Published in final edited form as:
Curr Protein Pept Sci. 2006 June ; 7(3): 217–227.

Advances in Homology Protein Structure Modeling

Zhexin Xiang*
Center for Molecular Modeling, Center for Information Technology, National Institutes of Health, Building 12A Room 2051, 12 South Drive, Bethesda, Maryland 20892-5624, USA

Abstract

Homology modeling plays a central role in determining protein structure in the structural genomics project. The importance of homology modeling has been steadily increasing because of the large gap that exists between the overwhelming number of available protein sequences and experimentally solved protein structures, and also, more importantly, because of the increasing reliability and accuracy of the method. In fact, a protein sequence with over 30% identity to a known structure can often be predicted with an accuracy equivalent to a low-resolution X-ray structure. The recent advances in homology modeling, especially in detecting distant homologues, aligning sequences with template structures, modeling of loops and side chains, as well as detecting errors in a model, have contributed to reliable prediction of protein structure, which was not possible even several years ago. The ongoing efforts in solving protein structures, which can be time-consuming and often difficult, will continue to spur the development of a host of new computational methods that can fill in the gap and further contribute to understanding the relationship between protein structure and function.

Keywords

homology modeling; structure refinement; sequence alignment; side-chain prediction; loop prediction; model assessment; colony energy; conformation sampling; structural genomics

1. INTRODUCTION

Understanding the mechanism of protein function generally requires knowledge of protein three-dimensional structure [1–2], which is ultimately determined by protein sequence [3]. Protein structure determination using experimental methods such as X-ray crystallography or NMR spectroscopy is time consuming and not successful with all proteins, especially membrane proteins [4]. Currently there are about 2 million protein sequences in Swissprot and TrEMBL (http://us.expasy.org/sprot/), among them about 30,000 proteins have had their structures solved experimentally (http://www.rcsb.org/pdb/). Although the rate of experimental structure determination will continue to increase, the number of newly discovered sequences grows much faster than the number of structures solved.

The huge gap between the number of available sequences and experimentally solved protein structures could possibly be resolved by computational methods. Theoretical structure prediction can be divided into two extreme camps: ab–initio method [5–7] and homology modeling [8–9]. The first approach has fold prediction from physical chemistry principles as one of its goals. The second method predicts the three-dimensional structure of a given protein sequence based primarily on its sequence similarity to one or more proteins of known structures. Fold recognition, an approach between the two extremes, has become an important

*Address correspondence to this author at the Center for Molecular Modeling, Center for Information Technology, National Institutes of Health, Building 12A Room 2051, 12 South Drive, Bethesda, Maryland 20892-5624, USA; E-mail: xiangz@mail.nih.gov.

tool that supplements sequence-based methods to detect remote homologs [10]. Actually it could be considered as a combination of the two because it samples the known protein conformations in PDB with scoring energies [11]. In the early 1990s, fold recognition was considered as a separate field. The recent CASPs (Critical Assessment of Structure Prediction) experiments arguably demonstrate that fold recognition has almost ceased to be a major problem, and thus, it is often considered as part of hard homology modeling [12]. The progress has come, mainly, from increased databases, profile analysis and better understanding of energetic determinants of protein stability [13].

For decades, structure prediction has fascinated the scientific community; it is an extremely important problem that is simple to define but difficult to solve. Although ab-initio methods have achieved astonishing progress in the recent years, it is still unrealistic for reliable application in the years to come [12]. The bottleneck is mainly due to the inaccuracy of force field and intractable enormous conformation sampling. The free energy difference between folding and unfolding is just several kcal/mol, equivalent to several atomic van der Waals interactions [14–15], which poses a daunting challenge to any existing energy functions. Furthermore, the inaccuracy of force field makes conformation sampling much more difficult, since it could guide energy minimization to the conformation of global energy minimum that is totally different from the true conformation. On the contrary, homology modeling has assumed an increasingly important role in protein structure prediction in the recent years with the advent of structural genomics initiatives around the world. This is because many protein sequences are evolutionarily related, and thus can be classified into different families. Proteins in the same families frequently have noticeable similarities and thus share three-dimensional architecture, which allows a structural description of all proteins in a family even when only the structure of a single member is known. This evolutionary relationship provides the rationale for structural genomics, a systematic and large-scale effort towards structural characterization of all proteins, where a representative protein in each family is chosen to be solved experimentally with the rest reliably predicted by a homology modeling method [16]. Currently, only 7677 protein families have been identified according to the Pfam database (http://pfam.wustl.edu/). Certainly, this number is strongly dependent on the sequence similarity cutoffs used to cluster the sequence space. If 30% sequence identity cutoff is used, which is generally considered as a threshold for successful homology modeling, statistical estimates place the number somewhere between 10000 and 30000 for all proteins in Nature [17], but only a fraction of which have distinct spatial arrangements [18]. Even though ab-initio method can not solve protein folding problem in the foreseeable future, structure prediction will nevertheless be solved by homology modeling method anyway with the completion of structural genomics project, which looks like an attainable goal in the next 10 years. In fact, if we assume that protein structure is of global energy minimum, homology modeling is a simple scheme to search the conformation space by minimally disturbing those existing solutions, i.e., the experimentally solved structures. The obvious advantage is that the homology modeling technique relaxes the stringent requirement of force field and enormous conformation searching, because it dispenses with the calculation of a physical chemistry force field and replaces it, in large part, with the counting of sequence identities.

Given a protein sequence, homology modeling usually consists of the following four steps [19–20]: 1) identify the homologue of known structure from the Protein Data Bank; 2) align the query sequence to the template structure; 3) build the model based on the alignment; 4) assess and refine the model. When the sequence identity is above 40%, the alignment is straight forward, there are not many gaps, and 90% of main-chain atoms could be modeled with an RMSD (root-mean-square distance) error of about 1 Å [20]. In this range of sequence identity, the structural difference between proteins mainly arises from loops and side-chains. When the sequence identity is about 30–40%, obtaining correct alignment becomes difficult, where insertions and deletions are frequent. For sequence similarity in this range, 80% of main-chain

backbone atoms can be predicted to RMSD 3.5 Å, while the rest of residues are modeled with larger errors, especially in the insertion and deletion regions [21–24]. Even in correctly aligned regions, loop modeling and side-chain placement pose difficulties [25]. When the sequence similarity is below 30%, the main problem becomes the identification of the homologue structures, and alignment becomes much more difficult.

Approximately 57% of all known sequences have at least one domain that is related to at least one protein of known structure [26]. The probability of finding a related known structure for a randomly selected sequence from a genome ranges from 30% to 65%, since a few genomes have received more research attention than others [19,27–28]. The percentage is steadily increasing because more distinct folds are discovered each year, and because the number of different structural folds that proteins adopt is limited. Current estimates suggest that there are between 1000 and 5000 folds in the universe of compact globular proteins, with about 200 new folds realized annually from the structure deposition [18]. Currently, over 1.1 million proteins can readily have at least one of their domains reliably predicted with homology modeling methods. Given the rate of experimental structural determination at approximately 6000 proteins each year (http://www.rscb.org), it is arguable homology modeling has already saved up to hundreds of years of human effort. Even if homology modeling is generally much less accurate than experimental methods, it can still be helpful in proposing and testing hypothesis in molecular biology, such as hypotheses about the location of ligand binding sites [29–30], substrate specificity [31–32], function annotation [33], and drug design [34]. It can also provide starting models for solving structures from X-ray crystallography, NMR and electron microscopy [35–36]. In the next 10 years, structural genomics will possibly discover all protein distinct folds in Nature, making homology modeling applicable to almost any protein sequence [37]. The usefulness of homology modeling is ever increasing when more proteins can be predicted with higher accuracy.

