Briefings in Bioinformatics, 18(6), 2017, 1021–1032

doi: 10.1093/bib/bbw070
Advance Access Publication Date: 25 August 2016
Paper

Recent advances in sequence-based protein structure prediction

Dukka B. KC

Corresponding author: Dukka B KC, Department of Computational Science and Engineering, North Carolina A&T State University, Greensboro, NC 27411, USA. Tel.: (336)285-3210; Fax: (336)256-1247; E-mail: dbkc@ncat.edu

Abstract

The most accurate characterizations of the structure of proteins are provided by structural biology experiments. However, because of the high cost and labor-intensive nature of the structural experiments, the gap between the number of protein sequences and solved structures is widening rapidly. Development of computational methods to accurately model protein structures from sequences is becoming increasingly important to the biological community. In this article, we highlight some important progress in the field of protein structure prediction, especially those related to free modeling (FM) methods that generate structure models without using homologous templates. We also provide a short synopsis of some of the recent advances in FM approaches as demonstrated in the recent Computational Assessment of Structure Prediction competition as well as recent trends and outlook for FM approaches in protein structure prediction.

Key words: protein structure prediction; free modeling; template-based modeling; contact prediction; evolutionary constraints; CASP.

Introduction

Owing to the significant improvement in genome sequencing technologies and efforts, the genomic sequences of a large number of organisms have now been determined. As of August 2015, 187 million sequences from over 500 000 organisms have been deposited in Genbank databases [1]. Among them, 50 million sequences have been translated into protein amino acid sequences and stored in the UniprotKB/TrEMBL database [2]. However, sequences alone do not provide insight into what each protein does in living cells, and the three-dimensional (3D) structure of these proteins is often important for interpreting their biological roles.

Structural biology techniques such as Nuclear magnetic resonance (NMR), X-ray crystallography and Cryo-EM provide the most accurate characterization of the protein structure. However, because of the technical difficulties associated with cost and time, the gap between the number of protein sequences and that of protein structures is rapidly expanding. As of August 2015, there are only ∼100 000 proteins whose structures have been experimentally solved in Protein Data Bank (PDB) [3], compared with50 million protein sequences in UniprotKB [2]. Therefore, solved structures only account for ∼0.2% of the known sequences. One promising approach to close this gap is the development of computational approaches that are able to generate high-resolution structural models for sequences that can be conveniently used by the biological community.

Types of computational approaches for protein structure prediction

Historically, computational approaches for protein structure prediction have been classified into three categories [4]: comparative modeling (CM or homology modeling) [5], threading [6] and free modeling (FM or *ab initio*) [7] approaches. In CM, the structure of a query protein sequence is constructed by first matching the query sequence to an evolutionarily related protein whose structure has already been solved, where the residue equivalency is obtained by aligning the sequences or sequence profiles [8–10]. Threading approaches are designed to match the query sequence directly to the 3D structures of solved proteins with the goal of recognizing similar protein folds even when there is no evolutionary relationship to the query. Finally, FM

D. B. KC is an assistant professor of Computational Science and Engineering at North Carolina A&T State University, working on developing computational tools to decipher relationship between protein sequence, structure, function and evolution.
Submitted: 24 March 2016; **Received (in revised form):** 27 June 2016

approaches are designed for query without structurally related solved proteins, where models are built from scratch by *ab initio* folding methods.

There are two kinds of qualitatively different classes of approach for structure modeling: comparative modeling and *de novo* methods. Comparative modeling approach in one way or the other make use of structural templates and in *de novo* methods attempt to predict the structure from 1st principles and without structural templates. The Computational Assessment of Structure Prediction (CASP) categorizes targets into two groups based on the availability of structural templates for a given target: (i) template-based modeling (TBM) and (ii) FM. TBM refers to the case where structural templates are available for the target, and FM refers to the case for which there are no template structures available [11].

FM category in CASP assesses methods that predict 3D structures from a given protein sequence without the explicit use of template structures available in PDB [12, 13]. It has to be noted that historically, the most successful methods for such *de novo* structure prediction from sequence used fragment assembly [14, 15]. However, the recent top modeling groups in CASP9 and CASP10 incorporated the use of remote templates or selection and refinement of server models [12, 13]. This trend continued in CASP11 where some of the best-performing methods like Baker-Rosetta server declared parent templates for 47% of their models and Zhang-Servers declared 17% parent templates in FM-only targets [16].

The classification of these computational approaches is becoming increasingly blurry. For example, both CM and threading methods use sequence profile alignments [17, 18] and FM approaches often use evolutionary and threading information [19, 20]. In this respect, various composite approaches have been developed that combine different methods from CM, threading and FM approaches. ROSETTA[21] and I-TASSER [22] are some typical examples of composite approaches to protein tertiary structure prediction.

The average accuracy of current protein structure prediction approaches is highly correlated to the evolutionary distance between target (query) sequence and template structure [23, 24]. When the query sequence has a sequence identity >50% to the template(s), models built by CM can have a backbone root mean squared distance (RMSD) as low as 1.0 Å. Similarly, when the target sequence has sequence identity between 30 and 50% to the templates, the models often have backbone RMSD of around 85% of the core regions as low as 3.5 Å, with the errors mainly in loop regions and tails [23]. Furthermore, when the sequence identity drops below 30% (commonly known as 'the twilight zone') [25], modeling accuracy sharply decreases because of the errors in alignment and lack of significant template hits.

In CASP8, there were altogether 13 (strictly 10 FM and 3 FM/TBM) targets and out of those 13, six were predicted well by a number of groups, whereas there was no satisfactory model for four of the targets and the prediction for three targets was fair (based on visual inspection and Global Distance Test (Total Score) (GDT_TS score)) [26]. Despite the increased difficulty (as assessed by target domain GDT_SCORE compared withCASP8), the methods in CASP9 performed better [13]. There were 19 FM targets in CASP10 and the most successful method submitted best models for only four of them [12]. In this regard, there has been a slow but steady progress in modeling the structure of protein sequences for which the sequence identity to be solved drops below 30%, as witnessed in community-wide blind CASP experiments [18, 27].

Two problems are crucial in modeling these proteins: (i) how to identify the correct templates for the sequences of similar structures in the PDB and, once identified, how to refine the template structure closer to that of the native structure and (ii) for the sequences without similar structures in PDB, how to build models of correct topology from scratch.

It should be noted that, starting from CASP7, the target sequences are divided into only two assessment categories: TBM and (*template*)-FM. This classification is based both on the evolutionary relationship between target sequences and the existing templates and their difficulty level indicated by server performance. Usually, the target sequences with detectable structure templates from the PDB, which usually have a higher accuracy by the automated server predictions, are classified as TBM, whereas other targets without templates in the PDB are generally classified as FM. Here, we will discuss recent advances in structure prediction using these two categories. We discuss some important recent advances in FM approaches by starting with a short synopsis on recent advances in TBM approaches.

Recent advances in TBM approaches

Probably, the origin of TBM approach can be traced back to 1969 when Browne and colleagues [28] tried to build structural models of the bovine alpha-lactalbumin using the solved hen egg-white lysozyme structure as a template. Yang *et al.* [29] attribute several factors for the improvement of TBM approaches. First, the development of PSI-BLAST[9] and the consequent profile-to-profile alignment techniques [6, 30, 31] significantly increased the accuracy of template identification and alignment relative to single sequence-based or manual alignment approaches. Second, composite structure assembly simulations combine multiple templates identified by meta-server threading alignments [32, 33], which can drive individual templates considerably closer to the native structures [34–38]. Finally, the rapid accumulation of experimental sequence and structure databases converted many non- or distant-homology targets to close-homology ones by providing close homology templates. We will briefly highlight two TBM approaches: MODELLER [39], ModBase [40] and I-TASSER [22] and some recent trends in TBM.

