Research Article **Proteins: Structure, Function and Bioinformatics**
DOI 10.1002/prot.22692

Current status of membrane protein structure classification

Sindy Neumann[1,*], Angelika Fuchs[1,*], Armen Mulkidjanian[2,3], and Dmitrij Frishman[1,#]

[1]Department of Genome Oriented Bioinformatics, Technische Universität München, Wissenschaftszentrum Weihenstephan, D-85354 Freising, Germany

[2]School of Physics, Universität Osnabrück, D-49069 Osnabrück, Germany

[3]A.N.Belozersky Institute of Physico-Chemical Biology, Moscow State University, Moscow, 119991, Russia

[*]These authors contributed equally

[#]Corresponding author, e-mail: d.frishman@wzw.tum.de, tel.: +49 8161 71 2134

Short Title: Classification of membrane proteins.

Received: Aug 25, 2009; Revised: Dec 02, 2009; Accepted: Jan 08, 2010

Abstract

For over two decades continuous efforts to organize the jungle of available protein structures have been underway. While a number of discrepancies between different classification approaches for soluble proteins have been reported, the classification of membrane proteins has so far not been comparatively studied due to the limited amount of available structural data.

Here, we present an analysis of α-helical membrane protein classification in the SCOP and CATH databases. In the current set of 63 α-helical membrane protein chains having between one and thirteen transmembrane helices we observed a number of differently classified proteins both regarding their domain and fold assignment. The majority of all discrepancies affect single transmembrane helix, two helix hairpin and four helix bundle domains while domains with more than five helices are mostly classified consistently between SCOP and CATH. It thus appears that the structural constraints imposed by the lipid bilayer complicate the classification of membrane proteins with only few membrane-spanning regions. This problem seems to be specific for membrane proteins as soluble four helix bundles, not restrained by the membrane, are more consistently classified by SCOP and CATH. Our findings indicate that the structural space of small membrane helix bundles is highly continuous such that even minor differences in individual classification procedures may lead to a significantly different classification. Membrane proteins with few helices and limited structural diversity only seem to be reasonably classifiable if the definition of a fold is adapted to include more fine-grained structural features such as helix-helix interactions and reentrant regions.

Key words: fold classification; four helix bundle; molecular evolution; protein fold; structure comparison

Introduction

Structure-based classification of proteins provides a helpful resource to reveal their evolutionary relationships and to obtain an easily accessible overview of the existing protein fold space and the number of naturally observed folds. In addition, it has found widespread application in many areas of structural bioinformatics, including homology modeling, fold recognition, and structural genomics. Several resources for structural classification of proteins exist, with SCOP[1] and CATH[2] being the most widely used ones. Both databases use a hierarchical classification system and rely on a largely similar definition of a protein fold which takes into account the number of secondary structure elements, their spatial orientation, and connectivity.[3,4] While SCOP and CATH incorporate different levels of manual supervision in their classification procedure, other classification approaches have been proposed that completely rely on large-scale structure comparisons and are therefore fully automated.[5-9]

Given the varying classification procedures it is not surprising that several studies have found remarkable differences between individual fold classifications.[6,10,11] The differences arise from variations in the applied domain assignment procedure, which generally is the first step for each classification approach since structural domains are used as classification entities. Furthermore, classification databases may disagree on their fold and homology assignments. Large folds in one database might be divided into several more specific folds within another classification system, leading to proteins belonging to the same fold in the first case but to different folds in the latter case. Even more drastic discrepancies can be observed where one database classifies two proteins into an evolutionary related family while another classification approach places the same pair of proteins into completely different folds due to the fact that proteins may be structurally diverse despite a common evolutionary origin.[10,12] Some of these disagreements clearly arise as a consequence of specific differences in

classification methodologies. At the same time, a recent hotly debated topic is the notion of a continuous protein structure space, which would naturally complicate the classification of proteins into discrete fold categories.[6,13-15] Rooted in the idea that short polypeptides (corresponding to structural motifs) form basic evolutionary units,[16,17] compact domains are said to be constructed from several such substructures leading to local structural similarity of one protein to several other proteins that are not globally related to each other. However, while discussions regarding the nature of the protein fold space are still ongoing,[12] classification approaches such as SCOP and CATH are well established in the scientific community and will continue to serve as valuable tools for structural biologists and bioinformaticians.

So far, comparative analyses of structure classification databases have generally been carried out on the full set of available PDB[18] proteins. Membrane proteins, which account for 20 – 30% of a species' proteome but make up only two percent of all PDB entries, were therefore never in the focus of any previous work. A specific analysis of membrane protein structures in classification approaches such as SCOP and CATH has become feasible only very recently. Historically, such an analysis was impeded by the small number of available experimentally solved structures. Therefore, large-scale approaches trying to estimate the structural variety of membrane proteins had to rely completely on computational methods such as structure prediction and sequence clustering.[19-23] Due to the technological progress in structure determination of membrane proteins[24-27] the number of available unique membrane protein structures has been quickly growing over the past years (179 unique structures as of January 2009, data from the Stephen White lab Web site) allowing for a first glimpse into the structural universe of membrane proteins based on experimentally determined structures. Furthermore, structural classification of membrane proteins has to cope with difficulties arising from several structural properties specific for this type of proteins. Due to the physical constraints imposed by the lipid bilayer, the range of possible membrane protein structures is

limited and, indeed, all integral membrane proteins so far were found to adopt either the α-helix bundle[28,29] or the β-barrel architecture[30,31]. Despite this overall structural uniformity, recently determined membrane protein structures revealed a number of unexpected structural features such as strongly tilted helices[32] and reentrant regions[33]. As the concept of a fold is completely qualitative and allows for varying interpretation even for soluble proteins, its definition with respect to membrane proteins is therefore even less clear, complicating structural classification. Furthermore, considering the idea of a continuous fold space 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.

