# Homology Modeling: Generating Structural Models to Understand Protein Function and Mechanism

**Srinivas Ramachandran and Nikolay V. Dokholyan**

## 1 Homology Models: Need and Applicability

Geneticists and molecular and cell biologists routinely uncover new proteins important in specific biological processes/pathways. However, either the molecular functions or the functional mechanisms of many of these proteins are unclear due to a lack of knowledge of their atomic structures. Yet, determining experimental structures of many proteins presents technical challenges. The current methods for obtaining atomic-resolution structures of biomolecules (X-ray crystallography and NMR spectroscopy) require pure preparations of proteins at concentrations much higher than those at which the proteins exist in a physiological environment. Additionally, NMR has size limitations, with current technology limited to the determination of structures of proteins with masses of up to 15 kDa. Due to these reasons, atomic structures of many medically and biologically important proteins do not exist. However, the structures of these proteins are essential for several purposes, including *in silico* drug design [1], understanding the effects of disease mutations [2], and designing experiments to probe the functional mechanisms of proteins. Comparative modeling has gained importance as a tool for bridging the gap between sequence and structure space, allowing researchers to build structural models of proteins that are difficult to crystallize or for which structure determination by

S. Ramachandran • N.V. Dokholyan (✉)
Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, NC 27599, USA

Molecular and Cellular Biophysics Program, University of North Carolina, Chapel Hill, NC 27599, USA
e-mail: ramachan@email.unc.edu

N.V. Dokholyan
Center for Computational and Systems Biology, University of North Carolina, Chapel Hill, NC 27599, USA
e-mail: dokh@unc.edu

N. Dokholyan (ed.), *Computational Modeling of Biological Systems: From Molecules to Pathways*, Biological and Medical Physics, Biomedical Engineering,
DOI 10.1007/978-1-4614-2146-7_5, © Springer Science+Business Media, LLC 2012

NMR spectroscopy is not tractable. Comparative modeling, or homology modeling, exploits the fact that two proteins whose sequences are evolutionarily connected display similar structural features [3]. Thus, the known structure of a protein (template) can be used to generate a molecular model of the protein (query) whose experimental structure is not known.

The applicability of comparative modeling in structural biology has been validated by the observations of several groups, e.g., that a limited number of protein folds are observed in nature [4, 5] and that nature is able to reuse similar folds for diverse protein functions [6]. Thus, several researchers have used the already available breadth of structural information to build structural models of many proteins whose experimental structures have not been determined. For example, ModBase [7] and SWISS-MODEL [8], repositories of comparative models generated using automated protocols, have structural models for 3.4 million and 2.2 million unique sequences respectively; for comparison, the repository for experimental structures, protein data bank (PDB [9]) has 67,728 experimental structures. The burgeoning number of structural models in repositories such as ModBase and SWISS-MODEL reflects the usefulness of comparative modeling in significantly closing the gap between the number of known sequences and known structures. To further close this gap, the protein structure initiative [10] aims to determine the experimental structures for representative members of protein families that do not yet have any structural templates in the PDB.

Structural models generated by homology modeling can be of direct medical and biological relevance. Structural models can be used to predict the effects of single nucleotide polymorphisms uncovered from genome-wide association studies, helping to delineate the molecular etiology of genetically transmitted diseases [2]. Homology-based structural models have already been used widely in *in silico* drug screening [11–13]. For biological experiments, structural models can be used to design mutations that lead to specific changes in the function or stability of the modeled protein [14, 15]. Importantly, homology models can be used as starting models for molecular replacement in X-ray crystallography [16], leading to better experimental structures. Furthermore, these structural models can be used in conjunction with methods such as FRET that provide interresidue distances and for mapping residue-level experimental data, such as accessibility measured through EPR [17] and H-D exchange mass spectrometry [18, 19].

Thus, to better understand the function and mechanism of a given protein of unknown structure, researchers can generate structural models using comparative modeling. In this chapter, we discuss the process of generating a homology-based structural model of a protein of interest. In particular, we focus on the critical controls and tests to be used at each step of model building to ensure that the final model is physically and biologically reasonable and, most importantly, to determine the extent to which the given model can be used in interpretations of experimental data. Comparative modeling involves several steps, such as identification of the template, sequence threading, processing insertions and deletions, model optimization, quality control, and finally, model interpretation (Fig. 1, Table 1). We discuss each of these steps in the following sections.

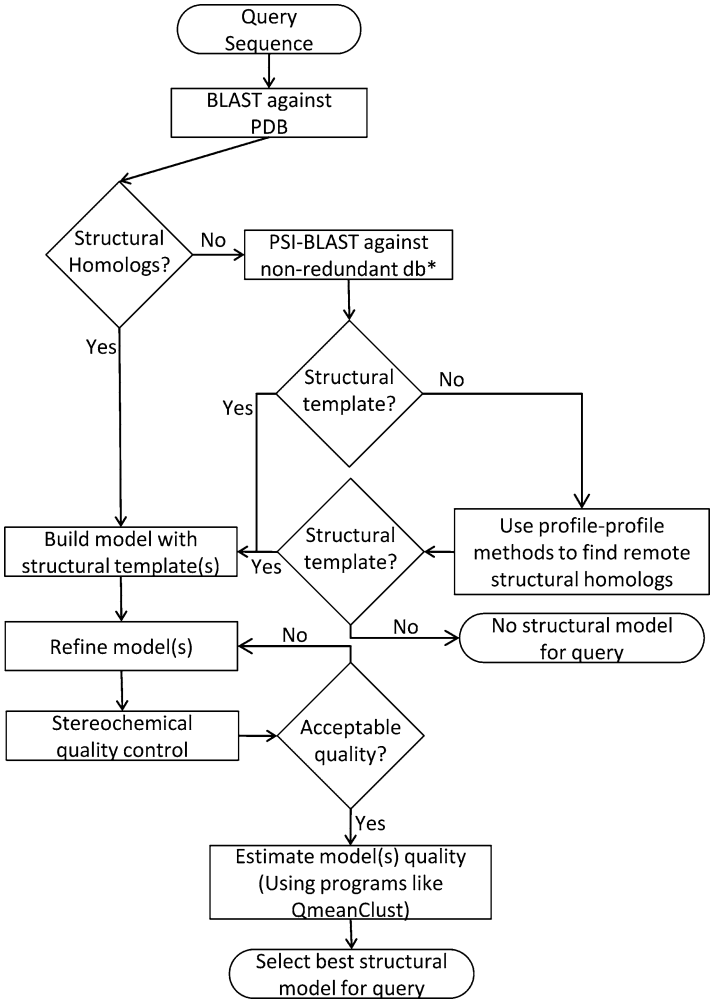


**Fig. 1** Flowchart of the steps followed in the construction of a comparative structural model (*database)

## 2 Template Identification

### *2.1 Domain Delineation*

One of the first steps to be performed with the query sequence is to determine the number of domains in the sequence. In many multidomain proteins, a single structural template covering the whole sequence may not be available. Instead, templates for each of the domains may be available. Many programs that employ

