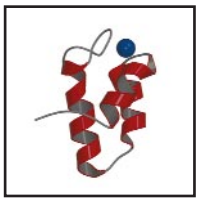
Protein structure modeling for structural genomics

Roberto Sánchez, Ursula Pieper, Francisco Melo, Narayanan Eswar, Marc A. Martí-Renom, M.S. Madhusudhan, Nebojša Mirković and Andrej Šali

The shapes of most protein sequences will be modeled based on their similarity to experimentally determined protein structures. The current role, limitations, challenges and prospects for protein structure modeling (using information about genes and genomes) are discussed in the context of structural genomics.



Evolution has produced families of proteins whose members share the same three-dimensional architecture and frequently have detectably similar sequences. This conservation allows a structural description of all proteins in a family even when only the structure of a single member is known. Evolution also provides the rationale for structural genomics, a systematic and large-scale effort towards structural characterization of all proteins[1–3]. Structural genomics will achieve its aim by focusing the techniques of X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy on proteins predicted to have sufficiently novel structures. The remaining protein sequences will then be modeled based on one or more of the defined structures. Thus, the new structures will put most protein sequences within a 'modeling distance' of at least one known structure while minimizing the total cost of the project. The number of modeled sequences will continue to be at least two orders of magnitude larger than the number of the experimentally determined protein structures (Fig. 1, Table 1).

The ultimate aim of structural genomics is not to obtain the structures or models for all proteins, but to contribute to biology and medicine through functional annotation[4–7], and through applications of protein structures such as virtual drug screening[8]. Proteins will be annotated by structural genomics based on evolutionary homology as well as information about their structures alone. Structural genomics will frequently establish homology from structure and then infer function from homology. Such structure-based transfer of functional information is preferred over the sequence-based extrapolations because (i) similarity in structure is generally more recognizable than similarity in sequence and (ii) because structure frequently allows a more judicious and informative transfer of functional description than sequence alone. In addition to improving homology-based arguments, structural genomics will contribute to functional annotation of proteins by allowing the use of methods that depend only on the structure of the protein to be characterized, such as the matching of three-dimensional patterns[5] and explicit docking of ligands[8]. The first step in most structure-based annotations will be calculation of a three-dimensional model, although there are of course trivial cases where modeling is not needed and difficult cases where modeling cannot yet be helpful. Given the needs of functional annotation, structural genomics should ideally result in all-atom models with few significant atomic errors larger than ~3 Å. This condition implies that the protein structure modeling method of choice for structural genomics is homology-based or comparative modeling[9,10] because it is the most detailed and accurate of all current protein structure prediction techniques. Comparative modeling can produce a model for a protein sequence if it is recognizably related to at least one known protein structure. Comparative modeling involves fold assignment, sequence-structure alignment, model building, and model evaluation (Box 1).

There are two additional approaches to modeling of protein structures: fold assignment or threading[11] and *ab initio* protein structure prediction[12] (Box 1). The fold assignment methods assign a fold to the target sequence by aligning it with the most compatible known protein structure from a set of alternatives. As such, the fold assignment methods are best seen as the first, and in many cases the most important, step in comparative protein

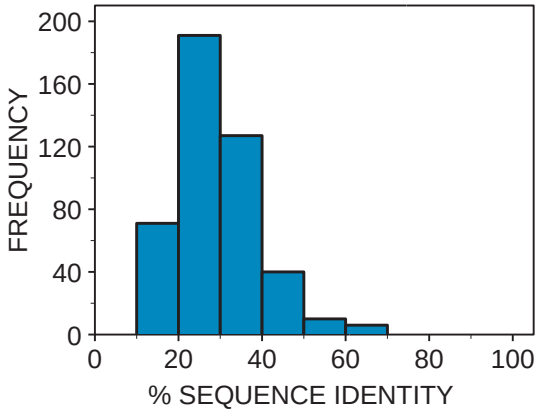
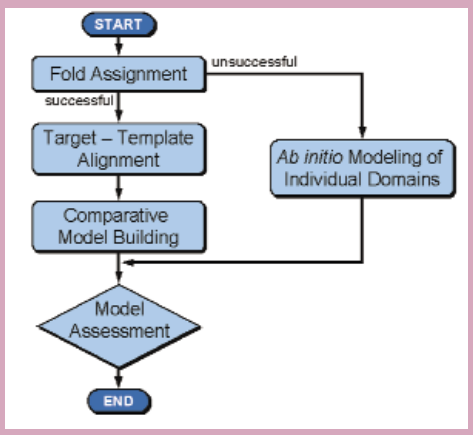


