Published in final edited form as:
Methods Mol Biol. 2010 ; 673: 73–94. doi:10.1007/978-1-60761-842-3_6.

Template-Based Protein Structure Modeling

Andras Fiser

Abstract

Functional characterization of a protein is often facilitated by its 3D structure. However, the fraction of experimentally known 3D models is currently less than 1% due to the inherently time-consuming and complicated nature of structure determination techniques. Computational approaches are employed to bridge the gap between the number of known sequences and that of 3D models. Template-based protein structure modeling techniques rely on the study of principles that dictate the 3D structure of natural proteins from the theory of evolution viewpoint. Strategies for template-based structure modeling will be discussed with a focus on comparative modeling, by reviewing techniques available for all the major steps involved in the comparative modeling pipeline.

Keywords

Homology modeling; Comparative protein structure modeling; Template-based modeling; Loop modeling; Side chain modeling; Sequence-to-structure alignment

1. Introduction

The class of methods referred to as template-based modeling includes both the threading techniques that return a full 3D description for the target and comparative modeling (1). This class of protein structure modeling relies on detectable similarity spanning most of the modeled sequence and at least one known structure. Comparative modeling refers to those template-based modeling cases where not only the fold is determined from a possible set of available templates, but a full atom model is also built (2). In practice, it means that if the structure of at least one protein in the family has been determined by experimentation, the other members of the family can be modeled based on their alignment to the known structure. It is possible because a small change in the protein sequence usually results in a small change in its 3D structure (3). It is also facilitated by the fact that 3D structure of proteins from the same family is more conserved than their amino-acid sequences (4). Therefore, if similarity between two proteins is detectable at the sequence level, then structural similarity can usually be assumed. The increasing applicability of template-based modeling is owing to the observation that the number of different folds that proteins adopt is rather limited and because worldwide Structural Genomics projects are aggressively mapping out the universe of possible folds (5–7).

Template-based approaches to structure prediction have their advantages and limitations. Comparative protein structure modeling usually provides high-quality models that are comparable with low-resolution X-ray crystallography or medium-resolution NMR solution structures. However, the applicability of these approaches is limited to those sequences that

can be confidently mapped to known structures. Currently, the probability of finding related proteins of known structure for a sequence picked randomly from a genome ranges approximately from 30 to 80%, depending on the genome. Approximately 70% of all known sequences have at least one domain that is detectably related to at least one protein of known structure (8). This fraction is more than an order of magnitude larger than the number of experimentally determined protein structures deposited in the Protein Data Bank (PDB) (9). As we will see, in practice, template-based modeling always includes information that is independent from the template, in the form of various force restraints from general statistical observations or molecular mechanical force fields. As a consequence of improving force fields and search algorithms, the most successful approaches often explore more and more template-independent conformational space (10, 11).

2. Methods

All current comparative modeling methods consist of five sequential steps: (1) to search for proteins with known 3D structures that are related to the target sequence, (2) to pick those structures that will be used as templates, (3) to align their sequences with the target sequence, (4) to build the model for the target sequence given its alignment with the template structures, and (5) to evaluate the model, using a variety of criteria.

There are several computer programs and web servers that automate the comparative modeling process (Table 1). While the web servers are convenient and useful (10, 12–14), the best results are still obtained by nonautomated, expert use of the various modeling tools (15). Complex decisions for selecting the structurally and biologically most relevant templates, optimally combining multiple template information, refining alignments in nontrivial cases, selecting segments for loop modeling, including cofactors and ligands in the model, or specifying external restraints require an expert knowledge that is difficult to fully automate (16), although more and more efforts on automation point to this direction (17, 18).

2.1. Searching for Structures Related to the Target Sequence

Comparative modeling usually starts by searching the PDB (9) for known protein structures using the target sequence as the query. This search is generally done by comparing the target sequence with the sequence of each of the structures in the database.

There are two main classes of protein comparison methods that are useful in fold identification. The first class compares the sequences of the target with each of the database templates by using pairwise sequence–sequence comparisons (such as FASTA and BLAST (19)) (20–22) and fold assignments (23). To improve the sensitivity of the sequence-based searches, evolutionary information can be incorporated in the form of multiple sequence alignment (24–28). These approaches begin by finding all sequences in a sequence database that are clearly related to the target and easily aligned with it (29, 30). The multiple alignment of these sequences is the target sequence profile, which implicitly carries additional information about the location and pattern of evolutionarily conserved positions of the protein. The most well-known program in this class is PSI-BLAST (27), which implements a heuristic search algorithm for short motifs. A further step to increase the

sensitivity of this approach is to precalculate sequence profiles for all the known structures and then use pairwise dynamic programming algorithm to compare the two profiles. This has been implemented, among other programs, in COACH (31) and FFAS03 (32, 33). The construction of profile-based Hidden Markov Models (HMM) is another sensitive way to locate universally conserved motifs among sequences (34). A substantial improvement in HMM approaches was achieved by incorporating information about predicted secondary structural elements (35, 36). Another development in this group of methods is the phylogenetic tree-driven HMM, which selects a different subset of sequences for profile HMM analysis at each node in the evolutionary tree (37). Locating sequence intermediates that are homologous to both sequences may also enhance the template searches (22, 38). These more sensitive fold identification techniques are especially useful for finding significant structural relationships when sequence identity between the target and the template drops below 25%. More accurate sequence profiles and structural alignments can be constructed with consistency-based approaches such as T-Coffee (39), PROMAL (and PROMAL3D for structures) (40, 41), and ProbCons (42).

The second class of methods relies on pairwise comparison of a protein sequence and a protein structure; the target sequence is matched against a library of 3D profiles or threaded through a library of 3D folds. These methods are also called fold assignment, threading, or 3D template matching (32, 43–47). These methods are especially useful when sequence profiles are not possible to construct because there are not enough known sequences that are clearly related to the target or potential templates.

Template search methods “outperform” the needs of comparative modeling in the sense that they are able to locate sequences that are so remotely related as to render construction of a reliable comparative model impossible. The reason for this is that sequence relationships are often established on short conserved segments, while a successful comparative modeling exercise requires an overall correct alignment for the entire modeled part of the protein.

2.2. Selecting Templates

Once a list of potential templates is obtained using searching methods, it is necessary to select one or more templates that are appropriate for the particular modeling problem. Several factors need to be taken into account when selecting a template.

2.2.1. Considerations in Template Selection—The simplest template selection rule is to select the structure with the highest sequence similarity to the modeled sequence. The construction of a multiple alignment and a phylogenetic tree (48) can help in selecting the template from the subfamily that is closest to the target sequence. The similarity between the “environment” of the template and the environment in which the target needs to be modeled should also be considered. The term “environment” is used here in a broad sense, including everything that is not the protein itself (e.g., solvent, pH, ligands, quaternary interactions). If possible, a template bound to the same or similar ligands as the modeled sequence should generally be used. The quality of the experimentally determined structure is another important factor in template selection. Resolution and R-factor of a crystal structure and the number of restraints per residue for an NMR structure are indicative of their accuracy. The

criteria for selecting templates also depend on the purpose of a comparative model. For example, if a protein–ligand model is to be constructed, the choice of the template that contains a similar ligand is probably more important than the resolution of the template.

