteins. A previous study showed that most of the membrane proteins are single-domain proteins, with eukaryotic proteins having a higher incidence of multiple domains [127]. However, since this analysis was based on only 26 genomes, it is not quite clear how representative these results are for a much-expanded set of genomes. Therefore, the abundance of single- and multi-domain membrane proteins was analyzed in the dataset using Pfam-A [145, 146] domain assignments. Each domain was assigned as either soluble (containing no TMHs), transmembrane (containing TMHs) or hybrid (containing both soluble and TMH regions; for more information see Materials and Methods). When no domain class is specified in the following, all three domain classes are referenced. Almost 83% of all membrane proteins were found to have at least one domain assignment. Among these proteins the frequency of multi-domain membrane proteins was 20.9% in prokaryota, 28.1% in eukaryota and 37.9% in viruses (for detailed information see Table 3.2 and Table 3.3).

Table 3.2: Occurrence of single-domain and multi-domain membrane proteins

	All	Eukaryota	Prokaryota		Viruses
			Archaea	Eubacteria	
Sequences	373,800	153,486	14,314	204,628	1,420
Sequences with assignment	310,138	126,684	10,332	172,377	778
Single-domain proteins					
Sequences with one domain[a]					
Absolute number	236,027	91,033	8,502	136,038	483
Percentage[b]	63.1/76.1	59.3/71.9	59.4/82.3	66.5/78.9	34.0/62.1
Sequences with one soluble domain					
Absolute number	14,495	6,286	405	7,737	70
Percentage[b]	3.9/4.7	4.1/5.0	2.8/3.9	3.8/4.5	4.9/9.0
Sequences with one transmembrane domain					
Absolute number	205,683	77,500	7,761	120,131	316
Percentage[b]	55.0/66.3	50.5/61.2	54.2/75.1	58.7/69.7	22.3/40.6
Sequences with one hybrid domain[c]					
Absolute number	15,849	7,247	336	8,170	97
Percentage[b]	4.2/5.1	4.7/5.7	2.3/3.3	4.0/4.7	6.8/12.5
Multi-domain proteins					
Sequences with multiple domains[a]					
Absolute number	74,111	35,651	1,830	36,339	295
Percentage[b]	19.8/23.9	23.2/28.1	12.8/17.7	17.8/21.1	20.8/37.9
Sequences with multiple transmembrane domains					
Absolute number	32,710	15,602	1,084	15,978	47
Percentage[b]	8.8/10.6	10.2/12.3	7.6/10.5	7.8/9.3	3.3/6.0

[a] Either soluble, transmembrane or hybrid[c].
[b] Two values are given: first value corresponds to all sequences and second value to all sequences having a domain assignment.
[c] Hybrid domains are defined as domains containing predicted TMHs and soluble regions longer than 120 residues.

Altogether, multi-domain proteins (irrespective of taxonomic origin) accounted for almost 24% of the dataset of membrane proteins with at least one domain assignment. Furthermore, eukaryotic membrane proteins more often contain multiple domains classified as transmembrane (12.3%) than prokaryotic (9.3%) and viral membrane proteins (6%). Although the abundance of multi-domain proteins was by far not as high as that of globular proteins (40-60% [297]), the trend among the kingdoms was the same: eukaryotic proteins were more often multi-domain proteins than prokaryotic proteins.

Table 3.3: Domain combinations in membrane proteins. For every subset of the membrane protein dataset (Archaea, Eubacteria, Eukaryota and Viruses), the absolute number of proteins containing the specified number of hybrid (H), transmembrane (T) and soluble (S) domains is given. Hybrid domains are defined as domains containing predicted TMHs and soluble regions longer than 120 residues.

	0 H	1 H	>1 H	
		Archaea		
0 T	3,982	336	2	0 S
0 T	405	52	0	1 S
0 T	136	71	0	>1 S
1 T	7,761	74	0	0 S
1 T	262	17	0	1 S
1 T	131	1	0	>1 S
>1 T	952	0	0	0 S
>1 T	63	49	0	1 S
>1 T	20	0	0	>1 S
		Eubacteria		
0 T	32,251	8,170	282	0 S
0 T	7,737	1,439	2	1 S
0 T	4,167	1,120	6	>1 S
1 T	120,131	541	9	0 S
1 T	8,787	129	1	1 S
1 T	3,810	68	0	>1 S
>1 T	14,110	26	4	0 S
>1 T	836	550	0	1 S
>1 T	448	4	0	>1 S
		Eukaryota		
0 T	26,802	7,247	140	0 S
0 T	6,286	1,813	176	1 S
0 T	3,088	827	79	>1 S
1 T	77,500	607	14	0 S
1 T	8,664	477	33	1 S
1 T	3,932	196	3	>1 S
>1 T	10,231	112	4	0 S
>1 T	2,068	614	3	1 S
>1 T	2,397	171	2	>1 S
		Viruses		
0 T	642	97	6	0 S
0 T	70	42	8	1 S
0 T	103	39	17	>1 S
1 T	316	7	0	0 S
1 T	15	0	0	1 S
1 T	11	0	0	>1 S
>1 T	12	0	0	0 S
>1 T	0	1	0	1 S
>1 T	23	7	4	>1 S

High percentage of multi-domain membrane proteins (24%) warrants the application of more sophisticated sequence clustering methods, as was done in CAMPS 2.0.

3.3.4 Structural classification of membrane proteins using meta-models

The procedure to identify SC-clusters (see Materials and Methods) designed to represent the protein structure level was applied to an initial dataset of 494,679 sequences with at least three TMHs. The resulting 1,353 SC-clusters comprised 413,714 sequences accounting for 83.6% of the initial dataset. The size of a SC-cluster ranges between 14,411 and 15 sequences (the minimum size to derive SC-clusters; see Materials and Methods) with a median size of 83. The vast majority of SC-clusters have fewer than 200 members (73.9%) and large SC-clusters are rare (Figure 3.7). The list of the 15 largest

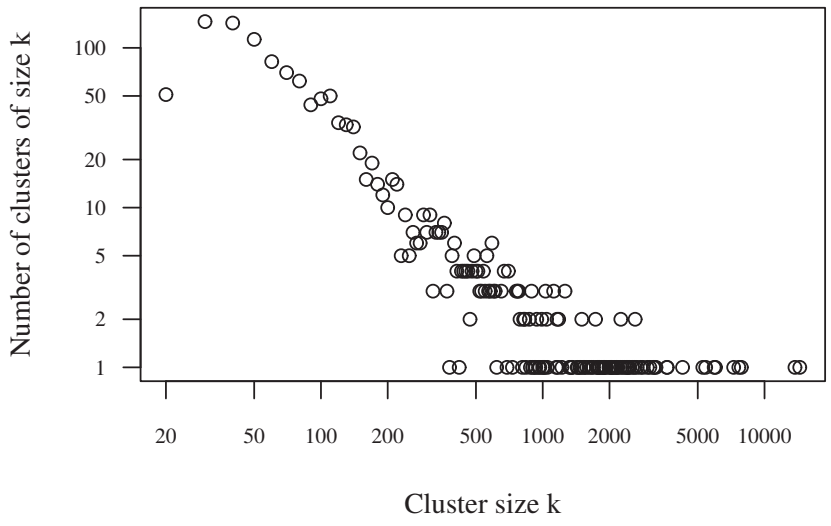

Figure 3.7: Double logarithmic plot of the SC-cluster size distribution. Most of the SC-clusters have less than 200 members and large SC-clusters are rare.

SC-clusters (Table 3.4) includes prominent superfamilies, such as the major facilitator superfamily and the ABC superfamily. Ten of the largest SC-clusters are already associated with known 3D structures. In total 53 out of 1,353 SC-clusters have at least one protein member whose structure is known, which means that 1,300 additional structures

Table 3.4: The 15 largest SC-clusters in CAMPS 2.0

Cluster	Description	Size	TMH range	Cores	Taxonomy[a]	Representative structure[b]
CMSC0001	Major facilitator superfamily #1	14,411	10-13	12	Eu+Pro	2cfqA
CMSC0002	Family A G protein-coupled receptor-like superfamily #1	13,716	6-8	7	Eu+Pro	-
CMSC0003	Major facilitator superfamily #2	7,855	10-13	11	Eu+Pro	1yg7A
CMSC0004	SC-cluster containing ABC transporter proteins	7,651	5-7	6	Eu+Pro	2hydA
CMSC0005	Major facilitator superfamily #3	7,258	10-13	10	Eu+Pro	1pw4A
CMSC0006	APC superfamily #1	6,006	10-13	13	Eu+Pro	3giaA
CMSC0007	Uncharacterized SC-cluster CMSC0007	5,925	8-11	10	Eu+Pro	-
CMSC0008	Binding-protein-dependent transport system inner membrane component #1	5,429	5-7	7	Eu+Pro	3d31C
CMSC0009	7 transmembrane receptor (rhodopsin family) #1	5,261	6-8	8	Eu	1u19A
CMSC0010	ABC-2-transporter-like clan	4,264	5-7	7	Eu+Pro	-
CMSC0011	Binding-protein-dependent transport system inner membrane component #2	3,635	5-7	6	Eu+Pro	3dhwA
CMSC0012	AcrB/AcrD/AcrF family	3,630	10-13	12	Eu+Pro	2j8sA
CMSC0013	SC-cluster containing calcium ATPase proteins	3,240	8-11	10	Eu+Pro	1wpgA
CMSC0014	Branched-chain amino acid transport system/ permease component #1	3,172	7-9	9	Eu+Pro	-
CMSC0015	Binding-protein-dependent transport system inner membrane component #3	3,157	4-6	5	Eu+Pro	-

[a] Eu: Eukaryota, Pro: Prokaryota.

[b] If a SC-cluster is associated with more than one structure, a representative structure was chosen, giving preference to structures determined by X-ray diffraction. If several X-ray structures were available, the one with the best resolution was selected.

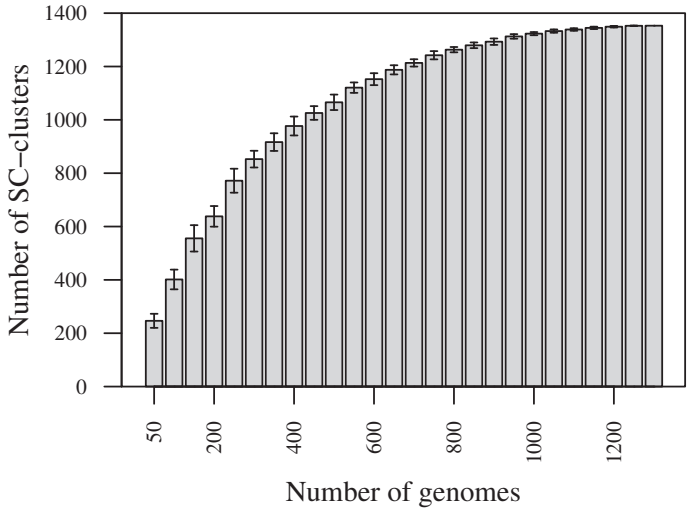

Figure 3.8: Dependence of the number of SC-clusters on the number of analyzed genomes. From the full set of 1,253 genomes, subsets containing 50 genomes, 100 genomes etc. were constructed randomly and the number of SC-clusters was counted. For each subset, the calculations were repeated 25 times. The mean values with their standard deviation (mean $\pm$ SD) are shown.

would be needed to provide full structural coverage for the sequence space investigated in this work. Figure 3.8 shows how the number of SC-clusters depends on the number of genomes analyzed. In the range of about 50 to 700 genomes, the number of SC-clusters increases rapidly as the number of genomes increases with almost 90% of all SC-clusters covered with 700 genomes. The curve shows clear saturation as the number of genomes approaches 1000. Thus, it seems that the current set of 1,353 SC-clusters is already representative for the entire membrane protein sequence space and is not likely to increase dramatically with the addition of new genomes. All but one SC-cluster in Table 3.4 are universal with respect to the superkingdoms of life, containing protein members from all three superkingdoms. Overall there are 179 universal SC-clusters and 770 SC-clusters that are specific for one superkingdom (Archaea: 13, Bacteria: 348, Eukarya: 409; Figure 3.5).

Using the dataset of 413,714 sequences covered by the SC-clusters, the loop length distributions were investigated among the different classes of α-helical membrane proteins (defining loops as regions connecting two TMHs), e.g. proteins with three TMHs

(3TMH), four TMHs (4TMH) and so forth. The most frequent classes are 6TMH (13.3%) and 7TMH (12.9%), corresponding to abundant families of ABC transporters and GPCR proteins (Figure 3.9A). The loop length distributions are relatively homogeneous among

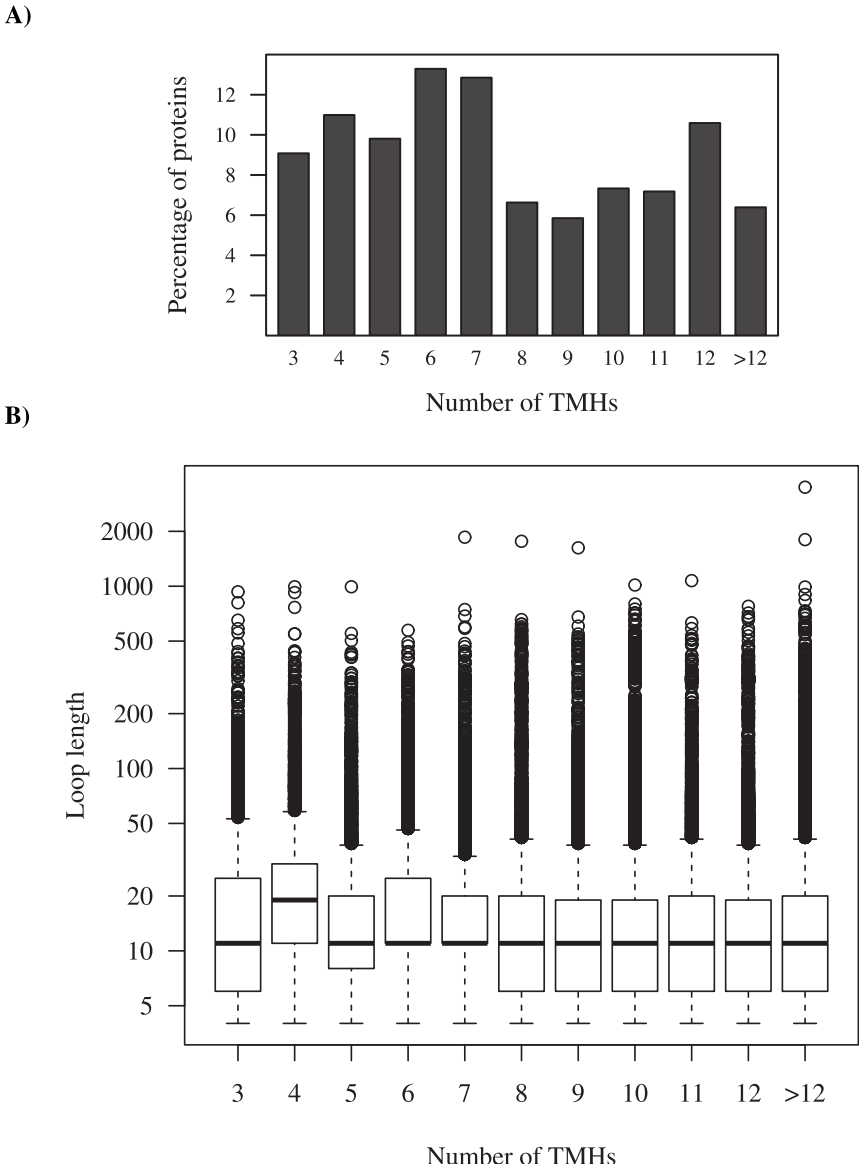

Figure 3.9: (A) Distribution of TMH number. (B) Distribution of loop lengths (shown as boxplot) for different classes of membrane proteins. Both distributions refer to membrane protein sequences covered by SC-clusters.

the different membrane protein classes (Figure 3.9B). Except for 4TMH, all classes have a median loop length of 11 residues. The distributions also show that in case of 3TMH and 4TMH 90% of all loops are shorter than 80 residues and for the other classes about

90-95% of all loops are not longer than 50 residues. The remaining loops vary greatly in length and can reach 2,000 residues with some of them containing soluble Pfam [145, 146] domains (about 47% of the loops longer than 300 residues enclose a soluble Pfam domain).

At this point, it has to be mentioned that this analysis is based on predicted TMHs and therefore is influenced by the methodology used in Phobius [100]. For example, when the loop length distributions were analyzed for proteins with experimentally derived topology information (SwissProt [139]), the median loop lengths were found to be longer than 11 residues (ranging from 13 to 18 residues) and they were not as uniform among the different membrane protein classes as in case of predicted topology. For comparison, the loop lengths were recalculated for the same SwissProt dataset using predicted TMHs and it was found that the median length was close to 11 residues (data not shown).

Because of the fact that the classification approach is substantially based on sequence similarity, it has to be assumed that the method can not deal with analogous structures resulting from convergent evolution [150]. That means that membrane proteins having the same fold, but very low sequence similarity will be assigned to different SC- clusters. Thus, it is likely that several SC-clusters describe the same fold. However, membrane proteins within the same SC-clusters typically share the same fold. Taken together, it can be concluded that the number of 1,353 SC-clusters is a reasonable upper bound for the number of existing membrane protein folds.

3.3.5 Comparison of SC-clusters with Pfam

CAMPS SC-clusters were compared to the most widely used functional classification of proteins, the Pfam [145, 146] database. In contrast to CAMPS, Pfam is a purely sequence-based approach and does not utilize structural features. Furthermore, the two databases also use different methods to build models for the protein families (Pfam) and SC-clusters (CAMPS), respectively. For each protein family in Pfam a seed alignment is constructed from a non-redundant sequence set. This alignment is then manually verified and a HMM is built from the final seed alignment [145]. While CAMPS also uses a non-redundant set of sequences to create an alignment for each SC-cluster, it operates with a series of inter-connected HMMs (see Materials and Methods) rather than with a single HMM, with each individual HMM corresponding to clearly defined structural features (TMHs and connecting loops). In addition to families, Pfam also provides a higher order classification called protein clans [149]. A clan is a group of Pfam families presumed to have a common evolutionary origin. The presence of related structures and significant

profile-profile comparison scores are the most important criteria of relatedness for Pfam families.

SC-clusters versus Pfam families

The comparison of SC-clusters with Pfam families involved 94,337 sequences, 654 SC-clusters, and 550 Pfam families (Table 3.5). By analyzing the distribution of Pfam family assignments within each SC-cluster, a perfect agreement could be found for 602 SC-clusters (1:1 relationships, i.e. all members of a SC-cluster were related to exactly one Pfam family). The remaining 52 SC-clusters (involving 14.7% of the sequences) were associated with two to five distinct Pfam family assignments each (1:n relationships). For each of these SC-clusters the Pfam clan assignments of the corresponding Pfam families were extracted and it was found that in those cases where at least two clan assignments were available (41 out of 52 cases), all Pfam families within a SC-cluster belonged to the same Pfam clan. By analyzing the distribution of SC-cluster assignments within each Pfam family, 452 1:1 relationships and 98 1:n relationships (involving 32.8% of the sequences) were found. Less stringent conditions were also tested for 1:1 relationships (requiring 90% rather that 100% of all sequences to be linked with the same SC-cluster) and another 30 cases of agreement were found. Accordingly, the number of 1:n relationships went down from 98 to 68 cases (involving 18.2% of the sequences).

Pairwise sequence identities for the sequences involved in the comparison between SC-clusters and Pfam families had a median value of 27% for the agreements (for both directions, i.e. CAMPS versus Pfam and Pfam versus CAMPS) and 23% for the disagreements (again for both directions; Table 3.5). Likewise, the median standard deviation of the number of TMHs (Table 3.5) was also determined. For the 1:1 relationships, the median was around 0.60 (for both directions) and for the 1:n relationships 0.86 (SC-clusters versus Pfam families) and 0.80 (Pfam families versus SC-clusters), respectively. Thus, the general trend here is that both values (sequence identity and TMH number) clearly differ between agreements and disagreements and that sequences from SC-clusters spanning multiple Pfam families are equally similar to each other (both with respect to sequence identity and TMH number) as sequences from Pfam families spanning several SC-clusters.

SC-cluster versus Pfam clans

In the next step, SC-clusters were compared to Pfam clans. At this level, 56,756 sequences, 293 SC-clusters and 43 Pfam clans were found to be involved (Table 3.6). Each

Table 3.5: Comparison of CAMPS 2.0 SC-clusters with Pfam families

	SC-clusters versus Pfam families	Pfam families versus SC-clusters
DB1[a]	SC-clusters	Pfam families
DB2[a]	Pfam families	SC-clusters
Sequences[b]	94,337	94,337
Groups(DB1)[c]	654	550
Groups(DB2)[c]	550	654
1:1 relationships[d]		
Occurrences	602	452
Involved sequences	80,487 (85.3%)	63,385 (67.2%)
1:n relationships[e]		
Occurrences	52	98
Involved sequences	13,850 (14.7%)	30,952 (32.8%)
Median sequence identity[f]		
1:1 relationships	27.4%	27.9%
1:n relationships	22.8%	23.6%
Median standard deviation of TMH number[g]		
1:1 relationships	0.60	0.62
1:n relationships	0.86	0.80

[a] For each group in DB1 the associated groups in DB2 were analyzed.
[b] Number of sequences involved in the comparison.
[c] In case of CAMPS groups denote SC-clusters and in case of Pfam families.
[d] One DB1 group assignment is associated with one DB2 group assignment.
[e] One DB1 group assignment is associated with multiple DB2 group assignments.
[f] For all sequences involved in 1:1/1:n relationships, the median of all pairwise sequence identities (taken from SIMAP) was calculated.
[g] For all sequences involved in 1:1/1:n relationships, the median of all standard deviations of the number of TMHs was calculated.

of the 293 SC-clusters was associated with one Pfam clan respectively (1:1 relationship). Conversely, the analysis of the distribution of SC-cluster assignments within each Pfam clan revealed that 1:1 relationships were much less frequent than 1:n relationships. Six clans perfectly agreed with SC-clusters, but 37 clans were linked to more than one SC-cluster (involving 99.1% of the sequences).

Analogously to the comparison at the family level, the median sequence identity and standard deviation of TMH number were measured (Table 3.6). In case of agreements, the median sequence identity reached 26.5% (SC-clusters versus clans) and 48.9% (clans versus SC-clusters), respectively, and in case of disagreements, 23.1%. For those cases, where one SC-cluster corresponded to one Pfam clan, the number of TMHs deviated by

Table 3.6: Comparison of CAMPS 2.0 SC-clusters with Pfam clans

	SC-clusters versus Pfam clans	Pfam clans versus SC-clusters
DB1[a]	SC-clusters	Pfam clans
DB2[a]	Pfam clans	SC-clusters
Sequences[b]	56,756	56,756
Groups(DB1)[c]	293	43
Groups(DB2)[c]	43	293
1:1 relationships[d]		
Occurrences	293	6
Involved sequences	56,756 (100%)	527 (0.9%)
1:n relationships[e]		
Occurrences	0	37
Involved sequences	0	56,229 (99.1%)
Median sequence identity[f]		
1:1 relationships	26.5%	48.9%
1:n relationships	-	23.1%
Median standard deviation of TMH number[g]		
1:1 relationships	0.83	0.37
1:n relationships	-	1.10

[a] For each group in DB1 the associated groups in DB2 were analyzed.
[b] Number of sequences involved in the comparison.
[c] In case of CAMPS groups denote SC-clusters and in case of Pfam clans.
[d] One DB1 group assignment is associated with one DB2 group assignment.
[e] One DB1 group assignment is associated with multiple DB2 group assignments.
[f] For all sequences involved in 1:1/1:n relationships, the median of all pairwise sequence identities (taken from SIMAP) was calculated.
[g] For all sequences involved in 1:1/1:n relationships, the median of all standard deviations of the number of TMHs was calculated.

a factor of 0.83 and for the reverse cases (one Pfam clan was linked to one SC-cluster) by a factor of 0.37. A much higher deviation (1.10) was found when one Pfam clan was associated with multiple SC-clusters. Compared to the family level, the first trend (sequence identity and TMH number clearly differ between agreements and disagreements) could also be observed at the clan level. In contrast, sequences from SC-clusters spanning one Pfam clan (forward comparison) were not as similar as sequences from Pfam clans spanning one SC-cluster (reverse comparison) with respect to sequence similarity (26.5% and 48.9%, respectively) and TMH number (0.83 and 0.37, respectively). Looking at SC-clusters and Pfam families, the median sequence identity varied only slightly

(27.4% and 27.9%, respectively) between both types of agreements (forward and reverse comparison). It seems that this is because the number of involved sequences was not comparable (56,756 in case of Pfam families and only 527 in case of Pfam clans). What is highly apparent is the clear difference between the median standard deviation of the number of TMHs at the family level and at the clan level (e.g. 0.60 in case one SC-cluster is associated with one Pfam family and 0.83 if one SC-cluster is associated with one Pfam clan). This observation reflects the fact that Pfam clans include more divergent sequences than Pfam families.

