J Struct Funct Genomics
DOI 10.1007/s10969-009-9069-8

A survey of integral α-helical membrane proteins

Libusha Kelly · Ursula Pieper · Narayanan Eswar · Franklin A. Hays · Min Li · Zygy Roe-Zurz · Deanna L. Kroetz · Kathleen M. Giacomini · Robert M. Stroud · Andrej Sali

Received: 16 February 2009 / Accepted: 21 August 2009

Abstract Membrane proteins serve as cellular gatekeepers, regulators, and sensors. Prior studies have explored the functional breadth and evolution of proteins and families of particular interest, such as the diversity of transport-associated membrane protein families in prokaryotes and eukaryotes, the composition of integral membrane proteins, and family classification of all human G-protein coupled receptors. However, a comprehensive analysis of the content and evolutionary associations between membrane proteins and families in a diverse set of genomes is lacking. Here, a membrane protein annotation pipeline was developed to define the integral membrane genome and associations between 21,379 proteins from 34 genomes; most, but not all of these proteins belong to 598 defined families. The pipeline was used to provide target input for a structural genomics project that successfully cloned, expressed, and purified 61 of our first 96 selected targets in yeast. Furthermore, the methodology was applied (1) to explore the evolutionary history of the substrate-binding transmembrane domains of the human ABC transporter superfamily, (2) to identify the multidrug resistance-associated membrane proteins in whole genomes, and (3) to identify putative new membrane protein families.

Keywords Membrane proteins · Superfamily analysis · Multidrug resistance · ABC transporters · Target selection

Abbreviations

CSMP	Center for the Structures of Membrane Proteins
PMT	Pharmacogenetics of Membrane Transporters
ABC	ATP-binding cassette
SLC	Solute carrier
IMG	Integral membrane genome
IMP	Integral membrane protein
TMH	Transmembrane helix
PSSM	Position-specific scoring matrix
DDM	*n*-dodecyl-β-D-maltopyranoside

Introduction

Integral membrane proteins perform a variety of critical cellular functions, such as protecting the cell from

Electronic supplementary material The online version of this article (doi:10.1007/s10969-009-9069-8) contains supplementary material, which is available to authorized users.

L. Kelly
Graduate Group in Bioinformatics, University of California at San Francisco, San Francisco, CA, USA

L. Kelly (✉) · U. Pieper · N. Eswar · D. L. Kroetz · K. M. Giacomini · A. Sali (✉)
Department of Bioengineering and Therapeutic Sciences, Department of Pharmaceutical Chemistry, and California Institute for Quantitative Biosciences, University of California at San Francisco, 1700 4th Street, San Francisco, CA 94158, USA
e-mail: libusha@salilab.org; libusha@gmail.com

A. Sali
e-mail: sali@salilab.org

M. Li · Z. Roe-Zurz · R. M. Stroud
Membrane Protein Expression Center, University of California at San Francisco, San Francisco, CA, USA

F. A. Hays · R. M. Stroud
Department of Biochemistry and Biophysics, University of California at San Francisco, San Francisco, CA, USA

R. M. Stroud
Center for the Structure of Membrane Proteins, University of California at San Francisco, San Francisco, CA, USA

Published online: 17 September 2009

external toxins, acting as the starting point of intracellular signaling cascades, and maintaining critical ion concentrations. They make up approximately 20–30% of an organism's genome [4, 27]. Bioinformatics methods allow us to predict with better than 90% accuracy [12] all α-helical transmembrane proteins in the wealth of sequenced genomes [21].

Membrane proteins are generally difficult to work with experimentally, which leads to incomplete annotation of their functions. Important prior work focusing on membrane protein families includes an analysis of membrane protein fold space [28], the computational analysis of the membrane protein space (CAMPS) database of membrane proteins from bacteria (http://webclu.bio.wzw.tum.de/binfo/proj/camps) [25], the transporter-specific databases TCDB (http://www.tcdb.org/) [33] and TransportDB (http://www.membranetransport.org/) [32], and the G-protein coupled receptor database GPCRDB (http://www.gpcr.org/7tm/) [17]. However, a comprehensive analysis of the content and evolutionary associations between membrane proteins and families in a diverse set of genomes is lacking.

Classification schemes for protein families vary; sequence and structure data can be used separately or in combination with each other to cluster sequences automatically (ProDom [35]) or manually (Pfam-A [11]; SCOP [1]). Less than 1% of the structures in the protein data bank (PDB) are high-resolution structures of membrane-bound proteins [40], providing only sparse data for structure-based functional annotation. In contrast, sequence databases are growing rapidly, thus expanding the universe of known membrane protein families and folds. Improved structural, functional, and evolutionary annotation of membrane proteins is essential to exploit this sequence information.

Here, an automated, sequence- and structure-based approach was taken to define the integral membrane protein set of an organism. In particular, the bioinformatics analysis underpins the Center for Structures of Membrane Proteins (CSMP; http://csmp.ucsf.edu) as well as the Center for Pharmacogenetics of membrane transporters (PMT; http://pharmacogenetics.ucsf.edu).

For CSMP, a list of protein targets was constructed for the structural genomics of membrane proteins (see the accompanying paper [23]). The project took a genome-wide approach, using yeast as a model organism, to tackle as many unique membrane protein families as possible. For PMT, the focus was on identifying and analyzing specifically the ATP-binding cassette (ABC) and solute carrier (SLC) transporter superfamilies. The overall goal is to leverage the cataloging and annotation of human genetic variation in these membrane transporters [22] by determining new structures using the CSMP crystallographic pipeline.

The annotation pipeline was used to provide target input for a structural genomics project that successfully cloned, expressed, and purified 61 of our first 96 selected targets. Furthermore, the methodology was applied to more general problems in membrane protein biology: First, to explore the evolutionary history of the substrate-binding transmembrane domains of the clinically important human ABC transporter superfamily. Second, to identify the multidrug resistance-associated (MDR) membrane proteins in whole genomes and compare the MDR capacity of different organisms. Finally, the pipeline was used to identify putative new membrane protein families.

Materials and Methods

Membrane protein annotation pipeline

The overall process consisted of five steps (Fig. 1).

Step 1: Identification of Pfam membrane protein families

All sequences from the curated Pfam database, Pfam-A [11], were run through the TMHMM2.0 program [19]. Pfam families with sequences containing three or more TMHs were identified as IM protein families.

