Coordinating the impact of structural genomics on the human α-helical transmembrane proteome

Ursula Pieper[1–3], Avner Schlessinger[1–3,16], Edda Kloppmann[4], Geoffrey A Chang[5], James J Chou[6], Mark E Dumont[7], Brian G Fox[8], Petra Fromme[9], Wayne A Hendrickson[10], Michael G Malkowski[11], Douglas C Rees[12], David L Stokes[10], Michael H B Stowell[13], Michael C Wiener[14], Burkhard Rost[4], Robert M Stroud[15], Raymond C Stevens[5] & Andrej Sali[1–3]

Given the recent successes in determining membrane-protein structures, we explore the tractability of determining representatives for the entire human membrane proteome. This proteome contains 2,925 unique integral α-helical transmembrane-domain sequences that cluster into 1,201 families sharing more than 25% sequence identity. Structures of 100 optimally selected targets would increase the fraction of modelable human α-helical transmembrane domains from 26% to 58%, providing structure and function information not otherwise available.

Integral membrane proteins are classified into two broad categories on the basis of the nature of their interaction with the membrane: integral monotopic proteins are attached to the lipid membrane from only one side, whereas integral bitopic and polytopic proteins—also known as transmembrane proteins—span the lipid bilayer once and more than once, respectively. The transmembrane proteins typically have either an α-helical fold or a multistranded β-barrel fold (**Supplementary Note 1**).

Several reliable methods to predict α-helical transmembrane domains from sequence are available[1]. Using such methods, a recent survey predicted that approximately a quarter of the human proteome is composed of proteins with at least one transmembrane α-helix[2]. The α-helical transmembrane domains are significantly more abundant than the β-barrel transmembrane domains and also appear to be functionally more diverse[3]. For example, α-helical transmembrane proteins are found in all biological membranes, whereas β-barrel transmembrane proteins only span the mitochondrial membrane in humans, the outer membranes of Gram-negative bacteria and chloroplast membranes in photosynthetic eukaryotes.

Integral membrane proteins are essential for cellular function. The α-helical transmembrane proteins comprise critical functional groups, including receptors, transporters, transceptors, ion channels, enzymes and others[2,4]. In particular, approximately 60% of current drug targets are membrane proteins[5], with G protein–coupled receptors (GPCRs) and ion channels alone accounting for 30% and 10% of primary drug targets, respectively. In addition, membrane transporters are important secondary drug targets for regulating the absorption, distribution, metabolism and excretion (ADME) of drugs targeting other proteins[6].

Determining the structure of membrane proteins is a powerful tool for understanding their diverse functions and discovering new drugs. However, in stark contrast to their frequency and importance, only 1,035 of the approximately 85,000 entries in the Protein Data Bank (PDB; 2 October 2012) describe α-helical transmembrane-protein structures[7,8], owing to extraordinary technical challenges involved in their purification and structure determination by X-ray crystallography, NMR spectroscopy or electron microscopy.

Progress has recently been made in *de novo* structure prediction of α-helical membrane-protein sequences by relying on a large number of homologous sequences determined by genomic sequencing[9]. Nevertheless, comparative or homology protein-structure modeling, which relies on experimentally determined structures of homologous proteins[10–12], remains the most accurate method for computing three-dimensional models of membrane protein sequences. Although human membrane proteins can sometimes be modeled with useful accuracy by comparative modeling on the basis of structures of their homologs from other organisms, the utility of prokaryotic structures for modeling of their human homologs is often limited[13], owing to low sequence and structure similarity.

To significantly increase the number of proteins with characterized structures, the US National Institutes of Health established the Protein Structure Initiative (PSI)[14], which includes, in the current PSI:Biology stage, four large-scale centers focused on globular proteins (http://sbkb.org/kb/psi_centers.html/) and nine specialized centers focused on membrane proteins (http://sbkb.org/kb/membprothub.html/). The PSI has maximized structural characterization of the protein-sequence space by an efficient combination of experimentation and prediction. Thus, selection of target

A full list of affiliations appears at the end of the paper.

