Pure Appl. Chem., Vol. 74, No. 6, pp. 907–914, 2002.

Modeling genome structure and function*

Ram Samudrala

Department of Microbiology, University of Washington, Seattle, WA 98195-7242, USA

Abstract: The ongoing genomics revolution has led to the creation and enumeration of all the genes encoded within several organisms. The next steps are to catalog all proteins, their structures, and their functions in different contexts. At the same time, scientists have been pursuing experimental and theoretical approaches to integrate this information to gain understanding of the behavior of entire systems. In this work, we provide a framework for obtaining structures for all tractable protein sequences encoded by a genome, and using the resulting structures to aid in understanding function. Our aim is to integrate the output produced with other genomic and proteomic data to create a comprehensive picture of systems and organismal biology.

INTRODUCTION: CREATING A “PARTS LIST” OF GENES AND PROTEINS

A fundamental biological challenge is to understand how the linear information in an organism’s genome is processed to produce the resulting behavior (phenotype). Given a gene, it is possible to directly determine the protein sequence using the Genetic Code. Understanding how the amino acid sequence is translated into a 3D structure, and how these structures interact with their environment to give rise to biological function, are among the most fascinating and important problems in the world today.

The nucleotide sequences of all the genes present in an organism have been determined for several organisms [1,2]. The sequencing of whole genomes heralds a new revolution in biology, both molecular and organismal. Sophisticated computational and experimental techniques can be used to parse an entire genome to create a “parts list” of all the genes present, and, consequently, all the proteins encoded by those genes.

Such a “parts list” is only the first step in understanding the correlation between genotype and phenotype. Once we have the sequences of all the proteins encoded by an organism’s genome, we need to understand how the proteins function in the context of their environment. To help achieve this goal, one possible approach is to determine the 3D folds these proteins adopt, and use that as a means of annotating function, since it is the structure that mediates function in nature. Once the structures and functions are obtained for the proteins, they can be integrated with the vast amount of individual sequence, genomic (microarray), and proteomic (mass spectrometry, genome-wide two-hybrid assay) data to provide a comprehensive picture of organismal behavior [3,4].

*Plenary lecture presented at the International Conference on Bioinformatics 2002: North–South Networking, Bangkok, Thailand, 6–8 February 2002. Other presentations are presented in this issue, pp. 881–914.

OBTAINING 3D STRUCTURE FOR PROTEINS

Protein structure determination

There are two primary experimental techniques used for determining a protein structure: X-ray crystallography and nuclear magnetic resonance (NMR) experiments. These techniques form the basis of global efforts in structural genomics projects currently underway [5]. While these methods produce the best models for the native structure of a protein, they are generally time- and labor-intensive. The continually increasing amount of DNA and protein sequence data from genome projects makes it infeasible for NMR and X-ray crystallography techniques to rapidly provide information about the 3D structures of all the sequences determined [6]. Thus, there is an urgent need for robust methods for predicting structure from amino acid sequence.

Protein structure prediction

There are two primary categories of methods for predicting protein structure from sequence: comparative and ab initio modeling. In the comparative modeling category, the methodologies rely on the presence of one or more evolutionarily related template protein structures that are used to construct a model; the evolutionary relationship can be deduced from sequence similarity [7–10] or by "threading" a sequence against a library of structures and selecting the best match [11–13]. In the ab initio category, there is no strong dependence on database information, and prediction methods are based on general principles that govern protein structure and energetics [14–18]. The categories vary in difficulty, and, consequently, methods in each of these categories produce models with different levels of accuracy relative to the experimental structure.

Protein structure prediction methods are rigorously evaluated by the Critical Assessment of Structure Prediction (CASP) experiments held every two years (special issues of *Proteins: Structure, Function, Genetics*, 1995, 1997, 1999, and 2002). These experiments evaluate prediction techniques by asking modelers to construct models for a number of protein sequences before the experimental result is known, over a period of 3–4 months. We have taken part in all four CASP experiments, including the most recent one (CASP4) that finished in December 2000 [19]. The CASP4 results provide a benchmark as to what level of model accuracy we can currently expect from our approaches.

At CASP4, we made predictions for all of the 40 targets for which an experimental answer was made available [20]. The CASP4 results show that within each of the general structure prediction categories, some methods, including ours, are able to produce models with a fair amount of accuracy.

