doi:10.1016/j.jmb.2005.08.058

J. Mol. Biol. (2005) **353**, 744–759

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Protein Family Clustering for Structural Genomics

Yongpan Yan[1,2] and John Moult[1]*

[1]*Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, 9600 Gudelsky Drive Rockville, MD 20850, USA*

[2]*Molecular and Cell Biology Program, University of Maryland, College Park, MD 20742, USA*

**Corresponding author*

A major goal of structural genomics is the provision of a structural template for a large fraction of protein domains. The magnitude of this task depends on the number and nature of protein sequence families. With a large number of bacterial genomes now fully sequenced, it is possible to obtain improved estimates of the number and diversity of families in that kingdom. We have used an automated clustering procedure to group all sequences in a set of genomes into protein families. Benchmarking shows the clustering method is sensitive at detecting remote family members, and has a low level of false positives. This comprehensive protein family set has been used to address the following questions. (1) What is the structure coverage for currently known families? (2) How will the number of known apparent families grow as more genomes are sequenced? (3) What is a practical strategy for maximizing structure coverage in future? Our study indicates that approximately 20% of known families with three or more members currently have a representative structure. The study indicates also that the number of apparent protein families will be considerably larger than previously thought: We estimate that, by the criteria of this work, there will be about 250,000 protein families when 1000 microbial genomes have been sequenced. However, the vast majority of these families will be small, and it will be possible to obtain structural templates for 70–80% of protein domains with an achievable number of representative structures, by systematically sampling the larger families.

Keywords: protein sequence clustering; SCOP benchmark; structural genomics; structure coverage; protein universe

Introduction

The ultimate goal of structural genomics is to provide structures for all biological proteins. Although there have been enormous improvements in experimental methods for determining structure,[1] these still lag behind sequencing methods by orders of magnitude, in both cost and speed. As a result, currently, only about 1% of proteins with known sequence also have an experimentally known structure. Fortunately, it is not essential to experimentally determine the structure of every protein; evolutionarily related proteins have similar structures,[2,3] and so comparative modeling methods can be used to obtain a structure for any protein with a detectable evolutionary relationship to one with an experimental structure. This strategy has been widely accepted.[3–8] The accuracy of comparative models depends on the closeness of the evolutionary relationships they are based on,[9] and is never as high as that of a high-quality X-ray structure. Nevertheless, these models are useful for many practical applications.[10]

The minimum number of experimental structures that will be needed in order to model all proteins using evolutionary relationships depends on the nature of protein sequence space. In particular, this number depends on how many families of evolutionarily associable proteins there are. The recent increase in fully sequenced genomes has made it possible to estimate this quantity more reliably than in the past. Here, we make use of knowledge of the full sequences for a set of 67 bacteria to obtain such an estimate.

No sequence-based method is able to detect all evolutionary relationships: experimental structure determinations reveal previously undetectable relationships in many cases. Thus, all

Abbreviations used: PSSM, position specific scoring matrix.

E-mail address of the corresponding author: jmoult@tunc.org sequence-based families are, in some sense, arbitrary, reflecting the effectiveness of current relationship detection algorithms rather than the number of independent evolutionary lines. From a structural genomics perspective, current methods are sufficiently powerful that they already represent very coarse-grained sampling of structure space, so that models based on one experimental structure per family are probably at the limit of useful accuracy.[11] A single family will also often embrace a number of functions.[12]

Clustering of proteins into families has long been used as a basis for extending function annotation, and so there is a history of algorithm development.[13–24] Many of the family sets have been developed for specific purposes, and there is so far no universally accepted comprehensive source. For example, Pfam A,[24] one of the best established sets, uses sensitive methods to detect remote evolutionary relationships, and is curated by hand, providing a high level of reliability. As a consequence, coverage is incomplete.

We have developed an automated family classification scheme, applicable to estimation of the number of experimental structures that will be needed for structural genomics. There are three main steps: identification of evolutionary relationships; parsing of the full proteins into probable structural domains; and clustering into families. Conventional PSI-BLAST searches are used to detect sequence relationships within a set of 67 fully sequenced bacterial genomes. Lists of relationships are sub-divided on the basis of a protein domain identification method. Lists are then merged into families with a multi-linkage clustering procedure. Although a relatively standard sequence search method is used, benchmarking with SCOP structural superfamilies[25,26] shows a slightly higher sensitivity than previously reported methods including profile and profile-profile methods†.[27–29] We attribute this to the robust clustering step and reasonably effective parsing into domains.

There have been a number of studies of the number of protein families in biology. Estimates vary from 1000 to 30,000.[4,7,8,30–36] As more genome sequences are completed, it becomes possible to improve the reliability of the estimate. Our study, focusing on recently available complete genome sequences, leads to an estimate for the prokaryotes that is substantially higher than previous ones: Clustering 178,310 sequences from 67 microbial genomes already generates 31,874 families. A recent study of five fully sequenced eukaryotic genomes has also led to a much larger number than previously suggested, 45,000 protein families.[37] A more relevant quantity for structural genomics is the number of detectable families there will be in future. We have developed a method of estimating growth in the number of families, and find there will be about 250,000 families when 1000 genomes are sequenced. Apparent singletons (proteins with no detectable relatives)[38] are the fastest growing category.

At first glance, these increased estimates are discouraging for the structural genomics goal of obtaining structures for all domains. However, because most sequences are in relatively large families, we estimate that it will still be possible to have coverage for 70% ~ 80% of domains within the next decade.

Procedures

Protein sequences

All identified protein sequences in 140 genomes were retrieved from Genbank‡. Of these, 67 were used for building the family estimate model, and the rest were reserved for testing the projections of the model. All downloaded and generated information were stored in a MySQL relational database running on a Linux server (Tables 1 and 2).

Generation of homolog lists

For each protein, a six-round PSI-BLAST search was performed against the set of all other sequences in the genome set.[39,40] Low-complexity regions were omitted, using the default SEG option,[41] covering 5.9% of the residues. Homologs with an E-value 10^{-4} or lower to the search sequence were collected, creating a homolog list for each protein.

Domain parsing

Each homolog list was examined for domain structures, as described below. A number of domain parsing methods have been developed.[16,21,42] In the present work, domain boundaries are identified on the basis of the location of indels in the PSI-BLAST sequence alignment. Indel locations are found by counting the number of sequences with an amino acid at each position in the alignment. Figure 1 shows an example.

Domain boundaries are defined as positions in the multiple alignment where there are relatively deep minima in the number of sequences with residues. The detailed procedure is as follows.

(1) Calculate the slope of the alignment count for each position in the alignment.
(2) Find all the turning points (positions where the sign of the slope changes).
(3) Discard the trough points that make a domain too short to be viable (less than 40 residues between turning points).
(4) Discard the trough points where a trough is not significantly lower than the surroundings

† http://supfam.mrc-lmb.cam.ac.uk/PRC/

‡ http://www.ncbi.nlm.nih.gov/Genomes/index.html

Table 1. The 67 fully sequenced microbial genomes used for protein family construction, and the number of proteins in each genome

Genome	Number of proteins
Aeropyrum pernix	2694
Agrobacterium tumefaciens str. C58 (Dupont)	5402
Aquifex aeolicus	1553
Archaeoglobus fulgidus	2407
Bacillus halodurans	4066
Bacillus subtilis	4100
Borrelia burgdorferi	1637
Brucella melitensis	3198
Buchnera sp. APS	574
Campylobacter jejuni	1629
Caulobacter crescentus	3737
Chlamydia muridarum	916
Chlamydophila pneumoniae AR39	1110
Chlamydophila pneumoniae CWL029	1052
Chlamydophila pneumoniae J138	1069
Clostridium acetobutylicum	3672
Clostridium perfringens	2723
Corynebacterium glutamicum	3040
Deinococcus radiodurans	3102
Escherichia coli	4289
Escherichia coli O157:H7	5361
Haemophilus influenzae Rd	1709
Halobacterium sp. NRC-1	2605
Helicobacter pylori 26695	1566
Helicobacter pylori J99	1490
Lactococcus lactis subsp. *lactis*	2266
Listeria innocua	3043
Listeria monocytogenes EGD-e	2846
Mesorhizobium loti	7275
Methanobacterium thermoautotrophicum	1869
Methanococcus jannaschii	1770
Mycobacterium leprae	1605
Mycobacterium tuberculosis	3869
Mycobacterium tuberculosis CDC1551	4187
Mycoplasma genitalium	480
Mycoplasma pneumoniae	688
Mycoplasma pulmonis	782
Neisseria meningitidis	2025
Neisseria meningitidis Z2491	2032
Nostoc sp. PCC 7120	6129
Pasteurella multocida	2014
Pseudomonas aeruginosa	5565
Pyrobaculum aerophilum	2605
Pyrococcus abyssi	1765
Pyrococcus horikoshii	2064
Ralstonia solanacearum	5116
Rhizobium sp. NGR234	416
Rickettsia conorii	1374
Rickettsia prowazekii	834
Salmonella enterica subsp. *enterica* serovar *Typhi*	4749
Salmonella typhimurium LT2	4553
Sinorhizobium meliloti	6205
Staphylococcus aureus subsp. *aureus* Mu50	2748
Staphylococcus aureus subsp. *aureus* N315	2624
Streptococcus pneumoniae	2094
Streptococcus pyogenes	1696
Sulfolobus solfataricus	2977
Sulfolobus tokodaii	2826
Synechocystis PCC6803	3169
Thermoplasma acidophilum	1478
Thermoplasma volcanium	1526
Thermotoga maritima	1846
Treponema pallidum	1031
Ureaplasma urealyticum	611
Vibrio cholerae	3828
Xylella fastidiosa	2831
Yersinia pestis	4039

There are 12 archaea, and 55 bacterial genomes. In total, there are 178,310 protein sequences.