Received 16 August 2012; accepted 9 January 2013; published online 5 February 2013; doi:10.1038/nsmb.2508

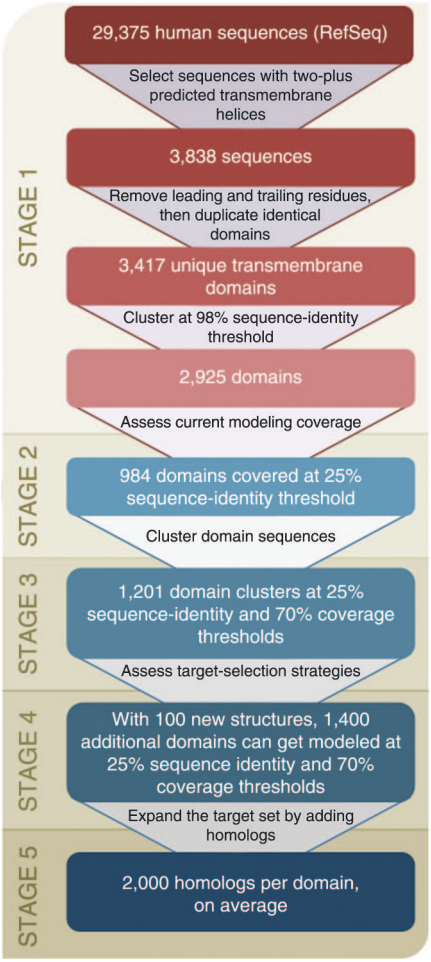

Figure 1 Flowchart of the analysis (**Supplementary Notes 1** and **2**). In stage 1, α-helical transmembrane regions of all human sequences from the RefSeq-37 database[32] were predicted by TMHMM2.0 (ref. 33; **Supplementary Fig. 1a**) and clustered with USEARCH[34] at 98% sequence identity, thus resulting in 2,925 unique α-helical transmembrane domains with at least two predicted transmembrane helices. In stage 2, the current modeling coverage of these domains was assessed by automated comparative modeling using ModPipe[35] (**Supplementary Fig. 1b**). In stage 3, the 2,925 unique α-helical transmembrane domains were clustered at 25% sequence identity and 70% coverage thresholds by using BLASTCLUST[36], thus resulting in 1,201 domain clusters (**Supplementary Table 2** and **Supplementary Fig. 2**). In stage 4, several target lists, including existing target lists of the nine PSI:Biology centers, were assessed by mapping the number and quality of models as a function of the number of targets; for example, if representative structures for the 100 largest clusters were determined, 1,400 additional α-helical transmembrane domains could be modeled on the basis of at least 25% sequence identity to the closest known structure (**Supplementary Fig. 3**). In stage 5, the target-sequence set was expanded by adding homologous sequences from other organisms, extracted from UniProtKB[23] (**Supplementary Fig. 4**).

proteins is key to maximizing coverage of the protein universe.

A number of target-selection schemes have been used in the past, ranging from focusing on only novel sequences or large families with no structural representative to selecting all proteins in a model genome[3,15–19]. These schemes organize the proteins of interest into clusters of related sequences, which are generally expanded by including additional homologs from other organisms. Any member of a cluster can then be a target for structure determination, because it is by definition sufficiently similar to the remaining cluster members to allow its experimental structure to serve as a template for comparative modeling[10]. As the sequence similarity between the target and its homologs increases, the accuracy of the resulting models also increases, but at the cost of having to determine a larger number of target structures. Consequently, a target-selection scheme needs to balance the effort needed for structure determination of all targets with the accuracy of resulting comparative models.

