Ratiocinative screen of eukaryotic integral membrane protein expression and solubilization for structure determination

Franklin A. Hays · Zygy Roe-Zurz · Min Li · Libusha Kelly · Franz Gruswitz · Andrej Sali · Robert M. Stroud

Received: 16 September 2008 / Accepted: 24 October 2008 / Published online: 22 November 2008

Abstract Persistent hurdles impede the successful determination of high-resolution crystal structures of eukaryotic integral membrane proteins (IMP). We designed a high-throughput structural genomics oriented pipeline that seeks to minimize effort in uncovering high-quality, responsive non-redundant targets for crystallization. This "discovery-oriented" pipeline sidesteps two significant bottlenecks in the IMP structure determination pipeline: expression and membrane extraction with detergent. In addition, proteins that enter the pipeline are then rapidly vetted by their presence in the included volume on a size-exclusion column—a hallmark of well-behaved IMP targets. A screen of 384 rationally selected eukaryotic IMPs in baker's yeast *Saccharomyces cerevisiae* is outlined to demonstrate the results expected when applying this discovery-oriented pipeline to whole-organism membrane proteomes.

Keywords Discovery-oriented screen · Membrane protein structure · Structural genomics · Eukaryotic integral membrane protein · *Saccharomyces cerevisiae*

Abbreviations

DDM	*n*-Dodecyl-β-D-maltopyranoside
IMAC	Immobilized metal affinity chromatography
SEC	Size exclusion chromatography
TMH	Transmembrane helices
LIC	Ligase independent cloning

Franklin A. Hays and Zygy Roe-Zurz have contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s10969-008-9046-7) contains supplementary material, which is available to authorized users.

F. A. Hays (✉) · R. M. Stroud (✉)
Department of Biochemistry and Biophysics, University
of California at San Francisco, San Francisco,
CA 94158-2517, USA
e-mail: haysf@msg.ucsf.edu

R. M. Stroud
e-mail: stroud@msg.ucsf.edu

Z. Roe-Zurz · M. Li · F. Gruswitz
Membrane Protein Expression Center, University of California
at San Francisco, San Francisco, CA 94158-2517, USA

L. Kelly
Graduate Group in Bioinformatics, University of California
at San Francisco, San Francisco, CA 94158-2517, USA

L. Kelly · A. Sali
Center for the Structure of Membrane Proteins, University
of California at San Francisco, San Francisco,
CA 94158-2517, USA

L. Kelly · A. Sali
Department of Pharmaceutical Chemistry, University
of California at San Francisco,
San Francisco, CA 94158-2517, USA

L. Kelly · A. Sali
California Institute for Quantitative Biosciences,
University of California at San Francisco,
San Francisco, CA 94158-2517, USA

A. Sali
Department of Bioengineering and Therapeutic Sciences,
University of California at San Francisco, San Francisco,
CA 94158-2517, USA

OG	*n*-Octyl-*β*-D-glucoside
FC-12	*n*-Dodecylphosphocholine
IMP	Integral membrane protein

Introduction

Can a high-output structural genomics style pipeline be deployed successfully to obtain structures of eukaryotic integral membrane proteins? If so, how would such a pipeline be developed and implemented? What leaks and bottlenecks should one expect?

These questions, and many others, provided some of the framework for discussions at an NIH sponsored workshop in April of 2008. A stated objective of the workshop was to address "the challenges and technical barriers to the high-throughput determination of protein structures" (http://www.team-psa.com/NIGMS-BottlenecksWorkshop/). Indeed, integral membrane proteins, because of their added complexity, have often fallen outside high-throughput pipelines within the Protein Structure Initiative (PSI). Efforts within PSI-2 specialized centers are starting to yield positive dividends for membrane proteins [1] yet generic protocols for the reliable expression, membrane extraction with detergent, purification and even crystallization of this important class of targets are still being developed. In general, discussions relating to IMP high-throughput structure determination focus both on efficient implementation of established methods and the development of novel tools and approaches. The current mini-review will focus on common impasses and hurdles along the road to eukaryotic IMP structure determination and the implementation of existing methodologies within a discovery-oriented pipeline. This pipeline was the topic of an oral presentation at the NIH Bottlenecks Workshop mentioned above.

