Structural genomics and its importance for gene function analysis

Jeffrey Skolnick[1]*, Jacquelyn S. Fetrow[2], and Andrzej Kolinski[1,3]

*[1]Laboratory of Computational Genomics, The Danforth Plant Science Center, 893 N, Warson Rd., St. Louis, MO 63141. [2]Geneformatics, Inc., 5830 Oberlin Drive, Suite 200, San Diego, CA 92121-3754. [3]Department of Chemistry, University of Warsaw, Pasteura 1, 02093 Warsaw, Poland. *Corresponding author (skolnick@danforthcenter.org).*

Received 8 October 1999; accepted 24 January 2000

Structural genomics projects aim to solve the experimental structures of all possible protein folds. Such projects entail a conceptual shift from traditional structural biology in which structural information is obtained on known proteins to one in which the structure of a protein is determined first and the function assigned only later. Whereas the goal of converting protein structure into function can be accomplished by traditional sequence motif-based approaches, recent studies have shown that assignment of a protein's biochemical function can also be achieved by scanning its structure for a match to the geometry and chemical identity of a known active site. Importantly, this approach can use low-resolution structures provided by contemporary structure prediction methods. When applied to genomes, structural information (either experimental or predicted) is likely to play an important role in high-throughput function assignment.

Keywords: Protein folding, threading, protein function prediction, structural genomics, functional motifs

Sequencing of the genomes of numerous organisms has forever changed the face of biology[1,2]. Currently, emphasis lies not on the study of individual molecules, but on the large-scale, high-throughput examination of the genes and gene products of an organism, with the aim of assigning their functions[1–3]. Sequence genomics, the determination of the entire DNA sequence of the organism of interest, has been incredibly successful at providing the raw material for this process[4]. But knowledge of the sequences of the genes and gene products alone does not provide insight into what each molecule does in the cell. High-throughput screening techniques and biological assays[5] will certainly provide some of this knowledge, but further insight comes when structural information is available[5]. The first age of genomics is the sequence era, but now structural biology is poised to play an important role.

In this review, we examine the nature and scope of experimental structural genomics, and then describe the theoretical approaches involved (i.e., tools that can predict protein structure and guide target selection). We also discuss the techniques for transforming protein structural information into functional information, and the biological implications of the use of structural information.

How many novel folds can proteins adopt?

Over the years, there have been numerous estimates as to how many novel folds there might be, with values ranging from 1,000 to 10,000–100,000[6–9]. A key issue is whether protein structure space is discrete, with a finite number of quantifiable folds[6], or continuous, where protein structures can morph one into the other. Most of our knowledge about protein structure comes from the subset of proteins that are water-soluble and that crystallize easily[10,11]. How representative these proteins are of the entire universe of protein structures is unclear. Furthermore, the clustering of folds is dependent on the metric of similarity used to compare structures[12–14]. This becomes especially important when one is classifying structures that share only a subset of global features.

One purely operational criterion for the requisite number of structures is that any protein sequence be within homology modeling distance of a known protein structure[15]. How close the final structure must be depends on the desired use of the model. At present, if one wishes to identify small molecules that bind to a protein of interest[16], the structure of the template protein must be quite close to that adopted by the sequence of interest. Alternatively, if identification of biochemical activity is the only goal, the structures need not be so close[17–20]. Thus, the determination of the number of distinct folds hinges on how the structure will be used. However, as modeling methods continue to improve, the mesh of protein structure space that needs to be sampled can become coarser. As we describe below, the determination of protein function from structure can be accomplished by a variety of means (see Fig. 1).

Experimental approaches

In the past, the function of a protein of interest was first identified and then its structure determined by means of time-consuming x-ray or NMR experiments. In contrast, structural genomics aims to first determine the structure of proteins, and then investigate their function later (if at all)[21–24]. As high throughput is an absolute necessity, one has to ask whether it makes sense to continue refining a given structure or to move on and solve a different structure, with the raw data of the partially unrefined model stored for a later day. The answer depends on the ratio of the time required for refinement versus that of new structure generation. Perhaps certain regions, such as guessed or known active sites, might be better refined, whereas other less important regions might be left to be of poorer quality.