**Table 1** Some representative methods for the different steps involved in the construction of a comparative structural model

| Procedure | Server |
|---|---|
| Identify homologous sequences | BLAST [22], PSI-BLAST [23] |
| Protein family classifications | Pfam [21], InterPro [20] |
| Profile-based | HMMER [25], HHSearch [27], SAM [26] |
| Threading + profile-based | FUGUE (based on structure profile created by HOMSTRAD) [32], PROSPECT [33], SPARKS2 [34] and SP3 [35] |
| Profile-based + secondary structure prediction | PPA [76] |
| Meta servers | TASSER [77], I-TASSER [40], Bioinfobank [39] |
| Stereochemical quality control | Gaia [62], WHAT IF [61], PROCHECK [65], MolProbity [64] |
| Estimating model quality | Qmean [55], QmeanClust [60] |

machine learning approaches can be used to delineate domain boundaries in a protein sequence and even identify the potential function of the identified domains. InterProScan [20] and Pfam [21] are two databases available online that one can use to find the domains present in the query sequence. For some multidomain query sequences, one may be able to find structural templates with similar domain architecture, which will be the ideal scenario, but in others one may have to model individual domains separately and look for experimental constraints to model domain–domain orientations.

## *2.2 Direct Sequence Homology: BLAST and PSI-BLAST*

BLAST (basic alignment and search tool) [22] is a powerful and efficient tool to discover the evolutionary connections of a given protein sequence. Given a protein sequence of interest, any current researcher will first and foremost employ BLAST to search for homologs in all available sequence databases to uncover the functional and evolutionary details of the protein sequence. In the context of comparative modeling, BLAST helps in the identification of the structural template on which to base the structural model for a given sequence. While using the protein BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins), one can specify the sequence databases that should be searched; for comparative modeling, one usually chooses the PDB. In the context of BLAST, the PDB sequence database contains all the sequences that have an associated experimental structure. The match between BLAST "hits" and a given sequence are described by three parameters: similarity, coverage, and expect value ($E$-value). All three parameters are important in selecting the best template for a given sequence. A minimum of 30% similarity between query and template is essential for unambiguous alignment that can be used for generating a homology model. For each domain, at least 70% sequence coverage

is required. The extent of coverage determines the number of residues that need to be modeled without prior knowledge of their backbone coordinates. There are exceptions for the lower bounds of both similarity and query coverage, which will be discussed under remote homology, but if one were to choose a template based on BLAST results alone, the lower bounds for similarity and coverage are to be followed strictly to obtain unambiguous structural models. The $E$-value provides the statistical significance of a "hit" and describes the number of hits that can be obtained by chance in a given database with a given score. Thus, lower the $E$-value, greater the significance of a given hit. Generally, $E$-values less than 0.01 are considered significant for generating homology models.

If no homologs in the PDB are detected using BLAST for a given sequence, the alternative strategy is to use position-specific iterated-BLAST (PSI-BLAST) [23]. PSI-BLAST constructs a position-specific scoring matrix (PSSM) using the multiple sequence alignment of BLAST hits detected above a certain threshold (based on $E$-value). The PSSM is then used for searching the database. The construction of the PSSM and the subsequent database search are performed iteratively for several rounds until no new sequences are found. By using information from all the BLAST hits of a given iteration, PSI-BLAST helps uncover distant homologs. In determining the optimal template using PSI-BLAST, one uses the same thresholds for sequence coverage, similarity, and $E$-value that were discussed for BLAST.

Once a suitable template is identified, it is worthwhile to closely analyze the sequence alignment between the query and template. Analysis from several rounds of critical assessment of structure prediction (CASP) [24] has shown that the sequence alignment between query and template is the most important step in comparative modeling. The most prominent inaccuracies in homology models arise from inaccurate sequence alignment rather than errors in subsequent steps of structure building. Significantly, BLAST scoring matrices and PSSMs may not incorporate subtle structural details pertinent to the given protein like the positioning of structurally important cysteine disulphide bridges, proline residues, residues important in protein function, etc. In cases where the positioning of these residues is known to be important based on experimental data, one should manually edit the alignment to ensure that these residue positions are preserved between the query and template. Thus, one should consider all available functional, biochemical, and structural data of all possible residues in the query sequence while scrutinizing and updating the sequence alignment between the query and the template.

## *2.3 Remote Homology*

If a template is not detectable with BLAST or PSI-BLAST, one needs to use programs that are capable of identifying distant evolutionary relationships. It has been shown that two proteins can share a high degree of structural similarity in spite of the lack of detectable sequence similarity [6]. The lack of sequence similarity in these cases highlights high divergence of the sequences and also

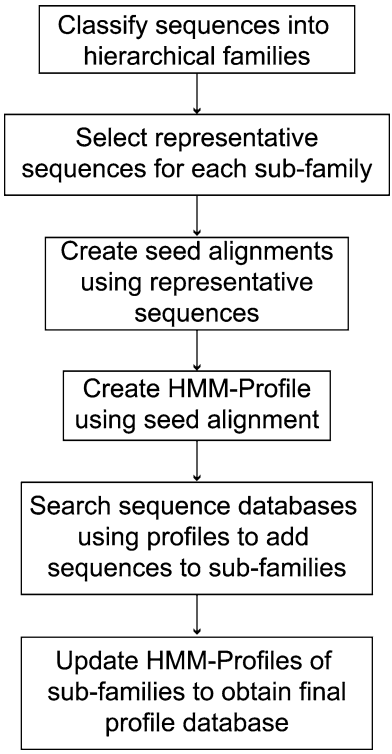


**Fig. 2** Construction of profile databases. The scheme illustrates the steps involved in constructing profile databases based on sequence alone

the weakness of our current metric, namely sequence similarity, in identifying distantly related sequences. To account for the observations of distant relationships between protein sequences and to utilize these relationships in protein structure and function prediction, many programs and servers have been developed that detect remote homologs. Even though most of these servers have easy to use interfaces that do not require any knowledge of the underlying computation, in order to discriminate between different identified templates (either by a single program or multiple programs), one needs to have a clear understanding of the algorithms and guiding principles used in these programs. Hence, we give a brief overview of the underlying principles of two important classes of bioinformatics approaches that are used in the detection of remote homologs: sequence-profile-based methods and structural-profile-based methods. Many subsequent approaches have combined the sequence-profile and structural-profile-based methods to increase the robustness of identifying and aligning distantly related proteins.