Principal of factor sparsity: hurdles in eukaryotic IMP structure determination

The Pareto Principle, whereby highest value derives from small numbers of most favorable cases, is used to describe the unequal relationship between inputs and outputs within systems [2]. This principle is applicable to efforts in pursuing the structures of novel eukaryotic IMPs since the majority of the effort leading to a structure is often spent troubleshooting. These diagnostic ventures typically include exploration of various expression systems, solubilization detergents, orthologous proteins, purification and buffer conditions, ligands and stabilizing mutations, and crystallization optimization. The current body of IMP structures bears this out with the broad range of methods employed to obtain, purify and crystallize the protein of interest. These broad and inclusive applications also require heroic effort, time and expense. All of the approximately 94 unique α-helical IMP structures (Supplementary Table) currently available have resulted from protracted efforts focused on a specific functional class or family of protein. Examples of this approach are the *β*2-adrenergic receptor [3–5], or the Kv1.2 potassium channel [6]. What characterizes this intensive approach is the progression and modification of methods to a specific target protein of known function. These efforts, though often time consuming and laborious, have been increasingly successful in producing novel IMP structures (Fig. 1). Unfortunately, such an intensive approach is inherently untenable in high- or medium-throughput pipelines as streamlining the process is not readily possible. Thus, if one is strictly interested in structural genomics-directed efforts for the elucidation of novel eukaryotic IMP structures then two apparent directions can be pursued: (1) develop novel methods and reagents to streamline the process or (2) use standard methods and reagents in a novel, streamlined way. In addressing the latter, we have asked what methods and systems have worked to date in generating novel structures, and how these can be used in the discovery-oriented pipeline of a structural genomics approach to IMPs.

A discovery-oriented pipeline can be constructed by harvesting successful methods from the IMP literature to develop pragmatic criteria that supervise the flow of

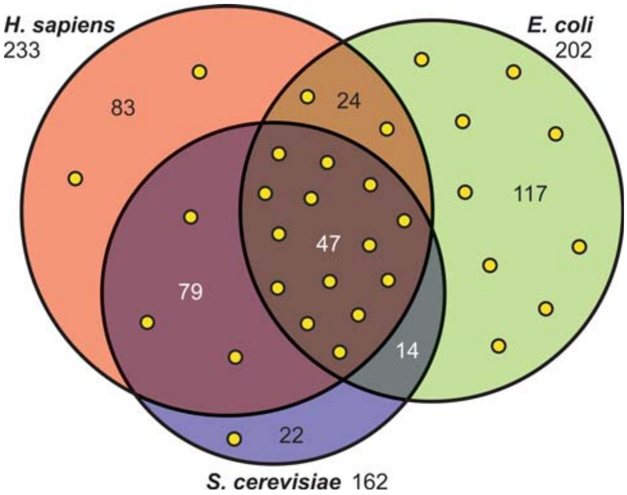

Fig. 1 Protein family homology for *S. cerevisiae*. Venn diagram with protein families (Pfam) for *H. sapiens* (*light red*), *E. coli* (*light green*) and *S. cerevisiae* (*light blue*) represented by each respective *circle*. Total number of membrane protein associated Pfams for each organism is listed beside the organism name of each while number of Pfams within each node are shown within the respective sections of the diagram. *Yellow circles* represent a Pfam with a solved representative x-ray structure within that node. Structures may be from an organism not explicitly included within the diagram if an associated Pfam is represented within the node

targets. This is a key distinction—a discovery-oriented pipeline constitutes a progression of targets (proteins) with standardized methods while a systems-oriented approach requires the development of methods and protocols to facilitate the progression of a specific target. In general, there are four typical hurdles to overcome in pursuing eukaryotic IMP structures: (1) producing sufficient quantities of functional protein, (2) finding an appropriate detergent for membrane extraction of the target protein, (3) purification of the target protein, and (4) optimizing initial crystal hits to obtain sufficient resolution data for structure determination. Finding a truly suitable formula that matches multiple IMPs is difficult. If one were to implement a structural genomics style pipeline for eukaryotic IMP targets, could compromises be identified that still result in high-output results while retaining information on the probability for increased success through additional tuning? We have shown that there are and published results serve as an instructive guide.