Summary

The classification of membrane proteins at the SC-cluster level resembles Pfam at the family rather than the clan level. The similarities are due to similar techniques employed by both resources (sequence similarity and HMMs). However, the comparison also showed that using additional information in the classification approach (namely the loop length and the number of TMHs) can result in a different classification, as was shown to be the case for 14.7% (CAMPS versus Pfam families; one SC-cluster corresponds to several Pfam families) and 32.8% (Pfam families versus CAMPS; one Pfam family corresponds to several SC-clusters) of the sequences. In the former case all Pfam families contained in a given SC-cluster belonged to the same Pfam clan implying that these SC-clusters are more inclusive than Pfam families. For the second type of discrepancy several explanations are possible. First, Pfam is more sensitive for finding divergent sequences since it uses an HMM-based search method. CAMPS also utilizes HMMs, but their application is preceded by an initial clustering based on FASTA alignments which are not necessarily capable of identifying very remote homologies. Furthermore, Pfam alignments are manually curated, while CAMPS is generated automatically. Thus, proteins sharing high structural similarity but very low sequence similarity may not be assigned to the same SC-cluster. Second, previous studies have shown that sequence families (such as Pfam) are not always connected to only one structural family [298, 299]. Third, CAMPS considers additional information for the classification leading to a further division of clusters since more criteria have to be met. Utilization of additional information (such as TMHs) increases the ability to find related membrane proteins [300, 301]. Traditional homology detection methods initially developed for globular proteins do not perform well for membrane proteins since TMHs are similar *per se* due to their biased amino acid sequence composition, which leads to a high rate of false positives. CAMPS utilizes different models for membrane-spanning

and extramembraneous regions.

Taken together, it was found that the SC-clusters reasonably agree with Pfam families. On the other hand, for almost 15% of the sequences (forward comparison), SC-clusters were found to be more inclusive meaning that SC-clusters were linked to multiple Pfam families. The reverse comparison revealed that about 18% of the sequences correspond to Pfam families that are associated with multiple SC-clusters. Differences between CAMPS and Pfam reflect different objectives. While CAMPS aims at structural families, Pfam is a functional classification that does not consider structural information at all. Even though significant sequence similarity is a strong indicator of structural similarity, sequence identities in the 'twilight zone' of 20-35% [184] are problematic. Proteins with very low sequence similarity can share structural similarity (see Sadekar *et al.*, 2006 [302] for an example), but this is not common [184]. Structural information may be helpful for finding distantly related globular proteins [303], and this is even more true for membrane proteins due to their biased amino acid composition and the ensuing higher chance similarity. Therefore, CAMPS classification is primarily geared towards the users interested in structural membrane protein families.

3.3.6 Comparison of SC-clusters with TCDB

Similarly, CAMPS was also compared with the hierarchical TCDB database [247] that uses both functional and phylogenetic information, but no sequence information. It is analogous to the numerical EC (Enzyme Commission) system that classifies enzymes by reaction type [179]. In contrast to Pfam and CAMPS, TCDB does not consider sequence similarity (except at the superfamily level; see below) and, unlike Pfam, TCDB is a membrane protein specific database. Similar to Pfam, TCDB also provides a higher classification level called superfamilies that comprises large families consisting of proteins highly divergent in sequence. The relationship between CAMPS and TCDB was investigated at both levels.

SC-clusters versus TCDB families

At the family level, 2,085 protein sequences, 337 SC-clusters and 228 TCDB families were considered for comparison (Table 3.7). When the distribution of TCDB family assignments for each individual SC-cluster was analyzed, 304 1:1 relationships and 33 1:n relationships (covering 13% of the sequences) were found. For the 33 SC-clusters that were linked to multiple TCDB families, the TCDB superfamily assignments were also

Table 3.7: Comparison of CAMPS 2.0 SC-clusters with TCDB families

	SC-clusters versus TCDB families	TCDB families versus SC-clusters
DB1[a]	SC-clusters	TCDB families
DB2[a]	TCDB families	SC-clusters
Sequences[b]	2,085	2,085
Groups(DB1)[c]	337	228
Groups(DB2)[c]	228	337
1:1 relationships[d]		
Occurrences	304	176
Involved sequences	1,814 (87%)	682 (32.7%)
1:n relationships[e]		
Occurrences	33	52
Involved sequences	271 (13%)	1,403 (67.3%)
Median sequence identity[f]		
1:1 relationships	26.8%	28.4%
1:n relationships	25.8%	22.8%
Median standard deviation of TMH number[g]		
1:1 relationships	0.71	0.64
1:n relationships	0.70	1.46

[a] For each group in DB1 the associated groups in DB2 were analyzed.
[b] Number of sequences involved in the comparison.
[c] In case of CAMPS groups denote SC-clusters and in case of TCDB families.
[d] One DB1 group assignment is associated with one DB2 group assignment.
[e] One DB1 group assignment is associated with multiple DB2 group assignments.
[f] For all sequences involved in 1:1/1:n relationships, the median of all pairwise sequence identities (taken from SIMAP) was calculated.
[g] For all sequences involved in 1:1/1:n relationships, the median of all standard deviations of the number of TMHs was calculated.

examined and it was found that all SC-clusters having at least two superfamily assignments (8 out of 33) were associated with the same superfamily. The inverse comparison revealed that 176 TCDB families have a perfect agreement with SC-clusters, while the remaining 52 families were associated with multiple SC-clusters (covering 67.3% of the sequences). For 2 out of 52 families CATH [164, 165] and SCOP [162, 163] fold assignments were found for two associated SC-clusters in each case. In the first case (TCDB family 3.A.1 ('ABC superfamily')), one SC-cluster was found to be linked with the CATH fold 1.10.3470 ('ABC transporter involved in vitamin B12 uptake, BtuC') and the SCOP fold f.22 ('ABC transporter involved in vitamin B12 uptake, BtuC') (Figure 3.10A) and the other SC-cluster with the CATH fold 1.20.1560 ('ABC transporter transmembrane

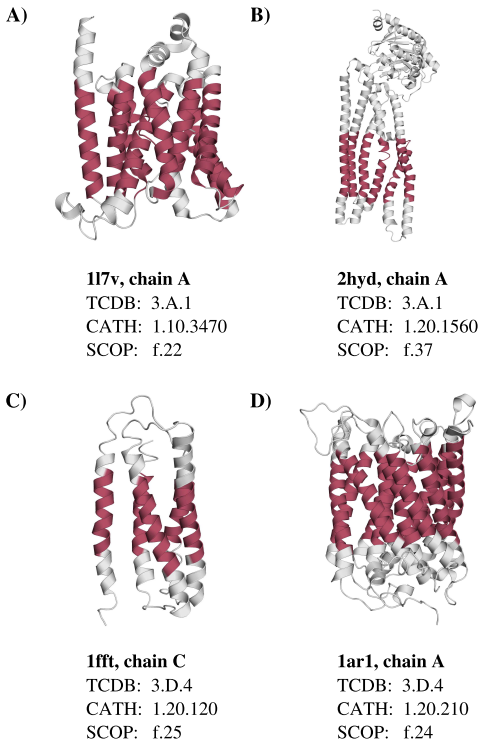

Figure 3.10: Examples of membrane proteins assigned to the same TCDB family, but to different SCOP and CATH folds. (A) Bacterial ABC transporter (PDB code: 1l7v, chain A) with 10 TMHs. (B) Bacterial ABC transporter (PDB code: 2hyd, chain A) with 6 TMHs. Both transporters are assigned to TCDB family 3.A.1 ('ABC superfamily'). (C) Ubiquinol oxidase from *E. coli* (PDB code: 1fft, chain C) with 5 TMHs. (D) Cytochrome **c** oxidase (PDB code: 1ar1, chain A) with 12 TMHs. Both oxidases belong to TCDB family 3.D.4 ('Cytochrome oxidase superfamily'). TMH are colored in red with coordinates extracted from PDBTM [26]. The figure was drawn using PyMOL [66].

region fold') and the SCOP fold f.37 ('ABC transporter transmembrane region') (Figure 3.10B). Similarly, the second TCDB family 3.D.4 ('Cytochrome oxidase superfamily') was connected with two different CATH and SCOP folds, respectively (Figures 3.10C and 3.10D). Again, the two CATH folds (1.20.120 'Four helix bundle' and 1.20.210 'Cytochrome *c* oxidase; chain A') and the two SCOP folds (f.25 'Cytochrome *c* oxidase subunit III-like' and f.24 'Cytochrome *c* oxidase subunit I-like') belonged to the same TCDB family (3.D.4), but to different SC-clusters.

By calculating the pairwise sequence identities of all sequences involved in 1:1 and 1:n relationships between CAMPS and TCDB, it was found that the median sequence identity for the cases of agreement was 26.8% (SC-clusters versus TCDB families) and 28.4% (TCDB families versus SC-clusters), respectively (Table 3.7). For those cases

where one SC-cluster was associated with multiple TCDB families, the median sequence identity was 25.8%. In contrast, considering the cases where one TCDB family was linked with more than one SC-cluster, this value dropped down to 22.8%. Similarly, the median standard deviation of the number of TMHs was also found to differ significantly between the cases of agreement (SC-clusters versus TCDB families: 0.71, TCDB families versus SC-clusters: 0.64)) and the 1:n relationships between TCDB families and SC-clusters (1.46; Table 3.7). The median standard deviation for the cases where one SC-cluster was associated with multiple TCDB families was in the same range as for the agreements (0.70).

SC-clusters versus TCDB superfamilies

The comparison at the superfamily level of TCDB comprised 647 sequences, 69 SC-clusters and 10 TCDB superfamilies (Table 3.8). As in the case of Pfam clans, 1:1 relationships were found for all 69 SC-clusters. Thus, no SC-cluster was associated with more than one superfamily. However, in the reverse case, only two superfamilies were found to perfectly agree with SC-cluster assignments (1:1 relationships) and eight superfamilies (involving 98.8% of the sequences) were spread over up to 20 SC-clusters.

When comparing the median sequence identities and the median standard deviations of the number of TMHs for the agreements and disagreements between SC-clusters and TCDB superfamilies (Table 3.8), the same trend as with TCDB families was observed (values for agreements clearly differ from 1:n relationships between TCDB and CAMPS). However, sequence identities are slightly lower and the standard deviations of the TMH number are slightly higher compared with TCDB families since superfamilies also include divergent sequences.

Summary

CAMPS and TCDB do correspond in many cases, but rather at the family and not at the TCDB superfamily level. As discussed above for Pfam, poor matches between SC-clusters and TCDB families are primarily due to different objectives. While SC-clusters are based on sequence and structural similarity and represent structural families, the TCDB database is a functional classification not considering sequence information at all. For 13% of the sequences involved in this analysis, the SC-clusters were found to be more inclusive (i.e. one SC-cluster was linked to multiple TCDB families), which is consistent with the finding that some folds can be connected to different functions (e.g. TIM barrel [195]; 'same fold, different functions' paradigm). Similarly, in many

Table 3.8: Comparison of CAMPS 2.0 SC-clusters with TCDB superfamilies

	SC-clusters versus TCDB superfamilies	TCDB superfamilies versus SC-clusters
DB1[a]	SC-clusters	TCDB superfamilies
DB2[a]	TCDB superfamilies	SC-clusters
Sequences[b]	647	647
Groups(DB1)[c]	69	10
Groups(DB2)[c]	10	69
1:1 relationships[d]		
Occurrences	69	2
Involved sequences	647 (100%)	8 (1.2%)
1:n relationships[e]		
Occurrences	0	8
Involved sequences	0	639 (98.8%)
Median sequence identity[f]		
1:1 relationships	24.3%	26.8%
1:n relationships	-	22.3%
Median standard deviation of TMH number[g]		
1:1 relationships	0.84	0.79
1:n relationships	-	1.54

[a] For each group in DB1 the associated groups in DB2 were analyzed.
[b] Number of sequences involved in the comparison.
[c] In case of CAMPS groups denote SC-clusters and in case of TCDB superfamilies.
[d] One DB1 group assignment is associated with one DB2 group assignment.
[e] One DB1 group assignment is associated with multiple DB2 group assignments.
[f] For all sequences involved in 1:1/1:n relationships, the median of all pairwise sequence identities (taken from SIMAP) was calculated.
[g] For all sequences involved in 1:1/1:n relationships, the median of all standard deviations of the number of TMHs was calculated.

cases (involving $> 67\%$ of the sequences) one TCDB family corresponded to several SC-clusters, consistent with the notion that the same function can be linked to different folds ('same function, different folds' paradigm). Indeed, two of these TCDB families (3.A.1 and 3.D.4) correspond to different CATH and SCOP folds, respectively (each corresponding to a different SC-cluster; Figure 3.10). Since the latter notion ('same function, different folds') was particularly apparent in the comparison between CAMPS and TCDB, a lot more SC-clusters than TCDB families were involved (337 versus 228).

3.3.7 Comparison of SC-clusters with SCOP and CATH

The SC-cluster classification approach aims at identifying structural membrane protein families whose members share the same fold. Thus, it was particularly interesting to evaluate how well the SC-clusters correlate with SCOP [162, 163] and CATH [164, 165] folds. To this end, membrane proteins covered by SC-clusters as well as by SCOP or CATH were identified.

In SCOP and CATH proteins are assigned to the same fold if their structures are similar in the overall shape and connectivity of secondary structure elements. Thus, at the fold level the classification approach of SCOP and CATH is solely based on structure. In contrast, CAMPS is mainly based on sequence information, but also exploits structural features. The major difference here is that while SCOP and CATH rely on tertiary structure information CAMPS uses predicted topology information.

Since membrane protein structures remain scarce only 54 proteins with known structure were involved in the comparison with CATH, spread over 21 CATH folds and 31 SC-clusters (Table 3.9). When each of the 31 SC-clusters was investigated separately and the distribution of CATH fold assignments within each of them was tested, a perfect agreement was found in all cases. By comparing the two databases in the reverse direction (i.e. by analyzing the distribution of SC-clusters within each CATH fold), 1:1 relationships were found for 16 out of 21 CATH folds. The five other folds (CATH codes 1.10.287 'Helix hairpins', 1.20.120 'Four helix bundle', 1.20.950 'Fumarate reductase cytochrome b subunit', 1.20.1070 'Rhodopsin 7-helix transmembrane proteins' and 1.20.1300 '3 helical TM bundles of succinate and fumarate reductases') were associated with two to four SC-clusters (Figure 3.11). Except for one case (fold 1.20.1070), all proteins involved in the disagreements had two to five TMHs (here the number of TMHs corresponds to the PDBTM [26] annotation). One explanation for the disagreements between CATH and CAMPS might be the fact that membrane proteins with few helices (< 6 TMHs) are difficult to classify in general, as was demonstrated in the comparative analysis of membrane protein classification in SCOP and CATH (see Chapter 2, page 23). Specifically, it was found that the fold space of membrane proteins with less than six TMHs is rather continuous, thus complicating their structural classification. Indeed, all CATH folds except 1.20.1070 involved in the disagreements with SC-clusters (1.10.287, 1.20.120, 1.20.950 and 1.20.1300) were already found to be involved in the disagreements between CATH an SCOP (see section 2.3.4, 35).

As was already done in the comparisons with Pfam and TCDB, the median sequence identity and the median standard deviation of the number of TMHs were calculated

Table 3.9: Comparison of CAMPS 2.0 SC-clusters with CATH folds

	SC-clusters versus CATH folds	CATH folds versus SC-clusters
DB1[a]	SC-clusters	CATH folds
DB2[a]	CATH folds	SC-clusters
Sequences[b]	54	54
Groups(DB1)[c]	31	21
Groups(DB2)[c]	21	31
1:1 relationships[d]		
Occurrences	31	16
Involved sequences	54 (100%)	36 (66.7%)
1:n relationships[e]		
Occurrences	0	5
Involved sequences	0	18 (33.3%)
Median sequence identity[f]		
1:1 relationships	34.5%	35.7%
1:n relationships	-	28.5%
Median standard deviation of TMH number[g]		
1:1 relationships	0.71	0.71
1:n relationships	-	0.58
Median structural similarity [Z-score][h]		
1:1 relationships	30.1	30.2
1:n relationships	-	7.5

[a] For each group in DB1 the associated groups in DB2 were analyzed.
[b] Number of sequences involved in the comparison.
[c] In case of CAMPS groups denote SC-clusters and in case of CATH folds.
[d] One DB1 group assignment is associated with one DB2 group assignment.
[e] One DB1 group assignment is associated with multiple DB2 group assignments.
[f] For all sequences involved in 1:1/1:n relationships, the median of all pairwise sequence identities (taken from SIMAP) was calculated.
[g] For all sequences involved in 1:1/1:n relationships, the median of all standard deviations of the number of TMHs was calculated.
[h] Similar to [f]. Structural similarity describes the overall similarity of 3D structures.

for the cases of agreement and disagreement. And since the comparison with CATH was based on three-dimensional structures, the median structural similarity was also measured using DaliLite [225] (Table 3.9). For those cases where one SC-cluster corresponded to one CATH fold and *vice versa* the median sequence identity was 34.5% and 35.7%, respectively. In case of disagreements, the sequence identity was only 28.5%. In

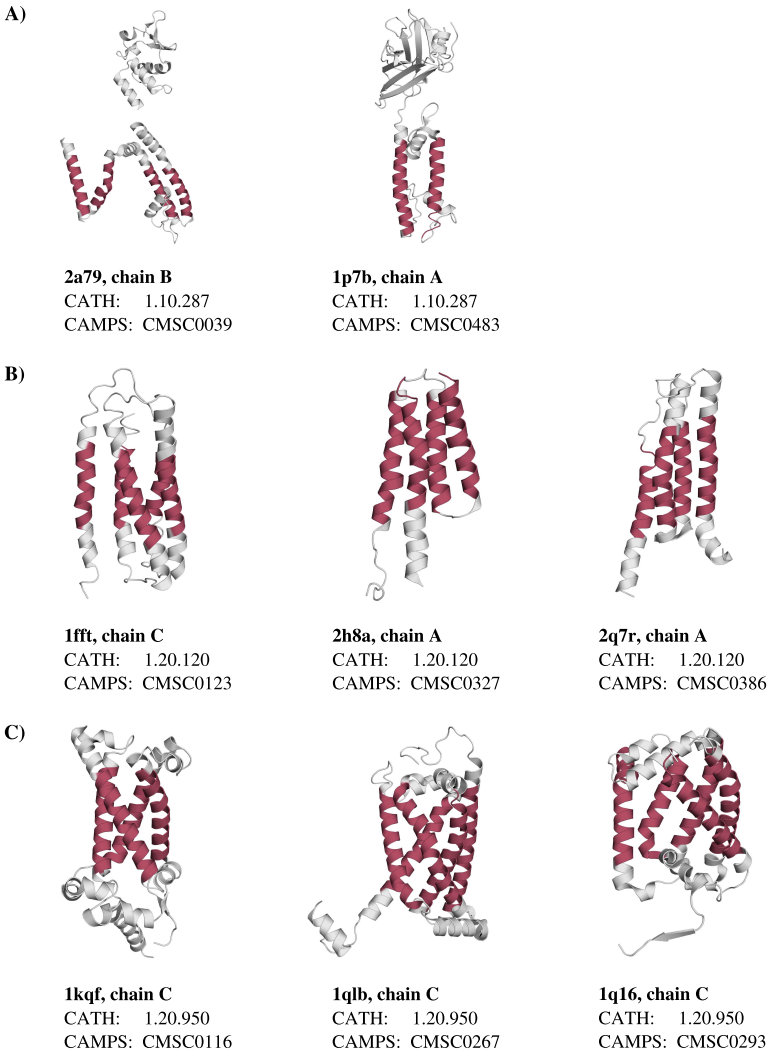

Figure 3.11: Examples of membrane proteins assigned to the same CATH fold, but to different SC-clusters. The proteins shown are assigned to CATH folds (A) 1.10.287, (B) 1.20.120, (C) 1.20.950, (D) 1.20.1070 and (E) 1.20.1300. The corresponding SC-clusters are given for each structure. TMH coordinates were extracted from PDBTM [26]. The figure was drawn using PyMOL [66]. (*Figure continues on next page.*)

terms of the number of TMHs, the median standard deviation was found to range between 0.58 (disagreements) and 0.71 (agreements). In contrast, structural similarity of proteins involved in agreement and disagreements differed significantly. While cases of agreement achieved a median Z-score of about 30, the median Z-score for those CATH folds corresponding to several SC-clusters was only 7.5. According to the authors of DaliLite Z-scores above 20 describe true homologies and those in the range of 8 and 20 correspond to probable homologues. Values in the range of 2 and 8 define the grey area,

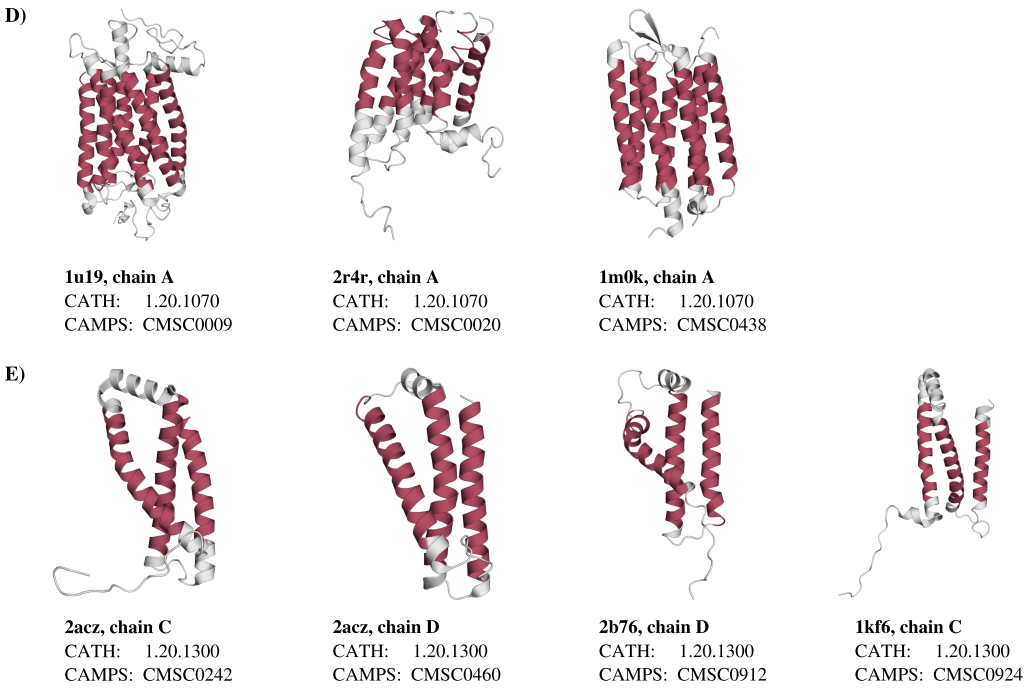

Figure 3.11: *Continued.*

and values below 2 are not significant.

The comparison of SCOP folds with SC-clusters, which involved 54 proteins, 37 SC-clusters and 25 SCOP folds (Table 3.10), yielded similar results. Each of the 37 SC-clusters mapped to one SCOP fold while 20 out of 25 SCOP folds had a perfect relationship with SC-clusters. Only five SCOP folds (SCOP codes f.13 'Family A G protein-coupled receptor-like', f.14 'Voltage-gated potassium channels', f.21 'Heme-binding four-helical bundle', f.38 'MFS general substrate transporter' and f.58 'MetI-like') were found to be spread over two to eight SC-clusters (Figure 3.12). 13 out of 17 proteins involved in the disagreements had between two and six TMHs. Again, it is assumed that the discrepancies arose due to the difficulty in classifying membrane proteins with few helices. Indeed, two SCOP folds (f.14 and f.21) belong to those where SCOP and CATH disagree (see section 2.3.4, page 35).