If the goal of structural genomics is to identify novel folds, one has to eliminate known folds from contention. This can be accomplished by simply eliminating sequences homologous to proteins for which the structures are already known. Once such a list, which may number in the hundreds if not thousands[25], has been compiled, which sequences should be prioritized for investigation?

One strategy is entirely opportunistic: select only those proteins that express well (often a rate-limiting step) and that crystallize and diffract well. Suitable candidates then undergo crystallographic

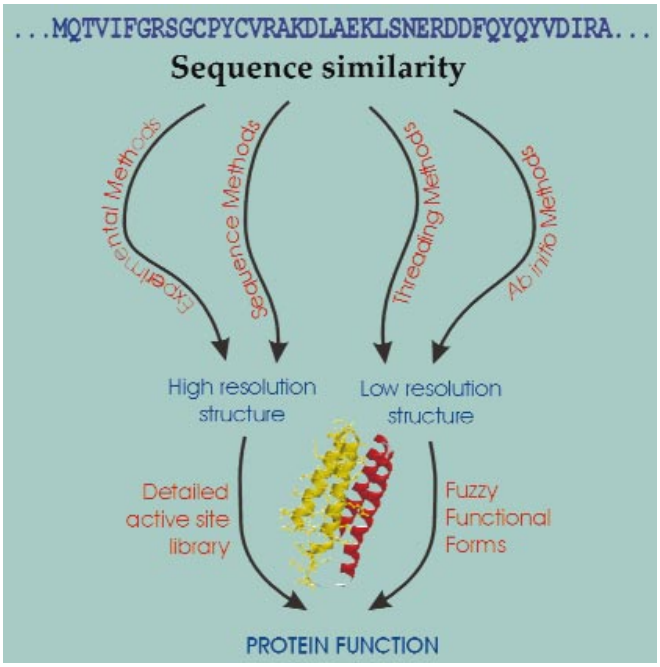

Figure 1. Schematic illustration of the various methods of determining protein structure using the sequence→structure→function paradigm for an entire genome.

study, the diffraction data are collected, and the structure is solved[26]. NMR will be especially helpful for determining the structure of those proteins that express well, but do not crystallize[22]. However, such proteins will still need to be soluble at concentrations required by NMR spectroscopic techniques.

A key problem will be determining the structure of membrane proteins, which may be present in roughly 20–30% of genomic sequences[27]. Since, with current technology, given individuals can perhaps determine at most five to ten structures per year, the strategy for target selection remains absolutely crucial if the yield of novel folds is to be enhanced.

Theoretical approaches

Various approaches to massive structure prediction employ different techniques and assume different levels of generality. Certainly, the most straightforward strategy for predicting structure is to use standard bioinformatics and molecular modeling tools in a semiautomatic way in order to screen massive amounts of genomic data[25]. This usually consists of three steps: first, search for sequence similarity to a member of a set of carefully selected sequences with known three-dimensional structure; second, use the detected structural template to build a molecular model; and third, carefully validate the resulting models.

Using standard sequence alignment algorithms and MODELLER—a comparative modeling engine that allows automatic homology modeling and model evaluation—Sanchez and Sali[25,28] recently have scanned a portion of the genome of the yeast *Saccharomyces cerevisiae*. They found homologous proteins of known structure for about 17% of the proteins (1,071 sequences) and built three-dimensional models for these yeast proteins. Only 40 of these modeled proteins had previously determined experimental structure, and 236 proteins were shown to be related to a protein of known structure for the first time.

An obvious limitation of the above approach is that it requires a homologous protein of known structure. Depending on the genome, 15–25% of all sequences now have a homologous protein of known structure[29]. This percentage is slowly increasing as new structures are being solved at an increasing rate. Interestingly, the majority of solved structures exhibit an already known fold. At this point, it is still uncertain whether this indicates that proteins can adopt a limited number of folds or if it simply indicates a bias toward proteins that crystallize.

Beyond sequence-based approaches

Threading methods[30,31] go a bit further than pure sequence-based alignment methods. They allow one to search for sequences that have a similar fold without apparent sequence similarity. They may also be useful for identifying distantly related pairs of proteins, and thereby increase the fraction of proteins for which an already known fold can be assigned. The recent CASP3 prediction experiment results suggest that some progress has been made in the prediction of medium-difficulty targets, but not much progress has been made on targets of greater difficulty[32]. Nevertheless, as demonstrated by Fischer and Eisenberg[33], such a sequence–structure matching approach noticeably increases the fraction of annotated proteins. For the *Mycoplasma genitalium* genome, the folds of 103 out of a total of 468 proteins were assigned by their threading algorithm, whereas traditional sequence methods identified only 75 proteins as being similar to known structures.