Step 2: Sequence collection and membrane protein sequence definition

Protein sequences from the genomes of 34 organisms representing each of the three major kingdoms of life were collected from UniProt [38], Ensemble [18], and sequencing projects (Fig. 2). The TMHMM2.0 program [19] was used to predict all sequences with three or more TMHs (queries).

There are many algorithms available for predicting transmembrane helices. A recent analysis of 13 transmembrane helix prediction programs found that four methods, including TMHMM2.0, were consistently high-performing in predicting transmembrane helices in membrane proteins of known structure [3]. TMHMM2.0 was selected for its relative accuracy as well as its speed and the option to install it on our computers, which was essential due to the large number of sequences processed.

The cutoff of three transmembrane helices was chosen to focus on integral membrane proteins rather than monotopic or membrane-associated proteins that may only have membrane anchoring helices or signal peptides. Such a restriction ignores some important classes of integral membrane proteins, such as one- and two-crossing proteins

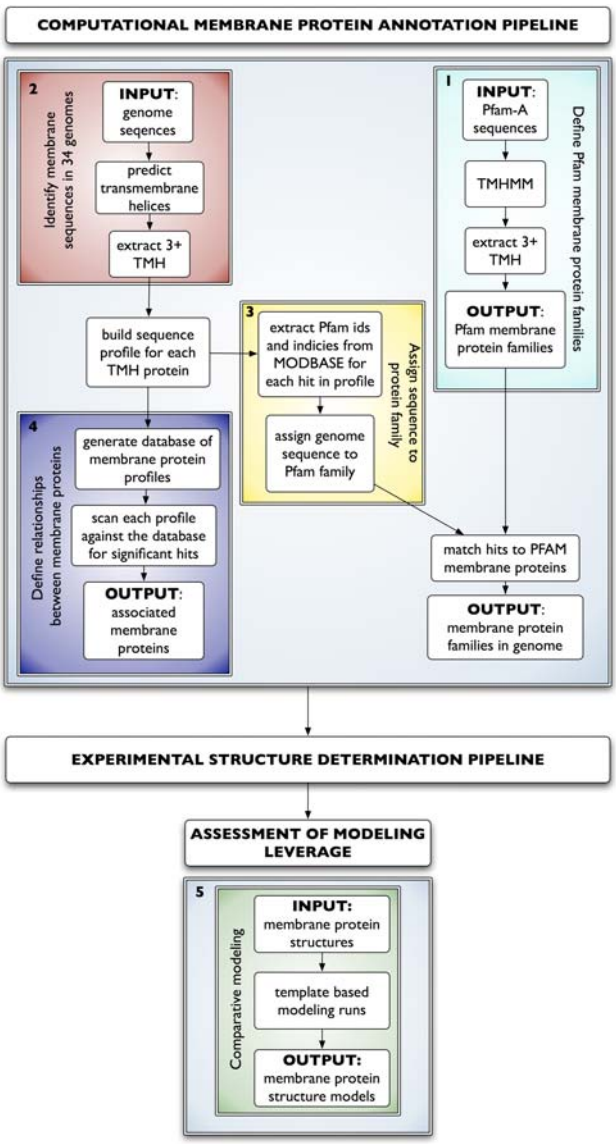

Fig. 1 Membrane protein annotation pipeline. All Pfam-A families with at least three predicted transmembrane helices (TMH) were used to identify membrane protein families in 34 genomes (*cyan*). Sequences predicted to have three or more TMHs in each genome were collected (*red*). In parallel, Pfam family membership was defined where available for each sequence profile (*yellow*). Automated multiple sequence alignment profiles were generated for each sequence. A database of profiles was constructed, and each profile was compared to all other profiles in the database to link membrane proteins (*blue*). The annotation pipeline can be generally used as input to an experimental structure determination pipeline. Finally, resulting structures can be used as templates to generate comparative models for all homologous sequences. The five steps are detailed in Methods

that oligomerize to form channels. However, robust automated identification of secretory signals for eukaryotes remains a difficult problem due to the diversity and lack of characterization of sorting signals to different compartments and incorrect identification of signal peptides as TMHs [9, 10]. Therefore, to restrict our work to integral membrane proteins, proteins with only one or two predicted TMHs were discarded.

Step 3: Sequence profiles, family and taxonomic classification

The automated comparative modeling pipeline MODPIPE was used to construct multiple sequence profiles. Sequences homologous to each of the query membrane proteins in each organism were identified by iteratively searching the UniProt database using the profile.build() module of MODELLER [29] with a threshold e-value of 0.01. Profile.build() uses dynamic programming for aligning profiles against sequences and an empirical definition of statistical significance based on the scores collected during the scan of the database. For each profile, every sequence was linked back to its NCBI taxonomy browser identifier and to Pfam protein family annotations (if available) through the MODBASE [30] database of comparative models. Each query without a Pfam identifier inherited the Pfam identifier of the most similar membrane protein in the same profile (parent); the query-parent alignment had to cover at least 75% of the parent. For some queries, this procedure does not result in a Pfam identifier.

Step 4: Relationships between membrane protein families

To link membrane protein families across the tree of life, a database of the target sequence profiles was generated. For each query, the profile database was scanned using the profile.scan() routine in MODELLER for significant hits, corresponding to e-values smaller than 1. Benchmarking indicates this cutoff results in a true positive rate of 64.5% and a false positive rate of 0.03%. The false positive rate indicates approximately 3 errors per 10,000 matches (Eashwar et al., in preparation). The scans link membrane proteins both within organisms and across organisms.

Step 5: Assessing modeling leverage

The MODWEB comparative modeling server was used to generate models for all sequences homologous to seven CSMP atomic resolution membrane protein structures. MODWEB (http://salilab.org/modweb) is an integral module of the MODBASE (http://salilab.org/modbase) database of comparative models. MODWEB calculates a profile for each identifiable sequence homolog in the UniProt database, followed by modeling these homologs based on detectable templates in the PDB [30].