Table 3.10: Comparison of CAMPS 2.0 SC-clusters with SCOP folds

	SC-clusters versus SCOP folds	SCOP folds versus SC-clusters
DB1[a]	SC-clusters	SCOP folds
DB2[a]	SCOP folds	SC-clusters
Sequences[b]	54	54
Groups(DB1)[c]	37	25
Groups(DB2)[c]	25	37
1:1 relationships[d]		
Occurrences	37	20
Involved sequences	54 (100%)	32 (59.3%)
1:n relationships[e]		
Occurrences	0	5
Involved sequences	0	22 (40.7%)
Median sequence identity[f]		
1:1 relationships	33.1%	35.1%
1:n relationships	-	25.0%
Median standard deviation of TMH number[g]		
1:1 relationships	0.45	0.58
1:n relationships	-	0
Median structural similarity [Z-score][h]		
1:1 relationships	29.6	29.6
1:n relationships	-	5.8

[a] For each group in DB1 the associated groups in DB2 were analyzed.
[b] Number of sequences involved in the comparison.
[c] In case of CAMPS groups denote SC-clusters and in case of SCOP folds.
[d] One DB1 group assignment is associated with one DB2 group assignment.
[e] One DB1 group assignment is associated with multiple DB2 group assignments.
[f] For all sequences involved in 1:1/1:n relationships, the median of all pairwise sequence identities (taken from SIMAP) was calculated.
[g] For all sequences involved in 1:1/1:n relationships, the median of all standard deviations of the number of TMHs was calculated.
[h] Similar to [f]. Structural similarity describes the overall similarity of 3D structures.

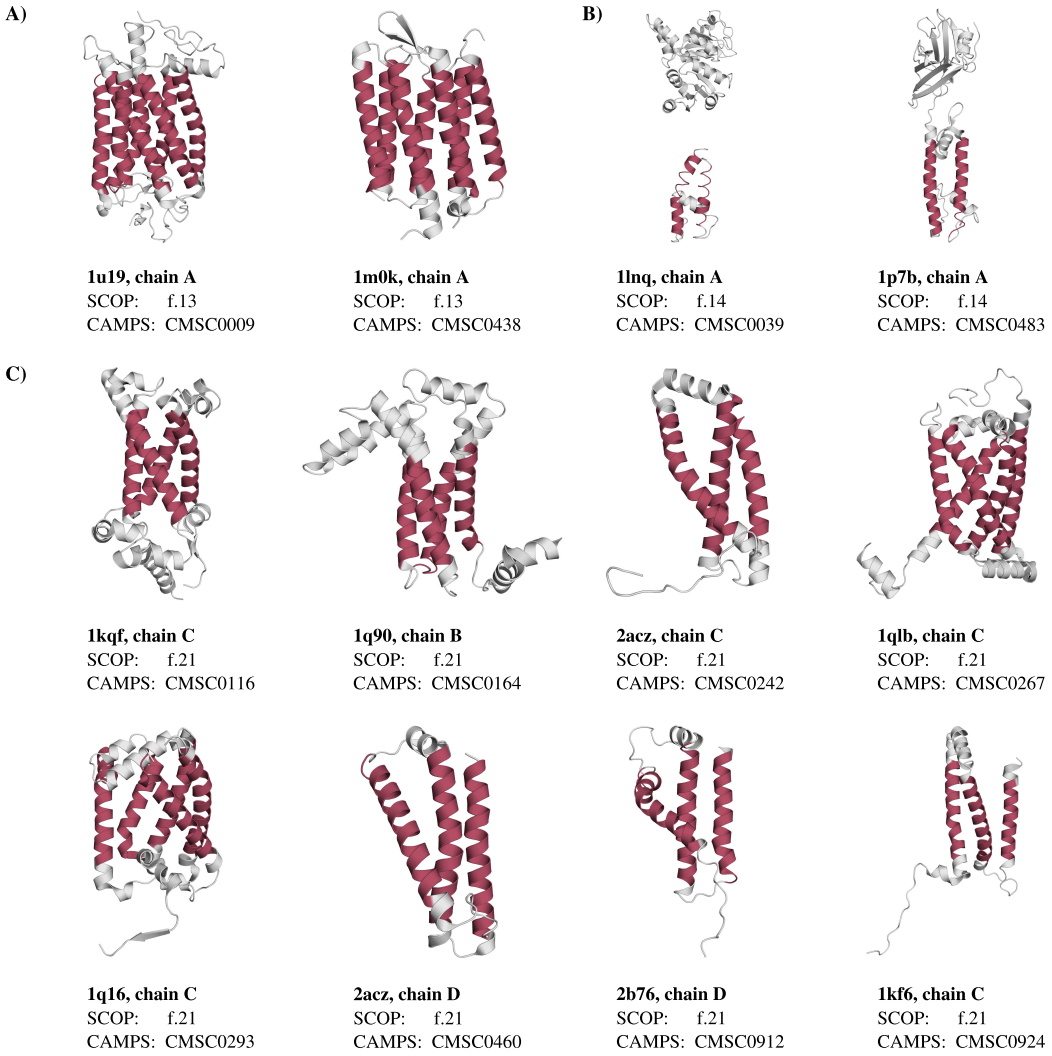

Figure 3.12: Examples of membrane proteins assigned to the same SCOP fold, but to different SC-clusters. The proteins shown are assigned to SCOP folds (A) f.13, (B) f.14, (C) f.21, (D) f.38 and (E) f.58. The corresponding SC-clusters are given for each structure. TMH coordinates were extracted from PDBTM [26]. The figure was drawn using PyMOL [66]. (*Figure continues on next page.*)

The median identity level between sequences covered by 1:1 relationships between CAMPS and SCOP ranged between 33.1% and 35.1% (Table 3.10). In case of disagreements, the median identity dropped down to 25%. The number of TMHs deviated by a factor of 0.45 and 0.58 (1:1 relationships) and 0 (1:n relationships), respectively (CATH: 0.71 and 0.58, see above). Again, a high degree of structural similarity could be observed in case of agreements (Z-score: 29.6) and a rather low similarity in case of disagreements (Z-score: 5.8).

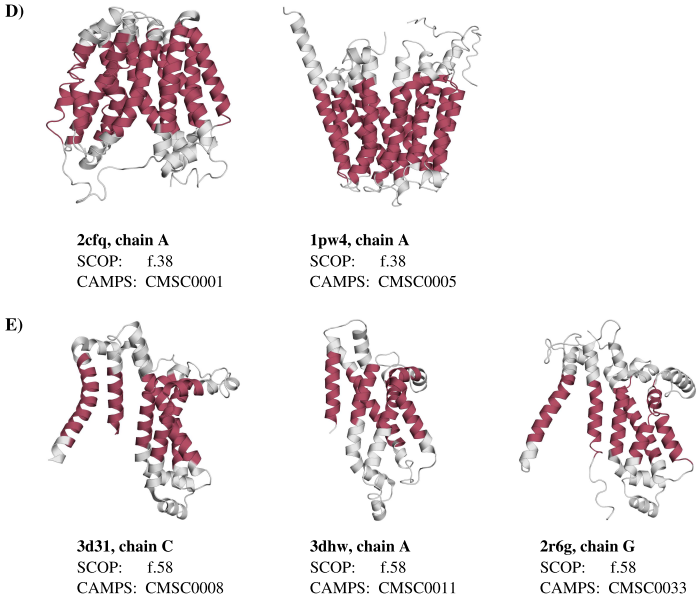

Figure 3.12: *Continued.*

In summary, CAMPS agrees reasonably well with CATH and SCOP in almost all cases. Differences in the classification occurred when the involved membrane proteins had fewer than six TMHs. For these proteins even CATH and SCOP do not agree in their classification (see section 2.3.4, page 35). One of the conclusions was that the fold definition has to be redefined for small membrane proteins that only have a limited structural diversity by integrating more fine-grained structural features. One such attempt is being undertaken here by using loop length as an additional structural determinant, and indeed the fact that some CATH and SCOP folds fall into several distinct SC-clusters suggests that the fold definition used here may offer advantages in dealing with minor differences between small membrane proteins.

3.3.8 Layered organization of the CAMPS database

The CAMPS 2.0 database provides a hierarchical organization of the membrane protein space. The first layer is composed of the SC-clusters that are designed to represent the structural level. The second layer consists of the MD- and FH-clusters representing the modeling distance and functional level, respectively. Thereby, each MD- and FH -cluster is associated with exactly one SC-cluster, while one SC-cluster is connected with one or more MD- and FH -clusters, respectively.

The first type of clusters (MD-clusters) composing the second layer of CAMPS, de-

scribes clusters whose members share a sequence identity of at least 30%. 22,360 MD-clusters were obtained covering 53% of the initial dataset with 7,047 MD-clusters having at least eight members.

By using the methodology described in the Materials and Methods section, 2,021 FH-clusters were derived (covering 75.7% of the initial dataset) with 1,319 of them having at least eight members based on the domain architecture of its members. Since FH-clusters were designed to represent the functional level and EC (Enzyme Commission) codes represent exact functions of enzymatic proteins, the FH-clusters were compared with ENZYME [181, 182] to assess the quality of the FH-clusters. As just a few membrane proteins are enzymes, this comparison applied only to a small subset of the sequences involving 127 FH-clusters and 106 EC codes. The EC code contains four numbers separated by dots, whereas the last number provides the most specific information about the catalyzed reaction. The analysis of the distribution of FH-clusters for each EC code showed that in 74 cases, EC codes agreed perfectly with FH- clusters (1:1 relationship). 32 EC codes were found to be associated with more than one FH-cluster (1:n relationship). In 11 out of 32 cases the FH-clusters were subclusters of the same SC-cluster, whereas in the remaining 21 cases the FH-clusters belonged to different SC-clusters. The reverse analysis (distribution of EC codes within each FH-cluster) revealed 101 1:1 relationships and 26 1:n relationships, whereas 16 FH-clusters were associated with several EC codes only differing in the last number of the code.

The comparison of the EC codes with the FH-clusters showed that the two systems concur well. However, there were also cases where the same code was found to be associated with several FH-clusters. In those cases where FH-clusters were connected with different SC-clusters, the most probable explanation is convergent evolution. An explanation for the other cases (several FH-clusters with same parental SC-cluster are associated with same EC code) is that homologous proteins with the same function can also be distantly related and methods with a strong reliance on sequence similarity, such as CAMPS, will not group all proteins to the same cluster. Similarly, in some cases several EC codes were found to be associated with the same FH-cluster. This may be due to the fact that not all homologous proteins having the same or similar domain architectures also have the same function.

3.3.9 Quality of the CAMPS clustering

It is important to note that several bioinformatics methods are incorporated in the CAMPS clustering procedure each influencing the quality of the final set of SC-, FH-

and MD-clusters. For example, it is well known that TMH prediction methods, such as Phobius [100], are not 100% reliable. In some cases, signal peptides are misclassified as TMHs and *vice versa*, two short TMHs are predicted as one TMH or one large TMH is split into two TMHs. Phobius has a prediction accuracy of about 68% [102] and errors in the prediction mainly affect the derivation of TMH cores that are used to generate the meta-models. The quality of the TMH cores also depends on the quality of the underlying alignments generated using ClustalW that achieves an accuracy of about 64% [304]. Large alignments are particularly challenging and usually not very accurate since ClustalW creates them progressively by a series of pairwise alignments and every time a new sequence is added new errors are likely to be introduced. Similarly, the accuracies of the FASTA and MCL algorithms determine the quality of the initial clustering. Taken together, each of the integrated methodologies can produce errors and these errors can accumulate causing a decrease in the quality of the CAMPS clustering. With more sophisticated methods to be incorporated in the future releases of CAMPS gradual progress towards significantly better clustering results may be achieved.

3.3.10 CAMPS website

Together with the release of CAMPS 2.0, a new website was developed that can be accessed at `http://webclu.bio.wzw.tum.de/CAMPS2.0/` (Figure 3.13). The browsing interface allows the user to explore the lists of all prokaryotic, eukaryotic and viral genomes that are contained in CAMPS and the lists of all SC-, FH-, and MD-clusters. In the search interface, the user can either search for specific membrane proteins or membrane protein clusters. For the protein search the user can specify different kinds of information, such as GenBank [296], UniProt [139] or Pfam [145, 146] accession numbers, organism names, GO [166, 180] terms, SCOP [162, 163] folds, etc. A similarity search against CAMPS sequences can be effected by BLAST [137]. For each individual search hit comprehensive information is being displayed such as the SC-/FH-/MD-clusters to which the protein is assigned, the topology (i.e. the sequence positions of the TMH and loop regions), the amino acid sequence, the 3D structure (if available), and numerous links to external databases (e.g. GO, Pfam, TCDB [247] etc.). At the cluster level, the user can specify an organism, a keyword, or a cross-reference (such as a Pfam accession number, CATH [164, 165] topology, SCOP fold, TCDB family etc.) to get CAMPS clusters (SC-, FH-, or MD-clusters) that contain membrane proteins that are associated with the specified search item. For each selected cluster CAMPS provides the TMH range (i.e. the range of TMH number among the cluster members), the list of associated

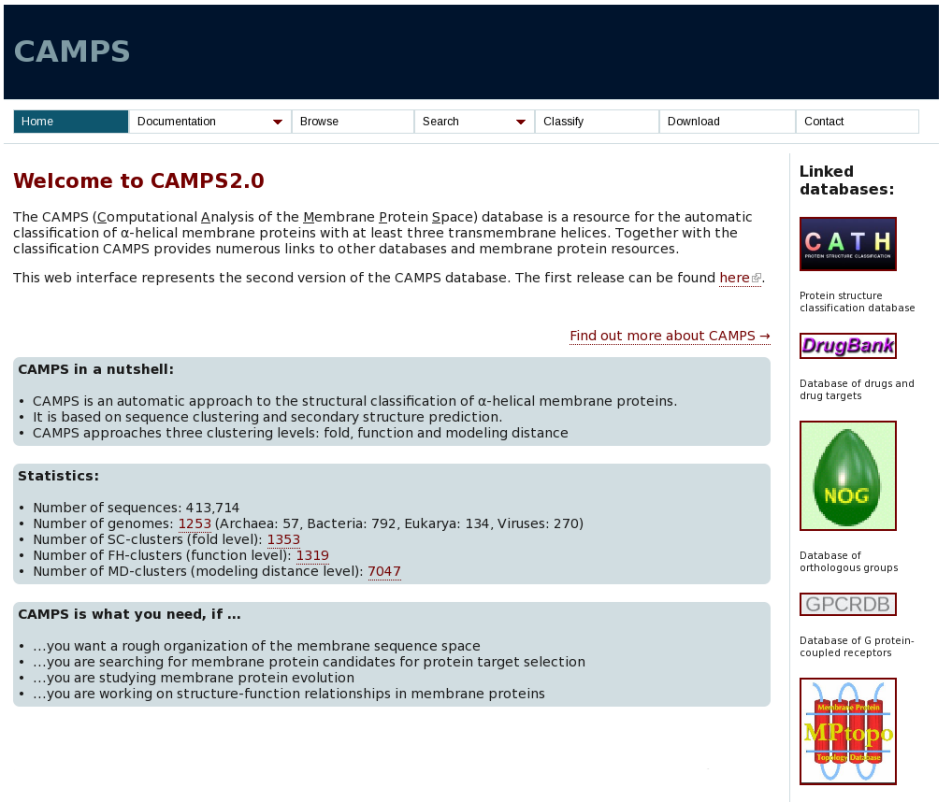

Figure 3.13: Screenshot of CAMPS 2.0 website.

genomes, the list of FH- and MD-clusters (in case of SC-clusters), the protein members, sequence alignments, the list of associated 3D structures (if available), and links to external databases. Finally, a user-submitted sequence that is not yet in CAMPS can be assigned to a SC-cluster using the full classification procedure described above.

3.4 Summary

- CAMPS provides an automatic hierarchical approach to the structural classification of α-helical membrane proteins (with at least 3 TMHs) using sequence clustering and secondary structure prediction
- CAMPS integrates structural and functional aspects
- Three clustering levels exist: fold (SC-clusters), function (FH-clusters) and modeling distance (MD-clusters)
- Major changes in new release: 1) incorporation of eukaryotic and viral genomes, 2) usage of meta-models for SC-cluster generation, 3) advanced fold determination (loop lengths)
- CAMPS classification is independent of known structures and relies on sequence similarity and topology conservation (with respect to the number of transmembrane helices and loop lengths)
- CAMPS is a comprehensive classification approach covering 413,714 membrane protein sequence from all superkingdoms
- 1,353 SC-clusters (only 53 of them with an associated structure) correspond roughly to membrane protein folds
- CAMPS agrees well with established databases (Pfam, SCOP/CATH)
- CAMPS serves as a membrane protein information resource (cross-references to many external databases such as UniProt, GO, GPCRDB, OMIM, TOPDB etc.)
- Applications: target selection, studies on membrane protein evolution

3.5 Clarification of contribution

The CAMPS 2.0 database was mainly developed by myself. Three people contributed to the development as follows. Within the context of his master thesis [278], Holger Hartmann implemented the meta-model framework that was later used for the construction of the new CAMPS database release. Angelika Fuchs supervised this master thesis. Antonio Martin-Galiano, who established the first release of the CAMPS database, assisted through constant advice. The establishment of the new release (except meta-model implementation), all comparisons against external databases and the development of the new website were done by myself. The results of this chapter were published in [305].

Chapter 4

Classification of membrane proteins based on helix-helix interactions

"Bioinformatics is a mixture of the mundane and the sublime."

(Nathan Siemers)

In the preceding chapter, a structural classification approach specifically tailored to membrane proteins was presented. In contrast to SCOP [162, 163] and CATH [164, 165], it is not based on three-dimensional structures, but on sequence similarity and topology conservation (with respect to the number of transmembrane helices and loop lengths). Hence, it offers a much more comprehensive classification since membrane protein structures are rare.

However, one shortcoming remains. Although sequence information is not the only considered fold determinant, it significantly influences the whole classification. Therefore, the approach can not deal with analogous structures resulting from convergent evolution [150]. Membrane proteins sharing the same fold, but almost no sequence similarity, will be assigned to different SC-clusters. This means that several SC-clusters may describe the same fold. To compensate this effect, the classification approach was further improved by considering helix-helix interactions as an additional fold determinant. This has been possible only recently, since existing methods for the prediction of helix-helix interactions are not suitable for membrane proteins and specific tools were missing so far (see section 4.1.2).

In this chapter, a method is described that allows to compare CAMPS SC-clusters according to their helix-helix interaction patterns visualized by helix interaction graphs.

In this way, it is possible to find and join SC-clusters with similar interaction patterns that are not related at the sequence level. After a short introduction to helix-helix interactions and their prediction, it is explained how the method is evaluated using a subset of all SC-clusters that are associated with an experimental determined structure. After defining the best parameters, the method is applied to all SC-clusters that fulfill certain criteria. Finally, the chapter closes with an analysis of the joined SC-clusters in terms of structural and functional aspects.

4.1 Introduction

4.1.1 Importance of helix-helix interactions

According to the 'two-stage' model [261], the packing of transmembrane helices is an important process in the folding of α-helical membrane proteins (see also section 3.1.2, page 47). While interhelical loops affect the helix packing of successive transmembrane helices ('short-range' interactions), helix-helix interactions can also have a 'long-range' effect in case of nonsuccessive helices [271]. The classic mode of helix-helix packing in membrane proteins is described by the 'knobs-into-holes' packing model that is already known from soluble coiled coils [25]. By analyzing a library of interacting helical pairs, it could be shown that this antiparallel motif with left-handed packing angles is the most common packing motif in membrane proteins (almost 30% of the library), which is often stabilized by the packing of small side chains appearing every seven residues [306]. There are different types of helix-helix interactions. The five most favorable are hydrogen bonds, salt bridges, aromatic interactions, closely packed small residues and closely packed valines, isoleucines, and leucines ([307] and references therein). One of the most known representatives of the fourth type is the GxxxG sequence motif, initially found in glycophorin A [308] and later shown to occur frequently in transmembrane helices in general [309].

By means of comparative analyses of helix-helix interactions in membrane and soluble proteins, Eilers *et al.* have shown that their interaction patterns differ in many aspects [310, 311]. Helices in membrane proteins are more tightly packed than in soluble proteins. Furthermore, membrane proteins have a higher diversity of residues with a distinct bias for closely packed small and polar residues. Finally, membrane proteins exhibit two general motifs for helix-helix interactions ('knobs-into-holes' and GxxxG) .

Taken together, helix-helix interactions significantly contribute to the folding of α-

helical membrane proteins. Further evidence comes from a recent study where a classification based on helix-helix interactions patterns [312] was shown to resemble the conventional fold classification of SCOP [162, 163] and CATH [164, 165]. Therefore, it is reasonable to include helix-helix interactions in the structural classification of membrane proteins as an additional fold determinant.

4.1.2 Prediction of helix-helix interactions

Since helix-helix interactions in membrane proteins differ remarkably from that of soluble proteins (see above), prediction methods that were developed for soluble proteins do not perform well on membrane proteins [313]. Hence, it is better to use a prediction method that is specifically tailored to transmembrane proteins. Only in the last four years, approaches to predict helix-helix interactions in membrane proteins became available [313–317].

The first developed method, called TMHcon [313], is a neural network based approach that uses several information such as correlated mutations, membrane protein topology and lipid-exposure scores to predict helix-helix contacts within the transmembrane helices of α-helical membrane proteins. TMHcon consists of two different neural networks, termed NN4 and NN4-D. While the former was trained on all contacts from all transmembrane helix pairs, the latter was specifically trained on long-range contacts from non-neighboring transmembrane helices. Based on a dataset of 62 membrane protein structures, TMHcon achieved a prediction accuracy of almost 26% which equals the accuracy of contact predictors available for soluble proteins. Furthermore, TMHcon allows to predict interacting transmembrane helices. This can be done by defining a threshold of required helix-helix contacts to predict a pair of helices as interacting (with one contact threshold for NN4 and one for NN4-D). Finally, if interacting helices are visualized by a so-called 'helix interaction graph' [312] (Figure 4.1), it is possible to compare the architectures of two α-helical membrane proteins.

4.2 Materials and Methods

4.2.1 Dataset

CAMPS 2.0 SC-clusters were used to apply the approach of predicted helix architectures to a comprehensive set of α-helical membrane proteins. For each of the 1,353 SC-clusters the most common number of transmembrane helices (TMHs) was defined

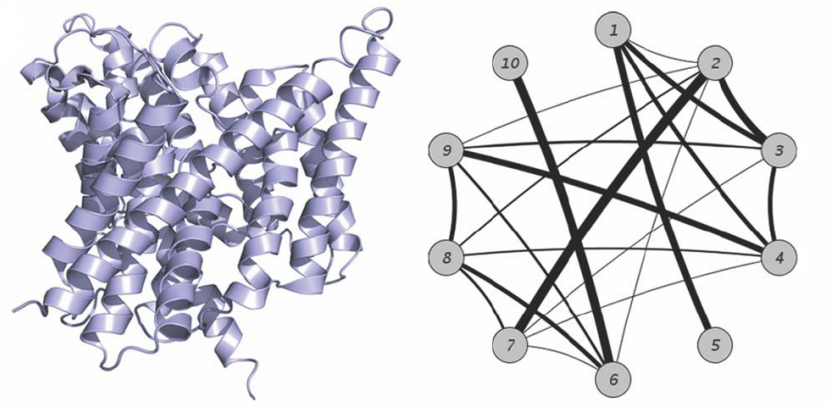

Figure 4.1: Example of a helix interaction graph. Left: Structure of the H(+)/Cl(-) exchange transporter clcA (PDB code: 1kpk) with 10 TMHs. Right: Corresponding helix interaction graph with nodes corresponding to TMHs and edges corresponding to interacting helices. Figure taken from [312].

among the members (further referred to as the representative TMH number) and those SC-clusters were selected with a representative TMH number of at least five and a structural homogeneity (reflecting the variation of the TMH number within the cluster, see [252]) of at least 0.80. The latter parameter was used to ensure that those proteins having the representative TMH number indeed represent the large majority of the corresponding cluster. From each of the 431 SC-clusters satisfying the above conditions, the 50 most divergent protein members (according to sequence identity) with a TMH number equaling the representative TMH number of the corresponding SC-cluster were retained for further consideration. In case fewer than 50 members were available, all of them were selected. By doing so, the final dataset contained 14,917 membrane protein sequences and will be further referred to as `CAMPS_SC`.

4.2.2 Prediction of consensus helix architectures

First of all, helix-helix contacts were predicted for all proteins from the 431 SC-clusters in `CAMPS_SC` using the TMHcon software [313]. The idea to derive consensus helix architectures from these contacts was as follows: first predict interacting helices for each protein using specific contact thresholds for NN4 and NN4-D (see section 4.1.2; resulting in individual helix architectures) and then generate consensus helix architectures for all 431 SC-clusters using the predicted individual helix architectures from the (50 or less; see above) selected cluster members (see Figure 4.2). Thereby, all helix interactions

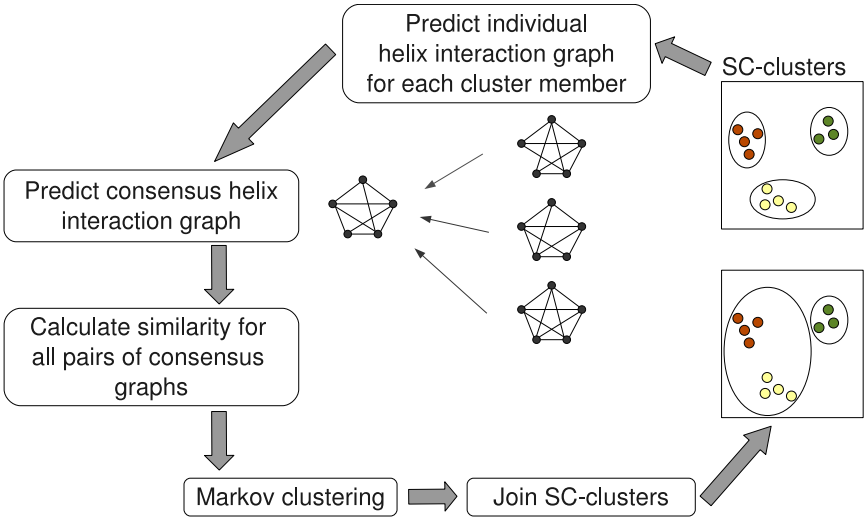

Figure 4.2: Workflow of joining SC-clusters with similar helix-helix interaction patterns using consensus helix architectures.

occurring in more individual architectures than a pre-set consensus threshold (con) are transferred to the consensus architecture.

