PROTEIN STRUCTURE MODELING

NARAYANAN ESWAR*, ANDREJ SALI*
Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, California Institute for Quantitative Biosciences, University of California at San Francisco, San Francisco, CA, USA

Abstract. Known protein sequences outnumber known protein structures by more than two orders of magnitude. Given this huge sequence-structure gap, most protein structures need to be predicted by computational methods rather than determined by experimental techniques. This chapter outlines various protein structure modeling approaches and associated resources.

Keywords: Comparative modeling, homology modeling, threading, integrative modeling, sequence-structure alignment

1. Introduction

Cellular functions are dependent on the three-dimensional (3D) structures of proteins and their complexes with small molecules and other macromolecules. Knowing the structures of the proteins is thus crucial for the understanding of cellular processes. The 3D structures of the proteins and their complexes are best determined by experimental methods that yield solutions at atomic resolution, such as x-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. However, despite recent advances in the application of such techniques in high-throughput mode, experimental structural characterization remains an expensive and time-consuming task (Chandonia and Brenner, 2006).

* To whom correspondence should be addressed. Andrej Sali or Eswar Narayanan, UCSF MC 2552, Byers Hall Suite 503B, 1700 4th Street, San Francisco, CA 94158-2330, USA; e-mails: sali@salilab.org, eswar@salilab.org

J.L. Sussman and P. Spadon (eds.), *From Molecules to Medicines*,

The publicly available Protein Data Bank (PDB) currently contains only ~50,000 structures (Berman et al., 2007). In contrast, rapid improvements in genome sequencing resulted in approximately five million protein sequences, including the complete genetic blueprints of humans and hundreds of other organisms (Bairoch et al., 2005; Benson et al., 2005). This wide sequence-structure gap can only be bridged by computational means.

Fortunately, domains in protein sequences are evolving gradually and can thus be clustered into a relatively small number of families with similar sequences and structures (Vitkup et al. 2001; Chandonia and Brenner 2005b). For instance, ~80% of all sequences in the UniProt database can be clustered into approximately 10,000 families (Bru et al., 2005; Letunic et al., 2006; Finn et al., 2008). Similarly, all the structures in the PDB can be classified into approximately 1,000 distinct folds (Andreeva et al., 2004; Pearl et al., 2005). Many computational methods for protein structure modeling seek to exploit these evolutionary relationships.

Computational approaches to protein structure prediction are greatly facilitated by the structural genomics initiative (Liu et al., 2007; Moult, 2008). Structural genomics aims to maximize the structural coverage of the sequence space by experimentally determining the representative structures for as many families as possible, thus allowing accurate modeling of the remaining members of these families (Sali, 1998). Currently, most targets for experimental structure determination are chosen from the largest protein families such that, in combination with computational methods, each new structure yields useful structural information for the largest possible fraction of sequences in the shortest possible time frame (Chandonia and Brenner 2005a).

2. Computational structure determination

There are three main types of computational protein structure modeling methods. First, *ab initio* methods aim to predict the structure of a target protein purely from its primary sequence using principles of physics that govern protein folding and/or using information derived from known structures but without relying on any evolutionary relationship to known folds (Simons et al., 1999; Das and Baker, 2008). Currently, these methods can only be applied to individual domains of less than approximately 150 residues (Baker and Sali, 2001).

Second, homology or comparative modeling methods rely on the fact that similar sequences adopt similar 3D structures. Comparative modeling consists of four main steps (Marti-Renom et al., 2000): (i) fold assignment

that identifies similarity between the target sequence of interest and at least one known protein structure (the template); (ii) alignment of the target sequence and the template(s); (iii) building an atomic model of the target based on the alignment with the chosen template(s); and (iv) predicting errors in the model. The first two steps, fold assignment and sequence-alignment, are frequently achieved by sequence-structure threading methods that seek to assess a coarse model derived by threading the target sequence through each structure within a library of protein folds (templates). The threading methods are most useful when the similarity between a target sequence and any of the known structures is not statistically significant (Godzik, 2003). Threading methods achieve higher sensitivity than sequence comparison methods by using structural information derived from the templates. Comparative modeling is the most accurate approach that can be easily applied on a large-scale to address the sequence-structure gap. Though these three classes of methods seem to address distinct regimes of the structure prediction problem, the divisions between them are increasingly being blurred. State-of-the-art modeling methods tend to employ the best features of each of these methods to improve the accuracy of the resulting models.