Here, we present the first comparative analysis of occurrence and classification of α-helical membrane proteins within the two most commonly used structure classification databases, SCOP and CATH. In contrast to the general survey of Hadley and Jones[10], we were specifically interested to learn how these two databases cope with the fact that α-helical membrane proteins share the overall structure of a largely parallel α-helix bundle, while at the same time comprising a significant variety due to specific structural features such as different helix interaction patterns or helix tilts[32]. Furthermore, we wanted to evaluate, based on the extent of classification similarities and discrepancies as well as quantitative structure comparisons, how continuous the currently known structure space of membrane proteins is in order to access the feasibility of structural classification of membrane proteins now and in the future. While, not surprisingly, already known classification discrepancies between SCOP and CATH such as differing domain assignments and the fold overlap problem[10] were observed for membrane proteins as well, we specifically noticed that the chance for a consistent fold assignment between SCOP and CATH clearly decreases for membrane proteins with fewer transmembrane helices. All cases of clear fold disagreements that were found in this work

involve membrane protein domains with one to five helices, with single transmembrane helix, two helix hairpin and four helix bundle domains being particularly frequently involved in these disagreements. In contrast, folds involving six or more transmembrane helices are in good agreement between SCOP and CATH. A comparison with soluble four helix bundle proteins further indicates that structural differences among membrane proteins of this type are less distinct than between their soluble analogs. We therefore conclude that membrane proteins with six or more transmembrane helices are sufficiently diverse to allow for a structural classification comparable to that of soluble proteins. However, for the current set of membrane protein domains with less than six transmembrane helices the structure space seems to be continuous enough to make structural classification into distinct folds a challenging task. We therefore suggest that future classification efforts for small membrane proteins will need to incorporate more fine-grained structural features such as helix-helix interactions, helix packings, helix tilt angles and reentrant regions.

Materials and Methods

Datasets of a-helical membrane proteins with known structure

An initial dataset of α-helical membrane proteins with experimentally determined structure was generated using the Protein Data Bank of Transmembrane Proteins (PDBTM[34]) 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 2673 amino acid chains. Using the cd-hit algorithm[35] 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[1] and CATH v3.2[2], 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 having different chain identifiers that are neither listed in PDB nor in SCOP or CATH; such duplicate chains were ignored. Model structures as well as those structures where the fraction of non-alpha carbon atoms is below 70% for which CATH does not provide classification[36] 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 since 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, since 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 while 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 (Supplementary Table I) and 28 CATH folds (Supplementary Table II).

Comparing domain assignments in SCOP and CATH

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, we directly compared the domain assignments between SCOP and CATH for all proteins in the MP_shared dataset. For those proteins where SCOP and CATH agreed in the number of domains we additionally calculated the extent of domain overlap. 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.

Comparing fold assignments in SCOP and CATH

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.

In order 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[37]. 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 due to 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

([38], available via http://www.cathdb.info/cgi-bin/SsapServer.pl). The functional consistency of SCOP and CATH folds was furthermore evaluated using GO annotations [39] as provided by the GOA group at the EBI (http://www.ebi.ac.uk/QuickGO/).

Analysis of four helix bundle proteins

In order to compare membrane proteins to α-helical soluble structures in terms of diversity and classification, a detailed analysis was conducted based on the class of four helix bundle proteins. Initial datasets of both soluble and membrane four helix bundles were constructed by manually selecting all folds and corresponding protein domains from the all-α and membrane protein classes in SCOP with a fold description containing the terms '4 helices' and 'bundle'. In total, 38 soluble and three transmembrane four helix bundle folds were identified (Supplementary Table III) containing 2601 and 78 protein domain entries, respectively. Applying the same domain redundancy filter as used for the full membrane protein dataset (removal of proteins with the same domain sunid in SCOP), 278 soluble and 14 membrane domains were kept for further analysis. It should be noted that the ATP synthase subunit *a* (1c17, chain M) was included in the dataset of membrane four helix bundle proteins although its structure was not solved experimentally. On the one hand, this was a necessary concession to the limited number of available non-redundant membrane protein structures. On the other hand, as both SCOP and CATH provide a classification for this protein despite their general exclusion of model structures, we thought that a comparison of these classifications would be of equal interest as for experimentally solved structures, although the structure in question may not necessarily be correct.

For every SCOP domain in the dataset of four helix bundle proteins, the corresponding protein domain within the CATH database was retrieved, if available. To this end, an initial set of matching domains was compiled by selecting all CATH domains having at least ten amino acids in common with any of the SCOP four helix bundle domains. As differing

domain positions may lead to different secondary structure composition (e.g. more or fewer helices or even additional beta strands) resulting in a matched CATH domain not representing a four helix bundle structure, the positional consistency of domain assignments between SCOP and CATH was imposed by focusing on only those domains where 90% of the SCOP positions were covered by the CATH domain definition and *vice versa*. While the latter highly restrictive criterion was used to detect those domains whose positions were defined with good agreement in SCOP and CATH, the initial domain set was additionally used to identify those protein chains where SCOP and CATH disagree significantly in their domain assignment. Using all CATH domains with good agreement to a SCOP four helix bundle domain, a list of CATH four helix bundle folds was compiled. In several cases, these folds contained additional protein domains whose classification in SCOP was traced back by a second matching step identifying corresponding SCOP domains for these CATH domains. Finally, all obtained domains were used to identify four helix bundle folds containing exactly the same proteins in SCOP and CATH as well as those folds spread over several folds in the respective other database.

Within a second analysis addressing the structural diversity of membrane and soluble four helix bundles, all-against-all protein structure comparisons were executed for all distinct protein domains with high (90%) positional agreement between SCOP and CATH domain boundaries using DaliLite v3.1[37]. Obtained similarity scores (Z-scores) as well as the fraction of aligned residues with respect to the smaller of the two compared structures (coverage) were used to compare proteins classified into the same fold in both SCOP and CATH to proteins classified either together within one database but separately in the other database or separately in both databases.