Fig. 1 Distribution of the % sequence identity between the known protein structures and proteins of *Mycobaterium genitalium*, modeled as in ref. 13 in February, 2000. Segments of at least 30 residues in 333 (69%) of the 479 sequences were possible to model or assign a fold. See Table 1 for definitions of a model and a fold assignment.

Laboratories of Molecular Biophysics, Pels Family Center for Biochemistry and Structural Biology, The Rockefeller University, 1230 York Avenue, New York, New York 10021, USA. Correspondence should be addressed to A.S. *email: sali@rockefeller.edu*

Box 1 Modeling protein structures



For a list of pointers to a large number of protein structure modeling tools, many of which are implemented as web servers, see *http://guitar.rockefeller.edu/tools/*, and for a more detailed review, see ref. 10. The first step in modeling of a protein sequence is to attempt to find related known protein structures in the Protein Data Bank for as many domains in the modeled sequence as possible (fold recognition or fold assignment). The folds of domains in the target sequence can be assigned by pairwise and multiple sequence similarity searches as well as by threading methods that rely explicitly on the known structures of the candidate template proteins.

While fold assignment predicts a structural relationship between two proteins, it does not produce an explicit three-dimensional model of the target sequence. Thus, fold assignment is generally followed by alignment of the target sequence with one or more template structures to establish the best possible correspondence between the residues in the target and template sequences. In the more difficult cases, semi-manual alignment of the whole family is necessary for the best results.

After the alignment, the next step is comparative model building that relies on the alignment and the template structures to produce explicit three-dimensional models of the aligned domains of the target protein. These models usually consist of all non-hydrogen atoms for both the main chain and side chains, including the insertions and deletions relative to the template structures.

Finally, the models need to be evaluated by considering structural and energetic criteria, not sequence similarity alone. Model evaluation helps to assess what information can be extracted from the model. If the model is unsatisfactory, it is possible to iterate through the cycle of fold assignment, alignment, modeling, and model evaluation in the search for a satisfactory model. In fact, a useful approach to fold assignment and alignment is to accept uncertain fold assignments and alignments, build a full atom comparative model of the target sequence, and make the final decision about whether or not the match and the alignment are accurate by evaluating the resulting comparative model.

If no suitable fold assignments, alignments and models are obtained, the only recourse are the *ab initio* protein structure prediction methods that depend solely on the sequence of the protein to be modeled[35–37]. Unfortunately, these methods are not yet generally applicable. However, in the hands of experts, a fraction of small proteins or domains can be modeled with accuracy that is comparable to that of the models implied by very difficult fold assignments[12].

structure modeling. The *ab initio* methods attempt to predict the native structure only from the sequence of the modeled protein. So far, the *ab initio* methods have produced models with the correct fold for only a few small protein domains. Nevertheless, recent progress in *ab initio* prediction raises hope that structural genomics will benefit from the *ab initio* modeling of long inserted loops (>15 residues) and domains (<150 residues) that are not accessible to experimental structure determination methods.

Target selection depends on errors in modeling

The targets for structure determination in structural genomics are likely to be individual domains rather than multi-domain proteins. The reason is that the structure of a single domain is usually easier to determine by X-ray crystallography or NMR spectroscopy than that of a more flexible multi-domain protein, although it would be beneficial to determine the structures of whole proteins whenever possible. To be effective, the targets for structural genomics should be chosen to allow calculation of useful models for most protein sequences in sequence databases while minimizing the total experimental effort. We first ask what is a useful level of accuracy for the models based on the experimental structures, and then estimate how many structures need to be determined experimentally to achieve the required level of accuracy. Thus, target selection is informed by the successes and failures of modeling.