Whereas threading methods can sometimes recognize remotely related proteins, the corresponding structures often differ substantially. In particular, the alignment of the probe sequence in the template structure is often very poor. Current model-building algorithms do not address these problems[33]. Recently, tools have been developed that can refine models with an initial backbone root mean square deviation (r.m.s.d.) from native that is in the range of 8–10 Å, to structures with an r.m.s.d. of 4–6 Å[34]. Nevertheless, a key question is whether or not such inexact models can provide insight into protein function. As discussed here, this is indeed the case.

Whereas the ability of threading (along with very sensitive sequence-based techniques) to identify proteins adopting known folds may be quite valuable in other contexts (see below), it is merely a strategy for eliminating candidates that have a novel fold, rather than one for identifying possible novel structures.

One way to identify novel folds is to employ ab initio approaches to protein structure prediction. The recent CASP3 evaluation of protein structure prediction methods indicates that progress in ab initio methods is very rapid and allows meaningful predictions[35–40]. Interestingly, over the range of comparable applicability (small proteins), results to date indicate that the quality of ab initio models is at least as good as those determined from threading. Furthermore, several ab initio folding groups have succeeded in identifying novel or near novel folds[40,41]. Although additional progress in ab initio folding is required (the ability to treat β-proteins in particular, and larger proteins in general, is lacking[40,42]), at least for small proteins, ab initio prediction is being employed to help identity possible novel folds, thereby assisting in target selection of the emerging structural genomics initiatives.

Determining function from structural information

Once the structure of a protein has been resolved, how does one determine its function? After all, functional analysis of gene products is a major goal of both the sequence and structural genomics projects[43]. With only sequence and no structure, researchers usually rely on sequence analysis, a method based on the underlying evolutionary relationships between the two sequences[44,45]. However, the inappropriate assignment of function between two proteins with significant sequence similarity has led to a number of errors in the annotation of genome sequences[46].

With the aim of extracting further information from protein sequences, sequence motif libraries have been developed[47–51]. One approach to creating more specific motifs is to use structural information. For instance, conserved sequence patterns can be com-

Structural genomics resources

At present, several pilot structural genomics projects are underway (see Table 1). As a proof of principle, Kim and coworkers[58] have solved the crystal structure of *Methanococcus jannaschii* MJ0577 protein, for which the function was previously unknown. The structure contains a bound ATP, suggesting MJ0577 is an ATPase or an ATP-mediated molecular switch; this was subsequently confirmed by biochemical experiments[58]. Importantly, efforts are also underway to minimize a duplication of efforts among the various structural genomics groups. For example, a very useful database, PRESAGE, has been assembled by Brenner and coworkers[59] that provides a collection of annotations reflecting current experimental status, structural assignments, models, and suggestions. Another similar resource is provided by the Protein Structure Initiative (http://www.structuralgenomics.org/).

Table 1. URLs for structural genomics pilot projects, computational tools, and key databases.