Using sequence-profiles (Fig. 2) for discovering remote homologs has been achieved using techniques such as hidden markov models (HMM), neural networks, and support vector machines. By far, HMM-based techniques have been used most frequently in direct template detection [25–27]. Other methods have been used in the prediction of subcellular localization [28], secondary structure prediction [29, 30], residue environment prediction [29], and identification of transmembrane segments

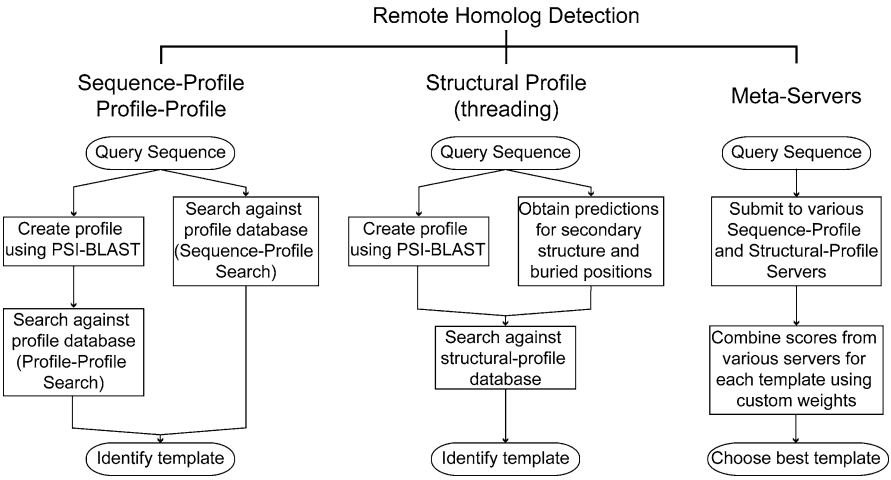


**Fig. 3** Remote homolog detection. An outline of different strategies involved in detecting remote homologs: sequence-based and structure-based methods and by using meta-servers. The steps involved in each of these strategies are also outlined as a flowchart

in protein sequence [31]. HMM-based methods rely on constructing an HMM-profile for any given sequence, based on a seed alignment generated either using BLAST or manually. Most of these methods have thousands of such profiles for all known sequences. Using these profiles, for any query sequence, sequence-profile and profile–profile matching can be performed to identify significant structural homologs. All known domains are usually arranged in hierarchical families based on either function or fold to enable quick retrieval of matches. A given sequence is searched against HMM-profiles of families that have at least one representative structure, a process called sequence-profile alignment (Fig. 3). A logical expansion of sequence-profile alignment is profile–profile alignment, where a profile is constructed based on evolutionary conservation of the query sequence. The seed alignment for constructing the profile for the query sequence is usually obtained using PSI-BLAST. Once an HMM-profile is generated for the query sequence using PSI-BLAST-based multiple sequence alignment, this profile is searched against other profiles that have at least one representative structure. Apart from providing a template structure for constructing a homology model, these profile–profile and sequence-profile alignments provide a quick means to predict domain boundaries and possible function of the sequence. For example, scanning the query sequence using Pfam [21] (a database of HMM-profile based domain families) will identify the different domains in the sequence as well as possible functional and structural information of the identified domains.

Structure-based threading [32–35] forms the basis of the second group of protocols (Fig. 4). We can observe high diversity in the specific protocol followed by each structure-based threading program to identify remote homologs. Since each

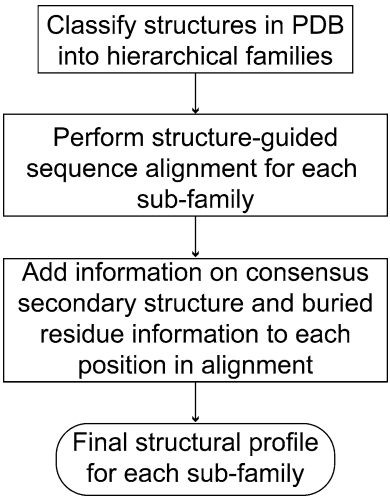


**Fig. 4** Construction of structural-profile databases. The scheme illustrates the steps involved in constructing structural-profile databases starting from the structures in PDB

threading program has its own optimized, intricate protocol and scoring system to identify structural templates, we only discuss the general principles underlying these programs rather than the scoring functions of specific programs. There are two groups of data available to the threading programs to generate an optimal alignment: data on the query sequence and data on all possible structural templates. Data on the query side consist of (Fig. 3) (1) the sequence-profile of the query sequence generated either using PSI-BLAST alone or PSI-BLAST and HMM programs and (2) the secondary structure propensity of each position of the query sequence, which can be determined using neural network or HMM-based programs such as PSIPRED [29] or Jpred [30]. Data on the template side are significantly richer. First, all known structures can be grouped into structural families based on structural similarity and a sequence alignment can be performed for sequences in each of these structural families. The sequence alignment, which is primarily based on the structural alignment, gives rise to residue propensities in each position of the fold, which we can denote as the structure-profile (Fig. 4). Second, one can obtain the secondary structure at each position of the fold using the dictionary of protein secondary structure(DSSP) program [36]. Third, one can obtain the environment of each position of the fold—whether it is buried or exposed, whether the backbone or side-chain are involved in any hydrogen bonds (Fig. 4). Fourth, distance or cut-off based residue–residue contact probability can be obtained in each structural family. These four pieces of information are used in a combinatorial fashion by different programs to match the two pieces of information available for the query sequence. Thus, each program uses a combination of terms that are optimally weighted to arrive at a final score that reflects the goodness of fit between a query sequence and a template structure (or a structural family, depending on the program). One way to align structure to sequence can be to match the structure-profile of the template (amino acid propensities in each position of the fold) to the

sequence-profile of the query (amino acid propensities at each position of the sequence based on evolution) using dynamic programming, with the gap-penalties at each template position set by the secondary structure at that position. It has been observed that insertions and deletions (which arise due to gaps in alignments) are minimal in positions that feature helix or strand [37], hence gap penalties can be set higher at positions in the template featuring helix or strand. Similarly, predictions for buried and exposed positions of the query sequence have also been used in sequence-structure alignments. To account for changes in fold topology, many programs also incorporate segmental threading [38] to arrive at a discontinuous sequence-structure alignment.

### *2.4 Meta Servers*

In the previous sections, we have introduced all the commonly used techniques for discovering structural templates. By far, the most successful approach in identifying a structural template for a given query sequence and in determining the best possible alignment between the sequence and structure has been to combine predictions from several diverse approaches. Servers such as 3D-Jury [39] and I-TASSER [40] have developed combined scoring functions that rate each structural template based on its scores in several profile–profile and structure-profile alignment programs to yield a consensus alignment (Fig. 3). For a researcher who is well versed in the biochemical and structural data connected to the query sequence, it is also possible to apply the available biochemical data as additional constraints in refining the consensus alignment between query and template.