Finally, a third group of methods, recently receiving a lot of attention, is the "integrative" or "hybrid" methods that combine information from a varied set of computational and experimental sources, including those listed above (Alber et al., 2008).

3. Geometrical accuracy of comparative protein structure models

We now focus on comparative protein structure modeling. The geometrical accuracy of comparative models can be estimated by building models for sequences with known structures and comparing them to their native structures. Specifically, a measure of accuracy is usually plotted as a function of the sequence identity of the target-template alignment that was used to calculate the target model (Fig. 1).

Based on such comparisons, sequence-structure relationships are coarsely classified into three different regimes in the sequence similarity spectrum: (i) the easily detected relationships characterized by >30% sequence identity, (ii) the "twilight zone" (Rost, 1999) corresponding to relationships with statistically significant sequence similarity in the 10–30% range, and (iii) the "midnight zone" (Rost, 1999) corresponding to statistically insignificant sequence similarity – the regime where threading methods show the greatest promise.

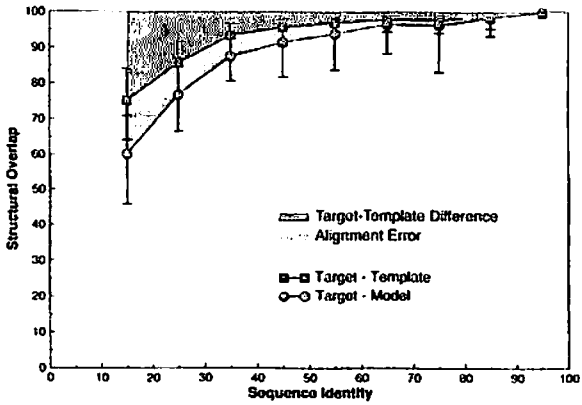

Figure 1. The median accuracy of comparative models plotted as a function of sequence identity. Structural overlap is defined as the fraction of equivalent C^{α} atoms. For the comparison of the model with the native structure (filled circles), two C^{α} atoms were considered equivalent if they belonged to the same residue and were within 3.5 Å of each other after least-squares superposition. For comparisons between the native structure and the template used for modeling (squares), two C^{α} atoms were considered equivalent if they were within 3.5 Å of each other after a structural alignment. The difference between the model and the actual target structure is a combination of the target-template differences (dark gray area) and the alignment errors (light gray area). The lower-panel indicates the most commonly seen model errors. The data was derived by analyzing approximately 1 million models produced by MODPIPE for sequences with known structures.

Models based on alignments with >30% sequence identity almost always have the correct fold. On average, such models also usually have >70–75% of the backbone atoms correctly modeled with a root-mean-squared-deviation (RMSD) of less than 3.5 Å (Fig. 1). However, as the sequence identity drops below 30%, even evolutionarily related proteins tend to show significant differences in their structures. These differences lead to errors in the alignment that, in turn, decrease the accuracy of the resulting model. Nevertheless, state-of-the-art alignment methods, including profile-sequence, profile–profile methods and structure-based environment dependent substitution matrices, have significantly improved the accuracy of such alignments (Shi et al., 2001; Wang and Dunbrack, 2004; Soding, 2005; Zhou and Zhou, 2005; Wu and Zhang, 2008). On average, it is not uncommon for models based on alignments with 20–30% sequence identity to have more than half the backbone atoms modeled accurately. For most alignments below 20% sequence identity, it still remains a challenge to calculate an accurate alignment. Getting a model that is close to the native structure, in this regime of sequence

identity, involves exploring the conformational space without reliance on the alignment. There have been recent reports of success in addressing this problem (Bradley et al., 2005; Misura et al., 2006; Chen and Skolnick, 2008; Zhang, 2008). However, such approaches involve computationally expensive search strategies that prevent their application on a large-scale.

