A limited universe of membrane protein families and folds

AMIT OBERAI,¹ YUNGOK IHM,¹ SANGUK KIM,² AND JAMES U. BOWIE¹

¹Department of Chemistry and Biochemistry, UCLA-DOE Institute for Genomics and Proteomics, Molecular Biology Institute, Los Angeles, California 90095-1570, USA

²Department of Life Science POSTECH, Pohang, Kyungbuk 790-784, Korea

(RECEIVED January 23, 2006; FINAL REVISION March 27, 2006; ACCEPTED March 27, 2006)

Abstract

One of the goals of structural genomics is to obtain a structural representative of almost every fold in nature. A recent estimate suggests that 70%–80% of soluble protein domains identified in the first 1000 genome sequences should be covered by about 25,000 structures—a reasonably achievable goal. As no current estimates exist for the number of membrane protein families, however, it is not possible to know whether family coverage is a realistic goal for membrane proteins. Here we find that virtually all polytopic helical membrane protein families are present in the already known sequences so we can make an estimate of the total number of families. We find that only ~700 polytopic membrane protein families account for 80% of structured residues and ~1700 cover 90% of structured residues. While apparently a finite and reachable goal, we estimate that it will likely take more than three decades to obtain the structures needed for 90% residue coverage, if current trends continue.

Keywords: structural genomics; protein structure; Pfam; SCOP

An important goal of structural biology and structural bioinformatics is to assign every part of every protein sequence to a function and to a unique fold for the independently structured regions (Holm 1998; Brenner and Levitt 2000; Heger and Holm 2000; Mittl and Grutter 2001; Chothia et al. 2003). While it is not possible to solve the structures of all folded proteins, the crude structure of most protein domains could be obtained through fold recognition and homology modeling methods if a representative structure is available for all sequences (Zhang and Kim 2003). Thus, one of the goals of structural genomics is to expand our library of folds with known structure (Vitkup et al. 2001). It is therefore of interest to know how many structures will be required to have a representative fold for most proteins.

To assign a sequence to a particular fold, there must be some way to recognize the compatibility of the sequence for the known structure. If the sequences of two proteins are similar, the structures are also very likely to be similar. Thus, an upper estimate of the number of structures required to span most protein sequences is the number of protein families that can be identified based on sequence similarity alone. Estimates for the number of soluble protein families have ranged widely (Orengo et al. 1994; Wolf et al. 2000; Kunin et al. 2003), but recent results suggest that there will be many more than 250,000 families (Yan and Moult 2005). Indeed, there are currently >250,000 families with at least two members in the ProDom family database (Servant et al. 2002). Different operational definitions and strategies for defining families account for some differences in estimates (Kunin et al. 2005), but the variation is also attributable to the observation that the number of soluble protein families is constantly increasing with the number of new sequences (Kunin et al. 2003; Yan and Moult 2005). Nevertheless, because of the uneven distribution of proteins in families, Yan and Moult predict that, for the first 1000 genomes

Reprint requests to: James U. Bowie, Department of Chemistry and Biochemistry, UCLA-DOE Institute for Genomics and Proteomics, Molecular Biology Institute, 611 Charles E. Young Dr. E, Los Angeles, CA 90095-1570, USA; e-mail: bowie@mbi.ucla.edu; fax: (310) 206-4749.

Article and publication are at <http://www.proteinscience.org/cgi/doi/10.1110/ps.062109706>.

~80% of all structured soluble protein domains should be represented by only 25,000 families (Yan and Moult 2005). This is a daunting, but reasonably achievable, goal for structural genomics.

Not all sequence families adopt different folds, however, so protein fold space is much smaller than the number of families (Murzin et al. 1995; Holm and Sander 1997; Orengo et al. 1997; Hou et al. 2003). There are only a few predominant protein folds in soluble proteins and many rare folds (Govindarajan et al. 1999). Similarly membrane proteins also appear to have an uneven distribution of folds (Ubarretxena-Belandia and Engelman 2001). To the extent that a structural relationship can be recognized in the absence of easily identifiable sequence similarity, the number of structures needed to span the sequences can be reduced. Thus, the minimum number of structures needed is the number of folds that exist. Several estimates of the number of soluble protein folds have been presented that utilize different models of how folds distribute over sequence families. The estimated number of folds ranges widely from 700 to ~8000 based on differences in assumptions and data sets (Orengo et al. 1994; Zhang and DeLisi 1998; Govindarajan et al. 1999; Wolf et al. 2000; Coulson and Moult 2002; X. Liu et al. 2004). The fact that there are already 898 soluble protein folds, according to the current release of the SCOP database (v1.69) (Murzin et al. 1995), indicates that the number will greatly exceed 1000.

Membrane proteins present special technical challenges, so the pace of membrane protein structure determination is much slower. While membrane proteins account for over a quarter of all proteins, they represent roughly 0.3% of the >34,000 structures currently deposited in the Protein Data Bank (PDB). Nevertheless, there may be considerably fewer membrane protein families and folds so that fewer structures may be needed to assign most of the sequence space to structures. In a study of 26 proteomes, Liu and colleagues (Y. Liu et al. 2002, 2004) found 214 polytopic membrane protein families and 1922 soluble protein families containing at least four members, and concluded that there are about 10 times more soluble protein families than membrane protein families. Without information on how this ratio scales with an increase in family number, however, these results do not yield an estimate of the total number of families.

To parse membrane protein sequence space into families, we developed a membrane protein specific family building algorithm, which differs in several ways from other algorithms that are optimized primarily for soluble proteins (Sonnhammer and Kahn 1994; Linial et al. 1997; Park and Teichmann 1998; Gouzy et al. 1999; Enright and Ouzounis 2000; Heger and Holm 2001; Krause et al. 2005; Yan and Moult 2005). First, for sequence similarity detection, sequences are normally masked to eliminate

low complexity regions. Masking often eliminates membrane spanning segments, however, so we adjusted masking to maintain them. Second, we utilized predictions of TM helices to help define domain boundaries. Finally, we made use of the recent finding that there are fewer multi-domain membrane proteins, which allowed us to be less restrictive in combining overlapping families than would be possible with soluble proteins (Y. Liu et al. 2004). Analysis of the families built by our algorithm suggests that there are indeed many fewer membrane protein families and folds than are found in soluble proteins.