As the required stringency of the consensus threshold is dependent on the sensitivity and selectivity of the preceding helix interaction prediction, a benchmark was performed to find optimal parameters for both the two contact thresholds (NN4 and NN4-D) for the individual architectures and the consensus threshold con. To this end, SC-clusters with a representative TMH number of at least five were selected that include a known PDB [27] structure. This was done by searching all protein sequences contained in these SC-clusters against PDBTM [26] (version 2.3) using BLAST [137] for matches with at least 95% sequence identity and at least 95% sequence coverage. Theoretical models and structures with a resolution worse than 4 Å were ignored. Furthermore, only structures were considered whose TMH number (according to TOPDB [290] or PDBTM, if the protein was not available in TOPDB) corresponds to the representative TMH number of the respective SC-cluster. If several structures were available for one SC-cluster the structure with the best resolution was chosen. By doing so, 28 SC-clusters were determined to be associated with a known structure. The 28 PDB proteins representing these SC-clusters will be further referred to as `CAMPS_TEST`.

For comparison, true helix interaction graphs were obtained for the structures in `CAMPS_TEST` by considering all TMH pairs (TMH annotations were taken from TOPDB/ PDBTM) with at least one helix-helix contact as interacting. A helix-helix contact was

defined as a residue pair (located on different TMHs) having a spatial distance of less than 5.5 Å. Subsequently, a consensus helix interaction graph was predicted for each SC-cluster in `CAMPS_TEST` using the 50 most divergent protein members (if the SC-clusters had fewer members all of them were selected). As a result, one individual helix architecture that was used as a reference and one predicted consensus architecture for each SC-cluster in `CAMPS_TEST` were available that were compared to each other. Different values for the two contact thresholds and the consensus threshold were tested and sensitivity and specificity were calculated. Thereby, sensitivity was defined as the proportion of all interacting helices in the true helix architectures obtained from the structures that were also present in the predicted consensus architectures. While specificity described the proportion of all true non-interacting helices that were also absent in the consensus architectures.

Using the parameter setting with the best sensitivity at a given specificity, consensus architectures were finally generated for all 431 SC-clusters in `CAMPS_SC` (in the same way as was done for the SC-clusters in `CAMPS_TEST`).

4.2.3 Large-scale classification of consensus helix architectures

To reveal consensus architectures representing similar membrane protein structures, all consensus architectures that were generated for the SC-clusters in `CAMPS_SC` were compared to each other using the HISS (Helix Interaction Similarity Score) scoring system [312] (see Figure 4.2). The HISS score is a measure that calculates the similarity of two helix architectures as represented by their helix interaction graphs. Thereby, only those pairs of consensus architectures with the same representative TMH number were considered for comparison. All HISS scores above different pre-defined thresholds were used to cluster the consensus architectures using the MCL algorithm [275] with varying inflation values. The inflation value is a MCL parameter that controls the granularity of the clustering (the higher the value the more fine-grained the clustering). Different combinations of the two parameters (HISS score threshold, inflation value) were tested and the final MCL clusters were validated using the Pfam-A [145, 146] annotations of the corresponding proteins. If a protein was classified to a Pfam-A family having a clan assignment, then the clan annotation was considered, otherwise the family annotation was used. Sensitivity was defined as the fraction of all proteins covered by the MCL clusters with the same Pfam-A annotation that were also found in the same MCL cluster. Similarly, specificity was calculated as the fraction of all proteins with different Pfam-A annotations that were also found in different MCL clusters. The parameter combination

with the best sensitivity at a given specificity was chosen for the final set of clustered consensus architectures.

4.2.4 GO enrichment analysis

To identify significantly enriched Gene Ontology [166, 180] (GO) terms within the set of proteins covered by the MCL clusters, the Ontologizer software [318] was used. The software requires a GO ontology file, an annotation file (with GO terms mapped to genes), so-called study sets (genes/proteins of interest) and a population set (reference set) as input. The GO slim [319] generic ontology (OBO v1.2; as of January 11, 2012) which is a subset of the whole GO containing high-level terms and the unfiltered UniProt [139] annotation file (as of December 13, 2011; both files were downloaded from `http://www.geneontology.org`) were chosen. Two separate enrichment analyses were conducted. In the first analysis (further referred to as protein class level enrichment analysis) all membrane proteins covered by the MCL clusters were grouped according to their TMH number and each group constituted a study set resulting in eleven study sets (5-15 TMHs). The union of all eleven study sets formed the population set. In the second analysis (further referred to as cluster level enrichment analysis) each MCL cluster itself described a study set using those members whose TMH number corresponded to the representative TMH number of the corresponding cluster yielding 151 study sets. The population set was again the sum of all study sets. For further consideration, all enriched GO terms were selected as returned by Ontologizer with an adjusted P-value better than or equal to 0.05. The adjusted P-values were calculated using the Bonferroni correction method (which is one of the optional parameters of the Ontologizer software).

4.3 Results and Discussion

4.3.1 Generation of consensus helix architectures

The intention of combining multiple predicted helix architectures from a number of structurally related membrane proteins into a consensus helix architecture is to have the means to represent the fold shared by a set of membrane proteins. Since structurally related membrane proteins are available by the SC-clusters from the CAMPS 2.0 database (see chapter 3, page 45), the generation of consensus helix architectures was applied to these membrane protein clusters. It is recalled that α–helical membrane proteins are classified in CAMPS 2.0 according to sequence similarity, the number of

transmembrane helices, and loop length patterns. Although sequence information is not the only feature in the classification, it has a major effect on the clustering result. Thus, it is possible that multiple SC-clusters describe similar structures originating from convergent evolution. By applying the method of consensus helix architectures to SC-clusters, it was aimed to find membrane proteins that share structural similarity, but lack sequence similarity.

First of all, optimal parameters for building consensus architectures were determined using a set of 28 SC-clusters containing known three-dimensional PDB [27] structures (dataset `CAMPS_TEST`). Consensus architectures with varying contact thresholds (for NN4 and NN4-D) and consensus thresholds con were generated for each of the 28 SC-clusters (as described in Materials and Methods) and compared to the observed individual helix architecture of the corresponding structure. Aiming at 80% specificity, the best parameter setting was achieved for C=11 (network NN4), C=11 (network NN4-D) and con=0.3 (consensus threshold) yielding a sensitivity of 61.6% (Table 4.1). For comparison, individual helix architectures were predicted for the proteins used for building consensus architectures as well and the average sensitivity and specificity was calculated. Similarly, helix architectures were predicted for the PDB proteins themselves and compared to the observed helix architectures. As can be seen in Table 4.1, consensus architectures reproduced observed helix architectures as good or even better as the average helix predictions and the predictions obtained for the PDB proteins. At 80% specificity, consensus architectures were 1.7% more sensitive than the PDB predictions and 2.2% more sensitive than the average predictions.

Using the optimal parameters at 80% specificity (C_{NN4}=11/$C_{NN4\text{-}D}$=11/con=0.3), consensus architectures were generated for all 431 SC-clusters from the `CAMPS_SC` dataset containing proteins with 5 to 15 TMHs. All pairs of consensus architectures representing the same number of TMHs (16,027 pairs in total) were compared to each other using the HISS scoring system [312]. Afterwards, all SC-cluster pairs (represented by their consensus architectures) above a given HISS score threshold were clustered using the MCL algorithm [275]. Trying different HISS score thresholds and different MCL inflation values (whereas the inflation value regulates the granularity of the clustering), the sensitivity and specificity of the MCL clusters with respect to Pfam-A [145, 146] annotations were calculated. Sensitivity was defined as the fraction of all protein pairs with the same Pfam annotation that were assigned to the same MCL cluster and specificity as the fraction of all proteins pairs with different Pfam annotations assigned to different MCL clusters. As can be seen in Table 4.2, the best parameter combination reaching

Table 4.1: Parameter optimization for generation of consensus helix architectures

Graph type	Contact threshold[a]	Consensus threshold[b]	Accuracy [%]	Sensitivity [%][c]	Specificity [%][d]
Consensus	C11/C11	0.3	71.8	61.6	80.1
	C12/C14	0.6	69.9	45.4	89.7
PDB[e]	C5/C12	-	70.9	59.9	79.8
	C12/C18	-	69.8	45.4	89.6
Average[f]	C6/C18	-	70.1	59.4	79.5
	C15/C27	-	66.0	41.3	89.5

[a] Contact threshold (NN4/NN4-D): number of required helix-helix contacts to predict a helix as interacting.
[b] Consensus threshold: fraction of individual helix architectures required to contain a helix interaction to transfer it to the consensus architecture.
[c] Sensitivity: fraction of known interacting helices that can also be found in the predicted architectures.
[d] Specificity: fraction of known non-interacting helices that are also absent in the predicted architectures.
[e] PDB: helix architectures derived from known PDB structures were compared against those that were predicted for these PDB proteins.
[f] Average: helix architectures were predicted for all proteins involved in the consensus architecture and compared against the helix architectures derived from the known PDB structures.

90% specificity for all 431 SC-clusters is given by a HISS score threshold of 0.86 and an inflation value of 2. This combination achieves a sensitivity of almost 52%. However, when sensitivity and specificity for SC-clusters with members having 5 to 7 TMHs and more than 7 TMHs were calculated separately, it was found that this parameter setting (0.86/2) is not optimal for both SC-cluster categories. For the former category (5-7 TMHs), a sensitivity of 79.7% and a specificity of 72.7% were obtained. For the latter category ($>$ 7 TMHs), specificity was already 98.5%. However, the sensitivity dropped down to 29.2%. This observation can be explained by the fact that membrane proteins with many TMHs (e.g. more than seven) have a higher potential of structural variability than small membrane proteins. Thus, different parameter settings were defined for the two categories both achieving similar values of sensitivity and specificity. Aiming at 90% specificity, a HISS score threshold of 0.95 and an inflation value of 1.1 were shown to perform best for the first category achieving almost 55% sensitivity. For the second category, the combination 0.75/1.1 performed with 90% specificity and almost 50% sensitivity (Table 4.2). It has to be noted that these values are based on both Pfam-A family and clan (= group of related families [149]) annotations (see Materials and Meth-

Table 4.2: Parameter optimization for MCL clustering of consensus helix architectures

SC-cluster dataset[a]	HISS score threshold	Inflation value	Sensitivity [%][b]	Specificity [%][c]
All	0.70	1.1	67.1	85.3[d]
	0.86	2	51.8	89.5
$\leq$ 7 TMHs	0.84	5	66.9	78.8
	0.95	1.1	54.9	91.2
$>$ 7 TMHs	0.70	1.1	54.2	84.3[d]
	0.75	1.1	49.9	89.5

[a] All: All SC-clusters from the classification dataset; $\leq$ 7 TMHs: SC-clusters with members having up to seven TMHs; $>$ 7 TMHs: SC-clusters with members having more than seven TMHs.
[b] Sensitivity: Fraction of all protein pairs having the same Pfam annotation that were assigned to the same MCL cluster using the respective HISS score threshold and inflation value.
[c] Specificity: Fraction of all protein pairs having different Pfam annotations that were assigned to different MCL clusters using the respective HISS score threshold and inflation value.
[d] Lower specificities were not obtained in the tested set of HISS score thresholds and inflation values.

ods). Calculating the values for family and clan annotations separately shows that an even higher sensitivity is achieved (for 0.95/1.1 and 0.75/1.1, respectively), when family assignments are considered alone ($\leq$ 7 TMHs: 74.8% sensitivity, 90.7% specificity; $>$ 7 TMHs: 93.1% sensitivity, 88.8% specificity). Using only clan assignments results in less than 50% sensitivity at about 90% specificity ($\leq$ 7 TMHs: 48.1% sensitivity, 93.5% specificity; $>$ 7 TMHs: 44.9% sensitivity, 89.6% specificity).

Accordingly, the parameter combinations 0.95/1.1 (for SC-clusters with up to 7 TMHs) and 0.75/1.1 (for SC-clusters with more than 7 TMHs) were used for the final clustering of all consensus architectures. By doing so, the 431 SC-clusters were joined into 151 MCL clusters, whereas 111 of them are singleton clusters (i.e. clusters only containing one SC-cluster) and 40 MCL clusters contain two or more SC-clusters (Table 4.3).

4.3.2 Validation of MCL clusters

Comparison with Pfam

Using the Pfam-A family and clan annotations of the membrane proteins involved in the clusters, the two clusterings (SC-clusters and MCL clusters (= joined SC-clusters)) were

Table 4.3: TMH distribution among SC-clusters and MCL clusters

Number of TMHs	Number of SC-clusters	Number of MCL clusters			Reduction
		Singleton[a]	Non-Singleton[b]	Total	factor[c]
5	97	25	18	43	2.3
6	121	26	8	34	3.6
7	68	28	3	31	2.2
8	24	1	1	2	12.0
9	12	3	1	4	3.0
10	37	16	4	20	1.9
11	27	0	1	1	27.0
12	30	6	1	7	4.3
13	5	0	1	1	5.0
14	8	4	2	6	1.3
15	2	2	0	2	1.0
Total	431	111	40	151	2.9

[a] MCL cluster containing only one SC-cluster.
[b] MCL cluster containing two or more SC-clusters.
[c] Number of SC-clusters divided by total number of MCL clusters.

compared to each other. It is important to mention at this point that it is not intended to fully reproduce the Pfam clustering. SC-clusters and MCL clusters are designed to represent sets of structurally similar membrane proteins likely to share the same fold, while Pfam is a sequence-based approach and does not consider structural features. Nevertheless, Pfam is used as a reference (and not SCOP [162, 163] or CATH [164, 165]) since Pfam annotations are available for a large majority of membrane proteins (about 70%; data not shown here). 66.5% of all protein pairs having the same Pfam-A family annotation were also found in the same SC-cluster (sensitivity), while 99.9% of all pairs with different annotations were assigned to different SC-clusters (specificity). Based on the clan annotations, a sensitivity of 11.3% and a specificity of 100% could be obtained for the SC-clusters. Through the further clustering of SC-clusters using helix architectures, the sensitivity could be improved significantly at the cost of a slightly reduced specificity. The MCL clusters approach a sensitivity of 81.6% (at 94.8% specificity) and 42.0% (at 96.1% specificity) at the family and clan level, respectively. The most important result here is that the MCL clusters are 30.7% more sensitive than SC-clusters when compared to Pfam clans. A clan is described as a set of related Pfam families that have arisen from a single evolutionary origin [149]. Based on structural information and profile-profile comparisons, some Pfam clans group together large, divergent families.

Thus, Pfam clans represent a perfect evaluation when investigating the clustering of structurally similar proteins with low sequence similarity. Based on these results, it can already be concluded that the further clustering of SC-clusters led to a considerable improvement, particularly as regards the clustering of divergent membrane proteins. Three cases will be further described in detail demonstrating the improvement.

Case 1: G protein-coupled receptors

In the first two cases, groups of SC-clusters being associated with different Pfam families from the same clan, could be found within the same MCL cluster. The first example involves 12 SC-clusters that all contain members being assigned to Pfam clan CL0192 ('Family A G protein-coupled receptor-like superfamily'). Through the further clustering process using helix architectures, these SC-clusters were grouped into the same MCL cluster due to their similar consensus architectures (Figure 4.3A). The G protein-coupled receptors (GPCRs) are known to be the largest and most diverse protein superfamily in the mammalian genome and are further divided into five main families [134, 320, 321] (with family A being the largest family of GPCRs). All GPCRs share a common structure of a seven transmembrane helix bundle [322], while sequence similarity is rather low among distant GPCRs. To verify whether it was possible to join SC-clusters that capture structural similarity beyond sequence similarity, the average pairwise sequence identity between all protein pairs of the same SC-cluster and all pairs from different SC-clusters (out of the set of 12 SC-clusters that were joined together; Figure 4.3A) was calculated. The average pairwise sequence identity was found to be 25.4% and 14.1% for protein pairs assigned to the same and different SC-clusters, respectively. Once again, these results confirm that the proposed method is able to identify membrane proteins with similar structures lacking significant sequence similarity.

Case 2: APC superfamily

Similarly, another five SC-clusters were found that are all linked with Pfam clan CL0062 ('APC superfamily') and are classified to the same MCL cluster (Figure 4.3B). While GPCRs are only present in eukaryotes, amino acid/polyamine/organocation (APC) transporters are numerous in all domains of life. The twelve families of the APC superfamily differ in their taxonomic distribution (some families are represented only in bacteria, other only in eukaryotes etc.) and the number of transmembrane segments. While most of them exhibit twelve segments, some APC transporters were also found

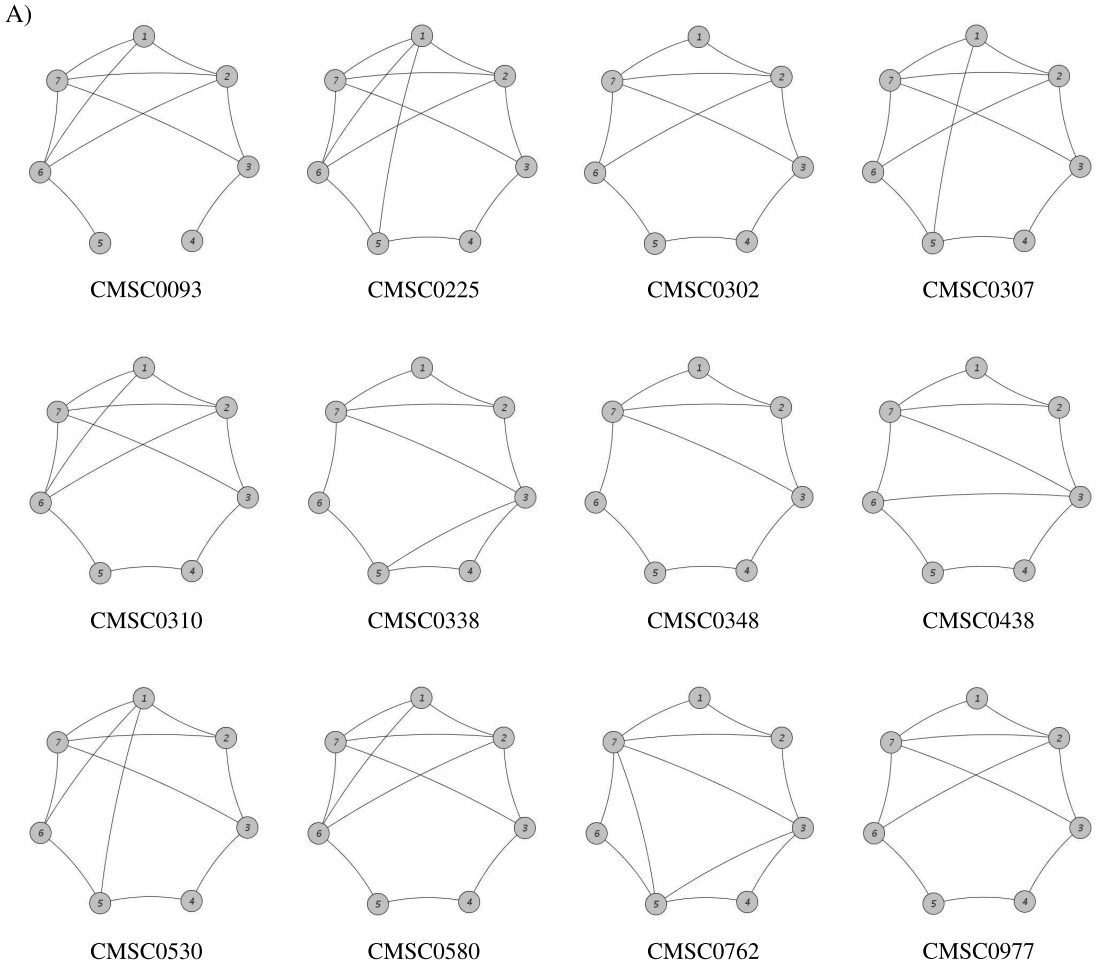

Figure 4.3: Consensus helix architectures from joined SC-clusters. (A) All corresponding SC-clusters belong to Pfam clan CL0192 ('Family A G protein-coupled receptor-like superfamily'). (B) All SC-clusters belong to clan CL0062 ('APC superfamily'). Nodes correspond to transmembrane helices, edges represent interacting helices. (*Figure continues on next page.*)

to contain 11 (which corresponds to the representative TMH number of the five SC-clusters), 12 or 14 transmembrane segments [323]. As in the previous case, the average pairwise sequence identity was calculated. For protein pairs originating from the same SC-cluster, the average identity was 27.0%. And for protein pairs assigned to different SC-clusters (joined into the same MCL cluster) the sequence identity was 14.5%.

Case 3: SC-clusters with similar 3D structures

The last case is special as it shows two SC-clusters linked with known 3D structures being similar to each other and that were joined into the same MCL cluster (Figure

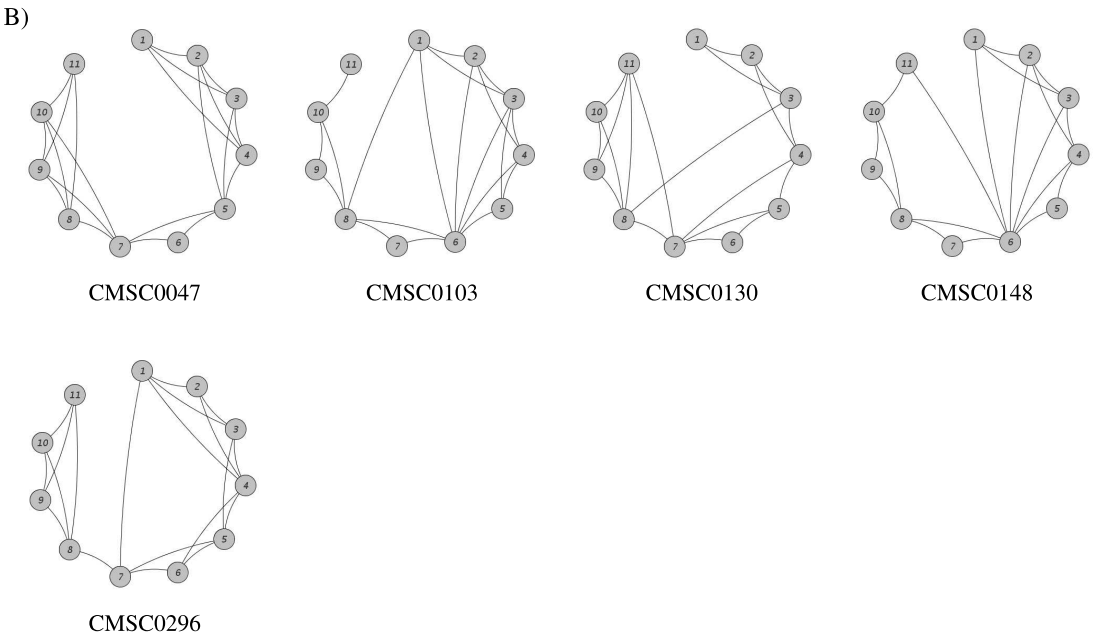

Figure 4.3: *Continued.*

4.4). SC-Cluster CMSC0058 contains the archaeal aquaporin AqpM (2f2b, chain A) and CMSC0180 a bacterial formate channel (3kly, chain A). By comparing the two structures using DaliLite [225], a high degree of structural similarity (Z-score: 17.6) was observed. While a Z-score of at least 2.0 indicates a common fold, a Z-score above 20 means that two structures are true homologues. At the same time, the sequence identity of the two channels is only 15.3%. It is interesting to note that both structures are also classified to the same CATH [164, 165] fold ('Glycerol uptake facilitator protein') and to the same OPM [288] superfamily ('Major Intrinsic Protein (MIP)/FNT superfamily'). Furthermore, Theobald and Miller also revealed that the two channels share a common structural fold in the absence of sequence similarity raising questions about the evolution of membrane proteins [324].

