Acta Pharmacologica Sinica (2014) 35: 11–23
 1671-4083/14
www.nature.com/aps

Open

Review

Deorphanizing the human transmembrane genome: A landscape of uncharacterized membrane proteins

Joseph J BABCOCK, Min LI*

The Solomon H Snyder Department of Neuroscience, High Throughput Biology Center and Johns Hopkins Ion Channel Center (JHICC), School of Medicine, Johns Hopkins University, Baltimore, MD 21205, USA

The sequencing of the human genome has fueled the last decade of work to functionally characterize genome content. An important subset of genes encodes membrane proteins, which are the targets of many drugs. They reside in lipid bilayers, restricting their endogenous activity to a relatively specialized biochemical environment. Without a reference phenotype, the application of systematic screens to profile candidate membrane proteins is not immediately possible. Bioinformatics has begun to show its effectiveness in focusing the functional characterization of orphan proteins of a particular functional class, such as channels or receptors. Here we discuss integration of experimental and bioinformatics approaches for characterizing the orphan membrane proteome. By analyzing the human genome, a landscape reference for the human transmembrane genome is provided.

Keywords: bioinformatics; human genome; membrane protein; orphan protein; ion channel; receptor; transporter; enzyme; drug target

Acta Pharmacologica Sinica (2014) 35: 11–23; doi: 10.1038/aps.2013.142; published online 18 Nov 2013

Introduction

The availability of the human genome sequence[1, 2] provided the blueprint for the diverse elements encoding the proteome. The exciting opportunity of comprehensively deciphering the function of these sequences remains a challenge. Traditionally, translating knowledge of a linear nucleic acid (amino acid) sequence into mechanistic insights requires a mixture of phenotypes obtained through genetic investigation, reconstituted biochemical assays, and structural determination. Though for any gene these studies may prove technically challenging, they are particularly so for membrane proteins at the cell surface or in intracellular organelle bilayers. Membrane proteins include receptors, ion channels, transporters, and enzymes. Constituting a significant fraction (20%–30%) of human genes[3], membrane proteins represent the targets of over half of known drugs[4, 5]. As the lipid membrane of the cell constitutes only 6%–12% of the cytosolic volume, with the plasma membrane representing only 2%–5% of this total[6], the biochemical environment necessary for transmembrane protein function is highly specialized. Furthermore, the chemical compositions of the two sides of the membrane are physiologically different; a membrane protein is thus theoretically situated in three biochemically distinct environments. In addition to the critical requirement of lipid environment, like soluble proteins functional characterization of membrane proteins faces other challenges including functional redundancy, macromolecular organization and dependence on physiological conditions[7–14].

Because of these challenges of characterizing a membrane protein, studies to understand the role of novel genes would benefit from the ability to narrow the potential number of candidates. In one scenario, molecular determinants are sought for a specific physiological process or disease phenotype that is hypothesized to involve membrane receptors, such as ion flux. Here, 'de-orphanizing' involves finding those genes whose presence or function correlates with this phenotype, through reverse genetics, transcriptional profiling, and other methods[15–17]. Alternatively, the phenotype of interest may not be known beyond a general category such as ion channel, and the challenge is to identify a plausible collection of uncharacterized genes that may share general functional similarity with known families[18, 19]. Consequently, 'de-orphanization' may also involve identifying the native ligand for novel receptors, ionic substrate for orphan channels or transporters, and physiological protein-protein interactions. Thus, it aids in the definition of their functional phenotype[20–22]. In all cases, it is helpful to leverage data on the functionally characterized portion of the genome to infer the biological roles of the unannotated set based on existing information. Traditionally, this idea is demonstrated by the use of nucleic acid (amino acid) sequence

* To whom correspondence should be addressed.
E-mail minli@jhmi.edu
Received 2013-07-28 Accepted 2013-09-08

similarity to infer possible functional homology. A popular heuristic algorithm for this problem is the Basic Local Alignment Search Tool (BLAST), which detects statistically significant matches between a query sequence and a database using a reference distribution of randomized sequence alignments as the 'null' comparison[23]. More complex approaches have also been proposed, such as hidden Markov models (HMM), which accommodate variations in insertion/deletion probability in different domains of a protein, instead of the position-agnostic gap penalty used by BLAST[24, 25]. Furthermore, innovations in statistical 'machine learning' models allow sequence data to be combined with other protein features and annotations to make functional predictions. As more information from large-scale functional and interaction studies becomes available, this kind of data integration will likely play an increasing role in prioritizing candidate lists of functionally uncharacterized genes as potential molecular determinants for phenotypes of interest.

In this perspective we review the roles that bioinformatics can play in deorphanizing the uncharacterized membrane proteins in the human genome. This task is outlined in Figure 1, which involves the two strategies outlined above. In the first scenario, a phenotype of interest is known, and genome-wide screens are used to generate candidate orphan genes which may be the molecular determinant for the process of interest. Here, bioinformatics approaches such as topology prediction are used to filter the results, with overall predictive accuracy of less concern than the detection of a single validated determinant for the phenotype. Alternatively, the objective is to iden-

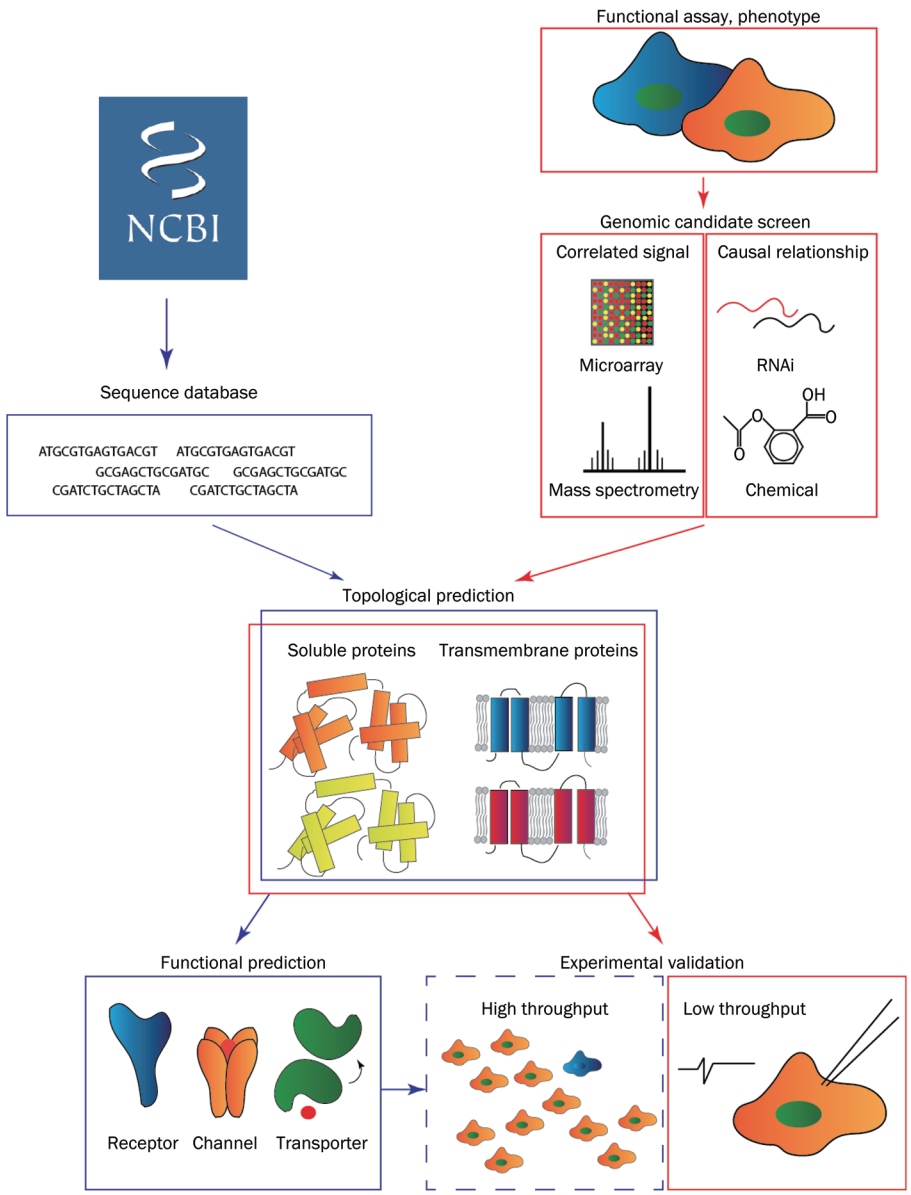

Figure 1. Deorphanization strategies. Left (Blue): *In silico* analyses of genomic sequences, topological prediction, and functional prediction. Right (Red): Phenotype of interest followed by genomic screen, bioinformatics evaluation of candidate list topology, and experimental validation.