4. Prediction of model accuracy

The accuracy of the predicted model determines the information that can be extracted from it. Thus, estimating the accuracy of a model in the absence of the known structure is essential for its interpretation. As discussed above, a model calculated using a template structure that shares more than 30% sequence identity is indicative of an overall accurate structure (i.e., RMSD of the backbone atoms when compared to the native structure is within ~0.5–3.0 Å).

It is generally useful to assess errors in (i) the choice of template structures, (ii) the alignment, (iii) the modeling of loops, (iv) rigid-body shifts and distortions, and (v) the packing of side-chains. Thus, a number of assessment scores have been developed that specialize in evaluating specific aspects of protein structure models, such as: (i) determining whether or not a model has the correct fold (Tanaka and Scheraga, 1976; Sippl, 1993; Miyazawa and Jernigan, 1996; Domingues et al., 1999; Melo et al., 2002); (ii) discriminating between the native and near-native states (Lazaridis and Karplus, 1999; Gatchell et al., 2000; Vorobjev and Hermans, 2001; Tsai et al., 2003; Zhang et al., 2004; Shen and Sali, 2006); and (iii) selecting the most native-like model in a set of decoys that does not contain the native structure (Shortle et al., 1998; Eramian et al., 2006). Different measures to predict errors in a protein structure perform best at different levels of accuracy. For instance, physics-based force-fields may be helpful at identifying the best model when all models are very close to the native state (<1.5 Å RMSD over all backbone C^{α} atoms, corresponding to ~85% target-template sequence identity). In contrast, coarse-grained scores such as atomic distance-dependent statistical potentials have been shown to have the greatest ability to differentiate between models in the ~3 Å C^{α} RMSD range. Tests show that such scores are often able to identify a model within 0.5 Å C^{α} RMSD of the most accurate model produced (Eramian et al., 2006).

5. Evaluation of protein structure modeling methods

It is crucial for method developers and users alike to assess the accuracy of their methods. An attempt to address this problem has been made by the

CASP (Critical Assessment of Techniques for Proteins Structure Prediction) experiments (Kryshtafovych et al., 2005). These biannual competitions acquire experimentally determined protein structures before they are released to the public and allow participants to predict the structures, which are then evaluated by human experts. However, the major limitation of this competition is that it can assess methods only over a limited number of target protein sequences (Bujnicki et al., 2001; Marti-Renom et al., 2002) and only once every 2 years. To overcome these limitations, two additional evaluation experiments have been described, LiveBench (Bujnicki et al., 2001) and EVA (Eyrich et al., 2001; Koh et al., 2003), which continuously evaluate participating modeling web-servers over a cumulative period of time. For example, the aims of EVA are (i) to evaluate continuously and automatically blind predictions by prediction servers, based on identical and sufficiently large data sets; (ii) to provide weekly updates of the method assessments on the web; and (iii) to enable developers, non-expert users, and reviewers to determine the performance of the tested prediction servers.

6. Genome-scale protein structure modeling and databases

6.1. LARGE-SCALE PROTEIN STRUCTURE MODELING

There are several automated modeling methods, available through the internet, as evidenced by the increasing number of web-servers that participate in the CASP competitions. However, bridging the widening sequence-structure gap requires the development of completely automated, stable, reliable and, most importantly, scalable modeling methods than can be applied to millions of sequences. Currently, there are at least three such large-scale efforts that have been applied to entire genomes, including SWISS-MODEL (Schwede et al., 2003), MODPIPE (Eswar et al., 2003), and FAMS (Takeda-Shitaka et al., 2005). Results of such large-scale calculations indicate that it is currently possible to model at least one domain for over half of all the sequences in most genomes.