## 3 From Alignment to a Structural Model

### *3.1 Model Construction*

Once a statistically significant alignment is obtained between query and template, construction of the homology model entails converting the template structure into a structural model for the query sequence. Model construction involves two important steps: first, in the regions of the template where alignment with the query exists, the sequence of the template has to be modified to the corresponding sequence of the query. Second, in the regions where alignment do not exist (insertions and deletions), either portions of the structure must be removed (deletion) or new structural fragments need to be built *de novo* (insertion). The first step requires only the modification of the side-chain atoms of the aligned positions as the amino acids in the template structure are morphed into the amino acids corresponding to the query sequence. With knowledge of the coordinates of the protein backbone,

positioning of a new side chain is straightforward [41]. Processing insertions is the most complicated step in model construction, since the positions of backbone atoms are not known for the inserted residues and must be modeled *de novo*. Insertions mostly involve internal loops that are longer in the query compared to the structural template. The methods to build these loops include ModLoop [42] (part of MODELLER [43]) which relies on satisfaction of spatial restraints, Hierarchical Loop Prediction with Surrounding Side chain optimization [44], kinematic closure protocol in Rosetta [45] and constrained all-atom DMD simulations [46]. Processing deletions involves the removal of residues in the template structure whose positions correspond to gaps in the query sequence in the sequence alignment. When a set of residues are removed, the ends of the deletion need to be connected by a peptide bond to ensure continuity of the protein chain. Several programs like all-atom DMD, MD, and Rosetta can be used to create the peptide bond between the ends of the deleted segment, with minimum perturbation to the backbone of the rest of the structure. Once the side chains are modified and the insertions and deletions are processed, one arrives at a complete (albeit initial) structural model for the query sequence. At this stage, if there are several templates that were identified, one can use the steps outlined above to construct several complete structural models for the query based on each of the template structures. In the case of several structural templates, one has to then choose among the many models for the same sequence, the structure that best represents the real structure corresponding to the query sequence, a process that we discuss in the section dealing with model quality. An illustration of the sequence/structure alignment and homology-based structural model is shown in Fig. 5.

## *3.2 Model Refinement*

Once a complete structural model for the query sequence is obtained, there are several possible steps by which the structural model can be refined to approach a physically accurate structure. Just modifying the side-chains and processing insertions and deletions as described above result in a model that has minimal changes from the structural template. However, with limited sequence identity (in many cases, only 30%), the structures of the template and the query will be expected to have significant conformational differences even though they share the same fold. For example, when transitioning from the template to the query sequence, many small to large amino acid changes in the core will require backbone perturbations to accommodate the large amino acid while retaining optimal packing in the core. Furthermore, homology models have been shown to have an excess of steric clashes and structural artifacts caused by unphysical overlap of newly positioned side-chain atoms with other side-chain and backbone atoms (Fig. 6a). Thus, most residues in the core will need to undergo concerted changes in the side-chain rotamers (changes in the $\chi$ angles), along with subtle changes in the protein backbone, to form a core that is optimally packed. Even though side-chain

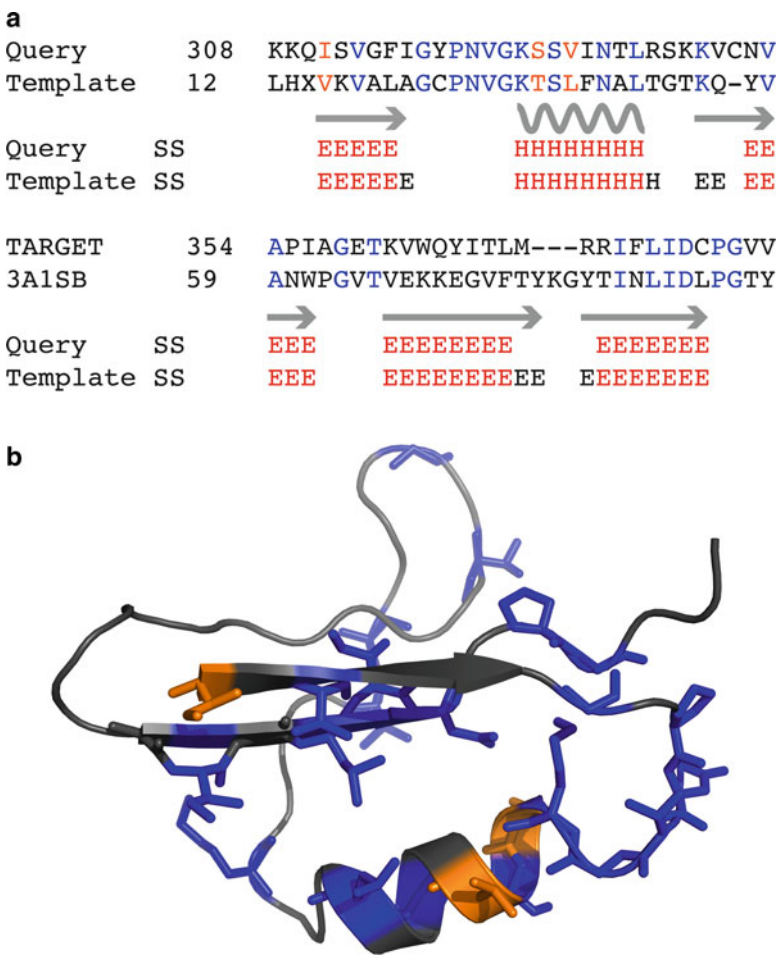


**Fig. 5** A homology-based structural model. The alignment (**a**) and the structural model (**b**) of the G-domain of nucleolar GTP-binding protein 2 (Uniprot ID Q13823, query) are shown as a representative homology-based structural model. The template chosen by HHSEARCH [27], as listed in the SWISS-MODEL database [8] has the PDB ID 3A1S. The residues identical between the query and template are colored *blue*. The similar residues are colored *orange*. The predicted secondary structure of the query and the observed secondary structure of the template are also shown; *H* denotes helix and *E* denotes strand. Note the high level of similarity in the predicted and observed secondary structure. The structural model retrieved from SWISS-MODEL database is rendered as a cartoon using PyMol (http://www.pymol.org), with the identical and similar residues rendered as sticks. The positions identical in the alignment are colored *blue* in the structure and the similar residues are colored *orange*

repacking can be performed with great accuracy and efficiency by many programs which fix the backbone position (using knowledge-based rotamer libraries), in this case, the core refinement is useful only when the repacking is coupled with subtle changes in backbone conformations.