Results and Discussion

Family size distribution in membrane proteins

We applied our family building algorithm, described in the methods section, to a starting data set of 86,033 membrane protein sequences derived from 95 genomes and the Swiss-Prot database. We found 4075 total polytopic families with two or more members. There were 8273 singletons (no hits better than an e-value of 10^{-5}), and 13,644 sequences were lost when monotopic membrane protein families were removed. The polytopic families include 64,116 sequences which account for 82.5% of the 77,760 membrane protein sequences, after excluding singletons.

The distribution of sequences is strongly weighted toward a few large families. Figure 1 shows the average size distribution of the families in blocks of 20, going from the largest to the smallest families. The average size of the 20 largest families for membrane proteins was 1275 members, and the average size of the next 20 largest dropped rapidly to 336. Thus, there are only a few very large families and many very small families in membrane proteins, similar to the observation for soluble proteins (Govindarajan et al. 1999; Chothia et al. 2003).

The largest families are listed in Table 1. Among the largest families are various signaling proteins, largely of eukaryotic origin, including GPCRs and potassium channels (Chou and Elrod 2002; Harte and Ouzounis 2002; Sansom et al. 2002). A number of different transporter families are also prominent, including the major facilitator family of secondary transporters and the ABC transporter family (Saier et al. 1999; Driessen et al. 2000; Schmitt and Tampe 2002). Proteins involved in energy production (cytochrome *b* and NADH ubiquinone oxidoreductases) are also found in the largest families (Berry et al. 2000; Friedrich and Bottcher 2004). The family of protein kinases represents an error in our automated method as they are single-pass proteins with large soluble domains. Most single-pass proteins are eliminated from

Figure 1. Distribution of family sizes. The average size of blocks of 20 families is shown, ordered from the largest to the smallest, for the first 500 families. There are a few very large families and many very small families.

our families, but the protein kinases are expressed with signal sequences, which were identified as a second TM domain in our algorithm.

Validation of family building algorithm: Families versus domains

With an ideal family building algorithm, family and folding domain would be equivalent units. In other words, all families would contain a single folding domain and different families would represent different folds. In reality, this is probably impossible to achieve because there is insufficient information in the sequence alone. For example, some domains are constructed from non-contiguous regions of the primary sequence and will tend to be separated in family building. Alternatively some structurally distinguishable domains are never found alone and are contiguous in sequence, so there is no indication in the sequence database that the domains are separable. Nevertheless, we can examine how close we have come in distinguishing domains by seeing how well our family building algorithm divides known folds into different families.

Since our algorithm is optimized for membrane proteins, we would ideally test it with known membrane protein folds. In particular, our family merging protocol was specifically designed with membrane proteins in mind. Membrane proteins more rarely contain multiple domains, so we were able to use a more liberal joining of overlapping families than would be allowable with

soluble proteins. As there are so few membrane protein structures, however, we decided to perform a less ideal test, using all protein classes.

We first constructed a sequence database containing known structural domains, starting with SCOP (v1.65), which organizes 20,619 structures from the PDB into folding domains. We obtained 9643 sequences for the SCOP structural domains from the Astral database (Chandonia et al. 2004), after removing sequences with >95% sequence identity with other members. From the original 9643 sequences, we then removed sequences with structural domains constructed from noncontiguous regions of sequence, as well as all sequences <75 residues, and were left with 7784 sequences. Many of these domains are elements of larger proteins, so we next retrieved the full length sequences available in the Swiss-Prot database and were able to retrieve 4931 sequences containing one or more of

Table 1. *The largest membrane protein families*

Family	Size of family
1. GPCR #1 (Rhodopsin-like)	5520
2. Major facilitator	3680
3. ABC transporter #1 (amino acid, phosphate, ferric, nitrate, nickel, taurine, sugar)	2469
4. GPCR #2 (serpentine)	1311
5. ABC transporter #2 (multidrug resistance)	1179
6. Potassium channel	1153
7. Receptor protein kinase ^a	1150
8. Transporter #1 (cationic acid, aromatic amino acid, choline, K ⁺ uptake)	1082
9. Cytochrome <i>b</i> #1 (N-terminal part)	1030
10. Cytochrome <i>b</i> #2 (C-terminal part)	942
11. ABC transporter #3 (branched amino acid, sugar, iron, zinc, manganese)	881
12. ATPase transporter	754
13. Transporter #2 (drug/metabolite, amino acid transport)	644
14. Transporter #3 (SecF protein export, cation multidrug efflux)	587
15. NADH ubiquinone oxidoreductase	584
16. Cytochrome P450 ^b	567
17. Transporter #4 (arsenic efflux, Na ⁺ /H ⁺ antiporter, gluconate, mannitol, Na ⁺ /sulfate symporter, Mg ²⁺ /citrate transporter)	551
18. Cytochrome <i>c</i> oxidase	479
19. ABC transporter #4 (heme, O-Ag, multidrug resistance, antibiotic resistance, polysaccharide export)	471
20. Transporter #5 (DNA damage inducible protein, virulence factor, polysaccharide export protein, Na ⁺ driven multidrug efflux)	468
21. Acetylcholine receptor	460
22. Adenylate cyclase	457
23. Permease (purines, pyrimidines, sulfate)	406
24. Acetyl transferase	399
25. Glutamate receptor	398

^a A monotopic membrane protein family that was not identified by our algorithm.

^b May be a monotopic family, but not known.

the original SCOP domains. To these sequences we added soluble protein sequences randomly picked from our genomic sequences to create a soluble protein data set similar in size to our membrane protein data set.

We applied our family building algorithm to a final set of 90,823 nonredundant sequences and obtained 11,339 families with at least two members. After discarding families that had no SCOP domains, 1704 families remained. In the 1704 families, 1315 contained only one SCOP domain and 389 contained more than one SCOP domain. Thus, under a quarter of the families included multiple SCOP domains. Most of these do not represent errors in the algorithm, however, but rather limitations in the starting data. In particular, if no isolated domains are present in the sequence database, there is no way for the family building algorithm to separate them. Indeed, for all but 112 of the families with more than one fold, the domains are contiguous in the sequence. As pointed out by Liu and colleagues (Y. Liu et al. 2004), joining domains will be much less problematic for membrane proteins, which are less often composed of multiple domains. Thus, these results likely represent the worst case scenario. Even so, the vast majority (>75%) of the families correspond to a single SCOP domain.

