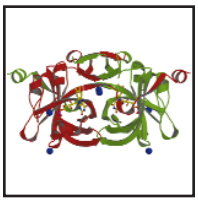
Target selection for structural genomics

Steven E. Brenner

Structural genomics aims to use high-throughput structure determination and computational analysis to provide three-dimensional models of every tractable protein. The process of choosing proteins for experimental structure characterization is known as target selection. In this nomenclature, the targets are regions of proteins to be studied by crystallography or NMR. Selection of the targets is principally a computational process of restricting candidate proteins to those that are tractable and of unknown structure, and prioritizing according to expected interest and accessibility.



To derive the most value from structural genomics efforts, projects attempt to maximize knowledge that will be gained from investment in experimental structural characterization. A primary benefit of structural genomics is expected to be the discovery of distant evolutionary relationships invisible from sequence, which may yield novel functional insights[1]. To this end, structural genomics involves ensuring that each family of proteins is represented by a known structure. Some aspects of the process are common to nearly all large-scale proposed projects, but to some extent the evaluation of what is most important reflects the particular insight and vision of the different consortia.

In the context of trying to provide a structure for every tractable protein, all projects employ exhaustively computational methods to predict structural characteristics and also to flag proteins that would lead to redundant effort or that are likely to prove impractical in a large-scale approach. In essence, these shared approaches involve exclusion of inappropriate protein families as targets. The aspects unique to each group involve the identification and prioritization of the remaining families.

While the broad variety of structural genomics efforts precludes any simple encompassing description, many projects can be generalized as having four distinct stages. These are: (i) realm identification (ii) family exclusion (iii) family prioritization, and (iv) protein and region selection.

Realm identification and family exclusion

Realm identification involves describing proteins that fall within the general universe of interest. For example, this might include proteins from a given organism, in which case the realm is simple to enumerate. Other cases, such as signaling proteins, may require considerable analysis to define. The realm of proteins is represented by the set of blue dots and stars in Fig. 1*a*, and it is described in conjunction with family prioritization, below.

The second stage, family exclusion, involves two major components. First is the identification of proteins that are likely to prove difficult or impossible to study by crystallography or NMR. In practice this means removing transmembrane and low-complexity regions (Fig. 1*b*). While three-dimensional information about membrane proteins is exquisitely valuable, the process of obtaining such structures is extraordinarily challenging and therefore not presently suitable for high-throughput analysis. Low complexity regions of sequences are segments that have comparatively little variation in residue content, such as a poly-serine stretch, and these have long been associated with unstructured regions[2]. Some groups also exclude proteins with post-translational modifications or which are thought to have obligate binding partners.

The second stage of family exclusion is removal of proteins that have known structures, can be computationally modeled, or are being studied by other groups. This process involves proteins beyond the realm of interest (Fig. 1*c*), with the goal of determining whether any protein of interest is evolutionally related to one of known structure. Since tertiary structure is better conserved than primary structure, if such a homology can be identified from sequence analysis, it is almost certain that the two proteins will share the same fold. In many cases this homology will allow modeling of the structures. Even if the relationship is too distant to allow effective modeling, once such a homology is found, it is clear that solving the structure will not reveal fundamentally new evolutionary relationships.

To make structural predictions, the proteins in the realm of interest are clustered with other homologous proteins into families using sequence analysis methods including BLAST, PSI-BLAST[3], intermediate sequence searching[4], and hidden Markov models[5,6]. These approaches are imperfect, and so results must be processed by clustering algorithms to define families. This remains a challenge, especially because multidomain proteins may need to simultaneously be in multiple distinct families. To help avoid these issues, many groups make use of well-curated protein family databases such as COGs[7] (*www.ncbi.nlm.nih.gov/COG*) and Pfam[8] (*pfam.wustl.edu*). Some groups also use threading and related methods of predicting structure. Finally, the families are labeled according to whether they contain a known structure; if so, then that family is excluded from the selection process.