Results and Discussion

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 (Supplementary Table I) and 28 CATH (Supplementary Table II) 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 (8 folds). Within the former class, four 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 four out of 28 CATH membrane protein folds (14.3%) are associated with more than one superfamily (Supplementary Table II). In SCOP, three membrane protein folds are further subdivided into more than one superfamily, and only five folds contain more than one family (Supplementary Table I), 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. While the number of newly identified folds for soluble proteins is steadily decreasing,[40] 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 (Supplementary Tables I and II). 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 [41,42] to allow any conclusions about the most prevalent membrane protein folds in nature. However, we note 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 [43-46]. On the other hand, it might also indicate that proteins with fewer transmembrane helices are harder to classify into different folds due to more limited degree of structural variation (see also below), 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 (Supplementary Tables I and II). 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 [47], although cytochrome b consists of eight transmembrane helices while cytochrome b_6 has only four helices. 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 while the C-terminal domain is classified separately.

Comparison of domain assignments in SCOP and CATH

Since domains are the basic units of protein structure classification in SCOP and CATH, the agreement in their assignments was first analyzed. SCOP and CATH use different methods to decompose 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.[48] Accordingly, previous analyses reported significant differences between SCOP and CATH in the resulting domain assignments considering all proteins in the respective databases.[10,11] However, since most of the membrane proteins are single-domain proteins,[49] 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. The observation that CATH identifies more multi-domain proteins than SCOP was already reported in the work of Hadley and Jones[10] and was found to be a direct result of the different domain definitions that are used in the two databases with CATH addressing geometrical aspects while SCOP also incorporating functional considerations.

In order to further elucidate differences in the domain assignments between SCOP and CATH we examined the separation into domains 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, while 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. Only in one case (cytochrome *b,* discussed above, Figure 1A), SCOP assigned two domains while CATH classified the full protein as one domain. In the remaining four cases, CATH separated a single domain protein from SCOP into two domains. Figure 1B displays one of these cases (subunit III of the cytochrome *c* oxidase) where the existence of a V-shaped cleft between two helix bundles seems to cause the domain split (further details regarding the domain assignment and classification of these cases are summarized in the Supplementary Material).

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.

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 [11]). 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 multidomain membrane proteins found in SCOP or CATH (2 and 6 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.

Comparison of fold assignments in SCOP and CATH

In order to determine the agreement between SCOP and CATH with respect to their fold

classification, we compared the fold assignments of all proteins in the MP_shared dataset. As MP_shared is a subset of MP_SCOP and MP_CATH it does not cover all known α-helical membrane protein folds (Supplementary Tables I and II). While 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, Supplementary Tables I and II).

Fold agreements

Nine folds were found to contain exactly the same domains in SCOP and CATH (Table I). 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 we observed a high degree of structural similarity among all proteins of the same fold. Average Z-scores for comparisons of the same fold (Table II) 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 ten transmembrane helices but classified consistently into two different folds (PDB 2exw, chain A

from fold f.20/1.10.3080 and PDB 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,[37] while a Z-score above 20 means that two structures are true homologs (see the DaliLite help file at http://www.ebi.ac.uk/Tools/dalilite/).

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[10] 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 I) and were already discussed above, 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.

Analyzing 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 II). 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 [10]. Inspecting the observed cases of fold overlaps more closely (for detailed descriptions see the Supplementary Material), 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, in the SCOP fold f.14 two different conformations of a potassium channel are classified together despite rather low similarity (Z-score 3.4; Figure 2, Table II). In contrast, CATH classifies both protein chains not only to different folds, but even to different architectures (1.10.287 and 1.20.120). The opposite effect is observed in the case of CATH fold 1.20.120 which covers proteins such as 1oedB (acetylcholine receptor, beta chain) and 1fftC (ubiquinol oxidase) with only a weak structural similarity (Dali Z-score of 8.1, SSAP score of 66 and SSAP RMSD as high as 7.63 Å) that are in fact not fulfilling the requirements of a common CATH classification (required SSAP score > 70). While these proteins end up in the same CATH fold due to the single-linkage clustering approach used during classification [50], SCOP places them in different folds (f.25 and f.36) which have homogenous functional GO assignments[39] 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 due to their short sequence length (Table II, folds 1.20.5, 4.10.81 and f.23). We therefore conducted additional comparisons using the SSAP algorithm [38] as well as manual structure inspection to analyze the structural consistency of bitopic

membrane protein folds in SCOP and CATH. Notably, SCOP collects most single helix domains of transmembrane proteins within fold f.23 irrespective of the presence of any additional extramembranous domains or their functional annotation (Figure 3). In contrast, CATH produces a structurally more meaningful classification of bitopic proteins in that all domains having either no globular parts or only globular stretches without secondary structure are classified within fold 1.20.5 while other bitopic membrane domains are split into several folds depending on their globular portions. 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) (for examples see Figure 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. While 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.

In summary, our comparative analysis of membrane proteins in SCOP and CATH leads us to conclude 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. Only for proteins with ten, four,

two and one helices more than one fold is observed. 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 can be seen from the fact that almost no cases of fold agreement can be detected for these proteins at the moment. Only three 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 [51,52]). According to the recent estimates, these features might occur quite frequently. For example, 10% of all polytopic membrane proteins are expected to contain reentrant regions.[33] Naturally, the possibility for structural modification and variation significantly increases with the number of available transmembrane helices. Accordingly, while 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 could be 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,[53] 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.[54] 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,[55] where each hairpin is stabilized by interactions with its neighbours. 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.

Comparison of membrane and soluble four helix bundle proteins

As four helix bundle proteins were found to be among those proteins posing problems for the structural classification of membrane proteins, we wondered whether this was due to some intrinsic properties of this particular architecture, or whether the structural restrictions imposed by the lipid bilayer additionally impede the classification. As described by Hadley and Jones (1999), discrepancies between SCOP and CATH occur frequently and at different levels and are therefore clearly not specific for membrane proteins. However, early studies already showed significant diversity among soluble four helix bundles despite the low number of helices.[56-58] Even with the limited number of available structures at that time, Harris and colleagues identified six classes of bundles depending on the pattern of interhelical angles, indicating a variety likely not possible for membrane proteins. In order to test whether this additional variety facilitates the classification, we compiled a non-redundant data set of 278 soluble four helix bundle domains from 38 SCOP folds (see *Materials and Methods*).

Classification and structural similarity among these domains was analyzed and compared to the analogous dataset consisting of 14 distinct membrane four helix bundle domains from three SCOP folds. Four helix bundle domains were chosen for this analysis over helix hairpin structures since SCOP contains only one soluble all-alpha fold with a 'helix hairpin' description where all hairpin structures of soluble proteins seem to be gathered.

