OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

More Than 1,001 Problems with Protein Domain Databases: Transmembrane Regions, Signal Peptides and the Issue of Sequence Homology

Wing-Cheong Wong[1]*, Sebastian Maurer-Stroh[1,2]*, Frank Eisenhaber[1,3,4]*

1 Bioinformatics Institute (BII), Agency for Science, Technology and Research (A*STAR), Singapore, **2** School of Biological Sciences (SBS), Nanyang Technological University (NTU), Singapore, **3** Department of Biological Sciences (DBS), National University of Singapore (NUS), Singapore, **4** School of Computer Engineering (SCE), Nanyang Technological University (NTU), Singapore

Abstract

Large-scale genome sequencing gained general importance for life science because functional annotation of otherwise experimentally uncharacterized sequences is made possible by the theory of biomolecular sequence homology. Historically, the paradigm of similarity of protein sequences implying common structure, function and ancestry was generalized based on studies of globular domains. Having the same fold imposes strict conditions over the packing in the hydrophobic core requiring similarity of hydrophobic patterns. The implications of sequence similarity among non-globular protein segments have not been studied to the same extent; nevertheless, homology considerations are silently extended for them. This appears especially detrimental in the case of transmembrane helices (TMs) and signal peptides (SPs) where sequence similarity is necessarily a consequence of physical requirements rather than common ancestry. Thus, matching of SPs/TMs creates the illusion of matching hydrophobic cores. Therefore, inclusion of SPs/TMs into domain models can give rise to wrong annotations. More than 1001 domains among the 10,340 models of Pfam release 23 and 18 domains of SMART version 6 (out of 809) contain SP/TM regions. As expected, fragment-mode HMM searches generate promiscuous hits limited to solely the SP/TM part among clearly unrelated proteins. More worryingly, we show explicit examples that the scores of clearly false-positive hits, even in global-mode searches, can be elevated into the significance range just by matching the hydrophobic runs. In the PIR iProClass database v3.74 using conservative criteria, we find that at least between 2.1% and 13.6% of its annotated Pfam hits appear unjustified for a set of validated domain models. Thus, false-positive domain hits enforced by SP/TM regions can lead to dramatic annotation errors where the hit has nothing in common with the problematic domain model except the SP/TM region itself. We suggest a workflow of flagging problematic hits arising from SP/TM-containing models for critical reconsideration by annotation users.

Citation: Wong W-C, Maurer-Stroh S, Eisenhaber F (2010) 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. doi:10.1371/journal.pcbi.1000867

Editor: Philip E. Bourne, University of California San Diego, United States of America

Received March 11, 2010; **Accepted** June 25, 2010; **Published** July 29, 2010

Funding: The authors thank A*STAR for financial support. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: wongwc@bii-sg.org (WCW); sebastianms@bii-sg.org (SMS); franke@bii-sg.org (FE)

Introduction

Following the request of a collaborator to hypothesize about the function of Eco1, an uncharacterized yeast gene at that time, the application of the full battery of sequence-based prediction tools [1,2] revealed an apparently significant hit to the Pfam domain PF00583 [3] in the local search mode Figure S1). This finding helped to identify a potential acetyl-CoA binding site and, subsequently, the hypothesis of Eco1's acetyltransferase function was proven experimentally [4]. At about the same time, another collaborator inquired about the function of the protein "Alt a 1" of the fungus *Alternaria alternata* (AAB40400). The same approach revealed an apparently significant hit to the Pfam domain PF00497 (Figure S2) indicating some relationship to bacterial extracellular solute-binding proteins. The initial hope of having found at least something to follow up faded away quickly when it became clear that the query has just a signal peptide (SP) in common with the proteins belonging to domain PF00497. This SP has artificially elevated the alignment score into the significance range and, thus, created the impression of functional relatedness. Why do the domain models perform so differently?

The theory of biomolecular sequence homology and its practical application for predicting function for uncharacterized genes by annotation transfer from well-studied homologues is one of the few achievements of theoretical biology that have significance for all fields of life science [5,6]. Similarity of amino acid sequences implies, to a certain degree, similarity in 3D structure and biological function [7–9]. Even apparently unrelated sequences with essentially zero sequence identity can adopt the same structural fold. This fact is rationalized by the conservation of the seemingly random, intricate hydrophobic pattern in the amino acid sequence of globular proteins that is required to form the tightly packed hydrophobic core of the tertiary structure [10]. This level of statistically significant sequence similarity is thought to arise from common ancestry under the pressure of selection at each step of mutational divergence with only rare instances of convergent evolution [11,12]. The corresponding evolutionarily favored amino acid exchanges tend to maintain side chain

Author Summary

Sequence homology is a fundamental principle of biology. It implies common phylogenetic ancestry of genes and, subsequently, similarity of their protein products with regard to amino acid sequence, three-dimensional structure and molecular and cellular function. Originally an esoteric concept, homology with the proxy of sequence similarity is used to justify the transfer of functional annotation from well-studied protein examples to new sequences. Yet, functional annotation via sequence similarity seems to have hit a plateau in recent years since relentless annotation transfer led to error propagation across sequence databases; thus, leading experimental follow-up work astray. It must be emphasized that the trinity of sequence, 3D structural and functional similarity has only been proven for globular segments of proteins. For non-globular regions, similarity of sequence is not necessarily a result of divergent evolution from a common ancestor but the consequence of amino acid sequence bias. In our investigation, we found that protein domain databases contain many domain models with transmembrane regions and signal peptides, non-globular segments of proteins having hydrophobic bias. Many proteins have inherited completely wrong function assignments from these domain models. We fear that future function predictions will turn out futile if this issue is not immediately addressed.

hydrophobicity, charge and side chain volume. Not surprisingly, it is these exchanges that score highly in the BLOSUM62 matrix [13] used in the BLAST/PSI-BLAST suite [14,15].

This general theme has received two variations. The first is introduced by the notion of the protein domain [16–18] and the existence of multi-domain proteins. Structurally, domains are protein sequence segments that form their own 3D structure with its independent hydrophobic core (and with a generally more polar surface); thermodynamically, they fold and melt independently; from the evolutionary point of view, these sequence segment are shuffled in the genome as independent units and are re-used in different contexts [5]. With respect to the homology search, the notion of domains leads to segmentation of protein sequences where the segments represent homologous members of a sequence family with the same type of domain. The family collection can become laborious; thus, protein domain libraries have appeared as a collective effort of the scientific community. Among the collections, there are PROSITE [19], BLOCKS [20], PRINTS [21], SUPERFAMILY [22], CDD [23], TIGRFAM [24], Panther [25], ProDom [26], EVEREST [27], the libraries of Y. Wolf and L. Aravind published with IMPALA [28] and, as the most systematically developed primary collections, Pfam [3] and SMART (Simple Modular Architecture Research Tool [29]).

The second issue is that many segments do not have globular structures at all [30–32]. They can be of fibrillar nature, transmembrane (TM) helices, disordered regions, etc. Typically, these regions have a clear amino acid compositional bias or a primitive repetitive pattern. Sequence similarity between two sequence segments of this type does not necessarily mean common ancestry but is obviously an enforced result of physico-chemical constraints. For example, long hydrophobic stretches such as transmembrane helices appear similar regardless of ancestry and, as in the introductory example, all signal peptides [33] but also GPI lipid anchor sites [34,35] or coiled coil regions [36] must look alike to a certain degree. Many types of polar non-globular regions, for example serine-rich segments, readily compensate for insertions/deletions or substitutions as long as the integral properties of the respective subsegments remain unchanged. Consequently, convergent evolution might have a more significant role for non-globular sequences.

Thus, sequence similarity can either be due to homology (common ancestry) or convergent evolution (common selective pressure). The criterion of sequence similarity for inferring homology is actually applicable only to globular segments and non-globular parts should be excluded from starting sequences in homology searches. The special case with amino acid compositional bias was recognized early and it was always advised to exclude those segments from similarity searches when hunting after distantly related proteins. For the BLAST/PSI-BLAST suite, the SEG program was advised to suppress at least the most obvious low complexity regions [14,37] besides the application of statistical corrections for compositional bias [37,38]. Sequence family searching heuristics should consider excluding also other types of non-globular segments such as coiled coil regions from the similarity search [39]. In the original concept of SMART [40], special care was paid to determine domain boundaries correctly, to include all secondary structural elements of globules, for example by matching the alignment section with known 3D structures, and to exclude all sequence parts such as polar or proline-rich linker regions that do not belong to the domain considered.

The unsupervised inclusion of transmembrane helices and signal peptide segments in homology searches is especially prone to erroneous addition of unrelated sequences to the sequence family under study since the systematic coincidence of hydrophobic positions creates the appearance of similarity in the hydrophobic pattern, otherwise the key to sequence homology among globular sequence segments [10]. The consequently generated high similarity score as in the introductory example of "Alt a 1" might support an otherwise unjustified annotation transfer and lead to wrongly predicted function if it were not detected by manual checks.

Similar precautions are generally out of scope when protein domain model libraries are applied for function prediction over query sequences, especially in a genome-wide mode. It is desirable to have systematic factors that might cause spurious annotations such as isolated similarities to signal peptides or some types of transmembrane helices be suppressed during the annotation workflow.

When checking domain databases for the inclusion of transmembrane helices and signal peptides into the domain model, we found more than thousand domain instances in Pfam and a couple of examples even in SMART. These hidden Markov models (HMMs) can be a systematic cause of spurious similarity hits especially if the HMM-based sequence scan is applied in the local search mode. In this work, we wish to emphasize that these domain models can also give rise to wrong hits even in the global search mode where the high score from the membrane-helical part can mask the absence of match for the associated globular domains. For support of the reader, database search results, alignments, domain library entry lists and files with "cleanup" domain models as referred to in the following text are provided as supplementary material at the associated WWW site http://mendel.bii.a-star.edu.sg/SEQUENCES/ProblemDomains-TM+SP/.

Results

Search for transmembrane helices and signal peptides included in SMART database alignments and validation of findings

Since the SMART database [29] is relatively small and the alignments are very well curated, its alignments were used as a test

ground for a SP/TM detection algorithm as described in detail in the Methods section.

In brief, we recovered the full length protein sequences that contained the segments in a given alignment of SMART version 6, applied 5 TM and 2 SP predictors published in the literature and we checked overlap of predicted SP/TM regions with the alignment segments. For an alignment position to be considered part of a predicted TM or SP region, the respective residue must be included into the predicted range in a critical number of sequences and by a certain number of prediction tools determined by a statistical criterion based on the binomial distribution (significance value 0.05).

For each predicted TM or SP region, we derive a score as the arithmetic mean of the logarithmic probabilities of SP/TM prediction over all alignment columns involved (Methods, equation 5). The false-positive prediction rate was assessed using the SCOP α-helical proteins and the SCOP membrane class (Structural Classification of Proteins [41,42]) to determine TM- and SP-score cutoffs with false-positive rates below 5%.

In contrast to the Pfam test described below, SMART version 6 alignments contain pleasantly few SP/TM regions. With a TM-score cutoff of ≥ -12 (FP rate of 4.67%) and SP-score cutoff of ≥ -1 (FP rate of 4.02%), the number of predicted TM helices and signal peptides are 40 and 5 respectively. At the domain level, this translates to 13 problematic domains with TMs and 5 with SPs, respectively (Table 1). Thus, the fraction of problematic domains is very low with 1.6% (13/809) having TMs and 0.6% (5/809) SPs segments.

These 18 predictions were manually validated: (i) If the respective predicted segments were indeed structural helices and not SPs/TMs, they should be part of one of the nearest globular domains in the sequence. The alignment sequences were searched against the sequences with known 3D structure from the Protein Data Bank (PDB) for any significant hits (with the generous Blast E-value$\leq$0.1) and we checked whether the predicted SP/TM region overlaps with the segment covered by the structure. If the predicted SP/TM region was missing in the structure or if it was described as a TM helix in the structural report, we considered the

Table 1. Summary of predicted/validated non-globular segments and supporting evidence for the 18 SMART version 6 domains.

Domain name	Type	Predicted segments	Validated Segments	Comments
SM00019 : SF_P (Pulmonary surfactant protein)	TM	33–58	1–58#	The N-terminal propeptide 1–58 of NP_003009 forms a TM when induced by a Brichos domain [99].
SM00157 : PRP (Major prion protein)	TM	117–140	112–135#	Latent transmembrane region in human prion protein BAG32277 [100,101].
SM00665 : B561 (Cytochrome B561/ferric reductase TM domain)	TM	4–146	N/a	Intrinsic membrane protein [102].
SM00714 : LITAF (LPS-induced tumor necrosis factor α factor)	TM	38–61	N/a	The LITAF domain appears to have a membrane-inserted motif (although without transmembrane segment) [103].
SM00724 : TLC (TRAM, LAG1 and CLN8 homology domains)	TM	10–76; 216–238; 287–307	N/a	Proof for 8 membrane-spanning segments in Lag1p (NP_011860) and Lac1p (NP_012917) [104]
SM00730 : PSN (Presenilin, signal peptide peptidase, family)	TM	5–27; 113–134; 214–285; 600–649	4–25#; 115–133#; 214–231#; 241–257#; 260–283#; 602–621#; 628–644#	Out of 10 TM regions shown for human presenilin-1 (AAB46371), 9 are in the domain alignment out of which 7 are predicted here [105].
SM00752 : HTTM Horizontally transferred transmembrane domain	TM	12–25; 75–95; 275–294; 338–357	N/a	Domain is known to have 4 TM regions [80].
SM00756 : VKc (catalytic subunit of vitamin K epoxide reductase)	TM	12–30; 104–192	13–32#; 142–189#	VKORC1 (Q9BQB6) is a membrane protein [106].
SM00780 : PIG-X (Mammalian PIG-X and yeast PBN1)	TM	230–248	230–252#	PBN1 (CAA42392) is a type I transmembrane protein in the endoplasmic reticulum [107].
SM00786 : SHR3_chaperone (ER membrane protein SH3)	TM	7–111; 167–186	N/a	Shr3p (NP_010069) has 4 membrane segments [108].
SM00793 : AgrB (Accessory gene regulator B)	TM	42–204	N/a	*S. aureus* ABW06464 is a membrane protein [109].
SM00815 : AMA-1 (Apical membrane antigen 1)	TM	522–527	515–602#	Segment missing in structure 1W81_A [110].
SM00831 : Cation_ATPase_N (Cation transporter/ATPase, N-terminus)	TM	72–90	65–94#	Segment maps to a TM helix of the ß-domain of 1KJU_A [111].
SM00190 : IL4_13 (Interleukin 4/13)	SP	1–20	1–23#	Annotated as secreted. Segment missing in structure 1ITL_A [112].
SM00476 : DNaseIc (deoxyribonuclease I)	SP	1–19	1–17#	Annotated as secreted. Segment missing in structure 1DNK_A [113].
SM00770 : Zn_dep_PLPC (Zn-dependent phospholipase C, α toxin)	SP	4–26	1–64#	Annotated as secreted. Segment missing in structure 1OLP_A [114].
SM00792 : Agouti	SP	1–19	1–89#	Annotated as secreted. Segment missing in structure 1Y7J_A [115].
SM00817 : Amelin (Ameloblastin precursor)	SP	11–28	1–26#	Protein AAG27036 [116] is secreted to enamel matrix.