Comparative modeling and fold recognition

Protein sequences that were determined to be evolutionarily related to sequences with known structure were modeled using comparative modeling techniques developed by us. The same procedure was used for comparative modeling and fold recognition targets. Target sequences related to sequences that have conformations determined by experiment were candidates for comparative modeling. Generally, alignments were obtained from the various servers available as part of the CAFASP meta-server [21]. Initial models were then constructed, and structure-based alignments were used in an iterative manner to refine alignments manually. Nonconserved side chains and main chains were constructed using a graph-theoretic approach with sampling provided by exhaustive and database searches. The final conformations were minimized by ENCAD [20].

Figure 1 shows some examples of the comparative modeling predictions with different difficulties made at CASP4. In the comparative modeling category, we made 29 predictions for targets that had sequence identities ranging from 50–10 % to the nearest related protein with known structure. For 23 of these proteins, we produced models ranging from 1.0–6.0 Å root mean square deviation (RMSD) for the Cα atoms between the model and the corresponding experimental structure for all or large parts of the protein, with model accuracies scaling fairly linearly with respect to sequence identity (i.e., the

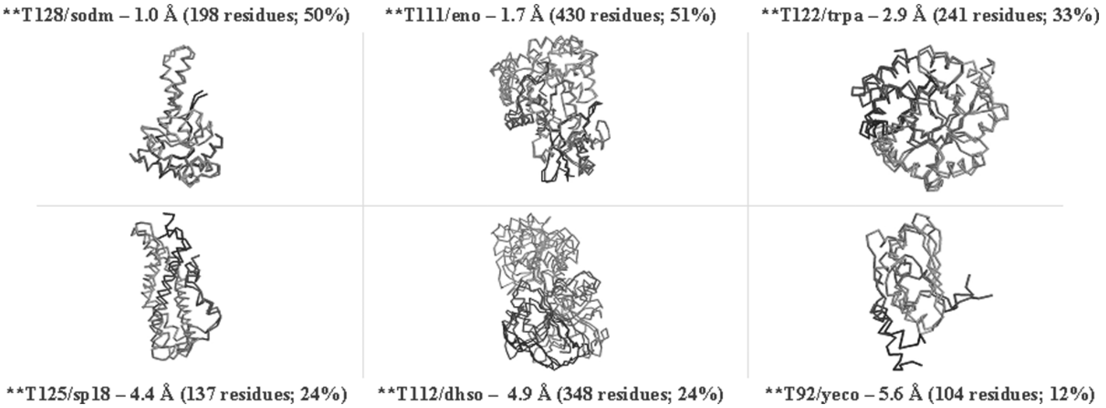

**T128/sodm – 1.0 Å (198 residues; 50%)
**T111/eno – 1.7 Å (430 residues; 51%)
**T122/trpa – 2.9 Å (241 residues; 33%)
**T125/sp18 – 4.4 Å (137 residues; 24%)
**T112/dhso – 4.9 Å (348 residues; 24%)
**T92/yeco – 5.6 Å (104 residues; 12%)

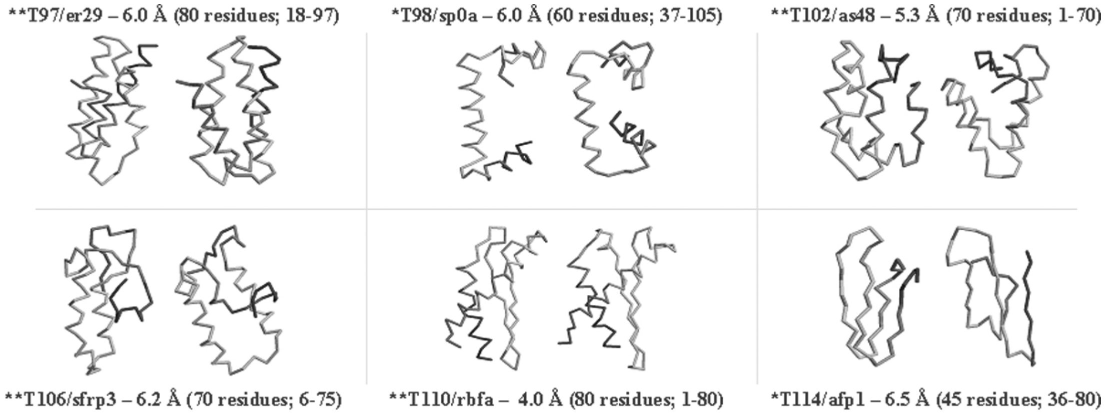
Fig. 1 Six examples of our comparative modeling predictions at CASP4 for targets with different difficulties. The superposition of the model and the experimental structures is shown, along with the C_α RMSD relative to the experimental structure and the percentage identity of the alignment between the target and template sequences. We made useful predictions for 23 out of 29 targets (**): sequences with high percentage identity to the template structures (≥50 %) were modeled well (1–2 Å RMSD) with model accuracy decreasing (4–6 Å RMSD) fairly linearly as the relationship becomes more tenuous (10–25 % sequence identity).