Table 2. The 73 recently sequenced microbial genomes used to test the family growth projections

Genome	Number of proteins
Agrobacterium tumefaciens str. C58 (Cereon)	5299
Bacillus anthracis str. Ames	5311
Bacillus cereus ATCC 14579	5255
Bacteroides thetaiotaomicron VPI-5482	4778
Bifidobacterium longum NCC2705	1729
Bordetella bronchiseptica	4994
Bordetella parapertussis	4185
Bordetella pertussis	3446
Bradyrhizobium japonicum USDA 110	8317
Brucella suis 1330	3264
Buchnera aphidicola str. Bp (Baizongiapistaciae)	504
Buchnera aphidicola str. Sg (Schizaphisgraminum)	546
Candidatus blochmannia floridanus	583
Chlamydia trachomatis	893
Chlamydophila caviae GPIC	1005
Chlamydophila pneumoniae TW-183	1113
Chlorobium tepidum TLS	2252
Chromobacterium violaceum ATCC 12472	4407
Clostridium tetani E88	2373
Corynebacterium efficiens YS-314	2950
Coxiella burnetii RSA 493	2009
Enterococcus faecalis V583	3113
Escherichia coli CFT073	5379
Escherichia coli O157:H7 EDL933	5349
Fusobacterium nucleatum subsp. *nucleatum* ATCC25586	2067
Haemophilus ducreyi 35000HP	1717
Helicobacter hepaticus ATCC 51449	1875
Lactobacillus plantarum WCFS1	3009
Leptospira interrogans serovar lai str. 56601	4727
Methanopyrus kandleri AV19	1687
Methanosarcina acetivorans C2A	4540
Methanosarcina mazei Goe1	3371
Mycobacterium bovis subsp. *bovis* AF2122/97	3920
Mycoplasma gallisepticum R	726
Mycoplasma penetrans	1037
Nitrosomonas europaea ATCC 19718	2461
Oceanobacillus iheyensis HTE831	3500
Pirellula sp.	7325
Porphyromonas gingivalis W83	1909
Prochlorococcus marinus str. MIT 9313	2265
Prochlorococcus marinus subsp. *marinus* str.CCMP137	1882
Prochlorococcus marinus subsp. *pastoris* str.CCMP13	1712
Pseudomonas putida KT2440	5350
Pseudomonas syringae pv. tomato str. DC3000	5471
Pyrococcus furiosus DSM 3638	2065
Salmonella enterica subsp. *enterica* serovar Typhi Ty2	4323
Shewanella oneidensis MR-1	4472
Shigella flexneri 2a str. 2457T	4068
Shigella flexneri 2a str. 301	4180
Staphylococcus aureus subsp. *aureus* MW2	2632
Staphylococcus epidermidis ATCC 12228	2419
Streptococcus agalactiae 2603V/R	2124
Streptococcus agalactiae NEM316	2094
Streptococcus mutans UA159	1960
Streptococcus pneumoniae R6	2043
Streptococcus pyogenes MGAS315	1865
Streptococcus pyogenes MGAS8232	1845
Streptococcus pyogenes SSI-1	1861
Streptomyces avermitilis MA-4680	7575
Streptomyces coelicolor A3(2)	8154
Synechococcus sp. WH 8102	2517
Thermoanaerobacter tengcongensis	2588
Thermosynechococcus elongatus BP-1	2475
Tropheryma whipplei str. Twist	808
Tropheryma whipplei TW08/27	783
Vibrio parahaemolyticus RIMD 2210633	4832
Vibrio vulnificus CMCP6	4537
Wigglesworthia glossinidia endosymbiont of Glossina	611
Wolinella succinogenes	503
Xanthomonas axonopodis pv. citri str. 306	4312
Xanthomonas campestris pv. campestris str. ATCC339	4181
Xylella fastidiosa Temecula1	2036
Yersinia pestis KIM	4090

In all, the 140 genomes code for 405,709 proteins.

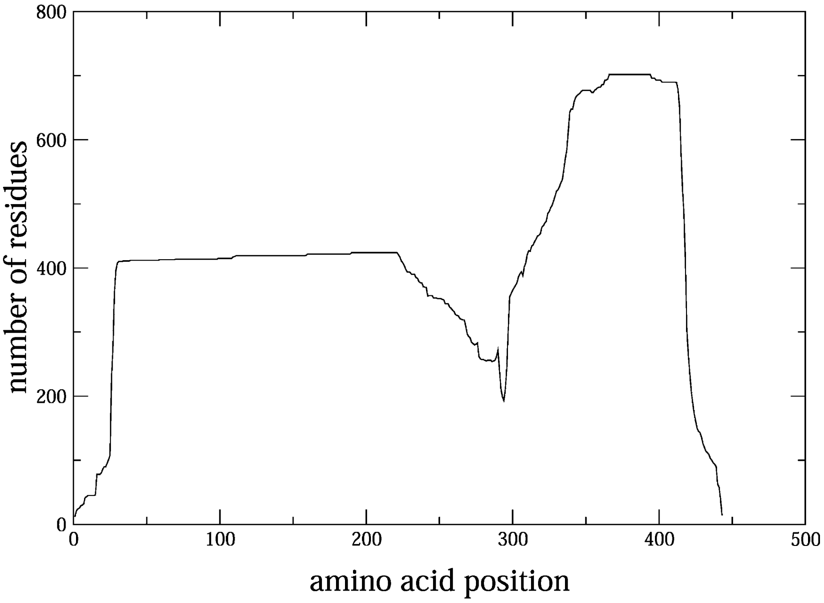

Figure 1. An example of domain parsing, for the multiple sequence alignment of *E. coli* ARGA (Swiss-Prot ID P08205, amino acid acetyltransferase). The domain-splitting algorithm produces two domains, residues 20–294 and 295–443. Pfam and InterPro also split this protein into two domains. Domain 1 (26–269) belongs to Pfam PF00696, an amino acid kinase family, and domain 2 (338–414) belongs to PF00583, an acetyltransferase family.

(trough height more than 60% of the peaks on either side).

(5) Divide the proteins in the homolog list into domains by cutting at each remaining trough point, to create homolog domain lists.

As described later, comparison of the results of this procedure with a set of PfamA domains in 50,000 randomly chosen Pfam sequences shows it is very conservative: 96% of single domain PfamA proteins are predicted as such, but only 24% of PfamA two domains proteins are predicted correctly. Other domain parsers adopt a different balance of false positives and false negatives.[43] While this and other parsers are far from satisfactory, domain parsing does improve the quality of the families.

Merging of domain lists

Domain lists are highly redundant, in that many domains appear in multiple lists. A key step is merging of the lists to form non-redundant, domain-based protein families. Merging also increases the range of evolutionary relationships that are clustered: A PSI-BLAST search starting from protein A may find a relative B, but not relative C. On the other hand, PSI-BLAST started from protein B may have found relative C, but not relative A. Merging of the A and B hit lists places A, B and C in one family.

The simplest clustering procedure is to iteratively merge all pairs of lists that contain at least one common domain, and then eliminate redundancies from the merged sets. Notoriously, this single linkage procedure leads to over-clustering, even when the false positive rate for inclusion of a domain in a single list is small. A number of strategies have been suggested for overcoming this problem.[22,44] We have developed a variable linkage procedure. Short domain lists are merged on the basis of a single common entry. The longer the lists, the more common entries are required.