Assignment of four helix bundle domains by SCOP and CATH

The four helix bundle structural motif occurs both independently and as part of larger structures.[57] Consistent domain assignments are therefore essential in order to guarantee that a domain classified as a four helix bundle in one database corresponds to a four helix bundle structure in another database. Comparing domain positions assigned by SCOP and CATH to the same PDB chain, 212 domains out of 278 soluble four helix bundle domain entries in SCOP (76.3%), were found to share ten or more sequence positions with a CATH domain while for the remaining 66 SCOP domains no matching CATH domain could be found. When a positional agreement of at least 90% between equivalent domains was required, the number of detected domains in CATH dropped to 188 corresponding to 67.6% of the full set of SCOP four helix bundle domains. Accordingly, in more than 30% of all soluble four helix bundle domains defined by SCOP, either no matching domain in CATH could be found at all, or domain positions in CATH were so strongly deviating that the secondary structure content of the matching domain was likely to be significantly altered.

Executing the same analysis for membrane proteins, 11 out of 14 distinct SCOP domains (78.6%) could be recovered in CATH with more than 90% identical positions. In fact, only one case, cytochrome *b*, was found where SCOP, but not CATH, classified a part of the structure as a four helix bundle, whereas only SCOP detects a heme-binding four helix bundle within the complete structure consisting of eight helices.

As differing domain assignments naturally interfere with consistent fold assignments, we used for all further analyses only those domains having equivalent positional assignment (90% identical positions) in both SCOP and CATH.

Classification of four helix bundle domains by SCOP and CATH

Given 188 soluble and 11 membrane domain entries from 28 and 3 different SCOP four helix bundle folds, respectively, the consistency between SCOP and CATH with respect to their fold classification was analyzed. To this end, we calculated the frequency of all domain pairs classified into the same fold in one database that are also assigned to the same fold in the other database. For soluble proteins, these percentages were remarkably high with 88% of all co-classified SCOP domains appearing also in the same fold in CATH and as many as 94% of all CATH co-classified domain pairs being in the same SCOP fold. However, the latter number disregards the fact that several of the analyzed CATH folds included additional domains either not present in SCOP at all or classified as a non four helix bundle fold by SCOP and hence not incorporated in this analysis. In total, eleven 1:1 fold relationships were observed where SCOP folds contain exactly the same domains as the respective CATH folds based on a domain matching criterion of 90% identical positions (Figure 4).

Despite the overall good agreement between SCOP and CATH, a number of fold disagreements could also be detected for soluble four helix bundle proteins. While six soluble SCOP folds were completely missing from CATH, three folds did not correspond to the mainly-α folds in CATH. In two cases these folds belonged to the 'few secondary structure' class (folds a.38 and a.53) and in one case (fold a.162) the corresponding fold in CATH was assigned to the 'mixed alpha-beta' class due to the presence of short beta-hairpins neglected by SCOP. Furthermore, several 1:N relationships ('fold overlaps') existed where either a SCOP or a CATH fold was acting as a superfold to several folds of the other database. Even

more complex cases of N:M relationships were also observed such that a SCOP fold corresponded to several CATH folds of which at least one could be mapped back again to several SCOP folds (Figure 4). In those cases where a single CATH fold covered more than one SCOP fold, it often also contained proteins not belonging to one of the original SCOP four helix bundle folds. Interestingly, even if SCOP and CATH domains shared 90% of identical positions, the remaining difference in domain positions can be enough for secondary structure elements to be cut off or added, changing the classification from a four helix bundle to either a bundle with different helix number or to a fold with few secondary structures or additional beta sheets. Nevertheless, only a minority of the soluble four helix bundle domains were affected by these discrepancies, while most of the proteins were classified consistently within the same fold in both databases.

As discussed above, the consistency between SCOP and CATH is much lower for membrane four helix bundles than for membrane folds with more transmembrane helices. For these proteins, from all domain pairs within the same SCOP four helix bundle fold, only 48% were also found in the same CATH fold while 71% of all CATH co-classified domain pairs were also in the same SCOP fold. No case of a complete fold agreement between SCOP and CATH could be found for membrane four helix bundles (Figure 4). It thus appears that membrane four helix bundle proteins are more difficult to classify than their soluble counterparts, although in principle such low degree of agreement between SCOP and CATH for this particular fold could be a statistical artifact caused by the paucity of available data.

Structural similarity among four helix bundle proteins

In order to explain the observed difficulties in the classification of membrane four helix bundle proteins, all-against-all structure comparisons were calculated using DaliLite and the observed structural variability of membrane and soluble four helix bundle domains was compared (Table III). To this end, all compared domain pairs were grouped into three

categories. The first category contained all comparisons concerning protein pairs consistently classified to the same fold within both CATH and SCOP, the second class covered all protein pairs classified to the same fold in only one of the two databases, and the third class was made up by protein pairs belonging to different folds in both databases. For soluble proteins, the last category was significantly different from the other two categories. For a large fraction of these comparisons (45.4%), DaliLite could not detect any similarity and hence did not produce any result. For the remaining comparisons, both the average fraction of aligned residues (coverage) and the average Z-score were clearly smaller than for proteins classified either in one or in both databases to the same fold indicating a structural diversity sufficient to discern individual folds. Analyzing specifically those folds containing exactly the same domains in SCOP and CATH (Supplementary Table IV) we observed that these folds in fact represent distinct regions of the structure space of four helix bundle domains. In all cases structure comparisons of proteins within each fold returned average Z-scores at least twice as high as comparisons of fold members with proteins not belonging to the respective fold. In most cases, the latter comparisons resulted in average Z-scores clearly below 2.0 and hence the analyzed proteins can be considered to share no significant similarity[37] despite the common four helix bundle architecture. Accordingly, these structurally isolated groups of proteins are assigned to individual folds in both SCOP and CATH.