higher the sequence identity, the better the prediction). Generally, on the models that were constructed to within 4.0 Å, side chains were predicted with an accuracy of 60–75 % correct for the $\chi 1$ angles, small loops were predicted within 1–2 Å, and larger loops were predicted to 1–3 Å. The accuracy of the loops and side chain building decreased as the relationship between the template and target sequences grew more distant.

Ab initio prediction

Target sequences without known homologs or analogs that were small in size and/or predicted to have largely helical content were modeled by our ab initio protocol. Such clusters can be subsequences of larger proteins, in which case they most likely represent domain boundaries [22]. Our general paradigm for predicting structure involves sampling the conformational space (or generating "decoys") such that native-like conformations are observed, and then selecting them using a hierarchical filtering technique with many different scoring functions. The two parts to our method are designed such that they are completely automated and readily extendable to the genome-wide level. Generally, we explore combinations of different representations/move sets with two search methods for exploring protein conformational space, and combinations of a variety of scoring function "filters" to identify biologically relevant conformations. We start with a sequence and generate conformations using two different move sets: fragments from a database with identical sequence and a 14-state ϕ,ψ model. Many trajectories are generated and minimized using two different protocols: Monte Carlo with simulated annealing and a genetic algorithm search. The minimization function is primarily an all-atom conditional probability discriminatory function, a hydrophobic compactness function, and a bad contacts function. Once a set of conformations is generated, a hierarchical filtering technique is applied using many different filters/scoring functions to produce one or a few final conformations [20].

Figure 2 illustrates some of our more successful predictions at CASP4 in the ab initio category. We made 11 predictions for targets that had no detectable sequence relationships. We produced 9 models with accuracies ranging from 4.0–6.0 Å C_α RMSD for 60–100 residue proteins (or large fragments of a protein).



**T97/er29 – 6.0 Å (80 residues; 18-97)

*T98/sp0a – 6.0 Å (60 residues; 37-105)

**T102/as48 – 5.3 Å (70 residues; 1-70)

**T106/sfrp3 – 6.2 Å (70 residues; 6-75)

**T110/rbfa – 4.0 Å (80 residues; 1-80)

*T114/afp1 – 6.5 Å (45 residues; 36-80)

Fig. 2 Examples of our ab initio predictions. Five of the examples were predictions submitted for CASP4; the sixth (T102/as48) is a "postdiction" using the actual secondary structure assignment that was available to all CASP predictors (our CASP4 submission for this target used predicted secondary structure, which was only 60 % accurate). The experimental structure is on the left, and the model is on the right. We were able to make topologically accurate predictions for 9 out of 11 targets modeled (**). Targets with largely helical content are modeled well, with predictions as accurate as 4.0 Å C_{α} RMSD for 80 residues.

MODELING STRUCTURES FOR ALL PROTEINS ENCODED BY A GENOME

Even though prediction methods need further development before they can produce models that can match experiment, they can be applied to large numbers of sequences relatively easily (costing only computational time), and the outputs produced could be thought of as an "educated guess" as to the protein's native structure. It is thus possible to construct a "genome prediction engine" using the computational resources available where we can take the protein sequences encoded by an organism's genome and attempt to predict their structures, and use the modeled structures to predict functions (Fig. 3).