Yeast is a viable system for homologous or heterologous overexpression of eukaryotic integral membrane proteins. To date, seven of the 13 heterologously expressed integral membrane protein structures were produced in some form of yeast [4, 6–11]. This may not be entirely surprising as “yeast”, especially *Pichia pastoris* or *Saccharomyces cerevisiae*, have a number of advantages for the production of eukaryotic IMP targets including: proper membrane targeting and insertion machinery, capabilities for post-translational modifications and a lower activation barrier and cost relative to other systems (e.g. sHEK293S or baculovirus). Furthermore, *S. cerevisiae* is attractive because of its applicability to high-throughput cloning and expression trials via episomal plasmids [12, 13]. An extensive library of non-lethal knockouts in the ‘Yeast Knockout’ deletion collection [14] makes for a well-characterized platform for downstream functional studies [15]. The observation that yeast is a suitable system for the overexpression of eukaryotic IMPs is in itself not new [12, 15, 16], yet worth reiterating for the purposes of this discussion. Thus, baker’s yeast is an appropriate system to utilize in a general eukaryotic IMP structure determination pipeline. The *S. cerevisiae* system is not expected to be optimal for every eukaryotic protein of interest but for generalized and streamlined screens it can produce reasonable indications of IMPs that seem more tractable. In addition, these results can be obtained relatively quickly—moving from initial cloning to expression/solubilization trials in just 2 weeks.

To date, DDM has been heavily favored in detergent-mediated extraction and purification of heterologously expressed eukaryotic IMP structures, accounting for nine out of the 13 known [4, 6, 8–10, 17–20]. Thus, if only one detergent is used, detergent *n*-dodecyl-*β*-D-maltoside (DDM) is an effective choice for broad screens of eukaryotic IMP targets both in its ability to solubilize specific targets and not impede initial crystallization. For pursuing prokaryotic targets an appropriate choice may be *n*-octyl-*β*-D-glucoside (OG). DDM is a compromise between shorter chain, and often less effective, detergents like OG and longer chain, and often more effective, zwitterionic detergents such as *n*-dodecylphosphocholine (FC-12). DDM is sufficiently strong to extract targets from the membrane while not being astringent enough to denature or inactivate them. In addition, DDM has been shown to encompass the larger area of Dumont’s Venn diagram [16], or the full range of target solubilization of other detergents that are more amenable to crystallization, namely DM, NG, and OG. Since detergent exchange can be considered later in the purification stage of the pipeline, DDM can comfortably be used as the sole detergent to initially solubilize targets while refining alternative detergent/lipid mixtures for crystallization. Using DDM further reduces the cost of these broad screens since it has a relatively low critical micelle concentration (CMC $\approx$ 0.12 mM in 200 mM NaCl) thereby requiring substantially less detergent in solubilization and purification buffers. Thus, using a single detergent for solubilization is both economic of protein source, and amenable to the high-throughput extensive phase of our discovery-oriented pipeline.

Implementation of streamlined purification protocols for eukaryotic IMPs is very challenging. Individual targets often require personalized attention to ensure stability and monodispersity within a given buffer condition and detergent. Fortunately, purification methods for IMP targets are largely the same as those for soluble proteins, with addition of detergent and lipids to protein containing micelles. Attendant complexities occur as a result of excess protein-free micelles, multiple aggregation states, instability, or precipitation of target protein.

Cleavable poly-histidine affinity tags greatly facilitate initial purification steps using immobilized metal affinity chromatography. Because of signal peptide processing considerations, carboxyl-terminal tags need to be chosen over amino-terminal ones in a moderately broad screen of IMPs. Enzyme inhibition in the presence of detergents may occur when using TEV protease [21]. Two proteases that work well in detergents for cleaving tags are human rhinovirus 3C and thrombin. Evaluating targets based on their yields post-IMAC (immobilized metal affinity chromatography) is a rigorous quantitative method with potentially more utility than expression levels based on GFP fluorescence, or western blot analysis. Size-exclusion chromatography (SEC) is a robust approach to evaluate protein homogeneity. Presence of the IMP target within the included volume of an SEC column following solubilization with a detergent signifies a non-aggregated state. Fully included targets can generally be purified to homogeneity and move on to crystallization trials. Ion exchange chromatography may be leveraged not only for purification but also for detergent exchange or to pre-concentrate protein without excess detergent prior to crystallization. When concentrating purified membrane proteins one must be very conscious of, and strive to eliminate, protein-free detergent micelles within the solution which may concentrate with the protein leading to heavy phase separation during crystallization. This is a major source of negative results at the crystallization stage.