In the following sections, we discuss the issues involved in homology modeling and the progresses made in sequence-structure alignment, model building, model refinement and model assessment.

2. HOMOLOGUE DETECTION AND ALIGNMENT

Homology modeling starts from selection of homologues with known structures from the PDB. If the query sequence has high sequence identity (>30%) to the structure, the homology detection is quite straightforward which is usually done by comparing the query sequence with all the sequences of the structures in the PDB. This can often be achieved simply with dynamic programming method [38] and its derivatives [39–40]. The most popular software is BLAST (http://www.ncbi.nlm.nih.gov/blast/) that searches sequence databases for optimal local alignments to the query. The BLAST program functions very well for alignment of sequences with high similarities. But when the sequence identity is well below 30%, homology hits from BLAST are not reliable. A number of alternative strategies have been developed. These include template consensus sequences [41–42] and profile analysis [43–45]. All these approaches, based on either multiple sequence or structure alignments, are more sensitive because the consensus sequences are better representative of the sequence family, and the profile reflects the conserved structural or functional preferences.

In the past several years, profile methods have emerged as the primary approach in distant homology detection. Position-specific profile search methods such as PSI-BLAST [46] and hidden Markov models (HMMs) [47], as implemented in the SAM [48] and HMMER (http://hmmer.wustl.edu) packages, have vastly improved the accuracy of sequence alignments and have extended the boundaries of detectable sequence similarity. Sequence profiles methods, e.g. PSI-BLAST, start from performing a pair-wise search of the database. The

significant alignments are then used by the program to construct a *position specific score matrix (PSSM)*. This matrix replaces the query sequence in the next round of database searching. The procedure may be iterated until no new significant alignments are found. The sequence profile method can be further enhanced by including evolutionary information for both the query and target proteins, where a profile from the query protein is compared with the profiles from the target proteins. Profile-profile alignments can be implemented in several fundamentally different ways [49–53]. The difference lies in how they calculate the score between two profile positions. A profile is a set of vectors, where each vector contains the frequency of each type of amino acid in a particular position of the multiple sequence alignment. In sequence-profile alignments, the score is calculated by extracting (the log of) the probability for an amino acid in this vector. However, in profile-profile alignments, two frequency vectors should be considered and this can be done in several different ways using a dot-product [49], a probabilistic model [54], or an information theoretical measure [50]. Ohlson and his coworkers [55] have demonstrated that profile-profile methods performed at least 30% better than standard sequence-profile methods both in their ability to recognize superfamily-related proteins and in the quality of the obtained alignments. In addition, profile-profile methods that use a probabilistic scoring function have an advantage over other methods as they can create good alignments and show a good fold recognition capacity using the same gap-penalties.

The sensitivity of profile method is often enhanced with information from multiple structure alignment, three-dimensional environment preferences, secondary structure prediction and solvent accessibility. Since the structural information is more conserved than sequence, it may represent the crucial requirement, in the process of evolution, of residues at specific positions with respect to the stability and function of the structure as a whole. Structure-derived sequence profiles that account for residue preferences in different structural environments are used widely [10,56–65]. The direct use of multiple structure alignments offers an intuitive approach for mapping structural information onto sequence. Sequence profiles generated from multiple structure alignments have been used to identify homologous core structures [66], residues conserved by evolution [23,67] and to derive structure-based substitution matrices [61,68–68]. Structure information has been shown to improve alignment quality in some studies. The Honig lab recently developed a series of hybrid multidimensional alignment profiles that combine sequence, secondary and tertiary structure information into hybrid profiles [65]. They demonstrated that both sequence-based and structure-based profiles contribute to remote homology detection and alignment accuracy, and that each contains some unique information. Przybylski & Rost [70] introduced a novel method that aligns generalized sequence and predicted structure profiles. Using predicted 1D structure (secondary structure and solvent accessibility) significantly improved over sequence-only methods, both in terms of correctly recognizing pairs of proteins with different sequences and similar structures and in terms of correctly aligning the pairs.

Although a major goal of the profile analysis has been remote homolog detection, an important side benefit has been significant improvement in alignment quality, even at levels of sequence identity for which pairwise alignment methods are known not to work. This, in turn, has had a positive impact on the starting alignments used in homology modeling, and thus has the potential to extend the applicability of homology modeling to increasingly lower levels of sequence similarity. There are many tools publicly available in this area, which can be found at the CASP6 website (http://predictioncenter.llnl.gov/casp6/) and in the previous publications [71].

3. MODEL BUILDING WITH ARTIFICIAL EVOLUTION

Given an alignment between the query sequence and the template, there are generally four methods in model building depending on how the information in the known structures is

transferred to the query sequence, i.e., rigid body assembly [72–75], segment matching [76], spatial restraint [77], and artificial evolution. Since the first three methods have already been reviewed extensively in literature [19–20], we are only describing our own approach, i.e., artificial evolution model building, which has been implemented in the NEST program, a module of JACKAL package. In the recent CASP5–6, homology models built by NEST have been ranked top 5% among all other participating groups. Since model building is a process that involves model refinement, we are also discussing the refinement module in NEST.

The alignment between the query and the template can be considered as a list of operations such as residue mutation, insertion or deletion. Suppose the template is the "parent structure", it would take Nature billions of years to evolve the template to the target. It is unlikely that Nature would finish the daunting task in one step. Instead, a more probable scenario is for Nature to evolve the template *via* multiple steps with minimal change to the template at each step. Accordingly, building a query model could be considered as a process of editing the template structure based on the alignment. Each operation, i.e., mutation, deletion or insertion, will disturb the template structure and thus involve an energy cost, either positive or negative. The model building starts from the operation with the least energy cost and so on. Each operation is followed by a slight energy minimization to remove atom clashes. The final structure is then subjected to more thorough energy minimization. The order for the first-round operation does not have to be determined by actually calculating the energy cost for each operation; instead, it can be conveniently estimated from an empirical point of view. For example, mutation is generally easier in evolution than insertion and deletion. As such, mutation operations on residues that are on the protein surface are usually performed first followed by mutation on buried small-sized residues and so on. The operation is considered successful if it does not cause a significant energy penalty to the structure; otherwise the operation is discarded and will return to the waiting-list. Insertion or deletion of multiple residues is considered as a group of operations, each operating on one residue. The operation starts from the middle residue of the segment with deletion preferred over insertion, since deletion is easier to be predicted. Similarly, operations with significant energy cost would also be considered unsuccessful and returned to the waiting-list for the next round operation. The next round operation actually works on the waiting-list, starting from the operation of the least energy cost that has been calculated from the previous round, but with a doubled energy cutoff. A number of rounds would finally accomplish the migration of the template structure to the model, which will be followed by a series of model refinement.