Table 1 Leveraging of experimental structures by comparative modeling[1]

Experimental Structure	Models or fold assignments	Models	Useful models	Less accurate models	Fold assignments only
P005	537	345	53	292	192
P007	42	40	28	12	2
P008	31	29	24	5	2
P018	172	50	11	39	122
P100	185	70	11	59	115
P102	26	25	22	3	1
P111	46	44	23	21	2
Total	1039	603	172	431	436

[1]A model is counted if it is at least 60 residues long and is assessed to have >30% of its Cα atoms within 3.5 Å of their true positions[13]. The models are subdivided into two classes. "Useful models" are defined to be based on >30% sequence identity to the known structure, while "Less accurate models" are based on <30% sequence identity. "Fold assignments only" denotes the number of proteins with a significant PSI-BLAST[33] relationship to a known structure ($E < 0.0001$) that failed to produce a reliable model. The calculations were performed in August, 2000.

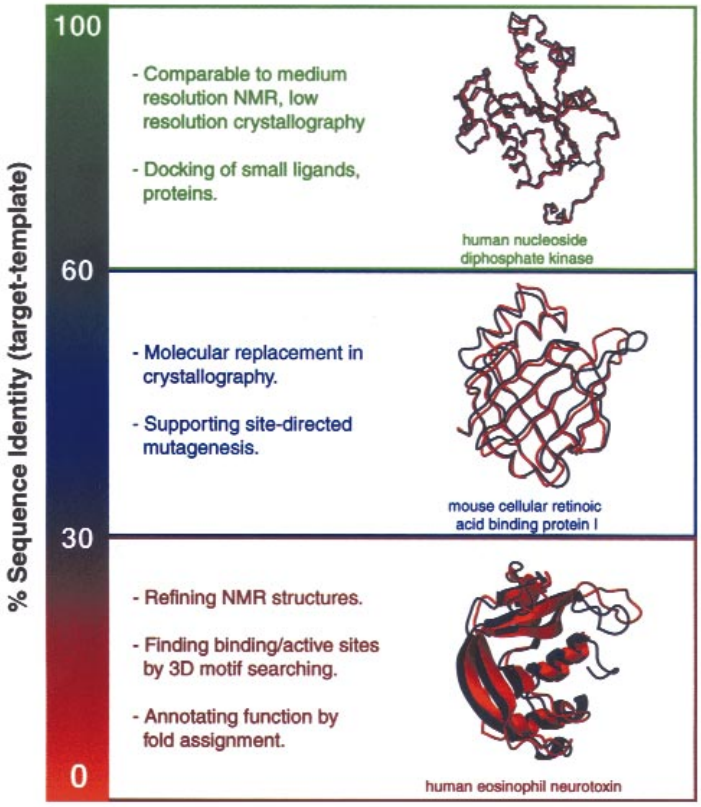


Fig. 2 Applications of comparative modeling. The potential uses of a comparative model depend on its accuracy. This in turn depends significantly on the sequence identity between the modeled sequence and the known structure on which the model was based. Sample models and corresponding experimental structures are shown on the right. Reproduced with permission from ref. 34.

The accuracy of comparative models tends to increase with the sequence similarity between the modeled sequence and the related known structures[13] (Fig. 2). To obtain a reasonable level of accuracy, the models must be based on alignments with few errors. Such alignments can usually be obtained when the sequence identity between the modeled sequence and at least one known structure is >30%. Thus, structural genomics should determine protein structures such that most sequences in the genome databases match at least one structure with an overall sequence identity of >30% (ref. 10). If this degree of sampling is achieved, most of the models will be based on sequence identity in the range of 30–50%. Such models tend to have >85% of the Cα atoms within 3.5 Å of their correct positions[10]. For functional analysis, the accuracy of the models is frequently higher, because the active site regions generally exhibit stronger structural conservation than the rest of the protein. The models based on >30% sequence identity are usually suitable for a number of applications[10], including the testing of ligand binding modes by designing site-directed mutants with altered binding capacity, and computational screening of databases of small molecules for potential inhibitors or lead compounds[8]. A fraction of the models will be based on >50% sequence identity. The average accuracy of such models approaches that of low resolution X-ray structures (3 Å resolution) or medium-resolution NMR structures (10 long-range restraints per residue)[10]. In addition to the applications listed above, these high quality models may be used for more reliable calculations of ligand docking and drug design, provided induced fit is not too large.