Crystallization is the last hurdle *en route* to structure by x-ray crystallography. Structural genomics efforts have produced significant insights into this bottleneck through development of novel methods and instrumentation [22]. Crystallization and screening in lipid mesophases or various microfluidic techniques are two such examples [23, 24]. Developments in instrumentation around precision liquid handling robots have driven down the time, cost and sample volume required to obtain initial crystal hits for purified targets. Within the Center for Structures of Membrane Proteins (CSMP.ucsf.edu), one of the specialized PSI-2 Centers focused on IMPs, hanging drop vapor diffusion screens have produced very positive results in generating initial crystal hits. The vast majority of these hits are derived from readily available commercial screens—some of which are tailored specifically for IMPs and have also resulted from structural genomics efforts [25].

Obtaining initial crystallization hits may often not present a hurdle. Rather, it is the optimization to obtain crystals of sufficient quality for structure determination that is almost always a matter of fine-tuning, especially with the added complexity of micelle size and homogeneity. This phenomenon is a reinforcement of Pareto's Principle in that the final crystallization conditions which produced a crystal for structure determination often resulted from very expansive efforts. In addition, these efforts are rarely linear in nature. Success is often very target specific (e.g. transporter versus channel) and inherently not very conducive to high-throughput workflows, though some established pipelines would likely result in measured success when applied to IMP targets (such as the RIKEN Spring-8 [26] or JCSG pipelines [22, 27]). Additionally, pre- or post-crystallization screens may be implemented to increase success rates from final purification to determined structure. These include: (1) ^{1}H NMR to characterize the folded state of the protein, (2) fluorescence-based thermal melting assays [28] to increase protein stability, (3) screening of orthologous proteins, (4) detergent screens, (5) deletion and/or truncation constructs for specific targets, and (6) light scattering (dynamic or static) to access dispersity in solution.

Principal of Least Effort: discovery-oriented screen in *S. cerevisiae*

Principal of Least Effort is a theory that postulates a path of least resistance will always be chosen when given a choice [29]. Naturally, one must ask if such an approach can be applied to vetting eukaryotic integral membrane proteins for downstream structure determination. With this mindset, a discovery-oriented pipeline was constructed to address the hurdles described above. This pipeline is constructed from strict empirical criteria that define the progression of targets with each step being dependent on the previous, and ultimately necessary for structure determination. These steps include using one expression system (*S. cerevisiae*), a single detergent for membrane extraction (DDM) and a single buffer for SEC void checks (20 mM TRIS–HCl pH 7.4 RT, 200 mM NaCl, 1 mM DDM and 10% w/v glycerol). The targets that express well, are soluble in DDM and included in SEC move along at high velocity through the pipeline, while targets that do not meet these criteria are retained. The benefit of the discovery-oriented pipeline is evident when moving into the intensive phase, where sets of workable and responsive targets that have *a posteriori* cleared most of the major hurdles are all that remain. How much effort does such an extensive phase take in identifying these targets, and can it significantly increase the output to input ratio for the pipeline?

We were able to establish a discovery-oriented pipeline to survey 384, or four sets of 96, rationally selected integral membrane proteins representing every IMP protein family within the *S. cerevisiae* genome [30]. To quickly identify proteins amenable to large-scale purification, functional characterization, and crystallization, targets were cloned into a protease deficient *S. cerevisiae* strain, W303-Δpep4, and grown in 500 ml cultures. Of the 351 targets cloned, 234 (67%) of these expressed and solubilized in DDM (Table 1). The 61 that expressed in the first cohort of 96 targets (64%) were followed by growths in three—one liter cultures and

Table 1 Results from extensive screen of *S. cerevisiae* IMP's

	Set 1	Set 2	Set 3	Set 4	
Cloned	87 (91%)	91 (95%)	85 (89%)	88 (92%)	91%
Expressed	72 (75%)	72 (75%)	62 (65%)	66 (75%)	73%
Solubilized	61 (64%)	58 (61%)	52 (54%)	63 (65%)	61%
SEC included	23 (24%)	N/A	N/A	N/A	24%

Three hundred and eight-four yeast IMP targets were divided into four sets of 96 targets each (Set 1–4). Shown is the number of targets that were successfully cloned, expressed, solubilized or present within the included volume on SEC for targets progressing through the pipeline from each group. In parentheses, for the noted step, is the percent success rate relative to the starting 96 targets. The far right column denotes the average rate of success for that step. Only targets from Set 1 were pushed through the SEC step evaluated by post IMAC expression levels and quality of size exclusion characteristics in one buffer condition. This resulted in the identification of 23 (24% of starting number) targets with high expression level (>0.5 mg protein/l of culture), soluble in DDM, and fully included by SEC. Furthermore, the extensive phase of the pipeline was divided into three general categories—target selection, expression plasmid construction and target prioritization.