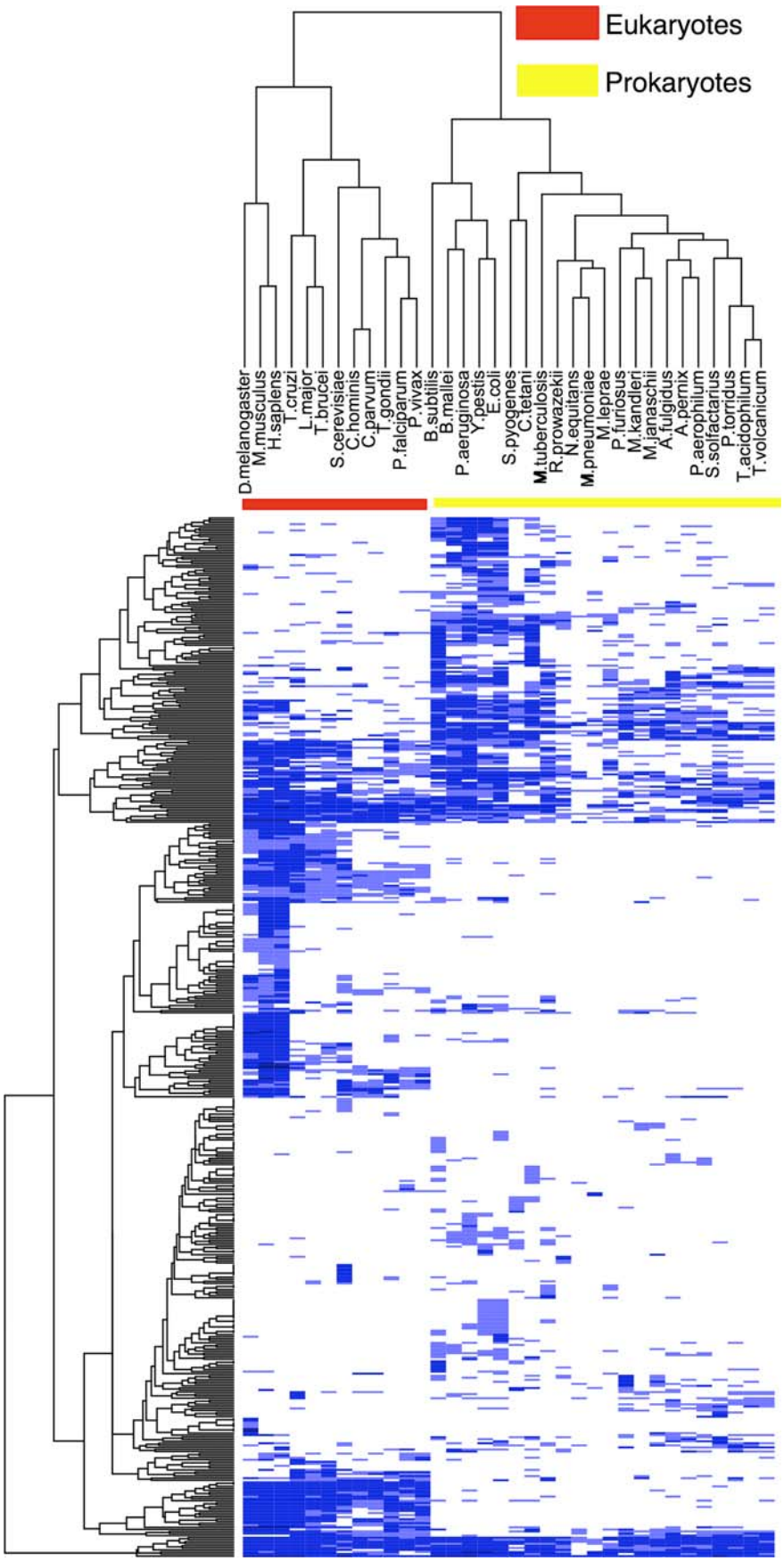

Fig. 2 PFAM membrane protein families in 34 organisms. Organism names are listed horizontally at the top (columns). 476 Pfam membrane protein families are on the vertical axis (rows). Colors indicate binning for the number of times a particular family appears in an organism. *White* indicates a particular family is not found in an organism; *light blue* means the family appears once; *medium blue*, between two and 49 times; and *dark blue* means the family is found 50 or more times. The *red* and *yellow bars* show the clustering of eukaryotes and prokaryotes, respectively. The heatmap was constructed using the R function "heatmap.2" with hierarchical clustering and default parameters [37]

Assessing membrane protein model quality

To ensure that the ZDOPE score, which was developed using a globular protein set, was suitable for use on membrane proteins, *Z*-score distributions for 145 and 36,786 known IM and globular protein structures, respectively (Supplementary Fig. 4) were compared. The averages are −0.63 and −1.56, respectively, suggesting that the *Z*-score is still informative about the accuracy of membrane protein structures. A *Z*-score cutoff of 0 was used to assess models of membrane proteins.

Two-dimensional visualization of membrane protein family associations

Each membrane protein profile is represented as a node. If two profiles are significantly similar, as defined above, an edge is drawn between them. Sets of similar profiles will cluster together because many profiles will be highly linked to each other. Cytoscape is used to visualize the links between membrane protein profiles and the layout algorithm implemented in yFiles organic is used to display the graph [36]. Distances between nodes on the graph are representative of the number of links between the nodes; sets of nodes which all link to each other tend to cluster closely in space.

Identification of unannotated homologs in seven membrane protein families related to multidrug resistance

The annotation pipeline was used to search for sequences in seven families with experimentally established links to multidrug resistance (MDR) in 34 organisms. The Pfam-A families associated with MDR included: three ABC transporter subtypes (PF06472, PF01061, and PF00664); Multidrug and toxic compound extrusion (MATE) transporters (PF01554); small multidrug resistance family transporters (PF00893), Golgi 4-TMH transporters (PF03821), and the AcrBDF family (PF00873) (Supplementary Table 3). All sequences annotated with one of these families were collected and examined for differences in MDR across organisms.

Identifying connections between human ABC transporter transmembrane domains

Most human ABC transporters contain both membrane and globular domains. To examine the evolutionary relationships between the substrate-binding membrane domains, 72 TMDs from 48 human ABC transporters were excised from the complete sequences for each transporter. Sequence profiles were generated for the membrane sequences using profile.build() as described above. The ABC TMD profiles were scanned against the IM database using profile.scan() to identify all related proteins in the 34 organisms.

Identification of new membrane protein families

Sequences that could not be annotated with a Pfam-A membrane protein family were identified. For each of the 4,389 profiles associated with these sequences, all hits in each profile were annotated with an NCBI taxonomy ID and the root taxonomy was recursively determined at the kingdom level. This procedure results in a classification of each profile as either exclusively related to one kingdom (i.e., all hits to the profile are eukaryotic, bacterial, or archaeal) or as some combination of the three kingdoms.