MODELLER

Although the main focus of this article is on the recent advances in FM methods, it is instructive to briefly review one of the most important early methods. MODELLER [39] is one of the most widely used TBM approach for protein structure prediction tool developed at Sali Lab. The underlying assumption for TMB approach is that 3D structure of proteins from the same family is more conserved than their primary sequences [41]. Thus, any detectable similarity in the sequence level implies structural similarity. In addition, because of various structural genomics project, the probability of finding related proteins of known structure for a new protein sequence is increasing. In this regard, TBM-based approaches will probably remain one of the most useful tools to fill the sequence–structure gap that currently exists.

TBM-based approaches including MODELLER generally consists of four steps: (i) searching for structures related to target sequence, (ii) aligning target sequence and the template(s), (iii) model building and (iv) evaluation of the model. MODELLER implements TBM by satisfaction of spatial restraints collected from various sources [39].

ModBase

The ModBase [40], which belongs to the TBM category, was also developed by Sali's group at UCSF. ModBase consists of a database and other associated resources for comparative protein structure models. The models are calculated automatically using ModPipe [42], a pipeline for comparative protein modeling that relies on number of modules of MODELLER [39] and other various sequence–sequence [43], sequence–profile [9] and profile–profile [30, 44], methods for sequence–structure alignment.

The templates that are used to build models are obtained using various fold assignment and sequence–structure alignment tools like PSI-BLAST, HHBlits [44] and HHsearch [30]. In addition to the model repository, the system also has a comparative modeling web server module called the ModWeb, where users can provide one or more FASTA sequences and obtain models of their sequences generated by ModPipe based on the templates found in the PDB. The system also has other resources for predicting a structural ensemble that fits a Small-angle X-ray scattering (SAXS) profile, for protein–protein docking and various others [40]. As a consequence, ModBase is one of the most comprehensive TBM resources. Indeed, as of November 2015, ModBase has >34 million models for domains in 5.7 million unique protein sequences spanning 65 genomes. Importantly, ModBase models can be accessed through external databases such as the Protein Model Portal [39], which is a repository for accessing protein structure models from a number of different resources.

I-TASSER

I-TASSER, developed in Zhang's group at the University of Michigan, is a composite TBM approach and has been one of the most successful TBM protein structure prediction approaches, as evidenced by results achieved in recent CASP assessments. I-TASSER is an improved version of TASSER [34] developed by Zhang and Skolnick, which refines the TASSER cluster centroid by Iterative Monte Carlo Simulations, thus the name I-TASSER. A protein is represented using reduced representation by its C-alpha atoms and side-chain centers of mass, called the CAS model [34] and the corresponding CAS potential. Another related reduced represented in use in other approaches is the CABS model developed by Kolinski where each residue is represented using four united atoms per residue: a-carbon, c-b and side-chain center of mass [45]. It is to be noted here that Kolinski's group has also developed a FM and consensus-based approach for prediction of protein structure based on CABS called as CABS-fold [46]. In essence, I-TASSER is an iterative fragment assembly approach where structural templates are identified for a target sequence using a meta-threading approach called LOMETS [32, 33]. Continuous fragments are then excised from the templates in the regions that are aligned by threading. These fragments are then used to reassemble full-length models by replica-exchange Monte Carlo Simulations (REMC).

Once reassembled, the structure trajectories are clustered to identify the low-free energy states. Starting from the clusters, a 2nd round of fragment assembly simulation is conducted to further refine the structural models. Finally, the models from the low-energy conformations are further refined by atomic-level simulations to obtain the final model. For interested readers, a retrospective report of the I-TASSER pipeline, which has a detailed description of the performance of I-TASSER in the last five CASP experiments (CASP7-11), can be found in Yang *et al.* [29]. In essence, the report highlights improving trends in the ability of structure refinement over the threading templates in I-TASSER.

One of the most recent developments in I-TASSER for CASP11 is the addition of QUARK [15] (an *ab initio*-based approach) as an intermediate step for TBM structure prediction. QUARK–TBM, which is an extension of QUARK to TBM, is included as an intermediate step in the pipeline where the modeling starts from the top 20 threading templates identified by LOMETS rather than from random conformations. Essentially, the new I-TASSER [29] pipeline starts with multiple structure templates identified by meta-threading programs, followed by the QUARK *ab initio* approach to generate initial full-length models under strong constraints from template alignments. Once the full-length models are generated, the protein structural models are constructed by reassembling continuous fragments excised from the top LOMETS threading alignments using I-TASSER. In addition to the spatial restraints from the threading templates, the restraints are also taken from the full-length models generated from QUARK-TBM. However, it should be noted that, because of its high computational cost, QUARK-TBM is only used for domains that are <300 residues in length. The structure decoys are then clustered by SPICKER, followed by fragment-guided molecular dynamics simulations to refine the SPICKER models at the atomic level. Finally, multiple Quality Assessment Programs are used to select the best model for the submission. As evidenced from the CASP11 assessment, the inclusion of QUARK simulation as an intermediate step improved the quality of the output models for TBM targets compared with the pipeline without using QUARK-TBM for the I-TASSER pipeline.

Use of sequential similarity of physical properties to identify homologs

In TBM, the structure of a target protein is modeled based on the known structure of protein whose sequence is most similar to the target. The standard procedures for calculating the degree of similarity between the target sequence and the known structure are methods based on sequence alignment, such as PSI-BLAST. Recently, new approaches have been developed to quantify the degree of similarity between two sequences, which is one of the key steps in TBM. One of them is the property factor method (PFM) [47], developed by Scherega's group, to identify 'sequence homologs', in which a residue equivalence metric based entirely on amino acid physical properties is used to identify the pairwise physical property similarities of the sequences. This approach is based on the notion that the relationship between protein sequence and structure arises entirely from amino acid physical properties and that the physical properties written into a protein sequence are the key determinants of the protein structure.

Essentially, in this approach, a protein sequence is represented by a factor matrix by appropriately specifying the properties of the 20 canonical amino acids in numerical form, as described in Kidera *et al.* [48]. The target sequence, as well as all the sequences in the PDB, is then converted to a property factor matrix before the degree of similarity of the target sequence and every sequence in the PDB is calculated using a fast-normalized cross-correlation algorithm. Subsequently, five sequences in the PDB with the highest normalized correlation coefficient are selected and the best candidate based on the structural alignment by using TM-align program [49] is selected.

To benchmark the method, the procedure was applied to 89 targets with known PDB structures from CASP10 and 51 targets

with known PDB structures from CASP11. It was observed that PFM [47] is better than PSI-BLAST [9] based on the benchmark. Furthermore, it was also shown that PFM outperforms PSI-BLAST in challenging targets.

It is hoped that through the advent and further development of methodologies such as this, TBM will enjoy increasing success. Essentially, as the PFM approach is able to detect sequence similarities arising from shared physical characteristics that may not be apparent to traditional alignment-based methods, this method may be used to find homology candidates where sequence-based homology methods do not work.

However, it should be noted that, like sequence-based homology methods (e.g. PSI-BLAST), the success of the method will depend on the properties of the PDB. For instance, if some targets have no candidates in the PDB with similar physical property distributions, then no match will be found for the target sequence.

Next, we will discuss some of the important recent developments in FM approaches for protein structure prediction.

Recent advances in FM approaches

FM refers to the approaches that seek to construct structural models for protein sequences that do not have a template detectable from the PDB. FM approaches are also called as *ab initio* or *de novo* structure prediction. There have been many recent advances in FM approaches. In this section, we will first briefly describe the fragment assembly approach, which is one of the important FM-based approaches. We will then turn our attention to some of the recent advances in FM approaches, focusing specifically on the use of evolutionary constraints, contact information, correlated mutation and other information in scoring functions to improve the prediction accuracy and increase the size of protein that these methods can handle.