In general, high-accuracy comparative models are based on more than 50% sequence identity to their templates and tend to have approximately 1-Å r.m.s. error for the main chain atoms[20]. Medium-accuracy models are based on 30–50% sequence identity and tend to have about 90% of the main chain atoms modeled with 1.5-Å r.m.s. error. Common errors in these models are side chain packing, core distortion and loop modeling errors, with occasional alignment errors. Comparative models based on less than 30% sequence identity are commonly considered low-accuracy comparative models because alignment errors tend to increase rapidly below this sequence-identity threshold. However, even at this low level of target-template similarity, the models can still be useful for some of the most demanding applications. For example, virtual screening against comparative models of a GPCR[21] and a solute carrier (SLC) transporter[22] on the basis of templates with less than 25% sequence identity were instrumental in discovering chemically novel small-molecule ligands. Thus, a useful threshold on sequence identity for selecting targets for experimental structure determination may be as low as 25%.

A fitting and unifying potential goal for the nine PSI membrane-protein centers involved in PSI:Biology is a comprehensive structural characterization of the human α-helical transmembrane domains. However, even for a large-scale effort, determining the structures of all human α-helical transmembrane proteins by X-ray crystallography, NMR spectroscopy or electron microscopy is not feasible in the foreseeable future. Therefore, an efficient target-selection strategy is useful and will have the largest impact on the broad scientific community. Here, we analyze several candidate target-selection schemes, thus assessing the feasibility of a comprehensive structural description of the human α-helical transmembrane proteome.

Analysis

The analysis was performed in five stages (**Fig. 1** and **Supplementary Notes 1–5**; http://salilab.org/membrane/). First, we identified the human α-helical transmembrane domain sequences (**Supplementary Fig. 1a**). Second, we assessed how many of them can currently be modeled by comparative modeling (**Supplementary Fig. 1b**). Third, we clustered the sequences (**Supplementary Fig. 2** and **Supplementary Table 1**). Fourth, we quantified the efficiency of two target-selection strategies (**Supplementary Fig. 3** and **Supplementary Table 2**) and compared the results with the current target lists of the nine PSI membrane-protein centers (**Supplementary Table 3**) to assess the degree to which structure determination following these strategies and lists would allow comparative modeling of the human membrane proteome. Fifth, to prepare lists of proteins that would enable structure determination of the most advantageous target proteins, we expanded the list of human domain targets by adding their homologs from the UniProtKB database[23] (**Supplementary Fig. 4**).

Target selection for structural genomics of human membrane proteins

Nine PSI membrane-protein centers have been funded to improve structure characterization of membrane proteins. Several of these centers focus on human membrane-protein families, and almost all centers aim to determine the structures of at least some human proteins or their homologs. This effort, combined with the broader membrane-protein structural biology community, should make a large impact on the structural coverage of the human transmembrane proteome, especially if a coordinated target-selection strategy is pursued. A coordinated approach to future target selection for the nine membrane-protein centers seems to be reasonable, particularly if the resulting increase in the structural coverage is significant.

For this study, we relied on established methods to predict α-helical transmembrane domains (**Supplementary Note 1**). The relatively low number of known structures of β-barrel transmembrane proteins, combined with a less prominent hydrophobic profile, makes it more difficult to develop

reliable computational methods for predicting β-barrel transmembrane domains on a genomic scale[1,24]. In addition, the estimated fraction of β-barrel transmembrane proteins in proteomes (2–3%)[3] is significantly lower than the fraction of α-helical transmembrane proteins (approximately 25%). For these reasons, we analyzed here only the human α-helical transmembrane-domain sequences.

Two key parameters of a target-selection strategy include the thresholds on the target-template sequence similarity and fraction of sequence modeled (modeling coverage). Because membrane-protein crystallization and structure determination is notoriously difficult, we considered a maximum possible accommodation on these thresholds, to maximize the number of sequences that can be modeled on the basis of a given number of new structures. For the human α-helical transmembrane domain sequences, accepting comparative models covering at least 60% of the domain sequence on the basis of at least 25% sequence identity to the closest template structure, 100 structures selected by the guided target selection would increase the number of modelable human α-helical transmembrane domains by more than a factor of two, from 26% to 58% (**Supplementary Note 2**).