The structural space for the respective membrane proteins on the other hand seems to be much more continuous. As expected given the structural limitations of the lipid bilayer in combination with the limited number of helices, structural differences among all proteins, no matter whether they are classified within SCOP and/or CATH to the same or to a different fold, are less pronounced. The average coverage and average Z-scores in all three analyzed categories are higher than for soluble proteins. Even proteins classified to different folds in both databases still have an average Z-score of 4.5 and an average coverage of 69.9% which is comparable to those soluble proteins classified to the same fold in one database but to

separate folds in the other database. Furthermore, DaliLite was able to detect at least a minimal similarity (Z-score > 2.0) among all proteins that were assigned to different folds, which was clearly not the case for their soluble counterparts where 45% of all comparisons did not retrieve any result and additional 30% resulted in a Z-score of less than two. As structural variations are more fine-grained for membrane four helix bundle proteins, their classification represents a harder problem as long as the same rules are applied as for soluble proteins. We expect that this problem will continue to persist as more solved structures become available unless a more specific fold definition for this type of proteins is at hand.

Conclusions

We present the first analysis of the classification of membrane proteins in SCOP and CATH. Thereby, the following topics were addressed: (a) the general occurrence of membrane proteins and folds within SCOP and CATH, (b) differences in their domain assignments, (c) differences in their fold assignments, and (d) a comparative analysis regarding the classification and diversity of soluble and membrane four helix bundle proteins in SCOP and CATH. Given the current census of membrane protein structures we observed a reasonable agreement between SCOP and CATH for all domains with six or more transmembrane helices. Discrepancies previously described for soluble proteins (differing domain assignments and fold overlap problems[10,11]) are so far to a large extent affecting 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. Interestingly, a comparison to soluble four helix bundle proteins revealed that this observation is not automatically tied to a possibly limited structural variability of four helix bundles *per se*, but rather is specific for membrane proteins. As membrane domains generally display a higher similarity among each other, their structural space can be interpreted as being more continuous than for soluble proteins, making their classification intrinsically more difficult.

Obviously, membrane proteins with more transmembrane helices are also structurally restricted and hence likely to be more similar between each other than soluble helix bundles. However, for membrane folds with more helices the spectrum of possible structural variations (such as tilted or reentrant helices) grows as well. We therefore assume that with the growing number of available structures the identification of distinct folds for these proteins will be possible and accordingly a classification similar to that for soluble proteins is feasible.

Nevertheless, in order to fully address the question how membrane proteins provide such a variety of different functions in the context of restricted structural variability, proteins with only few transmembrane helices need to be considered as well and hence the meaning of the term 'fold' should be redefined at least 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 helices, as well as functional features.

Acknowledgments

We would like to thank Erik Granseth for careful proofreading the manuscript. This work was supported by the DFG grant "A comprehensive analysis of structure-function relationships in membrane protein families" (FR 1411/5-1).

Figure legends

Figure 1: Domain assignment discrepancies between SCOP and CATH. **A**: Mitochondrial cytochrome *b* subunit of the cytochrome bc_1 complex (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 (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[34].

Figure 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 (1ors, chain C) **B**: KcsA potassium channel (2atk, chain C). The coordinates for the transmembrane helices of each domain (shown in red) were extracted from PDBTM[34].

Figure 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 bc1 (1be3, chain K). **B**: Photosystem 1 reaction centre subunit 3 (1jb0, chain F). **C**: Cytochrome c oxidase subunit 4 (2dyr, chain D). The coordinates for the transmembrane helices of each domain (shown in violet) were extracted from PDBTM[34].

Figure 4: Relationships between four helix bundle folds in SCOP (orange) and CATH (blue). Starting with a list of manually selected SCOP four helix bundle folds, CATH folds

containing at least one protein domain of these SCOP folds were selected. In case the extracted CATH folds covered additional protein domains, their occurrence within SCOP was detected by a second domain matching step. Depicted are both soluble (A) and membrane (B) four helix bundle folds. Dark colored folds are related to each other by domains having at least 90% of all positions in common, while light colored folds are only weakly connected by one or more domains sharing at least ten amino acids. In any case, folds may contain additional domains which are not present in the respective other database at all. $\overline{4HB}$ is used to denote SCOP folds not present in the original list of four helix bundle folds (see *Materials and Methods*). For membrane proteins, these folds (f.25 and f.14) are explicitly depicted.

References

1. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. Data growth and its impact on the SCOP database: new developments. Nucleic acids research 2008;36(Database issue):D419-425.

2. Cuff AL, Sillitoe I, Lewis T, Redfern OC, Garratt R, Thornton J, Orengo CA. The CATH classification revisited--architectures reviewed and new ways to characterize structural divergence in superfamilies. Nucleic acids research 2009;37(Database issue):D310-314.

3. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. Journal of molecular biology 1995;247(4):536-540.

4. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH--a hierarchic classification of protein domain structures. Structure 1997;5(8):1093-1108.

5. Holm L, Sander C. Touring protein fold space with Dali/FSSP. Nucleic acids research 1998;26(1):316-319.

6. Shindyalov IN, Bourne PE. An alternative view of protein fold space. Proteins 2000;38(3):247-260.

7. Dietmann S, Park J, Notredame C, Heger A, Lappe M, Holm L. A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. Nucleic acids research 2001;29(1):55-57.

8. Rogen P, Fain B. Automatic classification of protein structure by using Gauss integrals. Proceedings of the National Academy of Sciences of the United States of America 2003;100(1):119-124.

9. Sam V, Tai CH, Garnier J, Gibrat JF, Lee B, Munson PJ. Towards an automatic classification of protein structural domains based on structural similarity. BMC Bioinformatics 2008;9:74.

10. Hadley C, Jones DT. A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. Structure 1999;7(9):1099-1112.

11. Day R, Beck DA, Armen RS, Daggett V. A consensus view of fold space: combining SCOP, CATH, and the Dali Domain Dictionary. Protein Sci 2003;12(10):2150-2160.

12. Pascual-Garcia A, Abia D, Ortiz AR, Bastolla U. Cross-over between discrete and continuous protein structure space: insights into automatic classification and networks of protein structures. PLoS Comput Biol 2009;5(3):e1000331.

13. Yang AS, Honig B. 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. Journal of molecular biology 2000;301(3):665-678.

