# A Survey of Computational Methods for Protein Structure Prediction *

Hsin-Chuan Yuan[a],Chang-Biau Yang[a†], Chiou-Yi Hor[b]

[a] Department of Computer Science and Engineering
National Sun Yat-sen University
Kaohsiung 80424, Taiwan

[b] Iron and Steel Research & Development Department
China Steel Corporation
Kaohsiung 81233, Taiwan

**Abstract**

Currently, predicting protein structures is still a challenge job for structural researchers. Various types of techniques and resources have been developed in the past decades to address this problem from different perspectives. Among these methods, the first and most intuitive one is referred to the *ab initio* method, which intends to model the protein folding process by means of physical laws. Other methods, such as the comparative modeling methods, are based on the recognition of spatial motifs in protein database. These template-based methods can be further divided into the homology modeling and fold recognition methods. In this paper, we will review these techniques. Totally, five protein structure prediction servers and four protein backbone reconstruction methods will be surveyed.

***Keywords***— protein backbone, bioinformatics, protein structure prediction, ab initio method, comparative modeling methods

## 1 Introduction

Various types of proteins are endowed with different biochemical functions. The function of a protein is generally determined by its three-dimensional structure. It is estimated that a human body may contain over two million proteins. These proteins are crucial to the survival, functions and activities of living organisms. In addition, the essential properties of proteins are highly correlated with their folded structures. Consequently, many biologists are working on the determination of protein structures.

In recent years, many researchers are devoted to the experiments of X-ray crystallographic and nuclear magnetic resonance (NMR), which enhances our understanding of protein structures and functions at the atomic level. However, plenty of time and cost are required to obtain the precise measurement with high resolution. *Structural genomics* [7] aims to build tertiary structures of proteins encoded by their genomes. Launched by many institutions and facilities from high-technology countries, a large number of efforts have been made to establish the high-throughput method of structure prediction based on homology modeling approaches, and thus have eased the gap between the rates of protein identification and of solving structures.

The *protein data bank* (PDB) archive [8], founded in 1971, contains information about the three-dimensional structures of proteins and nucleic acids. With the aid of new generation techniques, such as synchrotron radiation sources and high-resolution NMR, PDB now contains more than 100,000 protein structures. With the increase of knowledge as well as the database in biochemistry, we are able to deduce the function of each protein and its role in human health.

After obtaining the information about the sequence of insulin [91], determining the relationship between the sequences of proteins, also known as primary structures, and their functions became an essential issue pursued by biochemists. During this period, American biochemist Christian Anfinsen, studying on bovine pancreatic ribonuclease (RNase), observed that the sequence of amino acids along the protein chain could fold spontaneously to *native conformation* [5]. He further demonstrated that after changing the environment to nature condition, the protein could refold to

*This research was partially supported by the National Science Council of Taiwan under contract NSC 102-2221-E-110-033-MY2.

†Corresponding author. E-mail: cbyang@cse.nsysu.edu.tw (Chang-Biau Yang).

its native conformation and repossess its endowed functions spontaneously [4].

Anfinsen's experiment results paved a new road for protein structure prediction. His theory indicates that the sequence of amino acids contains all of the information needed for determining the 3D structure of a protein. Therefore, scientists had come up with computational methods for the prediction of protein tertiary structures. In the early stage, researchers applied computer techniques to finding the minimum energy conformation which is stable, stationary and maintaining enough equilibrium in terms of thermodynamics. According to the Levinthal's paradox [54], despite the fact that proteins folds quickly to their native state in seconds or less, it still takes much time and resources to enumerate the huge volume of possible configuration space in order to arrive at the minimum energy configuration. This encourages the use of the accurate functions or algorithms to compute the energy and narrow down the searching space. The simulations using molecular dynamics provide a perspective of the motion of the atoms governed via the solution of Newton's equations. The interactions between atoms such as covalent bonds, van der Waals force, electrostatic interactions, hydrogen bonds and hydrophobic interactions are considered as parameters in the force fields. Many methods using energy minimization approaches were proposed and have been applied in practice that came to be known as the *de novo* or *ab initio* methods [47].

During the same period when the *ab initio* methods were developed, the *homology modeling* approaches [37] also received much attention. The homology modeling refers to constructing a protein structure based on other homologous proteins with known structures. Since the homologous proteins share very similar framework with the target protein, and the protein folding is usually more conserved than protein sequence during evolution, the models thus can be built by means of the templates of known folding patterns. Many potential templates were also explored as the modeling methods were develops. This facilitates the expansion of protein database as well as the collection of protein families. The protein threading methods were proposed in the 1990s for dealing with the distantly homologous templates [16]. The method is used to model those proteins that lack homologous templates, but share compatible structures.

The paper is organized as follows. In Section 2, we first introduce the evaluation of predicted model, including the root mean square deviation, GDT-TS (Global Distance Test - Total Score) [113] and TM-score [120]. Section 3 covers several protein structure prediction methods. In Section 4, the experimental results of the prediction methods are presented. Finally, the conclusions are given in Section 5.

## 2 Preliminaries

### 2.1 Evaluation Methods of Structure Prediction

The evaluation method plays an important role as a metric for assessing structure prediction methods. After the prediction groups release their predicted structures, these reference target structures are then evaluated by employing criteria, which are derived from scientific knowledge or algorithms, such as sequence consistency, structure disorder, and alternative side chain conformations. In spite of the manual methods, which are performed by the visual inspection, several automatic evaluation methods are proposed and currently competent enough to estimate the quality of the modeling results and to evaluate the effectiveness of the corresponding prediction methods. However, no one single metric works well for evaluating all cases, especially for the template-free targets, for which the predicted models might be quite far from the exact structure.

**Root Mean Square Deviation (RMSD)** The *root mean square deviation* (RMSD), used in CASP1, CASP2 and CASP3, is widely used in three-dimensional geometry of molecules to measure the average distance between two molecular structures. Typically, a method for calculating the optimal rotation is required to obtain the optimal alignment before the RMSD is invoked. The Kabsch's algorithm [42, 43] proposed in 1976 is used to compute the rotation matrix by means of an optimization approach. The RMSD is, however, not ideal for all comparing cases since its quadratic nature may penalize errors very severely, and thus, only a few large deviations can result in high RMSD values. In addition, the RMSD depends on the number of atom pairs, and tends to increase with protein size.

Given two protein structures $A$ and $B$, the RMSD is defined as follows:

$$RMSD(A,B) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i^A - X_i^B)^2}. \quad (1)$$

In Equation 1, $n$ denotes the length of the given proteins, $X_i^A$ and $X_i^B$ denotes the atomic coordinates of the $i$th atom in the backbone of protein $A$ and $B$, respectively. The lower RMSD indicates the higher similarity between the two structures.

In some cases, the $C_\alpha$ atoms in the protein backbone are already known and consistent in both structures, the RMSD can be calculated by directly measuring the distance between the two corresponding atoms without rotation, in which two proteins were superimposed in a residue-by-residue manner.