6.2. DATABASES OF PROTEIN STRUCTURE MODELS

Depositions to the PDB are restricted to atomic coordinates that are substantially determined by experimental measurements on specimens containing biological macromolecules (Berman et al., 2007). However, as mentioned above, several millions of comparative protein models have been generated for the protein sequences contained in the UniProtKB database using the experimentally determined structures in PDB. These models are disseminated to the community through individual databases such as MODBASE (Pieper

et al. 2006), SWISS-MODEL REPOSITORY (Kopp and Schwede 2006), and FAMSBASE (Yamaguchi et al., 2003). Databases of annotated comparative models increase the efficiency for expert users, allow cross-referencing with other (non-structure-centric) resources, and make comparative models accessible to non-experts.

The Protein Model Portal (http://www.proteinmodelportal.org) has recently been developed as part of the PSI Structural Genomics Knowledge Base to provide an integrated access to the various databases containing structural information and thereby implementing the first step of the community workshop recommendation (Kouranov et al. 2006) on archiving structural models of biological macromolecules. Currently, models calculated by the six structural genomics centers, MODBASE, and SWISS-MODEL Repository are accessible through a single search interface.

7. Integrative or hybrid modeling techniques

Biological function cannot be provided by a single protein molecule in isolation. It is the result of stable or transient interactions among individual proteins and other molecules in the cell. Most of these interactions remain uncharacterized by traditional structural biology techniques such as X-ray crystallography and NMR spectroscopy. This gap is being bridged by several emerging experimental approaches that vary in terms of the information they provide (Robinson et al., 2007). For example, the stoichiometry and composition of protein components in an assembly can be determined by methods such as quantitative immunoblotting and mass spectrometry. The shape of the assembly can be revealed by electron microscopy and small angle X-ray scattering. The positions of the components can be elucidated by cryo-electron microscopy and labeling techniques. Whether or not components interact with each other can be measured by mass spectrometry, yeast two-hybrid and affinity purification. Relative orientations of components and information about interacting residues can be inferred from cryo-electron microscopy, hydrogen/deuterium exchange, hydroxyl radical footprinting, and chemical-crosslinking (Alber et al., 2008) (Fig. 2).

When approaches dominated by a single source of information fail, simultaneous consideration of all available information about the composition and structure of a given protein or assembly, irrespective of its source, can sometimes be sufficient to calculate a useful structural model (Robinson et al., 2007). Even when the model resulting from such integrative or hybrid methods is of relatively low resolution and accuracy, it can still be helpful for studying the function and evolution of the modelled protein or assembly; it also provides the necessary starting point for a higher resolution study.

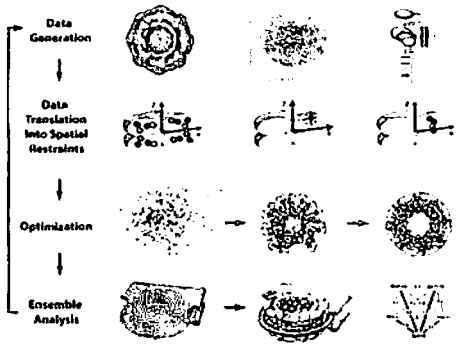

Figure 2. Integrative structure determination. The four steps of determining a structure of a protein or a macromolecular assembly by integration of varied data are illustrated with the example of the nuclear pore complex (Alber et al. 2007a, b; Robinson et al., 2007). First, structural data are generated by experiments, such as electron microscopy (left panel), immuno-electron microscopy (middle panel), and affinity purification of subcomplexes (right panel); many other types of information can also be added. Second, the data and theoretical considerations are expressed as spatial restraints ensuring the observed symmetry and shape of the assembly (electron microscopy, left panel), positions of constituent gold-labeled proteins (immuno-electron microscopy, middle panel), and proximity among the constituent proteins (affinity co-purification, right panel). Third, an ensemble of structural solutions that satisfy the data is obtained by minimizing the violations of the spatial restraints (from left to right). Fourth, the ensemble is clustered into sets of distinct solutions (left panel) as well as analyzed in different representations, such as protein positions (middle panel) and protein-protein contacts (right panel). The integrative approach to structure determination has several advantages: (i) It benefits from the synergy among the input data, minimizing the drawback of incomplete, inaccurate, and/or imprecise data sets (although each individual restraint may contain little structural information, the concurrent satisfaction of all restraints derived from independent experiments may drastically reduce the degeneracy of structural solutions); (ii) it can potentially produce all structures that are consistent with the data, not just one; (iii) the variation among the structures consistent with the data allows us to assess sufficiency of the data and the precision of the representative structure; (iv) it can make the process of structure determination more efficient by indicating what measurements would be the most informative. (This figure was reproduced from figure 5 inRobinson et al. (2007)).