14. Kolodny R, Petrey D, Honig B. Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction. Current opinion in

structural biology 2006;16(3):393-398.

15. Petrey D, Honig B. Is protein classification necessary? Toward alternative approaches to function annotation. Current opinion in structural biology 2009.

16. Lupas AN, Ponting CP, Russell RB. On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? J Struct Biol 2001;134(2-3):191-203.

17. Grishin NV. Fold change in evolution of protein structures. J Struct Biol 2001;134(2-3):167-185.

18. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic acids research 2000;28(1):235-242.

19. Jones DT. Do transmembrane protein superfolds exist? FEBS letters 1998;423(3):281-285.

20. Bowie JU. Helix-bundle membrane protein fold templates. Protein Sci 1999;8(12):2711-2719.

21. Remm M, Sonnhammer E. Classification of transmembrane protein families in the Caenorhabditis elegans genome and identification of human orthologs. Genome research 2000;10(11):1679-1689.

22. Oberai A, Ihm Y, Kim S, Bowie JU. A limited universe of membrane protein families and folds. Protein Sci 2006;15(7):1723-1734.

23. Martin-Galiano AJ, Frishman D. Defining the fold space of membrane proteins: the CAMPS database. Proteins 2006;64(4):906-922.

24. Grisshammer R, Tate CG. Overexpression of integral membrane proteins for structural studies. Q Rev Biophys 1995;28(3):315-422.

25. Torres J, Stevens TJ, Samso M. Membrane proteins: the 'Wild West' of structural biology. Trends Biochem Sci 2003;28(3):137-144.

26. Opella SJ, Marassi FM. Structure determination of membrane proteins by NMR spectroscopy. Chem Rev 2004;104(8):3587-3606.

27. Caffrey M. Crystallizing Membrane Proteins for Structure Determination: Use of Lipidic Mesophases. Annu Rev Biophys 2008.

28. Rees DC, Komiya H, Yeates TO, Allen JP, Feher G. The bacterial photosynthetic reaction center as a model for membrane proteins. Annual review of biochemistry 1989;58:607-633.

29. Wallin E, Tsukihara T, Yoshikawa S, von Heijne G, Elofsson A. Architecture of helix bundle membrane proteins: an analysis of cytochrome c oxidase from bovine mitochondria. Protein Sci 1997;6(4):808-815.

30. Weiss MS, Abele U, Weckesser J, Welte W, Schiltz E, Schulz GE. Molecular architecture and electrostatic properties of a bacterial porin. Science

1991;254(5038):1627-1630.

31. Wimley WC. The versatile beta-barrel membrane protein. Current opinion in structural biology 2003;13(4):404-411.

32. Yeagle PL, Bennett M, Lemaitre V, Watts A. Transmembrane helices of membrane proteins may flex to satisfy hydrophobic mismatch. Biochim Biophys Acta 2007;1768(3):530-537.

33. Viklund H, Granseth E, Elofsson A. Structural classification and prediction of reentrant regions in alpha-helical transmembrane proteins: application to complete genomes. Journal of molecular biology 2006;361(3):591-603.

34. Tusnady GE, Dosztanyi Z, Simon I. PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. Nucleic acids research 2005;33(Database issue):D275-278.

35. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics (Oxford, England) 2006;22(13):1658-1659.

36. Michie AD, Orengo CA, Thornton JM. Analysis of domain structural class using an automated class assignment protocol. Journal of molecular biology 1996;262(2):168-185.

37. Holm L, Kaariainen S, Rosenstrom P, Schenkel A. Searching protein structure databases with DaliLite v.3. Bioinformatics (Oxford, England) 2008;24(23):2780-2781.

38. Orengo CA, Taylor WR. SSAP: sequential structure alignment program for protein structure comparison. Methods Enzymol 1996;266:617-635.

39. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25(1):25-29.

40. Caetano-Anolles G, Wang M, Caetano-Anolles D, Mittenthal JE. The origin, evolution and structure of the protein world. Biochem J 2009;417(3):621-637.

41. Brenner SE, Chothia C, Hubbard TJ. Population statistics of protein structures: lessons from structural classifications. Current opinion in structural biology 1997;7(3):369-376.

42. Gerstein M. How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. Fold Des 1998;3(6):497-512.

43. Arkin IT, Brunger AT, Engelman DM. Are there dominant membrane protein families with a given number of helices? Proteins 1997;28(4):465-466.

44. Gerstein M. A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. Journal of molecular biology

1997;274(4):562-576.

45. Wallin E, von Heijne G. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. Protein Sci 1998;7(4):1029-1038.

46. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. Journal of molecular biology 2001;305(3):567-580.

47. Soriano GM, Ponamarev MV, Carrell CJ, Xia D, Smith JL, Cramer WA. 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 1999;31(3):201-213.

48. Jones S, Stewart M, Michie A, Swindells MB, Orengo C, Thornton JM. Domain assignment for protein structures using a consensus approach: characterization and analysis. Protein Sci 1998;7(2):233-242.

49. Liu Y, Gerstein M, Engelman DM. Transmembrane protein domains rarely use covalent domain recombination as an evolutionary mechanism. Proceedings of the National Academy of Sciences of the United States of America 2004;101(10):3495-3497.

50. Orengo CA, Martin AM, Hutchinson G, Jones S, Jones DT, Michie AD, Swindells MB, Thornton JM. Classifying a protein in the CATH database of domain structures. Acta crystallographica 1998;54(Pt 6 Pt 1):1155-1167.

51. Fleishman SJ, Ben-Tal N. Progress in structure prediction of alpha-helical membrane proteins. Current opinion in structural biology 2006;16(4):496-504.

52. Elofsson A, von Heijne G. Membrane protein structure: prediction versus reality. Annual review of biochemistry 2007;76:125-140.

53. Jekely G. Did the last common ancestor have a biological membrane? Biol Direct 2006;1:35.

54. Mulkidjanian AY, Galperin MY, Koonin EV. Co-evolution of primordial membranes and membrane proteins. Trends Biochem Sci 2009;34(4):206-215.