**Global Distance Test-Total Score** To overcome the RMSD shortcomings, GDT-TS [113] was introduced and it was first used in CASP4. The GDT-TS is used to measure the similarity between two protein structures with identical amino acid sequences. It first collects all associated $C_\alpha$ atoms whose distances are below some thresholds and then counts the number of these $C_\alpha$ pairs. The thresholds used for the GDT-TS are 1, 2, 4, and 8 Å. For the purpose of automatic evaluation for assessing the overall quality of a model, GDT-TS has become one of the commonly accepted measures and has been widely used since CASP4. A modification of GDT-TS, GDT-HA (high accuracy) [82], was also used with the thresholds of 0.5, 1, 2, and 4 Å since CASP6. With more rigorous cut-off distances, it identifies a better model in the backbone prediction. The GDT-TS is computed as:

$$GDT\text{-}TS = 100 \times \frac{C_1 + C_2 + C_3 + C_4}{4 \times N} \qquad (2)$$

where $C_1$, $C_2$, $C_3$, and $C_4$ denotes the number of residues with distance below 1, 2, 4, and 8 Å, respectively. And $N$ denotes the total number of residues.

**Template Modeling Score (TM-score)** The TM-score, developed by the Zhang's Lab [120], is a metric for measuring the structural similarity between two protein models. It is intended to reduce the impact of local structural variations and to eliminate the inherent protein size dependence. Although the TM-score is not applied as an automatic evaluation in CASP experiments, many protein structure prediction methods adopted the TM-score to assess the prediction quality or even to guide the prediction procedure. TM-score [120] is defined as follows:

$$TM\text{-}score = \frac{1}{L} \sum_{i=1}^{L} \frac{1}{(1 + d_i/d_0)^2}, \qquad (3)$$

where $L$ is the length of the two aligned structures, $d_i$ is the distance of $i$th $C_\alpha$ pair between two models after superposition, and $d_0$ is a scale to normalize the match difference, which is approximately equal to $1.24\sqrt[3]{L-15} - 1.8$. As TM-score weights small distances stronger than the large distances, it is more sensitive to the global topology than RMSD. TM-score ranges from 0 to 1, where 1 indicates two identical structures, TM-score $<$ 0.17 indicates randomly chosen unrelated pairs, and TM-score $>$ 0.5 indicates that two structures have approximately the same topology.

# 3 The Protein Structure Prediction Methodologies

## 3.1 *Ab Initio* Protein Structure Prediction

*Ab initio* methods [47] provide a natural approach to obtain structures from protein sequences without referring to any appropriate templates. This is because *ab initio* methods are based entirely on basic physics and quantum mechanics. However, lots of computation is required in *ab initio* techniques and it is not guaranteed that accurate structures could always be obtained. In addition, *ab initio* methods are limited to small proteins (less than 100 or 150 residues), while template-based methods may deal with long proteins and they are more accurate in general. Nevertheless, *ab initio* methods are requisite when no appropriate templates can be consulted and it is still necessary to build the non-homologous loop regions.

Realizing the difficulties in folding simulations with molecular dynamics, many researchers have simplified the computational approaches to reduce the computational complexity. Ignoring the side chain is a common way to reduce the size of the conformation space. A further simplification is made by fixing the peptide bond angle and considering the peptide bond as a plane, while leaving the $\psi$ and $\phi$ angels to represent the main chain. In 1975, Levitt and Warshel [55] adopted a simplified representation for a protein structure by reducing the peptide chain to the $C_\alpha$ trace and the centroid of side chain. The experiment simulated the folding process for bovine pancreatic trypsin inhibitor (BPTI) from an open chain into a folded conformation and achieved a backbone RMSD in the range of 6.5Å. The introduction of lattice methods further improves the efficiency of

folding simulation, in which the protein backbone is restricted to be placed on vertices of a lattice, while the restriction does not necessarily impose on the side chain. Despite the well-known primitive cubic system, some attempts have also been made, such as the torsional space [73], tetrahedral [36], and face-centered cubic [49] system. The HP model, proposed by Lau and Dill [52], is a most widely known type of lattice model, in which all of the residues are classified as either *hydrophobic* (H) or *polar* (P) in a simple cubic lattice.

The fragment assembly [10] method manages to narrow down the searching space by applying some conformation constraints. The idea is based on the observation that short sequence fragments have strong local structural biases. A library for each fragment of the target protein is extracted from protein structure database by using fold recognition techniques. The fragments are typically small and extracted from multiple known structures. Once the fragments are assembled into a new conformation, its energy is evaluated. To minimize the energy, parts of fragments are replaced and the energy is evaluated again. The entire process is repeated until some termination conditions are satisfied. Since the fragment assembly searches may stuck in a local minimum, the Monte Carlo simulated annealing (MCSA) [48] is usually applied, in which the energetically unfavorable moves are allowed. These uphill moves are controlled by the difference of the objective functions and the temperature. The temperature is gradually decreased during the search process to find the final conformation with lowest energy.

### 3.2 Protein Homology Modeling

Although the first article of homology modeling was published in the late 1970s [30], much attention has been attracted after 1990 when hundreds of protein structures have been deposited in the Protein Data Bank [8]. Many protein sequences are evolutionarily related, and proteins with similar sequences usually share similar structures [45, 101]. According to a statistic from PDB, about 90% of protein structures submitted to the PDB during 1997-2002 share similar folds and structures that already existed in PDB. Moreover, protein structures are generally more conserved than sequences. This suggests that many protein structures can potentially be solved by the comparative modeling method.

The first homology model was built by Browne and his co-workers in 1969 by using the structure of lysozyme derived from X-ray as template to model bovine $\alpha$-lactalbumin with about 39% sequence identity [14]. This original method laid the cornerstone in the field and is still the most widely used method, when the pairwise sequence identity is above 40%. The process of homology-based methods usually starts with identification of homologous protein structure from Protein Data Bank; then, an alignment is carried out between the query sequence and the template structure; next, the method builds an initial model by performing insertions, deletions and residue replacements from the template structures; finally, the model is assessed and optimized.

### 3.3 Protein Treading/Fold Recognition

Similar to the comparative modeling, the fold recognition [16] was introduced to construct protein structure from known templates of known structures even there is no homologous protein deposited in the Protein Data Bank. The term "threading" is stand for the process of aligning a protein sequence onto a backbone structure [15] and it evaluates the sequence-structure compatibility with a set of potential scores or quasi-energy functions. The concept is based on the observation that the number of unique protein folds in nature is much smaller than the number of proteins in nature. The unique protein folds, according to previous studies [12, 39, 56, 57, 58, 70, 72, 105, 114], may range from hundreds to few thousands. Accordingly, instead of comparing templates and target sequences, it is more sensible to align the template structures with the target sequences. The threading-based method, which models the proteins with the experimental structures as templates, is a different approach from the homology modeling in terms of the methodology.

Many proteins consist of more than one unique structures or structural folds called protein *domains* [84, 106]. A structural domain is a conserved and distinct structural unit that folds independently against the rest of the protein chain, and usually is responsible for a particular function or interaction within a protein. Several programs are designed to partition a protein structure into individual domains [1, 38, 97]. The domains that are compatible with the target can be applied as templates to build the target model.