Analyses of small genomes show that about 30–40 % of the proteins within the genome can be modeled by comparative modeling methods [23–26]. An additional 20–30 % of the sequences are (or contain) small domains with simple secondary structures that are viable candidates for ab initio structure prediction [27]. The remaining proteins are usually not amenable to structure prediction and sometimes even structure determination (a significant fraction of the latter are membrane proteins).

ANNOTATING FUNCTIONS FOR PROTEINS

The reason for obtaining structures for proteins encoded by a genome is so that they can be used to understand function and further our knowledge about the organism's biology. Given the different protein sequences, and corresponding predicted and experimental structures, it is possible to use a barrage of techniques to annotate functions. Even though structure prediction methods need further development, it is possible to produce models where functional hypotheses can be tested in a rational manner (for example, with mutagenesis experiments) through detailed analysis [28,29]. Additionally, structure comparisons can be used to detect functional homology that cannot be detected by sequence information alone [26], and microenvironment analyses that parse models for particular 3D motifs [30] can be used to discern molecular function. Both these structure-based approaches, used complementarily in conjunction with experimental data and sequence-only approaches like PROSITE [31], BLOCKS [32], and PRINTS [33], will enable us to better assign function to all or large parts of a proteome (Fig. 4).

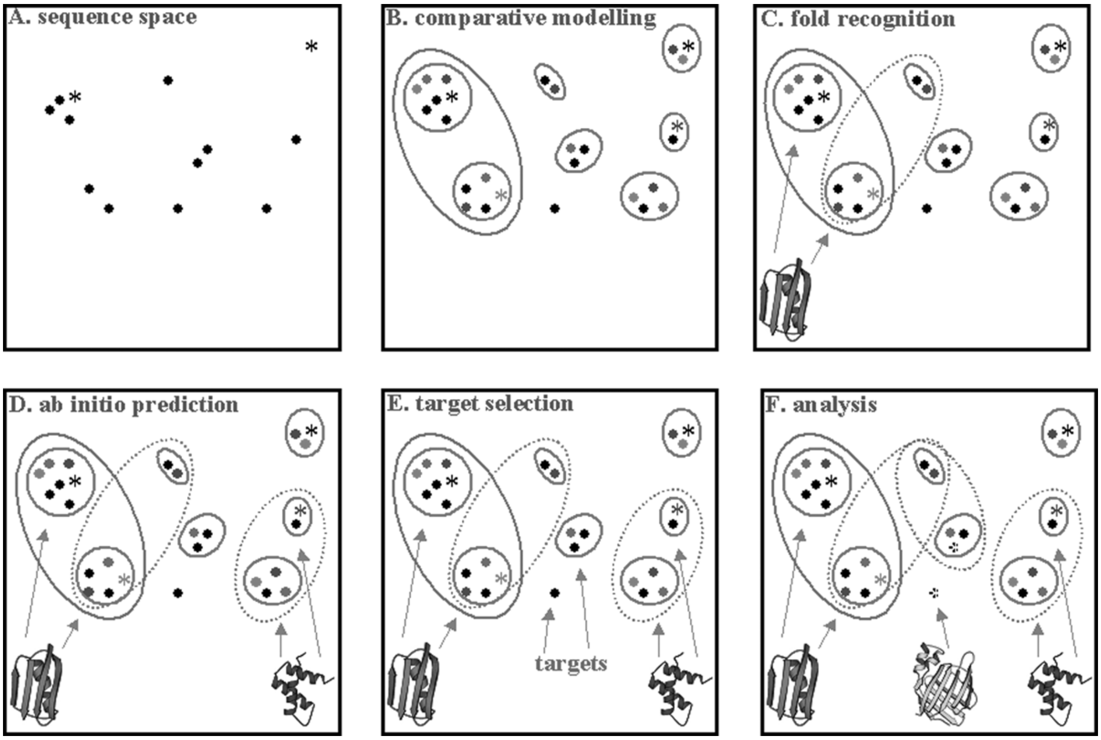


Fig. 3 Computational aspects of structural genomics. In the first steps (A & B, in an abstract hypothetical example above), we compare the protein sequences encoded by the genome of an organism to all known protein sequences and cluster them into families (solid circles). All members of families containing a known structure will be modeled using comparative modeling methods (B). The modeled structures will be used to cluster the proteins further, and the resulting alignments will be used to construct superior sequence recognizers and find more evolutionary distant relationships (C) (dotted circles). These will be used for further modeling. Proteins not modeled by comparative modeling techniques will be screened to determine if they can be predicted using ab initio methods (D). The remaining proteins will be candidates for experimental structure determination (E). Once models for all tractable proteins are obtained, this process can be iterated across different genomes, after further analysis of the available data (F). (Figure idea courtesy of Steven Brenner.)