To reduce the likelihood of spurious classification of new membrane protein families, the following restrictions were imposed: First, only single-kingdom classified sequences were considered when defining new families. All single-kingdom sequences were examined for hits to the IM database and clusters of sequences with more than four members were considered putative new families. Second, profile-profile alignments were calculated for all members in each putative cluster the overlap in each alignment of predicted TMHs was computed. To be defined as a new family, at least half of the TMHs in a profile had to have better than 60% coverage in all profile-profile alignments of the target against all other targets in the cluster. Sequences representing new families were run against the Pfam-A and Pfam-B databases to determine if either database had a corresponding family. Pfam-B is an automatically generated, uncurated clustering of domains identified using the ADDA program [15].

Results

Membrane protein annotation pipeline

The annotation pipeline consists of five main steps (Fig. 1, Methods). In the first step, we started with 1,779,528 sequences in Pfam-A (01/09/07) [11]. Of these, 172,079 are predicted by TMHMM [21] to have one or more transmembrane α-helices (TMHs); 99,937 have three or more TMHs. Because of the difficulty of accurately identifying signal peptides and possible errors in TMH prediction, only integral membrane protein sequences predicted to have at least three TMHs. These sequences belong to 598 unique Pfam families (Supplementary Table 1) were selected, of which 476 appear in at least one of the 34 organisms of interest (Fig. 2). Organisms were selected based on the following considerations:

completeness of the genome, model organism, relevance to human disease, diversity within each kingdom, and the availability of genomic DNA for cloning and expression. The 122 families with no representatives include: photosystem-related families (e.g., PF00421) that are not found in any of the selected organisms, families that are only found in a single organism (e.g., PF03303), and families with no characterized function (e.g., PF06836 and PF07099).

In the second step, sequences for the 34 genomes were collected from UniProt, Ensemble and organism-specific sequencing projects. In total, there were 21,379 proteins predicted to have at least three TMHs, corresponding to the "integral membrane genomes" (IMGs) of the 34 organisms (Supplementary Fig. 1). We were able to annotate between 41% (*Plasmodium falciparum*) and 93% (*Mus musculus*) of each IMG with Pfam family IDs; 16 of the genomes had more than 25% of their IMG unannotated, suggesting that there are many undiscovered membrane protein families.

In the third step, for each protein, an automated multiple sequence alignment of the corresponding superfamily was generated, represented by a position-specific scoring matrix (PSSM), and added to a database, which we call the integral membrane (IM) database (http://salilab.org/projects/integral_membrane_proteins/34pssmdb.txt.gz).

In the fourth step, each PSSM was compared to the whole IM database, using profile-profile alignments to identify related proteins.

The fifth step estimated the impact of newly solved atomic-resolution structures on the coverage of membrane protein sequence space. While it would be ideal to use crystal structures resulting from the target selection pipeline described here, our experimental pipeline based on yeast expression has not yet yielded any structures. To demonstrate the utility of this final step in the pipeline, we therefore use other recently solved membrane protein structures that were not part of our yeast target selection scheme. As of October 2008, the CSMP determined crystallographic structures for the following seven integral membrane proteins: the *Escherichia coli* Ammonia Channel Amtb (PDB ID: 2ns1), the *Nitrosomonas europaea* Rh50 Ammonium transport protein (PDB ID: 3bhs), two structures of *E. coli* lactose permease (PDB IDs: 2cfq, 2v8n), the archaeal aquaporin AqpM (PDB ID: 2f2b), a mutant structure of *E. coli* AqpZ (PDB ID: 2o9d), and an aquaglyceroporin from *Plasmodium falciparum* (PDB ID: 3c02).

These seven structures fall into three Pfam protein families: the ammonia channel and ammonium transport protein are members of the PF00909 ammonium transporter family, the two aquaporins and the aquaglyceroporin are members of the PF00230 major intrinsic protein (MIP) family, and lactose permease is a member of the PF01306 LacY proton/sugar transporter family. The structures were used as input to the final step of the computational pipeline to calculate how many sequences can be modeled based on these structures (i.e., the modeling leverage).

To assess the value of each new CSMP template for comparative modeling, models for sequences that could be modeled using both a CSMP structure and any non-CSMP membrane structure as templates were compared. To take partially modeled sequences into account, the comparison is performed at the residue level. Additionally, a "transmembrane region" was defined for each of the CSMP template structures that included all amino acid residues from the first TMH residue to the last TMH residue.

There were a total of 178,627 sequences for all membrane template-based modeling calculations, 13,317 of which were sequences based on CSMP templates. Of this set, 11,240 sequences with 1,108,633 residues in transmembrane regions could be modeled using both a CSMP template and another template. 18% of the residues (199,684) were modeled with higher target-template sequence identity with a CSMP template than any other available membrane protein structure, demonstrating the value of these additional structures for comparative modeling of membrane proteins. For individual models, the lactose permease structures had the best modeling leverage, with 24% of residues modeled with higher target-template sequence identity using the CSMP structures 2v8n and 2cfq. The aquaporin and aquaglyceroporin structures (2f2b, 2o9d, and 3c02) had less impact, with 8% of residues modeled the best with the CSMP template (Supplementary Table 2).

Calculating the impact of new membrane protein structures on coverage of membrane protein sequence space will aid in assessing target selection efforts by structural genomics consortia. Furthermore, this modeling approach is applicable to any new membrane protein structure.

Membrane protein family distribution in the three kingdoms of life

All sequences representing membrane protein families from each genome were collected and the number of times each family appeared in each genome was counted. Counts were assembled into a matrix (http://salilab.org/projects/integral_membrane_proteins/memb_counts.txt.gz). The counts ranged from 0 counts of a family in an organism to 1,468 for rhodopsin-like GPCRs in the mouse genome, demonstrating that some families are highly represented in multiple genomes and others are rare or restricted to only a few organisms. There are 13,139, 2,079, 1,956, and 30 families with 0, 1, 2–49, and 50–1468 representatives, respectively.

Target selection for the structural genomics of integral membrane proteins in yeast

Two subsets of target proteins for structural studies were selected. First, we aimed to maximize the coverage of the *Saccharomyces cerevisiae* IMG while minimizing the number of targets for expression. Second, we also selected a number of targets to further PMT's functional and clinical studies of ABC and SLC membrane transporters in drug disposition.

Target selection for sequence leverage

Pfam annotations were used to cover all membrane protein families in yeast and the associations between multiple sequence profiles were used to select sequences that are absent from Pfam (Methods). There are 621 predicted IM sequences in yeast and these were the input to our computational annotation pipeline. Of these, 490 sequences could be annotated with 165 unique Pfam membrane protein families and 131 could not be annotated with a Pfam identifier. Of the 165 annotated families, 79 were represented by a single sequence, meaning the family appeared only once in the yeast genome.