An example of a simple hybrid approach is building a pseudo-atomic model of a large assembly by fitting atomic structures of subunits into its cryo-electron microscopy map (Gao et al., 2003; Chandramouli et al., 2008; Topf et al., 2008). X-ray diffraction data has been combined with protein structure modelling to provide solutions for molecular replacement (Qian

et al., 2007). Unassigned or partially assigned NMR spectroscopy data and fragment-based modeling approaches have been combined to improve structure refinement in terms of its accuracy, efficiency, and success rate (Shen et al., 2008). A variety of different types of information, such as symmetry and protein proximity, have been used to characterize large symmetrical assemblies, including the nuclear pore complex (Alber et al., 2007b), EscJ from the type III secretion system (Andre et al., 2007), and the AAA+ ring complexes (Diemand and Lupas, 2006).

8. Future directions

8.1. PROTEIN STRUCTURE MODELING

Improvement in the accuracy of atomic comparative models will require methods that finely sample protein conformational space using a free energy or scoring function that has sufficient accuracy to distinguish the native structure from the non-native conformations. Despite many years of development of molecular simulation methods, attempts to refine models that are already relatively close to the native structure have met with relatively little success. This failure is likely to be due in part to inaccuracies in the scoring functions used in the simulations, particularly in the treatment of electrostatics and solvation effects. A combination of physics-based energy functions with the statistical information extracted from known protein structures may provide more accurate scoring functions. In addition to the scoring function, improvements in sampling strategies are also likely to be necessary.

8.2. INTEGRATIVE MODELING

Cryo-electron microscopy is emerging as a key technique for studying 3D structures of multi-component macromolecular complexes with masses larger than 250 kDa, such as membrane proteins, cytoskeletal complexes, ribosomes, quasi-spherical viruses, molecular chaperones, flagella, ion channels, and oligomeric enzymes. Electron cryo-tomography even enables the observation of macromolecules inside a living cell in its native state (Baumeister, 2004). Various modeling approaches are being developed that utilize cryo-electron microscopy density maps as a restraint in deriving a pseudo atomic model of the molecular components within a larger complex. Because of the significant likelihood of conformational differences between isolated domains and biological assemblies, additional research resulting in reliable hybrid modeling methods, which are able to correctly include structural information from various experimental sources of different resolution and reliability, is

essential. Structural information from hybrid models, generating a synoptic image of the heterogeneous information available for a given macromolecular system, is expected to increase sharply in the coming years.

ACKNOWLEDGEMENTS

We thank Drs. Ben Webb, Mallur S Madhusudhan, Marc A Marti-Renom, Min-yi Shen, Ursula Pieper and David Eramian for helpful discussions about comparative modeling. This article is partially based on papers by Eswar and Sali (2007) and Schwede et al. (2008). We also acknowledge funds from Sandler Family Supporting Foundation, U.S. National Institutes of Health (Grants R01-GM54762, R01-GM083960, U54-RR022220, U54-GM074945, P01-GM71790, U54-GM074929), U.S. National Science Foundation (Grant IIS-0705196), as well as Hewlett-Packard, Sun Microsystems, IBM, NetApp Inc. and Intel Corporation for hardware gifts.

References

Alber, F., Dokudovskaya, S., Veenhoff, L.M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B.T., et al. 2007a. Determining the architectures of macromolecular assemblies. *Nature* **450**: 683–694.