Validation of family building algorithm: Family boundaries

We sought sequence families representing domains that span the membrane-embedded segments of membrane proteins and do not include flanking soluble domains. We define ideal TM domain boundaries as the N-terminal residue that enters and the C-terminal residue that exits the membrane hydrocarbon core. To assess how well our families achieve this ideal, we examined how well the sequence family boundaries correspond to the membrane boundaries for a number of known membrane protein structures. The membrane boundary was defined from the structure by finding the most hydrophobic 30 Å slab in the membrane plane (Chamberlain and Bowie 2004).

We compared the family regions with the SCOP domains of 36 PDB structures consisting of 46 polytopic transmembrane chains, from the 40% nonredundant helical membrane protein domain entries in SCOP v1.69 (Murzin et al. 1995; Chandonia et al. 2004), provided by the Astral Web site. Table 2 compares the structure-defined membrane boundaries with our family boundaries. The average deviation from ideal membrane boundaries at either end is 22 ± 28 , which is about 8% of the average membrane region size of 258 residues. Some examples of how the family boundaries map on the structures are shown in Figure 2. It appears that our families largely capture the membrane-embedded domains and are reasonably effective at removing large soluble regions.

We did observe one clear failure of our automated algorithm, however. As noted above, the monotopic protein kinases should have been deleted but were included because the algorithm treats the signal sequence as a TM domain, and current methods do not reliably distinguish signal peptides from TM regions (Remm and Sonnhammer 2000).

Sequence space coverage

Given the rapid fall off in family size, it makes little sense to discuss the total number of families since not all families are equally important. For example, the largest ten families will account for a much larger fraction of sequence space than the smallest ten families. Moreover, we are interested in folded protein domains, and not all regions of sequence are structured. We assume that unstructured regions are less likely to be conserved and fall into protein families. Consequently we operationally define *structured sequence space* as those sequence regions that are part of families with two or more members (Vitkup et al. 2001) and ignore the ~10% of the sequence space made up of singletons. We believe this provides a more realistic view of structural coverage than the fraction of all residues (Fischer and Eisenberg 1999; Vitkup et al. 2001; Chothia et al. 2003). After removing singletons, 4075 families remain with at least two members.

As shown in Figure 3, the largest 100 families cover more than 50% of residues in families, while the largest 500 families cover almost 80% of the total structured sequence space. In other words, for any given residue in a membrane protein family derived from our current sequence database, there is an 80% chance that it will fall into one of about 500 families.

Membrane protein family space is largely saturated by currently known sequences

As we find more membrane protein sequences, will the number of families increase dramatically or have we already seen most of the extant families? To address this question, we examined how the number of families increases with increasing database size. We generated databases of 20,000, 40,000, 60,000, and 86,033 sequences by selecting at random from the full 86,033 sequence database of membrane proteins. We then applied our family building algorithm to each of these data sets and counted the number of families needed to account for 80% or 90% of all the family residues.

Using 86,033 membrane protein sequences we find 4075 polytopic families that include 64,116 sequences as described above. These families contain 83.9% of the total residues in the 64,116 sequences, indicating that a

Table 2. Comparison of membrane boundaries defined by structures against family boundaries

PDB code	Chain	Protein name	Family number	Family region	Membrane boundary	TM helices ^a
1m0l	a	Bacteriorhodopsin	305	23–249	9–224	7
1e12	a	Halorhodopsin	305	33–269	27–254	7
1l9h	a	Rhodopsin	1	39–306	35–327	7
1jb0	a	Apoprotein a1	97	56–746	65–754	11
1rzh	l	Light subunit photosynthesis reaction center	87	26–172	32–250	5
1rzh	m	Medium subunit photosynthesis reaction center	87	47–297	54–286	5
1v54	a	Mitochondrial cytochrome <i>c</i> oxidase subunit I	18	2–467	14–475	12
1v54	c	Mitochondrial cytochrome <i>c</i> oxidase subunit III	42	6–259	15–256	7
1v54	b2	Mitochondrial cytochrome <i>c</i> oxidase subunit II	34	34–220	40–143	2
1ehk	a	Mitochondrial cytochrome <i>c</i> oxidase subunit I	18	18–495	16–493	12
1fft	a	Mitochondrial cytochrome <i>c</i> oxidase subunit I	18	39–646	56–541	12
1fft	c	Mitochondrial cytochrome <i>c</i> oxidase subunit III	42	6–202	26–202	5
1fft	b2	Mitochondrial cytochrome <i>c</i> oxidase subunit II	34	22–229	38–110	2
1m56	b2	Bacterial cytochrome <i>c</i> oxidase subunit II	34	59–278	56–120	2
1c17	m	F1F0 ATP synthase subunit A	47	72–271	99–263	4
1j4n	a	Aquaporin	56	1–200	6–232	8
1rc2	a	Aquaporin	56	3–226	1–227	8
1fx8	a	Glycerol uptake facilitator protein	56	12–259	7–256	8
1u7g	a	Ammonium transporter	71	34–394	9–381	10
1pw4	a	Glycerol 3P transporter	2	30–441	25–440	12
1pv7	a	Lactose permease	2	10–401	7–398	12
1ots	a	Clc chloride channel	49	41–445	33–438	14
1r3j	c	Potassium channel protein	6	26–114	24–114	2
1lnq	a2	Potassium channel protein	6	26–105	19–98	3
1ors	c	KVAP Potassium channel protein	6	49–244	33–138	4
1msl	a	Gated mechanosensitive channel	184	1–121	14–90	2
1oed	a	Acetylcholine receptor	21	7–450	213–433	4
1oed	c	Acetylcholine receptor	21	8–507	226–480	4
1mxm	a3	Mechanosensitive MscS	50	24–167	28–103	4
1s7b	a	Multidrug resistance efflux transporter	124	4–105	4–100	4
1rh5	a	Translocase secY	63	6–422	24–414	10
1l7v	a	ABC transporter	11	11–323	10–323	10
1pf4	a2	ABC transporter	5	16–331	31–293	6
1oy8	a7,a8	Multidrug efflux transporter AcrB	14	8–1031	9–1019	6
1ezv	c2	Mitochondrial cytochrome <i>b</i> C-terminal	10	233–377	272–370	3
1ezv	c3	Mitochondrial cytochrome <i>b</i> N-terminal	9	33–280	28–247	5
1q90	d	Subunit IV of cytochrome <i>b</i> ₆ <i>f</i> complex	10	1–160	37–147	3
1q90	b	Cytochrome <i>b</i> ₆ subunit of cytochrome <i>b</i> ₆ <i>f</i> complex	9	37–206	32–209	5
1kqf	c	Formate dehydrogenase N, gamma subunit	532	14–208	12–175	4
1q16	c	Nitrate reductase 1, gamma chain	434	11–224	2–200	5
1qla	c	Fumarate reductase cytochrome <i>b</i> subunit	1176	24–232	21–235	5
1nek	c	Succinate dehydrogenase	322	1–129	22–128	3
1nek	d	Succinate dehydrogenase	693	10–113	12–114	3
1kff6	c	Fumarate reductase	936	8–129	20–129	3
1kff6	d	Fumarate reductase	911	1–129	12–116	3
1su4	a4	Ca ATPase	12	50–989	56–974	8