Family prioritization

After the exclusion steps, all remaining families are appropriate candidates for study in structural genomics. These families are then prioritized (Fig. 1*d*), although it is important to note that in some sense, prioritization is unnecessary. Since the ultimate goal is completeness, all of the approaches eventually contribute equally. However, prioritization has the advantage of ensuring that coherent and relevant information is made available rapidly,

Department of Plant & Microbial Biology, University of California, Berkeley, 111 Koshland Hall #3102, Berkeley, California 94720-3102, USA.
email: brenner@compbio.berkeley.edu

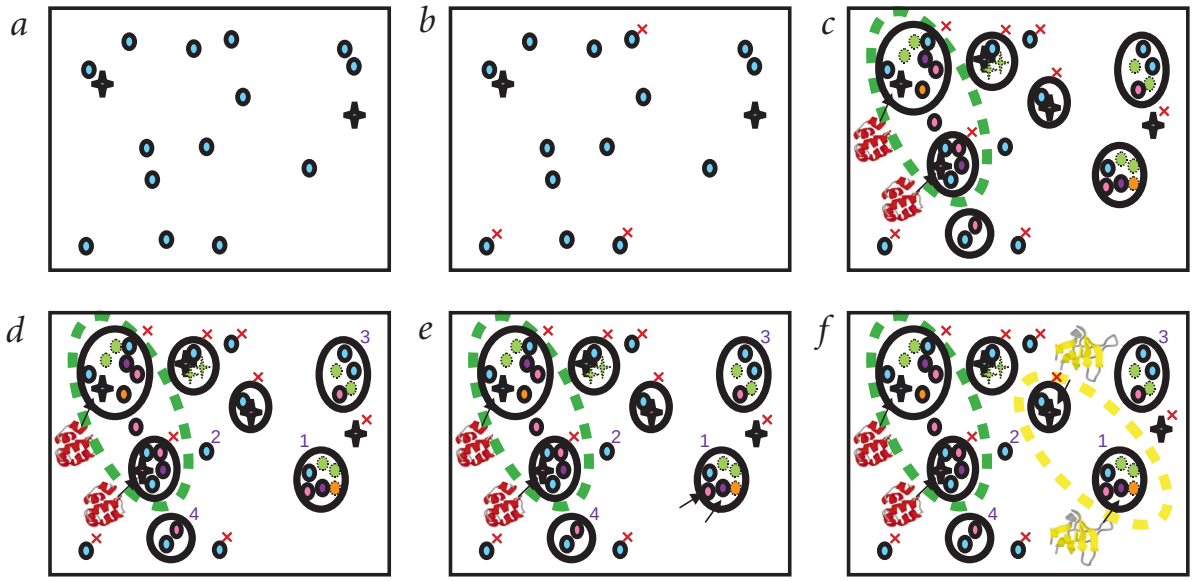


Fig. 1 ***a***, Realm of interest. Proteins within the realm of interest (in this case a genome) are plotted as blue dots in an arbitrary sequence space. Proteins of known structure are shown as stars, others as ovals. ***b***, Family exclusion: non-tractable. Transmembrane proteins and those with low complexity are excluded, as indicated by a red X. ***c***, Family exclusion: known structure. Homologs from other organisms (different colors) are considered and family relationships determined (black ovals). A superfamily, revealed from homologous similar structures among proteins without sequence similarity, is also found (green dotted oval). Families and superfamilies with a member of known structure are excluded, as indicated by a red X. ***d***, Family prioritization. Families are prioritized; in this case, a pervasive taxonomically diverse family is ranked highest and an ORFan is next. ***e***, Protein selection. Two proteins within the highest-priority family are chosen (arrows); note that neither is from the original realm of interest, but are homologous to such a protein. ***f***, Structure analysis and functional interpretation. The solved structure is similar to, and homologous to, another structure previously known. This means all of the proteins in the two families are homologous (indicated by the dotted yellow superfamily oval), and it may therefore be possible to make useful functional inferences.

and it also helps focus energies on posing the questions that we hope the molecular structures will answer.

The selection of realm and the prioritization are intertwined, and each structural genomics group has its own focus that combines the two. A popular prioritization approach is to look for proteins that are taxonomically dispersed — that is, found in many bacteria, archaea, and eukaryotes. Because such proteins are both ancient and conserved, it stands to reason that they represent important biological functions that structures might help explicate. (This selection criterion might subtly imply other criteria; three of the earliest such structures solved were eukaryotic translation initiation factors[9–11], since translation is one of the oldest conserved systems.) The New York structural genomics center and the Berkeley structural genomics consortium augment this consideration with a preference for large families that will permit many other proteins to be structurally modeled. Diametrically opposite to this approach is the cogently argued suggestion of solving structures of 'ORFans' — proteins found solely in a single organism and that lack obvious homology to any other protein[12]. The provenance of the ORFans is one of the great mysteries of evolutionary molecular biology, and structural biology can help discern whether the ORFans are ancient but rapidly evolving, truly new, or not genes at all. An intermediate phylogenetic approach, being pursued by the Midwest Structural Genomics Center, is to focus on those genes in higher eukaryotes, specifically those in *Caenorhabditis elegans* not found in *Saccharomyces cerevisiae*.