PRELIMINARY DATA AND RESULTS

Structure prediction

We have started implementing the above protocols for the genomes of three organisms: the opportunistic pathogen *Pseudomonas aeruginosa*, the model cereal *Oryza sativa* (rice), and human. The sample sizes we are currently dealing with are ~6000 coding sequences for the *P. aeruginosa* and ~60 000 for the rice and human genomes. For all three genome sets, 30–40 % of the sequences can be reliably modeled using comparative modeling and fold recognition methods. Another 10–20 % of the sequences (or domains within the sequences) can be modeled using ab initio methods. Our goal, after performing the modeling for all tractable proteins, is to use these modeled structures to annotate function.

For each of the three organisms, have produced models for two initial sets of 500 proteins using our predictive methodologies in each category. For modeling by homology, the proteins selected were those that had relatively high similarity to those with known structure, so that we could be assured of

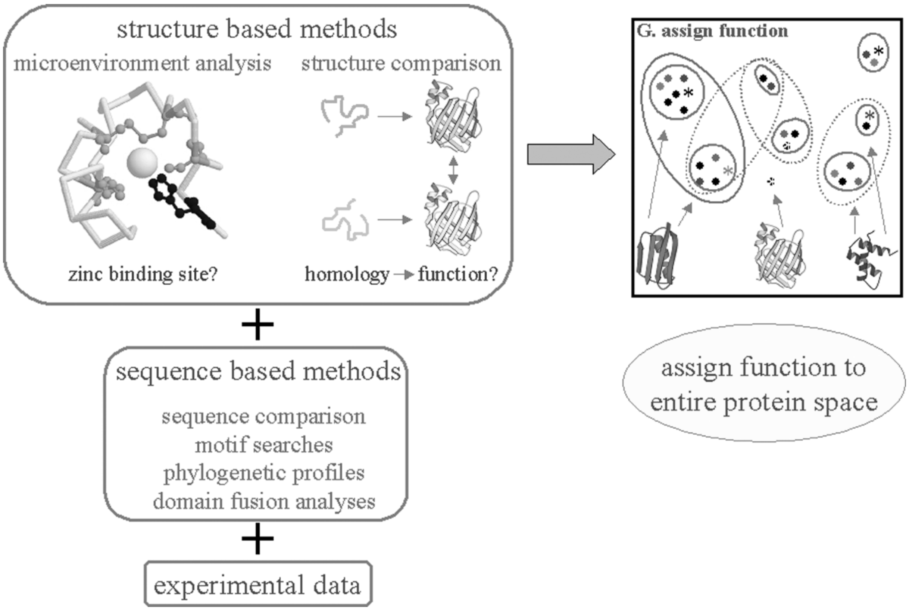


Fig. 4 Computational aspects of functional genomics. The structure-based methods we use include structure comparisons to infer function from homology, and microenvironment analysis to identify specific structural motifs corresponding to a particular function. Our goal is to use a barrage of computational techniques, both at the sequence and structure level, to annotate function in conjunction with available experimental data.

the high quality of the models and lack of an alignment error. These models are expected to be 1.0–2.0 Å RMSD from the corresponding native conformations based on our control studies where we modeled six proteins from each organism whose experimental structure was already determined, and our CASP results. In the case of our ab initio predictions, we selected the 100 proteins based on their having little or no structural or functional similarity to any other protein in the sequences databases, such that any annotation we make will provide information about a protein where none existed before.

Functional annotation

The initial sets of models produced using comparative modeling methods are not likely to be superior to straightforward sequence comparison for annotating function, but they will be useful for detailed studies where there is additional experimental data available.