Taken together, the results clearly indicate that using the proposed method of predicted helix architectures it is possible to identify structurally similar membrane proteins lacking sequence similarity. These cases are of significant importance as they provide insight into the previously poorly understood evolution of membrane proteins.

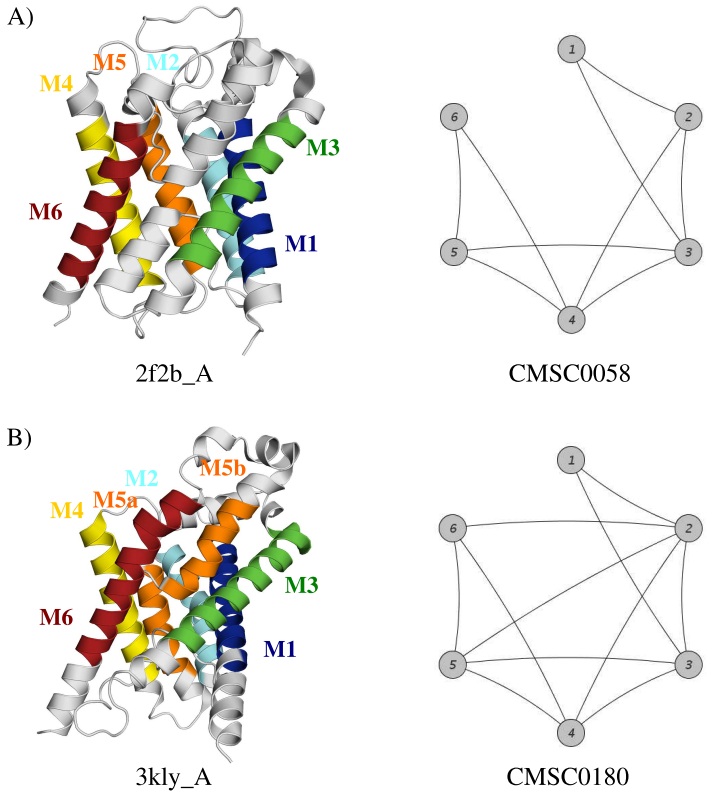

Figure 4.4: Example of two SC-clusters that were joined together. Both SC-clusters contain structures that show a very similar transmembrane helix packing. (A) Left panel: Representative structure (PDB code: 2f2b, chain A) of SC-cluster CMSC0058. Right panel: Consensus helix architecture for SC-cluster CMSC0058. (B) Left panel: Representative structure (PDB code: 3kly, chain A) of SC-cluster CMSC0180. Right panel: Consensus helix architecture for SC-cluster CMSC0180. Both structures contain six transmembrane helices (M1-M6) colored differently. In case of 3kly_A, the fifth helix is interrupted (M5a, M5b). Transmembrane helix coordinates were extracted from PDBTM [26].

4.3.3 Exploring the membrane protein structure space

Structural similarity of transporter families

Given the almost 3-fold decrease in membrane protein clusters after the clustering based on helix architecture similarity, it was interesting to further investigate whether proteins with a certain number of TMHs (further referred to as membrane protein class) where especially prone to be clustered together based on similar helix interaction patterns. This could give new insights into the structural similarity of a certain membrane protein class lacking sequence similarity.

As can be seen in Table 4.3, two membrane protein classes can be identified that

indeed contribute most substantially to the reduction of protein clusters, namely the 8 TMH and the 11 TMH class. Most noticeably, 23 out of 24 SC-clusters with members having eight TMHs were joined into one MCL cluster. By doing so, the MCL cluster grouped together different transporter proteins including ABC transporters, nickel transporters, NADH dehydrogenases and P-type ATPases, as well as many proteins of unknown function. Hence, it seems that these transporters have a common structural core, a structural similarity that could not be revealed using the SC-clustering approach. In fact, previous studies based on hydropathy profile analysis also revealed structural similarities between different families of secondary transporters not related in amino acid sequence indicating distant evolutionary relationships [325–328]. Therefore, it is hypothesized that structural similarity is likely to be found in other transporter families as well, and the MCL cluster (joining 23 SC-clusters) is one more case of structurally related transporters that arose either by divergent or convergent evolution [329]. The second remaining MCL cluster containing 8 TMH proteins is a singleton cluster comprising only one SC-cluster (CAMPS code CMSC0049). Given that CMSC0049 includes proteins of unknown function, it may be an interesting target for structural genomics.

In case of the 11 TMH class, all 27 SC-clusters were joined into a single MCL cluster (Table 4.3). Again, most of the SC-clusters represent different transporters, such as sulfate, ammonium, metal ion and amino acid transporters, as well as sodium/alanine and sodium/glutamate transporters. Furthermore, the grouped SC-clusters are linked with different Pfam clans (CL0062 'APC superfamily', CL0064 'CPA/AT transporter superfamily', CL0182 'IT superfamily'). Taken together, for the special case of 8 and 11 TMH proteins, it is assumed that (almost) all SC-clusters were combined into one MCL cluster because the considered transporter families share a common structural core.

Structural diversity of membrane protein classes

In a further analysis, it was examined whether certain membrane protein classes occur more often than others among the SC-clusters and MCL clusters. Given that SC-clusters and MCL clusters group together structurally related membrane proteins likely to share the same fold, this analysis allows to draw conclusions about the structural diversity of certain membrane protein classes. One factor that influences the structural diversity is the abundance of the membrane protein class itself, i.e. how many membrane proteins exist with a certain number of TMHs. Therefore, the distribution of membrane proteins with respect to their TMH number was studied first.

In 1998, Jones already examined the question of whether certain transmembrane

topologies occur more often than others [3]. This and other genome-wide analyses [4, 117, 131–133] have shown that proteins with 6 and 12 TMHs are predominant in uni-cellular organisms and constitute small-molecule transporters, sugar transporters and ABC transporters (see section 1.1.4, page 11). In contrast, proteins with 7 TMHs are abundant in *C. elegans* and human which can be explained by the high abundance of G-protein coupled receptors. For the membrane proteins in the dataset used in this work (`CAMPS_SC`) similar trends could be observed (see Figure 4.5A and Table 7.4 in the Appendix), except for the 12 TMH class. In the `CAMPS_SC` dataset, proteins with 12 TMHs were more abundant in eukaryotic than in prokaryotic proteins (Eukaryota: 16.0%, Archaea: 9.2%, Bacteria: 13.0%). This difference may be caused by the different datasets used. While previous analyses were based on not more than four eukaryotic genomes, the CAMPS database incorporates 134 eukaryotic genomes in total (see section 3.3.1, page 61). Hence, it can be expected that the distribution shown here (Figure 4.5A) is more representative.

Next, the distribution of TMH classes among the SC-clusters and MCL clusters was investigated (see Figure 4.5B and 4.5C and Table 7.4 in the Appendix). It is recalled that the distribution at the protein level reflects the abundance of the TMH classes, while the distribution at the cluster level rather displays their structural diversity. As can be seen in Figure 4.5A and Figure 4.5B, the distribution for proteins and SC-clusters are more or less the same. However, slight differences can be observed in the 8, 9 and 12 TMH class. While proteins with 8 and 9 TMHs are almost equally abundant as proteins with 10 TMHs, SC-clusters with 8 and 9 TMH proteins are less frequent than with 10 TMH proteins. Furthermore, the distribution in the range of 10 to 12 TMHs is more uniform at the SC-cluster level as at the protein level. In contrast to these minor differences, the distribution among MCL clusters differs significantly from both distributions among proteins and SC-clusters. The first striking difference between MCL and SC-clusters can be observed in the range of 5 to 7 TMHs. While the 6 TMH class is abundant among prokaryotic SC-clusters and the 7 TMH class among eukaryotic SC-clusters, this is not true for MCL clusters. On the contrary, there is a steady decrease in the number of MCL clusters and this relates to all superkingdoms. The second discrepancy can be found in the range of 8 to 15 TMHs. Compared to SC-clusters, MCL clusters with 8 and 11 TMH proteins are clearly less common (see above), as well as with 12 TMH proteins. Given that the applied clustering approach based on helix architectures is able to detect distant relationships, it is expected that this difference can be explained by the grouping of several SC-clusters that contain distantly related membrane proteins.

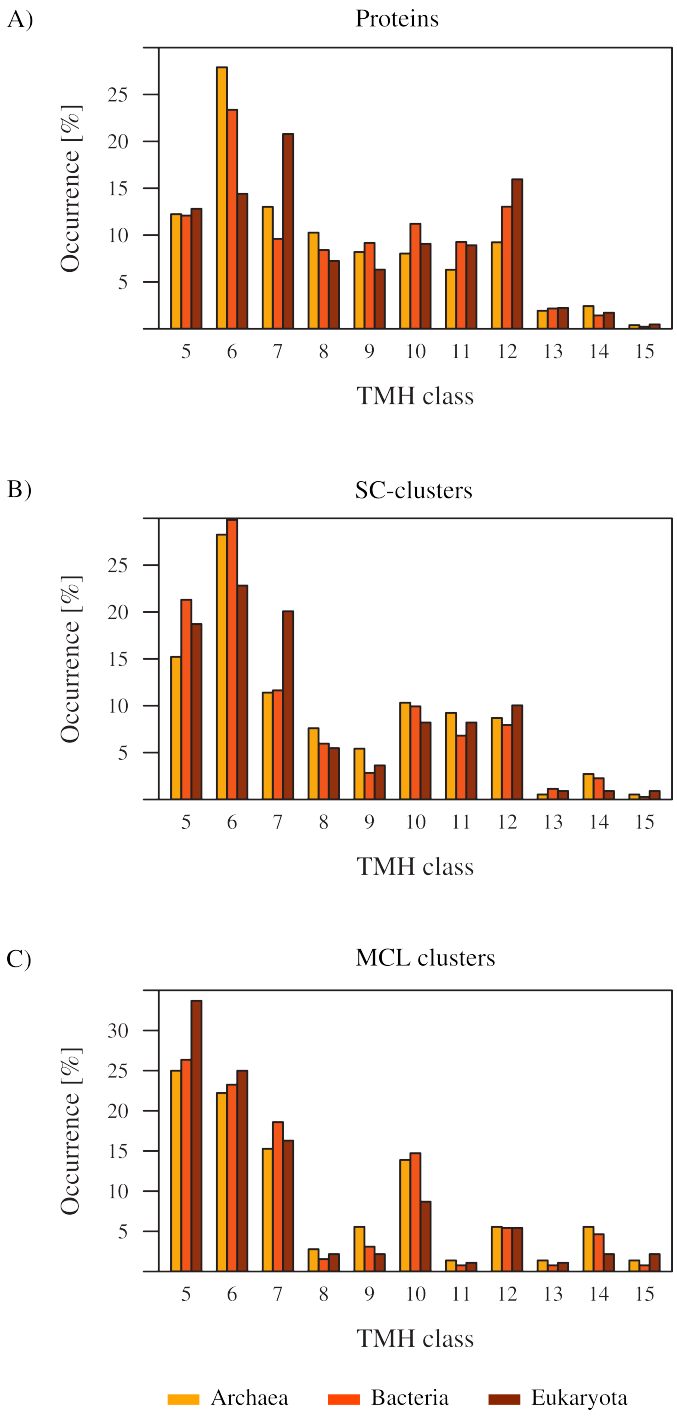

Figure 4.5: Occurrence of TMH classes among proteins, SC-clusters and MCL clusters. (A) Percentage of proteins with a certain number of TMHs. Percentage of SC-clusters (B) and MCL clusters (C) with a certain representative TMH number.

As the MCL clusters are the focus of this work, it was of particular interest why some TMH classes occur more often among MCL clusters than others and thus may exhibit a higher structural variability. Although the abundance of proteins with 8 to 12 TMHs is rather homogeneous (see Figure 4.5A), there are more MCL clusters with 10 and 12 TMH proteins (see Figure 4.5C) suggesting that these membrane protein classes are structurally more diverse. It is known from previous publications that internal gene duplications are a very common mechanism in membrane protein evolution [17, 120–125] (see section 1.1.3, page 8). According to these studies, proteins with 10 and 12 TMHs seem to have evolved through a complete gene duplication. Therefore, it is assumed here that proteins with 10 and 12 TMHs show a higher structural diversity as they originated from proteins with 5 and 6 TMHs that themselves are distributed among many different MCL clusters. Similarly, it is suggested that for the same reason proteins with 14 TMHs adopt more different structures than proteins with 13 TMHs.

Functional diversity of membrane protein classes and MCL clusters

Another explanation for the differences might be that some TMH classes are associated with more functions than others and thus structural variability correlates with functional diversity. To investigate whether this is true or not, a GO [166, 180] (Gene Ontology) term enrichment analysis was performed for both the TMH classes and the MCL clusters. Given that the assumption holds, it is expected that TMH classes associated with many MCL clusters also have several enriched GO terms. In fact, the highest number of distinct GO terms could be found for proteins with 5, 6, and 7 TMHs (having 6 to 7 terms; see Table 4.4 and Table 7.5 in the Appendix). So in this case, functional diversity seems to implicate structural diversity. The same applies to the 10 TMH and 14 TMH classes (5 and 4 terms, respectively) as compared to the remaining classes in the range of 8 to 15 TMHs. However, it has to be noted that the difference in the number of distinct GO terms is often too subtle in order to draw clear conclusions regarding the correlation between structural and functional diversity. For example, the 13 TMH and 14 TMH classes differ remarkably in their structural variability (according to the number of different MCL clusters), but only slightly in their functional variability (3 and 4 enriched GO terms, respectively). Although the number of GO terms for the 13 TMH class drops down to 2, if the terms ‘transport’ and ‘transmembrane transport’ are considered as only one enriched term, as the second is just a more detailed description of the first one.

GO enriched terms were used to draw conclusions about the functional diversity of

Table 4.4: Enriched GO terms in membrane protein (MP) classes. For each membrane protein class (i.e. set of proteins with a respective number of TMHs), the number of proteins that are annotated with the respective GO term is given (gray colored cells). If no number is given, the respective count is zero. All listed GO terms are enriched with a P-value ≤ 0.05.

	MP class											MP classes[c]
	5	6	7	8	9	10	11	12	13	14	15	
Size[a]	15,627	29,439	14,358	10,806	11,403	14,145	11,896	17,277	2,805	1,880	324	
Annotations[b]	4,283	13,741	4,472	3,997	4,702	5,763	5,436	11,389	1,495	556	67	
GO term	Number of proteins in MP class with annotated GO term											
Carbohydrate metabolic process							10					1
Cell cycle	137				252	427						3
Cell differentiation				6								1
Cell division	140				107	244						3
Cell proliferation			5									1
Cellular component assembly		132	56	152								3
Cellular nitrogen compound metabolic process										4		1
Embryo development						5						1
Lipid metabolic process	22	45	20									3
Locomotion			12									1
Photosynthesis						3						1
Protein complex assembly		132	56	152								3
Protein targeting	240	456				235						3
Reproduction			6									1
Response to stress								34	32			2
Signal transduction	153	275	200							53		4
Small molecule metabolic process										2		1
Transmembrane transport							2,707	8,866	1,221	391	40	5
Transport		11,303					5,068	11,236	1,429			4
Vesicle-mediated transport	11											1
GO terms[d]	6	6	7	3	2	5	3	3	3	4	1	

[a] Number of proteins in the membrane protein class.
[b] Number of proteins in the membrane protein class with GO annotations.
[c] Number of membrane protein classes that contain this enriched GO term.
[d] Number of enriched GO terms in this membrane protein class.

different membrane protein classes. However, it is clear that this approach is not necessarily straightforward. First, one protein can be associated with multiple GO terms (multi-functional protein). Second, one MCL cluster can be linked with multiple GO terms (multi-functional cluster). Therefore, the statement that the more GO terms can be found the higher the functional diversity might be misleading. To investigate the effect of multi-functional proteins, it was looked at the protein annotations and searched for GO terms that frequently occur in combination with other GO terms. This was found to occur in three cases. First of all, if proteins were annotated with the GO term 'cell division', the 'cell cycle' annotation was available as well. This case concerns the 5, 9 and 10 TMH classes. Secondly, the same applied to the combination 'protein targeting' and 'transport' concerning the 6 TMH class. Thirdly, the annotations 'transport', 'transmembrane transport' and 'response to stress' also occurred in combination affecting the 11, 12 and 13 TMH classes.

Similarly, to analyze the effect of multi-functional MCL clusters, a separate GO enrich-

ment analysis was performed (see Table 7.6 in the Appendix). Except for MCL clusters with 15 TMHs, significantly (i.e. P-value ≤ 0.05) enriched GO terms could be found for all MCL clusters. Compared to the protein class level enrichment analysis, additional GO terms were found to be enriched that were, however, not considered for further analyses. In total, seven MCL clusters were found to contain multiple enriched GO terms (mcl5_5tmh, mcl12_5tmh, mcl2_6tmh, mcl10_6tmh, mcl6_7tmh, mcl1_8tmh, mcl4_10tmh; see Table 7.6 in the Appendix), whereas terms that were already found to occur in combination were considered as only one term. Additionally, it was also observed that several MCL clusters from the same TMH class are associated with the same GO terms. For example, in case of the 7 TMH class, four clusters (mcl2_7tmh, mcl3_7tmh, mcl5_7tmh, mcl6_7tmh; see Table 7.6 in the Appendix) contain the term 'signal transduction'.

To summarize, taking into account the effects of multi-functional proteins, multi-functional MCL cluster and MCL clusters with the same functional annotations, it can be concluded that the structural diversity of MCL clusters can be explained by functional diversity, at least to some extent.

4.4 Summary

- Using consensus helix architectures it was aimed to find SC-clusters representing the same membrane protein fold
- Consensus helix architectures reproduce observed helix architectures better than individual helix architectures
- 431 SC-clusters were joined into 151 MCL clusters
- MCL clusters are 30.7% more sensitive than SC-clusters (if sensitivity is described as the fraction of all protein pairs with the same Pfam clan annotation that are also assigned to the same MCL/SC-cluster)
- Membrane proteins with similar structures but almost no sequence similarity could be grouped together using consensus helix architectures
- Some membrane protein classes show a higher structural diversity than others
- Structural diversity seems to correlate with functional diversity

4.5 Clarification of contribution

The idea of consensus helix architectures was initially developed by Angelika Fuchs. Barbara Hummel implemented the approach within the context of her bachelor thesis [330]. She was supervised by Angelika Fuchs, who revised the approach based on an intermediate version of the CAMPS 2.0 database. I applied and further developed this approach based on the latest CAMPS release. The benchmark, the application and the analysis of the joined SC-clusters were done by myself. Standalone tools for the prediction of consensus graph and their comparison were kindly provided by Angelika Fuchs.

Chapter 5

Conclusions and Outlook

> "The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!' (I found it!) but 'That's funny...'."
>
> *(Isaac Asimov)*

The overall objective of this thesis was to advance membrane protein research in the field of structure classification. A comprehensive structure classification system allows to explore structure-function relationships in membrane proteins that give valuable insights into evolution. To this aim, existing approaches originally developed for soluble proteins were first evaluated to figure out how they deal with membrane proteins. In a second step, a hierarchical large-scale classification approach was developed that covers both structural and functional aspects and is specifically tailored to α-helical transmembrane proteins. This chapter summarizes the main conclusions that can be drawn from this thesis and provides a short outlook on future work.

5.1 Structure space of small membrane proteins seems to be highly continuous

In chapter 2, the first analysis of the classification of membrane proteins in SCOP and CATH was presented. To this end, the general occurrence of membrane proteins and folds within the two databases was investigated, as well as the differences in their domain and fold assignments.

Given the current census of membrane protein structures, a reasonable agreement

between SCOP and CATH was observed for all domains with six or more transmembrane helices. Membrane proteins with more transmembrane helices are structurally restricted and hence likely to be more similar between each other than soluble helix bundles. However, the spectrum of possible structural diversity (such as tilted or reentrant helices) also increases with a growing number of transmembrane helices. Therefore, membrane proteins with six or more transmembrane helices seem to be sufficiently diverse to allow for a structural classification comparable to that of soluble proteins.

While domains with more than five helices are mostly classified consistently between SCOP and CATH, the majority of all discrepancies (such as differing domain assignments and fold overlap problems) affect proteins with less than six transmembrane helices. Single transmembrane helix, two helix hairpin and four helix bundle domains are among the most prevalent classes of membrane proteins in both SCOP and CATH, and their classification differs remarkably with almost no fold containing the same set of domains. These findings indicate that the structural space of small membrane proteins is highly continuous, making their classification intrinsically more difficult.

5.2 Fold definitions developed for soluble proteins are not equally suitable for membrane proteins

The comparative analysis of SCOP and CATH has shown that membrane proteins present a particular challenge for structure-based classification requiring a separate treatment. Given that soluble and membrane proteins differ remarkably, it is not surprising that the fold definition originally developed for soluble proteins can not automatically applied to membrane proteins as well. Although membrane proteins are structurally restricted by the lipid bilayer, they provide a variety of different functions. To fully address the question how this is accomplished, the meaning of the term ‘fold’ should be redefined, in particular for membrane proteins with a limited number of helices. This can be done, for example, by considering more fine-grained structural features of membrane protein structures such as helix-helix interactions and helix packing, the distribution of helix tilt angles, the presence of reentrant regions, as well as functional features.

5.3 CAMPS 2.0 reasonably agrees with sophisticated sequence and structure classification approaches

In chapter 3, a new release of the CAMPS database was described. The CAMPS database represents an automatic hierarchical classification approach tailored towards membrane proteins. The principal difference between CAMPS and other protein family databases is the reliance on both sequence information and predicted structural features. Three major changes have been implemented in the new release: i) eukaryotic and viral membrane proteins have been incorporated in the classification approach, ii) loop length patterns are used as an additional fold determinant, and iii) empirical rules to delineate structural families have been replaced by higher-order hidden Markov models (meta-models). As long as structure determination of membrane proteins lags far behind similar efforts for globular proteins, comparative sequence analysis and structure prediction will remain the principal tools for organizing the universe of membrane protein folds.

In contrast to the SCOP and CATH hierarchies, CAMPS 2.0 offers a large-scale classification not restricted to proteins with known structure. More importantly, for known structures the CAMPS 2.0 classification is in reasonable agreement with SCOP and CATH. Compared to purely sequence-based approaches, such as Pfam, CAMPS 2.0 provides more inclusive clusters, consistent with the notion that several functional families may share the same fold.

5.4 Existing membrane protein structures represent only a tiny fraction of the whole membrane protein fold space

Based on the currently available set of completely sequenced genomes, the classification approach applied in CAMPS 2.0 yielded 1,353 SC-clusters. Members of a SC-cluster are structurally related and are expected to share the same fold. Although several SC-clusters may describe the same fold, the number of SC-clusters can be used as an upper bound for the estimation of membrane protein folds. Given that only 53 out of 1,353 SC-clusters (4%) are associated with a known structure, it can be concluded that the membrane protein structures available in the PDB database represent only a small sample of the whole membrane protein fold space. Therefore, it can be expected

that with new structures becoming available in the future the spectrum of structural diversity will again increase considerably.

5.5 Incorporation of loop lengths and helix-helix interactions improves the classification of membrane proteins

The underlying principle of the CAMPS database is a hierarchical scheme with three clustering levels: fold (SC-clusters), function (FH-clusters) and modeling distance (MD-clusters). The SC-clusters are at the top level of the hierarchy and roughly correspond to membrane protein folds. In the initial release of CAMPS, sequence similarity and the number of transmembrane helices were used as fold determinants. In this thesis, the database was progressively developed by gradually considering additional fold determinants.

In the first round (see chapter 3, page 45), loop length patterns were included through the application of meta-models. Loops were included to the classification approach as it is known that they provide an additional source of structural variety. In fact, similar membrane protein architectures can have diverse loop lengths so that they can be further divided into subgroups. The comparison between CAMPS 1.0 and CAMPS 2.0 has shown that the incorporation of loop length patterns produces much more structurally homogeneous SC-clusters implying that loop lengths are an important determinant of membrane protein structure.

In the second round (see chapter 4, page 95), consensus helix architecture graphs were utilized to search for SC-clusters likely to describe the same fold due to convergent evolution. The joined SC-clusters were termed MCL clusters. By comparing the SC-clusters and MCL clusters against Pfam, it could be shown that the sensitivity (defined here as the fraction of protein pairs having the same Pfam annotation that were assigned to the same SC- or MCL cluster) could be improved remarkably. Most notably, the sensitivity increased by 30.7% when Pfam clans were used as a reference. As Pfam clans are known to group large divergent families, these results indicate that the further clustering of SC-clusters brought a significant improvement.

5.6 Some membrane protein topologies occur more often than others

In the last step, the MCL clusters were investigated in terms of structural and functional aspects. It appeared that in the range of 8 to 15 TMHs, the 10 TMH, 12 TMH and 14 TMH topologies are more prominent among the MCL clusters than the others implying that these topologies share a higher structural diversity. By performing a GO enrichment analysis, the question was addressed whether structural diversity correlates with functional diversity. However, the question could not be completely answered, since the number of distinct GO terms differed only slightly between the different topologies.