Fragment assembly approach

The fragment assembly approach has become one of the most popular approaches for FM protein structure prediction. The fragment assembly approach has its foundations in the work of Bowie and Eisenberg [50], who assembled new tertiary structures using small 9-mer fragments from other PDB proteins. Essentially, in the fragment assembly approach, models of protein structures are assembled from fragments of known protein structures. This idea of fragment assembly was later adopted by David Baker's group [14] for the development of ROSETTA. Subsequent work from David Baker's group, Bradley *et al.* [51], which was based on a high-resolution *ab initio*-based structure prediction for small proteins, popularized the method of fragment assembly. Other approaches such as I-TASSER [22] and QUARK [15] also use the fragment assembly approach. ROSETTA is one of the most popular approaches for FM approach. Hence, we discuss ROSETTA in this section.

ROSETTA

ROSETTA is perhaps one of the most actively developed tools for macromolecular modeling. The original development began in the laboratory of David Baker at University of Washington but now it is being codeveloped by >44 laboratories that are members of ROSETTACommons. In addition, a ROSETTA Conference to discuss the updates to the ROSETTA Source Code is held annually [52].

Although ROSETTA method was originally developed for *de novo* structure prediction, ROSETTA also has methods for homology modeling [53]. In this regard, it is a unified software package for protein structure prediction and functional design. The two common tasks for any structure prediction algorithms are as follows: (i) sampling of the conformational space and (ii) ranking of the models using the energy function. ROSETTA uses knowledge-guided Metropolis Monte Carlo Sampling in conjunction with knowledge-based energy functions.

Conformational sampling. ROSETTA has two sampling approaches: one for backbone and other for side-chain degrees of freedom. In addition, backbone conformational sampling is divided into large backbone conformational sampling and local backbone refinement. The large backbone conformational sampling is modeled by exchanging the backbone conformations of nine or three amino acid peptide fragments, whereas the local backbone refinement is performed by using Metropolis Monte Carlo sampling of ϕ and ψ. Please refer to Rohl *et al.* [21] for details.

Energy function. Based on the reduced representation or all-atom model, ROSETTA has two types of energy function: knowledge-based centroid energy function and knowledge-based all-atom energy function. In the centroid representation, the side chain is treated as centroid and the energy function includes solvation, electrostatics, hydrogen bonding and steric clashes [21]. The all-atom energy function includes the 6–12 Lennard-Jones potential for van der Waals forces, a solvation approximation, hydrogen bonding potential, electrostatics term and internal free energy term.

De novo structure prediction. The *de novo* structure prediction in ROSETTA begins with an extended peptide chain. Please refer to [21] for the details of *de novo* structure prediction using ROSETTA. This method works by first generating structurally diverse populations of low-resolution models. Protein models are represented by backbone atoms and the centroids of side chains during this step. These models are then refined in the context of an all-atom energy function by switching the structure representation to 'all-atom'. Using this method, the authors were able to achieve high-resolution structure prediction (<1.5 Å) for small proteins (<85 residues). Most recently, ROSETTA approach was able to successfully predict the structure of a 250-residue long protein that did not have any templates.

QUARK

QUARK [15] is an FM approach for protein structure prediction developed at Zhang's group at University of Michigan. QUARK starts by breaking query sequences into fragments of 1–20 residues where multiple fragment residues are retrieved at each position from unrelated experimental structures. Full-length structure models are then assembled from the fragments using REMC simulations guided by a composite knowledge-based force field. For force field development, QUARK takes a semi-reduced model to represent protein residues by the full backbone atoms and the side-chain center of mass. Initially, QUARK predicts a variety of structural features by using a neural network, and then the global fold is generated by REMC simulation by assembling the small fragments, as in ROSETTA [21] or I-TASSER [22].

The QUARK procedure can be divided into three steps: (i) multiple feature predictions and fragment generation starting from one query sequence, (ii) structure constructions using REMC and (iii) decoy structure clustering and full-atomic refinement. Based on a benchmark study, QUARK was able to correctly fold 31% of the mid-sized proteins (100–150 residues) with a TM-score >0.5. It was also noted that the cases where QUARK was not able to perform well were the proteins where the

structural topology was complex, such as β-proteins of complicated strand arrangement.

Fragment generation and fragment library

As discussed above, fragment-based approaches are one of the most successful approach for FM and they rely on accurate and reliable fragment libraries. Hence, accurate fragment library generation is important to the overall success of fragment-base approaches [54]. The quality of the fragments, hence are tied to the success of the fragment-based approaches. There are various approaches for generating fragments NNMake (ROSETTA's method for fragment library generation) [55], FRAGFOLD [56], HHFrag [57] SAFrag [58] and others.

Recently, some studies have been performed with the aim of designing a better fragment library for improving FM of protein structure. In this regard, De Oliviera *et al.* [44] developed Flib, a novel method to build better fragment library. Scoring fragments based on the predicted secondary structure of the fragment, (essentially α-helical fragments being predicted more accurately), Flib on a validation set of 41 proteins performed better than two state-of-the art methods NNMake and HHFrag.

In addition, not only the length of the fragment but also the number of fragments used per position is equally important for the success of fragment assembly-based approaches. In this regard, Xu and Zhang [59] performed systematic analysis of length of the fragments, number of fragments per position and how these factors affect the precision of the library and showed that this new fragment library developed based on the findings of these analysis performs better than the existing fragment libraries for the *ab initio* structure prediction.

Based on this study, it was also concluded that the optimal fragment length for structural assembly is around 10, and at least 100 fragments per position are required for reliable structure prediction. This also is in alignment with the fact that 9-mer fragments is the most popular choice for the size of fragments [54, 60]. In the case of ROSETTA, its fragment libraries contain 200 fragments per position and these fragments are typically three and nine residues long [55].

Physics-based FM methods

Some of the most successful FM methods based on various CASPs are not strictly *de novo* methods, as these methods use some form of template information. In this regard, purely Physics-based methods can be only strictly thought of as FM approaches. Recently, some impressive results have been achieved using Physics-based FM methods.

In this regard, using a specialized supercomputer (called Anton) that accelerated the execution of Molecular Dynamics (MD) simulations, Shaw and coworker [61] successfully conducted atomic-level molecular dynamics folding simulation of Bovine Pancreatic Trypsin Inhibitor (BPTI). In addition, Shaw's group reported successful atomic-level molecular dynamics simulations to understand the common principles underlying the folding landscape of 12 structurally diverse proteins with an average size of 50 residues [62].

More recently, the same group has successfully studied folding of ubiquitin, a 76-residue long protein, with a folding time of ∼3 ms timescale [63]. One of the major contributions of this work is that until this research, millisecond timescale-based MD simulations were not successful. Most recently, the same group has successfully completed atomic level of ubiquitin on the picosecond to millisecond timescale [64], which is a remarkable feat.

One thing to note here is that because of extremely long timescales required to reach the native structure, application of MD-based simulations for protein structure prediction still remains challenging. However, there have been some recent advances on accelerating the MD simulation. Like the recent trends of using contact-based restraints for other various approaches, Shaw group studied the extent in accelerating the prediction of protein structure using MD using residue–residue contact information [65]. It was observed that for ubiquitin, a speedup of more than an order of magnitude was observed using a relatively small number of restraints (=15) compared with unbiased simulations. We can expect to see other advancements in accelerating the MD-based simulation.

Use of evolutionary constraints from sequence homologs

Studies on Pfam family database have shown that residues in spatial proximity may coevolve across a protein family to maintain energetically favorable interactions [66]. This in turn might suggest that residue correlations (i.e. coevolution) could provide information about amino acid residues that are close in structure. In this regard, evolutionary constraints from sequence homologs have been used to predict 3D structures of proteins. Furthermore, because of the advent of high-throughput genomic sequencing technologies, it is possible to collect multitudes of homologous sequences. For example, Marks *et al.* [67] recently interrogated whether it is possible to infer evolutionary constraints from a set of sequence homologs of a protein. They then explored the idea of using information obtained from statistical analysis of multiple sequence alignments (MSAs) to predict protein structures. The primary challenge using this approach is to distinguish true evolution couplings from the noisy set of observed correlations. Essentially, the residue pair couplings were inferred using a maximum entropy model of the protein sequence with a global statistical model (Bayesian network framework). Finally, residue–residue contacts inferred from the evolutionary record are used to compute the structures from the sequence data alone. The overall method is called EVfold and it is available at http://EVfold.org. For a data set of 15 proteins whose lengths were in the range of 48–258 residues and which had different folds, including a transmembrane protein, EVfold was able to build a final model that had a TM-score of <0.7 for 11 out of 15 studied proteins. Based on the results, it can be concluded that coevolution signals provide some valuable information to determine accurate 3D structures.