Resource	Description	URL
Projects		
Center for Advanced Research in Biotechnology (Rockville, MD) and the Institute for Genomic Research (Rockville, MD)	Solve structures of unknown function in *Haemophilus influenzae*	http://structuralgenomics.org/
Brookhaven National Laboratory (Upton, NY), Rockefeller University (New York, NY), and Albert Einstein School of Medicine (New York, NY)	Pilot genomics project on yeast	http://proteome.bnl.gov/targets.html.
New Jersey Commission on Science and Technology and Rutgers University (Piscataway, NJ)	Metazoan organisms, human pathogen proteins	http://www-nmr.cabm.rutgers.edu/structuralgenomics/concept.html
Los Alamos National Laboratory and the University of California, Los Angeles	Thermophilic archeon *Pyrobaculum aerophilum*	http://www-structure.llnl.gov/PA/PA_intro.html
Argonne National Laboratory (Argonne, IL)	Technology for high-throughput structure determination	http://www.bio.anl.gov/research/structural_genomics.htm
PRESAGE	Structural genomics clearing house; coordination of efforts	http://presage.Stanford.edu/
Protein structure initiative	Structural genomics clearing house	http://structuralgenomics.org/
Tools		
Eisenberg group	Threading tools	http://www.doe-mbi.ucla.edu/People Eisenberg/Projects/
Expasy	Swiss-Prot site contains many sequence and structure searching tools	http://www.expasy.ch/
Gerstein group	Structure prediction of eight genomes comparative genomics	http://bioinfo.mbb.yale.edu/genome/
National Center for Biotechnology Information (Bethesda, MD)	BLAST sequence similarity search tool	http://www.ncbi.nlm.nih.gov/BLAST/
Sali group	Tools for protein structure modeling, including MODELLER	http://guitar.rockefeller.edu/subpages/programs.html
Skolnick–Kolinski group	Threading tools, *ab initio* folding tools, FFF library	http://bioinformatics.danforthcenter.org
Thornton group	Library of three-dimensional active site motifs	http://www.biochem.ucl.ac.uk/bsm/PROCAT/PROCAT.html
Databases		
Protein Data Bank	Database of solved protein structures	http://nist.rcsb.org/pdb/
Expasy	Swiss-Prot protein sequence and structure database	http://www.expasy.ch/
CATH	Protein structure classification database	http://www.biochem.ucl.ac.uk/bsm/cath/
SCOP	Murzin's database of protein structure classification	http://scop.mrc-lmb.cam.ac.uk.scop/

bined with information about the structural context[52]. Although addition of such information increases the specificity of motif searching methods, the method is still only applied to the one-dimensional sequences; overall three-dimensional structure of the unknown sequence is not required. Such methods thus simply extend sequence analysis; they do not constitute a true "structural genomics" approach in that they do not use structural information directly. However, at present, this is the most common way biologists use structural information, and sequence-based approaches certainly form the standard against which all other techniques must compete.

Because sequence analysis can only go so far, the next obvious step is the use of three-dimensional structural information; this may explain the onslaught of the structural genomics projects. In a series of major steps in structural classification, several databases of protein structures, organized by common folding arrangements, have been created (see Table 1; refs 12–14).

A recent analysis of protein sequences and structures demonstrates that functional information can be automatically derived from structural information only to a limited extent[53]. In this analysis, two functions were associated with seven folds each. Conversely, some folds can exhibit as many as 16 functions[54]. Obviously, structural information can aid in the detection of errors, can in some cases provide general functional information, and can augment any functional information provided by sequence analysis. However, knowledge of the overall structure or domain family is still not enough to confidently assign function, especially at a detailed biochemical level.

Atomic-resolution three-dimensional motifs require high-resolution structures. Clearly, the development of structural motifs for specific functional sites can aid in the identification of functional sites in structures. Additional analysis of common residue clusters or characteristic surface properties is necessary. Toward this end, Thornton and colleagues (http://www.biochem-.ucl.ac.uk/bsm/PROCAT/PROCAT.html) have developed a library of three-dimensional motifs that specify the relative positions in space of certain atoms involved in functional sites. To validate the method, they have developed a motif for the serine proteases[55]. This motif was able to identify all known serine proteases and triacylglycerol lipases in a set of protein structures. Furthermore, the structural motif identified two "non-esterase" triads, the significance of which is unclear. These types of structural motifs are essential for making the best use of high-resolution atomic structures. However, because the motifs require detailed knowledge of the positions of specific side chain atoms, they are useless in the analysis of predicted models. Predicted models rarely contain such detailed, high-quality information.

Toward functional information from modeling

As discussed, it takes considerable time and financial resources to accomplish high-throughput, experimental structure determination. This begs the question of what, if anything, can be learned from lower quality models of protein structures produced by the threading or ab initio methods described here. As a whole, in the field of structural biology, there has been a general belief that the lower quality models produced by protein structure prediction algorithms are not useful for functional analysis. However, recent work suggests that this is not the case[17–20].

Results from several groups suggest that active sites in proteins are well preserved and may be excised from crystal structures in order to describe the active site geometry at a level suitable for both low- and high-resolution models[25,53]. Toward this goal, we have begun to develop a structural motif library, populated by structural motifs called fuzzy functional forms (FFFs)[17]. These FFFs have been applied to both high-quality structures and to inexact models created from both ab initio and threading approaches. The resulting functional site analysis of these models is quite accurate[17,18,20].