Relevance and biological impact

To increase our confidence in the automatically computed clusters of sequences, we examined three well-defined superfamilies as examples: GPCRs, SLC transporters and claudins. On the basis of these examinations, we conclude that our automated clustering procedure reproduces previous detailed annotations, and thus the analysis of target selection is likely to be statistically robust.

G protein–coupled receptors. GPCR sequences were collected from GPCRDB[24]. This superfamily forms the largest cluster (616 sequences; **Supplementary Fig. 2a**), with the remaining 128 GPCR sequences forming several smaller clusters. Although a number of structures have recently been determined for this important class of membrane proteins[25], it is the diversity that is perhaps most intriguing and in need of further investigation. For example, the opioid receptors that have recently been structurally characterized differ only in a few residues in the orthosteric binding site but have drastically different pharmacological effects[26]. The key question from a structural-genomics perspective is what level of granularity (that is, how many structures) is needed to create reliable models of additional GPCRs. This superfamily serves as a control and a test case for answering the question about required structure-mapping granularity.

SLC transporters. Unlike the GPCRs, this important family of transporters is a very diverse set of sequences that are not all related by a common ancestor[27]; nevertheless, some members can share similar structural features despite weak sequence similarities. A total of 340 SLC sequences are included in 116 domain clusters with 1–14 cluster members (**Supplementary Fig. 2a**). The majority of the sequences in the clusters have been previously annotated as SLC members, with 5% being annotated as hypothetical protein, uncharacterized, fragment or similar. As with the GPCR superfamily, there is an important question about the granularity of the experimental structure set that is needed before other transporters can reliably be modeled.

Claudins. The claudin family cluster illustrates the impact that structural genomics can have on biology and medicine. The family of claudins in humans consists of 23 proteins, and three additional members were recently proposed[28]. Almost all of the 23 known human claudin sequences (20–27 kDa in size)[29] form a single cluster (**Supplementary Fig. 2a**) and contain four transmembrane helices with their two extracellular loops important for the formation of cell-cell interactions in tight junctions[30,31]. Despite the high importance of claudins, no structure has yet been determined for any claudin protein. Numerous studies of mutants, as well as freeze-fracture studies, have revealed a model of the function of individual domains[31]. The N terminus is located at the intracellular side and generally contains only seven residues. Despite the small size of the claudin monomers, claudins are challenging targets for structure determination because their transmembrane domains form multimeric complexes within the membrane. Even a single experimentally determined claudin structure is likely to result in a greatly increased understanding of the claudin function in molecular terms. Such a structure would also allow for the modeling of the other claudins, further increasing its impact.

Proposed target-selection strategy

To most efficiently bridge the gap between the human genomic sequences and membrane-protein-structure knowledge, we propose to collectively pursue structural studies from the largest 100 clusters in need of structural coverage. Such an effort would lead to a structural characterization of 100 additional protein families and increase the coverage of the human α-helical transmembrane proteome to 58% of all sequences (**Supplementary Note 2**).

The PSI:Biology mandate encourages membrane-protein centers to concentrate on a variety of important biological questions. Fortunately, the overlap between the proposed 100-target list and the current target lists of the individual membrane-protein centers is large (**Supplementary Table 3**); moreover, the largest membrane-protein families are also biologically important. Thus, the proposed 100-target list integrates the goals of maximizing structural characterization with the biological focuses of the individual centers.

Because determining structures of eukaryotic membrane proteins continues to be technically challenging, a target with several homologs from bacteria or archaea may be a good initial or alternate target, provided that it can be used to further biological research or to compute sufficiently accurate models of its eukaryotic homologs. Thus, the centers will routinely process several related homologous sequences through their structure-determination methods to maximize the likelihood of determining a structure from a family (**Supplementary Note 2** and **Supplementary Fig. 2a**).