Use of correlated mutation information

It has been long known that, given a sufficiently accurate list of contacts, the native fold of a protein can be deduced directly [68, 69]. However, accurate prediction of residue–residue contacts has remained a bottleneck. Residue–residue contacts can be inferred from the observations of correlated mutations in MSAs. These type of methods have an underlying hypothesis that any given contact that is critical for maintaining the fold of the protein will constrain the physicochemical properties of the two amino acids involved: if one or both contacting residues are mutated, then the stability of the native structure will be reduced. As a consequence, pairs of residues that coevolve are likely to be in close proximity to one another (i.e. in contact) in the native structure. This information is useful in mapping residue–residue contacts in the native structure.

Similarly, Jones' group [70] recently developed PSICOV, an algorithm that uses sparse inverse covariance estimation techniques to predict contacts accurately from sequence alignments. When sufficient homologous sequences are available, PSICOV [70] can predict long-range contacts (i.e. contacts separated by >23 residues) with an accuracy close to 80%. For instance, Jones' group [71] used the predicted contacts from PSICOV to identify the native fold for medium-sized (<200 residues) protein domains.

More recently, Nungent and Jones [72] developed FILM3, which uses a scoring function based on correlated mutations detected from multiple sequence alignment (i.e. only information derivable from the target sequence and its homologs) using PSICOV to produce a 3D model for larger α-helical transmembrane proteins. In this regard, it can be argued that the authors were able to replace the use of knowledge-based potentials or other statistically derived scoring function by correlated mutation-based scoring function. The method, FILM3, is able to predict the structure of proteins of up to 531 residues within a reasonable TM-score (<0.745 using a scoring function based on the estimated probabilities of residue–residue contacts predicted using PSICOV).

Similarly, David Baker's group [73] also developed GREMLIN—a method for predicting residue–residue contact based on the coevolutionary information. They then used this information to improve the prediction of structures of proteins. By going beyond the second-order approximation in the residue–residue covariation matrix inversion used by PSICOV, the authors improved the residue–residue contact prediction and then assessed the usefulness of contact prediction for protein structure prediction. In so doing, they interrogated how useful covariance-based contact predictions are for structure prediction when the homologous structure of the target protein is likely to be available. Based on analysis on a benchmark data set, it was suggested that the contact predictions are likely to be accurate when the number of aligned sequences is >5 times the length of the protein and that the predicted contacts are likely to be useful for structure prediction when the aligned sequences are more similar to the target protein than to the closest homologous structure of the target protein.

Recently, Hopf *et al.* [74] also used amino acid coevolution to define restraints on structural proximity of residue pairs and used this information to generate the *ab initio* structure of an insect odorant receptors, which does not have solved crystal structures or sequence similarity to other proteins. In this regard, coevolutionary information has been gaining a lot of attention for predicting structures of proteins.

Use of residue–residue contact information

Various experiments [75, 76] have been performed to determine if accurate protein structures can be reconstructed using true contacts. These studies suggest that contacts contain crucial information for structure prediction. However, recent work in protein structure modeling using NMR chemical shifts [77], sparse restraints [78] and Cryo-EM data [79] have shown that additional information can also significantly improve protein structure modeling. In this regard, protein contact map prediction information can be used to improve the modeling of protein structure. Moreover, recent studies suggest that predicted contacts could be used to reconstruct protein structures. Therefore, it will be important to exploit the new information obtained from protein residue–residue contact prediction methods.

One of the methods that uses residue–residue contacts to predict *ab initio* protein structure is the Residue-Residue Contact-guided ab initio Protein Folding (CONFOLD) method developed by Adhikari *et al.* [80]. This method uses predicted contacts and secondary structures to improve structure prediction for FM targets. Essentially, this approach consists of two stages. In the 1ststage, the initial contact-based distance restraints and secondary structure-based restraints are used to reconstruct protein models. In the second stage, the contact information, as well as the β-sheet information, is updated by analyzing the model having minimum energy in the 1ststage. In total, 400 models are reconstructed for a given protein in each stage and the 400 models in the second stage are considered as final predictions. Based on a comparative analysis on the EVFOLD benchmark [67] data set that comprises 15 proteins, the prediction accuracy of CONFOLD was similar to EVFOLD. Recently, I-TASSER used sequence-based contact predictions from SVMSEQ [81] to improve protein structure prediction since CASP8.

Number of contacts needed for reconstruction

Although there are various studies that use residue–residue contacts to improve protein structure prediction, until recently little work had been done to determine how much distance information is required to improve protein structure modeling. In this regard, Kim *et al.* [82] carried out a study to analyze how much contact information is needed to improve the modeling of protein structures. Based on the analysis and comparison of contact-assisted and non-assisted prediction using ROSETTA, Kim*et al.* [82] observed a consistent improvement over most of the best non-assisted predictions. Using an average amount of one correct contact per 12 residues, the group was able to model the correct topology for 15 out of 17 target domains in the CASP10 data set with a TM-score >0.5. Based on these observations, it was concluded that experimental, as well as bioinformatics, methods for obtaining contact information may only need to generate a limited number of accurate contacts (e.g. one correct contact for every 12 residues in the protein) to promote accurate topology-level modeling.

Adhikari *et al.* [80] also performed an analysis to estimate the number of contacts needed for reconstruction by scanning the structures in the PDB and found that 99% of known 3D structures have <3L (L: length of the protein) true contacts, and >50% of them have <2L true contacts. In their analysis, they found that 60% of the best models are reconstructed with the top 0.6L and that different proteins need different numbers of contacts to be folded well. Their analysis to determine the number of predicted contacts needed to obtain the best fold showed that different proteins need different numbers of contacts to be folded well, and they suggested that instead of fixing the number of contacts, predicting a range for the number of contacts would be useful for contact-based protein reconstruction.

Encouraging developments in contact prediction in CASP11 assessment

Contact prediction also has been a focus area in CASP. Especially, as highlighted by the fact that the successful prediction of a 250 residue target in CASP11 was attributed to contact prediction, contact prediction is gathering a lot of interest. However, until CASP10, the contact prediction accuracy of the participating methods was marginal (20%), at best [83]. CASP11 also included a category for prediction of contacts in FM targets, assigning a probability score between 0 and 1 to each contact. This served as a measure of the confidence of the assignment.

A pair of residues is defined to be in contact when the distance between their Cβ atoms is smaller than 8.0 Å.

Concentrating the assessment on long-range contacts (24 position separation between interacting residues) and ensuring fairness of the comparison by reducing the contacts per target (L/5), prediction of 29 groups were evaluated [83].

In this regard, Monastyrskyy *et al*. [83] did an assessment study on the improvement in contact prediction using CASP11 results. Based on the evaluation carried out on FM targets for which structural template could not be identified, and focusing on the long-range intradomain contacts (separation of the interacting residues of at least 24 positions along the sequence within the same domain), one of the major highlights of the study is the performance of CONSIP2 [84] with precision of 27% on target proteins that did not have templates. Based on these results and other assessments of CASP11 results, it can be concluded that encouraging developments have been observed in the area of contact prediction. It has to be noted that the precision of best methods for CASP9 was 21% and CASP10 was 20%.