Alber, F., Dokudovskaya, S., Veenhoff, L.M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B.T., et al. 2007b. The molecular architecture of the nuclear pore complex. *Nature* **450**: 695–701.

Alber, F., Forster, F., Korkin, D., Topf, M., and Sali, A. 2008. Integrating Diverse Data for Structure Determination of Macromolecular Assemblies. *Ann Rev Biochem.*

Andre, I., Bradley, P., Wang, C., and Baker, D. 2007. Prediction of the structure of symmetrical protein assemblies. *Proceedings of the Natl Acad Sci USA* **104**: 17656–17661.

Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J., Chothia, C., and Murzin, A.G. 2004. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* **32**: D226–229.

Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., et al. 2005. The Universal Protein Resource (UniProt). *Nucleic Acids Res* **33**: D154–159.

Baker, D., and Sali, A. 2001. Protein structure prediction and structural genomics. *Science* **294**: 93–96.

Baumeister, W. 2004. Mapping molecular landscapes inside cells. *Biol Chem* **385**: 865–872.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. 2005. GenBank. *Nucleic Acids Res* **33**: D34–38.

Berman, H., Henrick, K., Nakamura, H., and Markley, J.L. 2007. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res* **35**: D301–303.

Bradley, P., Misura, K.M., and Baker, D. 2005. Toward high-resolution de novo structure prediction for small proteins. *Science (New York)* **309**: 1868–1871.

Bru, C., Courcelle, E., Carrere, S., Beausse, Y., Dalmar, S., and Kahn, D. 2005. The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res* **33:** D212–215.

Bujnicki, J.M., Elofsson, A., Fischer, D., and Rychlewski, L. 2001. LiveBench-1: continuous benchmarking of protein structure prediction servers. *Protein Sci* **10:** 352–361.

Chandonia, J.M., and Brenner, S. 2005a. Update on the pfam5000 strategy for selection of structural genomics targets. *Conf Proc IEEE Eng Med Biol Soc* **1:** 751–755.

Chandonia, J.M., and Brenner, S.E. 2005b. Implications of structural genomics target selection strategies: Pfam5000, whole genome, and random approaches. *Proteins* **58:** 166–179.

Chandonia, J.M., and Brenner, S.E. 2006. The impact of structural genomics: expectations and outcomes. *Science (New York)* **311:** 347–351.

Chandramouli, P., Topf, M., Menetret, J.F., Eswar, N., Cannone, J.J., Gutell, R.R., Sali, A., and Akey, C.W. 2008. Structure of the Mammalian 80S Ribosome at 8.7 A Resolution. *Structure* **16:** 535–548.

Chen, H., and Skolnick, J. 2008. M-TASSER: an algorithm for protein quaternary structure prediction. *Biophys J* **94:** 918–928.

Das, R., and Baker, D. 2008. Macromolecular Modeling with Rosetta. *Ann Rev Biochem.*

Diemand, A.V., and Lupas, A.N. 2006. Modeling AAA+ ring complexes from monomeric structures. *J Structur Biol* **156:** 230–243.

Domingues, F.S., Koppensteiner, W.A., Jaritz, M., Prlic, A., Weichenberger, C., Wiederstein, M., Floeckner, H., Lackner, P., and Sippl, M.J. 1999. Sustained performance of knowledge-based potentials in fold recognition. *Proteins* **Suppl 3:** 112–120.

Eramian, D., Shen, M.Y., Devos, D., Melo, F., Sali, A., and Marti-Renom, M.A. 2006. A composite score for predicting errors in protein structure models. *Protein Sci* **15:** 1653–1666.

Eswar, N., and Sali, A. 2007. Comparative modeling of drug target proteins. In *Comprehensive Medicinal Chemistry II.* (ed. J.S. Mason), pp. 215–236. Elsevier, Oxford.

