Protein Science (2000), *9*: 197-200. Cambridge University Press. Printed in the USA.

Expectations from structural genomics

STEVEN E. BRENNER [1] and MICHAEL LEVITT [1]

[1] Department of Structural Biology, Stanford University, Fairchild Building D-109, Stanford, California 94305-5126

(Received January 29, 1999; Accepted October 8, 1999)

Reprint requests to current address: Steven E. Brenner, Department of Plant and Microbial Biology, University of California, 461A Koshland Hall, Berkeley, California 94720-3102;
e-mail:brenner@compbio.berkeley.edu.

Abstract

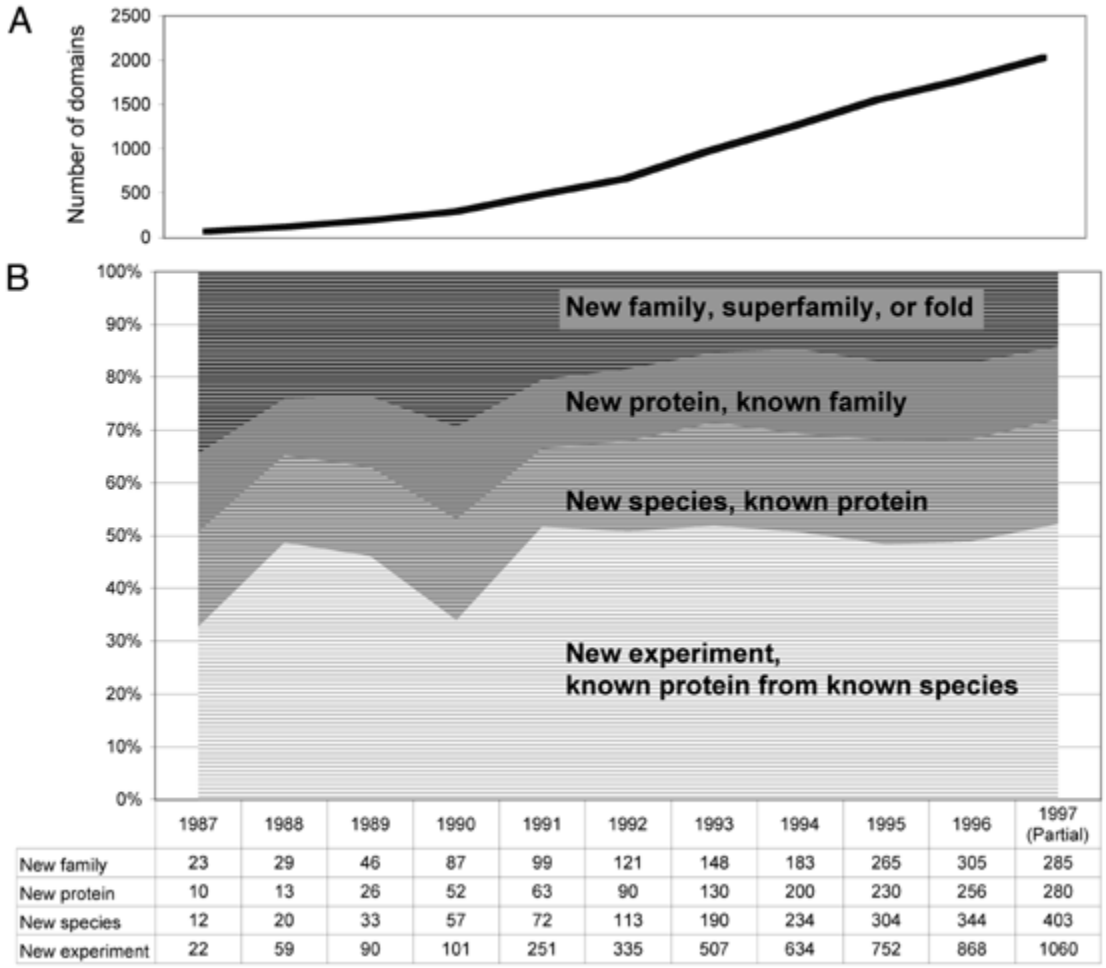
Structural genomics projects aim to provide an experimental structure or a good model for every protein in all completed genomes. Most of the experimental work for these projects will be directed toward proteins whose fold cannot be readily recognized by simple sequence comparison with proteins of known structure. Based on the history of proteins classified in the SCOP structure database, we expect that only about a quarter of the early structural genomics targets will have a new fold. Among the remaining ones, about half are likely to be evolutionarily related to proteins of known structure, even though the homology could not be readily detected by sequence analysis.