^aThe number of TM helices found in the known structure.

vast majority of the sequence space of membrane proteins is represented by the families. Similarly, we found 3107 families that include 43,470 sequences from the 60,000 membrane protein sequences. There were 6976 singletons (no hits for these sequences with *e*-value better than 10^{-5}). The remaining 9554 sequences, accounting for the difference in the number of sequences in polytopic families and the sequence database, were from families that contained less than at least two TM helices and were

thus removed. The families cover 83.5% of the total residues in the sequences. The 40,000 membrane protein sequence data set led to 2401 families that include 28,456 sequences from the 40,000 membrane protein sequences. There were 5703 singletons (no hits for these sequences with *e*-value better than 10^{-5}), and the remaining 5841 sequences were from families that contained less than at least two TM helices and were removed. The families cover 83.5% of the total residues in the sequences. The

smallest data set used to build families was 20,000 sequences, and we obtained 1401 polytopic families from

13,467 sequences. There were 4131 singletons, and the remaining 2402 sequences came from monotopic families

Figure 2. The family boundaries mapped onto some known structures. The regions of the structures that are included in the families are in red and the regions not included in the families are in green. (A) Fumarate reductase (PDB Code 1l0v) is comprised of four polypeptide chains (Iverson et al. 1999, 2002): Two chains (FrdA and FrdB) correspond to soluble domains and two chains (FrdC and FrdD) are membrane-spanning with three transmembrane helices each. In the crystal structure, two fumarate reductase complexes are related by twofold symmetry, with total of 12 TM helices. However, there is no evidence that a dimer is physiologically relevant. The membrane-bound region of chain C contains residues 26–127 and chain D contains residues 16–116. Our family region covered residues 8–129 of chain C and residues 1–119 of chain D, including all TM regions with a few residues extra on both the C and N termini. (B) MsbA (PDB code 1pf4) is an ATP binding cassette transporter. The protein is a homo-dimer with two identical chains that contain 6 TM helices and a soluble ATP binding domain at the C terminus, extending from residues 316 to 582 (Chang and Roth 2001; Chang 2003). The structure-defined TM boundary is between residues 31 and 293, while the family region is between residues 16 and 331, which is longer by 15 residues at the N terminus and by 38 residues at the C terminus. (C) MscS (PDB code 1mxm) is a heptamer of identical subunits, each with three TM helices and a cytoplasmic domain, extending from residues 113 to 280 (Bass et al. 2002). The structure-defined TM helices range between 28 and 103. The family region extends from residues 24 to 167, which extends beyond our definition of the membrane by four residues at the N terminus and 64 residues at the C terminus. (D) The MthK (PDB code 1lnq) calcium-gated potassium channel is a tetramer (Jiang et al. 2002a,b). Each monomer comprises a membrane spanning pore domain and an intracellular C-terminal RCK domain. The pore domain contains two transmembrane helices. The structure-defined transmembrane region ranges from residues 19 to 98 while the family region contains residues 26–105. (E) The Kv1.2 (PDB code 2a79) shaker voltage-gated potassium channel (Long et al. 2005). The structure for each subunit of the tetramer consists of Kv1.2 containing a transmembranous ion-conducting pore, a voltage sensor, a soluble T1 domain, and an associated soluble β -subunit. The Kv1.2 channel contains six transmembrane helices (S1–S6) that are connected to soluble T1 domain by T1–S1 linker. The T1–S1 linker, S1 helix, and S3 helix were modeled as 33, 19, and 21 amino acid poly-alanine sequences, respectively. The structure-defined TM boundary is 165–402. The family boundaries are residues 150–411. Thus, we add 15 residues at the N terminus and five residues at the C terminus.

Figure 3. Coverage of structured membrane protein sequence space by the largest families. A few of the largest families account for the majority of membrane protein residues.

that were removed. These families cover 82.9% of the total residues in the sequences.

As shown in Figure 4, as the number of sequences required for 80% and 90% residue coverage increases from 13,467 to 28,456, the number of membrane protein families increases, but essentially stops growing after that. These results suggest that, ignoring the rare families, we have already observed virtually all the membrane protein families that we can expect to see. The asymptote of a hyperbolic fit suggests that only 670 families will cover 80% of structured sequence space and 1720 families will cover 90% of structured sequence space for all extant polytopic membrane proteins.

How many membrane protein folds?

To address the number of membrane protein folds, given the number of families, it is necessary to know how folds are distributed across families. Here we define fold as something equivalent to the SCOP database definition, i.e., a distinct, compact unit of protein structure that is recognizable by a skilled human. According to the SCOP hierarchy, a family is a subset of a fold, and hence the number of families will define an upper bound on fold diversity. In soluble proteins, it is well known that some folds are very common and others are rare. The distribution of the folds over families has been modeled in various ways (Orengo et al. 1994; Zhang and DeLisi 1998; Govindarajan et al. 1999; Wolf et al. 2000; Coulson and Moulton 2002; X. Liu et al. 2004). At this point, however, there are too few structures to address this issue for membrane proteins. Consequently, to obtain a first

order approximation of the number of membrane protein folds, we make the assumption that the distribution of folds over families is similar to the distribution for soluble proteins. If this is the case, we can use the stretched exponential model based on the continuous distribution proposed by Govindarajan and colleagues (Govindarajan et al. 1999), which provides a reasonable description of the number of folds as a function of family number. By this measure, we estimate that only 550 folds cover 90% and 300 folds cover 80% of membrane protein structured sequence space. Considering the physical constraints imposed by the membrane bilayer environment (Bowie 1997a,b; Y. Liu et al. 2004), we expect this to be an overestimate.