2.2.2. Advantage of Using Multiple Templates—It is not necessary to select only one template. In fact, the optimal use of several templates increases the model accuracy (13, 17, 49, 50); however, not all modeling programs are designed to accept more than one template. The benefit of combining multiple template structures can be twofold. First, multiple template structures may be aligned with different domains of the target, with little overlap between them, in which case, the modeling procedure can construct a homology-based model of the whole target sequence. Second, the template structures may be aligned with the same part of the target and build the model on the locally best template.

An elaborate way to select suitable templates is to generate and evaluate models for each candidate template structure and/or their combinations. The optimized all-atom models can then be evaluated by an energy or scoring function, such as the Z-score of PROSA (46) or VERIFY3D (51). These scoring methods are often sufficiently accurate to allow selection of the most accurate of the generated models (52). This trial-and-error approach can be viewed as limited threading (i.e., the target sequence is threaded through similar template structures). However, these approaches are good only at selecting various templates on a global level.

A recently developed method M4T (Multiple Mapping Method with Multiple Templates) selects and combines multiple template structures through an iterative clustering approach that takes into account the “unique” contribution of each template, their sequence similarity among themselves and to the target sequence, and their experimental resolution (13, 17). The resulting models systematically outperformed models that were based on the single best template.

Another important observation from the same study was that below 40% sequence identity, models built using multiple templates are more accurate than those built using a single template only, and this trend is accentuated as one moves into more remote target–template pair cases. Meanwhile, the advantage of using multiple templates gradually disappears above 40% target–template sequence identity cases. This suggests that in this range, the average differences between the template and target structures are smaller than the average differences among alternative template structures that are all highly similar to the target (17).

2.3. Sequence-to-Structure Alignment

To build a model, all comparative modeling programs depend on a list of assumed structural equivalences between the target and template residues. This list is defined by the alignment of the target and template sequences. Many template search methods will produce such an alignment, and these sometimes can directly be used as the input for modeling. Often, however, especially in the difficult cases, this initial alignment is not the optimal target–template alignment. This is because search methods may be tuned for detection of remote relationships, which is often realized on a local motif and not on a full-length, optimal

alignment. Therefore, once the templates are selected, an alignment method should be used to align them with the target sequence. When the target–template sequence identity is lower than 40%, the alignment accuracy becomes the most important factor affecting the quality of the resulting model. A misalignment by only one residue position will result in an error of approximately 4 Å in the model.

2.3.1. Taking Advantage of Structural Information in Alignments—Alignments in comparative modeling represent a unique class because on one side of the alignment there is always a 3D structure, the template. Therefore, alignments can be improved by including structural information from the template. For example, gaps should be avoided in secondary structure elements, in buried regions, or between two residues that are far in space. Some alignment methods take such criteria into account (47, 53, 54).

When multiple template structures are available, a good strategy is to superpose them with each other first, to obtain a multiple structure-based alignment highlighting structurally conserved residues (55–57). In the next step, the target sequence is aligned with this multiple structure-based alignment. The benefits of using multiple structures and multiple sequences are that they provide evolutionary and structural information about the templates, as well as evolutionary information about the target sequence, and they often produce a better alignment for modeling than the pairwise sequence alignment methods (22, 58).

Multiple Mapping Method (MMM) directly relies on information from the 3D structure (14, 59). MMM minimizes alignment errors by selecting and optimally splicing differently aligned fragments from a set of alternative input alignments. This selection is guided by a scoring function that determines the preference of each alternatively aligned fragment of the target sequence in the structural environment of the template. The scoring function has four terms, which are used to assess the compatibility of alternative variable segments in the protein environment:(a) environment specific substitution matrices from FUGUE (47), (b) residue substitution matrix, Blosum (60), (c) A 3D–1D substitution matrix, H3P2, that scores the matches of predicted secondary structure of the target sequence to the observed secondary structures and accessibility types of the template residues (61), and (d) a statistically derived residue–residue contact energy term (62). MMM essentially performs a limited and inverse threading of short fragments: in this exercise the actual question is not the identification of a right fold, but identification of the correct alignment mapping, among many alternatives, for sequence segments that are threaded on the same fold. These local mappings are evaluated in the context of the rest of the model, where alignments provide a consistent solution and framework for the evaluation.

2.4. Model Building

When discussing the model building step within comparative protein structure modeling, it is useful to distinguish two parts: *template-dependent* and *template-independent* modeling. This distinction is necessary because certain parts of the target must be built without the aid of any template. These parts correspond to gaps in the template sequence within the target–template alignment. Modeling of these regions is commonly referred to as loop modeling problem. It is evident that these loops are responsible for the most characteristic differences

between the template and target, and therefore are chiefly responsible for structural and consequently functional differences. In contrast to these loops, the rest of the target, and in particular the conserved core of the fold of the target, is built using information from the template structure.

2.4.1. Template-Dependent Modeling

2.4.1.1. Modeling by Assembly of Rigid Bodies: A comparative model can be assembled from a framework of small number of rigid bodies obtained from the aligned template protein structures (63–65). The approach is based on the natural dissection of the protein structure into conserved core regions, variable loops that connect them, and side chains that decorate the backbone (66). A widely used program in this class is COMPOSER (67). The accuracy of a model can be somewhat increased when more than one template structure is used to construct the framework (68).

2.4.1.2. Modeling by Segment Matching or Coordinate Reconstruction: Comparative models can be constructed by using a subset of atomic positions from template structures as "guiding" positions, such as the Cα atoms, and by identifying and assembling short, all-atom segments that fit these guiding positions. The all-atom segments that fit the guiding positions can be obtained either by scanning all the known protein structures (69, 70) or by a conformational search restrained by an energy function (71, 72) or by a general method for modeling by segment matching (SEGMOD) (73). Even some side-chain modeling methods (74) and the class of loop construction methods based on finding suitable fragments in the database of known structures (75) can be seen as segment matching or coordinate reconstruction methods.

2.4.1.3. Modeling by Satisfaction of Spatial Restraints: The methods in this class begin by generating many constraints or restraints on the structure of the target sequence, using its alignment to related protein structures as a guide in a procedure that is conceptually similar to that used in determination of protein structures from NMR-derived restraints. The restraints are generally obtained by assuming that the corresponding distances between aligned residues in the template and the target structures are similar. These homology-derived restraints are usually supplemented by stereochemical restraints on bond lengths, bond angles, dihedral angles, and nonbonded atom–atom contacts that are obtained from a molecular mechanics force field (76). The model is then derived by minimizing the violations of all the restraints. Comparative modeling by satisfaction of spatial restraints is implemented in the computer program MODELLER (16, 77), currently the most popular comparative protein modeling program. In MODELLER, the various spatial relationships of distances, angles are expressed as conditional probability density functions (pdfs) and can be used directly as spatial restraints. For example, probabilities for different values of the main chain dihedral angles are calculated from the type of residue considered, from the main chain conformation of an equivalent template residue, and from sequence similarity between the two proteins. An important feature of the method is that the forms of spatial restraints were obtained empirically, from a database of protein structure alignments, without any user imposed subjective assumption. Finally, the model is obtained by optimizing the objective function in Cartesian space by the use of the variable target function method (78),

employing methods of conjugate gradients and molecular dynamics with simulated annealing (79).