Both the predicted and, if explicitly available in the literature, the validated segments of TM regions or signal peptides are provided in the sequence count of the respective SMART domain alignment. In cases marked with "#", the sequence positions are with respect to the reference sequence given in the comments.
doi:10.1371/journal.pcbi.1000867.t001

prediction as validated. (ii) Without structural hits, we searched the scientific literature for topological information about membrane embedding of reference sequence segments.

As the information collated in Table 1 confirms, none of the 18 cases is a false-positive SP/TM prediction. Thus, we conclude that the SMART domain database contains at least 18 problematic domain models. It is of interest to note that, except for 4 cases with accessions below SM00600, all other problematic domains have been added to SMART only in recent years (Figure 1).

Detection of more than a thousand domains in the Pfam database with SP/TM regions

Given that our SP/TM detection procedure provides statistical error measures for the prediction, it can be reasonably applied on the body of Pfam domain models. When this work was started, the available Pfam version was release 23 constructed with the HMMER2 package. About 19% (1937 out of 10340) of Pfam-A domains in release 23 [3,43–45] do not have more than 4 seed sequences in the alignment and, consequently, there is not enough statistical power for rejecting the null hypotheses even if the predicted SP/TMs are true (see Methods). In Figure 2, we show the distributions of the TM- and SP-scores per predicted SP/TM region for the alignments of the remaining 8403 domains of Pfam-A. Both histograms exhibit a bimodal distribution where true-positives cluster at high scores and false-positive predictions aggregate at low scores (see Methods). If we apply the same SP/TM-score cutoffs as in the SMART exercise (-12 and -1 respectively), the number of predicted TM helices and signal peptides are 3849 and 164 respectively.

At the domain level, this implies 1079 (10.4%) and 164 (1.6%) out of 10340 Pfam-A domains having TM or SP regions included into the domain alignment (Figure 3). The extent of the non-globular part introduced by TM regions together with the polar linkers between them in the domain alignments of Pfam can be huge (more than 500 positions). Whereas SMART strived for excluding non-globular parts from the domain alignments and included a few critical domains only recently, this has not been a matter for Pfam at all (Figure 1). The accumulation of problematic domains was even over all the history of Pfam. Interestingly, our conservative estimate of 10.4% (1079/10340) for TM-containing Pfam-A domains measures well against the estimated 16.7% (1365/8183) for Pfam-A release 19 reported by Bernsel *et al.* [46] who just applied TMHMM. It should be noted that their result is from a plain application of TMHMM without any additional false-positive hit suppression.

Among our 164 domains with SP predictions, we might expect 6.6 ($\sim$7) wrong predictions. On average, each domain with predicted TM regions contains about 3.6 (3849/1079) TM helices, out of which 0.17 (4.67% of 3.6) represent false-positive TM helices. We might expect that about 50 domains out of the 1079 domains are wrongly included into this list. Even if we remove those values from the total number of 1214 problematic domain models (1050 TM, 135 SP and 29 concurrent TM and SP errors), Pfam-A release 23 still contains more than 1001 critical cases as claimed in the title of this article.

Inclusion of non-globular sequences leads to false-positives in homology searches, a serious source of errors in protein function annotation

The domain alignments in Pfam and SMART are used for the derivation of hidden Markov models (HMMs) that, in turn, are applied for searching matches in query sequences with programs of the HMMER packages [47–49] with HMMER2 being the currently validated version. It should be noted that both the local and the global search modes for domain hits are available.

With SP/TM regions as part of the domain alignment, the respective HMMs are no longer useful for local mode searches since a match in the TM or SP region alone without any other sequence similarity to the query sequence can be sufficient to cause a false-positive fragmentary domain hit as in the introductory case of "Alt a 1". Further illustrative examples are provided in Table S1 and Figure 4. We especially searched for sequence examples having both hits with a SP/TM region containing domain model (with an alignment restricted to the SP/TM region only) as well as

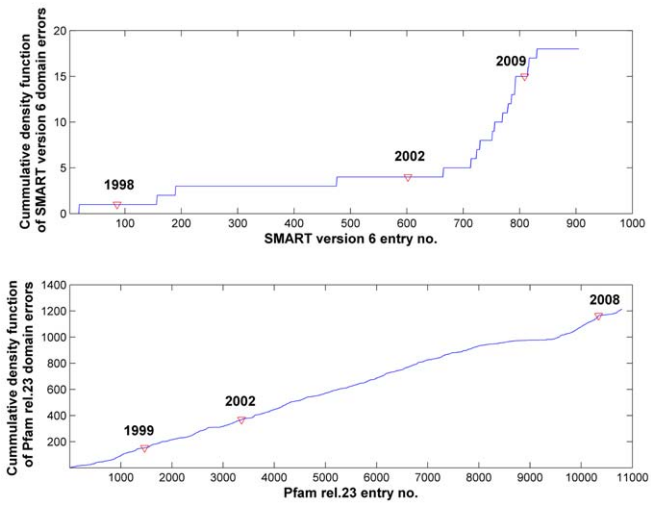

Figure 1. Cumulative plots of SMART version 6 and Pfam release 23 problematic domains. In SMART version 6, the total number of domains with predicted SP/TM segments peaks at 18, which made up 2.2% of 809 SMART domains (see top). Red triangles mark time points for the years 1998, 2002 and 2009 when the total number of domain models was 86, 600 and 809 respectively. In Pfam, the total number of problematic domains peaks at 1214, which made up 11.8% of 10340 Pfam domains (see bottom). Likewise, red triangles marked the years 1999, 2002 and 2008 with 1465, 3360 and 10340 Pfam entries respectively.
doi:10.1371/journal.pcbi.1000867.g001

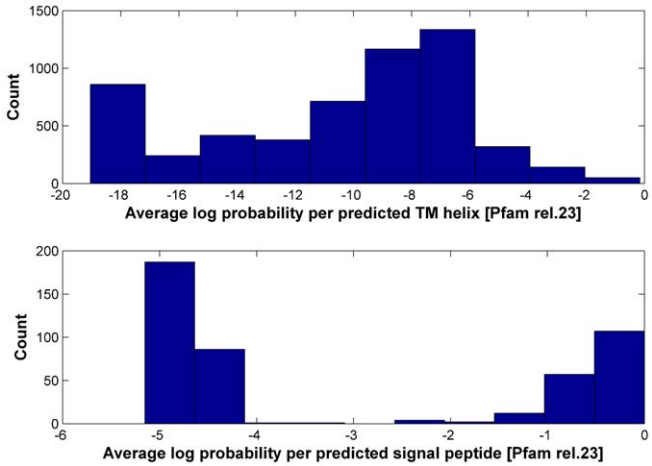

Figure 2. Histograms of average log probability per predicted transmembrane helix and per predicted signal peptide in Pfam release 23. The top part shows the histogram of average log probability per predicted transmembrane helix; the bottom part shows the same per predicted signal peptide. The log probability provided on the x-axis is calculated with equations 5 and 6. At the *TMcutoff* of ≥ -12 (false-positive rate 4.67%) and *SPcutoff* of ≥ -1 (false-positive rate 4.02%), the number of predicted TM helices and signal peptides are 3849 and 164 respectively.
doi:10.1371/journal.pcbi.1000867.g002

multiple other prediction tool hits that provide intrinsic annotation contradictions. Thus, we have two arguments supporting the idea that the SP/TM region containing hit is false-positive.

One of the referees brought up the argument that some of the sequences in Table S1 (and also in the subsequent Table S2) have become obsolete. In the revised Tables S1 and S2, we show that none of the sequence examples have disappeared; instead, the sequence entries have been updated and, in none of the cases of sequence edition, the computation results have been changed to the extent of compromising the conclusion. It needs to be emphasized that sequence-based prediction tools should be applicable to all types of sequences including naturally occurring ones, mutated versions, synthetic constructs as well as all types of hypothetical sequences. It is this ability of protein sequence analysis that makes it so powerful to conclude from genome sequences. For example, it should be noted that, sometimes, the absence of a domain hit is taken as indication of a sequence representing a non-coding RNA.

The model Herpes_glycop_D (PF01537.9) has a membrane-helix region that, together with its linkers on both side, are the sole part of a match in the fragmented search mode for a large variety of taxonomically and functionally diverse proteins out of which eight architectures are presented here. Similarly, the TM region (plus surrounding polar linkers) of model CDC50 (PF03381.7) significantly hits proteins with at least three different architectures in the fragmented HMM search.

For another 4 domain models Cation_ATPase_N (PF00690.1), GSPII_F (PF00482.11), PAP2 (PF01569.13) and HCV_NS4b (PF01001.11) provided as further illustration examples, the respective TM region hit a single TM helix segment of several seemingly unrelated proteins. In all cases, their alignment scores were above their family-wise gathering score thresholds.

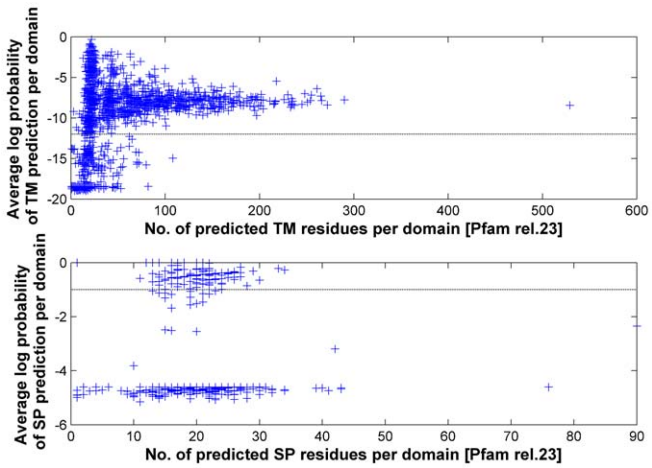

Figure 3. Average log probability plot of transmembrane helix and signal peptide predictions per domain. The top part shows the average log probability per predicted transmembrane helix calculated per domain; the bottom part shows the same per predicted signal peptide. Whereas the y-axis shows the log probability in accordance with equation 6 applied over all predicted segments for a given domain, the x-axis represents their cumulative length. At the *TMcutoff* of ≥ -12 and *SPcutoff* of ≥ -1 (horizontal dashed lines), the number of problematic TM and SP domains are 1079 and 164 respectively. The total number of problematic domains is 1214 (1050 TM, 135 SP and 29 concurrent TM and SP).
doi:10.1371/journal.pcbi.1000867.g003

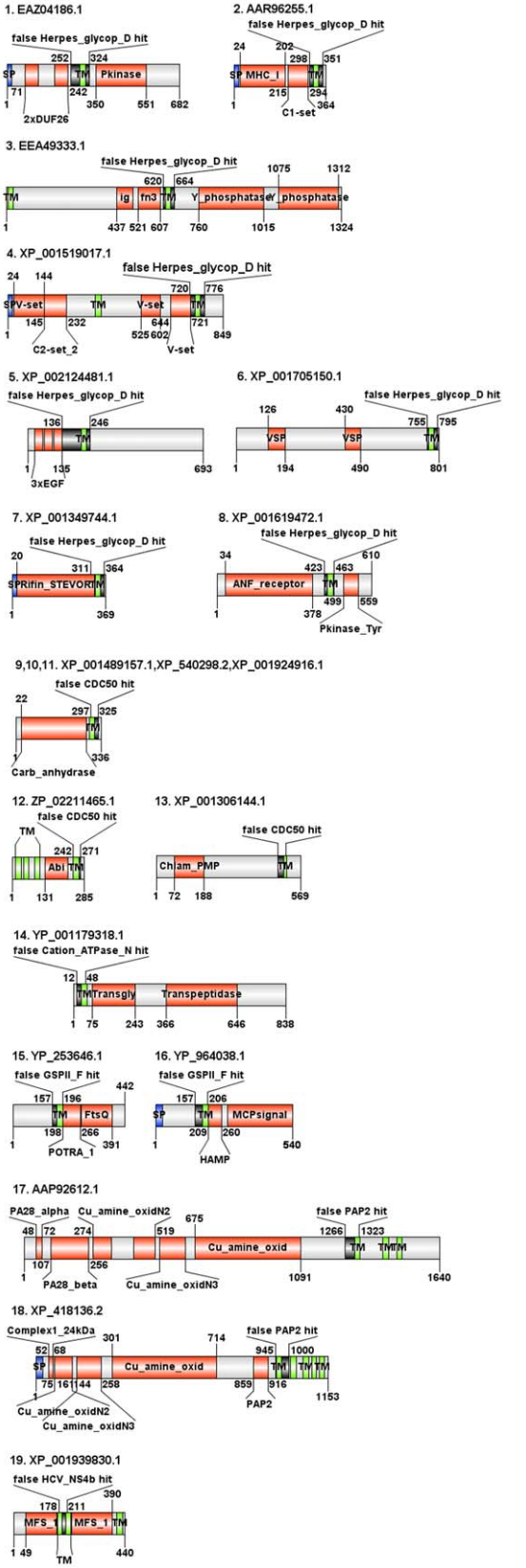

Figure 4. Examples of domain architectures of false-positive HMM hits caused by TM helices in the fragment-mode search. We show illustrative examples for six Pfam release 23 models: Herpes_glycop_D (PF01537.9), CDC50 (PF03381.7), Cation_ATPase_N (PF00690.18), GSPII_F (PF00482.11), PAP2 (PF01569.13) and HCV_NS4b (PF01001.11). The black boxes denote the problematic domain annotations in the respective sequences. Additional material such as hmmpfam outputs and alignments are available at the associated BII WWW site for this work. Domain architecture illustrations were created with DOG 1.5 [98].
doi:10.1371/journal.pcbi.1000867.g004

Not surprisingly, the global search mode that forces a complete match of the domain model over a subsegment of the query sequence is the standard regime for running hmmsearch and hmmpfam of the HMMER2 package. Typically, a positive hit is recognized either by a score above a so-called gathering threshold (which is supplied together with and determined empirically by the creator of the Pfam domain model) or an E-value below a trusted limit (such as 0.1, see page 23 of the HMMER2 user guide). It is particularly worrying that a number of domain models with SP/TM regions included generate quite convincing E-values for unrelated sequences even in the global search mode. In all these cases, matches of a hydrophobic region in the query with the hydrophobic segments of these validated SP/TM regions is the reason for the elevated score that frequently surpasses even the gathering score threshold.