For target selection, all *S. cerevisiae*'s predicted gene product sequences (~6600) were fed into the program TMHMM [31], which predicted 621 proteins to have three or more transmembrane helices (TMH). A minimum of three transmembrane helices was specifically chosen in order to omit secreted, monotopic or membrane associated proteins that are merely membrane-anchored or contain signal peptides. Although this minimum eschews important classes of IMPs, such as oligomeric channel-forming transmembrane proteins with fewer than three helices, the signal peptide prediction algorithms are simply not robust enough at the present time to identify two-TMH eukaryotic IMPs [32] with certainty. Parsing through the 621 targets evinced 162 unique Pfam membrane protein families represented in yeast. One target was chosen from each of the identified Pfams and, for those Pfams represented by more then one protein (83), we chose two targets. In addition, 131 proteins could not be annotated with a known Pfam. Each of these targets was placed in the pipeline generating a cohort of 384 protein targets (4 × 96 to facilitate cloning in 96-well format) representing all annotated and unannotated IMP families within *S. cerevisiae*.

Ligase independent cloning (LIC) was used to construct expression vectors in a high-throughput format. Genes were inserted into a 2 μ based *S. cerevisiae* LIC expression plasmid with an N-terminal FLAG tag followed by a 3C protease cleavage site and a C-terminal 10XHis tag preceded by a thrombin protease cleavage site. Meticulous control of protein induction was accomplished by selecting the GAL1 promoter. 351 out of 384 targets were cloned in the initial pass (91% success rate), with a throughput of up to 192 clones per week.

Target prioritization began with the growth of 500 ml culture volumes to test expression and detergent solubility for each of the 351 cloned constructs. Two hundred and seventy-two constructs expressed in the system and 234 were soluble in DDM, producing a 61% success rate after passing both expression and solubilization bottlenecks (Table 1). Targets were then prioritized by expression level and detergent solubilization based on qualitative analysis of western blots. Protein stability and integrity were ascertained for the 61 constructs that showed both expression and solubility in the first cohort of 96. These yeast IMPs contained the most diversity of the protein families within *S. cerevisiae* being single representatives of the selected Pfam families. Every target was grown in three liters of yeast culture, and screened in a single buffer condition: 20 mM TRIS–HCl pH 7.4 RT, 200 mM NaCl, 1 mM DDM and 10% w/v glycerol. Thirty-one out of 61 targets, or 51%, were found to reside mostly (>50%) in the SEC included volume. Twenty-three out of 61, or 38%, were fully included and of high quality for downstream studies (Table 1). This corresponds to 24% retention of targets through the extensive phase of our pipeline even while applying relatively strict and limited guidelines for target progression (single detergent, single SEC buffer, etc.). This 24% rate of retentiveness has parity with the 25% obtained for a set of globular prokaryotic targets in a recent systems-oriented screen [33]. For large Pfams with multiple members, such as the Major Facilitator Superfamily, we selected the most representative member of that family based upon multiple sequence alignment profiles [30].

Conclusions

Pursuing the structure of integral membrane proteins has, rather aptly, been referred to as "siege warfare" [34]. This depiction highlights the operose nature of membrane protein structure determination that, by many indications, is entering a Renaissance period of accelerated growth. The number of research groups pursuing IMP structures, deployment of novel methodologies and equipment, and ramping up of structural genomics efforts will all have a significant affect on the trajectory of this growth. The resulting biological insights from such efforts will likely be significant considering the current paucity of structures and relative importance membrane proteins play within biological systems. Laborious target-specific efforts will continue to pay dividends in producing structures. The outstanding question remains to what extent, and how, structural genomics efforts will play in the increased rate at which IMP structures are determined. The current mini-review focuses on the implementation of a streamlined pipeline designed to identify well-behaved targets early on to facilitate increased returns during crystallization and structure determination.