Fold recognition methods broadly fall into two categories. The first kind of method represents protein structures by labelling each amino acid

residue in the structure with the *environmental features*, e.g., the solvent accessibility of residues, the secondary structures, and so on. Then, the method aligns features between the target and template by making use of classical dynamic programming algorithm [11, 61]. The term "threading", first used by Bryant and Lawrence, was originally described the approach that adopted the statistical potential model in fold recognition [16]. The other kind of method considers the 3-D structure of the protein template based on a *statistical potential* [18, 93]. The structure-sequence compatibility is measured by the frequency of some or all of the residue pairs that are located at some spatial distances. Although this profile representation is much harder to use in calculating and assessing an alignment, it is still flexible in describing structures. For statistical potential models, the problem of identifying the alignment with highest compatibility is very difficult, especially when the pairwise interaction between residues is considered. Many programs [41, 109, 110] have been developed that adopt various heuristic strategies and some combinatorial algorithms such as the divide-and-conquer technique and linear programming to search for the optimal alignment.

The sequence-profile comparison method, conceived by Gribskov *et al.* [31], is another way to improve the sensitivity of template identification such as Position-Specific Iterative BLAST (PSI-BLAST) [3]. PSI-BLAST is used to search distant relatives of a protein by making use of the information of multiple sequences from same protein families. In PSI-BLAST, an amino acid substitution profile, called Position-Specific Scoring Matrix (PSSM), is generated in which residue substitution scores are given for each position of a protein sequence. Each entry in the matrix is a log-odd score calculated from the multiple sequence alignment constructed in a BLAST search. Positive scores indicate that the given residue is highly conserved and substitution occurs frequently; while negative scores indicate weakly conserved and substitution occurs less frequently. With the use of the evolutionary information, it not only improves the sensitivity of fold recognition, but also obscures the boundary between the homology modeling and fold recognition methods. Many studies now use PSI-BLAST as the first step to increasing the sensitivity of distant homology detection.

A significant improvement over profile-sequence comparison methods, introduced by Pietrokovski [79], was made by comparing profile to profile. With the creation of a profile database of sequences, each profile contains information on a protein family. The profile-profile alignment strategy can be implemented by aligning the profile of target with the profile of structural template.

# 4 A Review of Protein Structure Prediction Methods

## 4.1 Protein Structure Prediction Methods without Knowing $C_\alpha$ Trace

There are diverse protein structure prediction methods available, which adopt different approaches for constructing protein models. Five state-of-the-art and CASP-certified protein structure prediction servers are described in the following subsections.

### 4.1.1 I-TASSER

I-TASSER [87, 111, 119] is a typical meta-server for protein structure modeling built by the Yang Zhang Lab at the University of Michigan, Ann Arbor. Being a very prominent method, I-TASSER has been ranked as the top method and won first places in CASP7, CASP8, CASP10, and CASP11 competitions (came second in CASP9 competition).

The I-TASSER method can generally be divided into four stages: threading, structural assembly, model selection and refinement, and structure-based function annotation. In the first stage, the query is threaded through a PDB structure library by LOMETS [107] to identify structural templates. LOMETS was also developed by Yang Zhang Lab, which combines a number of profile-based fold recognition programs (FFAS-3D, HHsearch, MUSTER, pGenTHREADER, PPAS, PRC, PROSPECT2, SP3, and SPARKS-X) to generate protein models by collecting their target-template alignments. Each threading program has its unique features in threading process and therefore increases the diversity of the templates.

In the second stage, the query sequence is split into threading aligned regions and threading unaligned regions according to the alignment results from LOMETS. Continuous fragments of the threading aligned regions are extracted from the template structures, and are used to assemble new structural conformations with other regions,

while the unaligned regions (mainly loops/tails) are built by the *ab initio* modeling. In order to reduce the conformational searching space, a reduced model is adopted in which the protein chain are represented in terms of its $C_\alpha$ atoms and the side chain centers of mass. Moreover, the unaligned regions are restricted to a lattice system of grid size 0.87Å since the regions usually have a lower accuracy with the *ab initio* modeling. The threading regions are usually more accurate, therefore these regions are kept frozen in the simulation process to reserve the fidelity. The structure assembly is performed by a replica-exchange Monte Carlo (REMC) simulation technique. In the simulations, several replicas are performed in parallel at different temperatures. And the temperatures are exchanged between replicas periodically to avoid getting trapped in an attraction energy basin. The simulations are guided by a composite knowledge-based force field, which includes generic statistic potential derived from PDB (including hydrophobic interactions, $C_\alpha$ correlations and secondary structure propensity), hydrogen-bonding networks, and threading template-based restrains derived from LOMETS. In order to identify the models of lowest free-energy state, the structural decoys generated in the low-temperature replicas will be clustered by the SPICKER clustering program [121]. Then, by averaging the coordinates of the decoys within each SPICKER cluster, the centroid models will be obtained.

In the third stage, a second round of structural assembly simulation is performed against the selected centroid models to remove steric clashes and refine the global topology. In the assembly simulations, besides the constraints derived from centroid models, external constraints from LOMETS threading alignments and similar structures from PDB library (identified by TM-align [122]) are utilized. The decoys generated during the assembly simulation are clustered again and the structural models with lowest energy are then obtained. The final all-atom model is constructed by ModRefiner [108], which first builds the backbone from $C_\alpha$ trace and then constructs the side chain atoms from a rotamer library under the guide of a composite knowledge-based force field.

In the final stage, the structural model is matched against the BioLiP [71] database of ligand-protein interactions to identify homologous function templates. The function of the query protein is then inferred from the functional templates by three structure/function libraries of ligand-binding sites (LBS) [71], enzyme commission (EC) numbers [6], and Gene Ontology (GO) terms [29].

#### 4.1.2 Robetta

The Robetta [81, 96] server, developed by the Baker laboratory, provides both comparative modeling and *ab initio* approaches in the protein structure prediction problem. It has participated the CASP experiments as an automated prediction server since 2002, and it was reputed for providing accurate predicted protein models.

In the predicting process, targets are divided into individual domains which could fold independently. For domains with associated detectable templates or homologies, the Rosetta *de novo* protocol is then applied to modeling the protein structures. For domains without detectable templates or homologies, a comparative modeling, RosettaCM, is used to build the structures.

In the *de novo* process, structures are represented by using a reduced model in which the backbone are presented with its heavy atoms and the side chains are presented by its $C_\beta$ atom. For each query sequence, its sequence profile is generated by PSI-BLAST and its overlapping fragments with length of three and nine residues are extracted. In order to build a library of three- and nine-residue fragments corresponding to the fragments of query sequence, a profile-profiles similarity score over each fragment is calculated with $L_1$ norm regulation. In addition, the predicted secondary structure of query sequence is compared with the secondary structure of known structures in each fragment. This results in a ranking list of 200 nine-residue and 200 three-residue fragments for every fragments in the query.