This study also assessed the interdomain contact predictions as interdomain contacts help proper packing of the domains in multi-domain proteins. Based on the study, it was observed that the accuracy of predicting interdomain contacts is much lower than that of intradomain contacts. The highest precision achieved for CASP11 is below 6%, highlighting the dismal and little to no improvement compared with previous CASPs. In this regard, we can expect more methods focusing on improving the prediction accuracy of intradomain contacts.

Advances in FM approaches based on CASP10 assessment

Until recently, successes in FM approaches in predicting 3D structures of protein sequences without using template structures from experimentally solved proteins were limited to smaller proteins with lengths below 100 residues [27]. This is mainly attributed to two reasons: (1) lack of accurate force fields to describe the atomic interactions that can guide the protein-folding simulations and (2) insufficient sampling of the search space. However, there has been steady progress in the FM approaches, as evidenced by recent CASP assessments. We will first summarize assessment of FM approaches in CASP10 and then we will describe recent advances in FM approaches.

The Biennial CASP [85] assessment provides an objective and independent assessment of protein structure prediction methods. As discussed earlier, although much progress has been made during the last 20 years of CASP, the template-FM remains a challenge [12, 13, 26, 27]. Out of a total of 96 target proteins, there were 20 in the FM category of CASP10 [12]. Historically, the number of FM targets has been low: 13 in CASP8 and 30 in CASP9 [13]. Out of these FM targets, only two in CASP9, four in CASP9 and three in CASP10 belonged to potentially new folds. Owing to the fact that it is difficult to support a statistically meaningful evaluation of the FM techniques, the number of targets was increased by introducing the ROLL experiments, in which FM targets were rolled year-round.

Based on comprehensive analysis of the predictions for 11 FM and 19 ROLL server targets—where the lengths ranged from 58 to 533 residues—the Keasar group submitted the best models for four targets. QUARK was designated as the best server, with three best predictions, followed by the Zhang-Server and Baker-Rosetta servers [13]. Although the category is an FM category, CASP9 FM assessors [13] noted that 'meta-predictors' were also observed in CASP10 [12]. Ideally, it is difficult to find templates for these targets, but many predictors found templates and improved on those templates to produce the best models for the target. In this regard, even the most successful group submitted best models for only four of the 19 FM targets and eight of the 36 ROLL targets. This highlights the fact that prediction of structures without a template remains a challenge. As there were six or more groups that had at least one best model, more progress is expected in the future.

For CASP 10, Zhang's group participated using the QUARK pipeline as well as integrating QUARK (FM) with I-TASSER (TBM). The integration was based on the idea that the *ab initio* models built from scratch are usually different from experimental structures in the PDB, and that a close match between the templates and *ab initio* models is usually an indication of the correct fold adopted by the *ab initio* models. Essentially, I-TASSER (TBM) and QUARK (*ab initio* modeling) models were combined to find distant-homology proteins (in particular those that are longer than 100 residues). Initially, the LOMETS (threading) templates with the highest TM-score to the QUARK models (*ab initio*) were used as the initial models for I-TASSER simulations. This is based on the fact that any reasonable match between the *ab initio* folding simulations to the real protein may indicate a correct template hit. This method was shown to improve the quality and robustness of the final models for FM protein targets, as evidenced in CASP10 assessment [86].

Advances in FM approaches based on CASP11 assessment

The progress of FM approaches, although successful for smaller proteins, appeared to be stalled for larger proteins in the past decade until CASP11 [87] when an accurate 3D model of a large (256) residue protein was generated by David Baker's group.

The CASP11 FM target included largest number of targets for FM evaluation: 45 targets compared with30 targets in CASP9 and 20 in CASP10. Strictly speaking, FM approaches should not use template information. However, in the previous CASP9 and CASP10, the top-performing groups incorporated some use of templates or selection and refinement of server models. This trend continued in CASP11 where some of the best-performing methods like Baker-Rosetta server used parent templates for 47% and Zhang-servers used parent templates for 17% of FM-only targets [16]. It has to be noted here that QUARK declared 0% parent templates and it outperformed other template-based methods as servers. One of the important highlights of CASP11 is the target T0806-D1 (a 256-residue protein with no sequence similarity to existing template), for which Baker's group provided one of the outstanding models. It has to be noted that this case represents one of the largest correctly predicted FM models in the history of CASP. On investigation, it was observed that T0806 had a sufficiently large enough family for predicting coevolving residues from sequence alignments [16]. Overall, the success of the methods in this category in CASP11 is attributed to the incorporation of contact information.

Similarly, for the server model, Zhang's QUARK produced one of the outstanding models for target T0837-D1. This model maintained correct topology of the target fold over all seven α-helices, thus encompassing the entire fold. Overall, Baker's group and Zhang's group performed better in the overall prediction for CASP11 FM targets. Also for ROLL, the outstanding prediction was for the up and down α-helical bundle target R0034-D1 submitted by Zhang's group. One of the major advances in the methods in CASP11 compared with previous CASP is the improvement in prediction accuracy for larger protein

domains. Also, as highlighted by the not so successful prediction of multidomain targets like T0808-D2, it can be concluded that multi-domain targets are still challenging to FM prediction methods.

Some other notable methods in CASP11 is the method [88] from Kihara Lab at Purdue. The method used a new knowledge-based scoring function called Protein Residue Environment Score and helix interaction potential for selecting near-native structures from server models (made available for human predictors by CASP organizers) and then performed short structural refinement using MD simulation. This group was ranked 1st among all the participants in the FM category when the top one models were considered [88].

Protein–peptide complex prediction

Recently, some notable advancement has been achieved in the structure modeling of protein–peptide complexes. Kurcinski *et al.*[89] developed CABS-dock web server for the flexible docking of peptides to proteins. CABS-dock attempts to unify all three steps of computation protein–peptide docking: (i) prediction of the binding site on the receptor structure, (ii) initial modeling of the peptide backbone in the binding site(s) and (iii) refinement of the protein–peptide complexes to high resolution. This server when benchmarked against 103 bound and 68 unbound cases obtained that for over 80% of the cases, models with high or medium accuracy sufficient for practical applications where high quality is defined as ligand (peptide) RMSD <3 Å and medium accuracy is defined as 3 Å $<$ RMSD $<$ 5.5 Å.

Similarly, London *et al.* [90] developed Rosetta framework-based Rosetta FlexPepDock web server for the refinement of protein–peptide complex. Given a protein receptor structure and (possibly) inaccurate model of the peptide, FlexPepDock allows for full flexibility to the peptide and side-chain flexibility to the receptor using Monte Carlo-based minimization approach. On a benchmark data set that covers wide range of starting peptide conformations, Rosetta FlexPepDock server is able to create near-native models (peptide backbone RMSD $<$ 2 Å) in 91% cases for the bound receptor and rank them as one of the top five models in 78% of the cases when the initial backbone root mean square deviation is up to 5.5 Å.

Most recent trends, outlook and conclusion for FM approach

Recently, several trends have emerged in the FM-based protein structure prediction. For instance, to increase the accuracy of predictions, there has been a trend toward integrating TBM and FM techniques. Likewise, collaborative efforts for protein structure prediction have increased in recent years. In parallel with these efforts, there has also been an increase in the number of studies focused on improvement of structure modeling by large-scale model quality analysis. Below, we discuss some of these recent trends and the current outlook for FM-based approaches. We also summarize the discussed methods in the article in Table 1.

Integration of template-based and template-free protein structure modeling

Few methods have been proposed that integrates complementary modeling methods (e.g. template-based and template-free protein structure modeling methods). One of the examples of this type of tool is MULTICOM-NOVEL [91]. This server integrates the prediction capabilities of both TBM and template-FM to synergistically combine the two kinds of methods to improve protein structure prediction. In this work, the authors developed a new method by integrating several protein structure prediction methods, including their own template-based MULTICOM server [92], the *ab initio* contact-based protein structure prediction method CONFOLD [80], their multi-template-based model generation tool MTMG and locally installed external ROSETTA [21], I-TASSER [22] and RaptorX [93] protein structure prediction tools. This approach was ranked among the top 10 methods out of 44 servers in the recently concluded CASP11, which demonstrates the usefulness of integrating TBM-based approaches and FM-based approaches for advancing protein structure prediction.