Four hurdles have been addressed, each in a single 'most probable' fashion—expression, solubilization, purification and crystallization. The resulting discovery-oriented pipeline is a stripped-down way of trying to best circumvent the expression and solubilization hurdles leading up to crystallization. In selecting a single expression system and screening a large number of targets one can quickly select eukaryotic IMPs that express above a predefined threshold (e.g. 0.5 mg protein/l of culture for the culture). Solubility screens using DDM for each of the expressed IMPs allows for rapid identification of targets for subsequent scale-up. If required, downstream screens with shorter chain detergents can then be performed on targets that are solubilized by DDM, or longer chain detergents can be tested for those that are not extracted from the membrane with DDM. Determining if DDM soluble targets are within the included volume on SEC using standardized protocols allows for rapid and efficient initial characterization of each target. Comparing the expression, solubilization and initial SEC profiles for each target provides information for developing a target priority list for a subsequent production phase of the pipeline.

For the yeast screen this resulted in a list of 23 proteins of 61 attempted that expressed to appreciable levels, were extracted from the membrane in DDM and fully included on a SEC column [30]. Each of these proteins has now been pushed into a more intensive production phase oriented towards detailed characterization. Four of these targets have now entered crystallization trials; two of them have been shown to crystallize and one has diffracted below 3 Å resolution. Although this discovery-oriented approach has yet to be validated through the emergence of novel eukaryotic IMP structures, we feel the initial results are very promising. This approach comes at a cost of not being attached to specific protein identities, or functional class, until later in the pipeline. Also, attrition rates will be high initially as targets are vetted with the selected expression system, solubilization detergent and SEC buffer. Targets that are retained (i.e. fail to progress) within the first pass can then be pushed through a salvage pipeline using a new detergent or buffer condition for screening.

Is target selection needed for such a relatively simplistic screen? If using high-throughput techniques, including ligase independent cloning, combined with small scale expression and solubilization tests the answer may very well be no. Upwards of 200 targets can be pushed from initial cloning from genomic DNA stock to analysis of expression/solubilization data within three weeks using *S. cerevisiae*. At this rate entire integral membrane proteomes can be screened to rapidly identify the more tractable IMPs. This does not imply that the ones that fall 'below the bar' could not be expressed with other methods or expression systems that would be standard for the more traditional approach to a particular membrane protein class.

For large membrane proteomes the amount of work during the void check and initial purification stage may be significantly reduced by reducing the number of targets initially screened. In this case target selection may provide some significant benefits. With biomedically relevant starting sets (e.g. Table 2) even a modest yield of 10% for each organism would be incredibly insightful. For homologous overexpression within yeast the return is 24% of targets being soluble in DDM and fully included in SEC [30]. Significant attrition could be expected at each stage yet those results would feed back into the pipeline allowing for changes in detergent selection (e.g. FC-12 versus DDM) and screening buffers more broadly post-IMAC. For larger membrane proteomes, such as human, a ratiocinative selection that minimizes the number of targets while maximizing coverage of represented protein families would be beneficial.

Table 2 Number of integral membrane proteins with three or more transmembrane helices (TMH) in selected eukaryotes

Eukaryote	Total sequences	Predicted 3+ TMH
Cryptosporidium hominis	3886	310
Cryptosporidium parvum	3806	348
Plasmodium vivax	5334	378
Plasmodium falciparum	5342	513
Toxoplasma gondii	7787	540
Saccharomyces cerevisiae	6600	622
Leishmania major	8009	625
Trypansosoma brucei	8965	663
Trypanosoma cruzi	19245	1435
Drosophila melanogaster	17104	1805
Homo sapiens	32010	3158
Mus musculus	30133	3778

"Total Sequences" represents the total number of predicted proteins within the specified genome. "Predicted 3+ TMH" was determined by running the total number of coding sequences through TMHMM [31] in batch

Novel IMP structures derived from such extensive screens are likely to yield significant biological insights considering the current paucity of available IMP structures. Indeed, the Protein Data Bank, as of November 2007, contains only 94 unique α-helical IMP structures containing three or more transmembrane helices. Unique in this case is defined as greater then 95% sequence identity to remove point mutants and other slight sequence modifications. These 94 structures represent only 37 protein families, of the 598 identified with three or more transmembrane helices, with high resolution structural data available (or 6%) (Kelly et al. manuscript submitted). Using the *Homo sapiens*, *Saccharomyces cerevisiae*, and *E. coli* genomes as representative organisms, we analyzed the number of Pfam IMP families by organism to assess the structural coverage in each (Fig. 1). Of the 37 families, 21 are represented in the human genome, 25 in *E. coli* and 19 within yeast. Of these, 17 Pfams are in human and yeast, 16 in human and *E. coli*, 14 in yeast and *E. coli* and 14 are represented in all three organisms. Five Pfams with high resolution structures are not represented in any of the three organisms. Eight families are found only in *E. coli*, including the drug resistance-associated Acr transporter family (PF00873) and the disulfide bond formation protein family DsbB (PF02600).