Merging proceeds by selecting a first list, comparing it to all others, combining where the merge rules are satisfied, then picking a next so far unconsidered list, and so on, until all lists have been considered. The process is repeated a maximum of three times.

Further domain boundary checking

Incomplete domain parsing can occasionally lead to the merging of proteins that have no significant alignment. This is illustrated in Figure 2. An unparsed, two domain protein (b) in list I has a region of alignment with protein (c) in a second list, II. Sequence (c) shares no significant relationship with the primary domain in list I, but will be merged into that list. To reduce this effect, each candidate sequence in list II is checked for alignment overlap with the first sequence in list I.

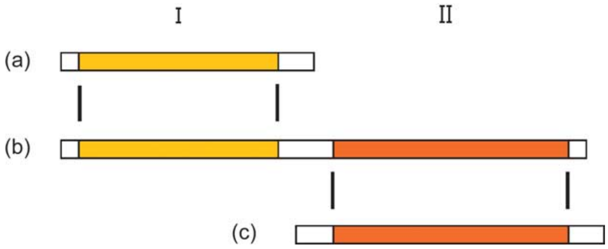

Figure 2. Domain merging check.

List II entries with an overlap of less than 40 residues are not merged.

Selection of parameter values

Results are dependent on a number of parameters. The parameters were optimized by building a set of families for proteins with domains in PfamA, varying parameter values to maximize the agreement between the generated and Pfam families, as others have done.[45] A set of 50,000 full-length SwissProt protein sequences including all the domains present in a 721 family subset of PfamA version 9.0 was used. Use of full-length protein sequences allows the domain parsing procedures to be tested.

Each generated family was compared with all the PfamA families, and the most similar one (most common sequences) was considered the best match. Two measures are used to assess the quality of the built families:

$$\text{False negative fraction } F_{\mathrm{N}} = (P - O)/P$$

$$\text{False positive fraction } F_{\mathrm{P}} = (M - O)/M$$

where P is the Pfam family size, M is the generated family size, and O is the common sequences between the two. F_{N} is the false negative rate for a generated family; the fraction of correct domains omitted. F_{P} is the false positive rate for a family; the fraction of incorrect domains included.

These ratios were determined for a range of PSI-BLAST conditions, with and without domain checks, and with different linkage rules, in order to optimize the procedure. Details are given in Results.

The final choice of parameters was up to six rounds of PSI-BLAST with an E-score threshold of 10^{-4}, and a maximum of three rounds of merging. Lists are merged into a family using the following merging rules:

(1) For lists with four or fewer members: at least one common entry required for merging.
(2) For lists with five to ten members: at least two common entries.
(3) For lists with more than ten members: at least 40% common entries.

Evaluation of domain family construction

Effectiveness of the family building procedure was assessed in terms of its ability to pair all members of SCOP superfamilies, and not to pair domains in different folds. (Note that effectiveness could not be evaluated against Pfam, since the method was tuned on that basis. SCOP† is a hierarchical organization of proteins based on evolutionary and structural relationships.[25,26] Since structural similarity provides a much more sensitive test of evolutionary relationships between proteins than does sequence, SCOP has been widely used as a benchmark for evaluating sequence alignment, clustering, and evolutionary relationship detection methods.[44,46] We have used SCOP40 (no sequence relationships higher than 40% identity) version 1.63, which contains 5226 domains, 1224 superfamilies, and 760 folds.

The 5226 domains were clustered into families, as described above: PSI-BLAST was run for each domain against the NR sequence database, augmented with the SCOP domain sequences. No domain parsing was performed, as SCOP is already domain-based. PSI-BLAST-generated homolog lists were merged using the linkage rules, to form a set of generated families.

The set of generated families was compared with the SCOP superfamilies in terms of all the possible pairwise relationships between domains. Any pair of domains found both in a generated family and a SCOP superfamily is considered a true positive. A pair of domains presented in a generated family set, but not assigned the same SCOP fold is considered a false positive, as it is unlikely to represent a homology relationship. SCOP40 version 1.63 was used, with 50,374 pairs of domains within the same superfamily, and more than 600,000 pairs of domains with each member in a different fold. True positives detected as a function of the false positives incurred were plotted in a ROC curve. A 1% false positive to true positive ratio was chosen as an overall measure of quality, as used by others.[47–50]

For comparison, several other alignment and family clustering methods were also tested using the same set of SCOP domains. These are BLAST, PSIBLAST, SAM-T99 (HMM)‡,[51] and PRC (a profile to profile method)§. Software was downloaded from the authors' web sites.

Programs and parameters used for SAM-99 were:

(1) target99 − seed [sequence fasta file] − out [output file] − db [nr + scop40] − iter 4
(2) fw0.7 [sequence.a2m file] [sequence.mod file]
(3) hmmscore [sequence name] − i [sequence.mod file] − sw 2 − db [scop40]

Programs and parameters used for PRC were:

Prc − Emax 10 [sequence.mod file] [mod library] [sequence name]

Transmembrane protein determination

Proteins with one or more transmembrane helical segments were identified using TMHMM‖.[52]

† http://scop.mrc-lmb.cam.ac.uk/scop/

‡ http://www.soe.ucsc.edu/research/compbio/sam2src/

§ http://supfam.mrc-lmb.cam.ac.uk/PRC/

‖ http://www.cbs.dtu.dk/services/TMHMM/

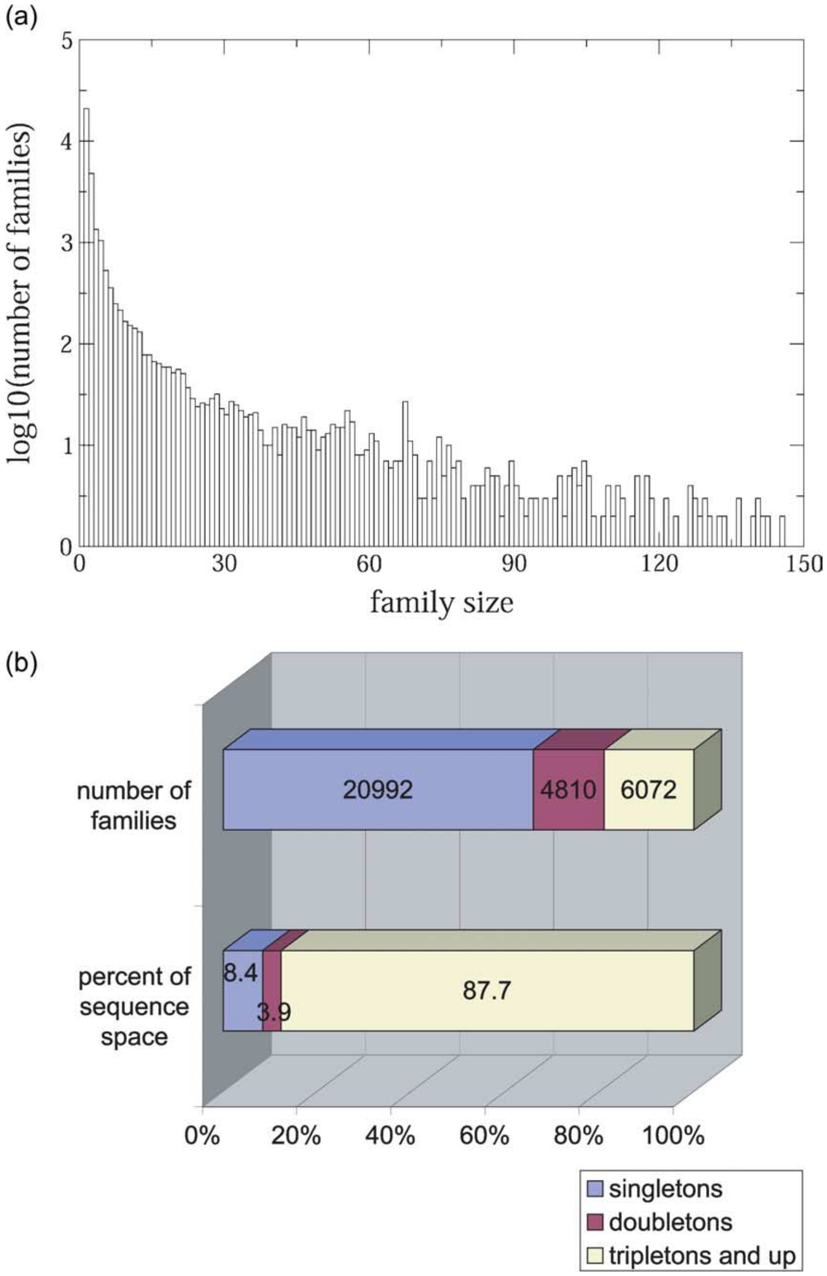

Figure 3. (a) Distribution of domain family size. Note the log scale. There is an approximately power law relationship between the number of families and family size: 20,992 of the 31,874 families have only a single member, while only 263 families are larger than 100. (b) Number of singletons (family size one), doubletons (family size two), tripletons and larger (top bar), and the percentage of sequence space covered by each of the three categories. Although there are 20,992 singletons and 4810 doubletons, these two categories represent only about 12% of sequence space. The 6072 larger families comprise the rest.