The requirement that each protein domain be at least 30% identical in sequence to a known structure determines the number of protein structures that need to be produced by structural genomics. The actual number is hard to estimate, partly because of the difficulties in defining domain families from sequence alone. The estimates for the total number of sequence families, which contain proteins with detectable sequence similarity, range from 5,000 (ref. 14) to 60,000 (ref. 15). The number of clusters of sequence domains that share at least 30% sequence identity with each other is several times larger than the number of sequence families, and is thus likely to be larger than 10,000.

Modeling leverages experimental protein structures — a case for structural genomics

Given the clear and demonstrated usefulness of structural biology, the justification for adding structural genomics to the traditional structural biology effort is the efficiency of large-scale, automated, parallel and industrialized structure determination. Thus, it is crucial that the success and productivity of structural genomics be measured quantitatively. Because structural genomics aims to put every protein sequence within a modeling distance of a known protein structure, appropriate benchmarks for structural genomics include the number and accuracy of comparative models that can be produced based on newly determined structures. The new structures will increase structural characterization in several ways. They will expand known 'structure space' either by revealing an entirely new fold or by associating a new sequence with an existing fold, and they will enable the generation of satisfactory models for previously unsatisfactorily modeled protein sequences.

These arguments are illustrated by an automated modeling exercise with seven structures determined by The New York Structural Genomics Research Consortium (Table 1) (*http://www.nysgrc.org/*). The seven new structures defined five new fold families that were not known at the time of target selection. Comparative modeling with all seven structures yielded useful models based on >30% sequence identity to the new structures for segments of 172 sequences and lower accuracy models for segments of 431 sequences in the non-redundant protein sequence database[16]. Overall, the seven new structures allowed for at least partial structural characterization of 1,039 proteins. Prior to the structural genomics effort, no structural information was known for these 1,039 proteins. These results highlight the power of combining experimental structure determination and comparative modeling, and strongly support the underlying premise of structural genomics.

Depending on a genome, the fraction of the protein sequences that have at least one segment detectably related to one or more known structures currently ranges from 20 to 69% (refs 13,17–23; Fig. 1). This fraction is currently growing with an annual rate of ~5–10% (Fig. 3). Approximately half of the models are relatively inaccurate because they are based on <30% sequence identity to the known structures. The coverage at the level of residues or domains as opposed to whole proteins is a factor of two smaller (Fig. 3). The relatively low coverage of domains, its relatively low growth rate, and relatively low average accuracy of the models justify an investment into structural genomics, in addition to the resources spent on traditional structural biology.

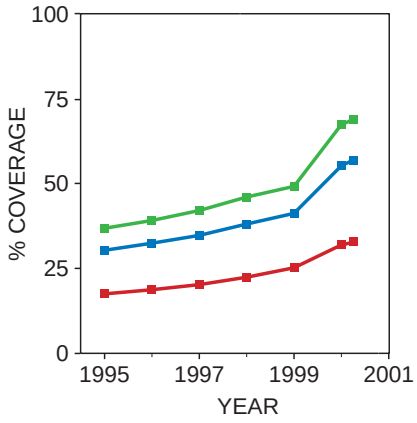


Fig. 3 Simulated effect of PDB growth on fold assignment and modeling coverage of the *Mycoplasma genitalium* proteins. The fraction of sequences that have a model or a PSI-BLAST fold assignment for a segment of at least 30 residues are in green. The fraction of sequences with a model only are in blue. The fraction of residues in the genome that occur in a model or a PSI-BLAST fold assignment are in red. The average fraction of a sequence that is modeled or assigned a fold is currently 48%. See Table 1 for definitions of a model and a fold assignment.