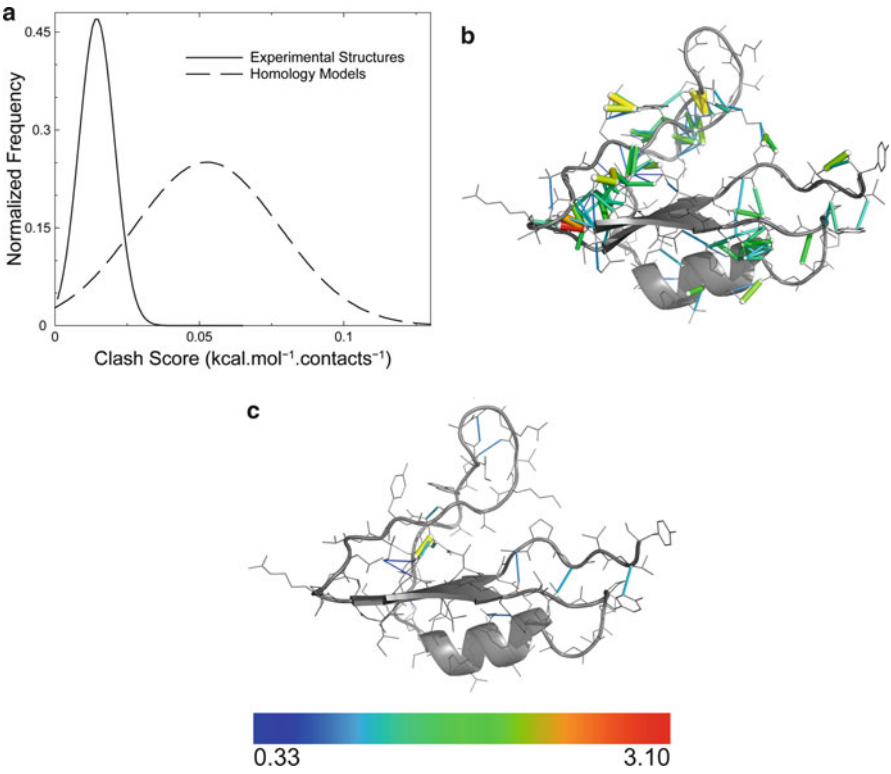


**Fig. 6** Steric clashes in homology models. Homology models on an average feature much higher extent of steric clashes when compared to experimental structures (**a**). The distribution of clash-scores (which is a normalized energetic parameter reflecting the extent of steric clashes in a protein structure [50]) of high-resolution crystal structures and representative homology models from SWISS-MODEL database are plotted. The efficacy of Chiron in minimizing clashes in protein structure is demonstrated for the homology model of Q13823, whose initial model (**b**) has a clash-ratio of 0.13, much higher than that seen in experimental structures. The protein structure is shown with the cartoon representation, rendered using PyMol (http://www.pymol.org). Clashes are denoted as *colored cylinders*, where both the colors and the thickness of the cylinders denote the van der Waals repulsion energy. The scale of the repulsion energy is shown as a gradient bar at the *bottom*, with the *numbers* at the ends indicating repulsion energy in kcal/mol. Note the large numbers of cylinders in the initial model, denoting excessive steric clashes. The minimized structure (**c**) has a clash-ratio of 0.018, within one standard deviation of the mean clash-score of high-resolution structures

The refined structural models by definition should feature physically reasonable backbone conformation and a well-packed core that has an acceptable extent of clashes. Minimal backbone perturbation to ensure ideal packing can be achieved by various means including “backrub” and knowledge-based backbone assembly (as used in Rosetta), all-atom DMD simulations and minimization using molecular mechanics forcefields. All these methods refine the structural model to the nearest local minima in the conformational space of the starting structure. Thus, if the

starting model is far away from the actual structure (>5Å root mean square deviation (RMSD)), these methods will be of limited utility in bringing the model closer to the actual structure. Steepest descent/conjugate gradient minimization using all-atom molecular mechanics force fields is the most widely used method to refine a structure while also resolving clashes. However, minimization using molecular mechanics may not resolve severe clashes in some cases, hampering subsequent molecular dynamics simulations. Use of molecular modeling tools such as Rosetta [47] is the alternate avenue for refining structures with severe clashes. These tools use knowledge-based potentials and small backbone moves to resolve clashes. However, these methods work best with smaller proteins (less than 250 residues in size). Tools such as MMTSB [48] and PULCHRA [49] have emerged for structure refinement, which includes removal of clashes during refinement. Chiron [50], an automated server evaluates the extent of clashes in a given structure and if required, minimizes these clashes to the levels seen in high-resolution X-ray structures. Chiron uses all-atom DMD [46, 51] with soft-core potentials. Additionally, Chiron uses a high coefficient of heat exchange of protein atoms with thermostat to ensure minimal perturbation of the protein backbone while resolving clashes in the protein. An example for clash minimization in a homology structural model using Chiron is shown in Fig. 6b, c. While Rosetta couples backbone moves to side-chain repacking, the other methods can also be used iteratively with side-chain repacking programs to achieve rigorous refinement. Side-chain repacking coupled with minimal backbone optimization to improve core packing transitions a complete structural model into a physically realistic model that can be used for further studies.

### *3.3 Estimating Model Quality*

The model quality can be classified into two types: (1) the stereochemical quality of the structural model and (2) the accuracy of the homology-based structural model with respect to its experimental structure. Given the lack of experimental structure for the query sequence, the real accuracy of a homology-based structural model cannot be assessed. To develop methods to predict this accuracy in the absence of known experimental structure, methods have been developed based on benchmark sets of structural models built for proteins with already existing experimental structures. In such cases where the experimental structure is known, there are several measures that estimate the model's quality. RMSD is the widely used measure to estimate the "structural similarity" between any two structures. However, large differences in positions of a small fraction of the proteins being compared can result in high values of RMSD, thus not reflecting the majority of the regions where the structures are highly similar. Thus, CASP competitions use another measure called global distance test (GDT), which is a measure of similarity of two structures with similar amino acid compositions but different tertiary structures. GDT is defined as the largest number of corresponding amino acids' alpha carbon atoms in the

compared structures that fall within a given distance cut-off. Usually, an average of GDT at 1, 2, 4, and 8 Å is used to measure structure quality, and is denoted as GDT_TS. Other variations include corrections to eliminate size-dependence [52] and also to include a negative term for incorrectly positioned amino acids (resulting in non-native contacts) [53]. There are various servers that have been developed to predict GDT_TS of a structural model in the absence of an experimental structure with which to compare.