Keywords: protein folds, SCOP; structural classification; structural genomics; superfamilies

Introduction

Structural genomics projects (Kim, 1998; Sali, 1998) are being driven by two fundamentally antithetical goals. One aim is to yield a complete representative set of protein folds. The other is to provide insights into the function of genome-encoded proteins, principally by recognizing homology between two structures with the same fold, but whose similarity could not be detected by sequence analysis. To what extent are each of these goals likely to be fulfilled? A study of the protein structures solved in recent years (Orengo et al., 1994; Holm & Sander, 1996; Brenner et al., 1997) can help answer this question.

The SCOP database (Murzin et al., 1995) organizes proteins according to their structural and evolutionary relationships. Figure 1 shows how SCOP 1.40s classifies protein domain structures submitted to the Protein Data Bank (PDB) (Bernstein et al., 1977) between 1987 and 1997. Even as the number of domains studied has grown dramatically, the nature of the sequences studied has been comparatively constant. Slightly more than half of the protein domains submitted to the PDB in 1997 represent a new experiment on a protein sequence identical or nearly identical to one already in the database, perhaps with some mutations, under different conditions, in a larger complex, or with bound ligands. A further 20% of the domains were from a protein for which a structure had already been solved from a different species, and 14% were new proteins for which there was a known structure of a homolog in the same family. In sum, more than 85% of the new protein domain structures experimentally determined were in the same SCOP family as a protein already in the PDB. Generally, relationships between these proteins could have been recognized by sequence comparison, and it should have

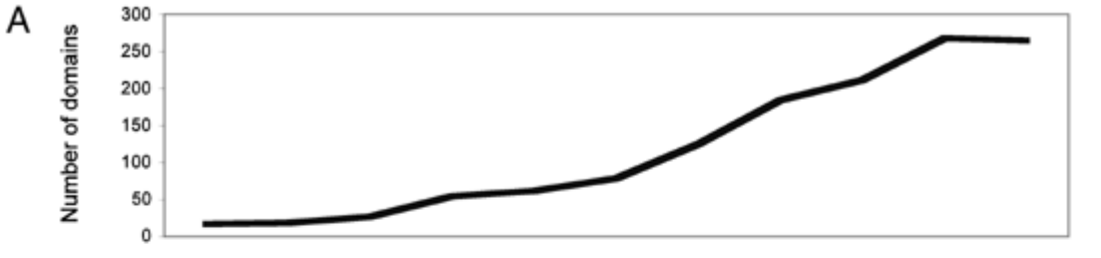


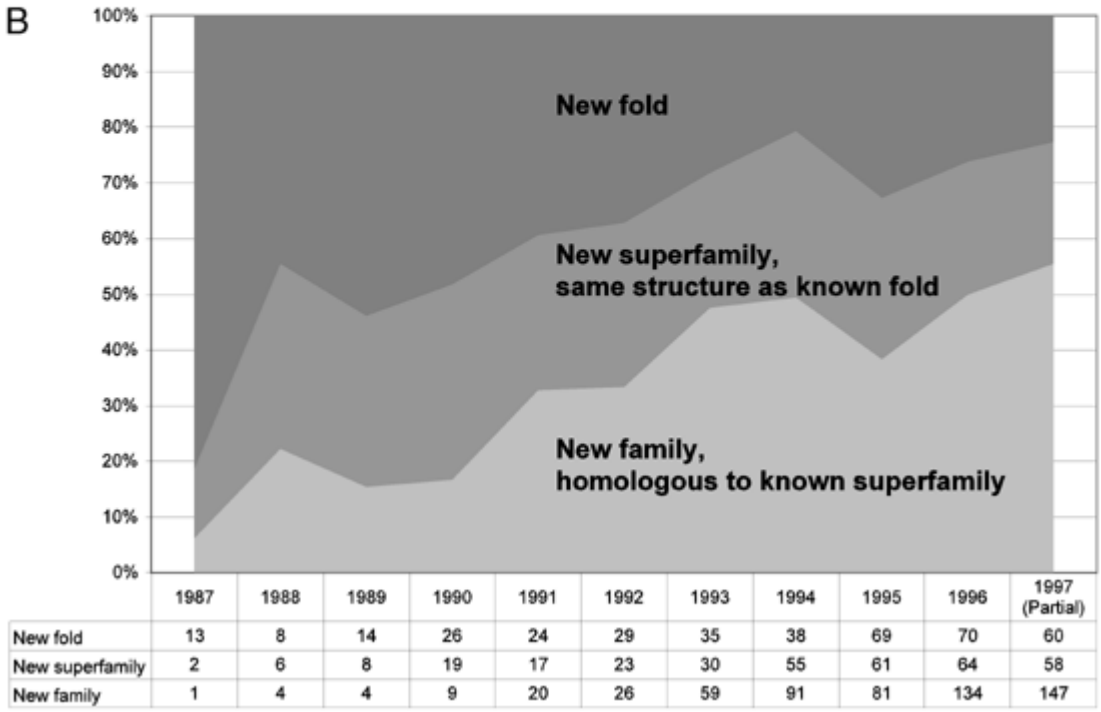
	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997 (Partial)
New family	23	29	46	87	99	121	148	183	265	305	285
New protein	10	13	26	52	63	90	130	200	230	256	280
New species	12	20	33	57	72	113	190	234	304	344	403
New experiment	22	59	90	101	251	335	507	634	752	868	1060