Problems of protein structure modeling

Protein structure modelers and their methods are tested bi-annually at meetings on Critical Assessment of Techniques for Protein Structure Prediction (CASP)[24]. Protein sequences of unknown three-dimensional structure are modeled and submitted to the organizers before the meeting. In parallel, the three-dimensional structures of the prediction targets are being determined by X-ray crystallography or NMR methods. They only become available after the models are calculated and submitted. Thus, a *bona fide* evaluation of protein structure modeling methods is possible. An important complement to the CASP meetings is an online evaluation of protein structure modeling web servers[25,26]. The online evaluation has the potential benefits of a much larger number of test structures, and of providing automatic and continuous feedback about the performance of the modeling servers.

To maximize the benefit of protein structure modeling for structural genomics, further advances are necessary in recognizing weak sequence-structure similarities, aligning sequences with structures, modeling of rigid body shifts, distortions, loops and side chains, as well as detecting errors in a model[10]. Interestingly, fold assignment would not be an important problem at the end of the structural genomics effort, if the target selection scheme suggested above applied and all structures could be determined. The reason is that all sequences would be related trivially at >30% sequence identity to a known structure. Thus, loop modeling and modeling of distortions and rigid body shifts, as well as side chain packing may be the most important challenges for protein structure modeling in the context of structural genomics. The need for accurate methods for assessing errors in protein structure models cannot be over-emphasized, since a model or a low resolution experimental structure with errors may still be used profitably if one is aware of its shortcomings.

Protein structure modeling for structural genomics must be applicable to whole genomes. Thus, there is a need to develop an automated, rapid, robust, sensitive, and accurate protein structure modeling pipeline. Automation makes it efficient for both the experts and non-experts to use the models, allowing them to spend more time designing experiments and interpreting information. It is important that the best possible models be easily accessible to the non-experts. Automation also encourages development of better methods and allows the frequent recalculation of the models that is needed because of the rapid growth of the sequence and structure databases.

In addition to making predictions, modelers are facing a more practical challenge of making others aware of their predictions[27]. Modeling methods need to be evaluated rigorously and be made accessible over the Internet. Users of models must learn how to interpret low-resolution and partially incorrect protein structure models.

The functions of proteins cannot be fully understood if we consider individual protein domains out of their cellular context. The precise nature of the assembly of domains within a larger protein is crucial, as is formation of multi-subunit complexes and transient protein–protein interactions. Because structural genomics will emphasize high throughput, it will most likely result in structures and models only for domains, not whole proteins. Thus, another need underlined by structural genomics is the need for domain docking that is robust with respect to flexibility and errors in the individual domain structures and their models. To the degree that such methods cannot be developed, applications of the models have to take these shortcomings into account.

Standard comparative modeling is not CPU time intensive. For example, a typical comparative model building calculation takes only a few minutes per model. All the operations needed for processing the yeast genome of ~6,400 proteins take only two days on a cluster of 200 Pentium III CPUs. However, application of more accurate, specialized methods for loop and side chain modeling is so time consuming that it is not yet possible to apply them on the genome scale. Another important methodological improvement, which will also require increased computer power and better algorithms, involves automating the cycle of alignment, modeling, and model evaluation for a single protein sequence[10]. This approach can decrease the effect of errors in the input alignment on the final model, but is computationally intensive, requiring from several hours to several days of CPU time for a single target

Box 2 IBM's 'Blue Gene' project

IBM recently announced a five year, $100 million initiative to build a supercomputer 500 times more powerful than the fastest computers available today (http://www.research.ibm.com/news/detail/bluegene.html). The new computer, nicknamed 'Blue Gene', is designed to perform more than one quadrillion operations per second (one petaflop). This performance will be achieved by more than one million processors, each approximately equivalent to a desktop PC.

IBM intends to apply Blue Gene to the *ab initio* protein folding problem. Its massive computing power will be needed to develop more accurate energy functions and protein representations, as well as to simulate molecular dynamics on a millisecond to second time scale. In addition, due to its parallel architecture, Blue Gene will be suited for calculations in bioinformatics where hundreds of thousands of proteins need to be processed essentially independently from each other. Computers such as Blue Gene will almost certainly be rapid enough to allow development of more accurate protein structure prediction methods (see main text), their application on the genomic scale, and timely updates of the models demanded by the rapidly growing input databases of protein sequences and structures.

sequence. Fold assignment by threading techniques on the genome scale is also not yet routinely possible. In addition, the application of long molecular dynamics simulations of a protein in solvent, guided by a complex energy function, promises to improve the accuracy of comparative modeling, but also at a cost of several days of CPU time per model[28,29]. In summary, fast computers are needed to allow development of more accurate modeling methods, their application on the genomic scale, and timely updates of the models demanded by the rapidly growing databases of protein sequences and structures.