There have been some other approaches for FM targets that integrate *ab initio* (FM) and TBM approaches. One of them is the method from Zhang's group [94]. In addition to the QUARK pipeline, for CASP11, Zhang's group [94] championed the idea of integrating TBM approaches with FM-based approaches a step further by integrating QUARK and I-TASSER to predict the FM targets. Essentially, threading templates identified by the meta-threading program LOMETS were sorted by their similarity (based on the TM-score) to QUARK (*ab initio*) models before being submitted to the I-TASSER pipeline. The spatial constraints were then collected from the new (sorted) LOMETS templates and the QUARK models, which were subsequently used by I-TASSER assembly simulations. Finally, the structural decoys were clustered using SPICKER and the models with the highest cluster density were further refined by fragment-guided molecular dynamic simulations. Compared with the QUARK pipeline alone, the integrated approach (Zhang-server) was able to model 60% more domains with length up to 204 residues. Despite these promising results, significant challenges still exist for FM of protein structures.

Large-scale model quality assessment

The two major challenges of protein structure prediction are conformational sampling and model quality assessment (MQA; or ranking). The aim of conformational sampling is to generate a number of conformations for a target protein, and the goal of MQA) is to assess the quality of these models and select the best ones as final predictions. In this regard, development of accurate MQA methods is indispensable for improving protein structure prediction. There are basically two main kinds of quality assessment methods: single-MQA methods [95] that evaluate the quality of one single model without using the information of other models; and multi-MQA methods [96] that uses the information of other models of the same protein to assess the quality of a model. Generally, protein structure prediction protocols use one or a few MQA methods. Some methods, like I-TASSER, also use clustering techniques to rank these models. The evaluation of model quality estimates in CASP10 [97] also highlights the fact that none of the methods can consistently select the best models. In this regard, protein MQA still needs some improvement.

One recent trend is to develop novel MQA methods, particularly large-scale MQA methods. To this end, Cao *et al.* [98] used a large-scale model combination approach to combine 14 MQA methods to improve the quality of protein model ranking. The method ranked 3rd out of all 143 human and server predictions for protein structure prediction in CASP 11 based on the sum of

Table 1. Protein structure prediction methods described in the article with short description and reference

Method	Type	Short description	Reference
Rosetta	FM/TBM	Fragment assembly	https://boinc.bakerlab.org/
I-TASSER	TBM	Fragment assembly	http://zhanglab.ccmb.med.umich.edu/I-TASSER/
ModBase	TBM	Comprehensive resource for TBM	http://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi
PMP	TBM	Resource for accessing protein structure models from various methods	http://www.proteinmodelportal.org/
TASSER	FM/TBM	MetaServer that combines various TASSER-based approaches	http://cssb.biology.gatech.edu/skolnick/webservice/TASSER/index.html
QUARK	FM	Fragment assembly developed at Zhang Laboratory	http://zhanglab.ccmb.med.umich.edu/QUARK/
Desmond	MD	Software suite for MD of biological systems	https://www.deshawresearch.com/resources_desmond.html
EVfold	Evolutionary couplings	Residue contact-based structure prediction	http://evfold.org/evfold-web/evfold.do
CONFOLD	FM	Residue contact-guided FM	http://protein.rnet.missouri.edu/confold/
CABS-dock	Protein–peptide complex	Protein–peptide complex server	http://biocomp.chem.uw.edu.pl/CABSdock
Rosetta FlexPepDock	Protein–peptide complex	Protein–peptide complex server	http://flexpepdock.furmanlab.cs.huji.ac.il/
AIDA	FM	*Ab initio* multi-domain server	http://ffas.sanfordburnham.org/AIDA/
MULTICOM	TBM		http://sysbio.rnet.missouri.edu/multicom_toolbox/
MODELLER	TBM	Available for download	http://salilab.org/modeller/
CABS-fold	FM	*De novo*, templates, distance restraints	http://biocomp.chem.uw.edu.pl/CABSfold/

PMP = Protein Model Portal.

z-score of the 1st models predicted for 78 CASP11 protein domains.

Collaborative effort: WeFold

Interdisciplinary collaboration is on the rise throughout the scientific community. CASP, which was started by Moult *et al.*[85] in 1994, also saw this new trend of 'coopetition' called WeFold [99] during the CASP10 experiments, where 13labs worldwide, using methods ranging from purely bioinformatics to Physics-based approaches, participated in a social media-based worldwide collaborative effort and competed in search of methodologies that are better than their individual parts. The 13 labs were arranged into five branches, each representing five independent protein structure prediction methods that combine different components from their contributing group.

Three of the branches produced one remarkable result each, and two of these results were featured by the assessors in the refinement and FM categories. However, none of the branches produced consistently good results.

The wfCPUNK branch that worked on the FM submitted model for four targets. The GDT_TS score for wfCPUNK was better than individual methods but not statistically significant (e.g. for target T0740 the GDT_TS of wfCPUNK was 32.1, whereas the highest GDT_TS of the individual methods was 30.81). Similarly, the WeFold branch, which attempted on the 43 human targets, performed comparably (11 targets) or better than one of the individual methods (TASSER (12 targets) [34]) in 53% of the cases. However, in 17 targets, TASSER significantly outperformed the WeFold branch that indicates that WeFold still needs a lot of improvement.

Nevertheless, this approach of coopetition (cooperative competition) to the difficult problem of protein structure prediction shows some promise and is an important step in the right direction. This collaboration has continued into CASP11, in which there were 18 different branches that each submitted their own prediction. As can be seen by the increased number of branches from CASP10 to CASP11, we hope to see more branches coming out of this collaborative effort.

Multi-domain protein structure prediction

The existing computational approaches for multi-domain protein structure-based prediction methods can be roughly divided into two classes: (i) *ab initio* methods and (ii) template-based methods. Existing *ab initio* methods can be further subdivided into two general approaches: (i) a docking approach, in which multi-domain structure prediction is treated as a docking problem [100, 101] and (ii) a domain assembly approach, in which the linker region is sampled iteratively [102]. Methods based on the domain assembly approach [102] are successful in only ~50% of the studied cases. Likewise, methods that use the docking approach yield an assembly within the top 10 solutions in only ~60% of the cases. This is likely because of the fact that the rigid docking models used by these approaches often cannot account for the flexibility in the linker region [100, 101]. Recently, Xu *et al.* [103] developed an energy minimization method, AIDA, which uses *ab initio* folding potential for domain assembly. However, this method also could only correctly predict 2 out of 15 multi-domain proteins from the CASP10 target. In this regard, multi-domain protein structure prediction is still challenging.

In addition, Monastyrskyy *et al.* also assessed the interdomain contact predictions because the interdomain contact helps proper packing of the domains in multi-domain proteins. Based on the study, it was observed that the accuracy of predicting interdomain contacts is much lower than that of intradomain contacts. The highest precision achieved for CASP11 is below 6% highlighting the disappointment, whereas the intradomain contact precision was 27% that also highlights the difficulty in multi-domain protein structure prediction. In this regard, despite the great strides that have been made in

single-domain protein structure prediction, multi-domain protein structure prediction remains a major challenge in the field.

Key Points

- Protein structure prediction can be classified according to TBM approaches and FM approaches. Classification of computational approaches for protein structure prediction is becoming increasingly blurry.
- Some of the new trends are to integrate TBM and FM approaches.
- Slow but steady progress has been made in modeling the structure of protein sequences for which the sequence identity to a solved drops below 30%, as witnessed in the community-wide blind CASP experiments.
- Use of sequence constraints, contact prediction and correlated mutation information shows some promise in improvement of FM approaches.
- Remarkable progress has been achieved in FM in CASP11 where a 250-residue long protein was successfully modeled.
- MD-based methods using specialized computers have been successful in folding.
- True FM approaches based on molecular dynamics simulation using specialized computers have succeeded in folding of ubiquitin, a 76-residue long protein, with a folding time of ~3 ms timescale.
- Likely that new method development in the prediction of the structure of multi-domain proteins and also more multi-domain targets will be seen in the CASP assessments.