The 79 singletons initiated our target list. For the remaining 83 annotated families, two sequences were selected from each family to improve the likelihood of successful structural characterization for that family. These two members were selected to ensure optimal coverage of each family (Methods), which is especially important for larger families. For example, the major facilitator family (MFS) has 57 sequences, the most of any membrane protein family in yeast. The MFS sequences fall into two major clusters, one with 44 MFS members and one with six. VBA1_YEAST, which is associated with 24 MFS-annotated sequences in the first cluster (55%) and MCH4_YEAST, which is associated with five MFS sequences in the second cluster (83%) were selected.

Of the 131 unannotated sequences, 16 were in two completely unannotated clusters of 8 sequences each, 62 hit no other sequences, six sequences fell in two unannotated clusters of three sequences each, and 14 fell into seven clusters of two sequences each. The remaining 33 sequences were associated with at least one other annotated sequence and were discarded. Because two sequences were selected from each unannotated cluster, there are an additional 98 targets. Complete coverage of the yeast genome therefore requires 347 targets out of the 621 IMG proteins. If a target fails in any stage of the experimental process, a similar yeast target can be selected for a subsequent trial [23].

The results of our computational annotation pipeline were entered into an experimental structure determination pipeline, as detailed in the Results and Discussion and a companion paper [23].

Target selection for biological significance

Two of the targets, the yeast genes STE6 and YN_99, code for ATP-binding cassette transporters that are homologous to human multidrug transporters in the B and G families, respectively. There are 48 characterized ABC transporters in the human genome and 18 are disease-associated [7, 8, 26]. There are many atomic structures available for isolated nucleotide binding domains (NBDs) from ABC transporters, and these structures have been successfully used to assess the role of interface-disrupting point mutants with clinical phenotypes in human ABC transporters [20]. However, a molecular level understanding of the clinical impact of genetic variation requires high-resolution structural data for the substrate-binding transmembrane domains (TMDs) of these proteins, providing the rationale for their inclusion into the target list.

In humans, ABCB1 (also known as MDR1) and other members of the B family, such as ABCB4 (MDR3) and ABCB11 (BSEP), are associated with multidrug resistance in cancer therapy. ABCB4 and ABCB11 are also associated with several forms of cholestasis [13]. Our collaborators at the PMT have identified 29 non-synonymous single nucleotide polymorphisms in these proteins. The STE6 structure would be particularly useful for structural modeling of sequence variations in humans because the domain organization of two TMDs and two nucleotide-binding domains NBDs is the same as in the human transporters ABCB1, ABCB11 and ABCB4 (Supplementary Fig. 2). The most similar structurally characterized homolog of the ABCB family is currently the *Staphylococcus aureus* transporter Sav1866 [5]. This transporter has only a single TMD and a single NBD that forms a homodimer; thus, it is not an ideal template for modeling the four domain multidrug resistance-associated transporters from the ABCB family.

The yeast nucleoside transporter target YAL022C (FUN26) [39] is homologous to the equilibrative nucleoside transporters ENT1 (SLC29A1) and ENT2 (SLC29A2). The PMT has identified two non-synonymous SNPs in SLC29A1as well as two non-synonymous SNPs and seven insertion/deletion mutations in SLC29A2 [22].

Additional structural data from these transporter families will be invaluable for interpreting the results of functional studies and suggesting molecular mechanisms for clinical phenotypes.

The final set of 384 targets was entered into the structural characterization pipeline of the CSMP [23]. Of these targets, 273 are significantly related to at least one human gene. In all, 1,249 human sequences are significantly

similar to the 273 yeast sequences, suggesting that about 40% of the human IMG has a corresponding gene in yeast (Supplementary Fig. 3a, b). Our clustering of the yeast IMG is generally in agreement with the manual "clans" clustering in Pfam (Supplementary Fig. 3c) [11].

Defining the scope of membrane protein structural genomics

The scope of structural genomics of membrane proteins is the number of target structures needed to achieve some desired coverage of the membrane protein sequence space. Current comparative modeling coverage of integral membrane protein sequences in UniProt [38] was examined first. Next, the total number of structures required for desired sequence coverage of the 598 Pfam integral membrane protein families described above was calculated.

The ModBase database [30] contains 806,266 models of 1,733,721 sequences the complete UniProt (as of 6/1/2005) for which the target-template identity is at least 30%. Of these, 61,749 models for 55,161 unique sequences are predicted by TMHMM to contain at least three TMHs. This estimate suggests that domains in only approximately 8% of integral membrane proteins can be currently modeled at reasonable accuracy (implied by the 30% target-template sequence identity) using available template structures (Supplementary Fig. 4).

To improve the coverage, it would be ideal to select sequences for structural characterization that yielded the greatest improvement in the number of modelable sequences based on the 598 Pfam integral membrane families. At 30% sequence identity, the 375,155 sequences in these families fall into 13,395 clusters. Thus, a representative structure from such a cluster provides a reasonable template for comparative modeling of the other sequences in its cluster. Using a target selection strategy where sequences from the largest clusters are selected for structural characterization first, 90% of the sequences in the currently known integral membrane families could be covered by 2,454 structures. In contrast, a random selection of crystallographic targets would require approximately eight times more structures (i.e., 20,000) to achieve the same coverage. For 70% coverage of sequence space, a more realistic goal, the ranking by cluster size requires 504 structures *versus* 2,500 for the random selection (Fig. 3).

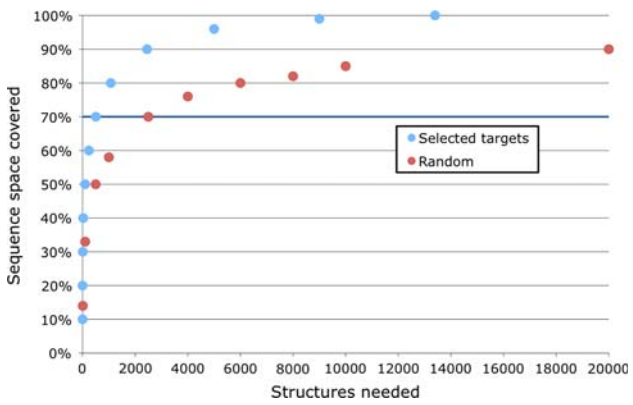

Fig. 3 Target selection for membrane protein structural genomics. Structural coverage of the known IMG (Integral Membrane Genome) sequence space was defined by taking sequences from 598 IM Pfam families and clustering them at 30% sequence identity

Applications of the membrane protein annotation pipeline

Identification of unannotated homologs in seven membrane protein families related to multidrug resistance

In the 34 genomes, 793 sequences were annotated as coming from one of seven Pfam-A families with experimentally established links to multidrug resistance (MDR). Of these sequences, 292 were not described by either the "Protein name" field in UniProt or the "DEFINITION" field in Genbank as MDR-related, but rather with descriptions such as "conserved membrane protein" or "uncharacterized protein".