5.7 Outlook

In the very last section of this thesis, some ideas are presented how the proposed structural classification approach may be further developed in the future.

To further extend the usability of the database, one can expand the membrane protein dataset in two possible ways: by integrating recently sequenced genomes and by also including membrane proteins with less than three helices. The restriction to proteins with at least three transmembrane helices was previously justified by the need to minimize the risk to include non-membrane proteins. However, with the increasing accuracy of newly developed membrane protein prediction tools, it is possible to include smaller membrane proteins as well.

The fold determinants used in this thesis are: sequence similarity, number of transmembrane helices, loop lengths and helix-helix interactions. The membrane protein fold definition can be further advanced by also taking into account irregular structures such as reentrant regions, tilted helices and interface helices (see also section 1.1.2, page 3).

The proposed structural classification method is based on full-length sequences. Traditional approaches, such as SCOP and CATH, use protein domains as the classification unit. CAMPS could be turned into a domain-based classification approach, if methods to predict domain boundaries from sequence would exist. To the best of my knowledge, no such method is currently available for membrane proteins. One possible solution to this problem is to use helix-helix interaction graphs. Given that interactions within a domain are more comprehensive than between domains [331], it may be possible to delineate domain boundaries by searching for highly connected subgraphs in the helix-helix interaction graph.

The usability of CAMPS can also be extended by regular updates. However, the development of a new release is a very time consuming process (that can take several months depending on the computational resources). Therefore, it would be necessary to establish an alternative update procedure that, for example, tests whether a new membrane protein sequence can be assigned to one of the existing SC-clusters or not. In the latter case, a new SC-cluster could be formed.

Chapter 6

Bibliography

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

[2] D. Frishman and H. W. Mewes. Protein structural classes in five complete genomes. *Nat Struct Biol*, 4(8):626–628, Aug 1997.

[3] D. T. Jones. Do transmembrane protein superfolds exist? *FEBS Lett*, 423(3):281–285, Feb 1998.

[4] 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.

[5] 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.

[6] M. Ahram, Z. I. Litou, R. Fang, and G. Al-Tawallbeh. Estimation of membrane proteins in the human proteome. *In Silico Biol*, 6(5):379–386, 2006.

[7] 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.

[8] A. M. Spiegel. Defects in g protein-coupled signal transduction in human disease. *Annu Rev Physiol*, 58:143–170, 1996.

[9] T. Suzuki, Y. Araki, T. Yamamoto, and T. Nakaya. Trafficking of alzheimer's disease-related membrane proteins and its participation in disease pathogenesis. *J Biochem*, 139(6):949–955, Jun 2006.

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

[11] J. Davey. G-protein-coupled receptors: new approaches to maximise the impact of gpcrs in drug discovery. *Expert Opin Ther Targets*, 8(2):165–170, Apr 2004.

[12] N. Hurwitz, M. Pellegrini-Calace, and D. T. Jones. Towards genome-scale structure prediction for transmembrane proteins. *Philos Trans R Soc Lond B Biol Sci*, 361(1467):465–475, Mar 2006.

[13] G. von Heijne. Principles of membrane protein assembly and structure. *Prog Biophys Mol Biol*, 66(2):113–139, 1996.

[14] W. C. Wimley. The versatile beta-barrel membrane protein. *Curr Opin Struct Biol*, 13(4):404–411, Aug 2003.

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

[16] 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.

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

[18] 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.

[19] 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.

[20] J. Deisenhofer, O. Epp, K. Miki, R. Huber, and H. Michel. Structure of the protein subunits in the photosynthetic reaction centre of rhodopseudomonas viridis at 3 Å resolution. *Nature*, 318:618–624, 1985.

[21] G. von Heijne. A day in the life of dr k. or how i learned to stop worrying and love lysozyme: a tragedy in six acts. *J Mol Biol*, 293(2):367–379, Oct 1999.

[22] R. Henderson and P. N. Unwin. Three-dimensional model of purple membrane obtained by electron microscopy. *Nature*, 257(5521):28–32, Sep 1975.

[23] R. Henderson, J. M. Baldwin, T. A. Ceska, F. Zemlin, E. Beckmann, and K. H. Downing. Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *J Mol Biol*, 213(4):899–929, Jun 1990.

[24] J. U. Bowie. Helix packing in membrane proteins. *J Mol Biol*, 272(5):780–789, Oct 1997.

[25] 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.

[26] 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.

[27] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res*, 28(1):235–242, Jan 2000.

[28] K. Lundstrom. Structural genomics for membrane proteins. *Cell Mol Life Sci*, 63(22):2597–2607, Nov 2006.

[29] S. J. Opella and F. M. Marassi. Structure determination of membrane proteins by nmr spectroscopy. *Chem Rev*, 104(8):3587–3606, Aug 2004.

[30] M. Caffrey. Crystallizing membrane proteins for structure determination: use of lipidic mesophases. *Annu Rev Biophys*, 38:29–51, 2009.

[31] R. M. Bill, P. J. F. Henderson, S. Iwata, E. R. S. Kunji, H. Michel, R. Neutze, S. Newstead, B. Poolman, C. G. Tate, and H. Vogel. Overcoming barriers to membrane protein structure determination. *Nat Biotechnol*, 29(4):335–340, Apr 2011.

[32] R. Ujwal and J. U. Bowie. Crystallizing membrane proteins using lipidic bicelles. *Methods*, Sep 2011.

[33] J. Torres, T. J. Stevens, and M. Sams. Membrane proteins: the 'wild west' of structural biology. *Trends Biochem Sci*, 28(3):137–144, Mar 2003.

[34] S. H. White. The progress of membrane protein structure determination. *Protein Sci*, 13(7):1948–1949, Jul 2004.

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

[36] G. E. Tusnady and I. Simon. Topology prediction of helical transmembrane proteins: how far have we reached? *Curr Protein Pept Sci*, 11(7):550–561, Nov 2010.

[37] E. Wallin, T. Tsukihara, S. Yoshikawa, G. von Heijne, and A. Elofsson. Architecture of helix bundle membrane proteins: an analysis of cytochrome c oxidase from bovine mitochondria. *Protein Sci*, 6(4):808–815, Apr 1997.

[38] G. von Heijne. Recent advances in the understanding of membrane protein assembly and structure. *Q Rev Biophys*, 32(4):285–307, Nov 1999.

[39] G. von Heijne. Proline kinks in transmembrane alpha-helices. *J Mol Biol*, 218(3):499–503, Apr 1991.

[40] M. S. Sansom. Proline residues in transmembrane helices of channel and transport proteins: a molecular modelling study. *Protein Eng*, 5(1):53–60, Jan 1992.

[41] S. Hong, K. S. Ryu, M. S. Oh, I. Ji, and T. H. Ji. Roles of transmembrane prolines and proline-induced kinks of the lutropin/choriogonadotropin receptor. *J Biol Chem*, 272(7):4166–4171, Feb 1997.

[42] H. Lu, T. Marti, and P. J. Booth. Proline residues in transmembrane alpha helices affect the folding of bacteriorhodopsin. *J Mol Biol*, 308(2):437–446, Apr 2001.

[43] E. R. Slepkov, S. Chow, M. J. Lemieux, and L. Fliegel. Proline residues in transmembrane segment iv are critical for activity, expression and targeting of the na+/h+ exchanger isoform 1. *Biochem J*, 379(Pt 1):31–38, Apr 2004.

[44] A. C. Conner, D. L. Hay, J. Simms, S. G. Howitt, M. Schindler, D. M. Smith, M. Wheatley, and D. R. Poyner. A key role for transmembrane prolines in calcitonin receptor-like receptor agonist binding and signalling: implications for family b g-protein-coupled receptors. *Mol Pharmacol*, 67(1):20–31, Jan 2005.

[45] R. Dutzler, E. B. Campbell, M. Cadene, B. T. Chait, and R. MacKinnon. X-ray structure of a clc chloride channel at 3.0 Å reveals the molecular basis of anion selectivity. *Nature*, 415(6869):287–294, Jan 2002.

[46] 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.

[47] P. L. Yeagle, M. Bennett, V. Lematre, and A. Watts. Transmembrane helices of membrane proteins may flex to satisfy hydrophobic mismatch. *Biochim Biophys Acta*, 1768(3):530–537, Mar 2007.

[48] A. Holt and J. A. Killian. Orientation and dynamics of transmembrane peptides: the power of simple models. *Eur Biophys J*, 39(4):609–621, Mar 2010.

[49] J. Vonck. A three-dimensional difference map of the n intermediate in the bacteriorhodopsin photocycle: part of the f helix tilts in the m to n transition. *Biochemistry*, 35(18):5870–5878, May 1996.

[50] J. A. Killian. Hydrophobic mismatch between proteins and lipids in membranes. *Biochim Biophys Acta*, 1376(3):401–415, Nov 1998.

[51] E. Granseth, G. von Heijne, and A. Elofsson. A study of the membrane-water interface region of membrane proteins. *J Mol Biol*, 346(1):377–385, Feb 2005.

[52] J. A. Killian and G. von Heijne. How proteins adapt to a membrane-water interface. *Trends Biochem Sci*, 25(9):429–434, Sep 2000.

[53] A. K. Chamberlain, Y. Lee, S. Kim, and J. U. Bowie. Snorkeling preferences foster an amino acid composition bias in transmembrane helices. *J Mol Biol*, 339(2):471–479, May 2004.

[54] J. Liang, L. Adamian, and R. Jr. Jackups. The membrane-water interface region of membrane proteins: structural bias and the anti-snorkeling effect. *Trends Biochem Sci*, 30(7):355–357, Jul 2005.

[55] J. P. R. O. Orgel. Sequence context and modified hydrophobic moment plots help identify 'horizontal' surface helices in transmembrane protein structure prediction. *J Struct Biol*, 148(1):51–65, Oct 2004.

[56] A. Kauko, K. Illergard, and A. Elofsson. Coils in the membrane core are conserved and functionally important. *J Mol Biol*, 380(1):170–180, Jun 2008.

[57] 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.

[58] D. J. Slotboom, I. Sobczak, W. N. Konings, and J. S. Lolkema. A conserved serine-rich stretch in the glutamate transporter family forms a substrate-sensitive reentrant loop. *Proc Natl Acad Sci U S A*, 96(25):14282–14287, Dec 1999.

[59] T. Iwamoto, A. Uehara, I. Imanaga, and M. Shigekawa. The na+/ca2+ exchanger ncx1 has oppositely oriented reentrant loop domains that contain conserved aspartic acids whose mutation alters its apparent ca2+ affinity. *J Biol Chem*, 275(49):38571–38580, Dec 2000.

[60] K. Murata, K. Mitsuoka, T. Hirai, T. Walz, P. Agre, J. B. Heymann, A. Engel, and Y. Fujiyoshi. Structural determinants of water permeation through aquaporin-1. *Nature*, 407(6804):599–605, Oct 2000.

[61] B. I. Kanner and L. Borre. The dual-function glutamate transporters: structure and molecular characterisation of the substrate-binding sites. *Biochim Biophys Acta*, 1555(1-3):92–95, Sep 2002.

[62] H. Korres and N. K. Verma. Topological analysis of glucosyltransferase gtrv of shigella flexneri by a dual reporter system and identification of a unique reentrant loop. *J Biol Chem*, 279(21):22469–22476, May 2004.

[63] J. S. Lolkema, I. Sobczak, and D.-J. Slotboom. Secondary transporters of the 2hct family contain two homologous domains with inverted membrane topology and trans re-entrant loops. *FEBS J*, 272(9):2334–2344, May 2005.

[64] T. Krupnik, I. Sobczak-Elbourne, and J. S. Lolkema. Turnover and accessibility of a reentrant loop of the na+-glutamate transporter glts are modulated by the central cytoplasmic loop. *Mol Membr Biol*, 28(7-8):462–472, 2011.

[65] C. Yan and J. Luo. An analysis of reentrant loops. *Protein J*, 29(5):350–354, Jul 2010.

[66] *The PyMOL Molecular Graphics System, Version 1.4, Schrödinger, LLC.*

[67] D. Yernool, O. Boudker, Y. Jin, and E. Gouaux. Structure of a glutamate transporter homologue from pyrococcus horikoshii. *Nature*, 431(7010):811–818, Oct 2004.

[68] M. A. Martí-Renom, A. C. Stuart, A. Fiser, R. Sánchez, F. Melo, and A. Sali. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct*, 29:291–325, 2000.

[69] 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.

[70] C. S. Reddy, K. Vijayasarathy, E. Srinivas, G. M. Sastry, and G. N. Sastry. Homology modeling of membrane proteins: a critical assessment. *Comput Biol Chem*, 30(2):120–126, Apr 2006.

[71] 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.

[72] P. C. Ng, J. G. Henikoff, and S. Henikoff. Phat: a transmembrane-specific substitution matrix. predicted hydrophobic and transmembrane. *Bioinformatics*, 16(9):760–766, Sep 2000.

[73] T. Müller, S. Rahmann, and M. Rehmsmeier. Non-symmetric score matrices and the detection of homologous transmembrane proteins. *Bioinformatics*, 17 Suppl 1:S182–S189, 2001.

[74] R. A. Sutormin, A. B. Rakhmaninova, and M. S. Gelfand. Batmas30: amino acid substitution matrix for alignment of bacterial transporters. *Proteins*, 51(1):85–95, Apr 2003.

[75] J. R. Hill, S. Kelm, J. Shi, and C. M. Deane. Environment specific substitution tables improve membrane protein alignment. *Bioinformatics*, 27(13):i15–i23, Jul 2011.

[76] M. Cserzö, J. M. Bernassau, I. Simon, and B. Maigret. New alignment strategy for transmembrane proteins. *J Mol Biol*, 243(3):388–396, Oct 1994.

[77] Y. Shafrir and H. R. Guy. Stam: simple transmembrane alignment method. *Bioinformatics*, 20(5):758–769, Mar 2004.

[78] W. Pirovano, K. A. Feenstra, and J. Heringa. Pralinetm: a strategy for improved multiple alignment of transmembrane proteins. *Bioinformatics*, 24(4):492–497, Feb 2008.

[79] S. Kelm, J. Shi, and C. M. Deane. Medeller: homology-based coordinate generation for membrane proteins. *Bioinformatics*, 26(22):2833–2840, Nov 2010.

[80] R. Sánchez, U. Pieper, F. Melo, N. Eswar, M. A. Martí-Renom, M. S. Madhusudhan, N. Mirković, and A. Sali. Protein structure modeling for structural genomics. *Nat Struct Biol*, 7 Suppl:986–990, Nov 2000.

[81] W. R. Taylor, D. T. Jones, and N. M. Green. A method for alpha-helical integral membrane protein fold prediction. *Proteins*, 18(3):281–294, Mar 1994.

[82] M. Pellegrini-Calace, A. Carotti, and D. T. Jones. Folding in lipid membranes (film): a novel method for the prediction of small membrane protein 3d structures. *Proteins*, 50(4):537–545, Mar 2003.

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

[84] 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.

[85] D. N. Langelaan, M. Wieczorek, C. Blouin, and J. K. Rainey. Improved helix and kink characterization in membrane proteins allows evaluation of kink sequence predictors. *J Chem Inf Model*, 50(12):2213–2220, Dec 2010.

[86] A. D. Meruelo, I. Samish, and J. U. Bowie. Tmkink: a method to predict transmembrane helix kinks. *Protein Sci*, 20(7):1256–1264, Jul 2011.

[87] 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.

[88] C. Papaloukas, E. Granseth, H. Viklund, and A. Elofsson. Estimating the length of transmembrane helices using z-coordinate predictions. *Protein Sci*, 17(2):271–278, Feb 2008.

[89] 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.

[90] H. Viklund and A. Elofsson. Octopus: improving topology prediction by two-track ann-based preference scores and an extended topological grammar. *Bioinformatics*, 24(15):1662–1668, Aug 2008.

[91] H. Viklund, A. Bernsel, M. Skwark, and A. Elofsson. Spoctopus: a combined predictor of signal peptides and membrane protein topology. *Bioinformatics*, 24(24):2928–2929, Dec 2008.

[92] T. Nugent and D. T. Jones. Transmembrane protein topology prediction using support vector machines. *BMC Bioinformatics*, 10:159, 2009.

[93] A. Kauko, L. E. Hedin, E. Thebaud, S. Cristobal, A. Elofsson, and G. von Heijne. Repositioning of transmembrane alpha-helices during membrane protein folding. *J Mol Biol*, 397(1):190–201, Mar 2010.

[94] L. J. McGuffin, K. Bryson, and D. T. Jones. The psipred protein structure prediction server. *Bioinformatics*, 16(4):404–405, Apr 2000.

[95] 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.

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

[97] D. T. Jones. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics*, 23(5):538–544, Mar 2007.

[98] E. L. Sonnhammer, G. von Heijne, and A. Krogh. A hidden markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol*, 6:175–182, 1998.

[99] G. E. Tusnady and I. Simon. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J Mol Biol*, 283(2):489–506, Oct 1998.

[100] L. Käll, 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.

[101] 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.

[102] L. Käll, 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.

[103] Z. Yuan, J. S. Mattick, and R. D. Teasdale. Svmtm: support vector machines to predict transmembrane segments. *J Comput Chem*, 25(5):632–636, Apr 2004.

[104] A. Lo, H.-S. Chiu, T.-Y. Sung, P.-C. Lyu, and W.-L. Hsu. Enhanced membrane protein topology prediction using a hierarchical classification method and a new scoring function. *J Proteome Res*, 7(2):487–496, Feb 2008.

[105] S. M. Reynolds, L. Käll, 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.

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

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

[108] L. A. Parodi, C. A. Granatir, and G. M. Maggiora. A consensus procedure for predicting the location of alpha-helical transmembrane segments in proteins. *Comput Appl Biosci*, 10(5):527–535, Sep 1994.

[109] J. Nilsson, B. Persson, and G. Von Heijne. Prediction of partial membrane protein topologies using a consensus approach. *Protein Sci*, 11(12):2974–2980, Dec 2002.

[110] P. D. Taylor, T. K. Attwood, and D. R. Flower. Bprompt: A consensus server for membrane protein prediction. *Nucleic Acids Res*, 31(13):3698–3700, Jul 2003.

[111] M. Arai, H. Mitsuke, M. Ikeda, J.-X. Xia, T. Kikuchi, M. Satake, and T. Shimizu. Conpred ii: a consensus prediction method for obtaining transmembrane topology models with high reliability. *Nucleic Acids Res*, 32(Web Server issue):W390–W393, Jul 2004.

[112] 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.

[113] M. Klammer, D. N. Messina, T. Schmitt, and E. L. L. Sonnhammer. Metatm - a consensus method for transmembrane protein topology prediction. *BMC Bioinformatics*, 10:314, 2009.

[114] P. L. Martelli, P. Fariselli, and R. Casadio. An ensemble machine learning approach for the prediction of all-alpha membrane proteins. *Bioinformatics*, 19 Suppl 1:i205–i211, 2003.

[115] R. Ahmed, H. Rangwala, and G. Karypis. Toptmh: topology predictor for transmembrane alpha-helices. *J Bioinform Comput Biol*, 8(1):39–57, Feb 2010.

[116] A. Sääf, M. Johansson, E. Wallin, and G. von Heijne. Divergent evolution of membrane protein topology: the escherichia coli rnfa and rnfe homologues. *Proc Natl Acad Sci U S A*, 96(15):8540–8544, Jul 1999.

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

[118] 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.

[119] 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.

[120] M. H. Jr. Saier. Tracing pathways of transport protein evolution. *Mol Microbiol*, 48(5):1145–1156, Jun 2003.

[121] T. Shimizu, H. Mitsuke, K. Noto, and M. Arai. Internal gene duplication in the evolution of prokaryotic transmembrane proteins. *J Mol Biol*, 339(1):1–15, May 2004.

[122] A. Hennerdal, J. Falk, E. Lindahl, and A. Elofsson. Internal duplications in alpha-helical membrane protein topologies are common but the nonduplicated forms are rare. *Protein Sci*, 19(12):2305–2318, Dec 2010.

[123] 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.

[124] A. Sääf, L. Baars, and G. von Heijne. The internal repeats in the na+/ca2+ exchanger-related escherichia coli protein yrbg have opposite membrane topologies. *J Biol Chem*, 276(22):18905–18907, Jun 2001.

[125] S. Choi, J. Jeon, J.-S. Yang, and S. Kim. Common occurrence of internal repeat symmetry in membrane proteins. *Proteins*, 71(1):68–80, Apr 2008.

[126] M. Rapp, S. Seppälä, E. Granseth, and G. von Heijne. Emulating membrane protein evolution by rational design. *Science*, 315(5816):1282–1284, Mar 2007.

[127] 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.

[128] J. Liu and B. Rost. Comparing function and structure between entire proteomes. *Protein Sci*, 10(10):1970–1979, Oct 2001.

[129] C. G. Knight, R. Kassen, H. Hebestreit, and P. B. Rainey. Global analysis of predicted proteomes: functional adaptation of physical properties. *Proc Natl Acad Sci U S A*, 101(22):8390–8395, Jun 2004.

[130] S. Mitaku, M. Ono, T. Hirokawa, S. Boon-Chieng, and M. Sonoyama. Proportion of membrane proteins in proteomes of 15 single-cell organisms analyzed by the sosui prediction system. *Biophys Chem*, 82(2-3):165–171, Dec 1999.

[131] M. Remm and E. Sonnhammer. Classification of transmembrane protein families in the caenorhabditis elegans genome and identification of human orthologs. *Genome Res*, 10(11):1679–1689, Nov 2000.

[132] M. Arai, M. Ikeda, and T. Shimizu. Comprehensive analysis of transmembrane topologies in prokaryotic genomes. *Gene*, 304:77–86, Jan 2003.

[133] H. Kim, K. Melén, 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.

[134] R. Fredriksson, M. C. Lagerström, L.-G. Lundin, and H. B. Schiöth. The g-protein-coupled receptors in the human genome form five main families. phylogenetic analysis, paralogon groups, and fingerprints. *Mol Pharmacol*, 63(6):1256–1272, Jun 2003.

[135] C. A. Orengo, D. T. Jones, and J. M. Thornton. Protein superfamilies and domain superfolds. *Nature*, 372(6507):631–634, Dec 1994.

[136] C. A. Ouzounis, R. M. R. Coulson, A. J. Enright, V. Kunin, and J. B. Pereira-Leal. Classification schemes for protein structure and function. *Nat Rev Genet*, 4(7):508–519, Jul 2003.

[137] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. Blast+: architecture and applications. *BMC Bioinformatics*, 10:421, 2009.

[138] W. Pearson. Finding protein and nucleotide similarities with fasta. *Curr Protoc Bioinformatics*, Chapter 3:Unit3.9, Feb 2004.

[139] The UniProt Consortium. Reorganizing the protein space at the universal protein resource (uniprot). *Nucleic Acids Res*, 40(D1):D71–D75, Jan 2012.

[140] E. V. Kriventseva, W. Fleischmann, E. M. Zdobnov, and R. Apweiler. Clustr: a database of clusters of swiss-prot+trembl proteins. *Nucleic Acids Res*, 29(1):33–36, Jan 2001.

[141] T. Meinel, A. Krause, H. Luz, M. Vingron, and E. Staub. The systers protein family database in 2005. *Nucleic Acids Res*, 33(Database issue):D226–D229, Jan 2005.

[142] C. J. A. Sigrist, L. Cerutti, E. de Castro, P. S. Langendijk-Genevaux, V. Bulliard, A. Bairoch, and N. Hulo. Prosite, a protein domain database for functional characterization and annotation. *Nucleic Acids Res*, 38(Database issue):D161–D166, Jan 2010.

[143] S. Hunter, P. Jones, A. Mitchell, and *et al.* Interpro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res*, 40(D1):D306–D312, Jan 2012.

[144] N. Rappoport, S. Karsenty, A. Stern, N. Linial, and M. Linial. Protonet 6.0: organizing 10 million protein sequences in a compact hierarchical family tree. *Nucleic Acids Res*, 40(1):D313–D320, Jan 2012.

[145] E. L. Sonnhammer, S. R. Eddy, and R. Durbin. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, 28(3):405–420, Jul 1997.

[146] M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, A. Bateman, and R. D. Finn. The pfam protein families database. *Nucleic Acids Res*, 40(D1):D290–D301, Jan 2012.

[147] R. L. Tatusov, E. V. Koonin, and D. J. Lipman. A genomic perspective on protein families. *Science*, 278(5338):631–637, Oct 1997.

[148] R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale. The cog database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41, Sep 2003.