Model refinement in NEST is performed in two steps. The first step is to increase alignment quality, and the second step is to directly refine the model itself. Metaservers, e.g., CAFASP (http://bioinfo.pl/cafasp/), are usually used to identify as many prospective templates as possible. In the absence of a unanimous template identified by all servers, all possible hits will be considered. For example, if multiple templates are identified but all servers point to the same structural family, all structures in the PDB from that family should be used as possible templates. For each template identified, a number of alignments are obtained either from different servers or from alternative alignments based on a particular alignment protocol. Because model building can be done rapidly using NEST, the ensemble of sequence alignments is readily converted to an ensemble of three-dimensional model structures. A genetic algorithm is used on the ensemble of models to repeatedly improve model quality. Specifically, all the models are superimposed based on sequence alignment and the regions of high variability are identified. The variable regions then try conformations from corresponding segments of other models. The segment of the lowest colony energy is chosen as the best choice. The colony energy is a simple scheme that favors conformations found in broad energy wells, thereby approximating entropic effects [78]. The process can be repeated until a stable model has been derived.

For each model, unaligned regions corresponding to gaps in the sequence alignment are modeled using the independent LOOPY program [78]. Specifically, two thousand initial conformations are randomly sampled and filtered against the consensus secondary-structure predictions from the CAFASP server. The 2000 conformations are then energy-minimized using our fast “direct tweak” method, and the 300 conformations of lowest energy are kept. An additional 300 are obtained from a fragment database using sequence similarity, secondary structure, and end-point geometry. The 600 conformations are subjected to additional energy minimization, and the conformation of lowest colony energy is selected. Side chains are modeled with the independent SCAP program [79], where the initial conformation starts from the NEST output. The final model will be further refined with constraints. The constraints include backbone hydrogen bonds and main-chain framework of the template structure, i.e., an energy penalty would be applied if the sampled structure violates an existing hydrogen bond or deviates significantly (more than 2 Å) from the original model. This is to guarantee that sampling only visits conformations close to the templates.

Table 1 shows the most widely used model building programs that are publicly available. Besides model building protocols, the programs also wildly differ from each other in the methods that are used for model refinement. Given the same alignment and template, it was generally believed there were no major differences between the best modeling programs. However, a recent study by Wallner and Elofsson [80] has shown that some programs performed better than others. In their study, a benchmark of six different homology modeling programs MODELLER, SEGMOD/ENCAD [76], SWISS-MODEL [81], 3D-JIGSAW [82], NEST, and BUILDER [83] is presented. Their study concluded that no single modeling program consistently outperformed the others in all tests. However, it is quite clear that three modeling programs, MODELLER, NEST, and SEGMOD/ENCAD, perform better than the others. Detailed analysis of these homology modeling programs revealed some interesting differences. For example, using a 1.4 GHz AMD XP processor, NEST needs 17s on average to build a model, while SEGMOD needs 6s, and MODELLER takes 43 to 430s in MODELLER6v2 and MODELLER6v2–10 respectively; MODELLER, SWISS-MODEL, and BUILDE produce more models that do not converge compared to the other programs. For sequence identities below 40% all modeling programs manage to bridge some gaps and build some loops correctly or incorrectly; therefore, accordingly, some models are better or worse than the template. In this region the MODELLER programs, NEST, SEGMOD/ENCAD, and SWISS-MODEL, improved 20% of the models. Only NEST rarely made the models worse, while all other programs deteriorated at least 5% of the models. The authors also found that NEST had more of its models “among best” than the other programs, thus, selecting a model from nest is almost always a good choice.

4. HOMOLOGY MODEL REFINEMENT

High-resolution refinement is a difficult task that requires an effective sampling strategy as well as an accurate energy function to guide the search through conformational space. Homology model refinement is primarily focused on tuning alignment and modeling loops and side chains. Loops are usually the most variable regions of a structure where insertion and deletion often occur. Correct alignment is the most important task for homology modeling, since the errors introduced into the model by misalignment are hard to remove in the later stages of refinement. When the sequence identity is above 40%, errors in the homology structure mainly comes from side chains; when the sequence identity is between 30–40%, loops and side chains become most problematic [20]. Given a good energy function, loop and side-chain refinement can in principle be applied repeatedly to relax the backbone closer to the native. Refinement on helix and beta sheet can be handled with similar methods as for loops, where proper hydrogen bond constraints should be applied to retain the secondary structure

definition [84]. Recent attempts have been made to use physical chemistry energy to refine side chains, loop and helix segments, target-template alignment, and the whole model.

4.1. Loop Prediction

The basic goal is to predict the conformation of a loop that is fixed at both ends by the protein backbone. Loop prediction is often regarded as a mini protein folding problem. Basically, the approaches fall into two main categories: *ab initio* (*de novo*) [78,85–88] and database methods [89–92]. *Ab-initio* methods of loop prediction involve the generation of a large number of randomly chosen candidate conformations; database approaches try to find a segment of a protein with known three-dimensional structure that fits the stem regions of a loop. In ab-initio method, information from database is often included, e.g. phi-, psi-maps of known loops [85, 93,94]. Once loops are generated in this way, energetic criteria (or sequence similarity for database approach) are often applied to select the most likely candidate.

It is obviously important that near-native conformations be present among the candidate conformations generated in the first step of loop modeling. For loops of less than 12 residues, adequate sampling does not appear to be a problem [87]. However, for database methods, loops longer than 5 residues often have problem identifying near-native conformations in the template library, which limits their utility for these cases [91]. Even if the Protein Data Bank has been significantly enlarged since then, recent research showed that the database search method is overtaken by the *ab initio* method at around six residues loop length [93]. The accuracy of loop modeling is highly dependent not only on the number of residues in the loop, but also on the distance between the loop stems. Generally, when the distance between the loop stems is shorter, the loop conformation is more like "Ω", and thus has more freedom to move around; therefore, it is more difficult to predict. If the template candidate comes from a protein structure of the same family as the target protein, database approach is usually more reliable, especially for longer loops.

A number of recent papers on sequence-dependent loop prediction using a database method achieved an average accuracy of more than 3.5Å rmsd for the backbone atoms of eight-residue loops [90,95]. By combining database and ab-initio approach, Deane & Blundell [93] further improved the prediction accuracy from 3.5 to 3.0 Å for 8-residue loops. Vlijmen and Karplus [92] used CHARMM to optimize initial conformations that were selected from the protein database. They report improved results for longer loops but their optimization procedure, which involves simulated annealing, effectively extends the range of conformation space searched beyond that provided by the database conformations. In this sense, their approach is closer to *ab-initio* loop generation. The most accurate loop prediction from database approach was reported by Michalsky et.al. [96], which has on average RMSD 1.35 Å for 8-residue loops. This was mainly achieved by a comprehensive compilation of backbone conformations found in the PDB, thus effectively relaxing the limitation of small template library on loop prediction.