Lastly, coordination of the PSI:Biology targets with the structural-biology and broader scientific communities will help to more efficiently advance the field where follow-up studies are conducted to understand specific systems in more depth. This analysis provides a comprehensive assessment of the human α-helical transmembrane proteome, the state of the field today and what the field can accomplish in the next few years. By identifying sequence families that can be better characterized through the solution of the structures of a few homologs (from either eukaryotic or prokaryotic sources), membrane biologists can select the best templates and models for any membrane-protein sequence of as-yet-undetermined structure or understand the reliability of the model with sufficient structural coverage. Structural biologists can readily identify the impact that further investigations of specific structures can have or that any particular new membrane protein structure will have on the knowledge of human membrane proteome.

Availability

A computational resource, the Membrane Protein Hub (http://sbkb.org/kb/membprothub.html/), has recently been established as part of the Structural Biology Knowledgebase in collaboration between PSI and Nature Publishing Group. The purpose is to disseminate the results of the nine PSI membrane-protein centers. All results of the current study are accessible in their own knowledgebase through the "human TM proteome" link on the Membrane Protein Hub home page and through http://salilab.org/membrane/ (**Supplementary Note 3**).

Note: Supplementary information is available at http://www.nature.com/doifinder/10.1038/nsmb.2508.

ACKNOWLEDGMENTS

We thank I. Wilson, H. Berman, J. Chin and P. Preusch for critical comments on the manuscript. Research was supported by the US National Institutes of Health PSI:Biology grants U54 GM094662 (A.S., U.P.), U54 GM094618 (R.C.S.), U54 GM094625 (R.M.S., A.S., U.P.), U54 GM094584 (B.G.F.), U54 GM094599 (P.F.), U54 GM094611 (M.C.W., M.E.D., M.G.M.), U54 GM094610 (G.A.C., D.C.R., M.H.B.S.), U54 GM094608 (J.J.C.), U54 GM095315 (W.A.H., B.R., E.K.) and U54 GM094598 (D.L.S.).