These quality assessment programs either assess a structural model by itself (single model) [54–56], or in the context of a large reference set of structural models generated for a given sequence using different methods (consensus) [57–59]. The single model methods can compare various structural parameters such as secondary structure and solvent accessibility (whether a given residue is buried or exposed) that are predicted for a given sequence with the corresponding structural parameters of the given model. The scoring functions used in single-model quality assessment programs also include knowledge-based potentials like a sequence-dependent torsion term (usually in the context of three consecutive residues), distance and cut-off based residue–residue interaction potentials and all-atom interaction potentials. Consensus quality assessment programs rely on the idea that if a diverse set of methods were used in generating many structural models for a given query, models that incorporate the best of all methods will be the ones closest to the experimental structure. How do the consensus methods select models based on these criteria? Using a large reference set of models for a given sequence (obtained from the various structure prediction servers), they determine the average distance of a given model to all other models. The distance measure used in most cases is GDT_TS, although specific servers apply various modifications to the distance to obtain better predictive power. Through several rounds of CASP, it is apparent that the models with least average distance to the rest of the reference set feature the best GDT_TS when compared to the experimental structure. We have to emphasize here that there are many mathematical formulations used in modifying the simple average distance to obtain better predictions, but these formulations may not always have a strong physical basis. Interestingly, weighting the average distances with single-model score yields very good prediction of GDT_TS of a model with respect to the experimental structure [60]. Thus, based on the experience gained from CASP competitions, the ideal strategy for constructing a homology-based structural model would entail generating several models using heterogeneous methods, which can also include human intervention during model building and the incorporation of known experimental constraints. Once a handful of models are obtained, one can use the quality assessment programs to obtain a prediction of how close the best model will be to the experimental structure. The quality score will in turn determine the use to which a structural model can be put (discussed below).

Apart from model accuracy, an important criterion for model quality is the stereochemical quality of the given model. The stereochemical quality here broadly defines the acceptable quality of the covalent geometry and the core-packing of a given structural model. The covalent geometry of a structural model is assessed by comparing all its bond lengths, bond-angles, and torsions to standard values.

The standard values of bond lengths and angles are obtained from studies on model small molecules [61] or through surveys of high-resolution crystal structures [62]. The backbone torsions comprise of the $\varphi$, $\psi$, and $\omega$ angles. The $\varphi - \psi$ of each residue can be compared to the allowed $\varphi - \psi$ map obtained from survey of high-resolution crystal structures to detect outliers. The side-chain torsions are again compared to the rotamer libraries [63] to detect outliers. All these measures ensure that the covalent geometry of the structural model is physically acceptable. The packing quality can be assessed based on the extent of steric clashes [50], the prevalence of voids, and the scaling of the solvent accessible surface area with protein length [62]. There are several servers that can compare these structural parameters of a model with benchmark distribution to indicate the areas of the protein structure that need further refinement to be physically acceptable [61, 62, 64, 65]. Importantly, the stereochemical quality of the structural model is essential for further studies including molecular simulations.

## 4 Experimental Constraints to Improve/Verify Homology Models

Any experimental data that can be used as a structural parameter, even indirectly, aid in building a better structural model based on homology [66]. Once a structural model is available, further experiments can be designed using insights from the model. Thus, designing experiments using structural models and building models that satisfy experimental constraints become an iterative process leading to better understanding of a protein's structure–function relationships. The experimental constraints that can be used in model building are diverse and we discuss several examples here. Usually, experimental constraints are sparse and by themselves not enough to lead to an unambiguous structural model. Thus, several models can satisfy a given set of constraints. However, the subset of models that do not satisfy a given experimental constraint can be eliminated from consideration. The experimental constraints can either be at the residue level or provide overall structural information. Some of the residue-level constraints include distance bounds between specific residues obtained by FRET and site-directed cross-linking. Iterative model building is possible using FRET and site-directed cross-linking, since a structural model allows probing a much smaller subset of residue-pairs for distance measurements as opposed to residue-pairs being chosen randomly [2, 67]. Furthermore, these distance measurements provide direct validation of a given structural model. Residue accessibilities obtained by EPR spectroscopy [17] and H-D exchange mass spectrometry [18, 19] also aid in model refinement. Experiments that provide information on the overall protein structure include small angle X-ray scattering (SAXS) [68], cryo-electron microscopy (CryoEM) [69], and circular dichroism (CD) spectroscopy, among others. SAXS provides the molecular envelope or the overall shape of the protein in solution, which can

help in discriminating between structurally diverse templates used in generating the structural model. The utility of CryoEM in atomic-level structural modeling depends greatly on the resolution of the electron density map obtained for a given protein. Recent advances in CryoEM have led to subnanometer resolutions density maps that can be used to directly refine all-atom structural models [70]. CryoEM densities are usually deposited at the electron microscopy data bank (http://www.emdatabank.org/), and programs to perform flexible docking of a structural model to EM densities have been developed [71]. Thus, high-resolution cryoEM currently offers the best alternative to X-ray crystallography and NMR for obtaining accurate atomic structure of a given protein. CD spectroscopy is used to determine the secondary structure content of a given protein and CD measurements can be used to assess the overall accuracy of secondary structure content of the structural model. Indirect structural constraints include mutational studies of the protein that assess changes in function and stability. These constraints can be included in model building only qualitatively, but still provide means to eliminate inaccurate models.

## 5 Conclusions

Comparative modeling of protein structures offers an efficient alternative to experimental structure determination in cases where there are difficulties in obtaining experimental structures for a given protein. Usually, if one can find a structural template more than 50% identical to the query sequence, a model with an estimated RMSD of 1 Å to the experimental structure can be obtained [72]. Thus, in cases where significant homology to a structural template exists, comparative modeling is a powerful technique to better understand the structure–function relationships and functional mechanisms of a given protein. Importantly, for clinically relevant proteins that are hard to crystallize, like G-protein coupled receptors (GPCRs) and ion channels, landmark structural studies have provided a sufficient number of templates to model many variants. The structural models of these variants have been instrumental in furthering our knowledge of different functional mechanisms (in $K^+$ channels [73,74]) and in virtual-ligand screening (GPCRs [75]). Advances in structural understanding of GPCRs and ion channels represent the most prominent impact of comparative structural models. These models have been used in numerous other cases to yield biologically useful insights [1]. During the process of model building, we need to undertake several precautions and assess model quality at each step. Most importantly, all structural models need some form of experimental validation to gain relevance. Thus, an iterative cycle of model building and experimental verification provides the best scenario for furthering our understanding of structural and functional aspects of many biological proteins, whose experimental structures remain unsolved.