Because conformational sampling does not appear to be a problem for loops of less than 12 residues, the quality of the scoring function used to evaluate loop conformations is the major determinant of loop-prediction accuracy. Rapp and Friesner [87] used the generalized Born solvation model and the AMBER94 force field to obtain low rmsd values for the two loops they studied. Fiser *et al.* [85] published an extensive *ab-initio* study on a data set of 40 loops and utilized a scoring function that included the CHARMM22 force field and statistical preferences taken from protein databases. They reported low rmsd less than 2 Å from known structures. Zhang et.al. [97] used DFIRE-based statistical potential for loop discrimination. Their results suggest that a single-term DFIRE-statistical energy function can provide accurate loop prediction as good as more rigorous physic-chemistry energies. Scoring functions based entirely on physical chemistry potentials and an accurate solvation model have the potential of identifying the native conformation as lowest in energy, but there are cases where lower

energy conformations appear [98]. One problem may be that essentially every approach seeks the lowest energy conformation, thus ignoring conformational entropy effects that will favor broad energy wells. We have recently implemented a procedure called “colony energy” that takes the shape of the energy well into account and yields highly accurate loop prediction (e.g. 1.4Å global rmsd for eight-residue loops) [78]. With crystal environments considered, Jacobson et.al. [88] achieved the best accuracy of 1.0 Å rmsd for 8-residue loops with a computing-intensive approach that combines OPLS all-atom energy function, efficient methods for loop buildup and side-chain optimization, and the hierarchical refinement protocol. Fogolari & Tosatto [99] demonstrated that molecular mechanics/Poisson-Boltzmann solvent accessible surface area, if combined with colony energy approach, is very effective in discriminating loop decoys.

Table 2 shows some loop modeling software that can be easily obtained from the web. Other loop modeling software only exists as internal components for model building packages listed in Table 1. Compared with database scanning method, most ab-initio loop prediction programs are very slow. For example, the loop prediction program by Fiser et.al. [85] requires about 40 hours for an 8-residue loop; while LOOPY needs only about 10 minutes with good accuracy [78]. In our recent participation in CASP6 (Critical Assessment of Structure Prediction), 2004, LOOPY was ranked the third in the category of loop assessment (http://predictioncenter.llnl.gov/casp6/meeting/presentations/).

4.2. Side-chain Prediction

The greatest success in the prediction of side-chain conformations has been achieved for core residues where packing constraints significantly simplify the problem. Even for core residues, the accuracy of side-chain prediction degrades when the structure of the backbone is itself not known to a high degree of accuracy. Many side-chain programs are based on rotamer libraries [100], which are generally defined in terms of side-chain torsional angles for preferred conformations of a particular side chain. The resolution of rotamer libraries has increased over time and rotamer libraries have been compiled simply by sampling all angles at some given level of resolution [101]. Since backbone conformation changes the frequency of the rotamers, backbone-dependent rotamer library is often used in side chain modeling [102–103]. The major advantage is to increase computing efficiency, since bad rotamers, e.g. clashing with the backbone, have been automatically removed during construction of the rotamer library. Baker and his coworkers [104] have developed a “solvated rotamer” approach that shows improvement on side chain packing at protein-protein interface. This approach extends current side-chain packing methods by using a rotamer library including solvated rotamers with one or more water molecules fixed to polar functional groups in probable hydrogen bond orientations, together with a simple energetic description of water-mediated hydrogen bonds. As the number of rotamers increases, however, so does the problem of sampling all possible conformations. There have been a variety of approaches developed to deal with the combinatorial problem in side-chain prediction [105–111].

In recent papers, accuracies of about 1Å rmsd have been reported for core residues in known structures where the backbone has been fixed in the native conformation [25,109,112,113]. A number of recent studies suggest that further improvements may still be possible. Mendes *et al.* [114] found, for example, that the use of an intrinsic torsional potential can improve prediction accuracy. Lovell *et al.* [115] recently reported a novel rotamer library in which internal clashes between side chain and backbone are removed. This library could, in principle, be used to improve prediction accuracy. We have recently shown that using a very detailed rotamer library, which is based on rotamers that use Cartesian coordinates taken from known structures rather than idealized bond lengths and angles, yields rmsd values relative to the native of only 0.62Å for core residues [79]. This appears to constitute a significant

improvement over existing procedures and demonstrates that the combinatorial problem, usually assumed to greatly complicate side-chain prediction, may in fact be of little consequence. This was later confirmed in more detailed study [116], which showed that low-order local minima for side chain prediction may be almost as accurate as the global minimum when evaluated against experimentally determined structures. Improvement on side chain prediction in recent years has mainly come from better energy functions. Eyal et.al. [117] showed that solvent accessibility and contact surface area are important on the accuracy of side chain prediction, particularly for modeling buried side chains. Liang and Grishin [118] have developed a new and simple scoring function for side chain prediction that consists of the following energy terms: contact surface, volume overlap, backbone dependency, electrostatic interactions, and desolvation energy. The weights of these energy terms were optimized to achieve the minimal average root mean square (rms) deviation between the lowest energy rotamer and real side-chain conformation on a training set of high-resolution protein structures. The derived scoring function combined with a Monte Carlo search algorithm was used to place all side chains onto a protein backbone simultaneously. The average prediction accuracy was 87.9% for chi(1), 73.2% for chi(1 + 2), and 1.34 Angstrom rms deviation for all side chains in a protein structure. As is the case for loop prediction, side-chain prediction accuracy depends sensitively on the accuracy to which the backbone conformation is known [119]. This suggests the possibility of developing procedures where side-chain and backbone conformation can be used iteratively to refine homology models.

Table 3 lists some publicly available side chain prediction programs and the methods they used. Earlier side chain predictions, e.g. RAMP [113], SMD [120], Confmat [112], etc., were usually based on small rotamer libraries; most recent programs were using very detailed rotamer libraries, e.g., SCAP [79], SCWRL [103], SMOL [118]. In our recent benchmark study of SCAP, SMOL and SCWRL, SCAP excelled in prediction for core and surface residues (manuscript submitted for publication). For partially buried residues, SMOL performed the best, which was due to its more sufficient conformation sampling and optimized scoring function. SCWRL performed reasonably well though not as accurate as the other two, but with much less CPU cost. On a 300MH SGI machine, SCWRL is very fast, 3 seconds for each protein, while SMOL needs 11700 seconds and SCAP needs 361 seconds. In the recent CASP6, SCAP was ranked the second in the category of side chain assessment for homology models (http://predictioncenter.llnl.gov/casp6/meeting/presentations/).

4.3. Other Advances in Model Refinement

Recent attempts on model refinement have been mainly achieved from increasing alignment accuracy. Almost all alignment software currently in use has to rely on one of the derivatives of dynamic programming. Although dynamic programming can obtain global optimal alignment for a given scoring matrix, it cannot account for non local residue-residue interactions. For example, double mutant effects can only be properly estimated when their spatial conformations are both available. As such, a cumbersome but effective method of refining alignment is to build multiple models based on the alternative alignments, with the best alignment corresponding to the model of the lowest physic-chemical energy. The rationale is the assumption that conformation of lower energy is more likely close to the native state. The method becomes possible due to the availability of more discriminatory energy functions and faster model building tools [121,122]. Tens of thousands of models can be built in a short time with linux clusters, each based on one variation of alignment. An effective scoring energy can be readily applied to the ensemble of models. These models can be further minimized with an approach similar to genetic algorithm, i.e., shuffling segments among different models by fixing other parts of protein, where the stems of the segment should have identical residues aligned with the template. Similarly, genetic algorithm is also an important tool to increase model quality based on multiple templates. The multiple models, each based on one template,

will be superimposed. Variable regions identified are exchanged and optimized among different models. In the optimization process, an rmsd constraint can be applied to restrict sampling to the conformational space close to the averaged framework of the original templates. The method has been utilized in the NEST program and produced satisfactory results in CASP6.