Structure coverage determination

Domain families with known structures were identified as follows. A sequence profile (position specific scoring matrix, PSSM) was obtained from the multiple sequence alignment of a protein family, using blastpgp.[39] Each protein sequence in the Protein Data Bank (June 15, 2003 release) was run against the set of family sequence profiles, using RPS-BLAST.[53] Any profile to sequence comparison with an E-value of 10^{-2} or lower was considered to represent a family that could be modeled on the basis of the corresponding structural template. Such families were considered to be structurally covered.

Results

Protein family clustering

Domain-based protein family set

Following the clustering procedure described in Procedures, 178,310 sequences from 67 sequenced prokaryotic genomes were parsed to 249,574 domain sequences and then clustered into 31,874 sequence families. Figure 3(a) shows the distribution of family sizes. Small families predominate. There are 20,992 singletons (families with only one member), about two-thirds of the total, and 4810 doubletons (family size two). At the other end of the spectrum, there are only 263 families larger than 100.

From the point of view of structural genomics, this result is discouraging: even this small number of genomes would require over 30,000 experimental structure determinations in order to provide templates for complete modeling. However, consideration of the large fraction of proteins in the larger families leads to a different view. Figure 3(b) compares the number of families of size one, two and larger with the total number of domains those categories contain. Although about two-thirds of the families are singletons, they represent only 8% of the domains. Families of size three and larger contain 88% of the domains, and there are only just over 6000 of those. Thus, 88% structural coverage of these 67 genomes would be provided by about 6000 experimental structure determinations.

Optimization of protein family construction

As discussed in Procedures, parameters for protein family construction were optimized by comparison of generated families with those in PfamA.[24]

Families were built for 50,000 full-length protein sequences covering 721 PfamA (release 9) families. The full sequences were clustered into new families, and each such family was best matched to a PfamA family. Each generated family was compared with the corresponding PfamA one using the false positive (F_{P}) and false negative (F_{N}) fractions. The smaller these values, the better the family building procedure.

Table 3 shows the level of agreement between the generated and PfamA families as a function of the E-value threshold for accepting PSI-BLAST relationships. Families are obviously over-clustered with a cutoff of 10^{-2}, judging by the high level of false positives (F_{P}). A threshold of 10^{-4} produces many fewer false positives than 10^{-2} and a lower number of false negatives than 10^{-6}, so was chosen as the final value. (Final values of the other parameters were used for these tests.)

Table 4 shows the agreement between the generated and PfamA families as a function of the linking procedure, domain parsing and checking, and the number of merging rounds. A maximum of six rounds of PSI-BLAST with an E-score threshold of 10^{-4} was used. Three rounds of single linkage clustering with no domain processing dramatically over-clusters compared with PfamA, compressing the sequences into 285 families, as opposed to the ideal 721, with a false positive rate of 79%. Domain parsing increases the number of families to 427, at the expense of a minor increase in false negatives, from 6.9% to 8.0%. Domain checking produces a further minor improvement.

Introduction of the family size-dependent linkage scheme further improves agreement with PfamA. Three rounds of merging generate 785 families with a false positive rate of 17.9%. Merging for five rounds increases the false positive rate slightly to 19.7%.

On the basis of these tests, the final protocol adopted was six rounds of all against all PSI-BLAST using a 10^{-4} threshold, followed by three rounds of hierarchical linkage. These conditions produce 785 families, of which 278 are identical with the corresponding PfamA families, with an average false positive rate of 17.9% and a false negative rate of 7.4%. PfamA families are assembled using sensitive sequence methods and are curated by hand to reduce false negatives, so that a good clustering method should have a low false negative rate, as seen here. The higher false positive rate may partly reflect the fact that Pfam does not cluster some real relationships.

Table 3. Agreement between generated and PfamA families, as a function of the PSI-BLAST E score threshold

PSI-BLAST E-score threshold	Number of families	F_{N}	F_{P}	Number of identical families
10^{2}	569	0.087	0.325	204
10^{4}	744	0.079	0.180	276
10^{6}	788	0.087	0.176	273

Table 4. Agreement between generated and PfamA families as a function of linkage protocol, domain parsing and checking, and the number of rounds of merging

Clustering method	Number of generated families	F_{N}	F_{P}	Number of identical families
Single linkage w/o domain splitting or domain check	285	0.069	0.792	154
Single linkage w/o domain check	427	0.080	0.415	244
Single linkage w/o domain check	480	0.079	0.377	255
Hierarchical merging, three rounds	785	0.074	0.179	278
Hierarchical merging, five rounds	744	0.079	0.197	276

Domain parsing, domain checking, and hierarchical linkage all improve the quality of the generated families. On the basis of these results, a protocol of three rounds of hierarchical merging, with domain parsing and checking, was adopted.

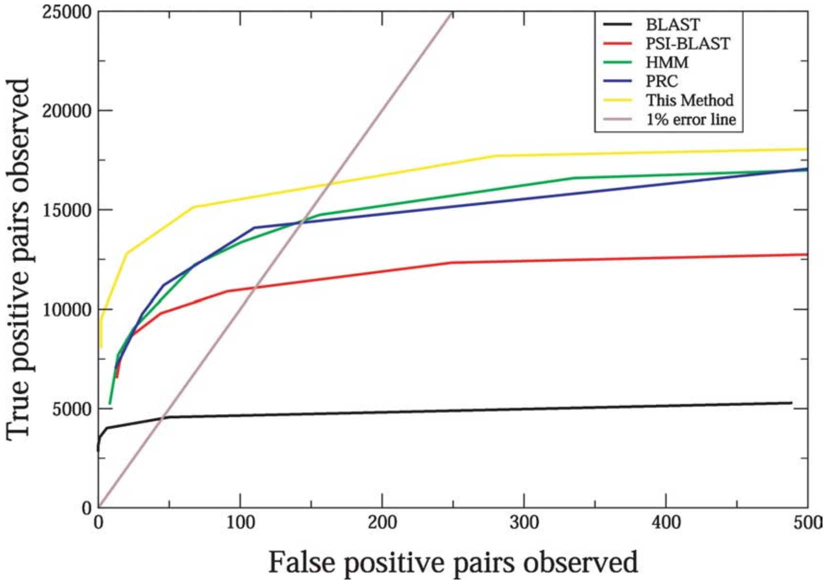

Figure 4. Benchmarking of the family building procedure, together with BLAST, PSI-BLAST, SAM-T99 and PRC, using SCOP40. True positive pairs are the fraction of pairwise relationships within superfamilies that are detected, out of 50,374 possible. False positive pairs are the fraction of apparent pairwise relationships between folds. The more true positives detected at a given false positive rate, the better the method. At a 1% ratio of false positives *versus* true positives, PSI-BLAST has approximately double the sensitivity of BLAST, the simple pairwise method. The hidden Markov model, SAM-T99 and the profile-profile method (PRC) improve the sensitivity to 28%. The new method achieves a modest but useful improvement to 32%. Improved sensitivity is attributed to the hierarchical linkage procedure.

Evaluation of the protein families

The final family building procedure was benchmarked against SCOP40 (a subset of SCOP containing no sequence identities greater than 40%) version 1.63. The SCOP set includes 5226 domain sequences grouped into 1226 superfamilies and 760 folds. As explained in Procedures, all pairwise detected relationships between proteins in the same superfamily were considered true positives, and all apparent relationships between proteins in different folds were considered false positives. Several other methods for detecting evolutionary relationships, BLAST, PSI-BLAST, SAM-T99 (a hidden Markov model method)[51] and PRC (a profile to profile method) were also evaluated.

The results are shown in Figure 4.

Overall, the new family building procedure delivers a higher fraction of true positives at a low false positive rate. At the commonly adopted threshold of 1% false positives/true positives,[47–50] BLAST detects only 9% of true positives. PSI-BLAST doubles the level of detection to 18%. SAM-T99 and PRC both detect about 28% of true positives. Our method finds 32%, a modest but useful improvement. Note that at a higher false positive rate (above 5%, not shown in the Figure), the profile–profile method performs the best. The results for BLAST, PSI-BLAST and SAM-T99 are very similar to those obtained by Park *et al*.[54] Their study showed that, using the PDBD40-J dataset (similar but smaller than SCOP40), BLAST is able to detect 14% of homologous relationships and the two profile methods, PSI-BLAST and SAM-T98, can detect 27% and 29%, respectively.