A similar comprehensive package is NEST that can build a homology model based on single sequence–template alignment or from multiple templates. It can also consider different structures for different parts of the target (55).

2.4.1.4. Combining Alignments, Combining Structures: It is frequently difficult to select the best templates or calculate a good alignment. One way of improving a comparative model in such cases is to proceed with an iteration of template selection, alignment, and model building, guided by model assessment, until no improvement in the model is detected (80, 81). Some of these approaches are automated (55, 82). In one example, this task was achieved by a genetic algorithm protocol that starts with a set of initial alignments and then iterates through realignment, model building, and model assessment to optimize a model assessment score. Comparative models corresponding to various evolving alignments are built and assessed by a variety of criteria, partly depending on an atomic statistical potential. In another approach, a genetic algorithm was applied to automatically combine templates and alignments. A relatively simple structure-dependent scoring function was used to evaluate the sampled combinations (18).

Other attempts to optimize target–template alignments include the Robetta server, where alignments are generated by dynamic programming using a scoring function that combines information on many protein features, including a novel measure of how obligate a sequence region is to the protein fold. By systematically varying the weights on the different features that contribute to the alignment score, very large ensembles of diverse alignments are generated. A variety of approaches to select the best models from the ensemble, including consensus of the alignments, a hydrophobic burial measure, low- and high-resolution energy functions, and combinations of these evaluation methods were explored (83).

Those metaserver approaches that do not simply score and rank alternative models obtained from a variety of methods but further combine them could also be perceived as approaches that explore the alignment and conformational space for a given target sequence (84).

Another alternative for combined servers is provided by M4T. The M4T program automatically identifies the best templates and explores and optimally splices alternative alignments according to its internal scoring function that focuses on the features of the structural environment of each template (17).

2.4.1.5. Metaservers: Metaserver approaches have been developed to take advantage of the variety of other existing programs. Metaservers collect models from alternative methods and either use them for inputs to make new models or look for consensus solutions within them. For instance, FAMS-ACE (85) takes inputs from other servers as starting points for refinement and remodeling after which Verify3D (51) is used to select the most accurate solution. Other consensus approaches include PCONS, a neural network approach that identifies a consensus model by combining information on reliability scores and structural

similarity of models obtained from other techniques (86). 3D-JURY operates along the same idea; its selection is mainly based on the consensus of model structure similarity (87).

2.4.2. Template Independent Modeling: Modeling Loops, Insertions—In comparative modeling, target sequences often have inserted residues relative to the template structures or have regions that are structurally different from the corresponding regions in the templates. Therefore, no structural information about these inserted segments can be extracted from the template structures. These regions frequently correspond to surface loops. Loops often play an important role in defining the functional specificity of a given protein framework, forming the functional, ligand-binding active sites. The accuracy of loop modeling is a major factor determining the usefulness of comparative models in applications such as ligand docking or functional annotation. Loops are generally too short to provide sufficient information about their local fold, and the environment of each loop is uniquely defined by the solvent and the protein that cradles it. In a few rare cases, it was shown that even identical decapeptides in different proteins do not always have the same conformation (88, 89).

There are two main classes of loop modeling methods: (1) the database search approaches and (2) the conformational search approaches (90–92). There are also methods that combine these two approaches (93–95).

2.4.2.1. Fragment-Based Approach to Loop Modeling: Earlier, it was predicted that it is unlikely that structure databanks will ever reach a point when fragment-based approaches become efficient to model loops (96), which resulted in a boost in the development of conformational search approaches from around 2000. However, many details of the fold universe have been explored during the last decade due to the large number of new folds solved experimentally, which had a profound effect on the extent of known structural fragments. Recent analyses showed that loop fragments are not only well represented in current structure databanks, but shorter segments are also possibly completely explored already (97). It was reported that sequence segments up to 10–12 residues had a related (i.e. at least 50% identical) segment in PDB with a known conformation, and despite the six-fold increase in the sequence databank size and the doubling of PDB since 2002, there was not a single unique loop conformation or sequence segment entered in the PDB ever since. Consequently, more recent efforts have been taken to classify loop conformations into more general categories, thus extending the applicability of the database search approach for more cases (98, 99). A recent work described the advantage of using HMM sequence profiles in classifying and predicting loops (100). An another recently published loop prediction approach first predicts conformation for a query loop sequence and then structurally aligns the predicted structural fragments to a set of nonredundant loop structural templates. These sequence–template loop alignments are then quantitatively evaluated with an artificial neural network model trained on a set of predictions with known outcomes (101).

ArchPred (98, 102), currently perhaps the most accurate database loop modeling approach, exploits a hierarchical and multidimensional database that has been set up to classify about 300,000 loop fragments and loop flanking secondary structures. Besides the length of the loops and types of bracing secondary structures, the database is organized along four

internal coordinates, a distance and three types of angles characterizing the geometry of stem regions (103). Candidate fragments are selected from this library by matching the length, the types of bracing secondary structures of the query and by satisfying the geometrical restraints of the stems and subsequently inserted in the query protein framework where their fit is assessed by the root mean squared deviation (RMSD) of stem regions and by the number of rigid body clashes with the environment. In the final step, remaining candidate loops are ranked by a Z-score that combines information on sequence similarity and fit of predicted and observed ϕ/ψ main chain dihedral angle propensities. Confidence Z-score cutoffs are determined for each loop length. A web server implements the method. Predicted segments are returned, or optionally, these can be completed with side-chain reconstruction and subsequently annealed in the environment of the query protein by conjugate gradient minimization.

In summary, the recent reports about the more favorable coverage of loop conformations in the PDB suggest that database approaches are now rather limited by their ability to recognize suitable fragments, and not by the lack of these segments (i.e., sampling), as thought earlier.

2.4.2.2. Ab Initio Modeling of Loops: To overcome the limitations of the database search methods, conformational search methods were developed. There are many such methods, exploiting different protein representations, objective function terms, and optimization or enumeration algorithms. The search strategies include the minimum perturbation method (104), molecular dynamics simulations (92), genetic algorithms (105), Monte Carlo and simulated annealing (106, 107), multiple-copy simultaneous search (108), self-consistent field optimization (109), and an enumeration based on the graph theory (110). Loop prediction by optimization is applicable to both simultaneous modeling of several loops and those loops interacting with ligands, neither of which is straightforward for the database search approaches, where fragments are collected from unrelated structures with different environments.