More recent research has been attempted to use MD simulation to model refinement. Lee *et al.* [123] used molecular dynamics simulations with an explicit solvent model to refine Rosetta models followed by scoring with the Poisson-Boltzman surface area solvation model. Their results showed that native structures could be distinguished from low-resolution models and that the native state is stable. Lu *et al.* [124] used a combination of local constraints, knowledge-based potentials and molecular dynamics approaches that showed promising improvements over previous studies using standard molecular dynamics methods. Fan & Mark [125] used classical molecular dynamics simulations with explicit water to refine homology models. A significant improvement over the model structures has been observed in a number of cases. The results indicate that homology models could be possibly refined with molecular dynamics simulations on tens to hundreds of nanoseconds time scale. Qian *et al.* [126] used the principal components of the variation of backbone structures within a homologous family to define a small number of evolutionarily favored sampling directions and showed that model quality can be improved by energy-based optimization along these directions. Li *et al.* [84] developed new hierarchical and multiscale algorithms to sample helices and flanking loops, which were evaluated with an all-atom protein force field (OPLS) and a Generalized Born continuum solvent model. This method, integrated with loop and side chain modeling technique, can potentially be used to refine homology structures iteratively. The next-generation structure modeling algorithms should be able to refine a protein structure closer to the native conformation, although this is not an easy task. The most critical part is to obtain an energy function that is sensitive enough to discriminate decoys from near-native conformations. Though conformation sampling is also difficult, computer clusters allow more thorough sampling of states around the original models.

5. MODEL ASSESSMENT

All models built by homology will have errors as discussed in the previous section. Verification of the model, and estimation of the likelihood and magnitude of errors has become one of the most important steps in advancing the state of the art of homology modeling. Errors of the model are usually estimated either from the energy of the model, or from the resemblance of a given characteristic of the model to real structures. The most critical component is the development of a scoring function that is capable of discriminating good and bad models, and therefore, would have enormous impact on the ability to predict protein conformations.

Scoring functions used for the evaluation of protein models generally fall into two broad categories. ‘Statistical’ effective energy functions [127] are based on the observed properties of amino acids in known structures, and have been widely used in fold recognition and homology modeling applications. A variety of statistical criteria have been used successfully to discriminate between deliberately misfolded and native structures. Most of them are directly or indirectly based on the analysis of contacts, either inter residue contacts, inter atom contacts, or contacts with solvent. For example, preferential distributions of polar and apolar residues inside or outside of protein can be used to detect completely misfolded models [128]; solvation potentials can detect local errors as well as complete misfolds [129]; packing rules have been implemented for structure evaluation [130]. Residue or atom contacts are discriminative because they are energetically favored. Real structures cannot tolerate too much unfavorable interaction. Thus for a model to be correct only a few infrequently observed atomic contacts are allowed. However, bond angles and bond lengths, though powerful in checking the quality

of experimental structures, are usually less useful for the evaluation of models because these factors have already been considered appropriately in the model building stage [19]. Although they are computational cheap, statistical energies are not sensitive to evaluate near native decoy structures, especially for segments of proteins, such as in the modeling of loop and side chains.

Physical effective energy functions [131] are based on a direct evaluation of the conformational free energy of a protein. Recent work has demonstrated that such a direct evaluation of the conformational free energy can be at least as successful as statistically based scoring functions in distinguishing the native structure of a protein from an incorrectly folded decoy, although generally at greater computational cost [121,132,133]. A distinct advantage of such physically derived functions is that they are based on well-defined physical interactions, thus making it easier to learn and to gain insight from their performance. Moreover, the success in CASP of *ab-initio* methods based on purely physical chemistry methods [134] suggests that our understanding of the forces that drive protein stability may have reached the point where it can be translated into widely applicable computational tools. One of the major drawbacks of accurate physical chemical description of the folding free energy of a protein is that the treatment of solvation required usually comes at a significant computational expense. Fast solvation models such as the Generalized Born [135] and a variety of simplified scoring schemes [121,136] may prove to be extremely useful in this regard.

A number of freely available programs can be used to verify homology models as shown in Table 4. They generally belong to one of two categories. The first category (e.g. PROCHECK and WHATIF) checks for proper protein stereochemistry, such as symmetry checks, geometry checks (chirality, bond lengths, bond angles, torsion angles, etc) and structural packing quality; the second category (e.g.VERIFY3D and PROSAII) checks the fitness of sequence to structure, and assigns a score for each residue fitting its current environment. A new graphics software called GRASP2 developed in the Honig lab is also extremely useful in model assessment [137]. The software can display alignments and template structures simultaneously for assessment of the alignment quality. For example, gaps and insertions can be mapped to the structures to verify that they make sense geometrically. Where residue substitutions occur, the user can verify that structural features such as hydrophobic packing are maintained and that active-site residues and other features of the target identified from the literature are conserved. The manual inspection should be combined with existing programs to further identify problems in the model.

6. SUMMARY

Protein structure plays a key role in understanding the mechanism of protein function. Experimentally solving protein structure is a tedious process that can not meet the demand resulted from the exponential growth of protein sequences. Protein structure prediction has been a dream for scientific community for decades, not only for computational chemist, but also for physicist, mathematician and computer scientist. The dream has not been viewed attainable until recently with the explosion of sequence and structural information and because of computational advances in many different areas including sequence profile analysis and better understanding of energetic determinants of protein stability. In the next 10 years, structural genomics project could possibly map out all the distinct folds in Nature, which makes it possible to solve the problem of protein structure prediction reliably using homology modeling method. The ever increasing databases of protein structure and sequence will be certain to spur the development of a host of new computational methods aimed at detecting new relationships between sequence, structure, and function. Continued progress in ab-initio modeling makes it possible to further refine homology models to higher accuracy. Such models will provide the basis for a more detailed analysis of structure and function relationships than

has been available in the past and, will provide powerful tools for the analysis of experimental data and for the design of new experiments.

Acknowledgements

I thank Drs. Peter Steinbach, Jan Norberg for their many useful comments and critical reading of the manuscript.