Fig. 1 Experimentalists' selection of proteins. The figure describes the category of new information, as described by SCOP 1.40s, provided by protein domains experimentally solved and submitted to the PDB in 1987-1997. Data for 1997 are not yet complete, as release holds mean that 189 PDB entries (10%) have not been classified. Obsolete PDB entries were classified in the same way as domains of their superseding entries. Only proteins in the main classes (1-7) of the SCOP classification were considered, and only a single representative of domains in a given PDB entry with identical classifications (e.g., homodimers or crystallographically related molecules) was included. Each entry is taken as a separate experiment. A: Number of domains considered for each year. B: Classification of new protein domains each year according to SCOP. The relative number of proteins in each category has been relatively constant, despite the immense growth in the absolute number.

been possible to structurally model the protein domains by computational methods. Presumably, these proteins were experimentally studied because of the need to obtain detailed structural information or knowledge of the domain's context. As the categorization in Figure 1 was recognizable from sequence in advance of structure determination, the distribution reflects the interests of the experimental structural biology community.

Figure 2 shows what was discovered from the proteins lacking significant pairwise sequence similarity to those already in the protein database. For these proteins, classification in SCOP requires knowledge of the structure; sequence would fail to predict these categories. In 1997, fewer than a quarter of such protein domains had a new fold, compared with about a half in 1990. Even when more sensitive sequence comparison methods are used, like PSI-BLAST in Figure 3, only 26% of unrecognizable sequences represent new folds.