To investigate the effects of SP/TM regions in homology searches, two separate HMM searches against the NR database were performed for each domain under study. The first run relied on an HMM using the original alignment. For the second run, we constructed a ''cleanup'' alignment via the removal of the predicted TM or SP segments. The two HMMs for the hmmls style of search (global with respect to the domain and local to the query sequence) were built from the alignments using the commands 'hmmbuild –F –amino model-file alignment-file' and 'hmmcalibrate –seed 0 –num 5000'. When contrasting the results of the two HMM runs at E-value$\leq$0.1, we assume all hits of the cleanup model as true-positives and scrutinize all additional hits of the original model as potential false-positive hits. We screened them for potentially contradictory annotation using sequence-analytic tools [1,2] and scientific literature. Below, we describe several representative cases (Table S2, Figure 5).

The model PIG-P (PF08510.4) includes a segment with TM helices (positions 1–91) and hydrophilic region (positions 92–208). In the global-mode search against the non-redundant database, the first 100 alignment positions of the model (i.e., the N-terminal part with the 2 TM helices) hit a pair of C-terminal TM helices in the four protein targets listed in Table S2 (Figure 5). The positions of the HMM covering the cytoplasmic part of Pig-P [50] correspond mostly to a single large gap in the alignment with any of the four hit sequences (and this gap has only a marginal influence on the total score). The E-values both with the HMMER2 and HMMER3 suites are very convincing (between e-27 and e-09) and the scores are all far above the gathering threshold. Nevertheless, these are certainly false-positive hits. Whereas, the Pig-Ps are endoplasmic reticulum proteins [50], EAY79580.1, EAZ17037.1 and XP_001842924.1 have nucleic acid binding domains and are most likely nuclear proteins and XP_761344.1 appears mitochondrial due to a CIA30 domain [51]. Just having two TMs and their short linker matching is a poor argument for common ancestry.

The PAP2 (type 2 phosphatidic acid phosphatase) domain model (PF 01569.13) hits the sequence XP_418136.2 (Table S2, Figure 5) in an internal segment. Inspection of the alignment shows that the only high scoring similarity regions belong to the two transmembrane segments and there are two large gaps corresponding to non-membrane segments in PAP2 proteins. Most

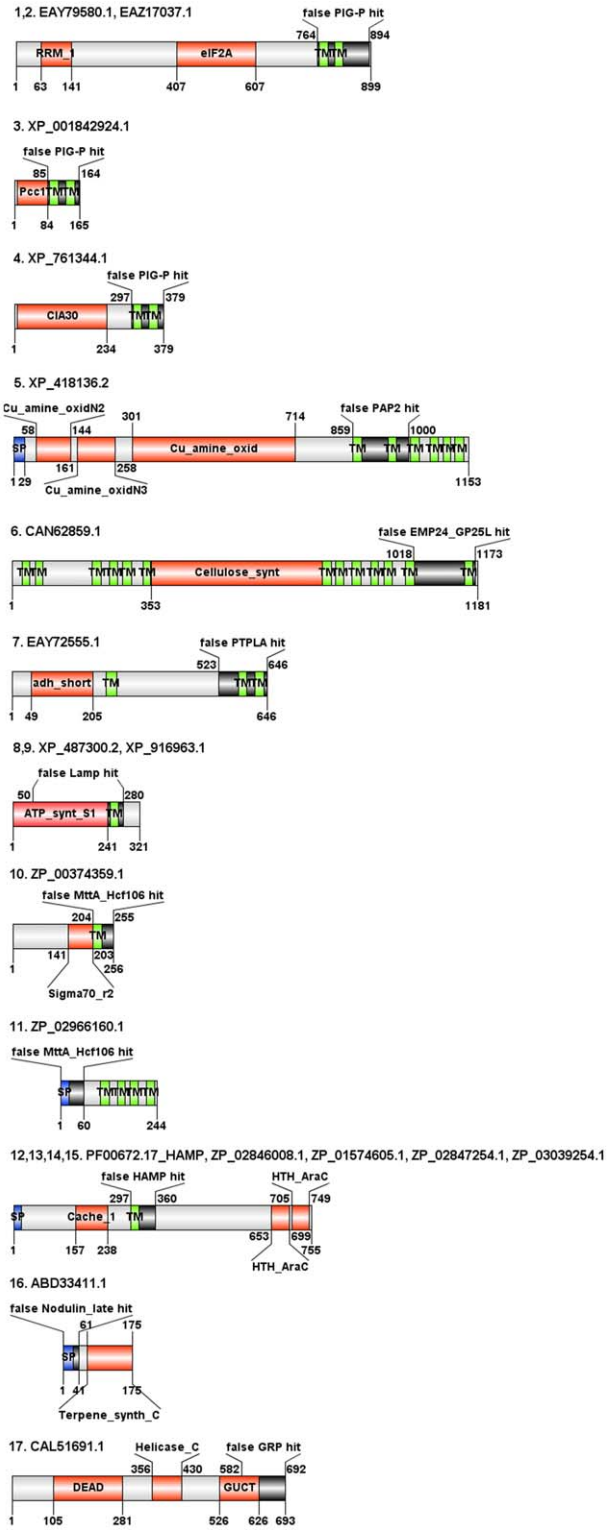

Figure 5. Examples of domain architectures of false-positive HMM hits caused by TM helices/signal peptdes in the global-mode search. Findings for nine Pfam release 23 models Pig-P (PF08510.4), PAP2(PF 01569.13), EMP24_GP25L (PF01105.15), PTPLA (PF04387.6), Lamp (PF01299.9), MttA_Hcf106 (PF02416.8), HAMP (PF00672.17), Nodulin_late (PF07127.3) and GRP (PF07172.3) are shown. The black boxes denote the problematic domain annotations in the respective sequences. Additional material such as hmmpfam outputs and alignments are available at the associated BII WWW site for this work. Domain architecture illustrations were created with DOG 1.5 [98].
doi:10.1371/journal.pcbi.1000867.g005

importantly, the two motifs A and C characteristic for PAP2 proteins are not conserved and the motif B is completely absent in the sequence hit [52]. Thus, this is a false-positive finding regardless of impressive scores and E-values.

The members of the EMP24_GP25L family (PF01105.15, Table S2) have a polar region, a coiled coil segment followed by a transmembrane part in their model [53]. Sequence CAN62859.1 (Figure 5) generates a significant, yet false-positive hit to the respective HMM although it does not have any traces of a heptade repeat in the sequence.

In the model for PTPLA (PF04387.6, Table S2), the first 30 N-terminal positions of the HMM contain the active site motif that is critical for function and, thus, for family membership [54,55]. The alignment of EAY72555.1 (Figure 5) with the respective HMM has a large gap in this region; nevertheless, the matches with two transmembrane conveniently shift the E-value into the region of statistical significance although, this time, the score is below the gathering threshold.

The Lamp domain (PF01299.9, Table S2) characteristic of lysosomal glycoproteins hits sequences XP_487300.2 and XP_916963.1 of ATP synthases (Figure 5) significantly both with regard to score and E-value. Inspection of the alignments shows that a segment of about 120 HMM positions out of 340 is absent in the sequence hits since the respective region is covered by three large gaps. As a result, several critical functional motifs (cysteines 1–5 and the cytoplasmic tail GY motif [56]) are missing in the hits. The total score is rescued by the transmembrane region match.

The typical architecture of MttA_Hcf106 (PF02416.8, Table S2) proteins (known to be involved in sec-independent translocation [57]) comprises of a TM segment followed by an amphipathic helix and an acidic domain. The alignment of the respective HMM with the false-positive hit sequences ZP_00374359.1, an RNA polymerase, and ZP_02966160.1 (Figure 5), a putative phosphatase, shows good match in the TM segment (in the case of ZP_02966160.1, with its signal peptide !!) followed by a moderate fitting to the amphipathic helix segment and an almost complete absence of the acidic part.

HAMP protein segments comprise of two α-helices connected with a linker having a characteristic motif [58]. In addition, the domain model PF00672.17 (Table S2) includes a preceding transmembrane segment which causes significant, yet false-positive global search HMM hits in four proteins (see Table S2, Figure 5) although none of them has traces of the linker region (covered by a gap in the alignment).

The architecture of the Nodulin_late domain (PF07127.3, Table S2) consists of a signal peptide followed by a region with two characteristic cysteine pairs [59]. The protein ABD33411.1 is annotated as a nodulin_late protein in the database and, indeed, the respective HMM produces a significant hit by any commonly used statistical criterion; yet, the hit is false-positive (Figure 5) since the alignment is good only in the signal peptide region but this match is followed by two large gaps and none of the cysteine pairs is conserved.

Further, the domain model GRP (PF07172.3, Table S2) for cell-wall related proteins comprises of an N-terminal signal peptide followed by a glycine-rich region [60]. The respective HMM matches the C-terminal part of CAL51691.1, a putative RNA helicase (Figure 5). Surprisingly, the signal peptide part of the GRP domain matches the C-terminal two secondary structural elements of the GUCT domain (PF08152.4) in CAL51691.1 (a β-strand and an α-helix in the homologous structure 2E29 chain A [61]).

Our final example illustrates the issue with multiple TM segments. If the linkers between them differ among query and model, the gap penalties offset some part of the score accumulated

by the hydrophobic position matches. The case of claudin proteins, small membrane glycoproteins with 4 TM helices and a length below 200 AA, is instructive in this respect. In a global search mode with the PMP22_claudin model (PF00822.12), the respective HMM hits numerous sequences of γ-subunits of voltage-dependent Ca-ion channels with E-values in the order of e-7. Closer inspection of the seed alignment showed that just a single channel sequence (CCG2_mouse) was included although they are not related to the family [62]. If we remove this entry, the new HMM still hits to 4TM γ-subunits of voltage-dependent Ca-ion channels (e.g., NP_542375.1, NP_542424.1) as well as to the 3TM XP_533601.2 (Natural killer cell protein) with E-values in the order of 0.08. In all cases, the sequence similarity is confined to matches of the hydrophobic segments.

Inclusion of non-globular sequences leads to false-negatives in homology searches thus decreases sensitivity of the domain model

The decrease in specificity of domain models harboring SP/TM regions is also accompanied by a decrease in sensitivity. In general, the need to have additional good alignment scores for the SP/TM pieces can become a burden for any true-positive sequences that are incompletely sequenced or missing the SP/TM-region pieces naturally.

By contrasting the HMM runs between the original and cleanup models, potential false-negatives were identified as hits that were found only by the cleanup models. Then (see Methods), we re-computed the scores/E-values for the original HMM as well as another set of scores/E-values using the same HMM and EVD parameters from the original model but without the SP/TM segments (cleanup case). Finally, the two sets of scores/E-values were compared to find hits where their original score/E-values were less significant than their re-computed ones (i.e. without SP/TM). These were considered as false-negatives.

In Table S3, we show selected false-negative examples of several domain models with validated SP/TM-regions where their re-computed scores/E-values drastically improved without their SP/TM segment scores. All re-computed hits' scores except for NP_848488.2 were clearly above their gathering score thresholds. Previously, all these hits would be treated as false-positives if gathering score thresholds were considered. In essence, the negative scores of the SP/TM segments (due to their absence in the corresponding sequence) had acted as heavy penalties on the total scores, thus, it was concluded that these hits were insignificant.

Significant rates of problematic function annotations in existing sequence databases due to SP/TM regions in domain models

It was already suggested in the literature that unsupervised annotation transfer based on spurious sequence similarities has created a myriad of false function annotations for sequences from genome projects [63–65]. If care is not exercised, the inclusion of SP/TM regions into domain models can become a perfect recipe for protein annotation disaster.

We explored this issue for PIR (Protein Identification Resource) iProClass v3.74 [66] and retrieved sequences with Pfam accession IDs for the problematic domains in Table 2. These sequences were re-annotated using HMMER2 hmmpfam in global-search mode (with parameter –null2). Interestingly, a number of sequences returned zero hmmpfam hits (searched for with a very permissive E-value $\leq$10) despite being annotated with the respective domains in the database and these are clearly false annotations (column 5). For each sequence with reproduced hit, summing up the match, insert and state transition log-odd scores (provided in the Pfam model) over its emitted HMM sequence allowed us to recalculate its total score as well as the SP/TM-region- (column 2) and non-SP/TM-segment-specific parts of the score log odd scores. We tagged a sequence as a potential false-positive hit if the total score was at least the gathering score threshold GA while its non-SP/TM-segment-specific score contribution was less than the expected non-SP/TM specific gathering score threshold $\overline{G}_{nonSPTM}$ (column 4, see Methods); thus, only the match to the SP/TM hydrophobic region carries the hit over the threshold. Surprisingly, the number of unjustified annotations is between 2.1 to 13.6% depending on the type of domain (column 6); thus, the annotation error due to spurious SP/TM matches can be quite substantial.

Sequence complexity of SP/TM-regions

The fact that signal peptide or transmembrane helix segments are of lower sequence complexity than their globular counterparts is not widespread general knowledge. To our current understanding, there is only a comment about this issue in the BAliBASE article of Bahr *et al.* [67] where the notion is considered "self-evident" without provision of any supporting data.

In brief, we extracted all sequences from Uni-Prot (release 14.4) with the feature keys "signal" and "transmem". Among the single-transmembrane proteins, we selected those characterized as "anchor" in a special group. For multi-TM region proteins, we selected those who have 5–9 annotated TM segments. Additionally, we got the experimentally verified α-helical TM regions as provided by TMPDB (release 6.3) [68]. As a reference point for helices in globular proteins, we took the set of alpha-helices in PDB (extracted from PDBFIND2.txt as of April 2010 [69]) with 14–28 amino acid residue length surrounded by coil regions. Within all sets, sequence redundancy was suppressed with Cd-hit and a 50% sequence identity threshold [70].