References

1. Blundell TL, Bedarkar S, Rinderknecht E, Humble RE. Proc Natl Acad Sci 1978;75:180–184. [PubMed: 272633]
2. Weber IT. Proteins 1990;7(2):172–184. [PubMed: 2158092]
3. Anfinsen CB. Science 1973;181:223–230. [PubMed: 4124164]
4. Johnson MS, Srinivasan N, Sowdhamini R, Blundell TL. Crit Rev Biochem Mol Biol 1994;29(1):1–68. [PubMed: 8143488]
5. Lesk AM. Proteins 1997;1:151–166. [PubMed: 9485507]Suppl
6. Zemla A, Venclovas C, Reinhardt A, Fidelis K, Hubbard TJ. Proteins 1997;1:140–150. [PubMed: 9485506]Suppl
7. Ingwall RT, Scheraga HA, Lotan N, Berger A, Katchalski E. Biopolymers 1968;6:331–368. [PubMed: 5641934]
8. Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J. Proteins 1999;37(4):592–610. [PubMed: 10651275]
9. Rost B, Fariselli P, Casadio R. Protein Sci 1996;5(8):1704–1718. [PubMed: 8844859]
10. Bowie JU, Luthy R, Eisenberg D. Science 1991;253:164–170. [PubMed: 1853201]
11. Xu Y, Xu D, Uberbacher EC. J Comp Biol 1998;5(3):597–614.
12. Venclovas C, Zemla A, Fidelis K, Moult J. Proteins 2003;53:585–595. [PubMed: 14579350]Suppl 6
13. Bissan A, Jung J, Xiang ZX, Honig B. Curr Opin Struct Biol 2001;5(1):51–56.
14. Wang C, Wan S, Xiang Z, Shi Y. J Phys Chem 1997;101:230–235.
15. Xiang Z, Xu Y, Shi Y. J Comp Chem 1995;16(2):200.1995
16. Goldsmith-Fischman S, Honig B. Protein Sci 2003;12(9):1813–1821. [PubMed: 12930981]
17. Liu X, Fan K, Wang W. Proteins 2004;54(3):491–499. [PubMed: 14747997]
18. Brenner SE, Chothia C, Hubbard TJ. Curr Opin Struct Biol 1997;7(3):369–376. [PubMed: 9204279]
19. Fiser, A.; Sali, A. Protein Structure. Chasman, D., editor. Marcel Dekker, Inc.; New York: 2003. p. 167-206.
20. Sanchez R, Sali A. JMolStruct (Theochem) 1997;398–399:489–496.
21. Harrison RW, Chatterjee D, Weber IT. Proteins 1995;23(4):1463–671.
22. Mosimann S, Meleshko R, James MN. Proteins 1995;23 (3):301–317. [PubMed: 8710824]
23. Yang AS, Honig B. J Mol Biol 2000;301(3):665–678. [PubMed: 10966776]
24. Sauder JM, Arthur JW, Dunbrack RL Jr. Proteins 2000;40(1):6–22. [PubMed: 10813826]
25. Bower M, Cohen FE, Dunbrack RL Jr. JMolBiol 1997;267:170–184.
26. Pieper U, Eswar N, Ilyin VA, Stuart A, Sali A. Nucleic Acids Res 2002;30:255–259. [PubMed: 11752309]
27. Kelley LA, MacCallum RM, Sternberg MJ. J Mol Biol 2000;299:499–520. [PubMed: 10860755]
28. Teichmann SA, Chothia C, Gerstein M. Curr Opin Struct Biol 1999;9:390–399. [PubMed: 10361097]
29. Francoijs CJ, Klomp JP, Knegtel RM. Protein Eng 2000;13(6):391–394. [PubMed: 10877848]
30. Zhou Y, Johnson ME. J Mol Recognit 1999;12(4):235–241. [PubMed: 10440994]
31. Jung JW, An JH, Na KB, Kim YS, Lee W. Protein Sci 2000;9(7):1294–1303. [PubMed: 10933494]
32. De Rienzo F, Fanelli F, Menziani MC, De Benedetti PG. J Comput Aided Mol Des 2000;14(1):93–116. [PubMed: 10702928]
33. Ginalski K, Rychlewski L, Baker D, Grishin NV. Proc Natl Acad Sci 2004;101(8):2305–2310. [PubMed: 14983005]
34. Takeda-Shitaka M, Takaya D, Chiba C, Tanaka H, Umeyama H. Curr Med Chem 2004;11(5):551–558. [PubMed: 15032603]

35. Talukdar AS, Wilson DL. IEEE Trans Med Imaging 1999;18 (7):604–616. [PubMed: 10504094]
36. Ceulemans H, Russell RB. J Mol Biol 2004;338(4):783–793. [PubMed: 15099745]
37. Vitkup D, Melamud E, Moult J, Sander C. Nat Struct Biol 2001;8:559–566. [PubMed: 11373627]
38. Needleman SB, Wunsch CD. J Mol Biol 1970;48(3):443–453. [PubMed: 5420325]
39. Smith TF, Waterman MS. J Mol Biol 1981;147(1):195–197. [PubMed: 7265238]
40. Gotoh O. J Mol Biol 1982;162(3):705–708. [PubMed: 7166760]
41. Taylor WR. J Mol Biol 1986;188(2):233–258. [PubMed: 3088284]
42. Chappey C, Danckaert A, Dessen P, Hazout S. Comput Appl Biosci 1991;2:195–202. [PubMed: 2059844]
43. Suyama M, Matsuo Y, Nishikawa K. J Mol Evol 1997;44:S163–173. [PubMed: 9071025]Suppl 1
44. Lolkema JS, Slotboom DJ. Mol Membr Biol 1998;15(1):33–42. [PubMed: 9595553]
45. Barton GJ, Sternberg MJ. J Mol Biol 1990;212(2):389–402. [PubMed: 2319605]
46. Altschul S, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Nucleic Acids Res 1997;25:3389–3402. [PubMed: 9254694]
47. Krogh A, Brown M, Mian I, Sjolander K, Haussler D. J Mol Biol 1994;235:1501–1531. [PubMed: 8107089]
48. Karplus K, Barrett C, Hughey R. Bioinformatics 1998;14:846–856. [PubMed: 9927713]
49. Rychlewski L, Jaroszewski L, Li W, Godzik A. Protein Sci 2000;9:232–241. [PubMed: 10716175]
50. Yona G, Levitt M. J Mol Biol 2002;315:1257–1275. [PubMed: 11827492]
51. Sadreyev R, Grishin NV. J Mol Biol 2003;326:317–336. [PubMed: 12547212]
52. Edgar R, Sjolander K. Bioinformatics 2003;19:1404–1411. [PubMed: 12874053]
53. Pei J, Sadreyev R, Grishin NV. Bioinformatics 2003;19:427–428. [PubMed: 12584134]
54. Mittelman D, Sadreyev R, Grishin N. Bioinformatics 2003;19:1531–1539. [PubMed: 12912834]
55. Ohlson T, Wallner B, Elofsson A. Proteins 2004;57(1):188–197. [PubMed: 15326603]
56. Johnson MS, Overington JP, Blundell TL. J Mol Biol 1993;231:735–752. [PubMed: 8515448]
57. Fischer D, Rice D, Bowie JU, Eisenberg D. FASEB J 1996;10:126–136. [PubMed: 8566533]
58. Rost B, Schneider R, Sander C. J Mol Biol 1997;270:471–480. [PubMed: 9237912]
59. Rice DW, Eisenberg D. J Mol Biol 1997;267:1026–1038. [PubMed: 9135128]
60. Hargbo J, Elofsson A. Proteins: Struct Funct Genet 1999;36:68–76. [PubMed: 10373007]
61. Shi J, Blundell TL, Mizuguchi K. J Mol Biol 2001;310:243–257. [PubMed: 11419950]
62. Jaroszewski L, Rychlewski L, Zhang B, Godzik A. Protein Sci 1998;7:1431–1440. [PubMed: 9655348]
63. Jones DT. J Mol Biol 1999;287:797–815. [PubMed: 10191147]
64. Panchenko AR, Marchler-Bauer A, Bryant SH. J Mol Biol 2000;296:1319–1331. [PubMed: 10698636]
65. Tang CL, Xie L, Koh IY, Posy S, Alexov E, Honig B. J Mol Biol 2003;334(5):1043–1062. [PubMed: 14643665]
66. Matsuo Y, Bryant SH. Proteins: Struct Funct Genet 1999;35:70–79. [PubMed: 10090287]
67. Mirny LA, Shakhnovich EI. J Mol Biol 1999;291:177–196. [PubMed: 10438614]
68. Ogata K, Ohya M, Umeyama H. J Mol Graph Model 1998;16:178–189. [PubMed: 10522237]
69. Blake JD, Cohen FE. J Mol Biol 2001;307:721–735. [PubMed: 11254392]
70. Przybylski D, Rost B. J Mol Biol 2004;341(1):2, 55–69.
71. Xu D, Xu Y, Uberbacher EC. Current Protein and Peptide Science 2000;1(1):1–21. [PubMed: 12369918]
72. Greer J. Proc Natl Acad Sci 1980;77(6):3393–3397. [PubMed: 6932026]
73. Greer J. JMolBiol 1981;153:1027.
74. Sutcliffe MJ, Haneef I, Carney D, Blundell TL. Protein Eng 1987a;1(5):377–384. [PubMed: 3508286]
75. Sutcliffe MJ, Hayes FR, Blundell TL. Protein Eng 1987b;1(5):385–392. [PubMed: 3508287]
76. Levitt M. JMolBiol 1992;226:507–533.
77. Sali A, Blundell TL. J Mol Biol 1993;234(3):779–815. [PubMed: 8254673]