Structural genomics analysis

Structure coverage of current protein families

A long-term aim of structural genomics is to obtain an experimentally determined structure for at least one protein in every family. We now ask to what extent that is already the case for the set of 67 bacterial protein families. We consider only families with three or more members, and exclude

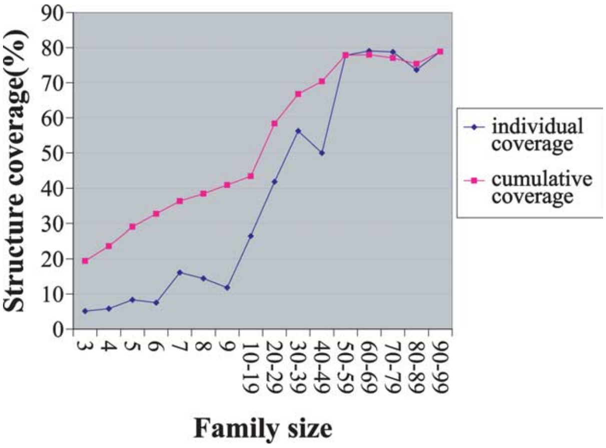

Figure 5. Fraction of the non-membrane protein families with three or more members for which there is at least one experimentally determined structure, as a function of family size (blue line). The purple line shows the coverage of all families that size and larger. Coverage is much larger for the larger families, approaching 80% for the biggest. The overall average coverage is 20%.

membrane protein families, since this class of structure is not yet amenable to high-throughput experimental techniques. There are 4907 non-membrane protein families with three or more members. Figure 5 shows the fraction of families with one or more known structures (the structure coverage). About 80% of families larger than 60 have a structural representative. Coverage drops with decreasing family size, to around 5% for families with only three members. Overall, 20% of all families size three or larger have one or more representative structures. A further 3926 structures would be required to complete the coverage.

Estimation of the number of families in a large number of genomes

The previous section provides an estimate of the number of structures needed to complete coverage of a set of already fully sequenced genomes. Of more interest in structural genomics is the number of structures that will be needed as a function of the number of sequenced genomes and, in particular, the limit of that quantity, i.e. the number of structures that will be needed to provide coverage of all protein families.

We have examined the increase in the number of detectable protein families as the number of fully sequenced genomes increases, using the following procedure. One of the 67 prokaryotic genomes is chosen at random, and the number of families it contains noted. A second genome is selected randomly, and the additional families present in that are added. This process is continued until all 67 have been selected. The whole procedure was repeated 100 times, and the average number of

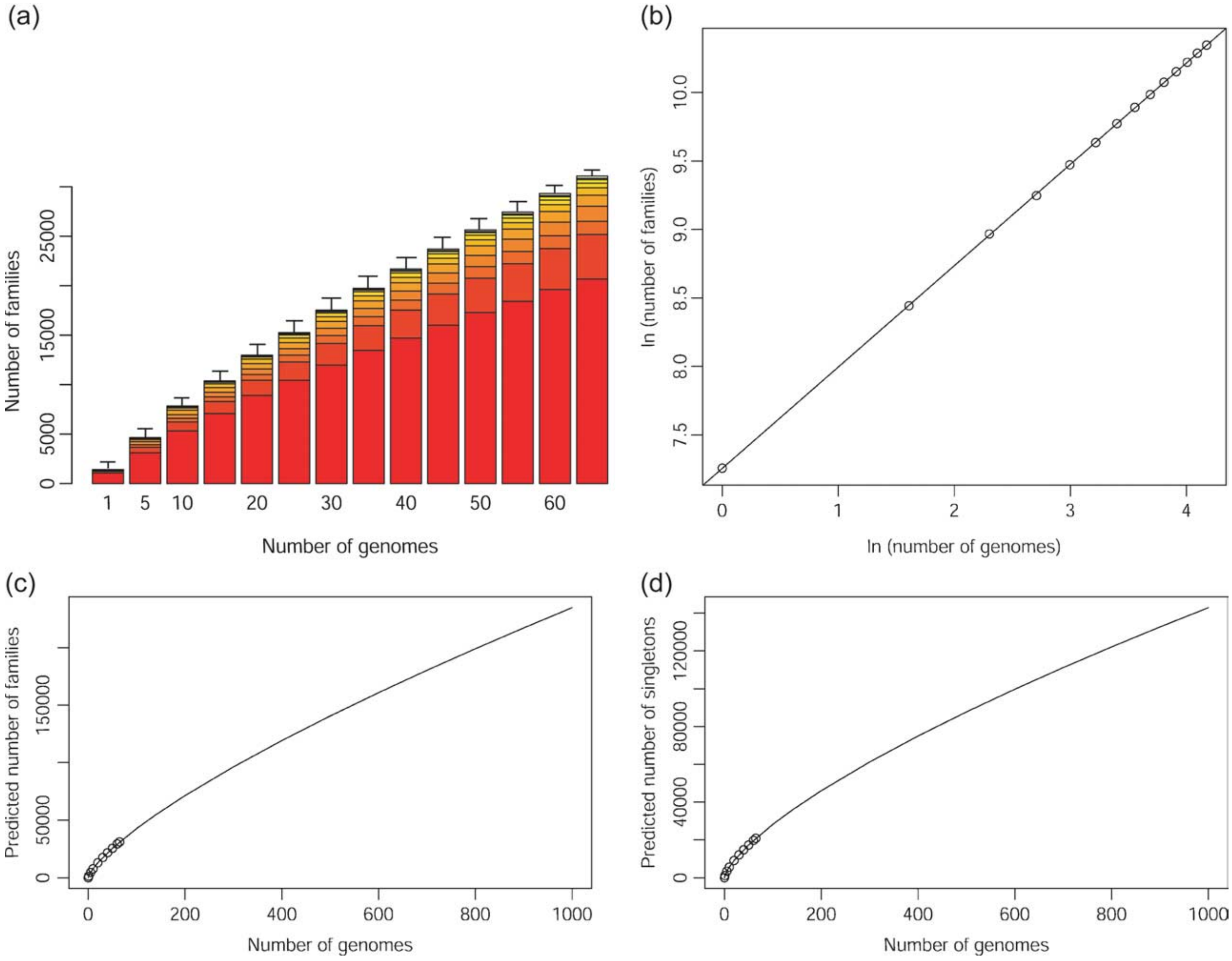

Figure 6. (a) Number of families as a function of the number of genomes. Full columns show the total families in the corresponding number of genomes, and subdivisions show the number of families in the following size ranges: 1, 2, 3, 4 or 5, 6–10, 11–20, 21–40, 41–70, 71–100, 101–1000. Smaller families are in the lower subdivisions. The total number of families is still increasing rapidly up to 67 genomes, and is far from saturation, though there is some decrease in the rate of growth. The singleton group is the fastest grower. The T bars show the standard deviations in the average number of families, over the 100 simulations. (b) Log–log view of the relationship between the number of families and number of genomes considered. A linear model gives an excellent fit to the data. (c) Predicted number of families as a function of the number of fully sequenced prokaryotic genomes, based on a log-linear fit to (b). The model predicts there will be about 250,000 families when 1000 genome sequences are available. (d) Predicted number of apparent singletons as a function of the number of fully sequenced prokaryotic genomes. The model predicts that there will be about 140,000 when the sequences of 1000 genomes are available.

families for each number of genomes calculated. The result is shown in Figure 6(a). Total bar heights represent all families in the corresponding number of genomes. Subdivisions show the number of families in different size ranges, with smallest families lowest. The number of protein families is still growing rapidly up to inclusion of 67 genomes, and is far from saturation, though the rate of increase is slowing. Clearly there will eventually be many more than 30,000 detectable families. A log-log representation of these data (Figure 6(b)) is close to linear, providing a basis for extrapolation to a larger number of genomes. Figure 6(c) shows the projected number of families up to a total of 1000 genomes, using that relationship. This model predicts a total of about 250,000 families at that point, a much higher estimate than any previous ones. There is of course a limit to how far this extrapolation can be made. At some point, as most newly sequenced genomes become phylogenetically close to known ones, the number of new families must begin to decrease more rapidly. Available data for up to 140 genomes show the model holding up remarkably well (see below).