In our calculations, we find that only 3% of residues in α-helices in globular domains are covered by hits of the quite stringent low complexity tool SEG (parameters window 12, 2.2, 2.5) [71] whereas this is the case for 18% for all residues in transmembrane helices extracted from TMPDB. Similarly, 24% of residues in signal peptides in UniProt are hit by the same SEG tool. Thus, SP and TM regions are more likely to be of low complexity than structural helices of comparable length.

Interestingly, the values for the Uni-Prot sets are 30% for single transmembrane proteins, 33% for single transmembrane proteins with the region annotated as "anchor" but only 12% for multi-transmembrane proteins. Thus, the problems with non-relevant matches in hydrophobic regions are more likely to occur, as a trend, in proteins having signal peptides or only a few transmembrane segments compared with cases of multi-membrane-spanning proteins.

Discussion

The notion of domain and the issue of SP/TM regions

There is no substitute for computational methods in large-scale functional annotation of sequence data and sequence similarity as surrogate for homology has to remain a decisive factor for function assignment [72]. E-value guided extrapolation of protein domain annotation has been a cornerstone for understanding completely sequenced genomes. There is about a decade of experience of using HMMER2 with a Pfam release 23-style or SMART domain library. These tools have indeed had tremendously high impact and have done a very good job.

Table 2. Unjustified annotation percentage of validated problematic domains in protein information resource (PIR) iproclass v3.74 (Global-mode search).

Domain Name	Type, validated region of model (size)	No. of retrieved sequences	No. of FP hits where $v \geq GA$, $v_{nonSPTM} < \bar{G}_{nonSPTM}$	No. of annotations without hmmpfam hits (E>10)	Total No. of unjustified hits (%)
PF00690.18 : Cation_ATPase_N (Cation transporter/ATPase, N-terminus), $GA = 18.90$, $\bar{G}_{SPTM} = 9.58$, $\bar{G}_{nonSPTM} = 18.79$, $c = -9.47$, $\bar{A} = -76.19$	TM,66–87 (87), ref.[111]	3684	74	3	77 (2.1%)
PF01105.15 : EMP24_GP25L (Endoplasmic reticulum and golgi apparatus trafficking proteins), $GA = -16.00$, $\bar{G}_{SPTM} = 13.82$, $\bar{G}_{nonSPTM} = -20.28$, $c = -9.54$, $\bar{A} = -208.58$	TM,141–167 (167), ref. [53]	1029	8	33	41 (4.0%)
PF01299.9 : Lamp (Lysosome-associated membrane glycoprotein), $GA = -87$, $\bar{G}_{SPTM} = 18.34$, $\bar{G}_{nonSPTM} = -95.80$, $c = -9.54$, $\bar{A} = -614.95$	TM,304–340 (340), ref. [56]	164	2	12	14 (8.5%)
PF01544.10 : CorA (CorA-like Mg2+ transporter protein) $GA = -61.3$, $\bar{G}_{SPTM} = 28.57$, $\bar{G}_{nonSPTM} = -80.17$, $c = -9.70$, $\bar{A} = -503.57$	TM,341–407 (407), ref. [117]	2717	15	71	86 (3.2%)
PF01569.13 : PAP2 (type 2 phosphatidic acid phosphatase) $GA = 8.3$, $\bar{G}_{SPTM} = 21.70$, $\bar{G}_{nonSPTM} = -3.92$, $c = -9.47$, $\bar{A} = -120.86$	TM,102–177 (177), ref. [52]	5231	108	19	127 (2.4%)
PF02416.8 : MttA_Hcf106 (sec-independent translocation mechanism protein) $GA = 7$, $\bar{G}_{SPTM} = 17.88$, $\bar{G}_{nonSPTM} = -1.30$, $c = -9.58$, $\bar{A} = -102.29$	TM,1–19 (74), refs. [57,118]	2085	283	0	283 (13.6%)
PF04387.6 : PTPLA (protein tyrosine phosphatase-like protein), $GA = 25$, $\bar{G}_{SPTM} = 13.59$, $\bar{G}_{nonSPTM} = 20.97$, $c = -9.56$, $\bar{A} = -291.27$	TM,89–168 (168), refs. [54,55]	277	3	3	6 (2.2%)
PF04612.4 : Gsp_M (General secretion pathway, M protein) $GA = 25$, $\bar{G}_{SPTM} = 24.68$, $\bar{G}_{nonSPTM} = 10.16$, $c = -9.85$, $\bar{A} = -247.83$	TM,1–40 (165), ref. [119]	401	19	6	25 (6.2%)
PF07127.3 : GRP (plant glycine rich proteins) $GA = 17.2$, $\bar{G}_{SPTM} = 14.64$, $\bar{G}_{nonSPTM} = 12.16$, $c = -9.59$, $\bar{A} = -173.44$	SP,1–49 (134), ref. [60]	207	12	4	16 (7.7%)
PF08294.3 : TIM21 (Mitochondrial import protein), $GA = -20.3$, $\bar{G}_{SPTM} = 0.19$, $\bar{G}_{nonSPTM} = -10.88$, $c = -9.61$, $\bar{A} = -309.20$	TM,1–36 (157), ref. [120]	118	7	1	8 (6.8%)
PF08510.4 : PIG-P (phosphatidylinositol N-acetyl-glucosaminyl transferase subunit P), $GA = -11.4$, $\bar{G}_{SPTM} = 40.20$, $\bar{G}_{nonSPTM} = -42.07$, $c = -9.53$, $\bar{A} = -233.36$	TM,1–67 (153), ref. [50]	143	4	0	4 (2.8%)

In the first column, we list selected Pfam domains with their accession, identifier, description and their gathering score (as in Pfam release 23) that have TM and/or SP regions included into the model. The region in the domain alignment that includes the validated SP/TM segments (together with interlinking loops as described in Methods) and the corresponding references are provided in the second column. The number of retrieved sequences from iProClass v3.74 with respect to each domain is given in the third column. The number of unjustified hits that returns results (and also satisfied the criteria) and without results are given in the next two columns. The last column gives the total and percentage of the unjustified hits with respect to the number of retrieved sequences. In addition, the log odd scores were re-derived from the match/insert/state transition scores provided by the respective HMM model. The reproduced scores v varied from the original scores at 0.57 ± 0.34. GA and $\bar{G}_{nonSPTM}$ (see equations 19 and 20) denote the domain gathering score threshold and the expected non-SP/TM-specific gathering score threshold respectively. Additional material such as hmmpfam outputs and alignments are available at the associated BII WWW site for this work.

doi:10.1371/journal.pcbi.1000867.t002

The fundamental consideration in this article, namely the difficulty to interpret sequence similarity as a result of similarity of non-globular segments, (especially signal peptides or transmembrane regions) within the current theory of sequence homology, the basis of annotation transfer, goes beyond the specific criticism for a few domain models. In this context, it appears necessary to recall what the notion of a protein domain implies. In the introduction of their article, Veretnik *et.al.* [18] provide a list of definitions extracted from the literature and applicable in a variety of research contexts. The criteria involve sequence or 3D structure similarity, structural compactness, assignment and atomicity of associated biological function; yet, not any conserved piece of sequence can be considered a domain.

In the special case of globular domains that have tertiary structure, sequence similarities imply sequence homology as well as fold and function similarity. If 3D structures are known, domains as compact (having an own hydrophobic core) and spatially distinct units of protein structures that share significant structural similarity can be grouped together (for example, in libraries such as SCOP [41,42] or CATH [73]). Structural domains are also units for folding and, in the thermodynamic sense, for melting [16]. It should be noted that, even for globular domains, sequence similarity does not guarantee the same structure and function, especially with sequence identities below 25% [7,8,74]. Whereas fold similarity is usually a consequence of hydrophobic pattern similarity, nevertheless, lots of the structural detail can be different affecting issues of conformational flexibility, binding specificity, catalytic activity, substrate preferences and, thus, biological function [1,5].

Although structure-based domain libraries aim at providing complete and well-defined annotation about a domain, the antecedent of requiring structural information and associated function makes it exclusive for only a small number of well-studied proteins. Thus, many more proteins in sequence databases remain difficult to characterize under this definition.

Meanwhile, a complementary domain definition based on the sequence homology also evolved independently. In the sequence-analytic context, domains as the basic components of proteins are families of sequence segments of minimal length (i) that are similar to each other with statistical significance, (ii) that provide for a specific biological function at the molecular level ("atom" of molecular function [5]) and (iii) that occur in different sequence domain contexts as they are reshuffled by evolution [75–77]. Indeed, this notion is the basic to the approach of sequence homology-based domain libraries like SMART and Pfam. Yet, there is a caveat: Because of the statistical significance criterion, similarity between sequences to be established requires them to be without any type of amino acid compositional bias or primitive repetitive pattern. This condition essentially brings together the structural and the sequence-analytic definition of domain since both, essentially, become applicable only to the globular domain type. The exclusion of sequential bias makes the application of the sequence homology theory to non-globular sequence segments (in contrast to globular segments) at least a borderline case and, often (certainly at low sequence identity), disables sequence similarity as argument for common ancestry, similarity of structure (if there is any 3D structure at all) and function.

It is crucial to note that similarity of sequences can either be due to homology (common ancestry) or convergent evolution (common selective pressure due to physical requirements or biological function). We wish to emphasize that generally applied sequence-statistical criteria for deducing homology have been derived from studies of globular domains. In these cases, conservation of an intricate, only apparently random hydrophobic pattern is necessary for composing the hydrophobic core and, thus, for fold conservation [1,10].

This condition is generally not fulfilled for non-globular segments (e.g., transmembrane helices, signal peptides, inter-domain linker regions, segments carrying lipid-attachment sites, etc.); thus, their functional annotation requires other methods than just annotation transfer based on position-wise sequence similarity. It appears likely that many types of non-globular segments re-occurring in evolutionary very distant proteins are rather the result of convergent evolution than common ancestry; for example, the likelihood of a *de novo* appearance of a phosphorylation site in a generally serine-rich stretch seems quite high in evolutionary time scales. This issue would deserve a more explicit study on its own.

In a generalized theme, SP/TM segments are usually the results of physico-chemical constraints and do not confer the specific biological function of the protein. Therefore, missing alignments in the SP/TM regions is less detrimental than that of the non-SP/TM regions if the membrane-embedded region is just used as translocation signal.

About the suitability of HMM-type models to infer homology from SP/TM-region containing sequences

To further the argument, in the framework of HMM, there is no clear demarcation of SP/TM and non-SP/TM regions towards the computation of the alignment scores. Hence, this questions the correctness of inclusion of SP/TM regions into the HMM or, at least, makes a separate consideration for them a matter of necessity in the context of the homology argument.

Our arguments raise the question whether position-specific scoring matrices (PSSM), HMMs or profiles are indeed the appropriate tool to classify all kinds of non-globular segments with regard to sequence homology. Matching the hydrophobic pattern alone is recognized insufficient for inferring homology among proteins with transmembrane helices. In previous reports [46,78], sequence similarity was attempted to be complemented with topology requirements. Anantharaman and Aravind [79] in their discussion with the reviewer list further arguments such as conservation of functional residue patterns, conservation of the number of TMs, the linker length, etc. Similar arguments are provided by Schultz [80]. If common ancestry is not a necessary requirement, PSSMs or HMMs are useful to test aspects of sequence similarity in context of physical pattern constraints (for example, as in the case of TMHMM [81] for the purpose of transmembrane helix prediction).

The case of SPs/TMs is of special importance since their hydrophobic stretches can create the false appearance of similarity to the respective hydrophobic core of the target template based on a hydrophobic pattern match. Alignments with many hydrophobic residues in the same columns generate high scores; thus, a SP/TM match can elevate an otherwise mediocre HMM score into the range of significance. The inclusion of a SP/TM into the domain model can compromise the selectivity of HMMs towards specific families and create hits not only to neighboring sequence families within the superfamily but also beyond. Whereas errors of the first kind might be considered not dramatic, we show with examples in Tables S1 and S2 that, most importantly, drastic cases of misannotation can happen.

Thus, the reliability in homology inference is greatly influenced by the amount of non-globular content in such domain library entries. We find that, even in the very well curated SMART domain collection (version 6), there are 18 domain models (out of 809) that include TMs or SPs. Based on our conservative approach, we find that clearly more than 1000 domains harbor SP/TM segments in Pfam release 23 (out of 10340 entries). To make matters worse, we observe a growing trend of addition of SP/TM region-containing domain models in Pfam and especially in SMART during the recent years (Figure 1).

In the Results section, we provide convincing examples that these domains have the potential to lead to annotation problems. They do not only cause promiscuous hits in fragment-mode HMM searches (Table S1). As we could see, the problems persist in the global-mode HMM searches by elevating the hits to significant levels beyond any normally applied E-value cutoffs or gathering score thresholds for a variety of SP/TM-region containing domain models (Table S2).

Therefore, our finding might suggest the mandatory removal of SPs/TMs from domain models. We do not recommend this at this stage. Such a strategy is not easy to implement due to several reasons. The required editing of domain libraries given their current status would be quite laborious and appears impractical in the short term. Then, there is also the issue with some multi-TM region protein domain models where there is little or no soluble globular component. Further, the biological significance of sequence similarity of proteins with TM regions and its relationship to homology has been studied only in a few cases [79,80,82,83].

Notably with regard to signal peptides, the Pfam team has conveyed to us the removal of signal peptides in most domain models for future releases (Alex Bateman, personal communication). Similarly, it appears reasonable to remove TM regions from models where they are not integral parts of the globular domain and, especially, where the domain occurs also outside the TM region context. An excellent match between SP/TM regions of non-relevant proteins is possible just because of their uniform hydrophobicity and this match will elevate scores in alignments. Often, this might be insufficient to overcome thresholds of significance but, as we see in our experience, it can happen and it happens systematically for some types of models. Most likely, the problems arise with domains having one or very few TM regions

which are the majority of cases in Pfam (366 with 1 TM helix, 170 with 2 TM helices, 127 with 3 TM helices, 416 with more than 4 TM helices as with our conservative estimates). As we have seen, the trend to low sequence complexity is especially strong for proteins segments representing a signal peptide or a single-TM anchor. Both the exclusion of signal peptides and of transmembrane helix anchors from domain models would remove the bulk (but not all) of the problems described in this article. Among all SP/TM regions, signal peptides, signal anchors and single TM regions have a trend to considerably more pronounced sequence complexity than TM regions in multi-TM proteins (see Results).