78. Xiang ZX, Csoto C, Honig B. Proc Natl Acad Sci 2002;99:7432–7437. [PubMed: 12032300]
79. Xiang ZX. Extending the accuracy limit of side-chain prediction. J Mol Biol 2001;311(2):421–430. [PubMed: 11478870]
80. Wallner B, Elofsson A. Protein Science 2005;14:1315–1327. [PubMed: 15840834]
81. Schwede T, Kopp J, Guex N, Peitsch MC. Nucleic Acids Research 2003;31:3381–3385. [PubMed: 12824332]
82. Bates PA, Kelley LA, MacCallum RM, Sternberg MJE. Proteins: Structure, Function and Genetics, Suppl 2001;5:39–46.
83. Koehl P, Delarue M. Curr Opin Struct Biol 1996;6(2):222–226. [PubMed: 8728655]
84. Li X, Jacobson MP, Friesner RA. Proteins 2004;55:368–382. [PubMed: 15048828]
85. Fiser A, Gian Do R, Sali A. Protein Sci 2000;9:1753–1773. [PubMed: 11045621]
86. Zheng Q, Kyle DJ. Proteins 1996;24(2):209–217. [PubMed: 8820487]
87. Rapp CS, Friesner RA. Proteins 1999;35(2):173–183. [PubMed: 10223290]
88. Jacobson MP, Pincus DL, Rapp CS, Day TJF, Honig B, Shaw DE, Friesner RA. Proteins: Struct Funct Genet 2004;55:351–367. [PubMed: 15048827]
89. Li W, Liu Z, Lai L. Biopolymers 1999;49:481–495. [PubMed: 10193195]
90. Wojcik J, Mornon JP, Chomilier J. J Mol Biol 1999;289(5):1469–1490. [PubMed: 10373380]
91. Fidelis K, Stern PS, Bacon D, Moult J. Protein Eng 1994;7:953–960. [PubMed: 7809034]
92. Van Vlijmen HW, Karplus M. J Mol Biol 1997;267:975–1001. [PubMed: 9135125]
93. Deane CM, Blundell TL. Protein Sci 2001;10:599–612. [PubMed: 11344328]
94. Tosatto SC, Bindewald E, Hesser J, Manner R. Protein Eng 2002;15(4):2, 79–86.
95. Deane CM, Blundell TL. Proteins 2000;40:135–144. [PubMed: 10813838]
96. Michalsky E, Goede A, Preissner R. Protein Eng 2003;16(12):979–985. [PubMed: 14983078]
97. Zhang C, Liu S, Zhou YQ. Protein Sci 2004;13(2):391–399. [PubMed: 14739324]
98. Smith KC, Honig B. Proteins 1994;18:119–132. [PubMed: 8159662]
99. Fogolari F, Tosatto SC. Protein Sci 2005;14(4):889–901. [PubMed: 15772305]
100. Ponder JW, Richard FM. JMolBiol 1987;193:775–791.
101. Maeyer MD, Desmet J, Lasters I. Fold Des 1997;2(1):53–66. [PubMed: 9080199]
102. Dunbrack RL Jr, Karplus M. J Mol Biol 1993;330(2):543–574. [PubMed: 8464064]
103. Canutescu AA, Shelenkov AA, Dunbrack RL Jr. Protein Sci 2003;12(9):2001–2014. [PubMed: 12930999]
104. Jiang L, Kuhlman B, Kortemme TA, Baker D. Proteins 2005;58(4):893–904. [PubMed: 15651050]
105. Lee C. J Mol Biol 1994;236(3):918–939. [PubMed: 8114102]
106. Lee C, Subbiah S. J Mol Biol 1991;217(2):373–388. [PubMed: 1992168]
107. Dahiyat BI, Mayo SL. Proc Natl Acad Sci 1997;94(19):10172–10177. [PubMed: 9294182]
108. Gordon DB, Mayo SL. Structure Fold Des 1999;7(9):1089–1098. [PubMed: 10508778]
109. Vasquez M. Curr Opin Struct Biol 1996;6:217–221. [PubMed: 8728654]
110. Kingsford CL, Chazelle B, Singh M. Bioinformatics 2005;21(7):1028–1036. [PubMed: 15546935]
111. Samudrala R, Huang ES, Koehl P, Levitt M. Protein Eng 2000;7:453–457. [PubMed: 10906341]
112. Koelh P, Delarue M. J Mol Biol 1994;239:249–275. [PubMed: 8196057]
113. Samudrala R, Moult J. Protein Eng 1998;11(11):991–997. [PubMed: 9876919]
114. Mendes J, Baptista A, Carrondo M, Soares CM. Biopolymers 1999;50:111–131. [PubMed: 10380336]
115. Lovell SC, Word JM, Richardson JS, Richardson DC. Proteins 2000;40(3):389–408. [PubMed: 10861930]
116. Desmet J, Spriet J, Lasters I. Proteins 2002;48(1):31–43. [PubMed: 12012335]
117. Eyal E, Najmanovich R, McConkey BJ, Edelman M, Sobolev V. J CompChem 2004;25(5):712–724.
118. Liang SD, Grishin NV. Protein Sci 2002;11(2):322–331. [PubMed: 11790842]