The log of the number of apparent singletons also grows approximately linearly with the log of the number of genomes, with a reliability coefficient of 0.9993, and a slope of 0.704. The projected number of singletons out of 1000 genomes is shown in Figure 6(d). For 1000 genomes, the estimate is 140,000 singletons.

The rapid growth of singletons in Figure 6(a) and the prediction made in Figure 6(d) clearly suggest their growth is also far from complete. This observation is contradictory to our earlier view that aggregation of homologs between genomes will lead to rapid disappearance of singletons.[8]

It should be borne in mind that, because of the limited sensitivity of sequence methods for detecting relationships, the large number of families does not imply a similar magnitude of independent evolutionary lines.

To test the extrapolation model, we have extended the study to include 140 prokaryotic genomes (the 67 used for the extrapolation plus 73 new ones, see Procedures) and built families for this set using the same procedure. The 405,709 sequences in these genomes produce 54,234 families, of which 36,457 are apparent singletons. The extrapolation models predict 54,910 families and 35,807 singletons, within 1% and 2%, respectively, of the actual values.

Structural coverage for the 67 genome set

The previous analysis shows that it will not be possible to obtain complete structural coverage of protein family space in the near future. However, as noted earlier, a relatively small fraction of the families contain a large fraction of all the sequences. For the 67 genome analysis, 19% of the families are size three and larger, but contain 88% of the proteins. This suggests a strategy of obtaining representative structures for the largest families first. Figure 7 shows an exploration of this idea for the 67 genome set. We assume that a representative structure is first obtained for the largest family, then the next largest, and so on. The blue curve shows the result for all non-membrane protein families with three or more members. The purple curve shows the number of structures needed, taking into account the already available structures. Because of existing high coverage, very few additional ones will be needed for large families. Altogether, about 4000 structures are required to obtain complete coverage of all families with three or more members, covering 88% of the domains in these genomes. (As discussed earlier, about 20% of these

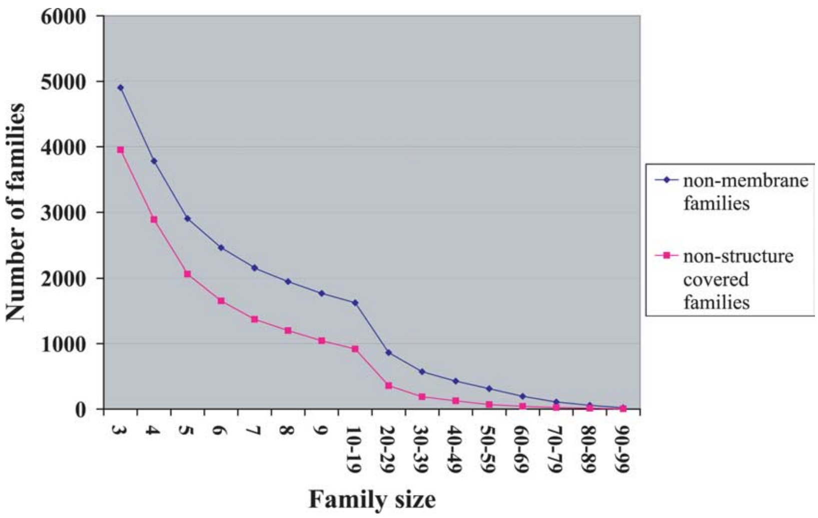

Figure 7. Cumulative number of experimental structures needed to obtain complete coverage of families size three and larger, starting with large families (right side of the plot). The blue curve is for all non-membrane protein families, and the purple curve is for the families with no current structural coverage. Very few additional structures are needed to complete coverage of large families: 1000 optimally selected ones would complete coverage of all non-membrane families larger than ten, including 80% of all the domains. About 4000 would be needed to provide one structure per family size three or larger, and would cover 88% of all the domains. (These numbers are for the set of 67 genomes analyzed in this work.)

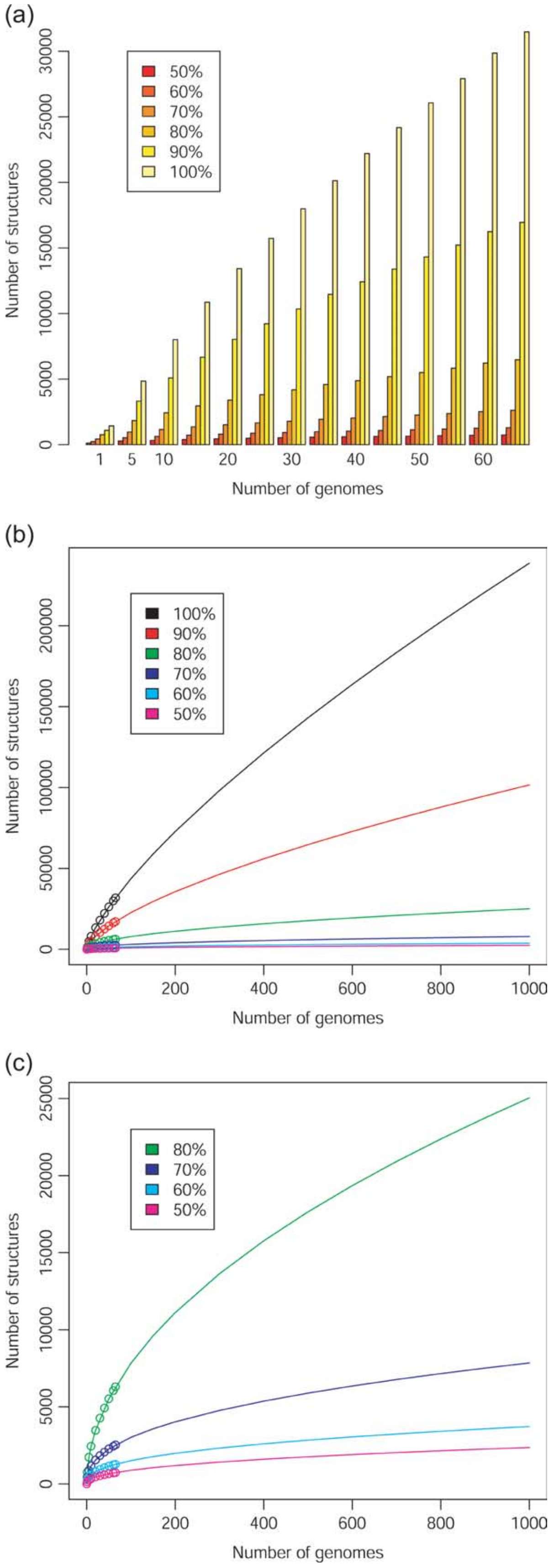

Figure 8. (a) Number of families with representative structures needed to provide structural coverage for different fractions of protein domains, as a function of the number of fully sequenced genomes. The lower the domain coverage required, the slower the growth in the number of families. (b) Projection of the number of families with representative structures needed to obtain structural coverage of different fractions of protein domains, up to 1000 genomes. A total of 250,000 structures would be required to obtain 100% coverage of these families, but 90% coverage would be obtained for less than half of that number. (c) Expansion of (b) for coverage between 50% and 80%. For 1000 genomes, approximately 8000 structures are needed to provide 70% domain coverage, achievable in the next decade, considering the rate of accumulation of solved structures.

families already have representative structures.) A total of 1000 structures will complete coverage for all non-membrane families with more than ten members, covering 80% of the domains in these genomes.

Achievable structural coverage for 1000 genomes

We now examine how many structures will be needed to achieve a given level of protein coverage, as the number of fully sequenced genomes grows. For that purpose, an extrapolation procedure similar to that described earlier was used. A genome was picked at random from the set of 67. The number of families was then calculated for that genome alone. The number of structures needed to obtain coverage of various fractions of all the proteins in that genome was calculated, assuming structures for the largest families are obtained first. Another genome was then selected randomly, and the number of structures needed to obtain various fractions of domain coverage for the two genomes was calculated, and so on, up to 65 genomes. The simulation was repeated 100 times, and the results averaged, to remove bias in genome order.

Figure 8(a) shows the results. Here, 100% coverage implies models for all domains in all families, 90% coverage implies that 90% of the domains will have models, and so forth. The general trend is that the lower the domain coverage required, the slower the growth of the number of structures needed, as a function of the number of genomes. The growth rates for 80% and 90% coverage are already decreasing when 65 genomes are considered, and growth has almost ceased for 50% and 60% coverage. Figure 8(b) shows the estimates for up to 1000 genomes, based on log-linear models. At that stage, less than half of the number of structures are needed for 90% coverage as for 100%, and the growth rate for 70% or lower coverage is slow. Figure 8(c) shows an expansion of the region below 80% coverage. Representative structures for about 8000 families will provide 70% coverage of all the domains in 1000 genomes. This is a reasonable expectation for the next decade, given the rate of accumulation of new experimental structures.