The MODLOOP module in MODELLER implements the optimization-based approach (111, 112). Loop optimization in MODLOOP relies on conjugate gradients and molecular dynamics with simulated annealing. The pseudoenergy function is a sum of many terms, including some terms from the CHARMM-22 molecular mechanics force field (76) and spatial restraints based on distributions of distances (113, 114) and dihedral angles in known protein structures. The performance of the approach later was further improved by using CHARMM molecular mechanic force field with Generalized Born (GB) solvation potential to rank final conformations (115). Incorporation of solvation terms in the scoring function was a central theme in several other subsequent studies (95, 116–118). Improved loop prediction accuracy resulted from the incorporation of an entropy like term to the scoring function, the “colony energy,” derived from geometrical comparisons and clustering of sampled loop conformations (119, 120). The continuous improvement of scoring functions delivers improved loop modeling methods. Two recent loop modeling procedures have been introduced that are utilizing the effective statistical pair potential that is encoded in DFIRE (121–123). Another method is developed to predict very long loops using the Rosetta approach, essentially performing a mini folding exercise for the loop segments (124). In the

Prime program, large numbers of loops are generated by using a dihedral angle-based building procedure followed by iterative cycles of clustering, side-chain optimization, and complete energy minimization of selected loop structures using a full-atom molecular mechanic force field (OPLS) with implicit solvation model (125).

2.5. Model Evaluation

After a model is built, it is important to check it for possible errors (see Note 1). The quality of a model can be approximately predicted from the sequence similarity between the target and the template and by performing internal and external evaluations.

Sequence identity above 30% is a relatively good predictor of the expected accuracy of a model. If the target–template sequence identity falls below 30%, the sequence identity becomes significantly less reliable as a measure of the expected accuracy of a single model (see Note 2). It is in such cases that model evaluation methods are most informative.

“Internal” evaluation of self-consistency checks whether or not a model satisfies the restraints used to calculate it, including restraints that originate from the template structure or obtained from statistical observations. Assessment of the stereochemistry of a model (e.g., bonds, bond angles, dihedral angles, and nonbonded atom–atom distances) with programs such as PROCHECK (126) and WHATCHECK (127) is an example of internal evaluation. Although errors in stereochemistry are rare and less informative than errors detected by methods for external evaluation, a cluster of stereochemical errors may indicate that the corresponding region also contains other larger errors (e.g., alignment errors).

“External” evaluation relies on information that was not used in the calculation of the model and as a minimum test whether or not a correct template was used. A wrong template can be detected relatively easily with the currently available scoring functions. A more challenging task for the scoring functions is the prediction of unreliable regions in the model. One way to approach this problem is to calculate a “pseudoenergy” profile of a model, such as that produced by PROSA (128) or Verify3D (51). The profile reports the energy for each position in the model. Peaks in the profile frequently correspond to errors in the model. Other recent approaches usually combine a variety of inputs to assess the models, either wholly (129) or locally (130). In benchmarks, the best quality assessor techniques use a simple consensus approach, where reliability of a model is assessed by the agreement among alternative models that are sometimes obtained from a variety of methods (131, 132).

3. Accuracy of Modeling Methods and Typical Errors in Template Based Models

3.1. Accuracy of Methods

An informative way to test protein structure modeling methods, including comparative modeling, is provided by the biannual meetings on Critical Assessment of Techniques for Protein Structure Prediction (CASP) (133). Protein modelers are challenged to model sequences with unknown 3D structure and to submit their models to the organizers before the meeting. At the same time, the 3D 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, although in these exercises it is not trivial to separate the contributions from programs and human expert knowledge. Alternatively a large-scale, continuous, and automated prediction benchmarking experiment is implemented in the program EVA – EValuation of Automatic protein structure prediction (134). Every week EVA submits prereleased PDB sequences to participating modeling servers, collects the results, and provides detailed statistics on secondary structure prediction, fold recognition, comparative modeling, and prediction on 3D contacts. The LiveBench program has implemented its evaluations in a similar spirit (135). After many years of operations, these benchmark platforms are not kept up to date lately, although their service would be essential to keep the user community well informed about latest developments and the best-performing techniques available. A rigorous statistical evaluation (136) of a blind prediction experiment illustrated that the accuracies of the various model-building methods, using segment matching, rigid body assembly, satisfaction of spatial restraints, or any combinations of these are relatively similar when used optimally (137, 138). This also reflects on the fact that such major factors as template selection and alignment accuracy have a large impact on the overall model accuracy, and that the core of protein structures is highly conserved.

3.2. Errors in Comparative Models

The overall accuracy of comparative models spans a wide range. At the low end of the spectrum are the low resolution models whose only essentially correct feature is their fold. At the high end of the spectrum are the models with an accuracy comparable to medium-resolution crystallographic structures (139). Even low-resolution models are often useful to address biological questions because function can many times be predicted from only coarse structural features of a model. The errors in comparative models can be divided into five categories: (1) Errors in side-chain packing, (2) Distortions or shifts of a region that is aligned correctly with the template structures, (3) Distortions or shifts of a region that does not have an equivalent segment in any of the template structures, (4) Distortions or shifts of a region that is aligned incorrectly with the template structures, and (5) A misfolded structure resulting from using an incorrect template. Approximately 90% of the main-chain atoms are likely to be modeled with an RMS error of about 1 Å when the overall sequence identity is above 40% (140). When sequence identity is between 30 and 40%, the structural differences become larger, and the gaps in the alignment are more frequent and longer; misalignments and insertions in the target sequence become the major problems. As a result, the main-chain RMS error rises to about 1.5 Å for about 80% of residues. When sequence identity drops below 30%, the main problem becomes the identification of related templates and their alignment with the sequence to be modeled. In general, it can be expected that about 20% of residues will be misaligned and consequently incorrectly modeled with an error larger than 3 Å, at this level of sequence similarity. To put the errors in comparative models into perspective, we list the differences among structures of the same protein that have been determined experimentally. A 1 Å accuracy of main-chain atom positions corresponds to X-ray structures defined at a low resolution of about 2.5 Å and with an R-factor of about 25% (141), as well as to medium-resolution NMR structures determined

from ten interproton distance restraints per residue. Similarly, differences between the highly refined X-ray and NMR structures of the same protein also tend to be about 1 Å (142). Changes in the environment (e.g., oligomeric state, crystal packing, solvent, ligands) can also have a significant effect on the structure (143). The performance of comparative modeling may sometimes appear overstated because what is usually discussed in the literature are the mean values of backbone deviations. However, individual errors in certain residues essential for the protein function, even in the context of an overall backbone RMSD of less than 1 Å, can still be large enough to prevent reliable conclusions to be drawn regarding mechanism, protein function, or drug design.

Acknowledgments

This review is partially based on our previous publications (1, 144).