tify the unknown function of these novel genes, with the only reference as their similarity to known proteins. In the first step of this class of investigation, genomic databases are used as a basis for global prediction of membrane proteins through topological models. Secondly, these membrane proteins are further clustered into functionally related groups, based on sequence homology, conserved motifs, and existing annotations. As discussed in the following sections, this analysis has traditionally consisted of solely *in silico* approaches, where accuracy is judged through retrospective analyses predicting the class of previously characterized proteins. However, one may speculate that such methods could effectively narrow the search space for novel membrane proteins in experimental studies, particularly in cases where a phenotype of interest is not known or well characterized. After reviewing examples of the methodologies and results from both approaches, we provide an analysis of the current landscape of characterized and orphan membrane proteins in the human genome that might be utilized as a broad guide for de-orphanization efforts. Finally, we discuss future challenges, particularly in integrating experimental and bioinformatics approaches in cases where the phenotype of a novel transmembrane protein is not known in advance.

Genomic prediction and validation of transmembrane protein function

The ability to survey the expression and activity of a large number of genes through microarrays, large-scale proteomics, and functional genetics screens has greatly aided the ability to survey and molecularly characterize diseases and signaling pathways[26–30]. Because these processes may involve cascades that begin at the plasma or organelle membrane, transmembrane proteins that are the primary drivers initiating these pathways may have a similar readout to downstream components in these assays. Thus, bioinformatics that can identify transmembrane proteins helps to narrow the number of candidates in which to invest follow-up experimental effort. More effectively, bioinformatics may even potentially focus the number of genes initially screened by identifying candidates using existing datasets for novel functions. Both cases are illustrated by recent examples summarized in Table 1.

The discovery of Leucine zipper-EF-hand containing transmembrane protein 1 (Letm1) as a mitochondrial Ca^{2+}/H^{+} antiporter demonstrates the use of bioinformatics to refine a list of candidates from genomic functional assays[15]. To identify proteins implicated in calcium transport across the inner mitochondrial membrane, the authors used a genome-wide RNA interference (RNAi) screen in Drosophila cells using fluorescent calcium and membrane potential-sensitive dyes to identify genes whose loss affected the ion homeostasis of the mitochondrial compartment. Having identified a list of candidates, they further filtered the results to include only those with predicted transmembrane segments, as soluble proteins might be members of signaling pathways that indirectly modulate but are not themselves directly implicated in ion transport. A subsequent homology search for related mammalian sequences yielded Letm1 as a Ca^{2+}/H^{+} antiporter.

Table 1. Novel experimentally validated membrane proteins.

Gene	Methods	References
Letm1	Genomewide RNAi, TM prediction	[15]
dBest1	Genomewide RNAi, TM prediction	[31]
Piezo1, 2	Enhanced expression in N2A cells	[33]
TMEM16a	IL-4 induced gene expression in microarray	[16]
MCU	Phylogeny, microarray, mass spectrometry	[17]

Similarly, the chloride-conductive 'swell' Drosophila Bestrophin 1 (dBest1) channel was identified using a fluorescence anion-sensitive dye in a flux assay combined with RNAi knockdown[31]. As with the Letm1 study, bioinformatics was used to eliminate candidate genes regulating cell volume and chloride conductance lacking predicted transmembrane spanning segments. A challenging aspect of this study is that chloride channels, unlike other better characterized channels, such as voltage-gated potassium channels[32], currently lack a signature sequence motif that might help to restrict the search space of possible membrane proteins involved in chloride conductance[31]. Thus, an unbiased genomic screen using a very specific phenotypic outcome was used to perform the bulk of candidate selection, with bioinformatics refining the hit list rather than defining the initial experimental scope.

These two studies used functional genomics to identify candidate genes whose loss is causally linked to the phenotype of interest. A related approach is to find genes that are correlated, through expression level, with this phenotype. This method was used in one of three studies reporting discovery of the calcium-sensitive chloride channel transmembrane protein 16 (TMEM16a)[11, 16, 19]. Here, the authors used microarray analysis of bronchial epithelial cells, which display increased calcium-activated chloride current following interleukin 4 (IL-4) treatment[16]. After identifying genes differentially expressed following IL-4 treatment, topological predictions to filter the hit list guided subsequent identification of TMEM16a[16]. A similar strategy of identifying differentially expressed genes correlated with a phenotype of interest was used to identify channels involved with mechanosensation. Unlike other tissues examined by the authors, mouse neuro 2a (N2A) neural crest cells displayed a mechanosensitive current, leading to the hypothesis that pressure sensitive channels would be represented among transcripts enriched in this cellular population[33]. Experimental studies of the resulting candidates identified Peizo1 and Peizo2 as mechanosensitive channels[33]. As with functional genomics approaches, the success of these studies appears to require very specific phenotypic queries that may be compared to large genomic space using profiling methodologies such as microarrays.

An example of integrated genomic analysis is the discovery of the mitochondrial calcium uniporter component MCU[17]. Here, the authors leveraged previous mass spectrometric profiling of the mitochondrial proteome[34], phylogenetic conservation of genes along an evolutionary tree, and tissue

coexpression[35] to identify genes with similar profiles across these three parameters compared to the uniporter regulator mitochondrial calcium uptake 1 (MICU1)[17]. This analysis identified MCU as a top candidate across all three parameters, a prediction verified by subsequent functional experiments. Unlike the previously described studies, bioinformatics played a key role in forming the initial 'hit list' for experimental validation, rather than refining a list that was primarily generated through unbiased screening with reference to a phenotype of interest. Another striking example of this purely bioinformatic discovery is the identification of the Ciona intestinalis voltage-sensitive phosphatase (Ci-VSP)[18]. In a 'perfect storm' of sequence homology, this gene was found to contain both a well-defined voltage sensor similar to ion channels and a phosphatase region. Thus, even though such a combination of modular units might not have been anticipated based on existing knowledge of these two protein families, unbiased computational screening allowed discovery of this novel transmembrane protein.

In most of the examples described, bioinformatic techniques have been utilized after unbiased, genome-wide analyses to filter candidate lists of potential membrane proteins underlying a phenomenon of interest, rather than identifying an initial, limited set for experimental evaluation. Also noteworthy is the fact that many of these studies utilize differences in tissue phenotypes such as ionic currents sensitive to particular stimuli, to identify candidate genes, rather than computational motifs. As noted above in analysis of chloride channels, the lack of well-defined functional motifs that might be used as an *in silico* filter necessitates this sort of approach. However, in the absence of a well-defined phenotype, how might novel membrane proteins be prioritized for characterization? How might the natural ligands, substrates and protein interaction partners of otherwise well-characterized orphan proteins be elucidated? We examine this question by first describing topological prediction algorithms, then methods for functional inference.

Prediction of membrane proteins and topology

The first level of discrimination in the bioinformatics analysis depicted in Figure 1 is to separate putative membrane proteins from soluble proteins. A number of algorithms have been reported for this task, as illustrated in Figure 2 and summarized in Table 2.