55. Meier T, Polzer P, Diederichs K, Welte W, Dimroth P. Structure of the rotor ring of F-Type Na+-ATPase from Ilyobacter tartaricus. Science 2005;308(5722):659-662.

56. Presnell SR, Cohen FE. Topological distribution of four-alpha-helix bundles. Proceedings of the National Academy of Sciences of the United States of America 1989;86(17):6592-6596.

57. Harris NL, Presnell SR, Cohen FE. Four helix bundle diversity in globular proteins. Journal of molecular biology 1994;236(5):1356-1368.

58. Kamtekar S, Hecht MH. Protein Motifs. 7. The four-helix bundle: what determines a fold? FASEB J 1995;9(11):1013-1022.

Table I. Comparison of SCOP and CATH with respect to α-helical membrane protein fold classification.

	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[a]	34	28
- with more than 1 superfamily	3	4
- with more than 1 family	5	-
General comparison using independent datasets (MP_SCOP, MP_CATH)		
Number of protein chains with fold assignment	156	110
Domain assignments		
- 1 domain per chain	152	101
- 2 domains per chain	4	9
Direct comparison using the shared dataset (MP_shared)		
Domain assignments (SCOP:CATH)		
- 1:1	58	
- 2:1	1	
- 1:2	4	
Fold agreements	f.3 ↔ 4.10.220 f.13 ↔ 1.20.1070 f.19 ↔ 1.20.1080 f.20 ↔ 1.10.3080 f.24 ↔ 1.20.210 f.29 ↔ 1.20.1130 f.30 ↔ 1.20.860 f.31 ↔ 1.20.1240 f.33 ↔ 1.20.1110	
Fold disagreement caused by domain disagreement	f.25 → 1.10.287 + 1.20.120 f.26 → 1.20.85 + 1.20.85 1.20.810 → f.21 + f.32	
Fold disagreement caused by fold overlap	f.14 → 1.10.287, 1.20.120 f.17 → 1.10.287, 1.20.20 f.21 → 1.20.810, 1.20.950, 1.20.1300 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 1.10.287 → f.14, f.17 1.20.5 → f.23, j.35, j.37 1.20.120 → f.14, f.25, f.36	

[a]Folds containing α-helical membrane proteins with at least one transmembrane helix

Table II. All-against-all structure comparisons between membrane proteins with agreeing and disagreeing fold assignments in SCOP and CATH using DaliLite.

Folds compared	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.

Table III. All-against-all structure comparisons of membrane and soluble four helix bundle domains using DaliLite.

		Category 1[a]	Category 2[b]	Category 3[c]
Soluble	Number of comparisons	1321	242	16015
	NA[d]	3.5%	4.1%	45.4%
	Average coverage[e]	74.3%	70.9%	57.1%
	Average Z-score	6.7	5.3	2.1
Membrane	Number of comparisons	10	15	30
	NA[d]	-	6.7%	-
	Average coverage[e]	87.5%	74.3%	69.9%
	Average Z-score	14.4	6.0	4.5

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

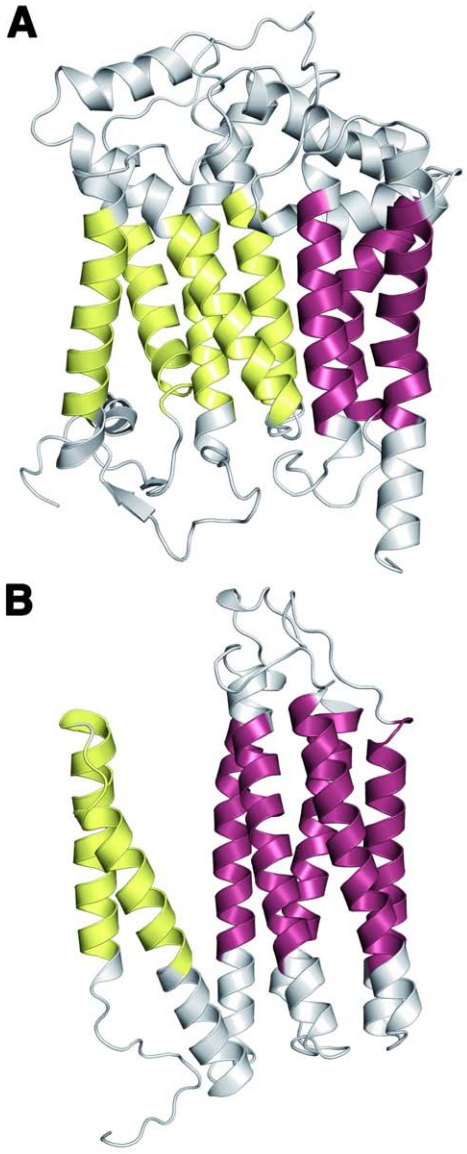

Figure 1: Domain assignment discrepancies between SCOP and CATH. A: Mitochondrial cytochrome b subunit of the cytochrome bc1 complex (1be3, chain C) is classified as a single-domain protein in CATH, but is divided into two domains in SCOP. B: Mitochondrial cytochrome c oxidase, subunit III (1oco, chain 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.
59x139mm (600 x 600 DPI)