[149] R. D. Finn, J. Mistry, B. Schuster-Böckler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S. R. Eddy, E. L. L. Sonnhammer, and A. Bateman. Pfam: clans, web tools and services. *Nucleic Acids Res*, 34(Database issue):D247–D251, Jan 2006.

[150] W. M. Fitch. Distinguishing homologous from analogous proteins. *Syst Zool*, 19(2):99–113, Jun 1970.

[151] T. Gabaldón and M. A. Huynen. Prediction of protein function and pathways in the genome era. *Cell Mol Life Sci*, 61(7-8):930–944, Apr 2004.

[152] D. A. Natale, U. T. Shankavaram, M. Y. Galperin, Y. I. Wolf, L. Aravind, and E. V. Koonin. Towards understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (cogs). *Genome Biol*, 1(5):RESEARCH0009, 2000.

[153] K. S. Makarova, L. Aravind, Y. I. Wolf, R. L. Tatusov, K. W. Minton, E. V. Koonin, and M. J. Daly. Genome of the extremely radiation-resistant bacterium deinococcus radiodurans viewed from the perspective of comparative genomics. *Microbiol Mol Biol Rev*, 65(1):44–79, Mar 2001.

[154] T. Håfström, D. S. Jansson, and B. Segerman. Complete genome sequence of brachyspira intermedia reveals unique genomic features in brachyspira species and phage-mediated horizontal gene transfer. *BMC Genomics*, 12:395, 2011.

[155] E. V. Koonin, K. S. Makarova, and L. Aravind. Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol*, 55:709–742, 2001.

[156] I. K. Jordan, I. B. Rogozin, Y. I. Wolf, and E. V. Koonin. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res*, 12(6):962–968, Jun 2002.

[157] A. G. Murzin and A. Bateman. Distant homology recognition using structural classification of proteins. *Proteins*, Suppl 1:105–112, 1997.

[158] C. Chothia and A. M. Lesk. The relation between the divergence of sequence and structure in proteins. *EMBO J*, 5(4):823–826, Apr 1986.

[159] M. Levitt and M. Gerstein. A unified statistical framework for sequence comparison and structure comparison. *Proc Natl Acad Sci U S A*, 95(11):5913–5920, May 1998.

[160] C. T. Zhang. Relations of the numbers of protein sequences, families and folds. *Protein Eng*, 10(7):757–761, Jul 1997.

[161] Z. X. Wang. A re-estimation for the total numbers of protein folds and superfamilies. *Protein Eng*, 11(8):621–626, Aug 1998.

[162] 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.

[163] 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.

[164] 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.

[165] A. L. Cuff, I. Sillitoe, T. Lewis, A. B. Clegg, R. Rentzsch, N. Furnham, M. Pellegrini-Calace, D. Jones, J. Thornton, and C. A. Orengo. Extending cath: increasing coverage of the protein structure universe and linking structure with function. *Nucleic Acids Res*, 39(Database issue):D420–D426, Jan 2011.

[166] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–29, May 2000.

[167] J. S. Richardson. The anatomy and taxonomy of protein structure. *Adv Protein Chem*, 34:167–339, 1981.

[168] J. Janin and S. J. Wodak. Structural domains in proteins and their role in the dynamics of protein function. *Prog Biophys Mol Biol*, 42(1):21–78, 1983.

[169] A. Andreeva, D. Howorth, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin. Scop database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res*, 32(Database issue):D226–D229, Jan 2004.

[170] L. Holm and C. Sander. The fssp database of structurally aligned protein fold families. *Nucleic Acids Res*, 22(17):3600–3609, Sep 1994.

[171] 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.

[172] 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.

[173] Z. Aung and K.-L. Tan. Automatic 3d protein structure classification without structural alignment. *J Comput Biol*, 12(9):1221–1241, Nov 2005.

[174] K. Chen and L. Kurgan. Pfres: protein fold classification by using evolutionary information and predicted secondary structure. *Bioinformatics*, 23(21):2843–2850, Nov 2007.

[175] 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.

[176] S. Dua and P. C. Kidambi. Protein structural classification using orthogonal transformation and class-association rules. *Int J Data Min Bioinform*, 4(2):175–190, 2010.

[177] P. Jain and J. D. Hirst. Automatic structure classification of small proteins using random forest. *BMC Bioinformatics*, 11:364, 2010.

[178] N. Daniels, A. Kumar, L. Cowen, and Matt Menke. Touring protein space with matt. *IEEE/ACM Trans Comput Biol Bioinform*, Apr 2011.

[179] Webb. *Enzyme Nomenclature 1992: Recommendations of the nomenclature committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes.* Academic Press, 1992.

[180] The Gene Ontology Consortium. The gene ontology: enhancements for 2011. *Nucleic Acids Res*, 40(D1):D559–D564, Jan 2012.

[181] A. Bairoch. The enzyme data bank. *Nucleic Acids Res*, 21(13):3155–3156, Jul 1993.

[182] A. Bairoch. The enzyme database in 2000. *Nucleic Acids Res*, 28(1):304–305, Jan 2000.

[183] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, Jul 1973.

[184] B. Rost. Twilight zone of protein sequence alignments. *Protein Eng*, 12(2):85–94, Feb 1999.

[185] A. R. Kinjo and K. Nishikawa. Eigenvalue analysis of amino acid substitution matrices reveals a sharp transition of the mode of sequence conservation in proteins. *Bioinformatics*, 20(16):2504–2508, Nov 2004.

[186] M. J. Sippl. Fold space unlimited. *Curr Opin Struct Biol*, 19(3):312–320, Jun 2009.

[187] M. Kosloff and R. Kolodny. Sequence-similar, structure-dissimilar protein pairs in the pdb. *Proteins*, 71(2):891–902, May 2008.

[188] H. H. Gan, R. A. Perlow, S. Roy, J. Ko, M. Wu, J. Huang, S. Yan, A. Nicoletta, J. Vafai, D. Sun, L. Wang, J. E. Noah, S. Pasquali, and T. Schlick. Analysis of protein sequence/structure similarity relationships. *Biophys J*, 83(5):2781–2791, Nov 2002.

[189] J. D. Watson, S. Sanderson, A. Ezersky, A. Savchenko, A. Edwards, C. Orengo, A. Joachimiak, R. A. Laskowski, and J. M. Thornton. Towards fully automated structure-based function prediction in structural genomics: a case study. *J Mol Biol*, 367(5):1511–1522, Apr 2007.

[190] Y. Loewenstein, D. Raimondo, O. C. Redfern, J. Watson, D. Frishman, M. Linial, C. Orengo, J. Thornton, and A. Tramontano. Protein function annotation by homology-based inference. *Genome Biol*, 10(2):207, 2009.

[191] M. Osadchy and R. Kolodny. Maps of protein structure space reveal a fundamental relationship between protein structure and function. *Proc Natl Acad Sci U S A*, 108(30):12301–12306, Jul 2011.

[192] E. Krissinel. On the relationship between sequence and structure similarities in proteomics. *Bioinformatics*, 23(6):717–723, Mar 2007.

[193] A. C. Martin, C. A. Orengo, E. G. Hutchinson, S. Jones, M. Karmirantzou, R. A. Laskowski, J. B. Mitchell, C. Taroni, and J. M. Thornton. Protein folds and functions. *Structure*, 6(7):875–884, Jul 1998.

[194] H. Hegyi and M. Gerstein. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J Mol Biol*, 288(1):147–164, Apr 1999.

[195] N. Nagano, C. A. Orengo, and J. M. Thornton. One fold with many functions: the evolutionary relationships between tim barrel families based on their sequences, structures and functions. *J Mol Biol*, 321(5):741–765, Aug 2002.

[196] O. C. Redfern, B. Dessailly, and C. A. Orengo. Exploring the structure and function paradigm. *Curr Opin Struct Biol*, 18(3):394–402, Jun 2008.

[197] C. Chothia. Proteins. one thousand families for the molecular biologist. *Nature*, 357(6379):543–544, Jun 1992.

[198] C. Zhang and C. DeLisi. Estimating the number of protein folds. *J Mol Biol*, 284(5):1301–1305, Dec 1998.

[199] S. Govindarajan, R. Recabarren, and R. A. Goldstein. Estimating the total number of protein folds. *Proteins*, 35(4):408–414, Jun 1999.

[200] 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.

[201] A. F. W. Coulson and J. Moult. A unifold, mesofold, and superfold model of protein fold use. *Proteins*, 46(1):61–71, Jan 2002.

[202] A. Grant, D. Lee, and C. Orengo. Progress towards mapping the universe of protein folds. *Genome Biol*, 5(5):107, 2004.

[203] X. Liu, K. Fan, and W. Wang. The number of protein folds and their distribution over families in nature. *Proteins*, 54(3):491–499, Feb 2004.

[204] L. H. Greene, T. E. Lewis, S. Addou, A. Cuff, T. Dallman, M. Dibley, O. Redfern, F. Pearl, R. Nambudiry, A. Reid, I. Sillitoe, C. Yeats, J. M. Thornton, and C. A. Orengo. The cath domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res*, 35(Database issue):D291–D297, Jan 2007.

[205] C. A. Orengo, I. Sillitoe, G. Reeves, and F. M. Pearl. Review: what can structural classifications reveal about protein evolution? *J Struct Biol*, 134(2-3):145–165, 2001.

[206] 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.

[207] C. Hadley and D. T. Jones. A systematic comparison of protein structure classifications: Scop, cath and fssp. *Structure*, 7(9):1099–1112, Sep 1999.

[208] I. N. Shindyalov and P. E. Bourne. An alternative view of protein fold space. *Proteins*, 38(3):247–260, Feb 2000.

[209] 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.

[210] G. Csaba, F. Birzele, and R. Zimmer. Systematic comparison of scop and cath: a new gold standard for protein structure analysis. *BMC Struct Biol*, 9:23, 2009.

[211] T. A. Holland, S. Veretnik, I. N. Shindyalov, and P. E. Bourne. Partitioning protein structures into domains: why is it so difficult? *J Mol Biol*, 361(3):562–590, Aug 2006.

[212] 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.

[213] J. Söding. Protein homology detection by hmm-hmm comparison. *Bioinformatics*, 21(7):951–960, Apr 2005.

[214] I. Friedberg and A. Godzik. Connecting the protein structure universe by using sparse recurring fragments. *Structure*, 13(8):1213–1224, Aug 2005.

[215] 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.

[216] 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.

[217] 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.

[218] 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.

[219] 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.

[220] R. I. Sadreyev, B.-H. Kim, and N. V. Grishin. Discrete-continuous duality of protein structure space. *Curr Opin Struct Biol*, 19(3):321–328, Jun 2009.

[221] J. Skolnick, A. K. Arakaki, S. Y. Lee, and M. Brylinski. The continuity of protein structure space is an intrinsic property of proteins. *Proc Natl Acad Sci U S A*, 106(37):15690–15695, Sep 2009.

[222] M. I. Sadowski and W. R. Taylor. On the evolutionary origins of 'fold space continuity': a study of topological convergence and divergence in mixed alpha-beta domains. *J Struct Biol*, 172(3):244–252, Dec 2010.

[223] W. Li and A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, Jul 2006.

[224] A. D. Michie, C. A. Orengo, and J. M. Thornton. Analysis of domain structural class using an automated class assignment protocol. *J Mol Biol*, 262(2):168–185, Sep 1996.

[225] L. Holm, S. Kääriäinen, P. Rosenström, and A. Schenkel. Searching protein structure databases with dalilite v.3. *Bioinformatics*, 24(23):2780–2781, Dec 2008.

[226] C. A. Orengo and W. R. Taylor. Ssap: sequential structure alignment program for protein structure comparison. *Methods Enzymol*, 266:617–635, 1996.

[227] G. Caetano-Anolles, M. Wang, D. Caetano-Anolles, and J. E. Mittenthal. The origin, evolution and structure of the protein world. *Biochem J*, 417(3):621–637, Feb 2009.

[228] S. E. Brenner, C. Chothia, and T. J. Hubbard. Population statistics of protein structures: lessons from structural classifications. *Curr Opin Struct Biol*, 7(3):369–376, Jun 1997.

[229] M. Gerstein. How representative are the known structures of the proteins in a complete genome? a comprehensive structural census. *Fold Des*, 3(6):497–512, 1998.

[230] I. T. Arkin, A. T. Brünger, and D. M. Engelman. Are there dominant membrane protein families with a given number of helices? *Proteins*, 28(4):465–466, Aug 1997.

[231] M. Gerstein. A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J Mol Biol*, 274(4):562–576, Dec 1997.

[232] 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 ca change le plus c'est la meme chose. *J Bioenerg Biomembr*, 31(3):201–213, Jun 1999.

[233] S. Jones, M. Stewart, A. Michie, M. B. Swindells, C. Orengo, and J. M. Thornton. Domain assignment for protein structures using a consensus approach: characterization and analysis. *Protein Sci*, 7(2):233–242, Feb 1998.

[234] S. Iwata, C. Ostermeier, B. Ludwig, and H. Michel. Structure at 2.8 Å resolution of cytochrome c oxidase from paracoccus denitrificans. *Nature*, 376(6542):660–669, Aug 1995.

[235] J. Deisenhofer and H. Michel. Nobel lecture. the photosynthetic reaction centre from the purple bacterium rhodopseudomonas viridis. *EMBO J*, 8(8):2149–2170, Aug 1989.

[236] W. R. Taylor and C. A. Orengo. Protein structure alignment. *J Mol Biol*, 208(1):1–22, Jul 1989.

[237] C. A. Orengo, A. M. Martin, G. Hutchinson, S. Jones, D. T. Jones, A. D. Michie, M. B. Swindells, and J. M. Thornton. Classifying a protein in the cath database of domain structures. *Acta Crystallogr D Biol Crystallogr*, 54(Pt 6 Pt 1):1155–1167, Nov 1998.

[238] Y. Jiang, A. Lee, J. Chen, V. Ruta, M. Cadene, B. T. Chait, and R. MacKinnon. X-ray structure of a voltage-dependent k+ channel. *Nature*, 423(6935):33–41, May 2003.

[239] S. B. Long, E. B. Campbell, and R. Mackinnon. Voltage sensor of kv1.2: structural basis of electromechanical coupling. *Science*, 309(5736):903–908, Aug 2005.

[240] Y. Jiang, A. Lee, J. Chen, M. Cadene, B. T. Chait, and R. MacKinnon. The open pore conformation of potassium channels. *Nature*, 417(6888):523–526, May 2002.

[241] 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.

[242] G. Jekely. Did the last common ancestor have a biological membrane? *Biol Direct*, 1:35, 2006.

[243] 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.

[244] T. Meier, P. Polzer, K. Diederichs, W. Welte, and P. Dimroth. Structure of the rotor ring of f-type na+-atpase from ilyobacter tartaricus. *Science*, 308(5722):659–662, Apr 2005.

[245] S. Neumann, A. Fuchs, A. Mulkidjanian, and D. Frishman. Current status of membrane protein structure classification. *Proteins*, 78(7):1760–1773, May 2010.

[246] B. Vroling, M. Sanders, C. Baakman, A. Borrmann, S. Verhoeven, J. Klomp, L. Oliveira, J. de Vlieg, and G. Vriend. Gpcrdb: information system for g protein-coupled receptors. *Nucleic Acids Res*, 39(Database issue):D309–D319, Jan 2011.

[247] M. H. Jr. Saier, M. R. Yen, K. Noto, D. G. Tamang, and C. Elkan. The transporter classification database: recent advances. *Nucleic Acids Res*, 37(Database issue):D274–D278, Jan 2009.

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

[249] T. Sadka and M. Linial. Families of membranous proteins can be characterized by the amino acid composition of their transmembrane domains. *Bioinformatics*, 21 Suppl 1:i378–i386, Jun 2005.

[250] 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.

[251] L. Kelly, U. Pieper, N. Eswar, F. A. Hays, M. Li, Z. Roe-Zurz, D. L. Kroetz, K. M. Giacomini, R. M. Stroud, and A. Sali. A survey of integral alpha-helical membrane proteins. *J Struct Funct Genomics*, 10(4):269–280, Dec 2009.

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

[253] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic dna. *J Mol Biol*, 268(1):78–94, Apr 1997.

[254] M. Delorenzi and T. Speed. An hmm model for coiled-coil domains and a comparison with pssm-based predictions. *Bioinformatics*, 18(4):617–625, Apr 2002.

[255] H. Nielsen and A. Krogh. Prediction of signal peptides and signal anchors by a hidden markov model. *Proc Int Conf Intell Syst Mol Biol*, 6:122–130, 1998.

[256] R. D. Finn, J. Clements, and S. R. Eddy. Hmmer web server: interactive sequence similarity searching. *Nucleic Acids Res*, 39(Web Server issue):W29–W37, Jul 2011.

[257] S. R. Eddy. Hidden markov models. *Curr Opin Struct Biol*, 6(3):361–365, Jun 1996.

[258] S. R. Eddy. What is a hidden markov model? *Nat Biotechnol*, 22(10):1315–1316, Oct 2004.

[259] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286, 1989.

[260] 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.

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

[262] D. M. Engelman, Y. Chen, C.-N. Chin, A. R. Curran, A. M. Dixon, A. D. Dupuy, A. S. Lee, U. Lehnert, E. E. Matthews, Y. K. Reshetnyak, A. Senes, and J.-L. Popot. Membrane protein folding: beyond the two stage model. *FEBS Lett*, 555(1):122–125, Nov 2003.

[263] R. Maggio, P. Barbier, F. Fornai, and G. U. Corsini. Functional role of the third cytoplasmic loop in muscarinic receptor dimerization. *J Biol Chem*, 271(49):31055–31060, Dec 1996.

[264] X.-Q. Ren, T. Furukawa, M. Yamamoto, S. Aoki, M. Kobayashi, M. Nakagawa, and S.-I. Akiyama. A functional role of intracellular loops of human multidrug resistance protein 1. *J Biochem*, 140(3):313–318, Sep 2006.

[265] L. Xu, Y. Li, I. S. Haworth, and D. L. Davies. Functional role of the intracellular loop linking transmembrane domains 6 and 7 of the human dipeptide transporter hpept1. *J Membr Biol*, 238(1-3):43–49, Dec 2010.

[266] Y. Sugiyama, N. Polulyakh, and T. Shimizu. Identification of transmembrane protein functions by binary topology patterns. *Protein Eng*, 16(7):479–488, Jul 2003.

[267] J. M. Otaki and S. Firestein. Length analyses of mammalian g-protein-coupled receptors. *J Theor Biol*, 211(2):77–100, Jul 2001.

[268] M. Wistrand, L. Käll, and E. L. L. Sonnhammer. A general model of g protein-coupled receptor sequences and its application to detect remote homologs. *Protein Sci*, 15(3):509–521, Mar 2006.

[269] J. M. Kim, P. J. Booth, S. J. Allen, and H. G. Khorana. Structure and function in bacteriorhodopsin: the role of the interhelical loops in the folding and stability of bacteriorhodopsin. *J Mol Biol*, 308(2):409–422, Apr 2001.

[270] J. S. Landin, M. Katragadda, and A. D. Albert. Thermal destabilization of rhodopsin and opsin by proteolytic cleavage in bovine rod outer segment disk membranes. *Biochemistry*, 40(37):11176–11183, Sep 2001.

[271] 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.

[272] J. Jeong, P. Berman, and T. Przytycka. Fold classification based on secondary structure–how much is gained by including loop topology? *BMC Struct Biol*, 6:3, 2006.

[273] T. Rattei, P. Tischler, S. Götz, M.-A. Jehl, J. Hoser, R. Arnold, A. Conesa, and H.-W. Mewes. Simap–a comprehensive database of pre-calculated protein sequence similarities, domains, annotations and clusters. *Nucleic Acids Res*, 38(Database issue):D223–D226, Jan 2010.

[274] K. D. Pruitt, T. Tatusova, W. Klimke, and D. R. Maglott. Ncbi reference sequences: current status, policy and new initiatives. *Nucleic Acids Res*, 37(Database issue):D32–D36, Jan 2009.

[275] A. J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*, 30(7):1575–1584, Apr 2002.

[276] R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. Thompson. Multiple sequence alignment with the clustal series of programs. *Nucleic Acids Res*, 31(13):3497–3500, Jul 2003.

[277] R. C. Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5):1792–1797, 2004.

[278] H. Hartmann. Design and analysis of meta-models for the classification of membrane proteins. Master's thesis, Technical University Munich, 2008.

[279] J. G. Henikoff, E. A. Greene, S. Pietrokovski, and S. Henikoff. Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res*, 28(1):228–230, Jan 2000.

[280] K. Sjölander, K. Karplus, M. Brown, R. Hughey, A. Krogh, I. S. Mian, and D. Haussler. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput Appl Biosci*, 12(4):327–345, Aug 1996.

[281] D. Wilson, R. Pethica, Y. Zhou, C. Talbot, C. Vogel, M. Madera, C. Chothia, and J. Gough. Superfamily–sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res*, 37(Database issue):D380–D386, Jan 2009.

[282] 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.

[283] K. Lin, L. Zhu, and D.-Y. Zhang. An initial strategy for comparing proteins at the domain architecture level. *Bioinformatics*, 22(17):2081–2086, Sep 2006.

[284] H. Hegyi and M. Gerstein. Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res*, 11(10):1632–1640, Oct 2001.

[285] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo, and D. S. Wishart. Drugbank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res*, 39(Database issue):D1035–D1041, Jan 2011.

[286] S. Jayasinghe, K. Hristova, and S. H. White. Mptopo: A database of membrane protein topology. *Protein Sci*, 10(2):455–458, Feb 2001.

[287] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*, 33(Database issue):D514–D517, Jan 2005.

[288] 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.

[289] L. Chen, R. Oughtred, H. M. Berman, and J. Westbrook. Targetdb: a target registration database for structural genomics projects. *Bioinformatics*, 20(16):2860–2862, Nov 2004.

[290] 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.

[291] J. Yang, L. Chen, L. Sun, J. Yu, and Q. Jin. Vfdb 2008 release: an enhanced web-based resource for comparative pathogenomics. *Nucleic Acids Res*, 36(Database issue):D539–D542, Jan 2008.

[292] Y. Inoue, M. Ikeda, and T. Shimizu. Proteome-wide classification and identification of mammalian-type gpcrs by binary topology pattern. *Comput Biol Chem*, 28(1):39–49, Feb 2004.

[293] S. Möller, J. Vilo, and M. D. Croning. Prediction of the coupling specificity of g protein coupled receptors to their g proteins. *Bioinformatics*, 17 Suppl 1:S174–S181, 2001.

[294] J. Muller, D. Szklarczyk, P. Julien, I. Letunic, A. Roth, M. Kuhn, S. Powell, C. von Mering, T. Doerks, L. J. Jensen, and P. Bork. eggnog v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res*, 38(Database issue):D190–D195, Jan 2010.

[295] T.-W. Chen, T. H. Wu, W. V. Ng, and W.-C. Lin. Dodo: an efficient orthologous genes assignment tool based on domain architectures. domain based ortholog detection. *BMC Bioinformatics*, 11 Suppl 7:S6, 2010.

[296] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. Genbank. *Nucleic Acids Res*, 36(Database issue):D25–D30, Jan 2008.

[297] D. Ekman, A. K. Björklund, J. Frey-Skött, and A. Elofsson. Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *J Mol Biol*, 348(1):231–243, Apr 2005.

[298] A. Elofsson and E. L. Sonnhammer. A comparison of sequence and structure protein domain families as a basis for structural genomics. *Bioinformatics*, 15(6):480–500, Jun 1999.

[299] S. B. Pandit, R. Bhadra, V. S. Gowri, S. Balaji, B. Anand, and N. Srinivasan. Supfam: a database of sequence superfamilies of protein domains. *BMC Bioinformatics*, 5:28, Mar 2004.

[300] M. Hedman, H. Deloof, G. Von Heijne, and A. Elofsson. Improved detection of homologous membrane proteins by inclusion of information from topology predictions. *Protein Sci*, 11(3):652–658, Mar 2002.

[301] W.-C. Wong, S. Maurer-Stroh, and F. Eisenhaber. More than 1,001 problems with protein domain databases: transmembrane regions, signal peptides and the issue of sequence homology. *PLoS Comput Biol*, 6(7):e1000867, 2010.

[302] S. Sadekar, J. Raymond, and R. E. Blankenship. Conservation of distantly related membrane proteins: photosynthetic reaction centers share a common structural core. *Mol Biol Evol*, 23(11):2001–2007, Nov 2006.

[303] C. Geourjon, C. Combet, C. Blanchet, and G. Delage. Identification of related proteins with weak sequence identity using secondary structure information. *Protein Sci*, 10(4):788–797, Apr 2001.

[304] J. D. Thompson, B. Linard, O. Lecompte, and O. Poch. A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PLoS One*, 6(3):e18093, 2011.

[305] S. Neumann, H. Hartmann, A. J. Martin-Galiano, A. Fuchs, and D. Frishman. Camps 2.0: Exploring the sequence and structure space of prokaryotic, eukaryotic, and viral membrane proteins. *Proteins*, Nov 2011.