Between 2% (mouse) and 8% (*Mycobacterium tuberculosis)* of the IMG of each organism is devoted to MDR. Furthermore, pathogenic organisms tend to have higher percentages of MDR membrane proteins in their genomes. For example, the pathogens *M. tuberculosis, Cryptosporidium parvum*, *Cryptosporidium hominis*, *Pseudomonas aeruginosa*, and *Leishmania major*, and the obligate parasite *Mycoplasma pneumoniae* all had more than 5% of their IMGs devoted to MDR.

Tracing the evolutionary history of human ABC transporters

ABC transporters are found in all three kingdoms of life. These proteins couple ATP binding, hydrolysis, and release to substrate transport across a membrane. They share a common architecture consisting of combinations of transmembrane domains (TMDs) and nucleotide-binding domains (NBDs). While the NBDs are well conserved, the TMDs, which contain the substrate binding sites, are more divergent.

The 72 TMDs in the human ABC transporter superfamily were associated with 669 unique sequences in the 34 organisms. In total, there were 16,503 connections between the human TMDs and sequences in the IM database (Fig. 4).

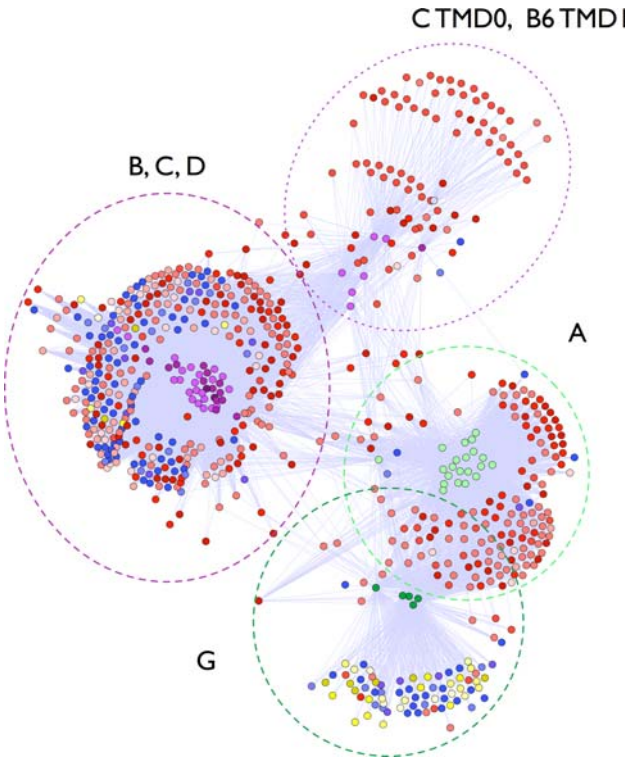

Fig. 4 Links between the transmembrane domains of human ABC transporters. The transmembrane domains of 48 human ABC transporter proteins were excised from their complete sequences. Profiles were generated for each transmembrane domain and run against the membrane protein profile database (Methods). Significantly related profiles are linked and colored according to organism with *red*, *blue*, and *yellow* representing eukaryotes, bacteria, and archaea, respectively. The two major clusters represent ABCA and G family members (*dark* and *light green*); and ABCB, C, and D family members (*purple*)

Identification of new membrane protein families

Finally, the analysis suggests that there exist additional unidentified membrane protein families. Out of the 21,385 sequences of membrane proteins in the selected genomes, 4,389 (21%) could not be annotated with a Pfam membrane protein family.

Of the 51 putative new families, 27 and 16 had one or more Pfam-B identifiers, respectively (11/10/08). Because groups of TMHs may act as a functional unit, a family definition needs to cover as long a stretch of conserved TMHs as possible; our analysis extends the membrane region coverage of 10 Pfam-A families. For example, in Eukaryotic cluster 14, the analysis indicates a conserved group of 3 TMHs, whereas the Pfam-A family hit Mpv17_PMP22 (PF04117) covers either one or two of the TMHs. Furthermore, the latest version of Pfam-A now includes new families, such as DuoxA (PF10204), Tmp39 (PF10271), and DUF2453 (PF10507) that each correlate with one of our newly identified eukaryotic families. Finally, Bacterial Cluster 4 has no Pfam-A or B classification in the conserved membrane region.

Discussion

Membrane protein family distribution in the three kingdoms of life

Simply using the family compositions of the IMGs, eukaryotic and prokaryotic organisms are clearly distinguishable (Fig. 2). This discrimination was robust with respect to binning and may be useful for identifying the kingdoms of the source organisms in large-scale environmental sequencing projects. Furthermore, some limited information on how organisms communicate with the environment is recovered. Two organisms, the bacterium *Mycoplasma pneumoniae* and the hyperthermophilic archaeon *Nanoarchaeum equitans*, cluster with each other rather than with their respective kingdoms. These organisms are obligate parasites, dependent on a host for survival, with small genomes and highly permeable membranes, thus presumably requiring similar IMGs [14, 24].

Target selection for the structural genomics of integral membrane proteins in yeast

Our annotation pipeline provided a comprehensive set of sequences covering all IM protein families in yeast as input to a structural genomics project. Despite high failure rates in membrane protein crystallography, 23 targets out of the first set of 96 targets were rapidly identified that expressed, were fully soluble in DDM, and were within the included volume of a size exclusion column. The first five of these targets were subsequently demonstrated to be stable within the assigned buffer and easily purified using established protocols. Finally, three of the top targets crystallized readily from standard sparse matrix screens [23].

A 24% return of identified targets from the original starting subset of 96 targets demonstrates that our combined computational and experimental pipeline can be successfully used to identify and prioritize eukaryotic integral membrane proteins for downstream crystallization and functional characterization. Future improvements could include considering a number of protein attributes, such as presence of disordered regions or protein interaction partners correlate with reduced odds of purification and crystallization.