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Punta, M. *et al. Methods* **41**, 460–474 (2007).
2. Fagerberg, L., Jonasson, K., von Heijne, G., Uhlen, M. & Berglund, L. *Proteomics* **10**, 1141–1149 (2010).
3. Punta, M. *et al. J. Struct. Funct. Genomics* **10**, 255–268 (2009).
4. UniProt Consortium. *Nucleic Acids Res.* **40**, D71–D75 (2012).
5. Hopkins, A.L. & Groom, C.R. *Nat. Rev. Drug Discov.* **1**, 727–730 (2002).
6. Giacomini, K.M. *et al. Nat. Rev. Drug Discov.* **9**, 215–236 (2010).
7. Rose, P.W. *et al. Nucleic Acids Res.* **39**, D392–D401 (2011).
8. Kloppmann, E., Punta, M. & Rost, B. *Curr. Opin. Struct. Biol.* **22**, 326–332 (2012).
9. Hopf, T.A. *et al. Cell* **149**, 1607–1621 (2012).
10. Baker, D. & Sali, A. *Science* **294**, 93–96 (2001).
11. Dunbrack, R.L. Jr. *Curr. Opin. Struct. Biol.* **16**, 374–384 (2006).
12. Liu, T., Tang, G.W. & Capriotti, E. *Comb. Chem. High Throughput Screen.* **14**, 532–547 (2011).
13. Granseth, E., Seppala, S., Rapp, M., Daley, D.O. & Von Heijne, G. *Mol. Membr. Biol.* **24**, 329–332 (2007).
14. Norvell, J.C. & Berg, J.M. *Structure* **15**, 1519–1522 (2007).
15. Vitkup, D., Melamud, E., Moult, J. & Sander, C. *Nat. Struct. Biol.* **8**, 559–566 (2001).
16. Marsden, R.L. & Orengo, C.A. *Methods Mol. Biol.* **426**, 3–25 (2008).
17. Raftery, J. *Methods Mol. Biol.* **426**, 37–47 (2008).
18. Büssow, K. *et al. Microb. Cell Fact.* **4**, 21 (2005).
19. Kelly, L. *et al. J. Struct. Funct. Genomics* **10**, 269–280 (2009).
20. Martí-Renom, M.A. *et al. Annu. Rev. Biophys. Biomol. Struct.* **29**, 291–325 (2000).
21. Carlsson, J. *et al. Nat. Chem. Biol.* **7**, 769–778 (2011).
22. Schlessinger, A. *et al. Proc. Natl. Acad. Sci. USA* **108**, 15810–15815 (2011).
23. UniProt Consortium. *Nucleic Acids Res.* **38**, D142–D148 (2010).
24. Vroling, B. *et al. Nucleic Acids Res.* **39**, D309–D319 (2011).
25. Katritch, V., Cherezov, V. & Stevens, R.C. *Trends Pharmacol. Sci.* **33**, 17–27 (2012).
26. Granier, S. *et al. Nature* **485**, 400–404 (2012).
27. Schlessinger, A. et al. *Protein Sci.* **19**, 412–428 (2010).
28. Mineta, K. *et al. FEBS Lett.* **585**, 606–612 (2011).
29. Escudero-Esparza, A., Jiang, W.G. & Martin, T.A. *Front. Biosci.* **16**, 1069–1083 (2011).
30. Angelow, S., Ahlstrom, R. & Yu, A.S. *Am. J. Physiol. Renal Physiol.* **295**, F867–F876 (2008).
31. Krause, G. *et al. Biochim. Biophys. Acta* **1778**, 631–645 (2008).
32. Larsson, T.P., Murray, C.G., Hill, T., Fredriksson, R. & Schioth, H.B. *FEBS Lett.* **579**, 690–698 (2005).
33. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E.L. *J. Mol. Biol.* **305**, 567–580 (2001).
34. Edgar, R.C. *Bioinformatics* **26**, 2460–2461 (2010).
35. Pieper, U. *et al. Nucleic Acids Res.* **39**, D465–D474 (2011).
36. Johnson, M. *et al. Nucleic Acids Res.* **36**, W5–W9 (2008).

[1]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, California, USA. [2]Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, California, USA. [3]California Institute for Quantitative Biosciences, University of California, San Francisco, San Francisco, California, USA. [4]Department of Bioinformatics and Computational Biology, Technical University Munich, Fakultät für Informatik, Garching, Germany. [5]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, California, USA. [6]Harvard University Medical School, Boston, Massachusetts, USA. [7]School of Medicine and Dentistry, University of Rochester Medical Center, Rochester, New York, USA. [8]Transmembrane Protein Center, University of Wisconsin–Madison, Madison, Wisconsin, USA. [9]Department of Chemistry and Biochemistry, Arizona State University, Tempe, Arizona, USA. [10]New York Structural Biology Center, New York, New York, USA. [11]Hauptman Woodward Medical Research Institute, Buffalo, New York, USA. [12]Division of Chemistry and Chemical Engineering, Howard Hughes Medical Institute, California Institute of Technology, Pasadena, California, USA. [13]Molecular, Cellular and Developmental Biology, University of Colorado Boulder, Boulder, Colorado, USA. [14]Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, Virginia, USA. [15]Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, California, USA. [16]Present address: Department of Pharmacology and Systems Therapeutics, Mount Sinai School of Medicine, New York, New York, USA. Correspondence should be addressed to A.S. (sali@salilab.org), R.C.S. (stevens@scripps.edu) or R.M.S. (stroud@msg.ucsf.edu).