Several groups, including a Japanese group, the TB Structural Genomics Consortium, and the Berkeley center aim to be truly genomic. These groups plan to ensure that for nearly all of their selected complete genome (*Pyrococcus horikoshii; Mycobacterium tuberculosis;* and *Mycoplasma genitalium*), every tractable protein will have an experimentally determined or predicted structure. On a smaller scale, looking at a single pathway, the Washington University Structural Genomics Consortium is studying the proteins involved in type IV secretion systems in pathogenic bacteria. Specific putative pathways for structural study might be found from use of context information such as homologous gene fusions, phylogenic profiles, and correlated expression patterns[13].

The Ontario Structural Proteomics Initiative takes a complementary approach. Rather than specifying the proteins of interest, their members are cloning all of the appropriate genes from *Methanobacterium thermoautotrophicum*. However, using what might be thought of as experimental target selection, they will only solve the structures of those proteins that prove easiest to study. This provides a maximal number of structures immediately but leaves gaps in our knowledge and only defers the necessary work on more challenging structures. A variation of this procedure is used by many other centers that will not necessarily study proteins from the genome they are interested in; often they will use more accessible homologs. In many cases, they may use the class-directed approach and study several members of a family in parallel, under the assumption that one will unpredictably prove far more tractable than the others[14].

A pervasive question underlying prioritization approaches is what one hopes to discover. A few groups have suggested looking for proteins with new folds[15,16], as this will help increase the repertoire of known protein shapes. More biologically informative is recognition of homology from structure that was invisible from sequence (Fig. 1*f*). But while homology has significant potential to yield insight into molecular function, it typically does not illuminate the overall cellular role. Thus, the Harvard structural genomics group aims to study proteins implicated in cancer in the hopes of augmenting the phenotypic effect with

molecular knowledge, and the San Diego Consortium will focus on signaling molecules. The German Protein Structure Factory intends to use suggested medical relevance as a key in selecting which cDNA clones it uses as starting points for structure analysis. Target selections for the German group, and also a Japanese group studying mouse cDNAs, the Berkeley Consortium, and the Northeast Structural Genomics Consortium, benefit from coordination with other genomic research efforts which provide complementary functional information.

Protein selection

After families have been prioritized, the final stage of target selection is identification of specific proteins or fragments to be experimentally characterized (Fig. 1*e*). As noted above, often these are not the original proteins of interest, but rather homologs that possess desirable characteristics such as small size, thermostability, appropriate pI, and methionine counts suitable for MAD phasing. Class-directed efforts may involve working not only on multiple members of a given family, but also on different protein fragments, constructs, and expression systems. At present, the selection of appropriate fragments is largely an intuitive and imprecise matter of trying to infer domain boundaries from multiple alignment data followed by careful experimentation. It is hoped that as large-scale efforts proceed, data harvesting will enable more successful quantitative approaches for selecting effective target proteins, constructs, and expression systems.

Collaborative approaches

Given the variety of prioritization criteria applied by different groups, it would seem that their selected targets should be quite distinct. However, among the very first 'structural genomics' structures solved were two groups' independent studies of homologous proteins related to eukaryotic translation initiation factor 5A (refs 10,11). With this early overlap as a warning, it is clear that as structural genomics programs scale up, coordination is necessary to avoid wholesale duplication of effort. In recognition of this, most structural genomics groups have been unusually open about the specific targets they have selected and their progress in cloning, expressing, purifying, concentrating or crystallizing, and structurally determining these proteins. In addition, two resources have been developed for community submissions. The PRESAGE database for structural genomics (*http://presage.berkeley.edu*) has nearly 400 targets from six major groups and several minor ones, with more submissions promised once the US National Institute of General Medical Sciences (NIGMS) centers begin operation and international groups ramp up[17]. The database at *www.structuralgenomics.org* has additional targets from at least two groups. In addition to allowing structural genomics researchers to avoid inadvertent duplication, these resources also serve as a convenient overview indicating progress of international structural genomics projects.

A separate incident also involving an early structural genomics experiment underscored another of the problems with standard target selection. In this instance, a solved protein's structure was not predicted by standard computational methods, but careful manual analysis in a different group had previously allowed prediction of its structure. Although the earlier structure prediction was published, correct, and clearly stated[18], it did not preclude the ensuing experiment, because neither the title nor abstract of the paper effectively suggested that it might contain this prediction. In essence, the prediction was inaccessible. The possibilities for unnecessary experimentation on correctly predicted structures has grown, as more than a dozen groups have attempted to make structural predictions of at least one complete genome. Therefore, the PRESAGE database has thousands of these predictions, both to aid structural genomics target selection as well as to make these data available to biologists interested in making use of structural information.