The Rosetta *de novo* protocol utilizes the Monte Carlo simulated annealing (MCSA) search to assemble compact structures. In the simulation, the junctions between fragments and backbone torsion angles of fragments ($\psi$ and $\phi$) remain adjustable in the simulation. The simulation is started with the protein in a fully extended conformation. In the beginning, a consecutive nine-residue fragment on the query is randomly selected and the torsion angles in the fragment chain are replaced with a randomly selected fragment from the top 25 fragments in the ranking list. Each movement follows the evaluation of a physically realistic energy function, which accounts for short-range interactions including van der Waals force, hydrogen bonding, and desolvation but neglects the long-range electrostatic interactions. These

terms are progressively added to the total potential. After the assembly with 9-residue fragments, 3-residue fragments are used for short refinement. By gradually reducing the temperature over the iterations from 2500 to 10, the conformations with the lowest free energy are then considered as the final model.

For query sequence with reliable templates detected, modeling is carried out by using RosettaCM protocol. In this protocol, two inputs are used: the alignment of the target with proteins of known structures and Rosetta de novo fragment sets. The alignment are generated by using remote homology recognition programs such as PSI-BLAST [3], HHsearch [95], RaptorX [46], and Sparks-X [112]. The fragment sets, generated by Rosetta, are used to construct the unaligned regions.

The RosettaCM constructs models in three stages. In the first stage, the query sequence is threaded onto the template structures to generate a set of threaded partial models. A base model is then randomly selected from the partial threads while the coordinates of remaining threads are transformed to superimpose onto the base thread in order to achieve the minimal RMSD. The partial threads are then further divided into fragments according to the secondary structure assigned by DSSP [44]. Continuous helices and strands with partial loops segments connecting to them are added into the fragment list. The full-chain models are constructed by using a Monte Carlo simulation, in which the global position of each segment is represents in Cartesian space, whereas the backbone and side chain conformation in each segments are represented in the torsion space. Two types of movements are implemented in the simulations: (1) Substitution of the backbone torsion angles from randomly selected Rosetta *de novo* modeling fragments. (2) Substitution of the coordinates of a randomly selected template segment. The Rosetta low-resolution energy function is applied during the simulations, the scoring function accounts for non-local interactions such as compactness, hydrophobic burial, $\beta$ strand pairing, the distance restraints derived from template structures, and a penalty for chain separation between adjacent residues.

In the second stage, in order to further improve the overall topology of the models generated in the first stage, and close the loops, a two-step Monte Carlo simulation is carried out: (1) Substitution for a randomly selected backbone region with either a *de novo* fragment or a template fragment with the superposition covering all corresponding residues. (2) Quasi-Newton minimization is performed over the entire protein in Cartesian space by using a smoothed version of the Rosetta low-resolution function [85]. Since regions such as loops and segment boundaries are lack of reliable backbone geometry, the *de novo* fragment substitutions could be used for promote closure.

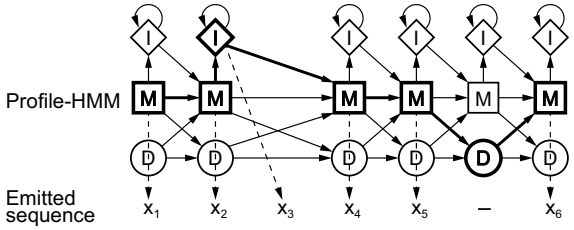


Figure 1: A profile HMM [95]. Match states indicate that the sequence has a character in the column, while deletion states do not. Insertion states allow an additional character between columns.

In the third stage, the side chain conformation is constructed by using the Rosetta Monte Carlo combinatorial optimization method, in which the strength of repulsive interaction is changed over the iterations. The annealing process, with Rosetta full-atom energy as objective function, is carried out to refine the backbone and side chain conformation.

#### 4.1.3 HHpred

Remote homology detection and sequence alignment are essential for the fold recognition. Hidden Markov Models (HMMs) have been exploited as protein modeling methods since 1990s [50]. A profile HMMs, used in sequence searching, is a left-to-right model with a series of nodes, where each node corresponds to a column in sequence or multiple sequence alignment (MSA) and it is comprised of a match state (M), an insertion state (I), and a deletion state (D), as shown in Figure 1. HHpred [83, 94, 95] first makes use of the HMM-HMM comparison to improve the sensitivity and alignment accuracy.

In the first step of HHpred, MSAs are built for the query sequence (or MSA) by an iterative HMM-HMM alignment method, named HHblits [83]. To further search for homologous templates in a large database, in the second step, a profile HMM is generated by HHsearch [94, 95] with the MSA from HHblits. Each HMM column contains the information about the 20 amino acid probabilities and the position-specific probabilities for

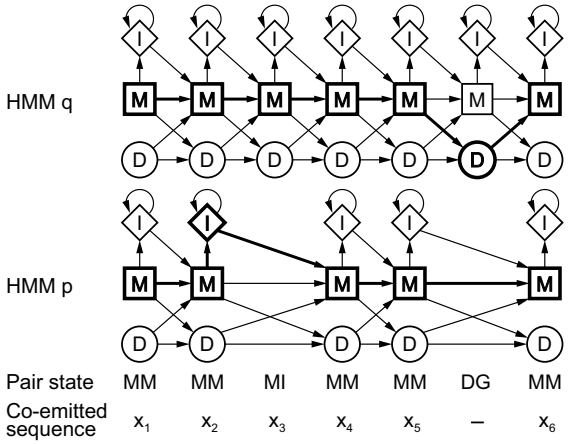


Figure 2: A pairwise alignment of profile HMMs by maximization of the log-sum-of-odd score [95].

insertion and deletion. Instead of log-odd scores, the log-sum-of-odd ones are adopted for the case of HMM-HMM comparison. The log-sum-of-odd score generalizes the log-odd score based on the co-emission probability. The alignment of two profile HMMs is shown in Figure 2.

In order to align two profile HMMs that maximize the log-sum-of-odds score,dynamic programming is used. By using dynamic programming for aligning the query HMM with each HMM in the selected HMM database, the optimal alignment is constructed with the maximum score.

In the final step, the HHpred alignments are used to construct the 3-D structural models by the MODELLER software [27, 89, 90]. The method starts by generating structural restraints that are derived from the multiple alignment of the target with the templates. It is based on the assumption that the distance between two residues in the target are generally close to the distance between the corresponding residues aligned in the template. In addition to the distance restraint, the stereochemical restraints such as bond angle, bond length, peptide bond dihedral angle and nonbonded van der Waals clash are also considered in the process. For loops or regions without a reliable template, the *ab initio* modeling are used, which largely rely on the law of physics. By optimizing the objective function that considers both spatial restraints and the CHARMM22 [64] force field terms, a proper model is built in Cartesian space.