Discussion

A principal goal of structure genomics is to obtain structures for a large fraction of naturally occurring proteins. This goal can be achieved by experimentally determining at least one structure for each protein family and building structure models for all other proteins, using comparative modeling methods.[55] The minimum number of experimental structures required for complete structural coverage of protein space is then equal to the number of apparent protein families. In a previous study,[8] we estimated this number by analyzing PfamA families,[24] and making a very simple extrapolation of likely future growth in the number of families.

In the present study, we have based the analysis on all families in a set of fully sequenced prokaryotic genomes, rather than the contents of PfamA. A major difference is the inclusion of all proteins, not just those in the larger families typically collected in PfamA. With this more realistic view of the protein universe, we find there are a very large number of such small families: for the set of 67 genomes analyzed, there are 25,802 families with only one or two members, out of a total of 31,874. Overall, there is an approximately power law relationship between the number of families and family size. In a comprehensive analysis of five eukaryotic genomes, Liu *et al*. also found a large number of small families, unique to that kingdom.[37]

Use of complete genome sequence sets has also allowed us to use a more realistic extrapolation method, in order to estimate the future growth in the number of families, as the number of fully sequenced genomes grows. We find that when 1000 genomes are available, there will be about 250,000 detectable protein families. Further, the number of families will still be growing at that point.

The large number of families makes it clear that complete structural coverage of protein space will not be possible in the near future. Nevertheless, it will be possible to obtain structural models for a large fraction of proteins. This is because most proteins belong to large families; for the 67 sequenced genomes, 88% of the proteins fall into just 6072 families. Further, the extrapolation model shows that this trend will continue, so that, considering all sequences, 80% structural coverage of the proteins in 1000 genomes can be obtained with 25,000 structures, and 70% coverage with 8000. The primary conclusion from this work is that a strategy of obtaining structural representatives for the largest families first will lead to a large fraction of structural coverage of protein space within the next decade. This strategy will also lead to early structural coverage of the families that perform more universal biological functions, and will provide the most leverage of experimental effort, by creating models for the largest number of proteins from each experimental structure. We envisage that when structures for proteins in small families are needed, they will typically be obtained one at a time, using conventional structural biology, rather than high-throughput methods.

The number of apparent protein families depends on the effectiveness of each of the steps in building them. There are three keys steps in our procedure. The first step uses PSI-BLAST to search for relatives of each protein. Other methods, in particular well-tuned hidden Markov models[51] and profile-profile methods,[27–29] are more sensitive for this purpose.[54,27] At low false positive rates, the later merging step compensates for the relative insensitivity of PSI-BLAST.

The second step of family building is parsing of proteins into domains. We have used a sequence profile-based approach, relying on the fact that the most insertions and deletions occur between domains.[16,21] We apply the procedure very conservatively to minimize splitting within domains. As a consequence, this step has many false negatives; it does not split at many domain boundaries that are obvious at the structure level. Additional procedures might further improve our method: mapping known structural domains and PfamA domains onto the proteins. We have not done that, because the majority of these families have not been studied structurally, and many are not in PfamA, so that use of these signals may distort the choice of parameters for family building.

The third step in family building is merging lists of related domains and filtering out redundant entries, to create domain families. As noted earlier, over-merging is a well known problem in protein family building; a small number of incorrect entries in the initial lists of relatives can easily lead to substantial over-merging. To avoid this, we use a procedure that requires an increasing number of common entries as a function of alignment size.

The rules for merging and other steps were tuned by reconstructing a set of PfamA domains from the corresponding full-length sequences, and comparing the generated families with the PfamA ones. The final procedure was benchmarked by comparing pairwise relationships within a set of generated families with those in a set of SCOP superfamilies. While these testing methods are very useful, they are not ideal. PfamA is a sequence-based family set, and so omits a large number of evolutionary relationships (placing related proteins in different families). A more sensitive method may therefore appear to have an excessively large number of false positives, and consequently may be detuned to reduce these. PfamA also focuses on larger families, whereas the genome data are dominated by smaller families. As a result, a better method for PfamA may not necessarily be optimum on genome data, and performance may be different from that suggested by quality measures on PfamA. Similarly, SCOP contains only proteins with known structure, and these may be unrepresentative of proteins in genomes as a whole; for example, not including proteins with significant inherent disorder, and under-representing proteins that form part of complexes. Nevertheless, PfamA and SCOP are probably the best training and test sets available. As in most of computational biology, the lack of a gold standard for methods development and evaluation is an inherent limitation.

According to our and other benchmarking, at a 1% ratio of false positives *versus* true positives, only about 30% of the pairwise evolutionary relationships implied by structure can be recovered with present sequence comparison methods. A consequence is that families built with those methods do not approximate independent evolutionary lines. As more structures are available, the number of families will decrease very substantially, because of merging on the basis of structural similarity, rather than sequence. For the purposes of structural genomics, a single representative structure for very large families containing very remote relatives is not particularly desirable. As the remoteness of the relationship between proteins increases, the quality of a model built on the basis of a relative with an experimental structure decreases. In particular, a substantial fraction of residues (up to 50%) will have no equivalent in the modeling template.[56] Thus, although families generated from sequence relationships are suboptimal from an evolutionary standpoint, they are very suitable for structural genomics.

Contrary to our earlier expectation,[8] the number of apparent singletons and other small families will continue to increase. Siew & Fischer also found the number of singletons is steadily growing, though the percentage of singletons as a fraction of all sequences is decreasing.[38] Because of the limited sensitivity of sequence methods, it is not possible to judge the biological significance of this at present. Many singletons may, in fact, have unrelated folds, as one estimate of the total number of folds suggests.[57] Or, most may turn out to be members of larger superfamilies, too remotely related for sequence methods to detect. A larger set of experimental structures of small families will settle this issue.

This work assessed how many experimental structures will be needed to provide models of a given fraction of all naturally occurring domains, based on one representative structure per family. Under this strategy, the majority of structures will be domain models, based on a single experimental structure within a protein family. While such a structure set will revolutionize our view of proteins in many ways, it is only the first step in providing complete structural information for natural proteins. Many proteins are multi-domain, particularly in higher eukaryotes,[58,59] and the function of a domain assembly is not always a simple combination of that in the constituent domains.[59] Generating reliable multi-domain structures will sometimes involve docking of domain models, requiring improvements in computational methods, or further experimental structures. Second, the relationships within families on which models will be based are often fairly distant, with levels of sequence identity well below 30%. Models based on such sequence relationships contain substantial errors, arising primarily from mistakes in aligning the sequence of interest with those of available templates, and because significant parts of the structures will differ from that of the templates.[11] Nevertheless, these low accuracy models will be adequate for establishing membership of a superfamily, and thus useful for a variety of purposes, including providing approximate molecular function information, guiding site-directed mutagenesis experiments, and choosing likely antigenic peptides. Other uses, such as identification of ligand specificity[60] and interpretation of the effect of disease-related mutations,[61] require greater accuracy, possible only by modeling against a template with 30% or greater sequence identity. Comprehensive structural information at that level will require many more structures.

Acknowledgements

We thank Eugene Melamud for help with many aspects of data management. This work was supported, in part, by NIH grant GM57890.

References

1. Berman, H. M. & Westbrook, J. D. (2004). The impact of structural genomics on the protein data bank. *Am. J. Pharmacogenomics*, **4**, 247–252.
2. Browne, W. J., North, A. C., Phillips, D. C., Brew, K., Vanaman, T. C. & Hill, R. L. (1969). A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme. *J. Mol. Biol.* **42**, 65–86.
3. Brenner, S. E. (2001). A tour of structural genomics. *Naure Rev. Genet.* **2**, 801–809.
4. Wolf, Y. I., Grishin, N. V. & Koonin, E. V. (2000). Estimating the number of protein folds and families from complete genome data. *J. Mol. Biol.* **299**, 897–905.
5. Orengo, C. A., Todd, A. E. & Thornton, J. M. (1999). From protein structure to function. *Curr. Opin. Struct. Biol.* **9**, 374–382.
6. Sternberg, M. J., Bates, P. A., Kelley, L. A. & MacCallum, R. M. (1999). Progress in protein structure prediction: assessment of CASP3. *Curr. Opin. Struct. Biol.* **9**, 368–373.
7. Liu, J. & Rost, B. (2002). Target space for structural genomics revisited. *Bioinformatics*, **18**, 922–933.
8. Vitkup, D., Melamud, E., Moult, J. & Sander, C. (2001). Completeness in structural genomics. *Nature Struct. Biol.* **8**, 559–566.
9. Tramontano, A. & Morea, V. (2003). Assessment of homology-based predictions in CASP5. *Proteins: Struct. Funct. Genet.* **53**, 352–368.
10. Baker, D. & Sali, A. (2001). Protein structure prediction and structural genomics. *Science*, **294**, 93–96.
11. Venclovas, C., Zemla, A., Fidelis, K. & Moult, J. (2003). Assessment of progress over the CASP experiments. *Proteins: Struct. Funct. Genet.* **53**, 585–595.