119. Huang ES, Koehl P, Levitt M, Pappu RV, Ponder JW. Proteins 1998;33(2):204–217. [PubMed: 9779788]
120. Tuffery P, Etchebest C, Hazout S, Lavery R. J Comp Chem 1993;14:790–798.
121. Petrey D, Honig B. Protein Sci 2000;9:2181–2191. [PubMed: 11152128]
122. Petrey D, Xiang XZ, Tang CL, Xie L, Gimpelev M, Mitors T, Soto CS, Goldsmith-Fischman S, Kernytsky A, Schlessinger A, Koh IYY, Alexov E, Honig B. Proteins 2003;53:430–435. [PubMed: 14579332]
123. Lee MR, Baker D, Kollman PA. J Am Chem Soc 2001;123:1040–1046. [PubMed: 11456657]
124. Lu H, Skolnick J. Biopolym 2003;70(4):575–584.
125. Fan H, Mark AE. Protein Sci 2004;13(1):211–220. [PubMed: 14691236]
126. Qian B, Ortiz AR, Baker D. Proc Natl Acad Sci 2004;101(43):15346–51. [PubMed: 15492216]
127. Sippl M. Curr Opin Struct Biol 1995;5:229–235. [PubMed: 7648326]
128. Baumann G, Froemmel C, Sander C. ProtEng 1989;2:329–334.
129. Holm L, Sander C. J Mol Biol 1992;225:93–105. [PubMed: 1583696]
130. Gregoret LM, Cohen FE. JMolBiol 1990;211:959–974.
131. Lazaridis T, Karplus M. J Mol Biol 1999;288:477–487. [PubMed: 10329155]
132. Janardhan A, Vajda S. Protein Sci 1998;7:1772–1780. [PubMed: 10082374]
133. Vorobjev Y, Almagro J, Hermans J. Proteins 1998;32:399–413. [PubMed: 9726412]
134. Lee J, Liwo A, Ripoll D, Pillardy J, Scheraga H. Proteins, 37, Suppl 1999;3:204–208.
135. Still W, Tempczyk A, Hawley R, Hendrickson T. J Am Chem Soc. 112 1990:6127–6129.
136. Huang E, Subbiah S, Levitt M. J Mol Biol 1995;252:709–720. [PubMed: 7563083]
137. Petrey D, Honig B. Methods in Enzymology 2003;374:492–509. [PubMed: 14696386]

Table 1

Homology Modeling Programs

Programs	Availability	Methods
NEST	http://trantor.bioc.columbia.edu/programs/jackal/	Artificial evolution
COMPOSER	http://www-cryst.bioc.cam.ac.uk/	Rigid-body assembly
Tripos (COMPOSER)	http://www.tripos.com/	Rigid-body assembly
CONGEN	http://www.congenomics.com/congen/congen_toc.html	Rigid-body assembly
MODELLER	http://guitar.rockefeller.edu/modeller/modeller.html	Spatial restraints
InsightII (MODELLER)	http://www.accelrys.com/	Spatial restraints
SWISS-MODEL	http://www.expasy.ch/swissmod/SWISS-MODEL.html	Rigid-body assembly
SCHRODINGER	http://www.schrodinger.com	Rigid-body assembly
WHATIF	http://swift.cmbi.kun.nl/whatif/	Rigid-body assembly
SEGMOD	Module in Look, sold to Celera in 1999	Segment matching
DRAGON	Contact Robin Munro at rmunro@nimr.mrc.ac.uk	Spatial restraints
ICM	http://www.molsoft.com/	Rigid-body assembly
3D-JIGSAW	http://www.bmm.icnet.uk/servers/3djigsaw/	Rigid-body assembly
Builder	Contact Koehl P at koehl@csb.stanford.edu	Self-Consistent Mean Field Approach
PrISM	http://trantor.bioc.columbia.edu/programs/PrISM/index.html	Rigid-body assembly

Table 2

Loop Modeling Program

Programs	Availability	Methods
LOOPY	http://trantor.bioc.columbia.edu/programs.html	Colony energy with ab-initio conformation sampling and torsional space minimization
PLOP	http://francisco.compbio.ucsf.edu/~jacobson/plop_manual/plop_overview.htm	Extensive conformation sampling, OPLS energy, sufficient energy minimization
COILS	http://www.ch.embnet.org/software/COILS_form.html	Scan database of known loops from PDB.
MODELLER (loop module)	http://guitar.rockefeller.edu/modeller/modeller.html	Ab-initio conformation sampling plus CHARMM force fields
CODA	http://www-cryst.bioc.cam.ac.uk/coda/	Combine database and ab-initio approach for loop modeling

Table 3

Side Chain Modeling Program

Programs	Availability	Methods
SCAP	http://trantor.bioc.columbia.edu/programs/jackal/	Colony energy method with simple energy and large Cartesian-coordinate rotamer library
SCWRL	http://dunbrack.fccc.edu/SCWRL3.php	Simple energy with backbone-dependent rotamer library
SMOL	Contact Grishin N.V. at Nikolai.Grichine@UTSouthwestern.Edu	Optimized scoring function with extended backbone-dependent rotamer library and Monte Carlo search method
SCCOMP	http://atlantis.weizmann.ac.il/~eyale/	Optimized scoring function and Gibbs sampling like algorithm
RAMP	http://www.ram.org/computing/ramp/ramp.html	knowledge based potentials and small rotamer library
SMD	http://condor.urbb.jussieu.fr/Smd.php	Flex force field, small rotamer library and dynamic cluster analysis of known structures
Confmat	Contact Koehl P at koehl@csb.stanford.edu	self-consistent mean field and small rotamer library
Maxsprout	http://www.ebi.ac.uk/maxsprout/	Rough energy function and small rotamer library

Table 4

Model Assessment Program

Programs	Availability	Quality to be checked
PROCHECK	http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html	Stereochemistry
WHATCHECK	http://www.sander.embl-heidelberg.de/whatcheck/	Stereochemistry, nomenclature, symmetry, packing, inside/outside profile, missing atoms/residues, hydrogen bonds, etc
ProsaII	http://www.came.sbg.ac.at	Mis-folded structures, faulty parts of structural models
VERIFY3D	http://www.doe-mbi.ucla.edu/Services/Verify_3D/	Residue's fitness in the model environment
ERRAT	http://www.doe-mbi.ucla.edu/Services/Errat.html	Statistical non-bonded atom-atom interactions
ANOLEA	http://www.fundp.ac.be/pub/ANOLEA.html	Non-local environment of heavy atoms
AQUA	http://www.nmr.chem.uu.nl/users/jurgen/Aqua/server/	Violation, completeness and redundancy of NOE distance restraints
Probe	http://kinemage.biochem.duke.edu/software/probe.php	atomic packing, either within or between molecules.
SQUID	http://www.ysbl.york.ac.uk/~oldfield/squid/	e analysis and display of data from crystallography and molecular dynamics
PROVE	http://www.ucmb.ulb.ac.be/UCMB/PROVE	Check the quality of the atomic model of a macromolecule structure based on the calculations of atomic volumes.
GRASP2	http://trantor.bioc.columbia.edu/programs.html	Graphic display model structure and sequence-template alignment