#### 4.1.4 RaptorX

Similar to HHpred, RaptorX [62, 75, 76, 77, 78] largely focuses on the quality of sequence-template alignment, and it also makes use of MODELLER [27, 89, 90] to construct 3-D models based on the multiple sequence alignment technique (described in the subsection of **HHpred**). In addition to the sequence profiles, RaptorX also exploits structure information making it excel in modeling of protein sequences without a large number of homologies. For each query sequence, it search for the Pfam database [26] in advance to examine whether it belongs to multiple domains or not. When at least one Pfam entry is identified, the query sequence is divided into several domains. RaptorX employs sequence information (sequence profile) and structure information of two residues being aligned and their neighborhood information to improve the detection sensitivity. In order to deal with the correlation among protein features (including both sequence profile and structure information), a conditional Neural Field (CNF) [74] method, called CNFpred, is proposed, which uses non-linear score function to guide the sequence-template alignment.

In the CNFpred, three states $M$, $I_t$ and $I_s$ are used to represent alignments of each residue. They are a match at both proteins, an insertion at the template protein, and an insertion at the target sequence, respectively. Since two residues are considered simultaneously, it conducts nine possible state transitions. Each log-likelihood of state transition is estimated by a corresponding scoring function, which is a neural network with 12 hidden neurons and its individual parameters. The concept of the CNFpred is illustrated in Figure 3. The optimal alignment can be calculated by Viterbi algorithm [100].

In order to avoid over-fitting while training the model parameters, a $L_2$-norm regulation, determined by a 5-fold cross validation, is applied to restrict the search space. The CNF model is trained by maximizing the expected TM-score [120] of a set of reference alignments, which takes the conservation information of aligned residue pairs into consideration. The expected TM-score is defined as follows:

$$Q = \frac{1}{N}\sum_i (w_i MAG_i) \tag{4}$$

$$w_i = \frac{1}{1+(d_i/d_0)^2}, \tag{5}$$

where $N$ is the shorter length of the aligned pair, $w_i$ is the local TM-score at alignment position $i$ and $MAG_i$ is the marginal alignment probability at position $i$. The local TM-score of the

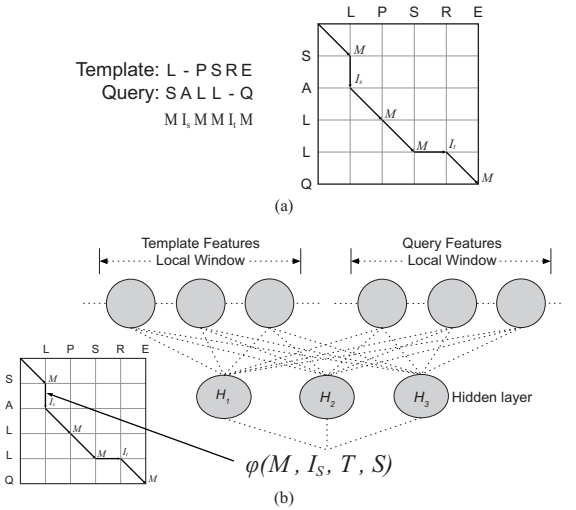


Figure 3: The concept of CNFpred [74]. (a) An example of a sequence-template alignment. Here, each path represents an alignment result, and the probability of the path is estimated by the CNF model. (b) The log-likelihood score of the state transition from $M$ to $I_s$, which is yielded by the neural network with one hidden layer.

aligned residues (of two superimposed proteins) at position $i$ are calculated with the distance deviation between aligned residues ($d_i$) and the normalization constant ($d_0$) depending only on protein length. The local TM-score ranges from 0 to 1 with higher values indicating more conserved positions. The score at the gap position is equal to 0. The marginal alignment is the accumulative probability of all possible alignments of two proteins in which these two residues are aligned together. The probability can be computed by the forward-backward algorithm [51]. Since the expected TM-score in Equation 4 is not concave, the function is then optimized by L-BFGS [60]. The optimization process is repeated with different starting positions to find a best solution.

### 4.1.5 MULTICOM

MULTICOM [20, 59, 102] is a server that takes advantage of diverse protein structure and structural feature prediction tools and systematically integrates the tools in its system architecture. Since a number of diverse tools are involved for prediction, each of which has its unique strength, MULTICOM has delivered outstanding performance and generated protein conformations of good quality in recent CASP competitions.

MULTICOM consists of five major steps for protein structure prediction. Each step is developed for specific tasks that other protein structure prediction methods specialize. The five steps are template identification, query-template alignment and combination, model generation, model assessment, and model refinement. The flowchart of the MULTICOM protein structure prediction system is illustrated in Figure 4. In order to support the template-based modeling, a number of information, including template sequence, secondary structure, tertiary structure, solvent accessibility (assigned by DSSP [44]), and sequence profiles are collected from multiple protein databases and tools.

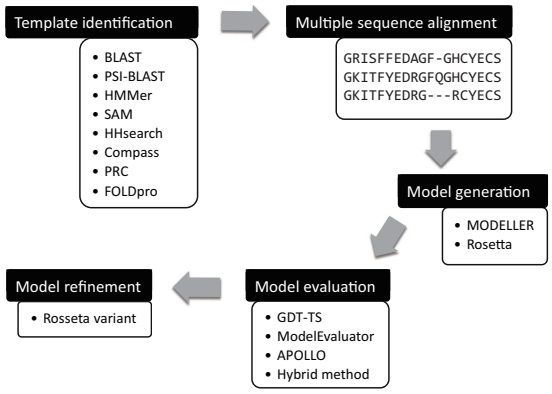


Figure 4: The five-step flowchart of MULTICOM system.

In the first step of MULTICOM, a query sequence is first searched by PSI-BLAST [3] through the non-redundant protein sequence database to identify its homologous sequences. Then, several query profiles (such as PSI-BLAST PSSM, HHsearch HMM [94], SAM HMM [40], HMMER HMM [25], PRC HMM [65], and COMPASS profile[88]) are generated from the query and its homologies, and is searched by corresponding searching tools against the template profiles in the libraries. Meanwhile, the query sequence is searched by using BLAST [2, 3] and CSI-BLAST [9]. Each searching tool may return a list of templates and the local sequence-template alignments with E-values describing the expected number of chance hits. In other words, a low E-value indicates that the probability of finding a match just by chance is small. The top-ranking templates and their query-template alignments are stored separately. In addition, a consensus list of templates frequently identified by each method is also stored for later combination.

In the second step, multiple template combi-

nation is performed, which is essential for more accurate predictions by removing structurally inconsistent templates. Three approaches are implemented for the multiple template combination. The first one is the central-star combination method [33, 34], which intends to integrate each list of templates generated by the searching tool. The second approach is *structure-alignment-driven profile alignment*. The third approach is a *hybrid alignment combination* approach developed in the MULTICOM system. It first takes each query-template alignment in the consensus list generated by PSI-BLAST as a seed. For the uncovered regions, the HHSearch alignment is added if it is available. Finally, the SPEM [123] global alignments are applied to the remaining uncovered regions.