Funding

KC DB is also partly supported by the National Science Foundation under cooperative agreement no. DBI-0939454 and this material is also partly based upon work supported by startup fund from College of Engineering of North Carolina A&T State University.

References

1. Benson DA, Karsch-Mizrachi I, Lipman DJ, *et al*. GenBank. *Nucleic Acids Res* 2003;**31**:23–7.
2. UniProt C. UniProt: a hub for protein information. *Nucleic Acids Res* 2015;**43**:D204–12.
3. Berman HM, Westbrook J, Feng Z, *et al*. The Protein Data Bank. *Nucleic Acids Res* 2000;**28**:235–42.
4. Zhang Y. Progress and challenges in protein structure prediction. *Curr Opin Struct Biol* 2008;**18**:342–8.
5. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 1993;**234**:779–815.
6. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 1991;**253**:164–70.
7. Liwo A, Lee J, Ripoll DR, *et al*. Protein structure prediction by global optimization of a potential energy function. *Proc Natl Acad Sci USA* 1999;**96**:5482–5.
8. Altschul SF, Gish W, Miller W, *et al*. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10.
9. Altschul SF, Madden TL, Schaffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
10. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci USA* 1987;**84**:4355–8.
11. Kryshtafovych A, Fidelis K, Moult J. CASP9 results compared to those of previous CASP experiments. *Proteins* 2011;**79(Suppl 10)**:196–207.
12. Tai CH, Bai H, Taylor TJ, *et al*. Assessment of template-free modeling in CASP10 and ROLL. *Proteins* 2014; **82(Suppl 2)**:57–83.
13. Kinch L, Yong Shi S, Cong Q, *et al*. CASP9 assessment of free modeling target predictions. *Proteins* 2011;**79(Suppl 10)**:59–73.
14. Simons KT, Kooperberg C, Huang E, *et al*. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J Mol Biol* 1997;**268**:209–25.
15. Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* 2012;**80**:1715–35.
16. Kinch LN, Li W, Monastyrskyy B, *et al*. Evaluation of free modeling targets in CASP11 and ROLL. *Proteins* 2015; doi:10.1002/prot.24973.
17. Zhou H, Zhou Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins* 2005;**58**:321–8.
18. Ginalski K, Rychlewski L. Protein structure prediction of CASP5 comparative modeling and fold recognition targets using consensus alignment approach and 3D assessment. *Proteins* 2003;**53(Suppl 6)**:410–7.
19. Bradley P, Chivian D, Meiler J, *et al*. Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. *Proteins* 2003;**53(Suppl 6)**:457–68.
20. Wu S, Skolnick J, Zhang Y. Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biol* 2007;**5**:17.
21. Rohl CA, Strauss CE, Misura KM, *et al*. Protein structure prediction using Rosetta. *Methods Enzymol* 2004;**383**:66–93.
22. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* 2010;**5**:725–38.
23. Baker D, Sali A. Protein structure prediction and structural genomics. *Science* 2001;**294**:93–6.
24. Tramontano A, Morea V. Assessment of homology-based predictions in CASP5. *Proteins* 2003;**53(Suppl 6)**:352–68.
25. Rost B. Twilight zone of protein sequence alignments. *Protein Eng* 1999;**12**:85–94.
26. Ben-David M, Noivirt-Brik O, Paz A, *et al*. Assessment of CASP8 structure predictions for template free targets. *Proteins* 2009;**77(Suppl 9)**:50–65.
27. Jauch R, Yeo HC, Kolatkar PR, *et al*. Assessment of CASP7 structure predictions for template free targets. *Proteins* 2007;**69(Suppl 8)**:57–67.
28. Browne WJ, North AC, Phillips DC, *et al*. A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme. *J Mol Biol* 1969; **42**:65–86.
29. Yang J, Zhang W, He B, *et al*. Template-based protein structure prediction in CASP11 and retrospect of I-TASSER in the last decade. *Proteins* 2015; doi:10.1002/prot.24918.
30. Soding J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* 2005;**21**:951–60.

31. Wu S, Zhang Y. MUSTER: improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins* 2008;**72**:547–56.
32. Ginalski K, Elofsson A, Fischer D, *et al*. 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* 2003;**19**:1015–18.
33. Wu S, Zhang Y. LOMETS: a local meta-threading-server for protein structure prediction. *Nucl. Acids. Res* 2007;**35**: 3375–82.
34. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc Natl Acad Sci USA* 2004;**101**:7594–9.
35. Yang J, Yan R, Roy A, *et al*. The I-TASSER Suite: protein structure and function prediction. *Nat Methods* 2015;**12**:7–8.
36. Joo K, Lee J, Sim S, *et al*. Protein structure modeling for CASP10 by multiple layers of global optimization. *Proteins* 2014;**82(Suppl 2)**:188–95.
37. Cheng J, Wang Z, Tegge AN, *et al*. Prediction of global and local quality of CASP8 models by MULTICOM series. *Proteins* 2009;**77(Suppl 9)**:181–4.
38. Misura KM, Chivian D, Rohl CA, *et al*. Physically realistic homology models built with ROSETTA can be more accurate than their templates. *Proc Natl Acad Sci USA* 2006;**103**:5361–6.
39. Webb B, Sali A. Comparative protein structure modeling using MODELLER. *Curr Protoc Bioinformatics* 2014;**47**:5 6 1–6 32.
40. Pieper U, Webb BM, Dong GQ, *et al*. ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res* 2014;**42**:D336–46.
41. Lesk AM, Chothia C. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J Mol Biol* 1980;**136**:225–70.
42. Eswar N, John B, Mirkovic N, *et al*. Tools for comparative protein structure modeling and analysis. *Nucleic Acids Res* 2003;**31**:3375–80.
43. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol* 1981;**147**:195–7.
44. Remmert M, Biegert A, Hauser A, *et al*. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* 2012;**9**:173–5.
45. Kolinski A. Protein modeling and structure prediction with a reduced representation. *Acta Biochim Pol* 2004;**51**:349–71.
46. Blaszczyk M, Jamroz M, Kmiecik S, *et al*. CABS-fold: server for the de novo and consensus-based prediction of protein structure. *Nucleic Acids Res* 2013;**41**:W406–11.
47. He Y, Rackovsky S, Yin Y, *et al*. Alternative approach to protein structure prediction based on sequential similarity of physical properties. *Proc Natl Acad Sci USA* 2015;**112**: 5029–32.
48. Kidera A, Konishi Y, Oka M, *et al*. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *J Protein Chem* 1985;**4**:23–55.
49. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* 2005;**33**: 2302–9.
50. Bowie JU, Eisenberg D. An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function. *Proc Natl Acad Sci USA* 1994;**91**:4436–40.
51. Bradley P, Misura KM, Baker D. Toward high-resolution de novo structure prediction for small proteins. *Science* 2005;**309**:1868–71.
52. Khare SD, Whitehead TA. Introduction to the Rosetta special collection. *PLoS One* 2015;**10**:e0144326.
53. Kaufmann KW, Lemmon GH, Deluca SL, *et al*. Practically useful: what the Rosetta protein modeling suite can do for you. *Biochemistry* 2010;**49**:2987–98.
54. Holmes JB, Tsai J. Some fundamental aspects of building protein structures from fragment libraries. *Protein Sci* 2004;**13**:1636–50.
55. Gront D, Kulp DW, Vernon RM, *et al*. Generalized fragment picking in Rosetta: design, protocols and applications. *PLoS One* 2011;**6**:e23294.
56. Jones DT, McGuffin LJ. Assembling novel protein folds from super-secondary structural fragments. *Proteins* 2003;**53 (Suppl 6)**:480–5.
57. Kalev I, Habeck M. HHfrag: HMM-based fragment detection using HHpred. *Bioinformatics* 2011;**27**:3110–16.
58. Shen Y, Picord G, Guyon F, *et al*. Detecting protein candidate fragments using a structural alphabet profile comparison approach. *PLoS One* 2013;**8**:e80493.
59. Xu D, Zhang Y. Toward optimal fragment generations for ab initio protein structure assembly. *Proteins* 2013;**81**:229–39.
60. Bystroff C, Simons KT, Han KF, *et al*. Local sequence-structure correlations in proteins. *Curr Opin Biotechnol* 1996;**7**:417–21.
61. Shaw DE, Maragakis P, Lindorff-Larsen K, *et al*. Atomic-level characterization of the structural dynamics of proteins. *Science* 2010;**330**:341–6.
62. Lindorff-Larsen K, Piana S, Dror RO, *et al*. How fast-folding proteins fold. *Science* 2011;**334**:517–20.
63. Piana S, Lindorff-Larsen K, Shaw DE. Atomic-level description of ubiquitin folding. *Proc Natl Acad Sci USA* 2013;**110**:5915–20.
64. Lindorff-Larsen K, Maragakis P, Piana S, *et al*. Picosecond to millisecond structural dynamics in human ubiquitin. *J Phys Chem B* 2016; doi: 10.1021/acs.jpcb.6b02024.
65. Raval A, Piana S, Eastwood MP, *et al*. Assessment of the utility of contact-based restraints in accelerating the prediction of protein structure using molecular dynamics simulations. *Protein Sci* 2016;**25**:19–29.
66. Miller CS, Eisenberg D. Using inferred residue contacts to distinguish between correct and incorrect protein models. *Bioinformatics* 2008;**24**:1575–82.
67. Marks DS, Colwell LJ, Sheridan R, *et al*. Protein 3D structure computed from evolutionary sequence variation. *PLoS One* 2011;**6**:e28766.
68. Gobel U, Sander C, Schneider R, *et al*. Correlated mutations and residue contacts in proteins. *Proteins* 1994;**18**:309–17.
69. Olmea O, Valencia A. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold Des* 1997;**2**:S25–32.
70. Jones DT, Buchan DW, Cozzetto D, *et al*. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 2012;**28**:184–90.
71. Taylor WR, Jones DT, Sadowski MI. Protein topology from predicted residue contacts. *Protein Sci* 2012;**21**:299–305.
72. Nugent T, Jones DT. Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proc Natl Acad Sci USA* 2012;**109**:E1540–7.
73. Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci USA* 2013;**110**:15674–9.
74. Hopf TA, Morinaga S, Ihara S, *et al*. Amino acid coevolution reveals three-dimensional structure and functional domains of insect odorant receptors. *Nat Commun* 2015;**6**:6077.