Some of the early studies in this field identified simple and effective heuristics for topology prediction. This is demonstrated by 'rules-based' methods such as Topology Prediction (TOPRED), which score each amino acid using the mean hydrophobicity of its surrounding residues, and calculate putative transmembrane regions and topology using the 'positive inside' rule[36] in which positively charged residues have a bias to face the cytoplasm[37]. Thus, topological predictions are generated in a manner analogous to a Doolittle Plot[38], by finding a threshold for hydrophobicity that will divide a protein's hydrophobicity profile into transmembrane and cytosolic elements. Similarly, alignment methods extend this idea by seeking supporting information across multiple proteins, such as dense alignment surface (DAS) and transmembrane multiple

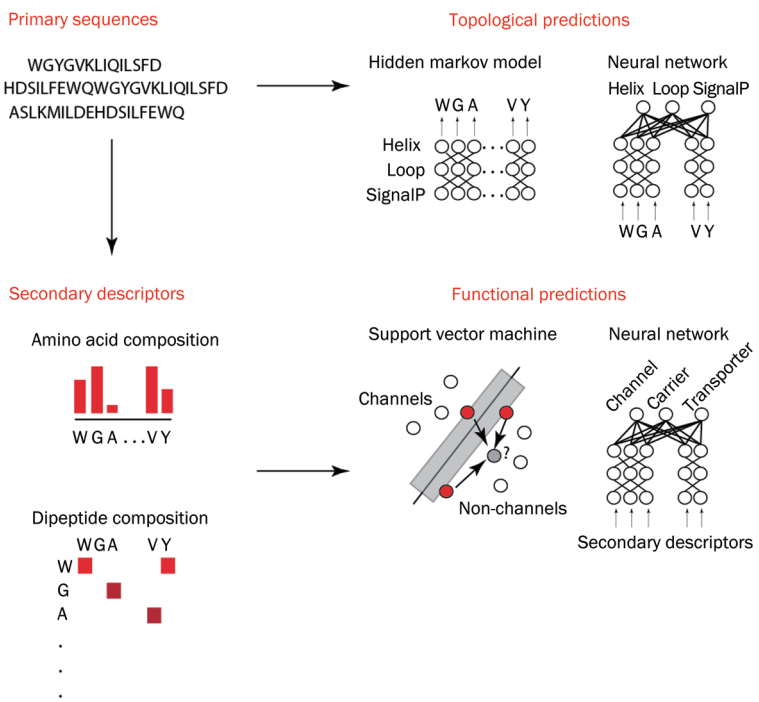

Figure 2. Algorithms for topological and functional prediction. Primary amino acid sequence (top left) is employed to predict secondary structure topology motifs (transmembrane helices, cytosolic loops, signal peptides) (top right), while secondary descriptors describing composition or substitution patterns of amino acids (bottom) are used for functional prediction for membrane proteins.

Table 2. Algorithms for predicting membrane proteins.

Algorithm class	Examples	References
Hidden markov models	TMHMM, phobius, SCAMPI, SPOCTOPUS, THUMBUP, HMMTOP1-2, TMpred	[43, 46–48, 85–88]
Neural network	MEMSAT3, SPOCTOPUS, PSIPRED, PHD	[47, 49, 89,90]
Support vector machine	MEMSAT-SVM	[50]
Consensus	BROMPT, ConPredII	[51, 52]
Rules based	TOPRED1-2, SOSUI, KKD	[37, 91–93]
Dynamic programming	MEMSAT	[44]
Alignment based	DAS, TMAP	[39, 40]

alignment prediction (TMAP), which generate consensus dot plots comparing the hydropathy profile of the protein of interest to a collection of background reference sequences or to multiple sequence alignments with homologs[39, 40].

Later developments have further explored the use of patterns present across databases of known proteins to identify useful statistical patterns for topological analysis. As with homology searches in genomic databases, hidden Markov models (HMMs) are a popular method to model the statistical properties of biological sequences. HMMs were developed for automated speech processing[41, 42], in which an observed audiogram is produced by a set of unknown words correlated with certain tonal patterns. The algorithm then statistically reconstructs the most likely word producing a given pattern of sounds over each time interval given these input properties. Similarly, an observed distribution of amino acids may be considered as an observed 'signal', with the hidden states being topological descriptions (such as transmembrane helix or cytoplasmic loop), which produce different distributions of observed amino acids[43]. This process, which cycles through each amino acid to find the optimal series of 'states' that explain the observed pattern, resembles earlier dynamic programming approaches which sought to find an optimal topological prediction by iteratively building up predictions from sub-sequences[44]. Additional complexity arises from the fact that type I membrane proteins possess a signal peptide directing them to the secretory pathway[45], a motif that resembles a transmembrane helix and thus may be misidentified by the algorithm. Thus, HMMs may be improved by incorporating 'signal peptide' as one of their hidden states, as implemented in the Phobius and signal peptide obtainer of correct topologies for uncharacterized sequences (SPOCTOPUS) programs[46, 47]. Other variations of this approach are possible, such as the scale-based method for prediction of integral membrane proteins (SCAMPI) program, which uses the predicted free energy of amino acids as the 'observed state' instead of the amino acids themselves[48]. Taken together, these prediction methods have demonstrated remarkable accuracies of 80%–97% in discriminating soluble from membrane proteins and predicting transmembrane helices in retrospective analyses[47, 48]. Additionally, as demonstrated by previously described studies, these methods' practical utility has been proven in successful filtration of hits lists from unbiased screens.

These successes are particularly notable given the inherently challenging nature of the problem they tackle. Indeed, one of the complexities of predicting protein topology from biological sequence is the inherent dependency between position and structure. Neural networks (NNs) are another approach that seeks to represent these nonlinearities by mapping a set of inputs [such as position-specific scoring matrices (PSSMs) representing the likelihood of residues in particular positions over a sliding window of a protein structure] to topological states such as transmembrane helices[49]. This mapping is performed by connecting the input data to the output through a series of 'neurons,' a set of logistic functions whose sigmoidal behavior in response to their inputs resembles activation thresholds in the mammalian nervous system. The observed toplogical states of a protein are thus modeled as a weighted combination of nonlinear activation functions, and the weights connecting the units are optimized to best reconstruct the desired output. This approach may be used independently, or combined with other algorithms. For example, the SPOCTOPUS program combines a NN and HMM, using the output from NN as an input to HMM[47], thus improving inference of the 'hidden states' in the HMM.

In addition to NNs, Support Vector Machines (SVMs) have also been utilized to perform a nonlinear mapping from input sequences to topological states. The 'support vector' in the name is derived from the fact that only a small subset of the data used to develop the model are used to generate parameters for future prediction. These 'support vectors' lie at the boundary between the classes of data, such as transmembrane helices and cytosolic loop sequences, which the algorithm seeks to classify. A strength of this approach is that the SVM may use a similarity function, such as the Gaussian distribution or a polynomial, to find a boundary separating these classes which may be intermingled in their original vector space. Like NNs, SVMs applied to topology prediction may also utilize as input a Position Specific Scoring Matrices (PSSM) for a sliding window over the protein sequence[50]. Generating this prediction over the whole length of the protein thus yields a predicted topology.

Given that each algorithm discussed above may have scenarios in which it performs better or worse, it seems reasonable to infer that combining some of these methods may overcome some of these individual shortcomings. This sort of combination has the benefit of offsetting weakness in a single method, and for potentially pooling weak evidence from multiple predictions to yield stronger collective evidence. For example the consensus prediction (ConPred) algorithm uses a heuristic rules system to average inputs from multiple topology prediction methods to derive a consensus[51]. Similarly, Bayesian prediction of membrane protein topology (BROMPT) uses a Bayesian belief network to combine evidence from five methods into a consensus[52], modeling this consensus as a

'child' node that receives weighted inputs from the five 'parent' methods.

The previously described algorithms, whether they employ amino acid frequencies, hydropathy, or folding free energy, primarily use information derived from the linear, primary structure of amino acid sequences. The resulting topology gives a 'flat' inference for tertiary or quaternary structure, but little guidance as to how the resulting helices are organized in a three-dimensional space. Such challenges have prompted the development of algorithms building on two dimensional topological predictions to infer three dimensional coordinates based on linear amino acid sequences, utilizing the population of previously solved x-ray crystal structures of membrane proteins to generate homology-based predictions[53–55]. In the absence of gold-standard structural data for most membrane proteins and channels, such techniques may represent the next-best option for tasks such as virtual small-molecule docking that require three-dimensional coordinates.

Functional sub-classification of transmembrane protein classes

After membrane proteins are identified and separated from soluble proteins using the topology prediction programs outlined above, the second level of classification in Figure 1 involves grouping the population of membrane proteins into individual functional classes, and to prospectively identify the function of characterized genes. Several methods have been reported to accomplish this task, which are summarized in Table 3 and visually diagrammed in Figure 2.

Table 3. Algorithms for predicting functional class of membrane proteins.