Is there evidence for soluble protein family saturation?

As discussed above, as the database size increases, we find that the number of membrane protein families needed to cover the vast majority of the database residues becomes saturated. Do soluble proteins exhibit a similar saturation with the current families? As our family building algorithm is optimized for membrane proteins, a direct comparison with soluble proteins is not valid. We, therefore, examined how the number of families required to cover structured sequence space changed in the PfamB families as the database size increased (Bateman et al.

Figure 4. Growth in membrane protein families as the number of sequences increases. The number of membrane protein families required to cover 80% and 90% of structured sequence space. The number of families grows for 80% and 90% sequence space coverage as the sequence database size increases from 13,467 to 28,456 sequences. A plateau is evident beyond 28,456 sequences, for both 80% and 90% structured sequence space coverage, indicating that the vast majority of membrane protein family space has been sampled.

2004). Like our family building algorithm, PfamB utilizes an automated family building algorithm, but it is optimized for soluble proteins. By examining PfamB from its inception (v0.2) in 1996 to the present (v18.0) (<http://www.sanger.ac.uk/Software/Pfam/ftp.shtml>), we found that the total number of PfamB families is growing at a nearly linear rate with increasing number of sequences (data not shown). As discussed previously, however, some families are much larger than others. We therefore determined how the number of families needed to cover 80% and 90% of all the residues in the families changes with increasing database size. We eliminated membrane protein families based on the criterion that at least half the members contained at least a single TM region. Like our membrane protein families, we also required that families contain at least two members. PfamB v17.0 had 129,746 families built from 336,353 sequences. The number of soluble families with at least two members is 96,905, derived from 268,144 sequences, and there are 7531 orphan soluble protein families. PfamB v7.0 had 78,233 total families built from 179,966 sequences. In PfamB v7.0, there are 2038 orphan families, and 60,638 families were soluble with at least two members, built from 143,017 sequences. PfamB v6.0 had 40,681 families built from 96,571 protein sequences. Of these, there are 1293 soluble orphan families and 31,168 soluble families with at least two members, built from 76,231 sequences.

Figure 5 shows the number of sequences in each of the three versions of PfamB versus the number of soluble protein families required to cover 80% or 90% of the family sequence space. Unlike membrane proteins, residue coverage does not appear to saturate when the sequences increase in number. This would suggest that there are many more soluble protein families left to be found. Even for 80% coverage, 50,000 families are currently needed, a result that is hard to reconcile with other much lower estimates quoted of total soluble families. By our analysis of the PfamB database, the number of important soluble protein families is at least 50,000 and, unless saturation suddenly occurs, it will likely grow even more. In their similar analysis of soluble protein families, Yan and Moulton (2005) also found little evidence for saturation in the currently known soluble protein families.

Time frame for obtaining structural representatives

How long will it take to obtain a representative structure for most membrane proteins? The answer will depend on how fast structures will be determined and whether structure determination is targeted to maximize sequence coverage.

To estimate the future pace of helical membrane protein structure determination, we followed the analysis of Stephen

Figure 5. Growth in soluble protein families as the number of sequences increases. Here we analyzed how PfamB families increase with the growth of the PfamB database. PfamB families of soluble proteins show much larger diversity than families based on membrane proteins. Unlike for membrane protein families, there is no significant decline in the slope as size of the sequence database increases. The number of families with at least two members is shown for PfamB v6.0, v7.0, and v17.0.

White (2004), who noted that the number of membrane protein structures determined has been increasing exponentially. He projected the future pace of structure determination by fitting the historical data to an exponential of the form $Y = e^{AX}$, where Y is the number of structures in a given year, X is the number of years since the first structure determination, and A is a fitted parameter. Because our families do not consider β -barrel membrane proteins, we have performed a similar analysis, counting only the helical membrane proteins, using the Membrane Proteins of Known 3D Structure Database (http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html). The number of new, unique structures determined each year since the first structure in 1985 is shown in Figure 6A. Fitting to the exponential equation yields the curve shown in Figure 6A and the parameter $A = 0.123$.

The most efficient route to covering membrane protein sequence space with the fewest structures would be to determine structures in order of family size. Figure 6B shows how structural coverage of our current sequence database would advance if we were to determine structures strictly in order of family size, assuming the pace of structure determination continues as modeled above. As a starting point, we use the structures that were known by the beginning of the year 2005. Indeed, we already have representative structures of many of the largest families,

Figure 6. The growth of membrane protein structural coverage. (A) The number of structures determined each year since 1985. The curve is a fit to an exponential equation as described in the text. (B) Ideal (solid line) and weighted-random (dotted line) estimation for coverage of sequence space in families by representatives of known structures, as a function of year, based on the observed exponential growth of structure determination seen in panel A.

including GPCRs, ABC transporters, potassium channels, cytochrome *b*, cytochrome *c* oxidase, etc., and therefore are already near 30% sequence coverage. By the ideal model, we would expect to achieve 80% coverage by 2020 and 90% sequence coverage by 2027.

The ideal model of careful structure targeting is unlikely to be practiced, however. We therefore modeled the rate of progress if proteins were selected at random for structure determination. Such a process would still be biased toward the larger families as the more members there are, the more likely a member will be selected for structure determination. As shown in Figure 6B, by a random model, we would not expect 80% residue coverage of our current sequence database until 2034, and 90% residue coverage would not occur until 2042. A more realistic view of structural coverage lies perhaps somewhere in the middle of the ideal and random estimates.

Conclusion

Our results indicate that most of the dominant membrane protein families are already present in our current sequence database. Thus, the boundaries of the membrane protein universe are visible. No doubt the overall number of families will continue to increase, but we expect the additional families to be minor contributors to the total structured sequence space.

The limited universe of membrane protein families allows us to provide an estimate for the number of structures needed to cover most of the structured membrane protein sequence space. For 90% coverage of structured sequence space we estimate a limit of ~1700 families will be required, which we estimate corresponds to an upper bound of ~550 folds. Similarly, for 80% coverage, only ~700 families will be required for which we estimate an upper bound of ~300 folds. To obtain representative structures for most sequences in the universe, the best case scenario is the total number of folds (~300–550). For this number of structures to provide useful models, however, it will be necessary to develop methods that can reliably detect whether a sequence adopts a known fold. A more realistic scenario with current technologies would require a structure for each distinct family (~700–1700).