12. Nagano, N., Orengo, C. A. & Thornton, J. M. (2002). One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol.* **321**, 741–765.
13. Dayhoff, M. O. (1976). The origin and evolution of protein superfamilies. *Fed. Proc. Fed. Am. Soc. Expt. Biol.* **35**, 2132–2138.
14. Haft, D. H., Selengut, J. D. & White, O. (2003). The TIGRFAMs database of protein families. *Nucl. Acids Res.* **31**, 371–373.
15. Servant, F., Bru, C., Carrere, S., Courcelle, E., Gouzy, J., Peyruc, D. & Kahn, D. (2002). ProDom: automated clustering of homologous domains. *Brief Bioinform.* **3**, 246–251.
16. Gracy, J. & Argos, P. (1998). Automated protein sequence database classification. II. Delineation of domain boundaries from sequence similarities. *Bioinformatics*, **14**, 174–187.
17. Gracy, J. & Argos, P. (1998). Automated protein sequence database classification. I. Integration of compositional similarity search, local similarity search, and multiple sequence alignment. *Bioinformatics*, **14**, 164–173.
18. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucl. Acids Res.* **30**, 1575–1584.
19. Mohseni-Zadeh, S., Brezellec, P. & Risler, J. L. (2004). Cluster-C, an algorithm for the large-scale clustering of protein sequences based on the extraction of maximal cliques. *Comput. Biol. Chem.* **28**, 211–218.
20. Gough, J. (2002). The SUPERFAMILY database in structural genomics. *Acta Crystallog. sect. D*, **58**, 1897–1900.
21. Heger, A. & Holm, L. (2001). Picasso: generating a covering set of protein family profiles. *Bioinformatics*, **17**, 272–279.
22. Yona, G., Linial, N. & Linial, M. (2000). ProtoMap: automatic classification of protein sequences and hierarchy of protein families. *Nucl. Acids Res.* **28**, 49–55.
23. Wu, C. H., Xiao, C., Hou, Z., Huang, H. & Barker, W. C. (2001). iProClass: an integrated, comprehensive and annotated protein classification database. *Nucl. Acids Res.* **29**, 52–54.
24. Sonnhammer, E. L., Eddy, S. R., Birney, E., Bateman, A. & Durbin, R. (1998). Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucl. Acids Res.* **26**, 320–322.
25. Hubbard, T. J., Murzin, A. G., Brenner, S. E. & Chothia, C. (1997). SCOP: a structural classification of proteins database. *Nucl. Acids Res.* **25**, 236–239.
26. Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., Chothia, C. & Murzin, A. G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucl. Acids Res.* **32**, D226–D229.
27. Ohlson, T., Wallner, B. & Elofsson, A. (2004). Profile-profile methods provide improved fold-recognition: a study of different profile–profile alignment methods. *Proteins: Struct. Funct. Genet.* **57**, 188–197.
28. Edgar, R. C. & Sjolander, K. (2004). COACH: profile-profile alignment of protein families using hidden Markov models. *Bioinformatics*, **20**, 1309–1318.
29. Sadreyev, R. & Grishin, N. (2003). COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.* **326**, 317–336.
30. Brenner, S. E., Chothia, C. & Hubbard, T. J. (1997). Population statistics of protein structures: lessons from structural classifications. *Curr. Opin. Struct. Biol.* **7**, 369–376.
31. Chothia, C. (1992). Proteins. One thousand families for the molecular biologist. *Nature*, **357**, 543–544.
32. Holm, L. & Sander, C. (1996). Mapping the protein universe. *Science*, **273**, 595–603.
33. Wang, Z. X. (1998). A re-estimation for the total numbers of protein folds and superfamilies. *Protein Eng.* **11**, 621–626.
34. Zhang, C. & DeLisi, C. (1998). Estimating the number of protein folds. *J. Mol. Biol.* **284**, 1301–1305.
35. Govindarajan, S., Recabarren, R. & Goldstein, R. A. (1999). Estimating the total number of protein folds. *Proteins: Struct. Funct. Genet.* **35**, 408–414.
36. Grant, A., Lee, D. & Orengo, C. (2004). Progress towards mapping the universe of protein folds. *Genome Biol.* **5**, 107.
37. Liu, J., Hegyi, H., Acton, T. B., Montelione, G. T. & Rost, B. (2004). Automatic target selection for structural genomics on eukaryotes. *Proteins: Struct. Funct. Genet.* **56**, 188–200.
38. Siew, N. & Fischer, D. (2003). Analysis of singleton ORFans in fully sequenced microbial genomes. *Proteins: Struct. Funct. Genet.* **53**, 241–251.
39. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
40. Altschul, S. F. & Koonin, E. V. (1998). Iterated profile searches with PSI-BLAST-a tool for discovery in protein databases. *Trends Biochem. Sci.* **23**, 444–447.
41. Wootton, J. C. (1994). Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem.* **18**, 269–285.
42. Liu, J. & Rost, B. (2004). CHOP: parsing proteins into structural domains. *Nucl. Acids Res.* **32**, W569–W571.
43. Liu, J. & Rost, B. (2004). Sequence-based prediction of protein domains. *Nucl. Acids Res.* **32**, 3522–3530.
44. Pipenbacher, P., Schliep, A., Schneckener, S., Schonhuth, A., Schomburg, D. & Schrader, R. (2002). ProClust: improved clustering of protein sequences with an extended graph-based approach. *Bioinformatics*, **18**, S182–S191.
45. Yona, G., Linial, N. & Linial, M. (1999). ProtoMap: automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. *Proteins: Struct. Funct. Genet.* **37**, 360–378.
46. Kahsay, R. Y., Wang, G., Dongre, N., Gao, G. & Dunbrack, R. L., Jr (2002). CASA: a server for the critical assessment of protein sequence alignment accuracy. *Bioinformatics*, **18**, 496–497.
47. Brenner, S. E., Chothia, C. & Hubbard, T. J. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.
48. Park, J., Teichmann, S. A., Hubbard, T. & Chothia, C. (1997). Intermediate sequences increase the detection of homology between sequences. *J. Mol. Biol.* **273**, 349–354.
49. Park, J., Holm, L., Heger, A. & Chothia, C. (2000). RSDB: representative protein sequence databases have high information content. *Bioinformatics*, **16**, 458–464.
50. Enright, A. J. & Ouzounis, C. A. (2000). GeneRAGE: a robust algorithm for sequence clustering and domain detection. *Bioinformatics*, **16**, 451–457.

51. Karplus, K. & Hu, B. (2001). Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set. *Bioinformatics*, **17**, 713–720.
52. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580.
53. Marchler-Bauer, A., Panchenko, A. R., Shoemaker, B. A., Thiessen, P. A., Geer, L. Y. & Bryant, S. H. (2002). CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucl. Acids Res.* **30**, 281–283.
54. Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. & Chothia, C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284**, 1201–1210.
55. Marti-Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., Melo, F. & Sali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291–325.
56. Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826.
57. Coulson, A. F. & Moult, J. (2002). A unifold, mesofold, and superfold model of protein fold use. *Proteins: Struct. Funct. Genet.* **46**, 61–71.
58. Vogel, C., Bashton, M., Kerrison, N. D., Chothia, C. & Teichmann, S. A. (2004). Structure, function and evolution of multidomain proteins. *Curr. Opin. Struct. Biol.* **14**, 208–216.
59. Bashton, M. & Chothia, C. (2002). The geometry of domain combination in proteins. *J. Mol. Biol.* **315**, 927–939.
60. DeWeese-Scott, C. & Moult, J. (2004). Molecular modeling of protein function regions. *Proteins: Struct. Funct. Genet.* **55**, 942–961.
61. Wang, Z. & Moult, J. (2001). SNPs, protein structure, and disease. *Hum. Mutat.* **17**, 263–270.

Edited by B. Honig

(Received 25 March 2005; received in revised form 18 August 2005; accepted 24 August 2005)
Available online 9 September 2005