One goal of structural genomics is to increase the number and variety of sequences that can be modeled with useful accuracy by comparative modeling [2]. Finding sequences that can be modeled based on a given template structure is the first step in this modeling process. The

multiple sequence profile for each target is used as a proxy for how many sequences could have at least three predicted TMHs modeled based on at least 30% sequence identity to the template structure (i.e., sequence leverage) [34]. The structures of our 384 selected yeast targets would enable the modeling of 63,584 UniProt sequences based on at least 30% sequence identity between the selected target and its homologs. Of these, 18,633 sequences were predicted to have at least three TMHs in the model. Thus, the sequence leverage of the 384 targets is 18,633. For comparison, 384 randomly selected yeast membrane protein structures would enable the modeling of a similar number of membrane sequences, but on average would cover 34 fewer families out of 162 total identified Pfam families in yeast.

Defining the scope of membrane protein structural genomics

Despite their relevance to human health and importance in cellular gatekeeping, regulation and sensing, the number of solved, high-resolution structures of membrane proteins is low. There are currently 94 α-helical integral membrane protein structures with less than 95% sequence identity to each other from 37 Pfam families with at least three TMHs. Of the structures solved, some, like the bacterial rhodopsins and the aquaporins, share the same fold. This dearth of structures and lack of diversity contribute to incomplete annotation of membrane protein sequences. Additional coordination in target selection for the structural genomics of membrane proteins, such as the selection proposed here, would greatly facilitate comprehensive accounting of membrane protein families and contribute to their functional annotation.

Other recent work predicted the number of structures required to cover membrane protein sequence space. Approximately 80% of membrane protein sequence space represented by sequences from 95 genomes can be organized into approximately 700 families (ignoring families with singleton members) [28]. A study of bacterial proteins predicted by TMHMM2.0 to have at least three TMHs suggested that 242 new structures would provide structural coverage at the fold level of approximately 70% of sequences belonging to the most populated prokaryotic membrane protein families [25].

For 80% coverage of membrane protein sequence space, our analysis requires approximately 1073 structures; this higher number is likely a consequence of including all genomes represented in Pfam instead of only 95 select genomes and highlights the importance of including as many identifiable protein families as possible. Our analysis suggests that 504 structures would cover 70% of membrane protein sequence space; this larger number is likely a result of including eukaryotic proteins in our analysis. The difficulty of membrane protein crystallization makes it likely that it will be many years before the goal of complete structural coverage of membrane protein families is achieved.

Application of the membrane protein annotation pipeline

Identification of multidrug resistance associated transporters

Identifying binding sites and predicting substrates for membrane proteins can be a more difficult problem than for their globular counterparts because they often lack clear surface features, such as pockets and grooves, which suggest binding sites. Many membrane proteins, including MDR transporters, are multispecific, transporting a variety of dissimilar substrates. In the case of a *Haemophilus influenzae* ABC transporter, the endogenous substrate of the protein remains unknown, even with a high-resolution crystal structure [31]. The combination of sequence-based evolutionary information with more diverse membrane protein structures will be a powerful tool for identifying putative substrates of uncharacterized membrane protein sequences, for example by computational ligand docking followed by experimental validation [16].

Tracing the evolutionary history of human ABC transporters

The sequence profile connections between the TMDs in human ABC transporters reveal two major clusters of sequences. The TMDs from ABC transporter families A and G clustered together and that families B, C, and D also clustered together (Fig. 4). Prior sequence analysis found that the nucleotide binding domains also cluster identically [7, 8]. This congruence suggests that the TMDs and NBDs of the human ABC transporters evolved from a common ancestor with both domains intact on the polypeptide chain, rather than evolving separately and then joining later in the evolution of the protein.

Next, the taxonomy of associated clusters of TMDs was analyzed to identify the breakdown of subfamilies by kingdom. The ABCG family had a significantly higher representation of archaeal sequences than either the B, C, D cluster or the A cluster (p-value $= 2 \times 10^{-12}$, 2-sample test for equality of proportions). The ABCA family had no archaeal hits at all and few bacterial hits (9 out of 199 total hits). This difference in taxonomic representation suggests a possibly more ancient origin for the ABCG family and a more recent origin for the ABCA family.

The ABCA family has undergone many gene duplication and loss events [6] and the complement of 12 ABCA

genes in the human genome may primarily transport molecules endogenous to eukaryotic organisms. One example is the photoreceptor cell-specific ABCA4 gene [6]. In contrast, the more taxonomically diverse sterol-transporting ABCG genes may represent more general mechanisms of lipid distribution. Due to their promiscuity, the full range of the ABC transporter substrates is currently not known. The membrane domain alignments described here will be used to suggest likely overlapping substrate specificities among these clinically relevant human transporters and their orthologs in other organisms.

Identification of new membrane protein families

While many of the clusters produced by the annotation pipeline had Pfam-B identifiers, Bacterial Cluster 4 has no Pfam-A or B classification in the conserved membrane region. The members in this cluster come from the pathogenic gram-negative bacteria *Pseudomonas aeruginosa*, *Burkholderia mallei*, *Yersinia pestis,* and *Rickettsia prowazekii.* The sequences contain between nine and 11 TMHs. The conserved region covers eight TMHs. Two of the sequences contain a non-transmembrane domain, Wzy_C (PF04932), that is involved in the synthesis of a lipopolysaccharide O-antigen found in the outer membrane of gram-negative bacteria. It is possible that the conserved membrane region found in this cluster is involved in a set of the transport process of synthesized lipopolysaccharides to the outer membrane in gram-negative bacteria.

The validity of this analysis is demonstrated by the agreement of our classification with both Pfam-A and B, alternative classifications that were done using different methods; the reclassification in this analysis to include complete transmembrane regions in putative families; and the identification of a bacterial family possibly involved in lipopolysaccharide synthesis and transport.

Conclusion

A membrane protein annotation pipeline was developed to define the integral membrane genome and associations between 21,379 proteins from 34 genomes; most, but not all of these proteins belong to 598 defined families. The annotation pipeline was used to provide target input for a structural genomics project that successfully cloned, expressed, and purified 61 of our first 96 selected targets. Furthermore, the methodology was applied to unsolved problems in membrane protein biology, including exploring the evolutionary history of the substrate-binding transmembrane domains of the human ABC transporter superfamily, locating sets of multidrug resistance-associated membrane proteins in whole genomes, and identifying putative new membrane protein families. While most predicted membrane proteins are assigned to an annotated protein family, a quarter or more of the membrane proteins in 16 of the 34 studied organisms remain unassigned, highlighting a need for large-scale structural and functional characterization of membrane proteins.