The universe of membrane protein families appears to be much more limited than found for soluble protein families. We find little evidence for family saturation using the Pfam database, and a recent study (Yan and Moult 2005) examining 67 genomes and over 400,000 total sequences showed only slight indications of saturation at 80% domain coverage. Their extrapolated results suggest that ~25,000 families will be needed for 80% coverage of 1000 genomes and that the number will continue to increase after that. This is ~50-fold greater than the total family diversity that we find for membrane proteins. Thus, the structural genomics goal of a structure for every sequence seems to be a more definable goal for membrane proteins.

If current trends continue, we estimated that it will take roughly three decades to obtain structural representatives for 90% of the current structured membrane protein sequence space. Even if a representative structure is available for all families, however, the structures will not be sufficiently similar to all members of a family to permit accurate sequence alignment and homology modeling (Vitkup et al. 2001; Yan and Moult 2005). As a result, many more structures will likely be needed for good sequence coverage and three decades is a severe underestimate. Thus, efforts to greatly improve the pace of membrane protein structure determination are much needed. Nevertheless, the results presented here strongly suggest that there is a finite universe of membrane protein folds so that the goal of at least one structure of a distant

relative for the vast majority of membrane protein sequences is both recognizable and achievable.

Materials and methods

The sequence database

A total of 95 genome sequences were downloaded from EMBL and EBI web sites (www.ebi.ac.uk) comprised of 13 eukarya, 16 archaea, and 66 eubacteria. These were further supplemented by Swiss-Prot 41.15 entries not already in the genomes (O'Donovan et al. 2002). Protein sequences smaller than 100 residues or greater than 5000 residues were rejected.

Proteins were classified as membrane or soluble using maximum hydrophobicity as described by Boyd and colleagues (Boyd et al. 1998). In this method, the most hydrophobic segment of the protein is identified and, if the maximum hydrophobicity is greater than a particular cutoff, the protein is deemed a membrane protein. We used the recommended JTT2 hydropathy scale, a segment length of 19 residues, and a hydrophobicity of >1.528 , which results in 95% confidence for membrane protein prediction. If the maximum hydrophobicity was <1.47 , the protein was classified as soluble. Protein sequences with a maximum hydrophobic segment between 1.47 and 1.528 were considered ambiguous and not considered further. The final data set contained 86,033 membrane protein sequences and 336,198 soluble protein sequences.

Transmembrane helix prediction

Transmembrane regions within membrane proteins were determined using the Eisenberg hydropathy scale and a fixed window length of 21 amino acid residues (Eisenberg et al. 1984). The MPTOPO data set of membrane proteins with determined structures was used for validation (Jayasinghe et al. 2001). Helical transmembrane protein prediction was validated against 139 α -helical TM proteins and 28 non- α -helical TM proteins. The Eisenberg scale correctly predicted 121 (87%) α -helical TM proteins and mistakenly identified 1 (3.6%) non- α -helical TM protein.

Our transmembrane helix predictions were validated using 101 α -helical membrane protein sequences with solved structure and containing 443 transmembrane segments. We considered a prediction correct if there was at least 50% overlap between the predicted and actual transmembrane region. By this criterion, we found a false-positive error rate of 13.3% and a false-negative error of 3.2% in transmembrane helix prediction.

Sequence comparisons

Sequence comparisons were done using BLAST 2.0 (Altschul et al. 1997). We refer to a detectable sequence relationship as a hit. To prevent spurious apparent hits caused by low complexity sequences, the low complexity regions were masked using SEG (Wootton and Federhen 1996). The standard parameters typically used for SEG often mask transmembrane segments, however, which are inherently low complexity. Indeed, using the default SEG parameters of 12 2.2 2.5, 22% of residues in the transmembrane regions were masked. We therefore adjusted the SEG parameters to 10 1.0 1.5, which minimized

transmembrane segment masking, while still masking obviously low-complexity regions.

Family building algorithm

We used an iterative clustering algorithm to define families (Linial et al. 1997; Park and Teichmann 1998; Gouzy et al. 1999). The overall scheme for building families is shown in Figure 7. Family members are evolutionarily related but this relation is not always evident by sequence homology due to limits in the sensitivity of sequence comparison programs. Consequently, two related members are linked directly if there is detectable sequence similarity, but in the absence of sequence similarity they can be linked indirectly by an intermediate sequence with detectable similarity to both. This method of indirect linkage can lead to the domain problem where two unrelated proteins may be related to a third protein, which has multiple domains. We deal with this by ensuring that all family member regions overlap one another.

We wanted each family to represent a single, independent folding domain. Reasoning that smaller sequences were more likely to contain a single domain, we organized sequences according to increasing size, with smallest sequences placed first. Subsequently, we organized BLAST hits hierarchically according to statistical significance: (1) e-value $< 10^{-10}$, (2) e-value $< 10^{-9}$, (3) e-value $< 10^{-7}$, (4) e-value $< 10^{-5}$. This ensured that we initiated families with the smallest, albeit most significant alignments. The statistical significance of BLAST hits is found to be exaggerated at higher e-values, and an e-value of 10^{-5} is

Figure 7. A schematic representation of the overall process in building polytopic membrane protein families.

found to correspond to an error rate of 1 in 500 (Brenner et al. 1998). We found clustering errors at e-values worse than 10^{-5} and these hits were not considered further in our analysis. In addition, we omitted hits which did not contain at least one transmembrane helix. We considered a hit region to contain a transmembrane helix if at least half of a predicted transmembrane region was part of the hit region.

The first family was initiated with a hit from the smallest sequence within the most significant hit group. We then searched down the hit list, in the order described, for additional significant hits between present family members and new sequences. New sequences were added to the family subject to the criteria discussed below. After all the hits were examined, this process was repeated to see if hits that had not been incorporated into families in the first round could now be added. Then the second family was initiated with the next smallest, most significant hit remaining and so forth.