References

1. Fiser A. Protein structure modeling in the proteomics era. Expert Rev Proteomics. 2004; 1:97–110. [PubMed: 15966803]
2. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct. 2000; 29:291. [PubMed: 10940251]
3. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J. 1986; 5:823. [PubMed: 3709526]
4. Lesk AM, Chothia C. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. J Mol Biol. 1980; 136:225. [PubMed: 7373651]
5. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res. 2008; 36:D419–D425. [PubMed: 18000004]
6. Chothia C, Gough J, Vogel C, Teichmann SA. Evolution of the protein repertoire. Science. 2003; 300:1701. [PubMed: 12805536]
7. Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, Redfern O, Pearl F, Nambudiry R, Reid A, et al. The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res. 2007; 35:D291–D297. [PubMed: 17135200]
8. Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS, Rossi A, Marti-Renom M, Karchin R, Webb BM, Eramian D, et al. MODBASE: a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res. 2006; 34:D291–D295. [PubMed: 16381869]
9. Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res. 2007; 35:D301–D303. [PubMed: 17142228]
10. Zhang Y. Template-based modeling and free modeling by I-TASSER in CASP7. Proteins. 2007; 69(Suppl 8):108–117. [PubMed: 17894355]
11. Das R, Qian B, Raman S, Vernon R, Thompson J, Bradley P, Khare S, Tyka MD, Bhat D, Chivian D, et al. Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@ home. Proteins. 2007; 69(Suppl 8):118–128. [PubMed: 17894356]
12. Battey JN, Kopp J, Bordoli L, Read RJ, Clarke ND, Schwede T. Automated server predictions in CASP7. Proteins. 2007; 69(Suppl 8):68–82. [PubMed: 17894354]
13. Fernandez-Fuentes N, Madrid-Aliste CJ, Rai BK, Fajardo JE, Fiser A. M4T: a comparative protein structure modeling server. Nucleic Acids Res. 2007; 35:W363–W368. [PubMed: 17517764]
14. Rai BK, Madrid-Aliste CJ, Fajardo JE, Fiser A. MMM: a sequence-to-structure alignment protocol. Bioinformatics. 2006; 22:2691–2692. [PubMed: 16928737]

15. Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T. Assessment of CASP7 predictions for template-based modeling targets. Proteins. 2007; 69(Suppl 8):38–56. [PubMed: 17894352]
16. Fiser A, Sali A. Modeller: generation and refinement of homology-based protein structure models. Methods Enzymol. 2003; 374:461. [PubMed: 14696385]
17. Fernandez-Fuentes N, Rai BK, Madrid-Aliste CJ, Fajardo JE, Fiser A. Comparative protein structure modeling by combining multiple templates and optimizing sequence-to-structure alignments. Bioinformatics. 2007; 23:2558–2565. [PubMed: 17823132]
18. Contreras-Moreira B, Fitzjohn PW, Offman M, Smith GR, Bates PA. Novel use of a genetic algorithm for protein structure prediction: searching template and sequence alignment space. Proteins. 2003; 53(Suppl 6):424. [PubMed: 14579331]
19. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res. 2001; 29:2994. [PubMed: 11452024]
20. Apostolico A, Giancarlo R. Sequence alignment in molecular biology. J Comput Biol. 1998; 5:173. [PubMed: 9672827]
21. Pearson WR. Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol. 2000; 132:185. [PubMed: 10547837]
22. Sauder JM, Arthur JW, Dunbrack RL Jr. Large-scale comparison of protein sequence alignment algorithms with structure alignments. Proteins. 2000; 40:6. [PubMed: 10813826]
23. Brenner SE, Chothia C, Hubbard TJ. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci U S A. 1998; 95:6073. [PubMed: 9600919]
24. Rychlewski L, Jaroszewski L, Li W, Godzik A. Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Sci. 2000; 9:232. [PubMed: 10716175]
25. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. Hidden Markov models incomputational biology. Applications to protein modeling. J Mol Biol. 1994; 235:1501. [PubMed: 8107089]
26. Henikoff JG, Pietrokovski S, McCallum CM, Henikoff S. Blocks-based methods for detecting protein homology. Electrophoresis. 2000; 21:1700. [PubMed: 10870957]
27. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389. [PubMed: 9254694]
28. Marti-Renom MA, Madhusudhan MS, Sali A. Alignment of protein sequences by their profiles. Protein Sci. 2004; 13:1071. [PubMed: 15044736]
29. Notredame C. Recent evolutions of multiple sequence alignment algorithms. PLoS Comput Biol. 2007; 3:e123. [PubMed: 17784778]
30. Edgar RC, Batzoglou S. Multiple sequence alignment. Curr Opin Struct Biol. 2006; 16:368–373. [PubMed: 16679011]
31. Edgar RC, Sjolander K. COACH: profile–profile alignment of protein families using hidden Markov models. Bioinformatics. 2004; 20:1309. [PubMed: 14962937]
32. Jaroszewski L, Rychlewski L, Zhang B, Godzik A. Fold prediction by a hierarchy of sequence, threading, and modeling methods. Protein Sci. 1998; 7:1431. [PubMed: 9655348]
33. Jaroszewski L, Rychlewski L, Li Z, Li W, Godzik A. FFAS03: a server for profile–profile sequence alignments. Nucleic Acids Res. 2005; 33:W284–W288. [PubMed: 15980471]
34. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics. 1998; 14:846. [PubMed: 9927713]
35. Karchin R, Cline M, Mandel-Gutfreund Y, Karplus K. Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. Proteins. 2003; 51:504. [PubMed: 12784210]
36. Karplus K, Katzman S, Shackleford G, Koeva M, Draper J, Barnes B, Soriano M, Hughey R. SAM-T04: what is new in protein-structure prediction for CASP6. Proteins. 2005; 61(Suppl 7): 135–142. [PubMed: 16187355]
37. Edgar RC, Sjolander K. SATCHMO: sequence alignment and tree construction using hidden Markov models. Bioinformatics. 2003; 19:1404. [PubMed: 12874053]

38. John B, Sali A. Detection of homologous proteins by an intermediate sequence search. Protein Sci. 2004; 13:54. [PubMed: 14691221]
39. Moretti S, Armougom F, Wallace IM, Higgins DG, Jongeneel CV, Notredame C. The M-Coffee web server: a meta-method for computing multiple sequence alignments by combining alternative alignment methods. Nucleic Acids Res. 2007; 35:W645–W648. [PubMed: 17526519]
40. Pei J, Kim BH, Grishin NV. PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res. 2008; 36:2295–2300. [PubMed: 18287115]
41. Pei J, Grishin NV. PROMALS: towards accurate multiple sequence alignments of distantly related proteins. Bioinformatics. 2007; 23:802–808. [PubMed: 17267437]
42. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S. ProbCons: probabilistic consistency-based multiple sequence alignment. Genome Res. 2005; 15:330. [PubMed: 15687296]
43. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol. 1999; 287:797. [PubMed: 10191147]
44. Finkelstein AV, Reva BA. A search for the most stable folds of protein chains. Nature. 1991; 351:497. [PubMed: 2046752]
45. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science. 1991; 253:164. [PubMed: 1853201]
46. Sippl MJ. Knowledge-based potentials for proteins. Curr Opin Struct Biol. 1995; 5:229. [PubMed: 7648326]
47. Shi J, Blundell TL, Mizuguchi K. FUGUE: sequence–structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol. 2001; 310:243. [PubMed: 11419950]
48. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 1981; 17:368. [PubMed: 7288891]
49. Venclovas C, Margelevicius M. Comparative modeling in CASP6 using consensus approach to template selection, sequence–structure alignment, and structure assessment. Proteins. 2005; 61:99–105. [PubMed: 16187350]
50. Sanchez R, Sali A. Evaluation of comparative protein structure modeling by MODELLER-3. Proteins. 1997; 1(Suppl, 50)
51. Eisenberg D, Luthy R, Bowie JU. VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzymol. 1997; 277:396. [PubMed: 9379925]
52. Wu G, McArthur AG, Fiser A, Sali A, Sogin ML, Mllerm M. Core histones of the amitochondriate protist, Giardia lamblia. Mol Biol Evol. 2000; 17:1156. [PubMed: 10908635]
53. Jennings AJ, Edge CM, Sternberg MJ. An approach to improving multiple alignments of protein sequences using predicted secondary structure. Protein Eng. 2001; 14:227. [PubMed: 11391014]
54. Blake JD, Cohen FE. Pairwise sequence alignment below the twilight zone. J Mol Biol. 2001; 307:721. [PubMed: 11254392]
55. Petrey D, Xiang Z, Tang CL, Xie L, Gimpelev M, Mitros T, Soto CS, Goldsmith-Fischman S, Kernytsky A, Schlessinger A, et al. Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. Proteins. 2003; 53(Suppl 6):430. [PubMed: 14579332]
56. Al Lazikani B, Sheinerman FB, Honig B. Combining multiple structure and sequence alignments to improve sequence detection and alignment: application to the SH2 domains of Janus kinases. Proc Natl Acad Sci U S A. 2001; 98:14796. [PubMed: 11752426]
57. Reddy BV, Li WW, Shindyalov IN, Bourne PE. Conserved key amino acid positions (CKAAPs) derived from the analysis of common substructures in proteins. Proteins. 2001; 42:148. [PubMed: 11119639]
58. Jaroszewski L, Rychlewski L, Godzik A. Improving the quality of twilight-zone alignments. Protein Sci. 2000; 9:1487. [PubMed: 10975570]
59. Rai BK, Fiser A. Multiple mapping method: a novel approach to the sequence-to-structure alignment problem in comparative protein structure modeling. Proteins. 2006; 63:644–661. [PubMed: 16437570]

60. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 1992; 89:10915–10919. [PubMed: 1438297]
61. Luthy R, McLachlan AD, Eisenberg D. Secondary structure-based pro-files: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. Proteins. 1991; 10:229–239. [PubMed: 1881879]
62. Rykunov D, Fiser A. Effects of amino acid composition, finite size of proteins, and sparse statistics on distance-dependent statistical pair potentials. Proteins. 2007; 67:559–568. [PubMed: 17335003]
63. Blundell TL, Sibanda BL, Sternberg MJ, Thornton JM. Knowledge-based prediction of protein structures and the design of novel molecules. Nature. 1987; 326:347. [PubMed: 3550471]
64. Browne WJ, North ACT, Phillips DC, Brew K, Vanaman TC, Hill RC. A possible three-dimensional structure of bovine lactalbumin based on that of hen's egg-white lysosyme. J Mol Biol. 1969; 42:65. [PubMed: 5817651]
65. Greer J. Comparative modeling methods: application to the family of the mammalian serine proteases. Proteins. 1990; 7:317. [PubMed: 2381905]
66. Topham CM, McLeod A, Eisenmenger F, Overington JP, Johnson MS, Blundell TL. Fragment ranking in modelling of protein structure. Conformationally constrained environmental amino acid substitution tables. J Mol Biol. 1993; 229:194. [PubMed: 8421300]
67. Sutcliffe MJ, Haneef I, Carney D, Blundell TL. Knowledge based modelling of homologous proteins, part I: three-dimensional frameworks derived from the simultaneous superposition of multiple structures. Protein Eng. 1987; 1:377. [PubMed: 3508286]
68. Srinivasan N, Blundell TL. An evaluation of the performance of an automated procedure for comparative modelling of protein tertiary structure. Protein Eng. 1993; 6:501. [PubMed: 8415577]
69. Claessens M, Van Cutsem E, Lasters I, Wodak S. Modelling the poly-peptide backbone with 'spare parts' from known protein structures. Protein Eng. 1989; 2:335. [PubMed: 2928296]
70. Holm L, Sander C. Database algorithm for generating protein backbone and side-chain co-ordinates from a C alpha trace application to model building and detection of co-ordinate errors. J Mol Biol. 1991; 218:183. [PubMed: 2002501]
71. Bruccoleri RE, Karplus M. Conformational sampling using high- temperature molecular dynamics. Biopolymers. 1990; 29:1847. [PubMed: 2207289]
72. van Gelder CW, Leusen FJ, Leunissen JA, Noordik JH. A molecular dynamics approach for the generation of complete protein structures from limited coordinate data. Proteins. 1994; 18:174. [PubMed: 8159666]
73. Levitt M. Accurate modeling of protein conformation by automatic segment matching. J Mol Biol. 1992; 226:507. [PubMed: 1640463]
74. Chinea G, Padron G, Hooft RW, Sander C, Vriend G. The use of position-specific rotamers in model building by homology. Proteins. 1995; 23:415. [PubMed: 8710834]
75. Jones TA, Thirup S. Using known substructures in protein model building and crystallography. EMBO J. 1986; 5:819. [PubMed: 3709525]
76. Brooks CL III, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy minimization and dynamics calculations. J Comput Chem. 1983; 4:187.
77. Sali A, Blundell TL. Comparative protein modeling by satisfaction of spatial restraints. J Mol Biol. 1993; 234:779–815. [PubMed: 8254673]
78. Braun W, Go N. Calculation of protein conformations by proton–proton distance constraints. A new efficient algorithm. J Mol Biol. 1985; 186:611. [PubMed: 2419572]
79. Clore GM, Brunger AT, Karplus M, Gronenborn AM. Application of molecular dynamics with interproton distance restraints to three-dimensional protein structure determination. A model study of crambin. J Mol Biol. 1986; 191:523. [PubMed: 3029386]
80. Guenther B, Onrust R, Sali A, O'Donnell M, Kuriyan J. Crystal structure of the ë-subunit of the clamp-loader complex of *E. coli* DNA polymerase III. Cell. 1997; 91:335. [PubMed: 9363942]
81. Fiser A, Filipe SR, Tomasz A. Cell wall branches, penicillin resistance and the secrets of the MurM protein. Trends Microbiol. 2003; 11:547. [PubMed: 14659686]