In the third step, for the combined query-template alignments in which one or more templates are found to cover almost the entire query, the template-based modeling tool MODELLER [27, 89, 90] is applied directly to generate a number of conformations (described in the **HHpred** subsection). For a query sequence without reliable templates or being only covered partially, a recursive protein modeling protocol [21] is used. This protocol first constructs the well-aligned regions of the query protein and keeps these regions fixed. A fragment-assembly tool Rosetta [53] is then used to sample the unaligned regions. In this step, a number of candidate models are produced for the query protein.

In the fourth step, a *support vector regression* (SVR) [22] is used to predict the GDT-TS of a model generated from the alignment. It takes the information of pairwise query-template alignment to generate the model (including the E-value of query-template alignment, the percentage of identical residue pairs in aligned position, the percent of aligned pairs of query in an alignment, and BLOSUM (blocks substitution matrix) [35] scores of all aligned residue pairs) as features in SVR. The input and output vectors for training are extracted from pairwise protein alignments for the query sequences and its corresponding GDT-TS. In addition to the SVR method, three model quality assessment methods (including ModelEvaluator [104], APOLLO [103], and a hybrid method) are applied. In this step, all models are ranked based on the scores described above.

In the final step, some poorly predicted regions (e.g., loop, tail regions) are re-sampled by a Rosetta variant [85], which keeps the reliable regions fixed while attempts to work over on the local uncertain regions. The model combination approach is essentially based on a model averaging process, which combines the top-ranking models that are globally similar to the query protein or possess some similar local regions. In the end of this refinement step, the most prominent models are delivered as the final model.

## 4.2 Protein Structure Prediction Methods with $C_\alpha$ Trace

For protein structure prediction methods that adopt reduced representation of proteins in the modeling simulations, only the $C_\alpha$ atoms of the backbone are retained (in some cases, the centroids of the side chains are also retained). In this situation, constructing protein atomic structures from $C_\alpha$ traces becomes an important intermediate step for the construction of full atomic models. In other words, it is assumed that the 3D coordinates of $C_\alpha$ are given. And the goal is to construct the 3D coordinates of other atoms. Four methods are described in the following subsections.

### 4.2.1 PULCHRA

PULCHRA (Protein Chain Reconstruction Algorithm) [86] is a tool, which can be installed locally, for the full-atom reconstruction from reduced representation of proteins. The method generally consists of three steps: optimization of $C_\alpha$ positions, backbone reconstruction and optimization, side chain reconstruction and optimization. Each step in the program is organized independently, which therefore can be adopted by the user for different applications. The backbone and side chain rotamer libraries are built with a list of 1351 experimental protein structures filtered by 35% sequence identity.

In the first step of PULCHRA, the $C_\alpha$ positions are optimized by removing irregular configurations. In the second step, the backbone reconstruction procedure is performed, which is based on the idea originally proposed by Purisima and Scheraga [80] and implemented by Milik *et al.* [68]. In the procedure, the fragments made up of four successive $C_\alpha$ are used to rebuild the peptide bond atoms between the second and third $C_\alpha$ atoms. It is based on the assumption that the atomic position distribution of the center peptide bond of the two fragments is similar if the structures of two fragments are also similar. Three distances between the first and third ($r_{13}$), second and fourth ($r_{24}$), and first and fourth ($r_{14}$) $C_\alpha$ atoms are re-

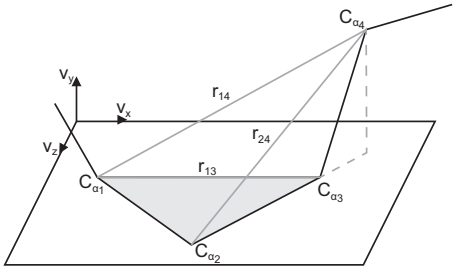


Figure 5: The structure of a fragment of consecutive $C_\alpha$ atoms. The three distances between the non-consecutive $C_\alpha$ atoms are used to form the backbone skeleton. The local coordinate system is used to rebuild the backbone plate between the second and third $C_\alpha$ atoms.

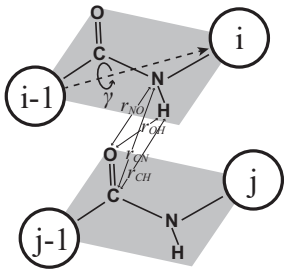


Figure 6: The procedure of hydrogen bond pattern optimization [86].

trieved, as shown in Figure 5. The sign of $r_{14}$ indicates the chirality of the chain. Subsequently, $r_{13}$ and $r_{24}$ are divided into 10 bins ranging from 4.5 to 7.5Å; $r_{14}$ is divided into 75 bins ranging from -11 to 11Å. A lookup table formed by these bins is used to select proper fragments from the backbone library. A local Cartesian coordinate system involving three $C_\alpha$-$C_\alpha$ vectors is used to represent the peptide bond:

$$v_x = \frac{v_{13}}{|v_{13}|}, \tag{6}$$

$$v_y = \frac{v_{23} \times v_{12}}{|v_{23} \times v_{12}|}, \tag{7}$$

$$v_z = v_x \times v_y, \tag{8}$$

where $v_{xy}$ denotes the vector pointing from the $x$th to the $y$th $C_\alpha$ atoms. To adjust all distorted hydrogen bonds in the reconstructed backbone, the hydrogen bond (C-O$\cdots$H-N) are rotated along the $C_\alpha$-$C_\alpha$ virtual bonds, and the hydrogen bond energy is calculated iteratively, as shown in Figure 6. This energy function is described in the DSSP program [44] and defined as follows:

$$E_{HB} = 332 q_1 q_2 (\frac{1}{r_{ON}} + \frac{1}{r_{CH}} + \frac{1}{r_{OH}} + \frac{1}{r_{CN}}), \tag{9}$$

where $q_1 = 0.42e$ and $q_2 = 0.20e$, with $e$ being the electron charge unit, $r_{XY}$ is the distance between atoms $X$ and $Y$ (Å), and $E_{HB}$ is the energy in kcal/mole. Meanwhile, for the energy calculation, the hydrogen atoms are reconstructed in this step.

In the final step, the side chains are reconstructed with the same reference frame. Each bin in the lookup table contains a list of possible side chain conformations. The conformation that is closest to the CM is adopted for reconstruction. If no CM is given on input, the conformation that has the highest occurrence in the rotamer library is adopted. If heavy atom clashes (distance between atoms $< 2$Å) are present, the second most frequently side chain conformation in the library is adopted. If the clashes still arises, the side chain is rotated around a virtual $C_\alpha$-CM bond until a legal conformation is obtained. After adding hydrogen atoms, the full-atom model is returned as the final prediction.

### 4.2.2 SABBAC

SABBAC [66] is a protein backbone reconstruction server based on the assembly of fragments from a reduced protein library. In order to select the appropriate fragments for the structure construction, Hidden Markov Model (HMM) is used to decompose the protein backbone conformation into series of overlapped fragments. Each fragment is made up by four successive $C_\alpha$ atoms, each of which is associated with a state in HMM. This results in a list of representative structural fragments, called *structural alphabet* (SA). A structural alphabet of 27 representative fragments (states) is obtained by using a statistical criterion (i.e. the Bayesian information criterion (BIC)), which optimally decomposes the local conformations of protein structures. The SA-27 alphabet is therefore used to describe the conformational variability [17]. The query $C_\alpha$ trace is encoded into a trajectory of letters in discrete structural alphabet space by using Viterbi algorithm [100]. After applying a standard dynamic clustering procedure, which encodes a number of similar prototype fragments with a letter, a collection of 155 fragments are yielded to describe the 27 letters [98, 99].