The two families only represented in human are the sodium:neurotransmitter symporter family (PF00209) and the inward rectifier potassium channels (PF01007). Finally, one family was only represented in yeast—the bacteriorhodopsin family that includes the fungal rhodopsin homologs (PF01036). Only six of the 184 eukaryote-only Pfams within this analysis have associated structures. The majority of structures within the 85 Pfams present within both *E. coli* and either eukaryote (human or yeast) are derived from prokaryotic homologs. This highlights the observation that eukaryotic IMPs are more experimentally difficult to pursue. Thus, considering 94% of the IMP protein families (three TMH or more) do not have a representative structure and, when available, that structure is often derived from a prokaryotic source any avenue to obtain eukaryotic structures more efficiently could be deemed beneficial.

Acknowledgements This review describes, in part, a broad screen which was published previously in The *Journal of Molecular Biology* [30]—we are grateful to all co-authors involved in bringing that project to fruition. This work was supported by the N.I.H. Roadmap Center Grant P50 GM073210 (to R.M.S.), Specialized Center for the Protein Structure Initiative Grant U54 GM074929 (to R.M.S. and A.S.), National Research Service Award F32 GM078754 (to F.A.H.) and a Sandler Biomedical Research postdoctoral fellowship (to F.A.H.).

References

1. Lunin VV, Dobrovetsky E, Khutoreskaya G, Zhang R, Joachimiak A, Doyle DA, Bochkarev A, Maguire ME, Edwards AM, Koth CM (2006) Nature 440:833–837. doi:10.1038/nature04642
2. Reed WJ (2001) Econ Lett 74:15–19. doi:10.1016/S0165-1765(01)00524-9
3. Cherezov V, Rosenbaum DM, Hanson MA, Rasmussen SG, Thian FS, Kobilka TS, Choi HJ, Kuhn P, Weis WI, Kobilka BK, Stevens RC (2007) Science 318:1258–1265. doi:10.1126/science.1150577
4. Rasmussen SG, Choi HJ, Rosenbaum DM, Kobilka TS, Thian FS, Edwards PC, Burghammer M, Ratnala VR, Sanishvili R, Fischetti RF, Schertler GF, Weis WI, Kobilka BK (2007) Nature 450:383–387. doi:10.1038/nature06325
5. Rosenbaum DM, Cherezov V, Hanson MA, Rasmussen SG, Thian FS, Kobilka TS, Choi HJ, Yao XJ, Weis WI, Stevens RC, Kobilka BK (2007) Science 318:1266–1273. doi:10.1126/science.1150609
6. Long SB, Campbell EB, Mackinnon R (2005) Science 309:897–903. doi:10.1126/science.1116269
7. Horsefield R, Norden K, Fellert M, Backmark A, Tornroth-Horsefield S, Terwisscha van Scheltinga AC, Kvassman J, Kjellbom P, Johanson U, Neutze R (2008) Proc Natl Acad Sci USA 105:13327–13332. doi:10.1073/pnas.0801466105
8. Long SB, Tao X, Campbell EB, MacKinnon R (2007) Nature 450:376–382. doi:10.1038/nature06265
9. Ago H, Kanaoka Y, Irikura D, Lam BK, Shimamura T, Austen KF, Miyano M (2007) Nature 448:609–612. doi:10.1038/nature05936
10. Pedersen BP, Buch-Pedersen MJ, Morth JP, Palmgren MG, Nissen P (2007) Nature 450:1111–1114. doi:10.1038/nature06417
11. Tornroth-Horsefield S, Wang Y, Hedfalk K, Johanson U, Karlsson M, Tajkhorshid E, Neutze R, Kjellbom P (2006) Nature 439:688–694. doi:10.1038/nature04316
12. Newstead S, Kim H, von Heijne G, Iwata S, Drew D (2007) Proc Natl Acad Sci USA 104:13936–13941. doi:10.1073/pnas.0704546104