Features	Algorithms	References
Amino acid composition	DISC-FUNCTION, VGIChan, VKCPred	[56–58]
Dipeptide composition	Transporter-RBF, VGIChan, ionchanPred, VKCPred	[56, 57, 60, 61]
Psi-blast PSSM	Transport targets, transporter-RBF	[61, 62]
Amino acid descriptors	Transport targets, transporter-RBF	[61, 62]
Pfam domains, gene ontology annotation	TransportTP	[68]

As with many topology prediction algorithms, these methods often require the amino acid sequence to be summarized in a quantitative fashion to compare two proteins. One such descriptor that has been successfully utilized is the fraction of a protein's sequence comprised of each of the twenty naturally occurring amino acids, a vector of length twenty that sums to one and is termed the 'amino acid composition'[56, 57]. The intuition behind this descriptor is that distinct classes of membrane proteins have a bias to include particular amino acids at greater frequency due to the structural requirements or constraints for their function. Refinements of the amino acid composition descriptor have also been proposed, such as using the un-normalized count of the twenty amino acids in a protein sequence, a method reported to be more effective as it also captures differences in the characteristic length of a protein family[58]. Similarly, expanding the normalized amino acid composition to a vector length sixty – twenty for composition of the whole protein, and twenty elements each for the amino acid composition of transmembrane and non-transmembrane segments - has also allowed better discrimination[59]. Like amino acid composition, dipeptide frequencies have also been successfully utilized as descriptors to discriminate membrane proteins of different classes[56, 57, 60]. The previously mentioned PSSM derived from Position-Specific Iterative BLAST (PSI BLAST), which measure the likelihood of a substitution from the observed to an alternate amino acid at a particular position based on substitution patterns between a protein and its homologous neighbors, have also been found to have high sensitivity as a descriptor[61]. More abstractly, numerical descriptors of folding energetics have also been employed in predictive models[61].

Just as the input descriptors to these algorithms are varied, so are the kinds of functional predictions produced in these studies. Several methods have been used to predict a query gene's family membership, such as classifying channels, transporters, and carriers from one another[58]. In greater detail, these methods have also been used to predict a protein's substrate, such as different metal ions for channels or protein/nucleic acids for transporters[62]. Predictions have also been targeted for functional parameters specific to particular classes of membrane proteins. For example, amino acid sequence has been used to predict the half-maximal activation potential of voltage gated channels[63], discriminate between channels based on their electrophysiological parameters[64], or identify channels that may serve as promising therapeutic targets[65].

These previously described methods, in essence, rely on the proximity of a query protein to a neighborhood of known proteins in the space of the descriptor used. Further refinements have been proposed, where this proximity measurement may be combined with other features such as Gene Ontology terms describing the biological processes, molecular functions, subcellular localization of a protein[66], presence of class-associated protein families (Pfam) domains[67], or the number of predicted transmembrane domains[68]. The resulting combination of annotated and raw sequence information may then be used in a prediction algorithm such as the previously discussed SVM[68]. Indeed, the ability of amino acid profiles to serve as relevant features for identifying functionally related proteins may suggest that families share specific motifs, and specific structural fragments and motifs have also been identified in related studies[69, 70].

Expanding these predictions based on two-dimensional structure correlated with classifications or functional parameters, methods have also been developed to directly infer function based on a three dimensional conformation. For example, the SLITHER program uses molecular modeling simulations to predict whether a putative substrate molecule may permeate the cavities or channels in a protein structure[71]. In cases

where the existence of a channel in a protein is unverified, the MolAxis program can be used to predict whether they exist using computational geometry[72]. Obviously, both of these methodologies require three-dimensional protein coordinates which are experimentally unavailable for most channels or other membrane proteins, but might be combined with homology-based three dimensional structure predictions described in the previous section to generate functional predictions for inferred three dimensional structures.

A related functional prediction is to identify the natural ligand, ion substrate or protein interaction partner of the novel proteins. Indeed, examples that highlight the challenge of deorphanizing a large number of seven transmembrane protein receptors, where the natural binding partner(s) of some otherwise well-characterized transmembrane receptors such as BRS-3 remains unknown[73]. Though not specifically developed to identify peptide – receptor interactions *in silico*, large-scale predictions of protein-protein interactions have been described using two and three-dimensional information[74–76]. Conceivably, such algorithms might be employed to identify novel interactions between peptide ligands and the subset of peptide-binding receptors. Direct bioinformatics identification of ligands such as neuropeptide precursors have also benefited from the increased availability of genome-wide proteomic and nucleotide data, as demonstrated by the computational prediction of more than 200 novel neuropeptides in the honeybee *Apis mellifera*, of which 100 were validated using peptidomics[77]. Related studies of the red flour beetle *Tribolium castaneum* have employed homology analysis to validate 30/41 predicted neuropeptide genes using mass spectrometry data, encoding 71 peptides[78]. Given the accuracy of the predictions in these studies using large genomics datasets, we speculate that such methods and information provide a promising pool of potential novel ligands that might be screened in functional assays against putative peptide-binding receptors.

A reference map of uncharacterized membrane proteins

In the previous sections we have provided an overview of experimental and computational methodologies used to deorphanize uncharacterized membrane proteins. Here we quantify how much of the transmembrane proteome has been characterized, and whether the coverage of the characterized regions is biased toward proteins with a particular topology by generating a reference map of the human transmembrane proteome.

This analysis is based on 35879 unique human RefSeq protein sequences downloaded from NCBI as GenBank records. To reduce bias in our analysis resulting from proteins with multiple isoforms, we collapsed this collection into unique gene symbols by retaining only the entry (for a given gene symbol) with the greatest number of annotated transmembrane segments among annotated sites in their GenBank fields, under the hypothesis that the sequence with the most annotated segments represents the most studied and highest-quality record for a particular gene. In cases where the gene has no transmembrane helices we simply kept the first occurring entry. Applying this filter left 19977 sequences. Because uncharacterized membrane proteins may lack annotated transmembrane segments, we utilized several of the previously described topology prediction programs to generate an estimated transmembrane segment count for these orphan proteins. The three programs used were TMHMM2.0[43], SCAMPI_multi[79], and PHOBIUS1.0.1[46], and the weights used to average the predictions were estimated using a linear regression against a count of known transmembrane segments.

We estimated the number of membrane proteins using a criterion of one or more predicted or annotated transmembrane segments. This analysis yielded 4991 of the 19977 sequences for unique genes passing this filter, corresponding to ~25% of the genome, a value in reasonable alignment with previous estimates[3]. To determine which of these 4991 membrane proteins were previously unannotated, we used two approaches. First, we selected a list of all RefSeq sequences lacking a Gene Reference into Function (GeneRIF) annotation, giving a set of 5723 unique proteins. While this filter can identify sequences that have previously been annotated for function, the lack of a hierarchy or sub-classification of these annotations by strength of evidence means that some of these sequences may actually by effectively uncharacterized. By manually examining many entries, we have indeed found that some GeneRIF entries describe presumed or inferred function without experimental support. While these may be useful for generating hypotheses, this ambiguity complicates our estimate of the number of uncharacterized membrane proteins. Thus, we also utilized the independent annotation in the Gene Ontology (GO) database. Following a similar methodology used to identify uncharacterized proteins in *Arabidopsis thaliana*[80], we identified all proteins either lacking GO annotation (2983 proteins) or having no data (ND) evidence code for Molecular Function (MF) annotation at the root node (the default assignment in the GO for uncharacterized proteins) (597 proteins), giving a total of 3580. These intersect with the GeneRIF-based set by 2431. The union of the uncharacterized sets gives 6872 proteins, of which ~25% (1533) are transmembrane. In contrast, only 216 of the intersecting set of 2431 are in our estimated transmembrane set, so we used the union of the estimated uncharacterized sets as a less conservative approach. A summary of all filters applied is given in Figure 3A. Many of the 4991 estimated membrane proteins in this analysis (3791, ~76%) have GO annotations for MF, including 1479 unique terms (as a single protein may have more than one MF annotation). The distribution of all MF terms assigned to more than ten proteins (167 terms) is shown in Figure 3B, indicating that G-protein coupled receptors, olfactory receptors, nucleotide binding receptors, and calcium interacting proteins dominate this list. To independently evaluate the quality of our inference, we used the same approach to predict the number of transmembrane proteins in *Saccharomyces cerevisiae*. The localization of approximately 75% of the yeast genome has been experimentally assessed using Green Fluorescent Protein (GFP)-tagged fusion proteins to determine presence/absence at twenty-two