In addition, we propose two other possible workarounds: First, one might process each query sequence with tools recognizing non-globular segments including those for SP/TM regions and mask them with X-runs before comparing the query with domain libraries. Yet, this would not exclude cases such as SPs in HMMs hitting structural helices (see the GRP example CAL51691.1 from Table S2). Alternatively, we offer a supplementary, "cleanup" version of Pfam release 23 (see the file "Pfam_rel23_globalHMM_cleanup.rar" at the WWW site for this article). In cases of problematic domain models with SPs/TMs, hits of query sequences both with the original HMM as well as with the HMM derived from the reduced alignment without the respective SPs/TMs are to be compared. We suggest considering collinear hits of both models as benign whereas hits from only the original HMM should be flagged as problematic pending manual check by the user of the annotation. For this purpose, we supply versions of the domain model that are cleaned from transmembrane helical and signal peptide inclusions (see associated WWW site for this work).

Whereas this work explores the issue of SPs/TMs in domain models mainly based on an analysis of HMMER2 and Pfam release 23, both have concurrently been updated to HMMER3 and Pfam release 24 [84]. We wish to underline that this revision does not resolve the problems described in this paper. For 16 out of the 17 sequence examples provided in Table S2, using HMMER3 with Pfam release 24 produces the same false-positive hits. In the remaining case of CAL51691.1 and domain model GRP (PF07172), the alignment of the respective domain entry has not changed and the absence of hit appears due to an increased gathering score (17.2 for global-mode and 15.9 for fragment-mode HMMER2-search in Pfam release 23 in contrast with 22.7 for HMMER3 and Pfam release 24). We do not think that the transition to HMMER3 resolves the problem of SPs/TMs included into seed alignments since SPs/TMs will contribute to the score similarly to buried structural helices regardless of any composition-based corrections. On the contrary, we have seen that the fragment-mode search with HMMER2 has essentially been useless in the E-value guided mode because of many false hits; for the current HMMER3 beta-release, this is the only search mode available so far.

E-value guided domain search *versus* gathering threshold criteria

As a remedy, switching from the E-value guided hit finding to gathering score thresholds is proposed. This is problematic from several viewpoints. The HMM concept has the beauty of a rigorous probabilistic formulation that allows a natural treatment for substitutions and gaps in the same formalized framework. Further, the introduction of E-values provides a handle to compare various types of predictions that hit the same sequence region. Unfortunately, the gathering score concept (an expert-defined domain-specific score threshold for homologous hit selection) brings in an arbitrary component into the prediction process.

Firstly, the determination of a gathering score is not guided by a fundamental consideration but, instead, depends on the data and literature situation at the time of seed alignment collection. Regardless, how carefully a gathering score is selected by the expert, it remains a subjective decision. The sequence with a true model hit with lowest score (as well as the false hit with the highest score) critically depends on the size of the non-redundant protein database, the variety of sequences therein and the quality of the seed alignment at the time of model construction. Sequence databases have a strong growth due to increasingly cheaper sequencing. With time, our biological knowledge grows and we know more about previously uncharacterized sequences. Not surprisingly, gathering thresholds have an inherent trend to be increased with time even if the underlying seed alignments do not change.

For example in the case of PF00583 (Acetyltransferase) in the introductory Eco1 example, the gathering scores have evolved the following way: Pfam5 (1999) with 6.5 (global mode/gm) and 6.5 (fragment- mode/fm), Pfam6 (2000) with 15 (gm) and 15 (fm) (with some shortening of the alignment compared with Pfam5), Pfam7 (2001) with 18.2 (gm) and 16.3(fm). The reader is invited to return to Figure S1 to verify that only the *Drosophila melanogaster* sequence AE003559 would make it over the gathering score threshold in 2000 and later whereas the Pfam5 gathering score would clearly support many homologues. Thus, the experimentally verified discovery of the Eco1 acetyltransferase might have been overseen after 2000 based on a gathering score criterion but it would never disappear from the radar in an E-value guided search at any time point. As for the other introductory example, the PF00497 (SBP_bac_3) model, the fragment-mode gathering thresholds have also been heavily changed over time: For Pfam5 (1999) and Pfam6 (2000) GA = −20 (fm), whereas, for Pfam7 (2001), GA = 49.9 (fm); thus, the sequence "Alt a 1" would have been a hit based on the gathering threshold criterion until the year 2000 but it would be suppressed with the more recent versions of Pfam. At the same time, the E-value generated by the example did not change.

Secondly, gathering scores hide the problem of balance between true-negative and false-positive hits. Although increasing gathering scores (as there is a trend in Pfam releases) reduce false-positive hit rates, this approach excludes a growing number of true hits and, thus, also limits the extrapolation power of domain models into the space of uncharacterized sequences. On the contrary, an E-value gives insights into the orders of magnitude of error rates when assuming the annotation transfer to be correct. The user of a gathering threshold guided assignment does have the illusion of dealing with ultimately correct hits; in contrast, an E-value provides a quantitative and typically non-zero statistical measure for annotation error.

Thirdly, gathering thresholds do not relate well with the statistics of hit distribution in the non-redundant database. In the HMMER2 manual, Sean Eddy says on page 22 "Calibrated HMMER E-values tend to be relatively accurate. E-values of 0.1 or less are, in general, significant hits". Further on page 43, he writes "The best criterion of statistical significance is the E-value. The E-value is calculated from the bit score. It tells you how many false positives you would have expected to see at or above this bit score. Therefore a low E-value is best; an E-value of 0.1, for instance, means that there's only a 10% chance that you would've seen a hit this good in a search of non-homologous sequences. *Typically, I trust the results of HMMER searches at about E = 0.1 and below, and I examine the hits manually down to E = 10 or so.*"

Whereas the E-values in the order of 0.1 are generally considered being below the significance threshold (and they are for many good domain models as we observed in our practice), we

find actually no general relationship between domain-specific gathering scores and E-value thresholds for Pfam release 23 (Figure 6). In fact, the gathering score thresholds can result in vastly different E-value thresholds (range 10^{-35} to 10^{5}). Nevertheless, E-value thresholds close to the empirical value of 0.1 are most frequent in Pfam (see bottom part of Figure 6 with the peak of the E-value threshold histogram at 0.07) and one wonders why there are domains at all where the E-value corresponding to the gathering score does dramatically differ from 0.1. There might be many reasons for this discrepancy and its resolution would require dedicated research. It would be of interest to see how the growth of sequence databases as well as of the biological knowledge (in contrast to the more static seed alignments and domain models) has an effect here. We also suggest that, among other factors, incompleteness of the seed alignments with regard to the actual sequence variety (due to sequences that became available after model construction), alignment length (actually involving several domains in one model instead of one), the presence of non-globular segments or other issues of alignment quality might play a role here.

Lastly, E-values are comparable since they are a statistical measure but gathering score thresholds are not and, therefore, scores calculated from different domain models or prediction tools cannot be compared. This makes decisions among domain models and other prediction tools hitting the same segment in the query difficult. For example, the sequence XP_001939830.1 (Table S1, entry 19) illustrates this point. It is a hit in the fragment mode both by MFS_1 (over positions 49 to 388, E = 1.9e-21, score = 79.3> gatheringscore = 25.4) and HCV_NS4b (over overlapping positions 178 to 211, E = 4.8e-5, score = 14.6>gathering score = 14.5). Whereas the first is a full domain hit, the second one covers essentially only a TM region. Although both are above gathering score, the E-value clearly supports finding the correct annotation.

We do not want to create the impression that we wish to nail down the Pfam team on, maybe, some unfortunately selected thresholds for previous releases. Also, the specific examples (rather the existence of such examples) are not relevant for the conclusions in this paper. We have to live with some error rate. In contrast, it is important that the theoretical fundamentals are reliable, that systematic causes for possibly questionable annotations are increasingly suppressed and that, together with the Pfam team, the community develops the theory.

About the state of automated annotation transfer in public databases

It is difficult to assess the total amount of wrong annotations currently persisting in public sequence databases since most of the protein sequences have never been a target of experimental study. With regard to theoretically derived function descriptions, the individual teams contributing to sequence databases, apparently, apply criteria with differing stringency and rigor. It appears that unrestrained annotation transfer justified by spurious sequence similarities is a major cause for annotation errors [64,65] and this process is facilitated by the convenience of automated annotation pipelines. Analogous to a self-replicating virus, any first annotation error perpetually propagates itself to any existing or new sequence database by the virtue of annotation transfer ironically [64,65].

In their analysis of database annotations for 37 enzyme families, Schnoes *et al.* [74] find approximately 40% of submitted sequences in 2005 were misannotated while none carried wrong annotation in 1993. It should emphasized that, in most cases, the misannotation involves an enzyme family or superfamily mix-up. To note, the fold as

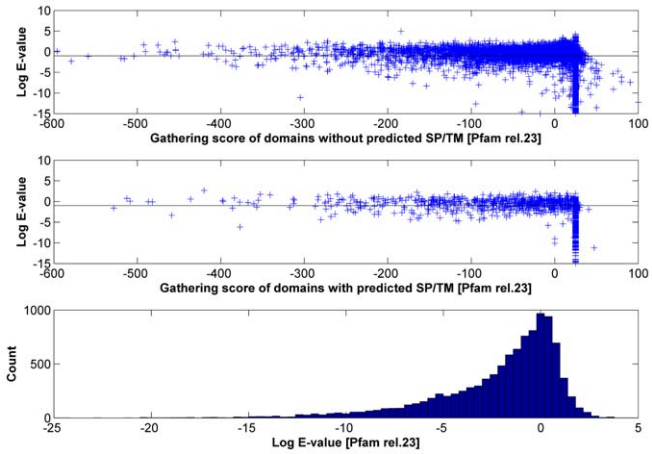

Figure 6. Relationship between the gathering score and the corresponding E-value threshold for Pfam domain library release 23. Whereas the y-axis shows the gathering score threshold (GA) for the global-mode search, x-axis shows the corresponding E-value threshold (in decimal log scale) calculated with the domain-specific extreme-value function with parameters provided in the corresponding HMM file (for an NR database size of 7365651 sequences) for this score. The upper plot represents the distribution for 9126 domains without detected SP/TM region, the middle part shows the same for the 1214 domains with SP/TM problems. Effectively, there is no clear correlation between gathering score and E-value threshold. If E-values close to 0.1 are considered significant, all dots should be close to the "−1" line (horizontal dashed lines) in this graph and, indeed, there is some agglomeration of data points in that area; yet, there are numerous outliers. Note that the E-values are computed using the equation

$$E = N\{1 - \exp[-\exp(-\lambda GA + \lambda\mu)]\}$$

where N is the database size, μ and λ are the extreme value distribution (EVD) parameters of the domain model. The bottom plot depicts the histogram of the 10340 domains in Pfam rel.23. The median of all log E-values that corresponded to the domain-specific GAs is found to be −1.16. This translates to an E-value of 0.07.
doi:10.1371/journal.pcbi.1000867.g006

well as the overall function have been recognized correctly. We want to caution that the disregard of non-globular segments in context of homology-based conclusions can contribute to annotation errors. This may mean not just missing the correct subfamily but leading function assignment far astray. In the examples provided in this article, the true function of the protein hits has nothing in common with the problematic domain model hit except for the occurrence of a hydrophobic region that matches the SP/TM segment(s).

Thus, the criteria for sequence homology in their present form appear not directly applicable to non-globular segments. SPs/TMs as part of domain models lead to pollution of database annotations as our PIR iProClass v3.74 analysis demonstrates. As a matter of fact, it is very difficult to prove wrong annotation for experimentally uncharacterized sequences otherwise than by detecting logical contradictions. Whereas the examples in Tables S1 and S2 have been carefully scrutinized manually against structural and literature information, the same approach is out of question for a database-scale study, even for selected domain models as in Table 2. Therefore, we applied a criterion based on score partition into the SP/TM-specific part and the remainder to estimate the amount of false-positive hits to get at least a lower boundary estimate for the scale of the problem. We did show the existence of problematic annotations from a few to over ten percent for a validated set of 11 Pfam domains that include SP/TM regions.

Conclusions

To conclude, sequence similarity among non-globular protein segments does not necessarily imply homology. Since matching of SPs/TMs creates the illusion of alignable hydrophobic cores, the inclusion of SPs/TMs into domain models without precautions can give rise to wrong annotations. We find that clearly more than 1001 domains among the 10340 models of Pfam release 23 suffer from this problem, whereas the issue is of relatively low importance for domains of SMART version 6 (18 out of 809). As expected, fragment-mode HMM searches generate promiscuous hits limited to solely the SP/TM part among clearly unrelated proteins for these models. More worryingly, we show explicit examples that the scores of clearly false-positive hits even in global-mode searches can be elevated into the significance range just by matching the hydrophobic runs. In the PIR iProClass database v3.74, we find that between 2.1% and 13.6% of its annotated Pfam hits appear unjustified for a set of validated domain models. We suggest a workflow of flagging problematic hits arising from SPs/TMs-containing models for critical reconsideration by annotation users. On the other hand, we have also seen that the inclusion of SP/TM regions into domain models can give rise to false negatives by imposing the need to have good scores over these regions in the query sequences when the actual domain occurs without the SP/TM context.

Materials and Methods

Assessment of false-positive detection of SP/TM segments by unsupervised prediction

It is well known that the problem of transmembrane helix prediction is not so much the detection of true hits as the suppression of false-positives [85]. In our context, it is important to have as few as possible wrong SP/TM predictions (and to carefully control their fraction) even on the expense of loosing true examples. Further, SP/TM prediction tools are designed for application to a single sequence, not to an alignment possibly polluted with gaps and/or shifts among predicted SP/TM regions among various sequences. Therefore, we developed the following procedure and statistical criteria for processing outputs of academically available SP/TM predictors.

In the general case, domain models are characterized by both seed and full alignments. We think that, in our context, operating with seed alignments is preferable since they are manually validated and are supposed to have lower levels of inclusions of unrelated sequences.