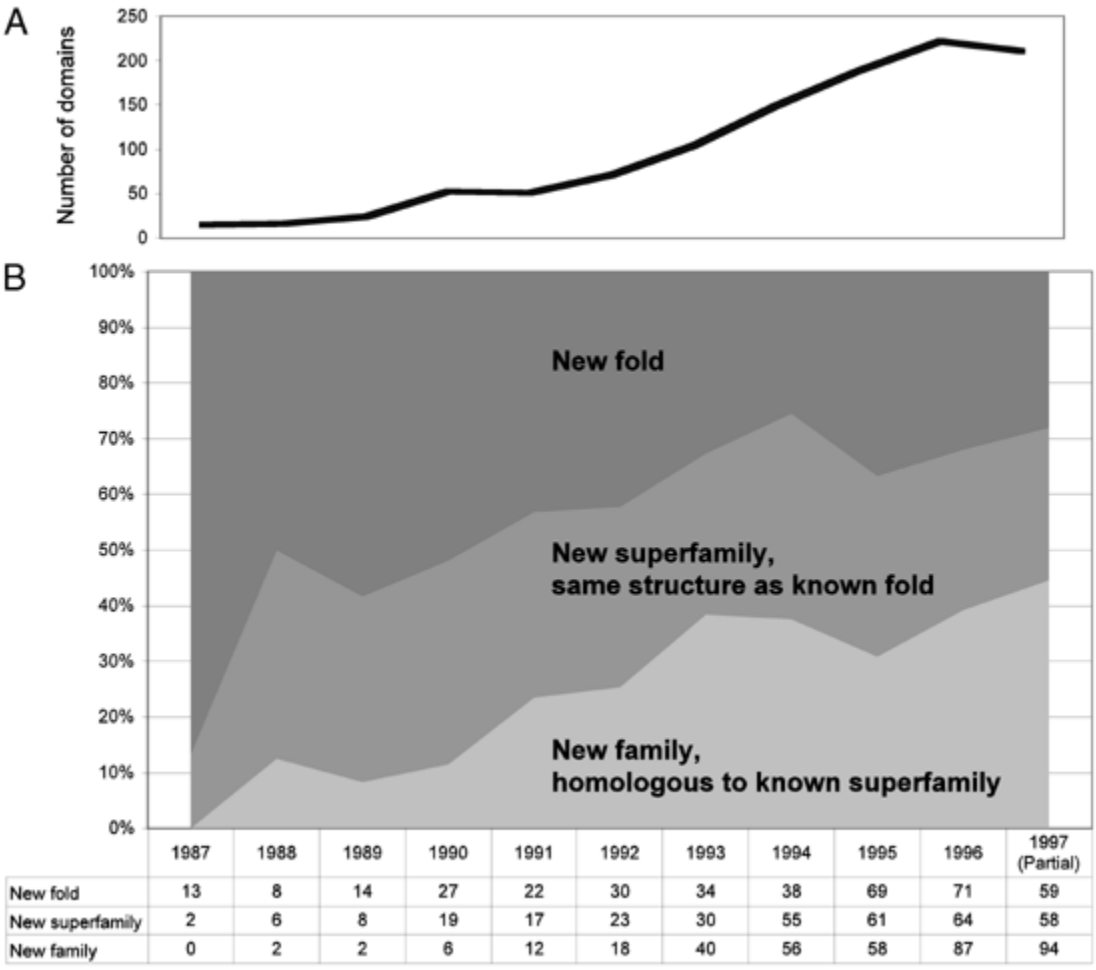




	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997 (Partial)
New fold	13	8	14	26	24	29	35	38	69	70	60
New superfamily	2	6	8	19	17	23	30	55	61	64	58
New family	1	4	4	9	20	26	59	91	81	134	147

Fig. 2 What was learned from proteins without significant pairwise sequence similarity to known structures? Protein domains without sequence similarity to those already in the PDB are shown according to the degree of novelty revealed by their solved structures. The proteins considered here roughly correspond to those in the "new family, superfamily, or fold" category in Figure 1; they were selected if they were in the ASTRAL (Brenner et al., 2000) set of sequences from SCOP 1.40s (i.e., classes 1-7 and having sequences more than 20 contiguous residues with few ambiguities), and they did not have a pairwise BLASTPGP 2.0.9 (Altschul et al., 1997) *E*-value score of <=0.01 to any other such sequence presently in the PDB on the accession date. Obsolete entries were not considered for this analysis. The family level in this graph incorporates all proteins that were homologous according to SCOP even if they were classified at a more specific level. A: Number of domains considered for each year. B: The fraction of new folds has shrunk over the years, while the number of homologs detected by structure detected has grown greatly.

This suggests that the 459 protein folds in the most recent SCOP incorporate a majority of the frequently occurring globular structures. From this trend, it might seem that all of the most common folds may soon be known. However, several analyses suggest that the frequency of different folds is highly skewed, so that new structures will continue to be found, albeit increasingly rarely, for a very long time to come (Brenner et al., 1997; Wang, 1998; Zhang & DeLisi, 1998; Govindarajan et al., 1999). Moreover, we still know little about those structures--such as membrane proteins--that are difficult to characterize structurally.



	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997 (Partial)
New fold	13	8	14	27	22	30	34	38	69	71	59
New superfamily	2	6	8	19	17	23	30	55	61	64	58
New family	0	2	2	6	12	18	40	56	58	87	94

Fig. 3 What was learned from proteins without significant PSI-BLAST sequence similarity to known structures? More sophisticated comparison methods permit the detection of additional homologs by sequence alone. When PSI-BLAST is used to select proteins without significant sequence similarity, the number of new folds and superfamilies stay effectively the same, while the number of evolutionarily related proteins detectable only by structure is reduced by about a third in years since 1994. Nonetheless, nearly half of the proteins not found similar already to known structures by PSI-BLAST can be recognized as homologous to another protein when using structure. The analysis was performed by searching each ASTRAL sequence from SCOP 1.40s against SEG-filtered (Wootton, 1994) SWISS-PROT, TREMBL, and updates through April 14, 1999 (Bairoch & Apweiler, 1999) with BLASTPGP 2.0.9 for 10 iterations or until convergence, with a matrix inclusion threshold of 1e-4. The output checkpoint files were then used to search the ASTRAL sequence database, and matches with *E*-value <=0.01 to domains already in the PDB were considered significant and excluded from this graph. An elaboration of the analysis in Brenner et al. (1998) determined that this procedures provides accuracy comparable to the pairwise BLASTPGP 2.0.9 *E*-value score of <=0.01 (S.E. Brenner, unpubl. obs.) used in Figure 2. A: Number of domains considered for each year. B: Fractions of proteins, not appearing similar to existing proteins of known structure with PSI-BLAST, whose structure reveals them to be a new fold, superfamily, or family.