Family members were added by the following procedure. There were two levels of clustering involved. The first incorporated only those hits with an e-value better than 10^{-9} . In this hit group the sequence was incorporated directly if it passed overlap and length tests. To prevent divergence in aligned regions and domain sizes that can occur between the first family member and the subsequently added members, we required that all members had similarity to each other with regard to size and region of the protein segment. This was implemented using two additional checks before adding a new sequence. (1) We required at least a 70% overlap in sequence of the aligned region with all other previously clustered members of the family. Percent overlap is defined as the length of the overlap of the region of the hit under consideration with the region already in the family, divided by the average size of the two regions. (2) We required that the ratio of the hit length under consideration and the family domain size was between one-half and two, where the domain size of the family was defined as the size of the first member of the family. The families in this high significance hit group were clustered until convergence or for a maximum of eight iterations.

The second level of clustering considered hits in the lower significance group, with e-value between 10^{-9} and 10^{-5} . Hits were added to each family built from the high significance hit group by clustering for a maximum of two iterations. In this group there is an increased danger of adding a false positive. Thus, the sequence was considered for inclusion subject to the following test: The sequence must show a hit with at least two different members that have already been incorporated into the family. Sequences were defined as different if they showed <80% sequence identity.

Post-processing of families

We organized the families according to their size, and removed those families that were monotopic. A family was judged to be monotopic if at least 25% of its members consisted of only a single transmembrane helix. Next, we processed the families to merge common domains that had been clustered separately. One family can be separated into two due to our stringent criterion for inclusion of a member into a family during the first phase of family building. Another situation where two families should be merged is when fragmentation or domain cutting by BLAST results in one domain broken over more than one family. One cause for this is incomplete sequence fragments in the sequence database, while another may be inherent weaknesses in BLAST. Such fragmentation of families was observed fre-

quently in the case of GPCR families. To correct for this family splitting, we merged a smaller family with a larger family if at least half of its members are common with the larger family, and if these common members overlap in sequence region by at least 50%.

Subsequent to merging families, sequence end-points for a family member were determined from all the merged hits for that member, based on the boundaries of the membrane domain. We compared hit regions for a given member in the merged family with the corresponding transmembrane helix regions predicted for that member. The initial and final hit regions that most closely incorporated the entire membrane domain were chosen as the family region. A region that was too large was always given preference to a region that was smaller than the actual membrane boundary.

Acknowledgments

We thank Salem Faham, Sehat Nauli, and Marisa Baron for helpful discussions and comments on the manuscript, and Blai Bonet for valuable suggestions on developing the algorithm. This work was supported by National Institutes of Health Grant GM3919 and a National Science Foundation Integrative Graduate Education and Research Traineeship predoctoral award (to A.O.). J.U.B. is a Leukemia and Lymphoma Society Scholar.

References

- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**: 3389–3402.
- Bass, R.B., Strop, P., Barclay, M., and Rees, D.C. 2002. Crystal structure of *Escherichia coli* MscS, a voltage-modulated and mechanosensitive channel. *Science* **298**: 1582–1587.
- Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., et al. 2004. The Pfam protein families database. *Nucleic Acids Res.* **32**: D138–D141.
- Berry, E.A., Guergova-Kuras, M., Huang, L.S., and Crofts, A.R. 2000. Structure and function of cytochrome *bc* complexes. *Annu. Rev. Biochem.* **69**: 1005–1075.
- Bowie, J.U. 1997a. Helix packing angle preferences. *Nat. Struct. Biol.* **4**: 915–917.
- . 1997b. Helix packing in membrane proteins. *J. Mol. Biol.* **272**: 780–789.
- Boyd, D., Schierle, C., and Beckwith, J. 1998. How many membrane proteins are there? *Protein Sci.* **7**: 201–205.
- Brenner, S.E. and Levitt, M. 2000. Expectations from structural genomics. *Protein Sci.* **9**: 197–200.
- Brenner, S.E., Chothia, C., and Hubbard, T.J. 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci.* **95**: 6073–6078.
- Chamberlain, A.K. and Bowie, J.U. 2004. Analysis of side-chain rotamers in transmembrane proteins. *Biophys. J.* **87**: 3460–3469.
- Chandonia, J.M., Hon, G., Walker, N.S., Lo Conte, L., Koehl, P., Levitt, M., and Brenner, S.E. 2004. The ASTRAL compendium in 2004. *Nucleic Acids Res.* **32**: D189–D192.
- Chang, G. 2003. Structure of MscA from *Vibrio cholera*: A multidrug resistance ABC transporter homolog in a closed conformation. *J. Mol. Biol.* **330**: 419–430.
- Chang, G. and Roth, C.B. 2001. Structure of MscA from *E. coli*: A homolog of the multidrug resistance ATP binding cassette (ABC) transporters. *Science* **293**: 1793–1800.
- Chothia, C., Gough, J., Vogel, C., and Teichmann, S.A. 2003. Evolution of the protein repertoire. *Science* **300**: 1701–1703.
- Chou, K.C. and Elrod, D.W. 2002. Bioinformatical analysis of G-protein-coupled receptors. *J. Proteome Res.* **1**: 429–433.
- Coulson, A.F. and Moul, J. 2002. A unfold, mesofold, and superfold model of protein fold use. *Proteins* **46**: 61–71.