Following the procedure described by Milik *et al.* [68], the coordinates of atoms between the second and third $C_\alpha$ atom are computed. A greedy algorithm [98, 99] is then applied to search the

optimal combination from a collection of candidate fragments and associates each position of the structure with one letter of structural alphabet. By adding one residue into each iteration along the N- toward C-terminus, possible assemblies are produced. These assemblies are then assessed by an energy criterion (see below) and ten assemblies of them are retained for the next iteration. In the end of the process, the assembly associated with the lowest score is returned as the final prediction of the structure.

For energy evaluation, the OPEP force field [92] is used, which is defined as follows:

$$E = E_{SC,SC}+E_{HB}+E_{VdW}+E_{PhiP}+E_{BB}+E_{Trans}, \quad (10)$$

where $E_{SC,SC}$ denotes the side chain-side chain interactions, $E_{HB}$ represents hydrogen bonds, $E_{VdW}$, $E_{PhiP}$, and $E_{BB}$ describe the van der Waals force, the $\phi$ positive contribution, and the valence angle distortion of $C_\alpha$, respectively. $E_{Trans}$ is the sum of the log-likelihood of all transitions between consecutive fragments.

### 4.2.3 BBQ

The BBQ (Backbone Building from Quadrilaterals) method [32] adopts the same approach for backbone reconstruction proposed by Milik *et al.* [68] (for the detailed description of the backbone reconstruction procedure, please refer to the subsection for **PULCHRA**). The library contains 1259 protein structures gathered from PDB. The proteins are then filtered with sequence identity higher than 90% and decomposed into fragments. This results in a library of 263,000 fragments. For each distance that is used to define a tetrapeptide, it is binned by 0.2Å. The distances $r_{13}$ and $r_{24}$ (refer to Figure 5) vary from 4.0 to 7.6Å, and the distance $r_{14}$ ranges from -11.0 to -4.0Å and 4.0 to 11.0Å. In some rare cases, a specific conformation is not in the lookup table, a number of neighboring bins are inspected. If all these bins are empty, then a fragment which is closest to the query one in term of Euclidean distance $r^{QD}$ is chosen from the library. The distance $r^{QD}$ is defined as follows:

$$r^{QD} = \sqrt{(r^Q_{13} - r^D_{13})^2 + (r^Q_{24} - r^D_{24})^2 + (r^Q_{14} - r^D_{14})^2}. \quad (11)$$

### 4.2.4 PD2 ca2main

The PD2 ca2main [69] is a tool for backbone reconstruction, which uses *Gaussian mixture models* (GMMs) to constructed a structural alphabet

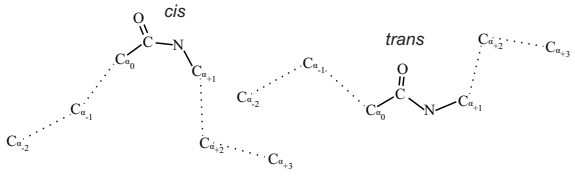


Figure 7: The *cis* and *trans* fragments of six consecutive residues [69].

(SA). By aligning the segment of the target model to the best fitted letter along the $C_\alpha$ trace, the central peptide bond of that segment can be reconstructed. A library of fragments derived from ASTRAL SCOP17.5A [19] are built, among which fragments with sequence identity and homology above 40% are filtered (identified by PSI-BLAST [3]) to avoid overfitting. Each fragment comprises six consecutive residues and is centered on an idealized peptide bond built from CHARMM definitions [13]. *Cis* and *trans* fragments are modeled with separate GMMs and are illustrated in Figure 7. In order to fit a GMM, the $C_\alpha$ coordinates of each fragment in the library are encoded into to a 12-dimensional data point $x$. Since the idealized peptide bond is fixed, the positions of $C_\alpha$ atoms connecting to it can only vary slightly. The above mentioned data point that contains four set of $C_\alpha$ coordinates with the $C_{\alpha_0}$ and $C_{\alpha_{+1}}$ excluded is given as follows:

$$x = (C^x_{\alpha_{-2}}, C^y_{\alpha_{-2}}, C^z_{\alpha_{-2}}, \ldots, C^x_{\alpha_{+3}}, C^y_{\alpha_{+3}}, C^x_{\alpha_{+3}}). \quad (12)$$

The hierarchical clustering is applied on a random sample of 2000 points to initiate the GMM for the subsequent expectation-maximization (EM) algorithm [23]. Then, the EM algorithm proceeds iteratively with the complete dataset until convergence is reached. The *Bayesian information criterion* (BIC) is used to determine the optimal number of model components and this results in a 528-component mixture model. The specific $C_\alpha$ coordinates of fragments in each component are averaged with fixed central peptide bond atom positions, resulting in the constitution of the structural alphabet. Given a target fragment, the alphabet member that minimizes a weighted $C_\alpha$ RMSD is then used to reconstruct the atoms in central peptide plate. The weighted RMSD technique is used to reduce the impact of some Ca with large coordinate deviations. That is, the inner $C_\alpha$ position gains more weight and vice versa. The weight ($W$) for each $C_\alpha$ position is given as

follows:

$$
\begin{aligned}
W(C_{\alpha_{-3}}) &= \frac{1}{100}, \\
W(C_{\alpha_{-2}}) &= \frac{1}{10}, \\
W(C_{\alpha_{-1}}) &= 1, \\
W(C_{\alpha_0}) &= 1, \\
W(C_{\alpha_{+1}}) &= \frac{1}{10}, \\
W(C_{\alpha_{+2}}) &= \frac{1}{100}.
\end{aligned} \tag{13}
$$

The letter (alphabet member) with the lowest weighted RMSD to the target fragment is then used to reconstruct the peptide bond on the target structure, in which $C_\beta$ and amide hydrogen atom coordinates are added with CHRAMM residue definitions.

In addition to the EM algorithm, an gradient energy minimization procedure is also provided optionally, which is guided by a simple backbone potential energy function [63]. The function considers local structure molecular mechanics terms derived from the OPLS-UA force field, a soft steric repulsive term for atom pairs separated by more than four bonds and statistical backbone hydrogen bonding potential terms.

## 5 Experimental Results

A comparison of five available servers that provide the structure prediction functions is given in Table 1. These servers are compared with respect to the following aspects: Dose the server predict secondary structures and/or tertiary structures? Dose the methodology adopt (distance) homology modeling or *ab initio* approach? Dose the server detect and assign domains for large target sequences? Dose the server collect models from other server and use that input to produce consensus structures (metaserver)? And dose the server provide functional annotations for predicted structures?