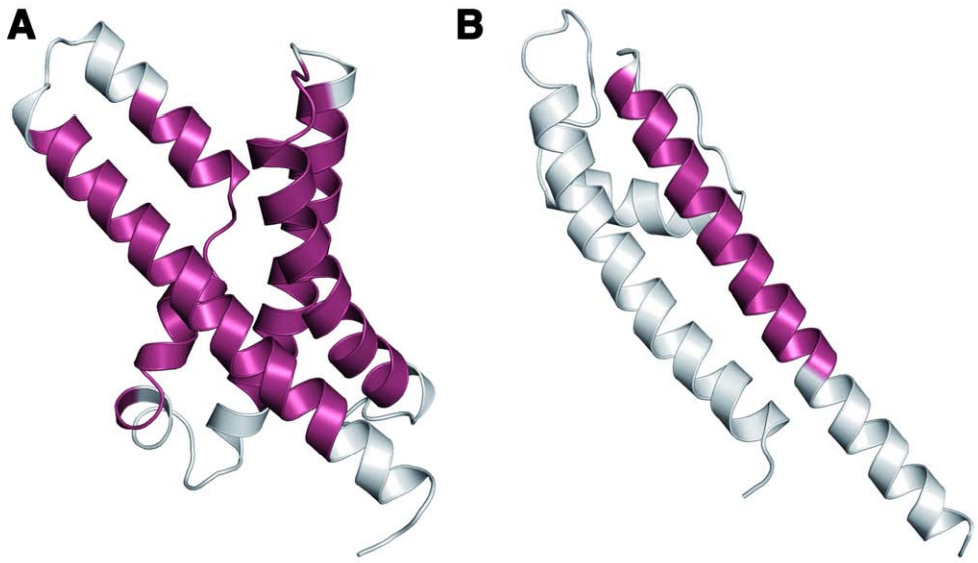

Figure 2: Potassium channels assigned to the voltage-gated potassium channel fold in SCOP (f.14) despite high structural diversity. A: KvAP potassium channel voltage sensor (1ors, chain C) B: KcsA potassium channel (2atk, chain C). The coordinates for the transmembrane helices of each domain (shown in red) were extracted from PDBTM.
122x69mm (600 x 600 DPI)

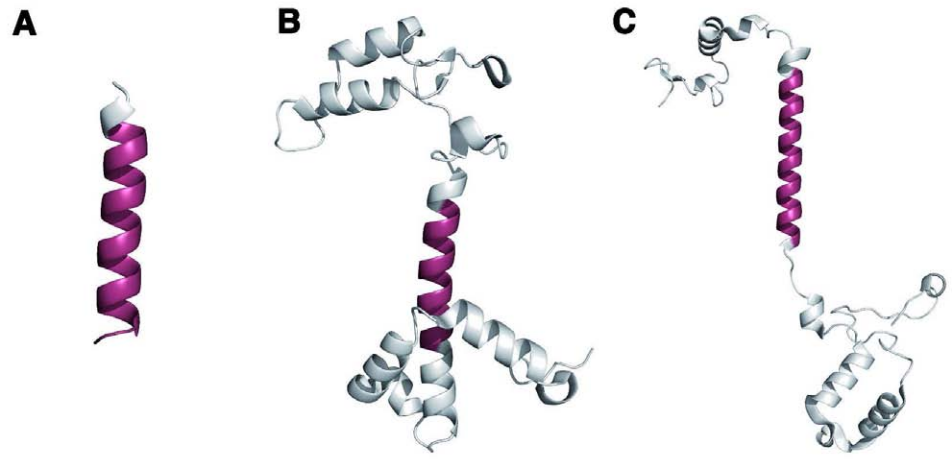

Figure 3: Common classification of bitopic membrane protein domains into SCOP fold f.23 despite different structural elements in their globular portions. In contrast, all three proteins are classified into different folds within CATH. A: Subunit of cytochrome bc1 (1be3, chain K). B: Photosystem 1 reaction centre subunit 3 (1jb0, chain F). C: Cytochrome c oxidase subunit 4 (2dyr, chain D). The coordinates for the transmembrane helices of each domain (shown in violet) were extracted from PDBTM.
146x70mm (600 x 600 DPI)

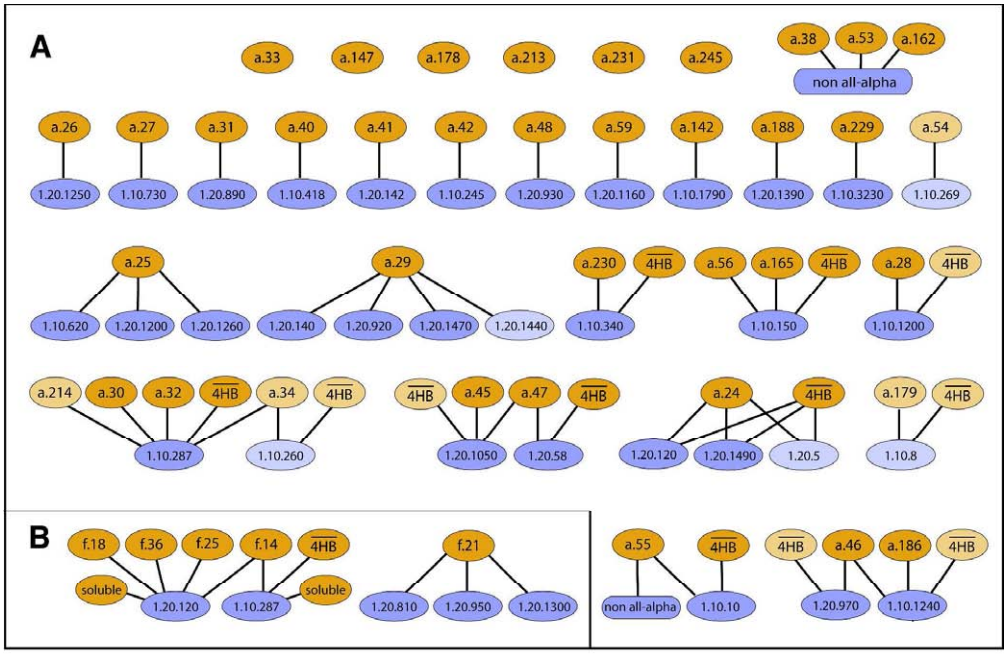

Figure 4: Relationships between four helix bundle folds in SCOP (orange) and CATH (blue). Starting with a list of manually selected SCOP four helix bundle folds, CATH folds containing at least one protein domain of these SCOP folds were selected. In case the extracted CATH folds covered additional protein domains, their occurrence within SCOP was detected by a second domain matching step. Depicted are both soluble (A) and membrane (B) four helix bundle folds. Dark colored folds are related to each other by domains having at least 90% of all positions in common, while light colored folds are only weakly connected by one or more domains sharing at least ten amino acids. In any case, folds may contain additional domains which are not present in the respective other database at all. 4HB is used to denote SCOP folds not present in the original list of four helix bundle folds (see Materials and Methods). For membrane proteins, these folds (f.25 and f.14) are explicitly depicted.
159x103mm (600 x 600 DPI)