## References


1. Cavasotto, C.N., Phatak, S.S.: Homology modeling in drug discovery: current trends and applications. Drug Discov. Today **14**, 676–683 (2009)
2. Serohijos, A.W., Hegedus, T., Aleksandrov, A.A., He, L., Cui, L., Dokholyan, N.V., Riordan, J.R.: Phenylalanine-508 mediates a cytoplasmic-membrane domain contact in the CFTR 3D structure crucial to assembly and channel function. Proc. Natl. Acad. Sci. U. S. A. **105**, 3256–3261 (2008)
3. Chothia, C., Lesk, A.M.: The relation between the divergence of sequence and structure in proteins. EMBO J. **5**, 823–826 (1986)
4. Finkelstein, A.V., Ptitsyn, O.B.: Why do globular proteins fit the limited set of folding patterns? Prog. Biophys. Mol. Biol. **50**, 171–190 (1987)
5. Zhang, Y., Hubner, I.A., Arakaki, A.K., Shakhnovich, E., Skolnick, J.: On the origin and highly likely completeness of single-domain protein structures. Proc. Natl. Acad. Sci. U. S. A. **103**, 2605–2610 (2006)
6. Todd, A.E., Orengo, C.A., Thornton, J.M.: Evolution of function in protein superfamilies, from a structural perspective. J. Mol. Biol. **307**, 1113–1143 (2001)
7. Pieper, U., Webb, B.M., Barkan, D.T., Schneidman-Duhovny, D., Schlessinger, A., Braberg, H., Yang, Z., Meng, E.C., Pettersen, E.F., Huang, C.C., et al.: ModBase, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res. **39**, D465–474 (2011)
8. Kiefer, F., Arnold, K., Kunzli, M., Bordoli, L., Schwede, T.: The SWISS-MODEL Repository and associated resources. Nucleic Acids Res., **37**, D387–392 (2009)
9. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The protein data bank. Nucleic Acids Res. **28**, 235–242 (2000)
10. Chandonia, J.M., Brenner, S.E.: The impact of structural genomics: expectations and outcomes. Science **311**, 347–351 (2006)
11. Becker, O.M., Dhanoa, D.S., Marantz, Y., Chen, D., Shacham, S., Cheruku, S., Heifetz, A., Mohanty, P., Fichman, M., Sharadendu, A., et al.: An integrated in silico 3D model-driven discovery of a novel, potent, and selective amidosulfonamide 5-HT1A agonist (PRX-00023) for the treatment of anxiety and depression. J. Med. Chem. **49**, 3116–3135 (2006)
12. Brylinski, M., Skolnick, J.: Q-Dock: low-resolution flexible ligand docking with pocket-specific threading restraints. J. Comput. Chem. **29**, 1574–1588 (2008)
13. Ekins, S., Mestres, J., Testa, B.: In silico pharmacology for drug discovery: applications to targets and beyond. Br. J. Pharmacol. **152**, 21–37 (2007)
14. Labro, A.J., Boulet, I.R., Choveau, F.S., Mayeur, E., Bruyns, T., Loussouarn, G., Raes, A.L., Snyders, D.J.: The S4-S5 linker of KCNQ1 channels forms a structural scaffold with the S6 segment controlling gate closure. J. Biol. Chem. **286**, 717–725 (2011)
15. Szklarz, G.D., Halpert, J.R.: Use of homology modeling in conjunction with site-directed mutagenesis for analysis of structure-function relationships of mammalian cytochromes P450. Life Sci. **61**, 2507–2520 (1997)
16. Claude, J.B., Suhre, K., Notredame, C., Claverie, J.M., Abergel, C.: CaspR: a web server for automated molecular replacement using homology modelling. Nucleic Acids Res. **32**, W606–W609 (2004)
17. Dong, J., Yang, G., McHaourab, H.S.: Structural basis of energy transduction in the transport cycle of MsbA. Science **308**, 1023–1028 (2005)
18. Chung, E.W., Nettleton, E.J., Morgan, C.J., Gross, M., Miranker, A., Radford, S.E., Dobson, C.M., Robinson, C.V.: Hydrogen exchange properties of proteins in native and denatured states monitored by mass spectrometry and NMR. Protein Sci. **6**, 1316–1324 (1997)
19. Engen, J.R., Smith, D.L.: Investigating protein structure and dynamics by hydrogen exchange MS. Anal. Chem. **73**, 256A–265A (2001)
20. Zdobnov, E.M., Apweiler, R.: InterProScan – an integration platform for the signature-recognition methods in InterPro. Bioinformatics **17**, 847–848 (2001)

21. Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., et al.: The Pfam protein families database. Nucleic Acids Res. **38**, D211–D222 (2010)
22. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. J. Mol. Biol. **215**, 403–410 (1990)
23. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**, 3389–3402 (1997)
24. Jauch, R., Yeo, H.C., Kolatkar, P.R., Clarke, N.D.: Assessment of CASP7 structure predictions for template free targets. Proteins **69**(Suppl 8), 57–67 (2007)
25. Finn, R.D., Clements, J., Eddy, S.R.: HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. **39**, W29–W37 (2011)
26. Karplus, K., Barrett, C., Hughey, R.: Hidden Markov models for detecting remote protein homologies. Bioinformatics **14**, 846–856 (1998)
27. Soding, J.: Protein homology detection by HMM-HMM comparison. Bioinformatics **21**, 951–960 (2005)
28. Nielsen, H., Engelbrecht, J., Brunak, S., von Heijne, G.: Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng. **10**, 1–6 (1997)
29. Jones, D.T.: Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. **292**, 195–202 (1999)
30. Cole, C., Barber, J.D., Barton, G.J.: The Jpred 3 secondary structure prediction server. Nucleic Acids Res. **36**, W197–W201 (2008)
31. Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L.: Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. **305**, 567–580 (2001)
32. Shi, J., Blundell, T.L., Mizuguchi, K.: FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J. Mol. Biol. **310**, 243–257 (2001)
33. Xu, Y., Xu, D.: Protein threading using PROSPECT: design and evaluation. Proteins **40**, 343–354 (2000)
34. Zhou, H., Zhou, Y.: Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins **55**, 1005–1013 (2004)
35. Zhou, H., Zhou, Y.: Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins **58**, 321–328 (2005)
36. Kabsch, W., Sander, C.: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers **22**, 2577–2637 (1983)
37. Pascarella, S., Argos, P.: Analysis of insertions/deletions in protein structures. J. Mol. Biol. **224**, 461–471 (1992)
38. Wu, S., Zhang, Y.: Recognizing protein substructure similarity using segmental threading. Structure **18**, 858–867 (2010)
39. Ginalski, K., Elofsson, A., Fischer, D., Rychlewski, L.: 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics **19**, 1015–1018 (2003)
40. Roy, A., Kucukural, A., Zhang, Y.: I-TASSER: a unified platform for automated protein structure and function prediction. Nat. Protoc. **5**, 725–738 (2010)
41. Dunbrack, R.L., Jr., Karplus, M.: Backbone-dependent rotamer library for proteins. Application to side-chain prediction. J. Mol. Biol. **230**, 543–574 (1993)
42. Fiser, A., Do, R.K., Sali, A.: Modeling of loops in protein structures. Protein Sci. **9**, 1753–1773 (2000)
43. Fiser, A., Sali, A.: Modeller: generation and refinement of homology-based protein structure models. Methods Enzymol. **374**, 461–491 (2003)
44. Sellers, B.D., Zhu, K., Zhao, S., Friesner, R.A., Jacobson, M.P.: Toward better refinement of comparative models: predicting loops in inexact environments. Proteins **72**, 959–971 (2008)