- Driessen, A.J., Rosen, B.P., and Konings, W.N. 2000. Diversity of transport mechanisms: Common structural principles. *Trends Biochem. Sci.* **25**: 397–401.
- Eisenberg, D., Schwarz, E., Komaromy, M., and Wall, R. 1984. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.* **179**: 125–142.
- Enright, A.J. and Ouzounis, C.A. 2000. GeneRAGE: A robust algorithm for sequence clustering and domain detection. *Bioinformatics* **16**: 451–457.
- Fischer, D. and Eisenberg, D. 1999. Finding families for genomic ORFans. *Bioinformatics* **15**: 759–762.
- Friedrich, T. and Bottcher, B. 2004. The gross structure of the respiratory complex I: A Lego system. *Biochim. Biophys. Acta* **1608**: 1–9.
- Gouzy, J., Corpet, F., and Kahn, D. 1999. Whole genome protein domain analysis using a new method for domain clustering. *Comput. Chem.* **23**: 333–340.
- Govindarajan, S., Recabarren, R., and Goldstein, R.A. 1999. Estimating the total number of protein folds. *Proteins* **35**: 408–414.
- Harte, R. and Ouzounis, C.A. 2002. Genome-wide detection and family clustering of ion channels. *FEBS Lett.* **514**: 129–134.
- Heger, A. and Holm, L. 2000. Towards a covering set of protein family profiles. *Prog. Biophys. Mol. Biol.* **73**: 321–337.
- . 2001. Picasso: Generating a covering set of protein family profiles. *Bioinformatics* **17**: 272–279.
- Holm, L. 1998. Unification of protein families. *Curr. Opin. Struct. Biol.* **8**: 372–379.
- Holm, L. and Sander, C. 1997. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.* **25**: 231–234.
- Hou, J., Sims, G.E., Zhang, C., and Kim, S.H. 2003. A global representation of the protein fold space. *Proc. Natl. Acad. Sci.* **100**: 2386–2390.
- Iverson, T.M., Luna-Chavez, C., Cecchini, G., and Rees, D.C. 1999. Structure of the *Escherichia coli* fumarate reductase respiratory complex. *Science* **284**: 1961–1966.
- Iverson, T.M., Luna-Chavez, C., Croal, L.R., Cecchini, G., and Rees, D.C. 2002. Crystallographic studies of the *Escherichia coli* quinol-fumarate reductase with inhibitors bound to the quinol-binding site. *J. Biol. Chem.* **277**: 16124–16130.
- Jayasinghe, S., Hristova, K., and White, S.H. 2001. MPTopo: A database of membrane protein topology. *Protein Sci.* **10**: 455–458.
- Jiang, Y., Lee, A., Chen, J., Cadene, M., Chait, B.T., and MacKinnon, R. 2002a. Crystal structure and mechanism of a calcium-gated potassium channel. *Nature* **417**: 515–522.
- . 2002b. The open pore conformation of potassium channels. *Nature* **417**: 523–526.
- Krause, A., Stoye, J., and Vingron, M. 2005. Large scale hierarchical clustering of protein sequences. *BMC Bioinformatics* **6**: 15.
- Kunin, V., Cases, I., Enright, A.J., de Lorenzo, V., and Ouzounis, C.A. 2003. Myriads of protein families, and still counting. *Genome Biol.* **4**: 401.
- Kunin, V., Teichmann, S.A., Huynen, M.A., and Ouzounis, C.A. 2005. The properties of protein family space depend on experimental design. *Bioinformatics* **21**: 2618–2622.
- Linial, M., Linial, N., Tishby, N., and Yona, G. 1997. Global self-organization of all known protein sequences reveals inherent biological signatures. *J. Mol. Biol.* **268**: 539–556.
- Liu, X., Fan, K., and Wang, W. 2004. The number of protein folds and their distribution over families in nature. *Proteins* **54**: 491–499.
- Liu, Y., Engelman, D.M., and Gerstein, M. 2002. Genomic analysis of membrane protein families: Abundance and conserved motifs. *Genome Biol.* **3**: research0054.1–research0054.12.
- Liu, Y., Gerstein, M., and Engelman, D.M. 2004. Transmembrane protein domains rarely use covalent domain recombination as an evolutionary mechanism. *Proc. Natl. Acad. Sci.* **101**: 3495–3497.
- Long, S.B., Campbell, E.B., and Mackinnon, R. 2005. Crystal structure of a mammalian voltage-dependent Shaker family K⁺ channel. *Science* **309**: 897–903.
- Mittl, P.R. and Grutter, M.G. 2001. Structural genomics: Opportunities and challenges. *Curr. Opin. Chem. Biol.* **5**: 402–408.
- Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**: 536–540.
- O'Donovan, C., Martin, M.J., Gattiker, A., Gasteiger, E., Bairoch, A., and Apweiler, R. 2002. High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief. Bioinform.* **3**: 275–284.
- Orengo, C.A., Jones, D.T., and Thornton, J.M. 1994. Protein superfamilies and domain superfolds. *Nature* **372**: 631–634.
- Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. 1997. CATH—A hierarchical classification of protein domain structures. *Structure* **5**: 1093–1108.
- Park, J. and Teichmann, S.A. 1998. DIVCLUS: An automatic method in the GEANFAMMER package that finds homologous domains in single- and multi-domain proteins. *Bioinformatics* **14**: 144–150.
- Remm, M. and Sonnhammer, E. 2000. Classification of transmembrane protein families in the *Caenorhabditis elegans* genome and identification of human orthologs. *Genome Res.* **10**: 1679–1689.
- Saier Jr., M.H., Beatty, J.T., Goffeau, A., Harley, K.T., Heijne, W.H., Huang, S.C., Jack, D.L., Jahn, P.S., Lew, K., Liu, J., et al. 1999. The major facilitator superfamily. *J. Mol. Microbiol. Biotechnol.* **1**: 257–279.
- Sansom, M.S., Shrivastava, I.H., Bright, J.N., Tate, J., Capener, C.E., and Biggin, P.C. 2002. Potassium channels: Structures, models, simulations. *Biochim. Biophys. Acta* **1565**: 294–307.
- Schmitt, L. and Tampe, R. 2002. Structure and mechanism of ABC transporters. *Curr. Opin. Struct. Biol.* **12**: 754–760.
- Servant, F., Bru, C., Carrere, S., Courcelle, E., Gouzy, J., Peyruc, D., and Kahn, D. 2002. ProDom: Automated clustering of homologous domains. *Brief. Bioinform.* **3**: 246–251.
- Sonnhammer, E.L. and Kahn, D. 1994. Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci.* **3**: 482–492.
- Ubarretxena-Belandia, I. and Engelman, D.M. 2001. Helical membrane proteins: Diversity of functions in the context of simple architecture. *Curr. Opin. Struct. Biol.* **11**: 370–376.
- Vitkup, D., Melamud, E., Moul, J., and Sander, C. 2001. Completeness in structural genomics. *Nat. Struct. Biol.* **8**: 559–566.
- White, S.H. 2004. The progress of membrane protein structure determination. *Protein Sci.* **13**: 1948–1949.
- Wolf, Y.I., Grishin, N.V., and Koonin, E.V. 2000. Estimating the number of protein folds and families from complete genome data. *J. Mol. Biol.* **299**: 897–905.
- Wootton, J.C. and Federhen, S. 1996. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266**: 554–571.
- Yan, Y. and Moul, J. 2005. Protein family clustering for structural genomics. *J. Mol. Biol.* **353**: 744–759.
- Zhang, C. and DeLisi, C. 1998. Estimating the number of protein folds. *J. Mol. Biol.* **284**: 1301–1305.
- Zhang, C. and Kim, S.H. 2003. Overview of structural genomics: From structure to function. *Curr. Opin. Chem. Biol.* **7**: 28–32.