Large investment in intensified development of protein structure modeling techniques is justified partly because structural genomics as a whole is a costly effort. The large number of targets and therefore the cost of the project could be reduced significantly by a relatively small improvement in the protein structure modeling techniques. The reasons are that: (i) the errors in models increase rapidly as the sequence identity to the known structures drops below 30% and (ii) most related protein pairs share less than 30–35% sequence identity (Fig. 1). For example, recent improvements of a large-scale comparative modeling pipeline[13] increased the coverage of a typical genome by ~20%, an effect that is equivalent to several years of growth of the Protein Data Bank (PDB)[30] (Fig. 3). If the current average model accuracy corresponding to 30% sequence identity is accepted as sufficient, a new comparative modeling method that is capable of delivering equally accurate models based on only 25% sequence identity would decrease the number of needed experimental structures by ~25%. On the scale of the structural genomics project, this may correspond to 5,000–10,000 structures and justifies a significant investment in the development of new modeling methods and multi-processor computers to run these methods (see Box 2).

New applications of protein structure models

The use of individual protein structure models in biology is already rewarding and increasingly widespread. However, just as the availability of many protein sequences and complete genomes made possible new and powerful methods of sequence analysis (see, for example, ref. 31), a database of many three-dimensional models that is complete at the level of a family, organism, or functional network is certain to encourage new kinds of applications. For example, a good drug target is a protein that is likely to have high ligand specificity; specificity is important because specific drugs are less likely to be toxic. Large-scale modeling facilitates imposing the specificity filter in target selection by enabling a structural comparison of the ligand binding sites of many proteins, either human or from other organisms. Such strategies are expected to work better than the current approaches that are usually based on a comparison of whole protein sequences[32]. For example, when a human pathogen needs to be inhibited, a good target may be a protein whose binding site shape is different from related binding sites in all of the human proteins. Similarly, when a human metabolic pathway needs to be regulated, the target identification could focus on the particular protein in the pathway that has the binding site most dissimilar from its human homologs.

No single experimental or computational approach is likely to result in accurate and complete models of proteins, protein assemblies, and pathways. Thus, a major challenge for modelers is to integrate structural models with other types of data, such as classical biochemical characterizations, sequence-based analyses, proteomics, and genome scale expression data and pairwise interaction maps.

Finally, there is hope that the protein structure prediction methods will benefit from structural genomics in a more fundamental way, not only from the existence of a larger number of templates for modeling. It is conceivable that the availability of many experimentally determined structures will increase our understanding of the physics and evolution of protein structures, and finally reveal the elusive code that links protein sequence to its structure. However, even without major conceptual or technological advances, it seems likely that we will witness a transition from knowing structures for only a fraction of all protein sequences to having structural information for most globular proteins within the next five to ten years.

Acknowledgments

We are most grateful to S.K. Burley, T. Gaasterland, J. Kuriyan, J. Bonanno, M. Chance, S. Almo, L. Shapiro, C. Lima and other members of the New York Structural Genomics Research Consortium for many discussions about structural genomics. We also thank J.P. Overington for comments on the manuscript. R.S. is a Howard Hughes Medical Institute predoctoral fellow. A.S. is an Alfred P. Sloan Research Fellow and an Irma T. Hirschl Trust Career Scientist. Support from The Merck Genome Research Institute, Mathers Foundation, the NSF, and the NIH is also acknowledged.

Associations with structural genomics

The authors are associated with the New York Structural Genomics Research Consortium. A.S. co-founded Prospect Genomics Inc. and consults for Structural Genomix Inc. and Molecular Simulations Inc.