Eswar, N., John, B., Mirkovic, N., Fiser, A., Ilyin, V.A., Pieper, U., Stuart, A.C., Marti-Renom, M.A., Madhusudhan, M.S., Yerkovich, B., et al. 2003. Tools for comparative protein structure modeling and analysis. *Nucleic Acids Res* **31:** 3375–3380.

Eyrich, V.A., Marti-Renom, M.A., Przybylski, D., Madhusudhan, M.S., Fiser, A., Pazos, F., Valencia, A., Sali, A., and Rost, B. 2001. EVA: continuous automatic evaluation of protein structure prediction servers. *Bioinformatics* **17:** 1242–1243.

Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L., et al. 2008. The Pfam protein families database. *Nucleic Acids Res* **36:** D281–288.

Gao, H., Sengupta, J., Valle, M., Korostelev, A., Eswar, N., Stagg, S.M., Van Roey, P., Agrawal, R.K., Harvey, S.C., Sali, A., et al. 2003. Study of the structural dynamics of the E coli 70S ribosome using real-space refinement. *Cell* **113:** 789–801.

Gatchell, D.W., Dennis, S., and Vajda, S. 2000. Discrimination of near-native protein structures from misfolded models by empirical free energy functions. *Proteins* **41:** 518–534.

Godzik, A. 2003. Fold recognition methods. *Methods Biochem Anal* **44:** 525–546.

Koh, I.Y., Eyrich, V.A., Marti-Renom, M.A., Przybylski, D., Madhusudhan, M.S., Eswar, N., Grana, O., Pazos, F., Valencia, A., Sali, A., et al. 2003. EVA: Evaluation of protein structure prediction servers. *Nucleic Acids Res* **31:** 3311–3315.

Kopp, J., and Schwede, T. 2006. The SWISS-MODEL Repository: new features and functionalities. *Nucleic Acids Res* **34:** D315–318.

Kouranov, A., Xie, L., de la Cruz, J., Chen, L., Westbrook, J., Bourne, P.E., and Berman, H.M. 2006. The RCSB PDB information portal for structural genomics. *Nucleic Acids Res* **34:** D302–305.

Kryshtafovych, A., Venclovas, C., Fidelis, K., and Moult, J. 2005. Progress over the first decade of CASP experiments. *Proteins* **61 Suppl 7:** 225–236.

Lazaridis, T., and Karplus, M. 1999. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J Mol Biol* **288:** 477–487.

Letunic, I., Copley, R.R., Pils, B., Pinkert, S., Schultz, J., and Bork, P. 2006. SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res* **34:** D257–260.

Liu, J., Montelione, G.T., and Rost, B. 2007. Novel leverage of structural genomics. *Nature Biotechnol* **25:** 849–851.

Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. 2000. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* **29:** 291–325.

Marti-Renom, M.A., Madhusudhan, M.S., Fiser, A., Rost, B., and Sali, A. 2002. Reliability of assessment of protein structure prediction methods. *Structure (Camb)* **10:** 435–440.

Melo, F., Sanchez, R., and Sali, A. 2002. Statistical potentials for fold assessment. *Protein Sci* **11:** 430–448.

Misura, K.M., Chivian, D., Rohl, C.A., Kim, D.E., and Baker, D. 2006. Physically realistic homology models built with ROSETTA can be more accurate than their templates. *Proc Natl Acad Sci USA* **103:** 5361–5366.

Miyazawa, S., and Jernigan, R.L. 1996. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol* **256:** 623–644.

Moult, J. 2008. Comparative modeling in structural genomics. *Structure* **16:** 14–16.

Pearl, F., Todd, A., Sillitoe, I., Dibley, M., Redfern, O., Lewis, T., Bennett, C., Marsden, R., Grant, A., Lee, D., et al. 2005. The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res* **33:** D247–251.

Pieper, U., Eswar, N., Davis, F.P., Braberg, H., Madhusudhan, M.S., Rossi, A., Marti-Renom, M., Karchin, R., Webb, B.M., Eramian, D., et al. 2006. MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res* **34:** D291–295.