82. John B, Sali A. Comparative protein structure modeling by iterative alignment, model building and model assessment. Nucleic Acids Res. 2003; 31:3982. [PubMed: 12853614]
83. Chivian D, Baker D. Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. Nucleic Acids Res. 2006; 34:e112. [PubMed: 16971460]
84. Kolinski A, Bujnicki JM. Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. Proteins. 2005; 61(Suppl 7):84–90. [PubMed: 16187348]
85. Terashi G, Takeda-Shitaka M, Kanou K, Iwadate M, Takaya D, Hosoi A, Ohta K, Umeyama H. Fams-ace: a combined method to select the best model after remodeling all server models. Proteins. 2007; 69(Suppl 8):98–107. [PubMed: 17894329]
86. Wallner B, Larsson P, Elofsson A. Pcons.net: protein structure prediction meta server. Nucleic Acids Res. 2007; 35:W369–W374. [PubMed: 17584798]
87. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics. 2003; 19:1015–1018. [PubMed: 12761065]
88. Mezei M. Chameleon sequences in the PDB. Protein Eng. 1998; 11:411. [PubMed: 9725618]
89. Fernandez-Fuentes N, Fiser A. Saturating representation of loop conformational fragments in structure databanks. BMC Struct Biol. 2006; 6:15. [PubMed: 16820050]
90. Shenkin PS, Yarmush DL, Fine RM, Wang HJ, Levinthal C. Predicting antibody hypervariable loop conformation. I. Ensembles of random conformations for ringlike structures. Biopolymers. 1987; 26:2053. [PubMed: 3435744]
91. Moult J, James MN. An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. Proteins. 1986; 1:146. [PubMed: 3130622]
92. Bruccoleri RE, Karplus M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. Biopolymers. 1987; 26:137. [PubMed: 3801593]
93. Deane CM, Blundell TL. CODA: a combined algorithm for predicting the structurally variable regions of protein models. Protein Sci. 2001; 10:599. [PubMed: 11344328]
94. van Vlijmen HW, Karplus M. PDB-based protein loop prediction: parameters for selection and methods for optimization. J Mol Biol. 1997; 267:975. [PubMed: 9135125]
95. de Bakker PI, DePristo MA, Burke DF, Blundell TL. Ab initio construction of polypeptide fragments: accuracy of loop decoy discrimination by an all-atom statistical potential and the AMBER force field with the Generalized Born solvation model. Proteins. 2003; 51:21. [PubMed: 12596261]
96. Fidelis K, Stern PS, Bacon D, Moult J. Comparison of systematic search and database methods for constructing segments of protein structure. Protein Eng. 1994; 7:953. [PubMed: 7809034]
97. Du P, Andrec M, Levy RM. Have we seen all structures corresponding to short protein fragments in the Protein Data Bank? An update. Protein Eng. 2003; 16:407. [PubMed: 12874373]
98. Fernandez-Fuentes N, Oliva B, Fiser A. A supersecondary structure library and search algorithm for modeling loops in protein structures. Nucleic Acids Res. 2006; 34:2085–2097. [PubMed: 16617149]
99. Michalsky E, Goede A, Preissner R. Loops in proteins (LIP) – a comprehensive loop database for homology modelling. Protein Eng. 2003; 16:979. [PubMed: 14983078]
100. Espadaler J, Fernandez-Fuentes N, Hermoso A, Querol E, Aviles FX, Sternberg MJ, Oliva B. ArchDB: automated protein loop classification as a tool for structural genomics. Nucleic Acids Res. 2004; 32:D185. Database issue. [PubMed: 14681390]
101. Peng HP, Yang AS. Modeling protein loops with knowledge-based prediction of sequence–structure alignment. Bioinformatics. 2007; 23:2836–2842. [PubMed: 17827204]
102. Fernandez-Fuentes N, Zhai J, Fiser A. ArchPRED: a template based loop structure prediction server. Nucleic Acids Res. 2006; 34:W173–W176. [PubMed: 16844985]
103. Oliva B, Bates PA, Querol E, Aviles FX, Sternberg MJ. An automated classification of the structure of protein loops. J Mol Biol. 1997; 266:814. [PubMed: 9102471]

104. Fine RM, Wang H, Shenkin PS, Yarmush DL, Levinthal C. Predicting antibody hypervariable loop conformations. II: minimization and molecular dynamics studies of MCPC603 from many randomly generated loop conformations. Proteins. 1986; 1:342. [PubMed: 3449860]
105. Ring CS, Cohen FE. Modeling protein structures: construction and their applications. FASEB J. 1993; 7:783. [PubMed: 8330685]
106. Abagyan R, Totrov M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. J Mol Biol. 1994; 235:983. [PubMed: 8289329]
107. Collura V, Higo J, Garnier J. Modeling of protein loops by simulated annealing. Protein Sci. 1993; 2:1502. [PubMed: 8401234]
108. Zheng Q, Rosenfeld R, Vajda S, DeLisi C. Determining protein loop conformation using scaling-relaxation techniques. Protein Sci. 1993; 2:1242. [PubMed: 8401209]
109. Koehl P, Delarue M. A self consistent mean field approach to simultaneous gap closure and side-chain positioning in homology modelling. Nat Struct Biol. 1995; 2:163. [PubMed: 7538429]
110. Samudrala R, Moult J. A graph-theoretic algorithm for comparative modeling of protein structure. J Mol Biol. 1998; 279:287. [PubMed: 9636717]
111. Fiser A, Sali A. ModLoop: automated modeling of loops in protein structures. Bioinformatics. 2003; 19:2500. [PubMed: 14668246]
112. Fiser A, Do RK, Sali A. Modeling of loops in protein structures. Protein Sci. 2000; 9:1753. [PubMed: 11045621]
113. Sippl MJ. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol. 1990; 213:859. [PubMed: 2359125]
114. Melo F, Feytmans E. Novel knowledge-based mean force potential at atomic level. J Mol Biol. 1997; 267:207. [PubMed: 9096219]
115. Fiser A, Feig M, Brooks CL III, Sali A. Evolution and physics in comparative protein structure modeling. Acc Chem Res. 2002; 35:413. [PubMed: 12069626]
116. Das B, Meirovitch H. Solvation parameters for predicting the structure of surface loops in proteins: transferability and entropic effects. Proteins. 2003; 51:470. [PubMed: 12696057]
117. Forrest LR, Woolf TB. Discrimination of native loop conformations in membrane proteins: decoy library design and evaluation of effective energy scoring functions. Proteins. 2003; 52:492. [PubMed: 12910450]
118. DePristo MA, de Bakker PI, Lovell SC, Blundell TL. Ab initio construction of polypeptide fragments: efficient generation of accurate, representative ensembles. Proteins. 2003; 51:41. [PubMed: 12596262]
119. Xiang Z, Soto CS, Honig B. Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. Proc Natl Acad Sci U S A. 2002; 99:7432–7437. [PubMed: 12032300]
120. Fogolari F, Tosatto SC. Application of MM/PBSA colony free energy to loop decoy discrimination: toward correlation between energy and root mean square deviation. Protein Sci. 2005; 14:889–901. [PubMed: 15772305]
121. Soto CS, Fasnacht M, Zhu J, Forrest L, Honig B. Loop modeling: sampling, filtering, and scoring. Proteins. 2007; 70:834–843. [PubMed: 17729286]
122. Zhang C, Liu S, Zhou Y. Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential. Protein Sci. 2004; 13:391–399. [PubMed: 14739324]
123. Soto CS, Fasnacht M, Zhu J, Forrest L, Honig B. Loop modeling: Sampling, filtering, and scoring. Proteins. 2008; 70:834–843. [PubMed: 17729286]
124. Rohl CA, Strauss CE, Chivian D, Baker D. Modeling structurally variable regions in homologous proteins with rosetta. Proteins. 2004; 55:656–677. [PubMed: 15103629]
125. Jacobson MP, Pincus DL, Rapp CS, Day TJ, Honig B, Shaw DE, Friesner RA. A hierarchical approach to all-atom protein loop prediction. Proteins. 2004; 55:351. [PubMed: 15048827]
126. Laskowski RA, Moss DS, Thornton JM. Main-chain bond lengths and bond angles in protein structures. J Mol Biol. 1993; 231:1049. [PubMed: 8515464]