1. Terwilliger, T.C. *et al. Protein Sci.* **7**, 1851–1856 (1998).
2. Burley, S.K. *et al. Nature Genet.* **23**, 151–157 (1999).
3. Brenner, S. & Levitt, M. *Protein Sci.* **9**, 197–200 (2000).
4. Koonin, E. *Trends Genet.* **16**, 16 (2000).
5. Skolnick, J., Fetrow, J., & Kolinski, A. *Nature Biotechnol.* **18**, 283–287 (2000).
6. Thornton, J., Orengo, C., Todd, A., & Pearl, F. *J. Mol. Biol.* **293**, 333–42 (1999).
7. Andrade, M. A. *et al. Bioinformatics* **15**, 391–412 (1999).
8. Makino, S., Ewing, T.J., & Kuntz, I.D. *J. Comput. Aided. Mol. Des.* **13**, 513–532 (1999).
9. Blundell, T. L., Sibanda, B.L., Sternberg, M. J. E. & Thornton, J. M. *Nature* **326**, 347–352 (1987).
10. Martí-Renom, M.A. et al. *Ann. Rev. Biophys. Biomolec. Struct.* **29**, 291–325 (2000).
11. Jones, D. *Current Opinion Structural Biol.* **10**, 371–379 (2000).
12. Baker, D. *Nature* **405**, 39–42 (2000).
13. Sánchez, R. & Sali, A. *Proc. Natl. Acad. Sci. USA* **95**, 13597–13602 (1998).
14. Wolf, Y., Grishin, N., & Koonin, E. *J. Mol. Biol.* **299**, 897–905 (2000).
15. Yona, G., Linial, N. & Linial, M. *http://protomap.stanford.edu/*
16. Benson, D. *et al. Nucleic Acids Res* **28**, 15–18 (2000).
17. Fischer, D. and Eisenberg, D. *Proc. Natl. Acad. Sci. USA* **94**, 11929–11934 (1997).
18. Rychlewski, L., Zhang, B., & Godzik, A. *Fold. Des.* **3**, 229–238 (1998).
19. Huynen, M., Doerks, T., Eisenhaber, F. & Orengo, C. *J. Mol. Biol.* **280**, 323–326 (1998).
20. Teichmann, S. A., Park, J., & Chothia, C. *Proc. Natl. Acad. Sci. USA* **22**, 14658–14663 (1998).
21. Jones, D. T. *J. Mol. Biol.* **287**, 797–815 (1999).
22. Guex, N., Diemand, A., & Peitsch, M.C. *Trends Biochem. Sci.* **24**, 364–367 (1999).
23. Sánchez, R. *et al. Nucl. Acids Res.* **28**, 250–253 (2000).
24. Moult, J., Hubbard, T., Fidelis, K. & Pedersen, J. T. *Proteins* **Suppl. 3**, 2–6 (1999).
25. Fischer, D. *et al.* Proteins **Suppl. 3**, 209–217 (1999).
26. Eyrich, V., Martí-Renom, M. A., Sali, A. & Rost, B. *http://dodo.cpmc.columbia.edu/~eval*
27. Brenner, S. E., Barken, D. & Levitt, M. *Nucl. Acids Res.* **27**, 251–253 (1999).
28. Duan, Y. and Kollman, P. A. *Science* **282**, 740–744 (1998).
29. Brooks III, C. L., Karplus, M. & Pettit, B. M. *Proteins: A theoretical perspective of dynamics, structure and thermodynamics.* (John Wiley & Sons, New York; 1988).
30. Berman, H.M. *et al. Nucleic Acids Res.* **28**, 235–242 (2000).
31. Sali, A. *Nature* **402**, 23–26 (1999).
32. Rosamond, J. & Allsop, A. *Science* **287**, 1973–1976 (2000).
33. Altschul, S.F. *et al. Nucleic Acids Res.* **25**, 3389–3402 (1997).
34. Sali, A. & Kuriyan, *J. Trends Biochem. Sci.* **22**, M20–M24 (1999).
35. Sippl, M. *Structure* **7**, R81–R83 (1999).
36. Sternberg, M. J. E., Bates, P. A., Kelley, L. A. & MacCallum, R. M. *Curr. Opin. Struct. Biol.* **9**, 368–373 (1999).
37. Koehl, P. & Levitt, M. *Nature Struct. Biol.* **6**, 108–111 (1999).