For a given domain model alignment, each sequence was subjected to sets of transmembrane (TM) and signal peptide (SP) segment predictors. We have used the following TM predictor tools – DASTM [85,86], TMHMM [81], HMMTOP [87], SAPS [88], PhobiusTM [89,90] and SP predictors – SignalP [29,33,91], PhobiusSig [89,90]. The variable M denotes the number of predictors in each set ($M=5$ and $M=2$ for TM and SP predictions respectively).

For each predictor m, only the positive or negative SP/TM predictions for each residue a_{ij} (where i is the sequence and j the alignment position) were considered, their respective prediction scores were ignored. Essentially, each positive/negative prediction can be seen as a Bernoulli random variable I_{ij} (an indicator variable assuming values one or zero). Collectively, a set of Bernoulli variables for each column j (made up by a number of sequences in the alignment) can be treated as a binomial random variable X_j having the value k (sum of I_{ij} over all sequences i).

To ensure that columns of domain alignments with an unequal number of sequences and/or gap instances are treated comparably, a hypothesis testing step is introduced [92]. Let n be the number of sequences (excluding gaps in the particular column) in the alignment. With p, we denote the actual (*a priori* unknown) probability of the residue a_{ij} to belong to a true SP/TM segment. For each test, one wishes to determine if each column is a SP/TM residue given the observed predictions under equal chance condition. Hence, the null and the alternative hypotheses are stated as $H_o : p \le 0.5, H_A : p > 0.5$. The type I error is defined as

$$P(X \ge k) = \sum_{x \ge k}^{n} \binom{n}{x} p^x (1-p)^{n-x} \tag{1}$$

We assume the null hypothesis is rejected at a significance level of $\alpha \le 0.05$. This means that, for alignments of four sequences and less, $P(X \ge k) \ge 1 - P(X \le 3) = 0.0625$ and, therefore, the null hypothesis is never rejected. The statistical test requires alignments of 5 sequences or more. For each rejected hypothesis, the corresponding expected positive predictions $k_{\exp}$ is calculated as

$$k_{\exp} = P(X \le k) \times k \tag{2}$$

Otherwise, $k_{\exp}$ is set to zero. Finally, the estimated probability $\hat{p}_{j,m}$ of column j to represent a residue of a true SP/TM segment is given as

$$\hat{p}_{j,m} = \begin{cases} \dfrac{k_{\exp}}{n} & if \quad k_{\exp} \ge 1 \\ 0.01 & if \quad k_{\exp} = 0 \end{cases} \tag{3}$$

The lower line in equation 3 is to avoid logarithms of zero in formulas below. Collectively, each domain alignment leads to a matrix of J column probabilities $\hat{\mathbf{p}}_j$ with M predictors for each segment type (TM or SP). The total logarithmic probability per column for either type of predictors is given as

$$\log \hat{p}_{j,total} = \sum_{m=1}^{M} \log \hat{p}_{j,m} \tag{4}$$

If $\log \hat{p}_{j,total} > M \log(0.01)$, we assume that position j belongs to a predicted SP/TM segment. We define the indicator functions F_j being unity in this case and zero in the other. Thus, a section of continuous alignment positions of unities in F_j is called a predicted TM (or SP) segment. The average logarithmic probability $\langle \hat{p} \rangle$ of this segment is given as

$$\log\langle \hat{p} \rangle = \frac{1}{R} \sum_{r}^{R} \log \hat{p}_{r,total} \tag{5}$$

where R is the total number of predicted residue columns for the given SP/TM segment and r is the starting position of this TM helix or signal peptide.

In practice, some of the predicted transmembrane helices and signal peptides can be fragmented due to small gaps in the alignment. In the case of signal peptide fragments, it is reasonable to assume that all the fragments come from a single signal peptide. Consequently, the average logarithm probability of SP prediction per domain is simply calculated using (5) summing over the smallest region that contains both the N-terminal alignment position and the C-terminal boundary of the most C-terminal predicted segment.

However, for the case of the fragmented TM helices, the situation can be complicated by occurrences of multiple transmembrane segments within the alignment. As indicator which fragments to unite into one segment, we use the raw TM predictions. The indicator function $Q_{j,m}$ is set to unity at position j where predictor m generates $k \geq 1$ (union of the column-wise TM predictions in all sequences); otherwise, it is equal to zero. The composite indicator function Q_j is set to unity only at positions j where $Q_{j,m} = 1$ for all predictors that produce overlapping hits (intersection of predicted TM segments among all predictors). Similarly to predicted segments in F_j, continuous runs of ones can be delineated in Q_j. If two predicted segments in F_j overlap with the same predicted segment in Q_j, the zero values of F_j in-between the two segments are restored to unity. The union operation preserves the continuity within a helix while the intersection operation maintains separation between helices. Finally, the average logarithm probability $\langle \hat{p} \rangle$ for a predicted TM segment consisting of G united fragments is given as weighted average

$$\log\langle \hat{p} \rangle = \frac{\sum_{g=1}^{G} R_g \log\langle \hat{p}_g \rangle}{\sum_{g=1}^{G} R_g} \tag{6}$$

where R_g is the total number of predicted TM residue columns in the g^{th} TM helix fragment. Only predicted segments with a $\log\langle \hat{p} \rangle$ above a cutoff (*TMcutoff* or *SPcutoff* respectively; see below) are considered in the further analysis; others are discarded and the respective positions in F_j are set to zero.

We have used our algorithm also to find SP/TM regions in α- and membrane proteins classified by SCOP [41,42] as a benchmark for finding TMcutoff. In this case, a single sequence and not an alignment is available; thus, we start with equation 3 and the conditions $k_{\text{exp}} = k$ and $n = 1$.

Specific considerations for transmembrane and signal peptide predictions

For the TM prediction problem, only the individual TM helix has been defined so far. To define a TM region that composes of one or more TM helices, adjacent TM helices separated by less than 40 amino acid residues are concatenated to form a region. The choice of 40 amino acids is based on the current knowledge that the smallest known globular domains such as Zinc fingers [93–97] are above 40 residues in length; thus, the inter-TM-helix residues just form some type of linker.

For the SP prediction problem, it is relevant that the actual N-terminus might be missing in the domain alignment. Thus, two rounds of SP predictions are necessary. After the initial round, the domain sequences with positive SP predictions are subjected to blastp runs (with parameters '-M BLOSUM62 -G 11 -E 1 -F F -I T') against NR database to retrieve their full sequence data. Only the full sequence data with percent identity ≥95% and Blast E-value ≤0.01 are then subjected to SP predictions. Finally, only overlapped SP predictions that are confirmed in both rounds are retained for further processing.

Determination of domain error cutoffs

The appropriate cutoff for predicted TM and SP segments in domain alignments have been determined with the help of the SCOP v1.75 [41,42] α protein, membrane class database and SMART version 6 database [29,40].

TM prediction hits among SCOP α class proteins are false-positives since the database contains predominantly structural helices. On the other hand, the membrane class contains mostly TM helices that made up the true-positive hits for these predictors. Figure 7 shows the histograms of the structural (top) and transmembrane (bottom) helices respectively. The clear separability between the two histograms strongly demonstrated that these two classes of helices are distinct. Table 3 gives the associated false-positive and false-negative rates of TM predictions at the various TM cutoffs.

In the case of the signal peptide prediction, both α- and membrane SCOP classes will deliver false-positive hits while the domain models from SMART with signal peptide are true positive hits. Figure 8 shows the histograms of false (top) and true signal peptides (bottom) respectively. In all, 45 out of 49 seed sequences for 5 SMART domains (SM00190 IL4_13, SM00476 DNaseIc, SM00770 ZN_dep_PLPC, SM00792 Agouti, SM00817 Amelin) were found to contain a predicted signal peptide. Out of them, predicted signal peptides for sequences from 4 domain models (except SM00817) were validated by their absence as a structural helix in the respective PDB entries (see Results, Table 1). Table 4 gives the associated false-positive and false-negative rates of SP predictions at different SP cutoffs.

Decomposition of HMM log odd scores into sequence segment specific components

In the following, the reader is assumed to be familiar with chapter three of [47] and our derivations starts with a reformulated version of their equation 3.6. Let the observed and hidden state sequences be Y and X. The joint probability of the observed and hidden state sequences is given as $P(Y,X) = P(Y_0 \ldots Y_L; X_0 \ldots X_L; a,b)$ where a and b are the emission and state transition probabilities of the model, and L is the length of the sequence. Upon expanding the equation, we get

$$\begin{aligned} P(Y,X) &= P(Y|X)P(X) \\ &= \prod_{i=0}^{L} P(Y_i|X_i) \times P(X_L|X_{L-1})P(X_{L-1}| \\ &\quad X_{L-2}) \ldots P(X_1|X_0)P(X_0) \\ &= \prod_{i=0}^{L} P(Y_i|X_i) \times P(X_0) \prod_{i=1}^{L} P(X_i|X_{i-1}) \end{aligned} \tag{7}$$

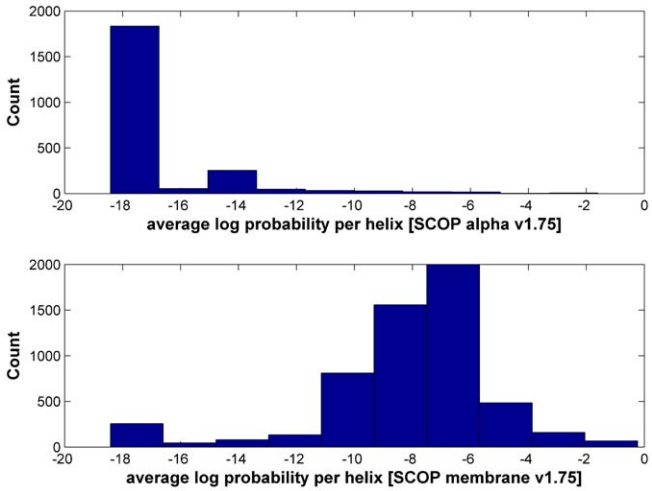

Figure 7. Histograms of average log probability per predicted transmembrane helix for SCOP v1.75 α-proteins class and membrane protein class. The top (average log probability per predicted transmembrane helix for SCOP v1.75 α-proteins class) and bottom (average log probability per predicted transmembrane helix for SCOP v1.75 membrane protein class) histograms represent the false-positive and true-positive distributions for TM predictions respectively. The total number of predicted structural and membrane helices is 2293 and 5592 respectively.
doi:10.1371/journal.pcbi.1000867.g007

The marginal probability of the observed sequence Y can be then be summed across all hidden sequence X as

$$P(Y)=\sum_{x} P(Y,X) \tag{8}$$

Often the most probable path given by X^* (given by the Viterbi algorithm) is a good approximation to $P(Y)$. Hence we have

$$P(Y)\approx P(Y,X^*) \tag{9}$$

In the HMM formalism, we use the log odd scores v for scoring sequences. Therefore, for an observed sequence Y, this is given as

$$v=\log_2\frac{P(Y;a_{HMM},b_{HMM})}{P(Y;a_{null},b_{null})} \tag{10}$$

Assume that $X^*=X$. Using (7) to (10), the log odd score v can be rewritten as

$$\begin{aligned} v&=\log_2\frac{P(Y,X;a_{HMM},b_{HMM})}{P(Y,X;a_{null},b_{null})}\\ &=\log_2\left[\frac{\prod_{i=0}^{L}P(Y_i|X_i;a_{HMM})}{\prod_{i=0}^{L}P(Y_i|X_i;a_{null})}\times\frac{P(X_0;b_{HMM})\prod_{i=1}^{L}P(X_i|X_{i-1};b_{HMM})}{P(X_0;b_{null})\prod_{i=1}^{L}P(X_i|X_{i-1};b_{null})}\right]\\ &=\sum_{i=0}^{L}\log_2\frac{P(Y_i|X_i;a_{HMM})}{P(Y_i|X_i;a_{null})}+\sum_{i=1}^{L}\log_2\frac{P(X_i|X_{i-1};b_{HMM})}{P(X_i|X_{i-1};b_{null})}+\\ &\quad\log_2\frac{P(X_0;b_{HMM})}{P(X_0;b_{null})}\\ &=\sum_{i=0}^{L}\log_2 e(Y_i|X_i)+\sum_{i=1}^{L}\log_2 t(X_i|X_{i-1})+\log_2 t(X_0) \end{aligned} \tag{11}$$

where e and t are the emission and state transition log odd scores. Thus, the total score is represented as a linear combination of

Table 3. FP and FN rates of TM predictions based on different TM cutoffs.

Average log probability of TM prediction	No. of FP	FP rate (%)	No. of FN	FN rate (%)
≥−6	21	0.91	4519	80.81
≥−7	37	1.61	3401	60.82
≥−8	45	1.96	2520	45.06
≥−9	47	2.04	1593	28.49
≥−10	72	3.14	910	16.27
≥−11	84	3.66	526	9.41
≥−12	107	4.67	418	7.48
≥−13	125	5.45	381	6.81
≥−14	206	8.98	362	6.47

The first column gives the various cutoffs for the average log probability of TM helix prediction (refer to equations 5 and 6). The next two columns denote the number and percentage of false-positive TM helices with respect to 2293 predicted helices from SCOP α-proteins based on the corresponding cutoff rate. Similarly, the last two columns describe the number and percentage of false-negative TM helices with respect to 5592 predicted helices from SCOP membrane proteins.
doi:10.1371/journal.pcbi.1000867.t003

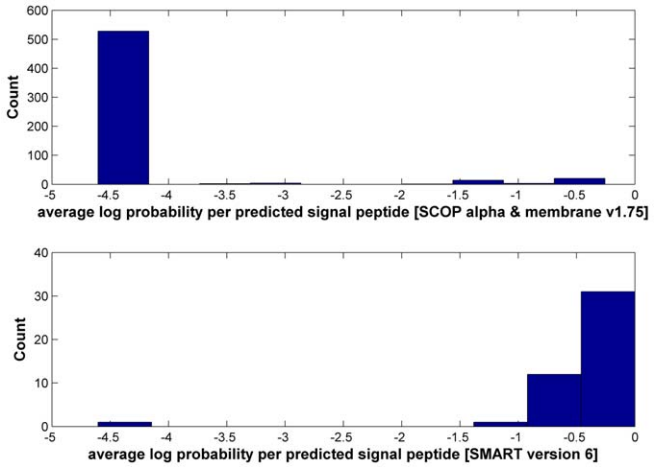

Figure 8. Histograms of average log probability per predicted signal peptide for SCOP v1.75 α- and membrane protein class and SMART version 6. The top (average log probability per predicted signal peptide for SCOP v1.75 α- and membrane protein class) and bottom (average log probability per predicted signal peptide for SMART version) histograms represent the false-positive and true-positive distributions for the SP predictions respectively. The total number of predicted signal peptides for SCOP α- and membrane proteins is 193 and 379 respectively, while the total number for SMART is 45. All except SM00817 Amelin (no available structure) were validated against their respective PDB entries.
doi:10.1371/journal.pcbi.1000867.g008

sequence position-specific terms (plus some position-independent constants and, what is not considered here, the so-called null2 correction). Therefore, the HMM log odd score can be decomposed into sequence segment-specific contributions, for example those arising from its globular and non-globular regions:

$$v = v_{glob} + v_{nonglob} + \log_2 t(X_0) \tag{12}$$

where $v_{glob} = \sum_{i=0}^{Lglob} \log_2 e(Y_i|X_i; a_{HMM}, a_{NULL}) + \sum_{i=1}^{Lglob} \log_2 t(X_i|X_{i-1}; b_{HMM,}, b_{NULL})$,

$$v_{nonglob} = \sum_{j=0}^{Lnonglob} \log_2 e(Y_j|X_j; a_{HMM}, a_{NULL}) + \sum_{j=1}^{Lnonglob} \log_2 t_{nonglob}(X_j|X_{j-1}; b_{HMM}, b_{NULL})$$

$L_{glob}, L_{nonglob}$ are the total lengths of the globular and non-globular segments respectively; a_{HMM}, a_{NULL} are the emission probabilities of the HMM and the null model respectively; b_{HMM}, b_{NULL} are the transition probabilities of the HMM and the null model respectively. In our work, we consider the SP/TM segments defined by $F_j = 1$ as non-globular part and the rest as globular.