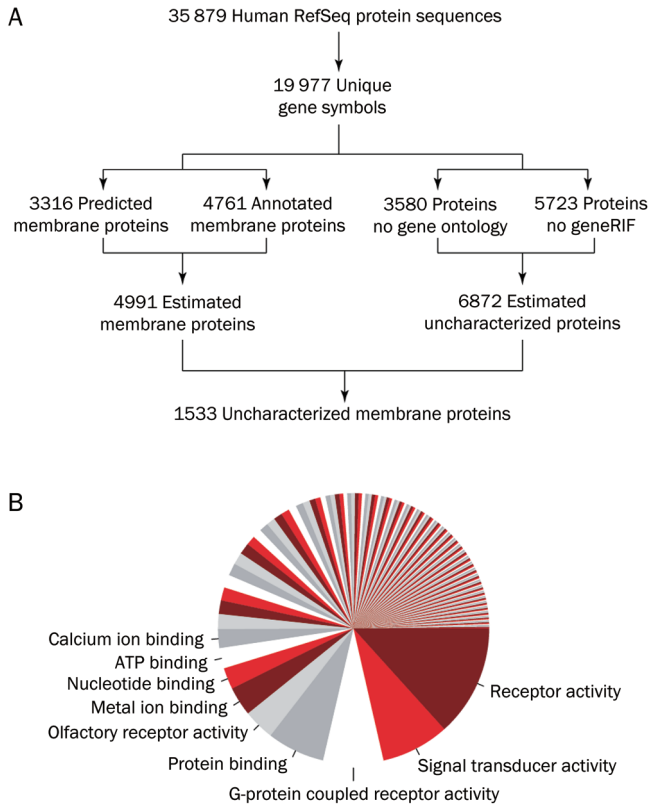

Figure 3. Estimating the number of uncharacterized human membrane proteins. (A) Human RefSeq protein sequences are collapsed to unique genes. Three topology prediction algorithms are averaged to generate a list of predicted membrane proteins, and merged with membrane proteins derived from GenBank transmembrane helix annotations to yield a combined population of estimated membrane proteins. Previous functional annotations are evaluated using GeneRIF fields and Gene Ontology (GO) records, which are merged to yield a combined population of estimated uncharacterized proteins. The intersection of the membrane and uncharacterized populations represent uncharacterized membrane proteins. (B) Distribution of top GO molecular function (MF) categories for all membrane proteins.

organelle sites[81], and we used this information to assess the accuracy of TM protein predictions. These analyses, shown in Figure 4, demonstrate that the predicted transmembrane proteins, which constitute ~20% of the yeast genome, are experimentally localized in the Endoplasmic Reticulum (ER), secretion pathway (vacuole) and cell periphery at higher rates (18, 13, and 7-fold respectively) than predicted soluble proteins, whose localization records are biased for the cytoplasm and nucleus (5 fold and 7.5-fold enrichment, respectively). While the discrimination is not perfect, the population of predicted TM proteins in yeast obtained using the predictive methodology from the human analysis is enriched for experimentally annotated localization at membrane sites, supporting the use of these topological predictions as a proxy for TM localization.

To gain a global overview of the distribution of the annotated and unannotated membrane proteins identified in our analysis, we generated a vector description of each sequence to allow systematic comparison. The first twenty elements of this vector contain the count of each of the twenty naturally occurring amino acids in the proteins sequence. The next twenty contain these counts restricted to the transmembrane regions, while the last twenty contain the counts for the cytosolic loops, giving a total length of sixty. All counts were calculated using the topological prediction output of TMHMM2.0 for transmembrane segments for consistency. The resulting vectors of length sixty were then embedded in a low-dimensional map using t-Stochastic Neighbor Embedding (t-SNE), an algorithm that produces coordinate maps of high-dimensional data which represent the pairwise similarity between objects[82]. This algorithm, compared to other nonlinear methods, has been shown to better separate images of handwritten digits and facial photographs into distinct clusters in two dimensional space[82]. To separate the resulting map into regions, we clustered the resulting coordinates from t-SNE using affinity propagation[83] using the squared Euclidean distance between the t-SNE coordinates and the maximum pairwise distance as the input preference for each datapoint to be a cluster center. The resulting map for membrane proteins with previously annotated function is displayed in Figure 5A, with colors representing the clusters defined by affinity propagation. The distribution of the estimated set of uncharacterized membrane proteins is shown in Figure 5B. While the range of space covered by the characterized and uncharacterized sets is comparable, the density of the uncharacterized membrane proteins is concentrated in a region of space occupied by seven transmembrane segment receptors (Figure 6, 5D) and reflecting orphan olfactory receptors. This is reflected quantitatively by the modest correlation coefficient of 0.20 between the grid-cell counts of the uncharacterized and characterized sequences in Figure 6.

While this analysis can distinguish broad structural classes of membrane proteins, as shown by the spatial localization of seven membrane proteins (Figure 5D), voltage-gated sodium and calcium channels are also intermingled with transporters in the lower left quadrant. While these proteins might be topologically similar, they are clearly functionally distinct. It thus remains unclear which class of sequence descriptors, if any, can best capture functional differences in this kind of analysis, and how to evaluate the accuracy of such features. From the perspective of future de-orphanization, it appears encouraging that the TMEM class of proteins is broadly distributed across the sequence space, suggesting that membrane proteins of many functional or topological classes may yet be elucidated.

Perspective

Review of the literature suggests a 'gap' between experimental and computational methods. While *in silico* functional predictions are primarily verified through retrospective accuracy, experimental studies with unbiased genomics approaches use bioinformatics as a way to pare down a candidate list, rather than restrict and guide the initial search space. Thus, we anticipate there are unrealized opportunities for predictive

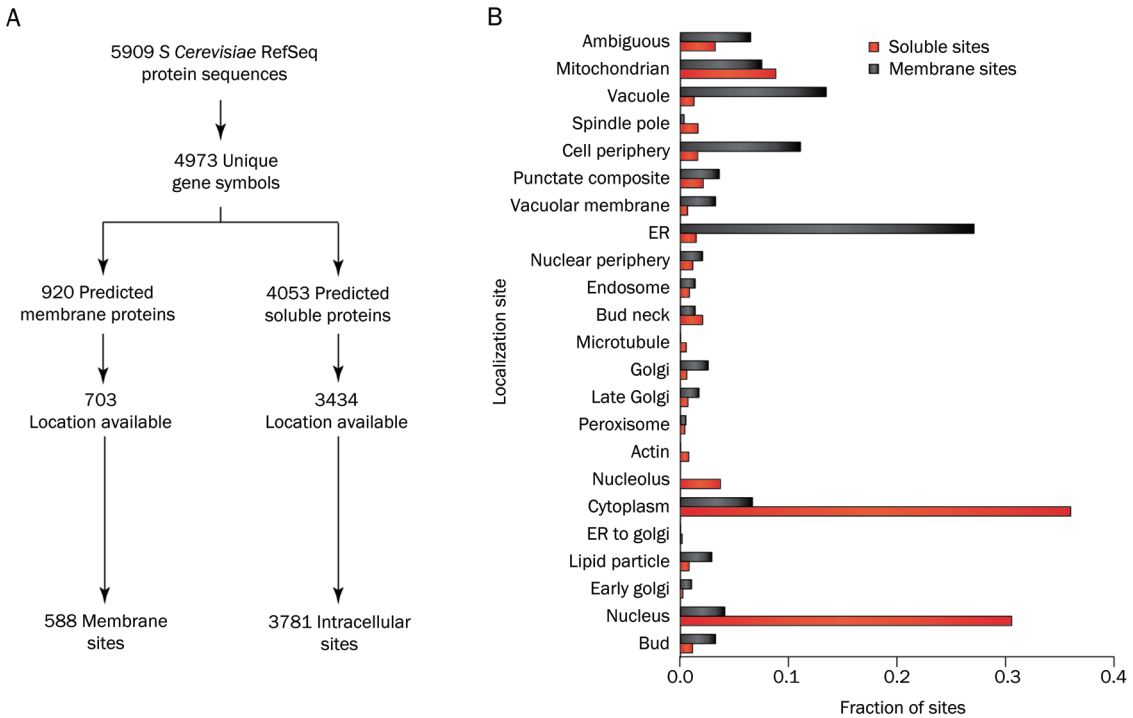

Figure 4. Subcellular localization of predicted membrane and soluble proteins in *S Cerevisiase*. GFP fusions of individual yeast proteins have been expressed, localized and annotated[81]. (A) Analytical pipeline for prediction of yeast membrane proteins beginning with 5909 RefSeq entries that are filtered and resulted in 4973 unique gene names. Topology algorithms for unique genes yield 920 putative membrane proteins and 4053 putative soluble proteins. Fractions of both groups possess experimentally determined subcellular locations. (B) The distribution of experimentally determined localization(s) for predicted membrane and soluble proteins in (A) among 22 cellular sites. Bar lengths are normalized to the total number of subcellular location sites available for predicted membrane and soluble proteins from (A).