75. Vassura M, Margara L, Di Lena P, *et al*. FT-COMAR: fault tolerant three-dimensional structure reconstruction from protein contact maps. *Bioinformatics* 2008;**24**:1313–5.
76. Duarte JM, Sathyapriya R, Stehr H, *et al*. Optimal contact definition for reconstruction of contact maps. *BMC Bioinformatics* 2010;**11**:283.
77. Cavalli A, Salvatella X, Dobson CM, *et al*. Protein structure determination from NMR chemical shifts. *Proc Natl Acad Sci USA* 2007;**104**:9615–20.
78. Schmitz C, Vernon R, Otting G, *et al*. Protein structure determination from pseudocontact shifts using ROSETTA. *J Mol Biol* 2012;**416**:668–77.
79. Chan KY, Trabuco LG, Schreiner E, *et al*. Cryo-electron microscopy modeling by the molecular dynamics flexible fitting method. *Biopolymers* 2012;**97**:678–86.
80. Adhikari B, Bhattacharya D, Cao R, *et al*. CONFOLD: residue-residue contact-guided ab initio protein folding. *Proteins* 2015;**83**:1436–49.
81. Wu S, Zhang Y. A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics* 2008;**24**:924–31.
82. Kim DE, Dimaio F, Yu-Ruei Wang R, *et al*. One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. *Proteins* 2014;**82(Suppl 2)**:208–18.
83. Monastyrskyy B, D'Andrea D, Fidelis K, *et al*. New encouraging developments in contact prediction: assessment of the CASP11 results. *Proteins* 2015; doi:10.1002/prot.24943.
84. Kosciolek T, Jones DT. Accurate contact predictions using covariation techniques and machine learning. *Proteins* 2015; doi:10.1002/prot.24863.
85. Moult J, Pedersen JT, Judson R, *et al*. A large-scale experiment to assess protein structure prediction methods. *Proteins* 1995;**23**:ii–v.
86. Zhang Y. Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. *Proteins* 2014;**82(Suppl 2)**:175–87.
87. Moult J, Fidelis K, Kryshtafovych A, *et al*. Critical assessment of methods of protein structure prediction (CASP) - progress and new directions in Round XI. *Proteins* 2016.
88. Kim H, Kihara D. Protein structure prediction using residue- and fragment-environment potentials in CASP11. *Proteins* 2015.
89. Kurcinski M, Jamroz M, Blaszczyk M, *et al*. CABS-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site. *Nucleic Acids Res* 2015;**43**:W419–24.
90. London N, Raveh B, Cohen E, *et al*. Rosetta FlexPepDock web server–high resolution modeling of peptide-protein interactions. *Nucleic Acids Res* 2011;**39**:W249–53.
91. Li J, Adhikari B, Cheng J. An Improved integration of template-based and template-free protein structure modeling methods and its assessment in CASP11. *Protein Pept Lett* 2015;**22**:586–93.
92. Li J, Deng X, Eickholt J, *et al*. Designing and benchmarking the MULTICOM protein structure prediction system. *BMC Struct Biol* 2013;**13**:2.
93. Kallberg M, Wang H, Wang S, *et al*. Template-based protein structure modeling using the RaptorX web server. *Nat Protoc* 2012;**7**:1511–22.
94. Zhang W, Yang J, He B, *et al*. Integration of QUARK and I-TASSER for ab initio protein structure prediction in CASP11. *Proteins* 2015; doi:10.1002/prot.24930.
95. Wang Z, Tegge AN, Cheng J. Evaluating the absolute quality of a single protein model using structural features and support vector machines. *Proteins* 2009;**75**:638–47.
96. McGuffin LJ, Roche DB. Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. *Bioinformatics* 2010;**26**:182–8.
97. Kryshtafovych A, Barbato A, Fidelis K, *et al*. Assessment of the assessment: evaluation of the model quality estimates in CASP10. *Proteins* 2014;**82(Suppl 2)**:112–26.
98. Cao R, Bhattacharya D, Adhikari B, *et al*. Large-scale model quality assessment for improving protein tertiary structure prediction. *Bioinformatics* 2015;**31**:i116–23.
99. Khoury GA, Liwo A, Khatib F, *et al*. WeFold: a coopetition for protein structure prediction. *Proteins* 2014;**82**:1850–68.
100. Inbar Y, Benyamini H, Nussinov R, *et al*. Combinatorial docking approach for structure prediction of large proteins and multi-molecular assemblies. *Phys Biol* 2005; **2**:S156–165.
101. Cheng TM, Blundell TL, Fernandez-Recio J. Structural assembly of two-domain proteins by rigid-body docking. *BMC Bioinformatics* 2008;**9**:441.
102. Wollacott AM, Zanghellini A, Murphy P, *et al*. Prediction of structures of multidomain proteins from structures of the individual domains. *Protein Sci* 2007;**16**:165–75.
103. Xu D, Jaroszewski L, Li Z, *et al*. AIDA: ab initio domain assembly for automated multi-domain protein structure prediction and domain-domain interaction prediction. *Bioinformatics* 2015;**31**:2098–105.