Estimation of the non-SP/TM component of the gathering score threshold

Here, equation (12) that denotes the total score v can be re-written as the sum of a non-SP/TM-specific $v_{nonSPTM}$, a SP/TM-specific v_{SPTM}, and a position-independent score c for a sequence as follows

$$v = v_{nonSPTM} + v_{SPTM} + c \tag{13}$$

In the following, we wish to derive the relative contribution of $v_{nonSPTM}$ and v_{SPTM} at scores v close to the gathering score GA. We assume that the proportion between $v_{nonSPTM}$ and v_{SPTM} as represented by the sequences from the seed alignment holds also for lower scores of true hits. Let the random variables V_{SPTM} and $V_{nonSPTM}$ denote the SP/TM-specific scores v_{SPTM} and non-SP/TM specific scores $v_{nonSPTM}$ of N seed sequences of the domain model. The sample mean $\overline{V}$ of the random variables are given as

Table 4. FP and FN rates of SP predictions based on different SP cutoffs.

Average log probability of SP prediction	No. of FP	FP rate (%)	No. of FN	FN rate (%)
≥−0.5	20	3.50	8	17.78
≥−1	23	4.02	1	2.2
≥−2	38	6.64	1	2.2
≥−3	38	6.64	1	2.2
≥−4	44	7.69	1	2.2

The first column gives the various cutoffs for the average log probability of SP prediction (refer to equation 5). The next two columns denote the number and percentage of false-positive SP with respect to 572 predicted SP from SCOP α- and membrane proteins based on the corresponding cutoff rate. Similarly, the last two columns describe the number and percentage of false-negative SP with respect to 45 predicted SP in seed sequences from SMART version 6 alignments.
doi:10.1371/journal.pcbi.1000867.t004

$$\overline{V}_{nonSPTM} = \frac{1}{N}\sum_{n=1}^{N} v_{n,nonSPTM} \tag{14}$$

$$\overline{V}_{SPTM} = \frac{1}{N}\sum_{n=1}^{N} v_{n,SPTM} \tag{15}$$

Here, we introduce a scaling factor in the form of random variable A as a shift factor in the logarithmic scale that relates the random variables V_{SPTM}, $V_{nonSPTM}$ and the constant c to the constant GA (gathering score threshold provided the domain model). The relationship can be written as

$$GA = V_{nonSPTM} + V_{SPTM} + c + A \tag{16}$$

Equation (16) can further be expressed in terms of two random variables $G_{nonSPTM}$ and G_{SPTM} that denote the SP/TM-specific and non-SP/TM-specific gathering score threshold means respectively.

$$\begin{aligned} GA &= \left[V_{nonSPTM} + \frac{L_{nonSPTM}}{L}A\right] + \left[V_{SPTM} + \frac{L_{SPTM}}{L}A\right] + c \\ &= G_{nonSPTM} + G_{SPTM} + c \end{aligned} \tag{17}$$

To obtain the mean of $G_{nonSPTM}$, we first need to solve for A by rewriting equation (16) in terms of A as given

$$A = GA - V_{nonSPTM} - V_{SPTM} - c \tag{18}$$

Consequently, taking the expectation of A (the sample mean over the seed alignment), we get

$$\overline{A} = GA - \overline{V}_{nonSPTM} - \overline{V}_{SPTM} - c \tag{19}$$

Finally, the non-SP/TM specific contribution $\bar{G}_{nonSPTM}$ to the gathering score threshold is given as

$$\overline{G}_{nonSPTM} = \overline{V}_{nonSPTM} + \frac{L_{nonSPTM}}{L}\overline{A} \tag{20}$$

Similarly, a SP/TM-specific threshold $\bar{G}_{SPTM}$ can be calculated. For the 11 domain models in Table 2, $\bar{A}$ is vastly negative and ranges from -76.19 (Cation_ATPase_N) to -614.95 (Lamp); thus, $v_{nonSPTM}$ is much larger than $\bar{G}_{nonSPTM}$ for any seed sequence.

Estimation of unjustified annotation instances in the database

For a set of sequences with a common problematic domain annotation, each sequence score can be represented by $(v, v_{nonSPTM})$. If we assume that all true hits must score above the gathering score GA and the threshold $\bar{G}_{nonSPTM}$ as derived in the previous section is truly the lower boundary for a score contribution from the non-SP/TM part of a correct domain hit, the validity of the annotation can be assessed by comparing $(v, v_{nonSPTM})$ with $(GA, \bar{G}_{nonSPTM})$. If $v \geq GA$ and $v_{nonSPTM} \geq \overline{G}_{nonSPTM}$, the domain hit is considered true-positive. If $v < GA$ and $v_{nonSPTM} \geq \overline{G}_{nonSPTM}$, the SP/TM part of the domain hit is degenerated; yet, the non-SP/TM part is well represented and we consider these hits false-negatives. In all cases with $v_{nonSPTM} < \overline{G}_{nonSPTM}$, we consider the annotation with the domain unjustified. Even if the total score is above the gathering score, formally, the shift to the significant range is only achieved by a large score from the SP/TM region.

We find that our derivation for $\bar{G}_{nonSPTM}$ is credible since it does not compromise the sensitivity of the domain models. The fraction of false-negative hits over the total retrieved sequences per problematic domain ranges between 0 to 5% (with the only outlier GRP at 10.1%).

Supporting Information

Figure S1 PF00583 hits leading to the Eco1 function discovery.
Found at: doi:10.1371/journal.pcbi.1000867.s001 (0.03 MB PDF)

Figure S2 False-positive hit of PF00497 in Alt a 1.
Found at: doi:10.1371/journal.pcbi.1000867.s002 (0.01 MB PDF)

Protocol S1 Mini-site with supplementary information, archive created with WinRAR (to be downloaded from http://www.rarlab.com/download.htm). Besides HMMER outputs, alignments, etc. for Tables S1, S2 and S3 and for Figures 4 and 5, we provide lists of affected Pfam models as well as HMMs for these domains without the respective SP/TM segments. The content of this file may also be found at http://mendel.bii.a-star.edu.sg/SEQUENCES/ProblemDomains-TM+SP/.
Found at: doi:10.1371/journal.pcbi.1000867.s003 (32.83 MB WinRAR)

Table S1 Summary of selected sequence hits with problematic domain annotations (fragment-mode search).
Found at: doi:10.1371/journal.pcbi.1000867.s004 (0.04 MB PDF)

Table S2 Summary of selected sequence hits with problematic domain annotations (global-mode search).
Found at: doi:10.1371/journal.pcbi.1000867.s005 (0.05 MB PDF)

Table S3 Summary of selected false-negative sequence hits with problematic domain annotations (global-mode search).
Found at: doi:10.1371/journal.pcbi.1000867.s006 (0.05 MB PDF)

Acknowledgments

The authors are grateful to Birgit Eisenhaber for critically reading this manuscript. It is also acknowledged that this work has been made available to the teams of Pfam (via Alex Bateman) and SMART (via Peer Bork) prior to publication. As a consequence, a considerable number of domain model revisions have been made leading, for example to the exclusion of signal peptides from models in future Pfam releases.

Author Contributions

Conceived and designed the experiments: WCW FE. Performed the experiments: WCW. Analyzed the data: WCW SMS FE. Contributed reagents/materials/analysis tools: WCW. Wrote the paper: WCW SMS FE.

References

1. Eisenhaber F (2006) Prediction of Protein Function: Two Basic Concepts and One Practical Recipe. In: Eisenhaber F, ed. Discovering Biomolecular Mechanisms with Computational Biology. Georgetown and New York: Landes Biosciences and Springer. pp 39–54.
2. Ooi HS, Kwo CY, Wildpaner M, Sirota FL, Eisenhaber B, et al. (2009) ANNIE: integrated de novo protein sequence annotation. Nucleic Acids Res 37: W435–W440.
3. Sammut SJ, Finn RD, Bateman A (2008) Pfam 10 years on: 10,000 families and still growing. Brief Bioinform 9: 210–219.
4. Ivanov D, Schleiffer A, Eisenhaber F, Mechtler K, Haering CH, et al. (2002) Eco1 is a novel acetyltransferase that can acetylate proteins involved in cohesion. Curr Biol 12: 323–328.
5. Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, et al. (1998) Predicting function: from genes to genomes and back. J Mol Biol 283: 707–725.