Although finding a new fold is exciting, it typically needs to be augmented with further experimental information to provide functional insight. Of the proteins submitted to the PDB in 1997 without significant sequence similarity to proteins in the database, about a quarter have an existing fold but do not appear homologous to proteins already in the database: these define a new superfamily. A much larger fraction--about half, depending upon the method used--create a new family because of lack of sequence similarity, but the tertiary structure reveals them to be evolutionarily related to other proteins of known structure. Structure is most valuable in elucidating function of an otherwise uncharacterized protein in cases such as these, when it reveals that two proteins are distant evolutionary relatives. Because the two proteins came from a single ancestor, it is likely that they retain some similarity in function (Martin et al., 1998; Hegyi & Gerstein, 1999). Moreover, the

structure may provide the information necessary to evaluate whether the functional site characteristics are indeed conserved. As shown in Figures 2 and 3, the fraction of proteins in this category has grown from almost none in 1987 to about half in 1997.

A large fraction of proteins in completed genomes cannot be effectively characterized by sequence comparison, so these proteins will be candidates for experimental work in structural genomics projects. We expect that the information to be learned from these proteins will be similar to that for those in Figure 3. Both sets of proteins have no significant sequence similarity to proteins of known structure. Although selection of proteins for experimental determination has been strongly biased in the past, these biases probably have little correlation with the probability of a protein having a new fold versus showing homology to another protein, because it is impossible to recognize these categories before the structure is solved.

Because experimental structural genomics projects will focus on proteins whose structure cannot be recognized by sequence analysis, the results in Figure 3 suggest that perhaps a quarter of the structures solved will have novel folds, and that this fraction will slowly decrease. Most importantly, it is likely that structure determination will reveal that nearly half of the proteins are homologous to a protein already in the database, despite absence of significant sequence similarity. Consequently, structure determination promises to be an increasingly effective and efficient means of detecting homology, and thus suggesting molecular function for proteins.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res 25*:3389- 3402.

Bairoch A, Apweiler R. 1999. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res 27*:49-54.

Bernstein FC, Koetzle TF, Williams GJ, Meyer EE Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535-542.

Brenner SE, Chothia C, Hubbard TJP. 1997. Population statistics of protein structures: Lessons from structural classifications. *Curr Opin Struct Biol 7*:369-376.

Brenner SE, Chothia C, Hubbard TJP. 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc Natl Acad Sci USA 95*:6073-6078.

Brenner SE, Koehl P, Levitt M. 2000. The ASTRAL compendium from protein structure and sequence analysis. *Nucleic Res* Forthcoming.

Govindarajan S, Recabarren R, Goldstein RA. 1999. Estimating the total number of protein folds. *Proteins 35*:408-414.

Hegyi H, Gerstein M. 1999. The relationship between protein structure and function: A comprehensive survey with application to the yeast genome. *J Mol Biol 288(1)*:147-164.

Holm L, Sander C. 1996. Mapping the protein universe. *Science 273*:595-603.

Kim SH. 1998. Shining a light on structural genomics. *Nat Struct Biol 5*:643-645.

Martin AC, Orengo CA, Hutchinson FG, Jones S, Karmirantzou M, Laskowski RA, Mitchell JB, Taroni C, Thornton JM. 1998. Protein folds and functions. *Structure 6(7)*:875-884.

Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol 247*:536-540.

Orengo CA, Jones DT, Thornton JM. 1994. Protein superfamilies and domain superfolds. *Nature 372*:631-634.

Sali A. 1998. 100,000 protein structures for the biologist. *Nat Struct Biol 5*:1029-1032.

Wang ZX. 1998. A re-estimation for the total numbers of protein folds and superfamilies. *Protein Eng 11*:621-626.

Wootton JC. 1994. Non-globular domains in protein sequences: Automated segmentation using complexity measures. *Comput Chem 18*:269-85.

Zhang C, DeLisi C. 1998. Estimating the number of protein folds. *J Mol Biol 284*:1301-1305.