13. Kim H, Melen K, Osterberg M, von Heijne G (2006) Proc Natl Acad Sci USA 103:11142–11147. doi:10.1073/pnas.0604075103
14. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, Chu AM, Connelly C, Davis K, Dietrich F, Dow SW, El Bakkoury M, Foury F, Friend SH, Gentalen E, Giaever G, Hegemann JH, Jones T, Laub M, Liao H, Liebundguth N, Lockhart DJ, Lucau-Danila A, Lussier M, M'Rabet N, Menard P, Mittmann M, Pai C, Rebischung C, Revuelta JL, Riles L, Roberts CJ, Ross-MacDonald P, Scherens B, Snyder M, Sookhai-Mahadeo S, Storms RK, Veronneau S, Voet M, Volckaert G, Ward TR, Wysocki R, Yen GS, Yu K, Zimmermann K, Philippsen P, Johnston M, Davis RW (1999) Science 285:901–906. doi:10.1126/science.285.5429.901
15. Bill RM (2001) Curr Genet 40:157–171. doi:10.1007/s002940100252
16. White MA, Clark KM, Grayhack EJ, Dumont ME (2007) J Mol Biol 365:621–636. doi:10.1016/j.jmb.2006.10.004
17. Ferguson AD, McKeever BM, Xu S, Wisniewski D, Miller DK, Yamin TT, Spencer RH, Chu L, Ujjainwalla F, Cunningham BR, Evans JF, Becker JW (2007) Science 317:510–512. doi:10.1126/science.1144346
18. Jasti J, Furukawa H, Gonzales EB, Gouaux E (2007) Nature 449:316–323. doi:10.1038/nature06163
19. Nishida M, Cadene M, Chait BT, MacKinnon R (2007) EMBO J 26:4005–4015. doi:10.1038/sj.emboj.7601828
20. Standfuss J, Xie G, Edwards PC, Burghammer M, Oprian DD, Schertler GF (2007) J Mol Biol 372:1179–1188. doi:10.1016/j.jmb.2007.03.007
21. Mohanty AK, Simmons CR, Wiener MC (2003) Protein Expr Purif 27:109–114. doi:10.1016/S1046-5928(02)00589-2
22. Lesley SA, Wilson IA (2005) J Struct Funct Genomics 6:71–79. doi:10.1007/s10969-005-2897-2
23. Cherezov V, Clogston J, Papiz MZ, Caffrey M (2006) J Mol Biol 357:1605–1618. doi:10.1016/j.jmb.2006.01.049
24. Li L, Mustafi D, Fu Q, Tereshko V, Chen DL, Tice JD, Ismagilov RF (2006) Proc Natl Acad Sci USA 103:19243–19248. doi:10.1073/pnas.0607502103
25. Newstead S, Ferrandon S, Iwata S (2008) Protein Sci 17:466–472. doi:10.1110/ps.073263108
26. Sugahara M, Asada Y, Shimizu K, Yamamoto H, Lokanath NK, Mizutani H, Bagautdinov B, Matsuura Y, Taketa M, Kageyama Y, Ono N, Morikawa Y, Tanaka Y, Shimada H, Nakamoto T, Sugahara M, Yamamoto M, Kunishima N (2008) J Struct Funct Genomics. doi:10.1007/s10969-008-9042-y
27. Abola E, Carlton DD, Kuhn P, Stevens RC (2007) In: Jhoti H, Leach A (eds) Structure-based drug discovery. Springer, Netherlands, pp 1–26
28. Vedadi M, Niesen FH, Allali-Hassani A, Fedorov OY, Finerty PJ Jr, Wasney GA, Yeung R, Arrowsmith C, Ball LJ, Berglund H, Hui R, Marsden BD, Nordlund P, Sundstrom M, Weigelt J, Edwards AM (2006) Proc Natl Acad Sci USA 103:15835–15840. doi:10.1073/pnas.0605224103
29. Zipf GK (1949) Human behavior and the principle of least effort: an introduction to human ecology. Addison-Wesley, Reading
30. Li M, Hays FA, Roe-Zurz Z, Vuong L, Kelly L, Robbins R, Ho C, Pieper U, O'Connell J III, Giacomini KM, Sali A, Stroud RM (2008) J Mol Biol (in press)

31. Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) J Mol Biol 305:567–580. doi:10.1006/jmbi.2000.4315

32. Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Nat Protoc 2:953–971. doi:10.1038/nprot.2007.131

33. Lewinson O, Lee AT, Rees DC (2008) J Mol Biol 377:62–73. doi:10.1016/j.jmb.2007.12.059

34. Wiener MC (2004) Methods 34:364–372. doi:10.1016/j.ymeth.2004.03.025