Acknowledgments We are grateful to Ms. Mutsuko Yamada and Drs. David Eramian, Jason Gow, Leslie Chinn and Mark Breidenbach for helpful discussions. This work was supported by NIH (U54 GM074929, RMS, KMG, and AS; U01 GM61390, KMG and AS; P50 GM073210, RMS; F32 GM072403, FAH; and R01 GM54762, AS), the Sandler Family Supporting Foundation (AS, FAH), Hewlett-Packard, NetApps, IBM, and Intel (AS).

References

1. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG (2008) Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res 36:D419–D425
2. Baker D, Sali A (2001) Protein structure prediction and structural genomics. Science 294:93–96
3. Cuthbertson JM, Doyle DA, Sansom MS (2005) Transmembrane helix prediction: a comparative evaluation and analysis. Protein Eng Des Sel 18:295–308
4. Daley DO, Rapp M, Granseth E, Melen K, Drew D, von Heijne G (2005) Global topology analysis of the Escherichia coli inner membrane proteome. Science 308:1321–1323
5. Dawson RJ, Locher KP (2006) Structure of a bacterial multidrug ABC transporter. Nature 443:180–185
6. Dean M, Annilo T (2005) Evolution of the ATP-binding cassette (ABC) transporter superfamily in vertebrates. Annu Rev Genomics Hum Genet 6:123–142
7. Dean M, Hamon Y, Chimini G (2001) The human ATP-binding cassette (ABC) transporter superfamily. J Lipid Res 42:1007–1017
8. Dean M, Rzhetsky A, Allikmets R (2001) The human ATP-binding cassette (ABC) transporter superfamily. Genome Res 11: 1156–1166
9. Elofsson A, von Heijne G (2007) Membrane protein structure: prediction versus reality. Annu Rev Biochem 76:125–140
10. Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2:953–971
11. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A (2006) Pfam: clans, web tools and services. Nucleic Acids Res 34:D247–D251
12. Fleishman SJ, Ben-Tal N (2006) Progress in structure prediction of alpha-helical membrane proteins. Curr Opin Struct Biol 16:496–504
13. Gillet JP, Efferth T, Remacle J (2007) Chemotherapy-induced resistance by ATP-binding cassette transporter genes. Biochim Biophys Acta 1775:237–262
14. Hasselbring BM, Jordan JL, Krause RW, Krause DC (2006) Terminal organelle development in the cell wall-less bacterium Mycoplasma pneumoniae. Proc Natl Acad Sci USA 103:16478–16483

15. Heger A, Holm L (2003) Exhaustive enumeration of protein domain families. J Mol Biol 328:749–767
16. Hermann JC, Marti-Arbona R, Fedorov AA, Fedorov E, Almo SC, Shoichet BK, Raushel FM (2007) Structure-based activity prediction for an enzyme of unknown function. Nature 448:775–779
17. Horn F, Bettler E, Oliveira L, Campagne F, Cohen FE, Vriend G (2003) GPCRDB information system for G protein-coupled receptors. Nucleic Acids Res 31:294–297
18. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Ouverdin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E (2007) Ensembl 2007. Nucleic Acids Res 35:D610–D617
19. Kall L, Krogh A, Sonnhammer EL (2004) A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338:1027–1036
20. Kelly L, Karchin R, Sali A (2007) Protein interactions and disease phenotypes in the ABC transporter superfamily. Pac Symp Biocomput 5:1–63
21. Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305: 567–580
22. Leabman MK, Huang CC, DeYoung J, Carlson EJ, Taylor TR, de la Cruz M, Johns SJ, Stryke D, Kawamoto M, Urban TJ, Kroetz DL, Ferrin TE, Clark AG, Risch N, Herskowitz I, Giacomini KM (2003) Natural variation in human membrane transporter genes reveals evolutionary and functional constraints. Proc Natl Acad Sci USA 100:5896–5901
23. Li M, Hays FA, Roe-Zurz Z, Vuong L, Kelly L, Ho C-M, Robbins R, Pieper U, Oconnell J, Miercke L (2008) Selecting optimum eukaryotic integral membrane proteins for structure determination by rapid expression and solubilization screening. J Mol Biol 385:820–830
24. Makarova KS, Koonin EV (2005) Evolutionary and functional genomics of the Archaea. Curr Opin Microbiol 8:586–594
25. Martin-Galiano AJ, Frishman D (2006) Defining the fold space of membrane proteins: the CAMPS database. Proteins 64:906–922
26. McKusick-Nathans Institute of Genetic Medicine, (J. H. U. B., MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD) (2008)
27. Miller JP, Lo RS, Ben-Hur A, Desmarais C, Stagljar I, Noble WS, Fields S (2005) Large-scale identification of yeast integral membrane protein interactions. Proc Natl Acad Sci USA 102: 12123–12128
28. Oberai A, Ihm Y, Kim S, Bowie JU (2006) A limited universe of membrane protein families and folds. Protein Sci 15:1723–1734
29. Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS, Rossi A, Marti-Renom M, Karchin R, Webb BM, Eramian D, Shen MY, Kelly L, Melo F, Sali A (2006) MODBASE: a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res 34:D291–D295
30. Pieper U, Eswar N, Webb BM, Eramian D, Kelly L, Barkan DT, Carter H, Mankoo P, Karchin R, Marti-Renom MA, Davis FP, Sali A (2009) MODBASE, a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res 37:D347–D354
31. Pinkett HW, Lee AT, Lum P, Locher KP, Rees DC (2007) An inward-facing conformation of a putative metal-chelate-type ABC transporter. Science 315:373–377
32. Ren Q, Chen K, Paulsen IT (2007) TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. Nucleic Acids Res 35:D274–D279
33. 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
34. Sali A (1998) 100,000 protein structures for the biologist. Nat Struct Biol 5:1029–1032
35. Servant F, Bru C, Carrere S, Courcelle E, Gouzy J, Peyruc D, Kahn D (2002) ProDom: automated clustering of homologous domains. Brief Bioinform 3:246–251
36. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504
37. R Development Core Team (2008) R: a language and environment for statistical computing
38. UniProt Consortium (2007) The Universal Protein Resource (UniProt). Nucleic Acids Res 35:D193–D197
39. Vickers MF, Yao SY, Baldwin SA, Young JD, Cass CE (2000) Nucleoside transporter proteins of Saccharomyces cerevisiae. Demonstration of a transporter (FUI1) with high uridine selectivity in plasma membranes and a transporter (FUN26) with broad nucleoside selectivity in intracellular membranes. J Biol Chem 275:25931–25938
40. White S (2007) http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html