For our ab initio predictions, about 10 % of the proteins modeled have strong structural hits to proteins in the SCOP database [34]. Even though structure comparison can be used to detect homology not observed by performing sequence comparisons, it is not clear whether modeled structures would find the same matches. To test this in a preliminary manner, we compared both the structures and our best predicted ab initio models of six CASP targets to determine if they identified the correct homolog. These proteins did not display a sequence relationship to any protein with known structure, but the structure comparisons provide strong evidence for an evolutionary relationship. In all six cases, the top scoring protein, measured using the Z-score provided by the CE program [35], was identical regardless of whether the predicted model or the actual experimental structure was used. This would suggest that structure comparisons of the models can be used with a reasonable expectation that they would produce similar results if the corresponding experimental result were used.

The best structural alignments can be used to create a comparative model, which can then be used to screen the entire sample of conformations generated by our ab initio protocol. We have performed this type of an analysis for the type III chaperone, Invb, from *Salmonella typhimurium*, allowing us to map mutations affecting binding to its effector proteins in a robust manner. The final model produced after iterating this process results in a mapping where the mutations are clustered strongly and provide strong evidence of putative binding sites. This alignment and subsequent model is superior to any produced by publicly available popular fold recognition/threading servers.

ASSIGNMENT OF CONFIDENCES AND DISSEMINATION OF DATA

We can assign preliminary estimates of confidence for structure and function for these sets of proteins based on the consensus results of many different sources, but more predictions are necessary to develop a better model for assigning confidences (for example, if the results of the modeling of two homologous proteins agree, then the prediction is more likely to be correct). Currently, it takes about a month on a farm of 64 1 GHz Pentium III processors to produce 200 comparative and ab initio models, and also to run the sequences through a barrage of sequence-only methods to gain more evidence. Our goal is to continue the modeling process in an iterative manner, continuing to add discrete sets of proteins to our modeling queue. The results of the modeling, along with summaries of the predictions made in conjunction with experimental data, will be made available via the Bioverse Web server [36].

INTEGRATING MOLECULAR AND GENOMIC DATA

Proteins in a cell do not work in isolation of one another. Thus, to understand the function of multiprotein complexes, or whole proteomes, it is necessary to have a structural and functional model for many proteins encoded by the genome of an organism. The CASP results indicate that structure prediction methods have matured to a point where they can be applied on a genome-wide scale, and that these structures can be used with novel but straightforward approaches to understand molecular function [28,30,37].

As technology develops, the sequencing of genomes for specific members of a population will become routine. However, raw sequences offer little information on their own. Obtaining structures for all tractable proteins encoded by an organism's genome, through computational and experimental techniques, combined with other genomic/proteomic data, gene expression arrays [38], genome-wide two-hybrid experiments [39], and other proteomics studies [40], will provide us with a dynamic picture of organismal structure, function, and evolution [41].

ACKNOWLEDGMENTS

This work was supported in part by a Burroughs Wellcome Fund Fellowship from the Program in Mathematics and Molecular Biology to Ram Samudrala. The author thanks John Moult and Michael Levitt and past and present members of their groups for their intellectual guidance and support.

REFERENCES

1. C. Fraser, J. Eisen, S. Salzberg. *Nature (London)* **406**, 799–803 (2000).
2. TIGR Gene Indices <http://www.tigr.org/tdb/tgi.shtml>.
3. D. Searls. *Annu. Rev. Genomics Hum. Genet.* **1**, 251–279 (2000).
4. H. Ge, Z. Lui, G. Church, M. Vidal. *Nat. Genet.* **29**, 482–486 (2001).
5. R. Stevens, S. Yokoyama, I. Wilson. *Science* **294**, 89–92 (2001).
6. A. May, M. Johnson, S. Rufino, H. Wako, Z. Zhu, R. Sowdhamini, N. Srinivasan, M. Rodionov, T. Blundell. *Phil. Trans. Roy. Soc. Lond.* **344**, 373–381 (1994).