6. Eisenhaber F (2006) Bioinformatics: Mystery, Astrology or Service Technology. In: Eisenhaber F, ed. Discovering Biomolecular Mechanisms with Computational Biology. Georgetown and New York: Landes Biosciences and Springer. pp 1–10.
7. Devos D, Valencia A (2000) Practical limits of function prediction. Proteins 41: 98–107.
8. Sander C, Schneider R (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9: 56–68.
9. Todd AE, Orengo CA, Thornton JM (2001) Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 307: 1113–1143.
10. Bork P, Gibson TJ (1996) Applying motif and profile searches. Methods Enzymol 266: 162–184.
11. Gough J (2005) Convergent evolution of domain architectures (is rare). Bioinformatics 21: 1464–1471.
12. Doolittle RF (1994) Convergent evolution: the need to be explicit. Trends Biochem Sci 19: 15–18.
13. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89: 10915–10919.
14. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
15. Altschul SF, Gertz EM, Agarwala R, Schaffer AA, Yu YK (2009) PSI-BLAST pseudocounts and the minimum description length principle. Nucleic Acids Res 37: 815–824.
16. Eisenhaber F, Bork P (1998) Sequence and Structure of Proteins. In: Schomburg D, ed. Recombinant proteins, monoclonal antibodies and theraeutic genes. Weinheim: Wiley-VCH. pp 43–86.
17. Holland TA, Veretnik S, Shindyalov IN, Bourne PE (2006) Partitioning protein structures into domains: why is it so difficult? J Mol Biol 361: 562–590.
18. Veretnik S, Bourne PE, Alexandrov NN, Shindyalov IN (2004) Toward consistent assignment of structural domains in proteins. J Mol Biol 339: 647–678.
19. Hulo N, Bairoch A, Bulliard V, Cerutti L, Cuche BA, et al. (2008) The 20 years of PROSITE. Nucleic Acids Res 36: D245–D249.
20. Henikoff JG, Greene EA, Taylor N, Henikoff S, Pietrokovski S (2002) Using the blocks database to recognize functional domains. Curr Protoc Bioinformatics Chapter 2: Unit.
21. Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, et al. (2003) PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res 31: 400–402.
22. Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, et al. (2009) SUPERFAMILY–sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Res 37: D380–D386.
23. Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, Weese-Scott C, et al. (2009) CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res 37: D205–D210.
24. Selengut JD, Haft DH, Davidsen T, Ganapathy A, Gwinn-Giglio M, et al. (2007) TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. Nucleic Acids Res 35: D260–D264.
25. Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, et al. (2005) The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res 33: D284–D288.
26. Bru C, Courcelle E, Carrere S, Beausse Y, Dalmar S, et al. (2005) The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acids Res 33: D212–D215.
27. Portugaly E, Linial N, Linial M (2007) EVEREST: a collection of evolutionary conserved protein domains. Nucleic Acids Res 35: D241–D246.
28. Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L, et al. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. Bioinformatics 15: 1000–1011.
29. Letunic I, Doerks T, Bork P (2009) SMART 6: recent updates and new developments. Nucleic Acids Res 37: D229–D232.
30. Eisenhaber B, Eisenhaber F (2005) Sequence complexity of proteins and its significance in annotation. In: Subramaniam S, ed. "Bioinformatics" in the Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics. New York: Wiley Interscience;DOI:10.1002/047001153X.g403313.
31. Eisenhaber B, Eisenhaber F (2007) Posttranslational modifications and subcellular localization signals: indicators of sequence regions without inherent 3D structure? Curr Protein Pept Sci 8: 197–203.
32. Tompa P, Dosztanyi Z, Simon I (2006) Prevalent structural disorder in E. coli and S. cerevisiae proteomes. J Proteome Res 5: 1996–2000.
33. Bendtsen JD, Nielsen H, von HG, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340: 783–795.
34. Eisenhaber B, Bork P, Eisenhaber F (1998) Sequence properties of GPI-anchored proteins near the omega-site: constraints for the polypeptide binding site of the putative transamidase. Protein Eng 11: 1155–1161.
35. Eisenhaber B, Bork P, Eisenhaber F (1999) Prediction of potential GPI-modification sites in proprotein sequences. J Mol Biol 292: 741–758.
36. Gruber M, Soding J, Lupas AN (2006) Comparative analysis of coiled-coil prediction methods. J Struct Biol 155: 140–145.
37. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, et al. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 29: 2994–3005.
38. Stojmirovic A, Gertz EM, Altschul SF, Yu YK (2008) The effectiveness of position- and composition-specific gap costs for protein similarity searches. Bioinformatics 24: i15–i23.
39. Schneider G, Neuberger G, Wildpaner M, Tian S, Berezovsky I, et al. (2006) Application of a sensitive collection heuristic for very large protein families: evolutionary relationship between adipose triglyceride lipase (ATGL) and classic mammalian lipases. BMC Bioinformatics 7: 164.
40. Schultz J, Milpetz F, Bork P, Ponting CP (1998) SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci U S A 95: 5857–5864.
41. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, et al. (2008) Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res 36: D419–D425.
42. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536–540.
43. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, et al. (2000) The Pfam protein families database. Nucleic Acids Res 28: 263–266.
44. Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, et al. (1999) Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. Nucleic Acids Res 27: 260–262.
45. Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res 26: 320–322.
46. Bernsel A, Viklund H, Elofsson A (2008) Remote homology detection of integral membrane proteins using conserved sequence features. Proteins 71: 1387–1399.
47. Durbin R, Eddy S, Krogh A, Mitchison G (1998) Biological sequence analysis: Probabilistic models of proteins and nucleic acids.
48. Eddy SR (2004) What is a hidden Markov model? Nat Biotechnol 22: 1315–1316.
49. Eddy SR (2008) A probabilistic model of local sequence alignment that simplifies statistical significance estimation. PLoS Comput Biol 4: e1000069.
50. Watanabe R, Murakami Y, Marmor MD, Inoue N, Maeda Y, et al. (2000) Initial enzyme for glycosylphosphatidylinositol biosynthesis requires PIG-P and is regulated by DPM2. EMBO J 19: 4402–4411.
51. Janssen R, Smeitink J, Smeets R, van Den HL (2002) CIA30 complex I assembly factor: a candidate for human complex I deficiency? Hum Genet 110: 264–270.
52. Sun L, Gu S, Sun Y, Zheng D, Wu Q, et al. (2005) Cloning and characterization of a novel human phosphatidic acid phosphatase type 2, PAP2d, with two different transcripts PAP2d_v1 and PAP2d_v2. Mol Cell Biochem 272: 91–96.
53. Ciufo LF, Boyd A (2000) Identification of a lumenal sequence specifying the assembly of Emp24p into p24 complexes in the yeast secretory pathway. J Biol Chem 275: 8382–8388.
54. Kihara A, Sakuraba H, Ikeda M, Denpoh A, Igarashi Y (2008) Membrane topology and essential amino acid residues of Phs1, a 3-hydroxyacyl-CoA dehydratase involved in very long-chain fatty acid elongation. J Biol Chem 283: 11199–11209.
55. Uwanogho DA, Hardcastle Z, Balogh P, Mirza G, Thornburg KL, et al. (1999) Molecular cloning, chromosomal mapping, and developmental expression of a novel protein tyrosine phosphatase-like gene. Genomics 62: 406–416.
56. Fukuda M (1991) Lysosomal membrane glycoproteins. Structure, biosynthesis, and intracellular trafficking. J Biol Chem 266: 21327–21330.
57. Settles AM, Yonetani A, Baron A, Bush DR, Cline K, et al. (1997) Sec-independent protein translocation by the maize Hcf106 protein. Science 278: 1467–1470.
58. Aravind L, Ponting CP (1999) The cytoplasmic helical linker domain of receptor histidine kinase and methyl-accepting proteins is common to many prokaryotic signalling proteins. FEMS Microbiol Lett 176: 111–116.
59. Scheres B, van EF, van der KE, van de WC, van KA, et al. (1990) Sequential induction of nodulin gene expression in the developing pea nodule. Plant Cell 2: 687–700.
60. de Oliveira DE, Seurinck J, Inze D, Van MM, Botterman J (1990) Differential expression of five Arabidopsis genes encoding glycine-rich proteins. Plant Cell 2: 427–436.
61. Ohnishi S, Paakkonen K, Koshiba S, Tochio N, Sato M, et al. (2009) Solution structure of the GUCT domain from human RNA helicase II/Gu beta reveals the RRM fold, but implausible RNA interactions. Proteins 74: 133–144.
62. Burgess DL, Gefrides LA, Foreman PJ, Noebels JL (2001) A cluster of three novel Ca2+ channel gamma subunit genes on chromosome 19q13.4: evolution and expression profile of the gamma subunit gene family. Genomics 71: 339–350.
63. Ouzounis CA, Karp PD (2002) The past, present and future of genome-wide re-annotation. Genome Biol 3: COMMENT2001.
64. Gilks WR, Audit B, de AD, Tsoka S, Ouzounis CA (2002) Modeling the percolation of annotation errors in a database of protein sequences. Bioinformatics 18: 1641–1649.
65. Gilks WR, Audit B, de AD, Tsoka S, Ouzounis CA (2005) Percolation of annotation errors through hierarchically structured protein sequence databases. Math Biosci 193: 223–234.

66. Wu CH, Huang H, Nikolskaya A, Hu Z, Barker WC (2004) The iProClass integrated database for protein functional analysis. Comput Biol Chem 28: 87–96.
67. Bahr A, Thompson JD, Thierry JC, Poch O (2001) BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. Nucleic Acids Res 29: 323–326.
68. Ikeda M, Arai M, Okuno T, Shimizu T (2003) TMPDB: a database of experimentally-characterized transmembrane topologies. Nucleic Acids Res 31: 406–409.
69. Hooft RW, Sander C, Scharf M, Vriend G (1996) The PDBFINDER database: a summary of PDB, DSSP and HSSP information with added value. Comput Appl Biosci 12: 525–529.
70. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658–1659.
71. Wootton JC, Federhen S (1996) Analysis of compositionally biased regions in sequence databases. Methods Enzymol 266: 554–571.
72. Bork P, Koonin EV (1998) Predicting functions from protein sequences–where are the bottlenecks? Nat Genet 18: 313–318.
73. Cuff AL, Sillitoe I, Lewis T, Redfern OC, Garratt R, et al. (2009) The CATH classification revisited–architectures reviewed and new ways to characterize structural divergence in superfamilies. Nucleic Acids Res 37: D310–D314.
74. Schnoes AM, Brown SD, Dodevski I, Babbitt PC (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol 5: e1000605.
75. Ponting CP, Schultz J, Copley RR, Andrade MA, Bork P (2000) Evolution of domain families. Adv Protein Chem 54: 185–244.
76. Ponting CP, Russell RR (2002) The natural history of protein domains. Annu Rev Biophys Biomol Struct 31: 45–71.
77. Copley RR, Letunic I, Bork P (2002) Genome and protein evolution in eukaryotes. Curr Opin Chem Biol 6: 39–45.
78. Hedman M, Deloof H, von HG, Elofsson A (2002) Improved detection of homologous membrane proteins by inclusion of information from topology predictions. Protein Sci 11: 652–658.
79. Anantharaman V, Aravind L (2010) Novel eukaryotic enzymes modifying cell-surface biopolymers. Biol Direct 5: 1.
80. Schultz J (2004) HTTM, a horizontally transferred transmembrane domain. Trends Biochem Sci 29: 4–7.
81. Sonnhammer EL, von HG, Krogh A (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol 6: 175–182.
82. Saier MH, Jr., Tran CV, Barabote RD (2006) TCDB: the Transporter Classification Database for membrane transport protein analyses and information. Nucleic Acids Res 34: D181–D186.
83. Yen MR, Choi J, Saier MH, Jr. (2009) Bioinformatic analyses of transmembrane transport: novel software for deducing protein phylogeny, topology, and evolution. J Mol Microbiol Biotechnol 17: 163–176.
84. Finn RD, Mistry J, Tate J, Coggill P, Heger A, et al. (2010) The Pfam protein families database. Nucleic Acids Res 38: D211–D222.
85. Cserzo M, Eisenhaber F, Eisenhaber B, Simon I (2002) On filtering false positive transmembrane protein predictions. Protein Eng 15: 745–752.
86. Cserzo M, Eisenhaber F, Eisenhaber B, Simon I (2004) TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter. Bioinformatics 20: 136–137.
87. Tusnady GE, Simon I (1998) Principles governing amino acid composition of integral membrane proteins: application to topology prediction. J Mol Biol 283: 489–506.
88. Brendel V, Bucher P, Nourbakhsh IR, Blaisdell BE, Karlin S (1992) Methods and algorithms for statistical analysis of protein sequences. Proc Natl Acad Sci U S A 89: 2002–2006.
89. Kall L, Krogh A, Sonnhammer EL (2004) A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338: 1027–1036.
90. Kall L, Krogh A, Sonnhammer EL (2007) Advantages of combined transmembrane topology and signal peptide prediction–the Phobius web server. Nucleic Acids Res 35: W429–W432.
91. Nielsen H, Engelbrecht J, Brunak S, von Heijne G (1997) A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Int J Neural Syst 8: 581–599.
92. Zar JH (10-18-1998) Biostatistical analysis. Upper Saddle River: Pearson Prentice Hall.
93. Krishna SS, Majumdar I, Grishin NV (2003) Structural classification of zinc fingers: survey and summary. Nucleic Acids Res 31: 532–550.
94. Klug A, Schwabe JW (1995) Protein motifs 5. Zinc fingers. FASEB J 9: 597–604.
95. Iuchi S (2001) Three classes of C2H2 zinc finger proteins. Cell Mol Life Sci 58: 625–635.
96. Leon O, Roth M (2000) Zinc fingers: DNA binding and protein-protein interactions. Biol Res 33: 21–30.
97. Alberts IL, Nadassy K, Wodak SJ (1998) Analysis of zinc binding sites in protein crystal structures. Protein Sci 7: 1700–1716.
98. Ren J, Wen L, Gao X, Jin C, Xue Y, et al. (2009) DOG 1.0: illustrator of protein domain structures. Cell Res 19: 271–273.
99. Johansson H, Eriksson M, Nordling K, Presto J, Johansson J (2009) The Brichos domain of prosurfactant protein C can hold and fold a transmembrane segment. Protein Sci 18: 1175–1182.
100. Shin JI, Shin JY, Kim JS, Yang YS, Shin YK, et al. (2008) Deep membrane insertion of prion protein upon reduction of disulfide bond. Biochem Biophys Res Commun 377: 995–1000.
101. Tompa P, Tusnady GE, Cserzo M, Simon I (2001) Prion protein: evolution caught en route. Proc Natl Acad Sci U S A 98: 4431–4436.
102. Verelst W, Asard H (2003) A phylogenetic study of cytochrome b561 proteins. Genome Biol 4: R38.
103. Ponting CP, Mott R, Bork P, Copley RR (2001) Novel protein domains and repeats in Drosophila melanogaster: insights into structure, function, and evolution. Genome Res 11: 1996–2008.
104. Kageyama-Yahara N, Riezman H (2006) Transmembrane topology of ceramide synthase in yeast. Biochem J 398: 585–593.
105. Nakai T, Yamasaki A, Sakaguchi M, Kosaka K, Mihara K, et al. (1999) Membrane topology of Alzheimer's disease-related presenilin 1. Evidence for the existence of a molecular species with a seven membrane-spanning and one membrane-embedded structure. J Biol Chem 274: 23647–23658.
106. Tie JK, Nicchitta C, von HG, Stafford DW (2005) Membrane topology mapping of vitamin K epoxide reductase by in vitro translation/cotranslocation. J Biol Chem 280: 16410–16416.
107. Ashida H, Hong Y, Murakami Y, Shishioh N, Sugimoto N, et al. (2005) Mammalian PIG-X and yeast Pbn1p are the essential components of glycosylphosphatidylinositol-mannosyltransferase I. Mol Biol Cell 16: 1439–1448.
108. Kota J, Ljungdahl PO (2005) Specialized membrane-localized chaperones prevent aggregation of polytopic proteins in the ER. J Cell Biol 168: 79–88.
109. Zhang L, Ji G (2004) Identification of a staphylococcal AgrB segment(s) responsible for group-specific processing of AgrD by gene swapping. J Bacteriol 186: 6706–6713.
110. Pizarro JC, Vulliez-Le NB, Chesne-Seck ML, Collins CR, Withers-Martinez C, et al. (2005) Crystal structure of the malaria vaccine candidate apical membrane antigen 1. Science 308: 408–411.
111. Xu C, Rice WJ, He W, Stokes DL (2002) A structural model for the catalytic cycle of Ca(2+)-ATPase. J Mol Biol 316: 201–211.
112. Smith LJ, Redfield C, Boyd J, Lawrence GM, Edwards RG, et al. (1992) Human interleukin 4. The solution structure of a four-helix bundle protein. J Mol Biol 224: 899–904.
113. Weston SA, Lahm A, Suck D (1992) X-ray structure of the DNase I-d(GGTATACC)2 complex at 2.3 A resolution. J Mol Biol 226: 1237–1256.
114. Clark GC, Briggs DC, Karasawa T, Wang X, Cole AR, et al. (2003) Clostridium absonum alpha-toxin: new insights into clostridial phospholipase C substrate binding and specificity. J Mol Biol 333: 759–769.
115. McNulty JC, Jackson PJ, Thompson DA, Chai B, Gantz I, et al. (2005) Structures of the agouti signaling protein. J Mol Biol 346: 1059–1070.
116. Krebsbach PH, Lee SK, Matsuki Y, Kozak CA, Yamada KM, et al. (1996) Full-length sequence, localization, and chromosomal mapping of ameloblastin. A novel tooth-specific gene. J Biol Chem 271: 4431–4435.
117. Lunin VV, Dobrovetsky E, Khutoreskaya G, Zhang R, Joachimiak, et al. (2006) Crystal structure of the CorA Mg2+ transporter. Nature 440: 833–837.
118. Weiner JH, Bilous PT, Shaw GM, Lubitz SP, Frost L, et al. (1998) A novel and ubiquitous system for membrane targeting and secretion of cofactor-containing proteins. Cell 93: 93–101.
119. Abendroth J, Rice AE, McLuskey K, Bagdasarian M, Hol WG (2004) The crystal structure of the periplasmic domain of the type II secretion system protein EpsM from Vibrio cholerae: the simplest version of the ferredoxin fold. J Mol Biol 338: 585–596.
120. Albrecht R, Rehling P, Chacinska A, Brix J, Cadamuro SA, Volkmer R, et al. (2006) The Tim21 binding domain connects the preprotein translocases of both mitochondrial membranes. EMBO Rep 7: 1233–1238.