algorithms to be used to identify novel membrane proteins and suggest possible phenotypes for functional validation. Furthermore, such studies will help computational researchers to better understand which models and descriptions of protein structure are most successful in predicting the results of these experimental validations, and thus iteratively improve the underlying bioinformatics algorithms.

We also speculate that predictions of three-dimensional structure have not been fully exploited for these kinds of studies. Indeed, the small fraction of transmembrane drug targets with crystal structures derived from DrugBank[84] indicated in Table 4 suggests that this is a role in which bioinformatics may fill a large existing knowledge gap. As more membrane proteins are crystallized and homology-based three dimensional coordinate prediction methods become more mature, it is intriguing to speculate that tertiary structure predictions might generate functional predictions using substrate docking, in a manner similar to virtual screening of small molecule ligands. Such approaches might complement existing predictors based on amino acid sequence alone.

Table 4. Structural characterization of human drug targets.

Description	Total number	Soluble	Trans-membrane
Unique genes	19 977	14 986 (75%)	4 991 (25%)
Unique drug targets	2 048	1 357 (68%)	636 (32%)
Unique drug targets with structure	791	555 (72%)	216 (28%)

Note: The 'total number' column doesn't add up to the soluble and transmembrane counts for the last two rows because the gene symbols in DrugBank don't map to 100% of the RefSeq entries.

An additional challenge comes from the fact that deorphanization often involves identification of unknown functions. Indeed, while many of the experimental studies discussed here have sought to generate candidate lists based on a specific phenotype, the challenge may often lie in assessing a completely unknown function. Indeed, even in cases where bioinformatics has perfectly identified a novel protein, such as Ci-VSP, which contains both a voltage-sensing domain and phosphatase catalytic domain[18], the substrate of this enzyme, and thus its biological role, was not immediately apparent from the initial characterization. Therefore, in the absence of functional knowledge – for example, lack of knowledge of a channel's presumed triggering stimulus that generates current – modified screening approaches will be needed to probe the function of unannotated membrane proteins. Ion channels as a class share the properties of conducting ionic currents, a general feature that may be exploited. Recent innovations in high-throughput patch clamping may allow a matrix of different

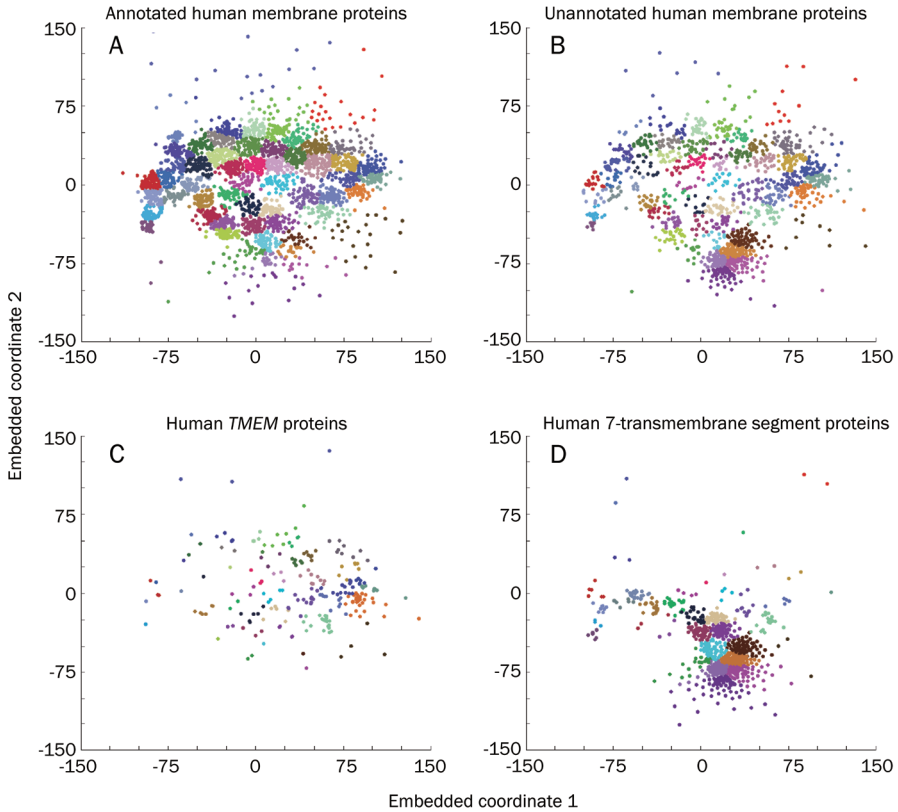

Figure 5. Landscape of human membrane protein diversity. Two dimensional embedded coordinates are generated from vectors counting the number of each of the twenty amino acids in a whole protein sequence, transmembrane segments, and cytosolic segments for 4991 estimated human membrane proteins, using the t-stochastic neighbor embedding (t-SNE) algorithm. Colors represent groups identified by applying affinity propagation clustering to the embedded coordinates. (A) Embedded coordinates and cluster identity of subset of human membrane proteins with previous functional annotation. (B) As in (A), for uncharacterized membrane proteins. (C) As in (A), for *TMEM* proteins. (D) As in (A), for sequences with seven transmembrane segments as denoted by RefSeq annotations or averaged predictions of three topology algorithms.

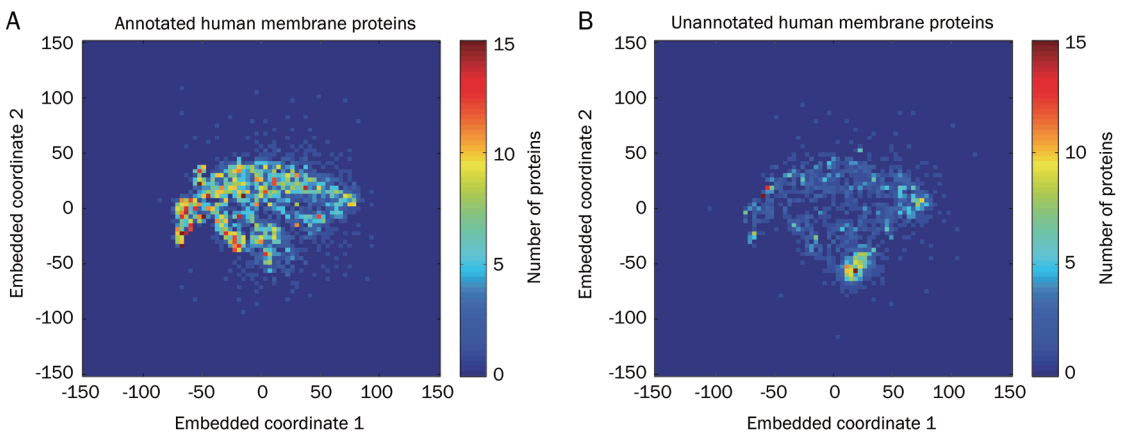

Figure 6. Density profile of landscape of human membrane proteins. Embedded two-dimensional coordinates generated from vectors containing counts of each of the twenty amino acids in a whole protein sequence, membrane segments, and cytosolic segments, using the t-stochastic neighbor embedding (t-SNE) algorithm for 4991 estimated human membrane proteins. (A) Count per coordinate grid representing the number of sequences (colorbar) for the subset of human membrane proteins with previous annotation. (B) As in (A), for uncharacterized membrane proteins.

potential stimuli and buffer conditions to be tested, allowing rapid functional profiling. Such approaches, combined with high-throughput imaging to determine localization, have the potential to systematize the characterization of novel mem-

brane proteins.