7. R. Doolittle. *Science* **214**, 149–159 (1981).
8. J. Greer. *Proteins: Struct., Funct., Genet.* **7**, 317–334 (1990).
9. C. Sander and R. Schneider. *Proteins: Struct., Funct., Genet.* **9**, 56–68 (1991).
10. A. Murzin and A. Bateman. *Proteins: Struct., Funct., Genet.* **29S**, 105–112 (1997).
11. J. Bowie, R. Luthy, D. Eisenberg. *Science* **253**, 164–170 (1991).
12. D. Jones, W. Taylor, J. Thornton. *Nature* **258**, 86–89 (1992).
13. H. Flockner, F. Domingues, M. Sippl. *Proteins: Struct., Funct., Genet.* **S1**, 129–133 (1997).
14. J. Lee, A. Liwo, D. Ripoll, J. Pillardy, J. Scheraga. *Proteins: Struct., Funct., Genet.* **S3**, 204–208 (1999).
15. A. Ortiz, A. Kolinkski, P. Rotkiewicz, B. Ilkowski, J. Skolnick. *Proteins: Struct., Funct., Genet.* **S3**, 177–185 (1999).
16. D. Osguthorpe. *Proteins: Struct., Funct., Genet.* **S3**, 186–193 (1999).
17. R. Samudrala, Y. Xia, E. Huang, M. Levitt. *Proteins: Struct., Funct., Genet.* **S3**, 194–198 (1999).
18. K. Simons, R. Bonneau, I. Ruczinski, D. Baker. *Proteins: Struct., Funct., Genet.* **S3**, 171–176 (1999).
19. Critical Assessment of protein Structure Prediction methods <http://predictioncenter.llnl.gov/>.
20. R. Samudrala and M. Levitt. "A comprehensive analysis of 40 blind protein structure predictions" (2002). To be submitted.
21. J. Bujnicki, A. Elofsson, D. Fischer, L. Rychlewski. *Bioinformatics* **17**, 750–751 (2001).
22. J. Gouzy, F. Corpet, D. Kahn. *Comp. Chem.* **23**, 333–340 (1999).
23. R. Sanchez and A. Sali. *Proc. Natl. Acad. Sci. USA* **95**, 13597–13602 (1998).
24. D. Jones. *J. Mol. Biol.* **287**, 797–815 (1999).
25. M. Martin-Renom, A. Stuart, A. Fiser, R. Sanchez, F. Melo, A. Sali. *Annu. Rev. Biophy. Biomol. Struct.* **29**, 291–325 (2000).
26. S. Brenner and M. Levitt. *Protein Sci.* **9**, 197–200 (2000).
27. R. Bonneau and D. Baker. *Annu. Rev. Biophy. Biomol. Struct.* **30**, 173–189 (2001).
28. R. Samudrala, Y. Xia, M. Levitt, N. Cotton, E. Huang, R. Davis. In *Proceedings of the Pacific Symposium on Biocomputing*, R. Altman, A. Dunker, L. Hunter, T. Klein, K. Lauderdale (Eds.), pp. 179–189 (2000).
29. C. Van Loy, E. Sokurenko, R. Samudrala, S. Moseley. *Mol. Microbiol.* (2002). In press.
30. L. Wei, E. Huang, R. Altman. *Structure* **7**, 643–650 (1999).
31. K. Hofmann, P. Bucher, L. Falquet, A. Bairoch. *Nucleic Acids Res.* **27**, 215–219 (1999).
32. J. Henikoff, E. Green, S. Pietrokovski, S. Henikoff. *Nucleic Acids Res.* **28**, 228–230 (2000).
33. T. Attwood, M. Croning, D. Flower, A. Lewis, J. Mabey, P. Scordis, J. Selley, W. Wright. *Nucleic Acids Res.* **28**, 225–227 (2000).
34. T. Hubbard, A. Murzin, S. Brenner, C. Chothia. *Nucleic Acids Res.* **25**, 236–239 (1997).
35. I. Shindyalov and P. Bourne. *Protein Eng.* **11**, 739–747 (1998).
36. Bioverse <http://bioverse.compbio.washington.edu>.
37. A. Barnes and C. Wynn. *Proteins: Struct., Funct., Genet.* **4**, 182–189 (1988).
38. E. Lander. *Nat. Genet.* **21**, 3 (1999).
39. B. Schwikowski, P. Uetz, S. Fields. *Nature Biotechnol.* **18**, 1242–1243 (2000).
40. S. Gygi, B. Rist, S. Gerber, F. Turecek, M. Gelb, R. Aebersold. *Nature Biotechnol.* **17**, 994–999 (1999).
41. T. Ideker, V. Thorsson, J. Ranish, R. Christmas, J. Buhler, J. Eng, R. Bumgarner, D. Goodlett, R. Aebersold, L. Hood. *Science*. **292**, 929–934 (2001).