While the approach needs to be developed further, such an automated structure-based analysis could enhance the sequence-based analysis of genome databases. It provides a more detailed and specific functional analysis of these genome sequences, largely because it combines the advantages of both sequence- and structure-based information. Most importantly, such structure-based approaches exhibit fewer false positive results than sequence-based approaches (N. Siew, J. Skolnick & J. Fetrow, unpublished data). For example, the sequences in eight genomes were analyzed for disulfide oxidoreductase function using the disulfide oxidoreductase FFF, the thioredoxin Block 00194, and the glutaredoxin Block 00195 obtained from the Blocks web site for sequence motifs[49]. Assuming that those sequences identified by both the FFF and Blocks are "true positives," by way of example, 13 such sequences exist in the *Bacillus subtilis* genome. (It should be stated that experimental evidence validating all of these "true positives" is lacking; thus, they are more correctly termed "consensus positives.") To find these 13 "consensus positive" sequences, the FFF hits 7 false positives. On the other hand, blocks hits 23 false positives. These data, and others, while very encouraging, do not yet establish the general applicability of the method and represent a work in progress.

A step beyond evolutionary relationships

Structural analysis of specific functional sites in proteins takes the researcher a step beyond the limitations of orthologous and paralogous evolutionary relationships, because functional sites can be re-created and re-used in different protein folds and families. The best-known example of this observation is the active site geometry of the eukaryotic and bacterial serine proteases; the protease active site in these two proteins is quite similar, although the overall protein structures of each family are quite different[56].

In another case, an RNA binding site was recognized in the structure of a viral protease that exhibited a trypsin-like protein fold[57]. In a more recent example, we have predicted a redox regulatory site very similar to the active sites of the glutaredoxin/thioredoxin oxidoreductase family, in the fold of the serine-threonine phosphatase-1 subfamily[19]. The crystal structure of one member of this family was solved in 1995, but the location of the putative redox regulatory site was not identified until 1999, a result that emphasizes the need to develop methods for automatic, but biologically relevant, functional site analysis as part of the structural genomics initiatives.

Conclusions

The key question is what insight, if any, into biology can structural genomics provide? We believe that this is not just "postage stamp" collecting on a genomic scale. If one knew all protein folds, the protein folding problem would be solved by brute force. Furthermore, structural genomics will increase our understanding of the design principles of proteins and may have applications to protein engineering.

However, the greatest payoff for biology will come from coupling the resulting structural information with biochemical functional information. If functional site libraries of all protein biochemical functions are built, then it would become possible to carry out functional threading: given a library of known functions, one could search the protein structure for a constellation of residues that matches a known active site. If a similar analysis were applied to known binding regions, then having a structure would go a long way to providing new insight into its function. Moreover, having structure would allow one to deal, in part, with the multilevel aspect of protein function. For example, proteins can add additional functions during evolution, so that even knowing the primordial function of a protein may not permit a full characterization of its charac-

teristics. Screening against a binding region/active site library can greatly assist in this process. Having the structure will enable researchers to engage in high-throughput inhibitor design. Thus, structure, whether predicted or experimentally determined, will play a very important role in high-throughput, biologically relevant function prediction.