127. Hooft RW, Vriend G, Sander C, Abola EE. Errors in protein structures. Nature. 1996; 381:272. [PubMed: 8692262]
128. Sippl MJ. Recognition of errors in three-dimensional structures of proteins. Proteins. 1993; 17:355. [PubMed: 8108378]
129. Eramian D, Shen MY, Devos D, Melo F, Sali A, Marti-Renom MA. A composite score for predicting errors in protein structure models. Protein Sci. 2006; 15:1653–1666. [PubMed: 16751606]
130. Fasnacht M, Zhu J, Honig B. Local quality assessment in homology models using statistical potentials and support vector machines. Protein Sci. 2007; 16:1557–1568. [PubMed: 17600147]
131. Wallner B, Elofsson A. Prediction of global and local model quality in CASP7 using Pcons and ProQ. Proteins. 2007; 69(Suppl 8):184–193. [PubMed: 17894353]
132. Wallner B, Elofsson A. Pcons5: combining consensus, structural evaluation and fold recognition scores. Bioinformatics. 2005; 21:4248–4254. [PubMed: 16204344]
133. Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol. 2005; 15:285–289. [PubMed: 15939584]
134. Eyrich VA, Marti-Renom MA, Przybylski D, Madhusudhan MS, Fiser A, Pazos F, Valencia A, Sali A, Rost B. EVA: continuous automatic evaluation of protein structure prediction servers. Bioinformatics. 2001; 17:1242. [PubMed: 11751240]
135. Bujnicki JM, Elofsson A, Fischer D, Rychlewski L. LiveBench-1: continuous benchmarking of protein structure prediction servers. Protein Sci. 2001; 10:352. [PubMed: 11266621]
136. Marti-Renom MA, Madhusudhan MS, Fiser A, Rost B, Sali A. Reliability of assessment of protein structure prediction methods. Structure (Camb). 2002; 10:435. [PubMed: 12005441]
137. Wallner B, Elofsson A. All are not equal: a benchmark of different homology modeling programs. Protein Sci. 2005; 14:1315–1327. [PubMed: 15840834]
138. Dalton JA, Jackson RM. An evaluation of automated homology modelling methods at low target template sequence similarity. Bioinformatics. 2007; 23:1901–1908. [PubMed: 17510171]
139. Baker D, Sali A. Protein structure prediction and structural genomics. Science. 2001; 294:93–96. [PubMed: 11588250]
140. Sanchez R, Sali A. Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. Proc Natl Acad Sci U S A. 1998; 95:13597. [PubMed: 9811845]
141. Ohlendorf DH. Accuracy of refined protein structures. Comparison of four independently refined models of human interleukin 1 beta. Acta Crystallogr D Biol Crystallogr. 1994; D50:808. [PubMed: 15299347]
142. Clore GM, Robien MA, Gronenborn AM. Exploring the limits of precision and accuracy of protein structures determined by nuclear magnetic resonance spectroscopy. J Mol Biol. 1993; 231:82. [PubMed: 8496968]
143. Faber HR, Matthews BW. A mutant T4 lysozyme displays five different crystal conformations. Nature. 1990; 348:263. [PubMed: 2234094]
144. Fiser, A. From Protein Structure to Function with Bioinformatics. Ridgen, DJ., editor. Springer; 2008. p. 57-81.

Table 1

Names and www addresses of some online tools useful for various aspects of comparative modeling

Template search and alignments	
BLAST/PSI-BLAST	http://www.ncbi.nlm.nih.gov/BLAST/
FastA/SSEARCH	http://www.ebi.ac.uk/fasta33
FASS03	http://www.ffas.ljcrf.edu/ffas-cgi/cgi/ffas.pl
PSIPRED	http://www.bioinf.cs.ucl.ac.uk/psipred/
123D	http://www.123d.ncifcrf.gov
UCLA-DOE	http://www.doe-mbi.ucla.edu/Services/FOLD/
PHYRE/3D-PSSM	http://www.sbg.bio.ic.ac.uk/~3dpssm
FUGUE	http://www.cryst.bioc.cam.ac.uk/~fugue
LOOPP	http://www.cbsuapps.tc.cornell.edu/
MUSTER	http://www.zhang.bioinformatics.ku.edu/MUSTER/
SAM-T06	http://www.soe.ucsc.edu/research/compbio/SAM_T06/T06-query.html
Prospect	http://www.compbio.ornl.gov/structure/prospect
Smith–Waterman	http://www.jaligner.sourceforge.net/
ClustalW	http://www.ebi.ac.uk/clustalw/
MUSCLE	http://www.drive5.com/lobster/
T-COFFEE	http://www.tcoffee.vital-it.ch/
PROMALS	http://www.prodata.swmed.edu/promals/promals.php
PROBCONS	http://www.probcons.stanford.edu
Homology modeling, loop and side-chain modeling	
MMM	http://www.fiserlab.org/servers/MMM
M4T	http://www.fiserlab.org/servers/M4T
MODELLER	http://www.salilab.org/modeller/modeller.html
MODWEB	http://www.modbase.compbio.ucsf.edu/ModWeb20-html/modweb.html
I-TASSER	http://www.zhang.bioinformatics.ku.edu/I-TASSER/
HHPRED	http://www.toolkit.tuebingen.mpg.de/hhpred
3D-JIGSAW	http://www.bmm.icnet.uk/servers/3djigsaw/
CPH-MODELS	http://www.cbs.dtu.dk/services/CPHmodels/
COMPOSER	http://www.cryst.bioc.cam.ac.uk
SWISSMODEL	http://swissmodel.expasy.org/workspace/
FAMS	http://www.pharm.kitasato-u.ac.jp/fams/
WHATIF	http://www.cmbi.kun.nl/whatif/
PUDGE	http://www.wiki.c2b2.columbia.edu/honiglab_public/index.php/Software
3D-JURY	http://www.meta.bioinfo.pl
RAPPER	http://www.mordred.bioc.cam.ac.uk/~rapper
ESYPRED3D	http://www.fundp.ac.be/sciences/biologie/urbm/bioinfo/esypred/
CONSENSUS	http://www.structure.bu.edu/cgi-bin/consensus/consensus.cgi
PCONS	http://www.pcons.net

SCWRL	http://www.dunbrack.fccc.edu/SCWRL3.php
WLOOP	http://www.bioserv.rpbs.jussieu.fr/cgi-bin/WLoop
ARCHPRED	http://www.fiserlab.org/servers/archpred
MODLOOP	http://www.salilab.org/modloop
Model evaluation	
PROCHECK	http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html
WHATCHECK	http://www.swift.cmbi.ru.nl/gv/whatcheck/
Prosa-web	http://www.prosa.services.came.sbg.ac.at/prosa.php
VERIFY3D	http://www.nihserver.mbi.ucla.edu/Verify_3D
ANOLEA	http://www.protein.bio.puc.cl/cardex/servers/anolea/
AQUA	http://www.urchin.bmrb.wisc.edu/~jurgen/Aqua/server/
PROQ	http://www.sbc.su.se/~bjornw/ProQ/ProQ.cgi