In addition to sharing information about targets and predictions, it would seem natural to also share methods for the common aspects of target selection, rather than having independent implementations of largely similar software. Already this happens in an indirect manner, with several groups making use of databases like Pfam[8], COGs[19], and BioSpace[20] (*biospace.stanford.edu*); already Pfam and BioSpace explicitly incorporate structure information, and the COGs database will be adding careful structure predictions soon. Beyond this, several groups have discussed sharing tools, and *www.structuralgenomics.org* has a web interface allowing one to select targets meeting certain criteria. Groups at The Rockefeller University have proposed developing target selection resources explicitly intended for distribution to the community.

As structural genomics evolves, with projects growing larger and the number of unexplored families diminishing, it is likely that efforts will become increasingly coordinated. At some point, the sharing of data and resources may yield to a measure of centralization of target selection. Until that time, the varied but coordinated approaches to target selection will allow several different swaths of molecular biology to be structurally explored. As structural genomics of family representatives becomes complete, the structural information it has revealed and the resources it develops are likely to motivate new high-throughput approaches to structural biology with even more diverse approaches for selection of targets.

Acknowledgments

I am grateful to the many groups who, in the collaborative spirit of structural genomics, shared information about their projects and target selection methodologies. Supported by grants from the NIH and NSF.

Associations with structural genomics

S.E.B. is a member of the Berkeley Structural Genomics Consortium and is an author of the PRESAGE database.

1. Brenner, S.E. & Levitt M. *Protein Sci.* **9**, 197–200 (2000).
2. Wootton, J.C. & Federhen S. *Methods Enzymol.* **266**, 554–571 (1996).
3. Altschul, S.F. *et al. Nucleic Acids Res.* **25**, 3389–3402 (1997).
4. Teichmann, S.A., Chothia, C., Church, G.M. & Park J. *Bioinformatics* **16**, 117–124 (2000).
5. Eddy, S.R. *Bioinformatics* **14**, 755–763 (1998).
6. Krogh, A., Brown, M., Mian, I.S., Sjolander, K. & Haussler, D. *J. Mol. Biol.* **235**, 1501–1531 (1994).
7. Tatusov, R.L., Galperin, M.Y., Natale, D.A. & Koonin, E.V. *Nucleic Acids Res.* **28**, 33–36 (2000).
8. Bateman, A. *et al. Nucleic Acids Res.* **28**, 263–266 (2000).
9. Cort, J.R., Koonin, E.V., Bash, P.A. & Kennedy, M.A. *Nucleic Acids Res* **27**, 4018–4027 (1999).
10. Kim, K.K., Hung, L.W., Yokota, H., Kim, R. & Kim, S.H. *Proc. Natl. Acad. Sci. USA* **95**, 10419–10424 (1998).
11. Peat, T.S., Newman, J., Waldo, G.S., Berendzen, J. & Terwilliger, T.C. *Structure* **6**, 1207–1214 (1998).
12. Fischer, D. *Protein Eng.* **12**,1029–1030 (1999).
13. Huynen, M, Snel, B, Lathe, W & Bork, P. *Curr. Opin. Struct. Biol.* **10**, 366–370 (2000).
14. Terwilliger, T.C. *et al. Protein Sci.* **7**, 1851–1856 (1998).
15. Mallick, P., Goodwill, K.E., Fitz-Gibbon, S., Miller, J.H. & Eisenberg, D. *Proc. Natl. Acad. Sci. USA* **97**, 2450–2455 (2000).
16. Portugaly, E. & Linial, M. *Proc. Natl. Acad. Sci. USA* **97**, 5161–5166 (2000).
17. Brenner, S.E., Barken, D. & Levitt, M. *Nucleic Acids Res.* **27**, 251–253 (1999).
18. Koonin, E.V., Tatusov, R.L. & Rudd, K.E. *Proc. Natl. Acad. Sci. USA* **92**, 11921–11925 (1995).
19. Tatusov, R.L., Koonin, E.V. & Lipman, D.J. *Science* **278**, 631–637 (1997).
20. Yona, G. & Levitt, M. *Proc. of Intelligent Systems for Mol. Biol.* **8**, 395–406 (2000).