1. Clark, M.S. Comparative genomics: the key to understanding the Human Genome Project. *Bioessays* **21**, 121–130 (1999).
2. DellaPenna, D. Nutritional genomics: manipulating plant micronutrients to improve human health. *Science* **285**, 375–379 (1999).
3. Wiley, S.R. Genomics in the real world. *Curr. Pharm. Des.* **4**, 417–422 (1998).
4. Lin, J. et al. Whole-genome shotgun optical mapping of *Deinococcus radiodurans*. *Science* **285**, 1558–1562 (1999).
5. Carulli, J. P. et al. High throughput analysis of differential gene expression. *J. Cell Biochem. Suppl.* **31**, 286–96 (1998).
6. Chothia, C. & Finkelstein, A. The classification and origins of protein folding patterns. *Annu. Rev. Biochem.* **59**, 1007–1039 (1990).
7. Murzin, A.G., Lesk, A.M. & Chothia, C. Principles determining the structure of beta-sheet barrels in proteins. II. The observed structures. *J. Mol. Biol.* **236**, 1382–1400 (1994).
8. Chothia, C., Hubbard, T., Brenner, S., Barns, H. & Murzin, A. Protein folds in the all-beta and all-alpha classes. *Annu. Rev. Biophys. Biomol. Struct.* **26**, 597–627 (1997).
9. Sali, A. 100,000 protein structures for the biologist (see comments). *Nat. Struct. Biol.* **5**, 1029–1032 (1998).
10. Holm, L. & Sander, C. Protein folds and families: sequence and structure alignments. *Nucleic Acids Res.* **27**, 244–247 (1999).
11. Dodge, C., Schneider, R. & Sander, C. The HSSP database of protein structure-sequence alignments and family profiles. *Nucleic Acids Res.* 26, 313–315 (1998).
12. Holm, L. & Sander, C. Dali/FSSP classification of three-dimensional folds. *Nucleic Acids Res.* **25**, 231–234 (1997).
13. Orengo, C.A. et al. CATH—a hierarchic classification of protein domain structures. *Structure* **5**, 1093–1108 (1997).
14. Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. Scop: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).
15. Sanchez, R. & Sali, A. Evaluation of comparative protein structure modeling by MODELLER-3. *Proteins* Suppl. 50–58 (1997).
16. Briem, H. & Kuntz, I. D. Molecular similarity based on DOCK-generated fingerprints. *J. Med. Chem.* **39**, 3401–3408 (1996).
17. Fetrow, J.S. & Skolnick, J. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/ thioredoxins and T1 ribonucleases. *J. Mol. Biol.* **281**, 949–968 (1998).
18. Fetrow, J.S., Godzik, A. & Skolnick, J. Functional analysis of the *Escherichia coli* genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *J. Mol. Biol.* **282**, 703–711 (1998).
19. Fetrow, J.S., Siew, N. & Skolnick, J. Structure-based functional motif identifies a potential disulfide oxidoreductase active site in the serine/threonine protein phosphatase-1 subfamily. *FASEB J.* **13**, 1866–1874(1999).
20. Zhang, L., Godzik, A., Skolnick, J. & Fetrow, J.S. Functional analysis of *E. coli* proteins for members of the a/b hydrolase family. *Folding and Design* 3, 535–548 (1998).
21. Orengo, C.A., Todd, A.E. & Thornton, J.M. From protein structure to function. *Curr. Opin. Struct. Biol.* **9**, 374–382 (1999).
22. Montelione, G.T. & Anderson, S. Structural genomics: keystone for a Human Proteome Project (news). *Nat. Struct. Biol.* **6**, 11–12 (1999).
23. Kim, S.H. Shining a light on structural genomics. *Nat. Struct. Biol.* **5** Suppl, 643–645 (1998).
24. Gaasterland, T. Structural genomics taking shape. *Trends Genet.* **14**, 135 (1998).
25. Sanchez, R. & Sali, A. Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl. Acad. Sci. USA* **95**, 13597–13602 (1998).
26. Terwilliger, T.C. & Berendzen, J. Automated MAD and MIR structure solution. *Acta Crystallogr. D* **55**, 849–861 (1999).
27. Wallin, E. & Heijne, G.V. Genome-wide analysis of intergral membrane proteins from eubacterial, archaen, and eukaryotic organismc. *Prot. Sci.* **7**, 1029–1038 (1998).
28. Goffeau, A. et al. Life with 6000 genes (see comments). *Science* **274**, 546, 563–567 (1996).
29. Elofsson, A. & Sonnhammer, E.L. A comparison of sequence and structure protein domain families as a basis for structural genomics. *Bioinformatics* **15**, 480–500 (1999).
30. Rost, B., Schneider, R. & Sander, C. Protein fold recognition by prediction-based threading. *J. Mol. Biol.* **270**, 471–480 (1997).