45. Mandell, D.J., Coutsias, E.A., Kortemme, T.: Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. Nat. Methods **6**, 551–552 (2009)
46. Ding, F., Tsao, D., Nie, H., Dokholyan, N.V.: Ab initio folding of proteins with all-atom discrete molecular dynamics. Structure **16**, 1010–1018 (2008)
47. Kaufmann, K.W., Lemmon, G.H., Deluca, S.L., Sheehan, J.H., Meiler, J.: Practically useful: what the Rosetta protein modeling suite can do for you. Biochemistry **49**, 2987–2998 (2010)
48. Feig, M., Karanicolas, J., Brooks, C.L. III.: MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. J. Mol. Graph. Model **22**, 377–395 (2004)
49. Rotkiewicz, P., Skolnick, J.: Fast procedure for reconstruction of full-atom protein models from reduced representations. J. Comput. Chem. **29**, 1460–1465 (2008)
50. Ramachandran, S., Kota, P., Ding, F., Dokholyan, N.V.: Automated minimization of steric clashes in protein structures. Proteins **79**, 261–270 (2011)
51. Dokholyan, N.V., Buldyrev, S.V., Stanley, H.E., Shakhnovich, E.I.: Discrete molecular dynamics studies of the folding of a protein-like model. Fold Des. **3**, 577–587 (1998)
52. Zhang, Y., Skolnick, J.: Scoring function for automated assessment of protein structure template quality. Proteins **57**, 702–710 (2004)
53. Sadreyev, R.I., Shi, S., Baker, D., Grishin, N.V.: Structure similarity measure with penalty for close non-equivalent residues. Bioinformatics **25**, 1259–1263 (2009)
54. Eramian, D., Eswar, N., Shen, M.Y., Sali, A.: How well can the accuracy of comparative protein structure models be predicted? Protein Sci. **17**, 1881–1893 (2008)
55. Benkert, P., Tosatto, S.C., Schomburg, D.: QMEAN: a comprehensive scoring function for model quality assessment. Proteins **71**, 261–277 (2008)
56. Wallner, B., Elofsson, A.: Can correct protein models be identified? Protein Sci. **12**, 1073–1086 (2003)
57. Wallner, B., Elofsson, A.: Pcons5: combining consensus, structural evaluation and fold recognition scores. Bioinformatics **21**, 4248–4254 (2005)
58. McGuffin, L.J., Roche, D.B.: Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. Bioinformatics **26**, 182–188 (2010)
59. Cheng, J., Wang, Z., Tegge, A.N., Eickholt, J.: Prediction of global and local quality of CASP8 models by MULTICOM series. Proteins **77**(Suppl 9), 181–184 (2009)
60. Benkert, P., Schwede, T., Tosatto, S.C.: QMEANclust: estimation of protein model quality by combining a composite scoring function with structural density information. BMC Struct. Biol. **9**, 35 (2009)
61. Hooft, R.W.W., Vriend, G., Sander, C., Abola, E.E.: Errors in protein structures. Nature **381**, 272–272 (1996)
62. Kota, P., Ding, F., Ramachandran, S., Dokholyan, N.V.: Gaia: automated quality assessment of protein structure models. Bioinformatics **27**, 2209–2215 (2011)
63. Dunbrack, R.L., Jr., Cohen, F.E.: Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci. **6**, 1661–1681 (1997)
64. Davis, I.W., Leaver-Fay, A., Chen, V.B., Block, J.N., Kapral, G.J., Wang, X., Murray, L.W., Arendall, W.B. III, Snoeyink, J., Richardson, J.S., et al.: MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucliec. Acids Res. **35**, W375–W383 (2007)
65. Laskowski, R.A., MacArthur, M.W., Moss, D.S., Thornton, J.M.: PROCHECK: a program to check the stereochemical quality of protein structures. J.Appl. Crystallogr., **26**, 283–291 (1993)
66. Alber, F., Forster, F., Korkin, D., Topf, M., Sali, A.: Integrating diverse data for structure determination of macromolecular assemblies. Annu. Rev. Biochem. **77**, 443–477 (2008)
67. Tung, C.S., Wall, M.E., Gallagher, S.C., Trewhella, J.: A model of troponin-I in complex with troponin-C using hybrid experimental data: the inhibitory region is a beta-hairpin. Protein Sci. **9**, 1312–1326 (2000)

68. Schneidman-Duhovny, D., Hammel, M., Sali, A.: FoXS: a web server for rapid computation and fitting of SAXS profiles. Nucleic Acids Res. **38**, W540–W544 (2010)
69. Baker, M.L., Zhang, J., Ludtke, S.J., Chiu, W.: Cryo-EM of macromolecular assemblies at near-atomic resolution. Nat. Protoc. **5**, 1697–1708 (2010)
70. Cong, Y., Baker, M.L., Jakana, J., Woolford, D., Miller, E.J., Reissmann, S., Kumar, R.N., Redding-Johanson, A.M., Batth, T.S., Mukhopadhyay, A., et al.: 4.0-A resolution cryo-EM structure of the mammalian chaperonin TRiC/CCT reveals its unique subunit arrangement. Proc. Natl. Acad. Sci. U. S. A. **107**, 4967–4972 (2010)
71. Rusu, M., Birmanns, S., Wriggers, W.: Biomolecular pleiomorphism probed by spatial interpolation of coarse models. Bioinformatics **24**, 2460–2466 (2008)
72. Bordoli, L., Kiefer, F., Arnold, K., Benkert, P., Battey, J., Schwede, T.: Protein structure homology modeling using SWISS-MODEL workspace. Nat. Protoc. **4**, 1–13 (2009)
73. Silva, J.R., Pan, H., Wu, D., Nekouzadeh, A., Decker, K.F., Cui, J., Baker, N.A., Sept, D., Rudy, Y.: A multiscale model linking ion-channel molecular dynamics and electrostatics to the cardiac action potential. Proc. Natl. Acad. Sci. U. S. A. **106**, 11102–11106 (2009)
74. Smith, J.A., Vanoye, C.G., George, A.L. Jr., Meiler, J., Sanders, C.R.: Structural models for the KCNQ1 voltage-gated potassium channel. Biochemistry **46**, 14141–14152 (2007)
75. Katritch, V., Rueda, M., Lam, P.C., Yeager, M., Abagyan, R.: GPCR 3D homology models for ligand screening: lessons learned from blind predictions of adenosine A2a receptor complex. Proteins **78**, 197–211 (2010)
76. Wu, S., Skolnick, J., Zhang, Y.: Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol. **5**, 17 (2007)
77. Zhou, H., Skolnick, J.: Ab initio protein structure prediction using chunk-TASSER. Biophys. J. **93**, 1510–1518 (2007)