In summary, while the characterization of the transmembrane genome has witnessed many informatics and experimental successes, our analysis shows that almost one-third of the membrane proteins still lack functional annotation. Given the current seeming lack of overlap between bioinformatics and unbiased screening approaches, we speculate there are opportunities for predictive algorithms to further refine screening studies, and for new profiling technologies to validate these predictive algorithms. This combination of smarter analytics and broader experimental methodology may thus help deorphanize the remaining membrane proteins in the genome, offering potential drug targets as well as greater understanding of these genes' biological roles.

Acknowledgements

We thank the Min LI laboratory for valuable discussions and Alison NEAL for editorial assistance. This work is supported by grants to Min LI from the National Institutes of Health (MH084691).

References

1 Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, *et al*. Initial sequencing and analysis of the human genome. Nature 2001; 409: 860–921.
2 Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, *et al*. The sequence of the human genome. Science 2001; 291: 1304–51.
3 Fagerberg L, Jonasson K, von Heijne G, Uhlen M, Berglund L. Prediction of the human membrane proteome. Proteomics 2010; 10: 1141–9.
4 Bakheet TM, Doig AJ. Properties and identification of human protein drug targets. Bioinformatics 2009; 25: 451–7.
5 Yildirim MA, Goh KI, Cusick ME, Barabasi AL, Vidal M. Drug-target network. Nat Biotechnol 2007; 25: 1119–26.
6 Alberts B. Molecular biology of the cell. 4th ed. New York Garland Science; 2002.
7 Dietrich A, Kalwa H, Storch U, Mederos y Schnitzler M, Salanova B, Pinkenburg O, *et al*. Pressure-induced and store-operated cation influx in vascular smooth muscle cells is independent of TRPC1. Pflugers Arch 2007; 455: 465–77.
8 Drew LJ, Rohrer DK, Price MP, Blaver KE, Cockayne DA, Cesare P, *et al*. Acid-sensing ion channels ASIC2 and ASIC3 do not contribute to mechanically activated currents in mammalian sensory neurones. J Physiol 2004; 556: 691–710.
9 Rosati B, McKinnon D. Regulation of ion channel expression. Circ Res 2004; 94: 874–83.
10 Arreola J, Begenisich T, Nehrke K, Nguyen HV, Park K, Richardson L, *et al*. Secretion and cell volume regulation by salivary acinar cells from mice lacking expression of the Clcn3 Cl^- channel gene. J Physiol 2002; 545: 207–16.
11 Schroeder BC, Cheng T, Jan YN, Jan LY. Expression cloning of TMEM16A as a calcium-activated chloride channel subunit. Cell 2008; 134: 1019–29.
12 Eroglu C, Allen NJ, Susman MW, O'Rourke NA, Park CY, Ozkan E, *et al*. Gabapentin receptor alpha2delta-1 is a neuronal thrombospondin receptor responsible for excitatory CNS synaptogenesis. Cell 2009; 139: 380–92.
13 Barhanin J, Lesage F, Guillemare E, Fink M, Lazdunski M, Romey G. K_VLQT1 and IsK (minK) proteins associate to form the I_{Ks} cardiac potassium current. Nature 1996; 384: 78–80.
14 Sanguinetti MC, Curran ME, Zou A, Shen J, Spector PS, Atkinson DL, *et al*. Coassembly of K_VLQT1 and minK (IsK) proteins to form cardiac I_{Ks} potassium channel. Nature 1996; 384: 80–3.
15 Jiang DW, Zhao LL, Clapham DE. Genome-wide RNAi screen identifies Letm1 as a mitochondrial Ca^{2+}/H^+ antiporter. Science 2009; 326: 144–47.
16 Caputo A, Caci E, Ferrera L, Pedemonte N, Barsanti C, Sondo E, *et al*. TMEM16A, a membrane protein associated with calcium-dependent chloride channel activity. Science 2008; 322: 590–4.
17 Baughman JM, Perocchi F, Girgis HS, Plovanich M, Belcher-Timme CA, Sancak Y, *et al*. Integrative genomics identifies MCU as an essential component of the mitochondrial calcium uniporter. Nature 2011; 476: 341–5.
18 Murata Y, Iwasaki H, Sasaki M, Inaba K, Okamura Y. Phosphoinositide phosphatase activity coupled to an intrinsic voltage sensor. Nature 2005; 435: 1239–43.
19 Yang YD, Cho H, Koo JY, Tak MH, Cho Y, Shim WS, *et al*. TMEM16A confers receptor-activated calcium-dependent chloride conductance. Nature 2008; 455: 1210–5.
20 Fontanilla D, Johannessen M, Hajipour AR, Cozzi NV, Jackson MB, Ruoho AE. The hallucinogen *N*,*N*-dimethyltryptamine (DMT) is an endogenous sigma-1 receptor regulator. Science 2009; 323: 934–7.
21 van der Horst E, Peironcely JE, Ijzerman AP, Beukers MW, Lane JR, van Vlijmen HW, *et al*. A novel chemogenomics analysis of G protein-coupled receptors (GPCRs) and their ligands: a potential strategy for receptor de-orphanization. BMC Bioinformatics 2010; 11: 316.
22 Ben-Shlomo I, Rauch R, Avsian-Kretchmer O, Hsueh AJ. Matching receptome genes with their ligands for surveying paracrine/autocrine signaling systems. Mol Endocrinol 2007; 21: 2009–14.
23 Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997; 25: 3389–402.
24 Durbin R. Biological sequence analysis: probabalistic models of proteins and nucleic acids. UK New York Cambridge University Press; 1998.
25 Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 2011; 39: W29–37.
26 Brass AL, Dykxhoorn DM, Benita Y, Yan N, Engelman A, Xavier RJ, *et al*. Identification of host proteins required for HIV infection through a functional genomic screen. Science 2008; 319: 921–6.
27 Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, *et al*. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 2003; 34: 267–73.
28 Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001; 98: 5116–21.
29 Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, *et al*. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. Nature 2006; 440: 637–43.
30 Gunsalus KC, Ge H, Schetter AJ, Goldberg DS, Han JD, Hao T, *et al*. Predictive models of molecular machines involved in *Caenorhabditis elegans* early embryogenesis. Nature 2005; 436: 861–5.
31 Stotz SC, Clapham DE. Anion-sensitive fluorophore identifies the Drosophila swell-activated chloride channel in a genome-wide RNA interference screen. PLoS One 2012; 7: e46865.
32 Papazian DM, Timpe LC, Jan YN, Jan LY. Alteration of voltage-dependence of Shaker potassium channel by mutations in the S4 sequence. Nature 1991; 349: 305–10.
33 Coste B, Mathur J, Schmidt M, Earley TJ, Ranade S, Petrus MJ, *et al*.