[306] R. F. S. Walters and W. F. DeGrado. Helix-packing motifs in membrane proteins. *Proc Natl Acad Sci U S A*, 103(37):13658–13663, Sep 2006.

[307] S. E. Harrington and N. Ben-Tal. Structural determinants of transmembrane helical proteins. *Structure*, 17(8):1092–1103, Aug 2009.

[308] 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.

[309] 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.

[310] 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.

[311] 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.

[312] A. Fuchs and D. Frishman. Structural comparison and classification of alpha-helical transmembrane domains based on helix interaction patterns. *Proteins*, 78(12):2587–2599, Sep 2010.

[313] 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.

[314] A. Lo, Y.-Y. Chiu, E. A. Rødland, P.-C. Lyu, T.-Y. Sung, and W.-L. Hsu. Predicting helix-helix interactions from residue contacts in membrane proteins. *Bioinformatics*, 25(8):996–1003, Apr 2009.

[315] T. Nugent and D. T. Jones. Predicting transmembrane helix packing arrangements using residue contacts and a force-directed algorithm. *PLoS Comput Biol*, 6(3):e1000714, Mar 2010.

[316] T. Nugent, S. Ward, and D. T. Jones. The mempack alpha-helical transmembrane protein structure prediction server. *Bioinformatics*, 27(10):1438–1439, May 2011.

[317] X.-F. Wang, Z. Chen, C. Wang, R.-X. Yan, Z. Zhang, and J. Song. Predicting residue-residue contacts and helix-helix interactions in transmembrane proteins using an integrative feature-based random forest approach. *PLoS One*, 6(10):e26767, 2011.

[318] S. Bauer, S. Grossmann, M. Vingron, and P. N. Robinson. Ontologizer 2.0–a multifunctional tool for go term enrichment analysis and data exploration. *Bioinformatics*, 24(14):1650–1651, Jul 2008.

[319] M. A. Harris, J. Clark, A. Ireland, and *et al.*, Gene Ontology Consortium. The gene ontology (go) database and informatics resource. *Nucleic Acids Res*, 32(Database issue):D258–D261, Jan 2004.

[320] E. S. Lander, L. M. Linton, B. Birren, and *et al.*, International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001.

[321] J. C. Venter, M. D. Adams, E. W. Myers, and *et al.* The sequence of the human genome. *Science*, 291(5507):1304–1351, Feb 2001.

[322] K. Palczewski, T. Kumasaka, T. Hori, C. A. Behnke, H. Motoshima, B. A. Fox, I. Le 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.

[323] D. L. Jack, I. T. Paulsen, and M. H. Saier. The amino acid/polyamine/organocation (apc) superfamily of transporters specific for amino acids, polyamines and organocations. *Microbiology*, 146 (Pt 8):1797–1814, Aug 2000.

[324] D. L. Theobald and C. Miller. Membrane transport proteins: surprises in structural sameness. *Nat Struct Mol Biol*, 17(1):2–3, Jan 2010.

[325] J. S. Lolkema and D. J. Slotboom. Estimation of structural similarity of membrane proteins by hydropathy profile alignment. *Mol Membr Biol*, 15(1):33–42, 1998.

[326] J. S. Lolkema and D. J. Slotboom. Classification of 29 families of secondary transport proteins into a single structural class using hydropathy profile analysis. *J Mol Biol*, 327(5):901–909, Apr 2003.

[327] J. S. Lolkema and D.-J. Slotboom. The major amino acid transporter superfamily has a similar core structure as na+-galactose and na+-leucine transporters. *Mol Membr Biol*, 25(6-7):567–570, Sep 2008.

[328] R. Ter Horst and J. S. Lolkema. Membrane topology screen of secondary transport proteins in structural class st[3] of the memgen classification. confirmation and structural diversity. *Biochim Biophys Acta*, 1818(1):72–81, Jan 2012.

[329] M. H. Jr. Saier. Convergence and divergence in the evolution of transport proteins. *Bioessays*, 16(1):23–29, Jan 1994.

[330] B. Hummel. Generation and analysis of helix-interaction consensus models for alpha-helical membrane proteins. Master's thesis, Technical University Munich, 2008.

[331] D. B. Wetlaufer. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc Natl Acad Sci U S A*, 70(3):697–701, Mar 1973.

Chapter 7

Appendix

Table 7.1: SCOP folds containing membrane proteins with at least one transmembrane helix

Fold	Description	Domains[a]	Super-families	Families	Min (TMS)[b]	Max (TMS)[b]	Domains in `MP_shared`
c.3	FAD/NAD(P)-binding domain	1	1	1	1	1	0
f.3	Light-harvesting complex subunits	1	1	1	1	1	1
f.13	Family A G protein-coupled receptor-like	6	1	2	2	7	6
f.14	Voltage-gated potassium channel	5	1	1	1	4	2
f.16	Gated mechanosensitive channel	1	1	1	2	2	0
f.17	Transmembrane helix hairpin	6	3	3	1	2	4
f.19	Aquaporin-like	4	1	1	6	6	4
f.20	Clc chloride channel	1	1	1	10	10	1
f.21	Heme-binding four helical bundle	9	3	5	3	5	7
f.22	ABC transporter involved in vitamin B12 uptake, BtuC	1	1	1	10	10	0
f.23	Single transmembrane helix	30	29	29	1	1	18
f.24	Cytochrome *c* oxidase subunit I-like	4	1	1	12	13	4
f.25	Cytochrome *c* oxidase subunit III-like	3	1	1	5	7	3

(Table continues on next page.)

Table 7.1: —*Continued.*

Fold	Description	Domains[a]	Super-families	Families	Min (TMS)[b]	Max (TMS)[b]	Domains in `MP_shared`
f.26	Bacterial photosystem II reaction centre, L and M subunits	2	1	1	5	5	2
f.29	Photosystem I subunits PsaA/PsaB	2	1	1	11	11	2
f.30	Photosystem I reaction center subunit X, PsaK	1	1	1	2	2	1
f.31	Photosystem I reaction center subunit XI, PsaL	1	1	1	3	3	1
f.32	Domain/subunit of cytochrome bc_1 complex (Ubiquinol-cytochrome c reductase)	2	1	1	3	3	1
f.33	Calcium ATPase, transmembrane domain M	1	1	1	10	10	1
f.34	Mechanosensitive channel protein MscS (YggB), transmembrane region	1	1	1	3	3	0
f.35	Multidrug efflux transporter AcrB transmembrane domain	1	1	1	6	6	0
f.36	Neurotransmitter-gated ion-channel transmembrane pore	4	1	1	4	4	4
f.37	ABC transporter transmembrane region	1	1	1	6	6	0
f.38	MFS general substrate transporter	2	1	2	12	12	0
f.41	Preprotein translocase SecY subunit	1	1	1	10	10	0
f.42	Mitochondrial carrier	1	1	1	6	6	0
f.43	Chlorophyll a-b binding protein	1	1	1	3	3	0
f.44	Ammonium transporter	1	1	1	11	11	0
f.49	Proton glutamate symport protein	1	1	1	8	8	0
f.51	Rhomboid-like	2	1	1	6	6	0
i.5	Photosystems	3	1	1	1	12	0
i.18	Membrane ion ATPase	1	1	1	10	10	0
j.35	Transmembrane helical fragments	13	1	1	1	2	1
j.37	Phospholamban fragments	1	1	1	1	1	1

[a] Number of distinct domains according to SCOP unique identifiers (sunid) for protein domains.
[b] Minimal and maximal number of transmembrane segments (TMS) according to PDBTM annotation.

Table 7.2: CATH folds containing membrane proteins with at least one transmembrane helix

Fold	Description	Domains[a]	Super-families	Min (TMS)[b]	Max (TMS)[b]	Domains in `MP_shared`
1.10.8	Helicase, Ruva protein, domain 3	1	1	1	1	1
1.10.287	Helix hairpin	14	6	2	3	6
1.10.442	Cytochrome *c* oxidase, chain D	1	1	1	1	1
1.10.3080	Clc chloride channel	2	1	10	10	1
1.20.5	Single alpha-helices involved in coiled-coils or other helix-helix interfaces	26	12	1	1	10
1.20.20	F_1F_0 ATP synthase	3	1	2	2	1
1.20.85	Photosynthetic reaction center, subunit *m* domain 1	13	1	2	3	4
1.20.120	Four helix bundle (Hemerythrin (Met), subunit *a*)	9	3	4	5	8
1.20.210	Cytochrome *c* oxidase, chain A	5	1	12	13	4
1.20.810	Cytochrome bc_1 complex, chain C	7	1	4	8	2
1.20.860	Alpha-t-alpha	1	1	2	2	1
1.20.950	Fumarate reductase cytochrome *b* subunit	2	2	4	5	2
1.20.1050	Glutathione S-transferase Yfyf (Class Pi), chain A, domain 2	1	1	4	4	0
1.20.1070	Rhodopsin 7-helix transmembrane protein	9	1	7	7	6
1.20.1080	Glycerol uptake facilitator protein	8	1	6	6	4
1.20.1110	Calcium-transporting ATPase, transmembrane domain	1	1	9	10	1
1.20.1130	Photosystem I p700 chlorophyll A apoprotein A1	2	1	11	11	2
1.20.1240	Photosystem 1 reaction centre subunit XI, chain L	1	1	3	3	1
1.20.1300	3 helical TM bundles of succinate and fumarate reductases	3	1	3	3	3
1.20.1450	Particulate methane monooxygenase, chain B	1	1	7	7	0
4.10.49	Cytochrome *c* oxidase, chain L	1	1	1	1	1
4.10.51	Cytochrome *c* oxidase, chain K	1	1	1	1	1

(Table continues on next page.)

Table 7.2: —*Continued.*

Fold	Description	Domains[a]	Super-families	Min (TMS)[b]	Max (TMS)[b]	Domains in `MP_shared`
4.10.81	Cytochrome *c* oxidase, chain M	2	1	1	1	2
4.10.91	Cytochrome *c* oxidase, chain J	1	1	1	1	1
4.10.93	Cytochrome *c* oxidase, chain I	1	1	1	1	1
4.10.95	Cytochrome *c* oxidase, chain G	1	1	1	1	1
4.10.220	Light-harvesting protein	3	1	1	1	1
4.10.540	Photosynthetic reaction center, chain H, domain 1	3	1	1	1	1

[a] Number of distinct domains according to a representative set of CATH domains at 95% sequence identity.
[b] Minimal and maximal number of transmembrane segments (TMS) according to PDBTM annotation.

Table 7.3: Number of initial clusters for each threshold round and additional information

Threshold	Clusters	Clusters with TMH cores	Sequences in clusters	Singletons
1E-5	6,278	1,300	484,104	10,575
1E-6	6,942	1,331	482,589	12,090
1E-7	7,427	1,362	481,612	13,067
1E-8	7,811	1,385	480,893	13,786
1E-9	8,175	1,432	480,220	14,459
1E-10	8,550	1,467	479,557	15,122
1E-11	8,910	1,514	478,940	15,739
1E-12	9,283	1,546	478,373	16,306
1E-13	9,678	1,586	477,737	16,942
1E-14	10,064	1,626	477,077	17,602
1E-15	10,462	1,676	476,428	18,251
1E-16	10,808	1,718	475,791	18,888
1E-17	11,214	1,754	475,185	19,494
1E-18	11,599	1,774	474,490	20,189
1E-19	11,992	1,794	473,849	20,830
1E-20	12,397	1,801	473,146	21,533
1E-22	13,133	1,857	471,811	22,868
1E-24	13,849	1,897	470,309	24,370
1E-25	14,252	1,927	469,575	25,104
1E-26	14,661	1,942	468,769	25,910
1E-28	15,488	1,999	467,225	27,454
1E-30	16,311	2,032	465,621	29,058
1E-35	18,397	2,084	461,265	33,414
1E-40	20,330	2,142	456,547	38,132
1E-45	22,178	2,171	451,384	43,295
1E-50	23,953	2,188	445,799	48,880
1E-55	25,620	2,216	439,842	54,837
1E-60	27,057	2,251	433,695	60,984
1E-70	29,722	2,182	420,169	74,510
1E-80	32,076	2,105	406,051	88,628
1E-90	34,170	2,021	391,874	102,805
1E-100	35,868	1,862	377,101	117,578

Table 7.4: Distribution of TMHs among proteins, SC-clusters and MCL clusters

	Number of TMHs										
	5	6	7	8	9	10	11	12	13	14	15
						Proteins[a]					
Archaea	615	1,402	654	516	412	404	317	464	97	122	20
Bacteria	13,770	26,622	10,939	9,585	10,449	12,774	10,556	14,852	2,482	1,635	263
Eukaryota	3,231	3,636	5,245	1,829	1,598	2,289	2,250	4,026	562	437	122
						SC-clusters[b]					
Archaea	28	52	21	14	10	19	17	16	1	5	1
Bacteria	75	105	41	21	10	35	24	28	4	8	1
Eukaryota	41	50	44	12	8	18	18	22	2	2	2
						MCL clusters[c]					
Archaea	18	16	11	2	4	10	1	4	1	4	1
Bacteria	34	30	24	2	4	19	1	7	1	6	1
Eukaryota	31	23	15	2	2	8	1	5	1	2	2

[a] These values represent the absolute number of proteins with the given number of TMHs.
[b] These values correspond to the absolute number of SC-clusters containing members with the given number of TMHs.
[c] These values correspond to the absolute number of MCL clusters containing members with the given number of TMHs.

Table 7.5: Significantly enriched GO terms for protein class level enrichment analysis.

GO term	Description	Study count	Population count	Adjusted *P*-value
		5 TMH		
GO:0006605	Protein targeting	240	1,128	1.14E-46
GO:0006629	Lipid metabolic process	22	97	4.20E-05
GO:0007049	Cell cycle	137	961	2.08E-13
GO:0007165	Signal transduction	153	708	3.44E-34
GO:0016192	Vesicle-mediated transport	11	13	3.25E-10
GO:0051301	Cell division	140	588	4.43E-36
		6 TMH		
GO:0006461	Protein complex assembly	132	390	1.59E-04
GO:0006605	Protein targeting	456	1,128	7.86E-28
GO:0006629	Lipid metabolic process	45	97	3.86E-05
GO:0006810	Transport	11,303	47,363	4.07E-03
GO:0007165	Signal transduction	275	708	5.59E-18
GO:0022607	Cellular component assembly	132	390	1.59E-04
		7 TMH		
GO:0000003	Reproduction	6	11	3.14E-03
GO:0006461	Protein complex assembly	56	390	3.43E-04
GO:0006629	Lipid metabolic process	20	97	2.25E-03
GO:0007165	Signal transduction	200	708	5.23E-58
GO:0008283	Cell proliferation	5	7	2.26E-03
GO:0022607	Cellular component assembly	56	390	3.43E-04
GO:0040011	Locomotion	12	15	7.92E-10
		8 TMH		
GO:0006461	Protein complex assembly	152	390	8.14E-68
GO:0022607	Cellular component assembly	152	390	8.14E-68
GO:0030154	Cell differentiation	6	15	1.85E-02
		9 TMH		
GO:0007049	Cell cycle	252	961	7.80E-57
GO:0051301	Cell division	107	588	1.07E-11
		10 TMH		
GO:0006605	Protein targeting	235	1,128	5.46E-26
GO:0007049	Cell cycle	427	961	1.61E-167
GO:0009790	Embryo development	5	7	7.19E-03
GO:0015979	Photosynthesis	3	3	3.93E-02
GO:0051301	Cell division	244	588	7.71E-87
		11 TMH		
GO:0005975	Carbohydrate metabolic process	10	16	2.10E-05
GO:0006810	Transport	5,068	47,363	3.93E-41
GO:0055085	Transmembrane transport	2,707	18,420	2.36E-107

(Table continues on next page.)

Table 7.5: —*Continued.*

GO term	Description	Study count	Population count	Adjusted *P*-value
		12 TMH		
GO:0006810	Transport	11,236	47,363	2.09E-185
GO:0006950	Response to stress	34	88	2.06E-02
GO:0055085	Transmembrane transport	8,866	18,420	0
		13 TMH		
GO:0006810	Transport	1,429	47,363	3.40E-15
GO:0006950	Response to stress	32	88	2.26E-25
GO:0055085	Transmembrane transport	1,221	18,420	2.11E-297
		14 TMH		
GO:0007165	Signal transduction	53	708	1.54E-26
GO:0034641	Cellular nitrogen compound metabolic process	4	8	2.25E-05
GO:0044281	Small molecule metabolic process	2	5	2.72E-2
GO:0055085	Transmembrane transport	391	18,420	6.43E-75
		15 TMH		
GO:0055085	Transmembrane transport	40	18,420	1.57E-03

Table 7.6: Significantly enriched GO terms for cluster level enrichment analysis.

Cluster	GO term	Description	Study count	Population count	Adjusted *P*-value
		5 TMH			
mcl1_5tmh	GO:0006629*	Lipid metabolic process	1	64	1.26E-02
mcl2_5tmh	GO:0016192*	Vesicle-mediated transport	1	10	4.50E-03
	GO:0030154	Cell differentiation	3	7	5.43E-09
mcl3_5tmh	GO:0007165*	Signal transduction	2	518	1.68E-02
mcl4_5tmh	GO:0016192*	Vesicle-mediated transport	6	10	2.99E-21
mcl5_5tmh	GO:0007049*	Cell cycle	132	562	5.48E-220
	GO:0007165*	Signal transduction	9	518	5.60E-03
	GO:0051301*	Cell division	132	401	1.23E-242
mcl6_5tmh	GO:0006457	Protein folding	1	5	1.05E-02
mcl7_5tmh	GO:0051301*	Cell division	3	401	4.50E-05
mcl8_5tmh	GO:0007165*	Signal transduction	13	518	9.72E-20
mcl9_5tmh	GO:0007067	Mitosis	1	1	2.60E-03
	GO:0007155	Cell adhesion	2	5	7.77E-06
mcl10_5tmh	GO:0040011	Locomotion	1	14	4.96E-02
mcl11_5tmh	GO:0006950	Response to stress	2	59	2.88E-04
	GO:0007568	Aging	2	4	1.01E-06
	GO:0008283	Cell proliferation	1	3	5.10E-03
mcl12_5tmh	GO:0007165*	Signal transduction	21	518	2.70E-29
	GO:0016192*	Vesicle-mediated transport	3	10	1.83E-09
mcl13_5tmh	GO:0006629*	Lipid metabolic process	10	64	6.48E-27
	GO:0008219	Cell death	1	3	7.90E-03
		6 TMH			
mcl1_6tmh	GO:0040007	Growth	1	12	7.90E-03
	GO:0040011	Locomotion	1	14	9.20E-03
mcl2_6tmh	GO:0006810*	Transport	1,205	28,687	1.97E-04
	GO:0007165*	Signal transduction	37	518	1.52E-02
	GO:0055085	Transmembrane transport	1,205	12,033	0
mcl3_6tmh	GO:0006810*	Transport	1,360	28,687	1.24E-35
mcl4_6tmh	GO:0006457	Protein folding	1	5	9.90E-03
mcl5_6tmh	GO:0006629*	Lipid metabolic process	40	64	4.90E-114
mcl6_6tmh	GO:0007165*	Signal transduction	95	518	5.25E-172
mcl7_6tmh	GO:0006412	Translation	3	5	1.32E-02
	GO:0006810*	Transport	1,540	28,687	3.93E-40
mcl8_6tmh	GO:0006810*	Transport	1,260	28,687	5.87E-33

(Table continues on next page.)

Table 7.6: —*Continued.*

Cluster	GO term	Description	Study count	Population count	Adjusted *P*-value
		6 TMH			
mcl9_6tmh	GO:0006605*	Protein targeting	301	691	0
	GO:0006810*	Transport	308	28,687	1.33E-07
mcl10_6tmh	GO:0006461*	Protein complex assembly	130	286	1.78E-41
	GO:0022607*	Cellular component assembly	130	286	1.87E-41
mcl11_6tmh	GO:0006605*	Protein targeting	16	691	2.87E-24
mcl12_6tmh	GO:0007165*	Signal transduction	25	518	6.71E-45
mcl13_6tmh	GO:0006810*	Transport	1,631	28,687	6.21E-43
mcl14_6tmh	GO:0007165*	Signal transduction	16	518	2.37E-28
mcl15_6tmh	GO:0055085	Transmembrane transport	13	12,033	9.50E-03
		7 TMH			
mcl1_7tmh	GO:0006810	Transport	114	28,687	6.00E-03
mcl2_7tmh	GO:0007165*	Signal transduction	2	518	1.40E-03
mcl3_7tmh	GO:0007165*	Signal transduction	5	518	2.81E-09
mcl4_7tmh	GO:0055085	Transmembrane transport	278	12,033	2.15E-105
	GO:0006810	Transport	278	28,687	1.35E-05
mcl5_7tmh	GO:0007165*	Signal transduction	2	518	1.40E-03
mcl6_7tmh	GO:0000003*	Reproduction	6	8	5.92E-10
	GO:0006629*	Lipid metabolic process	8	64	5.01E-06
	GO:0007165*	Signal transduction	174	518	5.12E-241
	GO:0040007	Growth	5	12	1.77E-06
	GO:0040011*	Locomotion	12	14	1.04E-21
mcl7_7tmh	GO:0055085	Transmembrane transport	8	12,033	9.60E-03
		8 TMH			
mcl1_8tmh	GO:0006461*	Protein complex assembly	152	286	1.63E-110
	GO:0022607*	Cellular component assembly	152	286	1.63E-110
mcl2_8tmh	GO:0040007	Growth	1	12	2.13E-02
		9 TMH			
mcl1_9tmh	GO:0006810	Transport	1,265	28,687	3.86E-17
		10 TMH			
mcl1_10tmh	GO:0007049*	Cell cycle	423	562	0
	GO:0051301*	Cell division	239	401	0
mcl2_10tmh	GO:0006810	Transport	248	28,687	4.70E-06
	GO:0055085	Transmembrane transport	248	12,033	6.61E-94

(Table continues on next page.)

Table 7.6: —*Continued.*

Cluster	GO term	Description	Study count	Population count	Adjusted *P*-value
		10 TMH			
mcl3_10tmh	GO:0006457	Protein folding	2	5	5.44E-06
	GO:0007049*	Cell cycle	4	562	4.63E-05
	GO:0051301*	Cell division	4	401	1.21E-05
mcl4_10tmh	GO:0006605*	Protein targeting	235	691	2.93E-84
	GO:0006810	Transport	2,120	28,687	1.74E-16
	GO:0009790*	Embryo development	5	6	1.85E-04
mcl5_10tmh	GO:0040007	Growth	1	12	2.00E-03
		11 TMH			
mcl1_11tmh	GO:0006810*	Transport	1,102	28,687	1.27E-05
	GO:0055086*	Transmembrane transport	566	12,033	2.49E-09
		12 TMH			
mcl1_12tmh	GO:0055085*	Transmembrane transport	44	12,033	1.43E-16
mcl2_12tmh	GO:0006810*	Transport	1,524	28,687	8.69E-40
mcl3_12tmh	GO:0006810*	Transport	8,168	28,687	8.73E-228
	GO:0055085*	Transmembrane transport	8,141	12,033	0
mcl4_12tmh	GO:0006810*	Transport	346	28,687	7.54E-09
		13 TMH			
mcl1_13tmh	GO:0006810*	Transport	570	28,687	2.29E-14
	GO:0006950*	Response to stress	28	59	8.59E-32
	GO:0055085*	Transmembrane transport	570	12,033	3.43E-218
		14 TMH			
mcl1_14tmh	GO:0055085*	Transmembrane transport	67	12,033	3.77E-25
mcl2_14tmh	GO:0007165*	Signal transduction	53	518	5.87E-95

[a] Asterisks indicate GO terms that were already found to be enriched by the protein class level enrichment analysis (see Table 7.5).

List of publications

Publications included in this thesis:

- **S. Neumann***, A. Fuchs*, A. Mulkidjanian and D. Frishman.
  Current status of membrane protein structure classification.
  Proteins 2010, **78(7)**:1760-1773.

- **S. Neumann**, H. Hartmann, A. J. Martin-Galiano, A. Fuchs and D. Frishman
  CAMPS 2.0: exploring the sequence and structure space of prokaryotic, eukaryotic and viral membrane proteins.
  Proteins 2012, **80(3)**:839-857.

- **S. Neumann**, A. Fuchs, B. Hummel and D. Frishman.
  Classification of α-helical membrane proteins using predicted helix architectures.
  prepared for submission.

Publications not included in this thesis:

- **S. Neumann** and D. Langosch
  Conserved conformational dynamics of membrane fusion protein transmembrane domains and flanking regions indicated by sequence statistics.
  Proteins 2011, **79**:2418-2427.

* These authors contributed equally.