31. Jones, D.T. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* **287**, 797–815 (1999).
32. Marchler-Bauer, A. & Brenner, S. Comparison of prediction quality in the three CASPs. *Proteins Suppl.* **3**, 218–225 (1999).
33. Fischer, D. & Eisenberg, D. Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*. *Proc. Natl. Acad Sci USA* **94**, 11929–11934 (1997).
34. Kolinski, A., Rotkiewicz, P., Ilkowski, I. & Skolnick, J. A method for the improvement of threading based protein models. *Proteins* **37**, 592–610 (1999).
35. Lee, J., Liwo, A., Ripoll, D.R., Pillardy, J. & Scheraga, H.A. Calculation of protein conformation by global optimiation of a potential energy function. *Proteins Suppl.* **3**, 204–208 (1999).
36. Simons, K.T., Bonneau, R., Ruczinski, I. & Baker, D. Ab initio structure prediction of CASP III targets using ROSETTA. *Proteins Suppl.* **3**, 171–176 (1999).
37. Ortiz, A., Kolinski, A., Rotkiewicz, P., Ilkowski, B. & Skolnick, J. Ab intio folding of proteins using restraints derived from evolutionary information. *Proteins Suppl.* **3**, 177–185 (1999).
38. Osguthorpe, D.J. Improved ab initio predictions with a simplified, flexible geometry model. *Proteins Suppl.* **3**, 186–193 (1999).
39. Samudrala, R., Xia, Y., Huang, E. & Levitt, M. Ab initio protein structure prediction using a combined hierarchical approach. *Proteins Suppl* **3**, 194–198 (1999).
40. Orengo, C., Bray, J.E., LoConte, L. & Sillitoe, I. Analysis and assessment of ab initio three-dimensional prediction, secondary structure and contacts prediction. *Proteins Suppl.* **3**, 149–170 (1999).
41. Murzin, A. Structure classification-based assessement of CASP3 prediction for the fold recognition targets. *Proteins Suppl.* **3**, 88–103 (1999).
42. Venclovas, C., Zemla, A., Fidelis, K. & Moult, J. Some measures of comparative performance in the three CASPs. *Proteins Suppl.* **3**, 231–227 (1999).
43. Brutlag, D.L. Genomics and computational molecular biology. *Curr. Opin. Microbiol.* **1**, 340–345 (1998).
44. Altschul, S.F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
45. Pearson, W.R. Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* **276**, 71–84 (1998).
46. Brenner, S.E. Errors in genome annotation. *Trends Genet.* **15**, 132–133 (1999).
47. Attwood, T.K. et al. Novel developments with the PRINTS protein fingerprint database. *Nucleic Acids Res.* **25**, 212–216 (1997).
48. Bairoch, A. Prosite: a dictionary of sites and patterns in proteins. *Nucleic Acids Res.* Suppl, **19**, 2241–2245 (1991).
49. Henikoff, J.G., Henikoff, S. & Pietrokovski, S. New features of the Blocks database servers. *Nucleic Acids Res.* **27**, 226–228 (1999).
50. Hofmann, K., Bucher, P., Falquet, L. & Bairoch, A. The Prosite database, its status in 1999. *Nucleic Acids Res.* **27**, 215–219 (1999).
51. Pietrovski, S., Henikoff, J.G. & Henikoff, S. The Blocks database—a system for protein classification. *Nucleic Acids Res.* **24**, 197–200 (1996).
52. Yu, L., White, J.V. & Smith, T.F. A homology identification method that combines protein sequence and structure information. *Protein Sci.* **7**, 2499–2510 (1998).
53. Kasuya, A. & Thornton, J.M. Three-dimensional structure analyis of Prosite patterns. *J. Mol. Biol.* **286**, 1673–1691 (1999).
54. Hegyi, H. & Gerstein, M. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288**, 147–164 (1999).
55. Wallace, A.C., Laskowski, R.A. & Thornton, J.M. Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci.* **5**, 1001–1013 (1996).
56. Fischer, D., Wolfson, H., Lin, S.L. & Nussinov, R. Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similarities: potential implications to evolution and to protein folding. *Protein Sci.* **3**, 769–778 (1994).
57. Matthews, D.A. et al. Structure of human rhinovirus 3C protease reveals a trypsin-like polypeptide fold, RNA-binding site, and means for cleaving precursor polyprotein. *Cell* **77**, 761–771 (1994).
58. Zarembinski, T.I. et al. Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. *Proc. Natl. Acad. Sci. USA* **95**, 15189–15193 (1998).
59. Brenner, S.E., Barken, D. & Levitt, M. The PRESAGE database for structural genomics. *Nucleic Acids Res.* **27**, 251–253 (1999).