Qian, B., Raman, S., Das, R., Bradley, P., McCoy, A.J., Read, R.J., and Baker, D. 2007. High-resolution structure prediction and the crystallographic phase problem. *Nature* **450:** 259–264.

Robinson, C.V., Sali, A., and Baumeister, W. 2007. The molecular sociology of the cell. *Nature* **450:** 973–982.

Rost, B. 1999. Twilight zone of protein sequence alignments. *Protein Eng* **12:** 85–94.

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

Schwede, T., Kopp, J., Guex, N., and Peitsch, M.C. 2003. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res* **31:** 3381–3385.

Schwede, T., Sali, A., Eswar, N., and Peitsch, M.C. 2008. Protein Structure Modeling. In *Computational Structural Biology*. (eds. T. Schwede, and M.C. Peitsch). World Scientific Publishing, Singapore.

Shen, M.Y., and Sali, A. 2006. Statistical potential for assessment and prediction of protein structures. *Protein Sci* **15:** 2507–2524.

Shen, Y., Lange, O., Delaglio, F., Rossi, P., Aramini, J.M., Liu, G., Eletsky, A., Wu, Y., Singarapu, K.K., Lemak, A., et al. 2008. Consistent blind protein structure generation from NMR chemical shift data. *Proc Natl Acad Sci USA* **105:** 4685–4690.

Shi, J., Blundell, T.L., and Mizuguchi, K. 2001. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol* **310:** 243–257.

Shortle, D., Simons, K.T., and Baker, D. 1998. Clustering of low-energy conformations near the native structures of small proteins. *Proc Natl Acad Sci USA* **95:** 11158–11162.

Simons, K.T., Bonneau, R., Ruczinski, I., and Baker, D. 1999. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins* **Suppl 3:** 171–176.

Sippl, M.J. 1993. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J Comp-Aid Mol Design* **7:** 473–501.

Soding, J. 2005. Protein homology detection by HMM-HMM comparison. *Bioinformatics (Oxford, England)* **21:** 951–960.

Takeda-Shitaka, M., Terashi, G., Takaya, D., Kanou, K., Iwadate, M., and Umeyama, H. 2005. Protein structure prediction in CASP6 using CHIMERA and FAMS. *Proteins* **61 Suppl 7:** 122–127.

Tanaka, S., and Scheraga, H.A. 1976. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules* **9:** 945–950.

Topf, M., Lasker, K., Webb, B., Wolfson, H., Chiu, W., and Sali, A. 2008. Protein structure fitting and refinement guided by cryo-EM density. *Structure* **16:** 295–307.

Tsai, J., Bonneau, R., Morozov, A.V., Kuhlman, B., Rohl, C.A., and Baker, D. 2003. An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins* **53:** 76–87.

Vitkup, D., Melamud, E., Moult, J., and Sander, C. 2001. Completeness in structural genomics. *Nat Struct Biol* **8:** 559–566.

Vorobjev, Y.N., and Hermans, J. 2001. Free energies of protein decoys provide insight into determinants of protein stability. *Protein Sci* **10:** 2498–2506.

Wang, G., and Dunbrack, R.L., Jr. 2004. Scoring profile-to-profile sequence alignments. *Protein Sci* **13:** 1612–1626.

Wu, S., and Zhang, Y. 2008. MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins*.

Yamaguchi, A., Iwadate, M., Suzuki, E., Yura, K., Kawakita, S., Umeyama, H., and Go, M. 2003. Enlarged FAMSBASE: protein 3D structure models of genome sequences for 41 species. *Nucleic Acids Res* **31:** 463–468.

Zhang, Y. 2008. I-TASSER server for protein 3D structure prediction. *BMC Bioinformat* **9:** 40.

Zhang, C., Liu, S., Zhou, H., and Zhou, Y. 2004. An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. *Protein Sci* **13:** 400–411.

Zhou, H., and Zhou, Y. 2005. SPARKS 2 and SP3 servers in CASP6. *Proteins* **61 Suppl 7:** 152–156.