Piezo1 and Piezo2 are essential components of distinct mechanically activated cation channels. Science 2010; 330: 55–60.
34 Pagliarini DJ, Calvo SE, Chang B, Sheth SA, Vafai SB, Ong SE, *et al*. A mitochondrial protein compendium elucidates complex I disease biology. Cell 2008; 134: 112–23.
35 Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, *et al*. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A 2004; 101: 6062–7.
36 Vonheijne G, Gavel Y. Topogenic signals in integral membrane-proteins. Eur J Biochem 1988; 174: 671–78.
37 Vonheijne G. Membrane-protein structure prediction-hydrophobicity analysis and the positive-inside rule. J Mol Biol 1992; 225: 487–94.
38 Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982; 157: 105–32.
39 Cserzo M, Wallin E, Simon I, von Heijne G, Elofsson A. Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. Protein Eng 1997; 10: 673–6.
40 Persson B, Argos P. Prediction of membrane protein topology utilizing multiple sequence alignments. J Protein Chem 1997; 16: 453–7.
41 Baker J. The DRAGON system — an overview. IEEE Transactions on Acoustics Speech and Signal Processing 1975; 23: 24–29.
42 Jelinek F, Bahl L, Mercer R. Design of a linguistic statistical decoder for the recognition of continuous speech. IEEE Transactions on Information Theory 1975; 21: 250–56.
43 Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 2001; 305: 567–80.
44 Jones DT, Taylor WR, Thorton JM. A model recognition approach to the prediction of all-helical membrane-protein structure and topology. Biochemistry 1994; 33: 3038–49.
45 Blobel G, Dobberstein B. Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma. J Cell Biol 1975; 67: 835–51.
46 Kall L, Krogh A, Sonnhammer EL. A combined transmembrane topology and signal peptide prediction method. J Mol Biol 2004; 338: 1027–36.
47 Viklund H, Bernsel A, Skwark M, Elofsson A. SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. Bioinformatics 2008; 24: 2928–9.
48 Bernsel A, Viklund H, Falk J, Lindahl E, von Heijne G, Elofsson A. Prediction of membrane-protein topology from first principles. Proc Natl Acad Sci U S A 2008; 105: 7177–81.
49 Jones DT. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. Bioinformatics 2007; 23: 538–44.
50 Nugent T, Jones DT. Transmembrane protein topology prediction using support vector machines. BMC Bioinformatics 2009; 10:159.
51 Arai M, Mitsuke H, Ikeda M, Xia JX, Kikuchi T, Satake M, *et al*. ConPred II: a consensus prediction method for obtaining transmembrane topology models with high reliability. Nucleic Acids Res 2004; 32: W390–3.
52 Taylor PD, Attwood TK, Flower DR. BPROMPT: a consensus server for membrane protein prediction. Nucleic Acids Res 2003; 31: 3698–700.
53 Kelm S, Shi JY, Deane CM. MEDELLER: homology-based coordinate generation for membrane proteins. Bioinformatics 2010; 26: 2833–40.
54 Durell SR, Hao Y, Guy HR. Structural models of the transmembrane region of voltage-gated and other K^+ channels in open, closed, and inactivated conformations. J Struct Biol 1998; 121: 263–84.
55 Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS. Three-dimensional structures of membrane proteins from genomic sequencing. Cell 2012; 149: 1607–21.
56 Saha S, Zack J, Singh B, Raghava GP. VGIchan: prediction and classification of voltage-gated ion channels. Genomics Proteomics Bioinformatics 2006; 4: 253–8.
57 Chen W, Lin H. Identification of voltage-gated potassium channel subfamilies from sequence information using support vector machine. Comput Biol Med 2012; 42: 504–7.
58 Gromiha MM, Yabuki Y. Functional discrimination of membrane proteins using machine learning techniques. Bmc Bioinformatics 2008; 9: 135.
59 Schaadt NS, Helms V. Functional classification of membrane transporters and channels based on filtered TM/non-TM amino acid composition. Biopolymers 2012; 97: 558–67.
60 Lin H, Ding H. Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition. J Theor Biol 2011; 269: 64–9.
61 Ou YY, Chen SA, Gromiha MM. Classification of transporters using efficient radial basis function networks with position-specific scoring matrices and biochemical properties. Proteins 2010; 78: 1789–97.
62 Chen SA, Ou YY, Lee TY, Gromiha MM. Prediction of transporter targets using efficient RBF networks with PSSM profiles and biochemical properties. Bioinformatics 2011; 27: 2062–7.
63 Li B, Gallin WJ. Computational identification of residues that modulate voltage sensitivity of voltage-gated potassium channels. BMC Struct Biol 2005; 5: 16.
64 Fernandez M, Fernandez L, Abreu JI, Garriga M. Classification of voltage-gated K^+ ion channels from 3D pseudo-folding graph representation of protein sequences using genetic algorithm-optimized support vector machines. J Mol Graph Model 2008; 26: 1306–14.
65 Huang C, Zhang R, Chen Z, Jiang Y, Shang Z, Sun P, *et al*. Predict potential drug targets from the ion channel proteins based on SVM. J Theor Biol 2010; 262: 750–6.
66 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000; 25: 25–9.
67 Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, *et al*. The Pfam protein families database. Nucleic Acids Res 2012; 40: D290–301.
68 Li HQ, Benedito VA, Udvardi MK, Zhao PX. TransportTP: A two-phase classification approach for membrane transporter prediction and characterization. BMC Bioinformatics 2009; 10: 418.
69 Marsico A, Henschel A, Winter C, Tuukkanen A, Vassilev B, Scheubert K, *et al*. Structural fragment clustering reveals novel structural and functional motifs in alpha-helical transmembrane proteins. BMC Bioinformatics 2010; 11: 204.
70 Marsico A, Scheubert K, Tuukkanen A, Henschel A, Winter C, Winnenburg R, *et al*. MeMotif: a database of linear motifs in alpha-helical transmembrane proteins. Nucleic Acids Res 2010; 38: D181–9.
71 Lee PH, Kuo KL, Chu PY, Liu EM, Lin JH. SLITHER: a web server for generating contiguous conformations of substrate molecules entering into deep active sites of proteins or migrating through channels in membrane transporters. Nucleic Acids Res 2009; 37: W559–64.
72 Yaffe E, Fishelovitch D, Wolfson HJ, Halperin D, Nussinov R. MolAxis: a server for identification of channels in macromolecules. Nucleic Acids Res 2008; 36: W210–5.
73 Majumdar ID, Weber HC. Biology and pharmacology of bombesin receptor subtype-3. Curr Opin Endocrinol Diabetes Obes 2012; 19:

3–7.

74 Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, *et al*. Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature 2012; 490: 556–60.

75 Hue M, Riffle M, Vert JP, Noble WS. Large-scale prediction of protein-protein interactions from structures. BMC Bioinformatics 2010; 11: 144.

76 Stein A, Mosca R, Aloy P. Three-dimensional modeling of protein interactions and complexes is going 'omics. Curr Opin Struct Biol 2011; 21: 200–8.

77 Hummon AB, Richmond TA, Verleyen P, Baggerman G, Huybrechts J, Ewing MA, *et al*. From the genome to the proteome: uncovering peptides in the Apis brain. Science 2006; 314: 647–9.

78 Li B, Predel R, Neupert S, Hauser F, Tanaka Y, Cazzamali G, *et al*. Genomics, transcriptomics, and peptidomics of neuropeptides and protein hormones in the red flour beetle *Tribolium castaneum*. Genome Res 2008; 18: 113–22.

79 Bernsel A, Viklund H, Falk J, Lindahl E, von Heijne G, Elofsson A. Prediction of membrane-protein topology from first principles. Proc Natl Acad Sci U S A 2008; 105: 7177–81.

80 Horan K, Jang C, Bailey-Serres J, Mittler R, Shelton C, Harper JF, *et al*. Annotating genes of known and unknown function by large-scale coexpression analysis. Plant Physiol 2008; 147: 41–57.

81 Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, *et al*. Global analysis of protein localization in budding yeast. Nature 2003; 425: 686–91.

82 Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Machine Learning Res 2008; 9: 85.

83 Frey BJ, Dueck D. Clustering by passing messages between data points. Science 2007; 315: 972–6.

84 Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, *et al.* DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res 2011; 39: D1035–41.

85 Sonnhammer EL, von Heijne G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol 1998; 6: 175–82.

86 Zhou HY, Zhou YQ. Predicting the topology of transmembrane helical proteins using mean burial propensity and a hidden-Markov-model-based method. Protein Sci 2003; 12: 1547–55.

87 Tusnady GE, Simon I. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. J Mol Biol 1998; 283: 489–506.

88 Tusnady GE, Simon I. The HMMTOP transmembrane topology prediction server. Bioinformatics 2001; 17: 849–50.

89 Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999; 292: 195–202.

90 Rost B, Fariselli P, Casadio R. Topology prediction for helical transmembrane proteins at 86% accuracy. Protein Sci 1996; 5: 1704–18.

91 Hirokawa T, Boon-Chieng S, Mitaku S. SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics 1998; 14: 378–9.

92 Claros MG, Vonheijne G. Toppred-Ii — an improved software for membrane-protein structure predictions. Comput Appl Biosci 1994; 10: 685–6.

93 Klein P, Kanehisa M, Delisi C. The detection and classification of membrane-spanning proteins. Biochim Biophys Acta 1985; 815: 468–76.