Similarly, for the four backbone reconstruction tools, a comparison of their properties is shown in Table 2. The reconstruction procedures of these four methods include identifying proper templates from their library and assembling the templates into complete conformations. SABBAC and PD2 ca2main are available in terms of web services. PULCHRA and BBQ are two local-installed programs. Besides, PULCHRA and PD2 ca2main provide additional functionalities for the reconstruction of side chain atoms.

Table 3: Average scores of 172 targets in CASP8, the data were retrieved from Zhang lab website [117]. Note that a server with higher TM-score, GDT-TS, GDT-HA and lower RMSD is considered more accurate.

| CASP8 | TM-score | GDT-TS | GDT-HA | RMSD |
|---|---|---|---|---|
| I-TASSER | 0.7067 | 66.65 | 49.73 | 8.4 |
| Robetta | 0.6746 | 63.53 | 46.92 | 10.5 |
| HHPred | 0.6659 | 62.55 | 46.58 | 13.4 |
| RaptorX | 0.6803 | 64.38 | 48.19 | 9.6 |
| MULTICOM | 0.6695 | 63.15 | 47.17 | 11.0 |

Table 4: Average scores of 144 targets in CASP9, the data were retrieved from Zhang lab website [118]. Note that a server with higher TM-score, GDT-TS, GDT-HA and lower RMSD is considered more accurate.

| CASP9 | TM-score | GDT-TS | GDT-HA | RMSD |
|---|---|---|---|---|
| I-TASSER | 0.6689 | 61.53 | 44.47 | 7.3 |
| Robetta | 0.6259 | 57.19 | 40.99 | 9.0 |
| HHPred | 0.6360 | 58.78 | 43.06 | 9.1 |
| RaptorX | 0.6554 | 60.26 | 43.36 | 7.2 |
| MULTICOM | 0.6381 | 58.74 | 42.31 | 9.1 |

To compare the performance of the five servers, we utilize the statistics adopted in 8th, 9th, 10th, and 11th CASP competitions and made publicly available by the Zhang lab [115, 116, 117, 118]. The quality of the prediction are assessed with the TM-score, GDT-TS, GDT-HA, and RMSD (only calculates $C_\alpha$ atoms). In general, the TM-score tends to be more sensitive to the global topology, while GDT-TS and GDT-HA tend to be more sensitive to the quality of local structures [108]. As shown in Tables 3, 4, 5, and 6, most servers basically achieves the similar level of accuracy with the exception that I-TASSER outperforms the other servers over the four CASP competitions. Since I-TASSER collects structures generated from other protein structure prediction servers, which adopts a variety of fold recognition techniques, it thus can take advantage of their merits.

To assess the performance of the four backbone reconstruction methods, the $C_\alpha$ trace of targets in CASP competitions are extracted. Each $C_\alpha$ structure is input to the reconstruction methods and RMSDs are calculated. Since the $C_\alpha$ atoms in the backbone reconstruction are fixed, only heavy atom pairs along protein backbones (including N,

Table 1: Comparison of five protein structure prediction servers.

| Name | Availability | SSP | TSP | TM/FM | DP | M/S | FA |
|---|---|---|---|---|---|---|---|
| I-TASSER | http://zhanglab.ccmb.med.umich.edu/I-TASSER/ | Yes | Yes | TM, FM | Yes | M | Yes |
| Robetta | http://robetta.bakerlab.org/ | Yes | Yes | TM, FM | Yes | S | No |
| HHPred | http://toolkit.tuebingen.mpg.de/hhpred | Yes | Yes | TM | No | S | No |
| RaptorX | http://raptorx.uchicago.edu/ | Yes | Yes | TM | Yes | S | No |
| MULTICOM | http://sysbio.rnet.missouri.edu/multicom_cluster/ | No | Yes | TM, FM | No | S | No |

SSP, secondary structure prediction; TSP, tertiary structure prediction; FM, template-free modeling;
TM, template-based modeling; DP, domain parsing; M/S, metaserver or single server; FA, functional annotation.

Table 2: Comparison of four servers/programs for backbone reconstruction from $C_\alpha$ trace.

| Name | Availability | TM/FM | SC |
|---|---|---|---|
| PULCHRA | downloadable program | TM | Yes |
| SABBAC | http://bioserv.rpbs.jussieu.fr/SABBAC.html | TM | No |
| BBQ | downloadable program | TM | No |
| PD2 ca2main | http://www.sbg.bio.ic.ac.uk/∼phyre2/PD2_ca2main/ | TM | Yes |

FM, template-free modeling; TM, template-based modeling; SC, side chain reconstruction.

Table 5: Average scores of 123 targets in CASP10, the data were retrieved from Zhang lab website [115]. Note that a server with higher TM-score, GDT-TS, GDT-HA and lower RMSD is considered more accurate.

| CASP10 | TM-score | GDT-TS | GDT-HA | RMSD |
|---|---|---|---|---|
| I-TASSER | 0.7283 | 67.73 | 48.98 | 7.0 |
| Robetta | 0.6837 | 64.27 | 46.94 | 8.1 |
| HHPred | 0.6316 | 58.89 | 43.56 | 9.0 |
| RaptorX | 0.6334 | 59.28 | 43.72 | 7.5 |
| MULTICOM | 0.6608 | 61.76 | 45.02 | 7.9 |

Table 6: Average scores of 117 targets in CASP11, the data were retrieved from Zhang lab website [116]. Note that a server with higher TM-score, GDT-TS, GDT-HA and lower RMSD is considered more accurate.

| CASP11 | TM-score | GDT-TS | GDT-HA | RMSD |
|---|---|---|---|---|
| I-TASSER | 0.5760 | 52.22 | 36.88 | 5.7 |
| Robetta | 0.5204 | 47.54 | 34.12 | 6.5 |
| HHPred | 0.5053 | 45.96 | 32.40 | 8.0 |
| RaptorX | 0.5186 | 47.03 | 33.32 | 7.7 |
| MULTICOM | 0.5211 | 47.52 | 33.77 | 6.8 |

Table 7: Average RMSDs for datasets ranging from CASP8 to CASP11.

| | CASP8 | CASP9 | CASP10 | CASP11 |
|---|---|---|---|---|
| PULCHRA | 0.5951 | 0.5989 | 0.5261 | 0.5476 |
| SABBAC | 0.4934 | 0.5551 | 0.4402 | 0.4705 |
| BBQ | 0.4585 | 0.428 | 0.3717 | 0.3774 |
| PD2 ca2main | 0.4042 | 0.4096 | 0.3386 | 0.3491 |

C, and O atoms) are involved for the RMSD calculation. Except for PD2 ca2main, all other three methods adopts the similar procedures [68]. The experimental results are presented in Table 7. It is shown that the overall results of PD2 ca2main are better than the other methods.

## 6 Conclusions

One of the ultimate goals for protein structure prediction is to predict the complete 3-D structure of a protein. This facilitates practitioners to extend their knowledge of the structures and functions of proteins without resorting to the experimental methods. Higher sensitivity for template detection and folding simulation can be achieved as the advancement of protein structure prediction tools as well as the growth of the protein databases. This can in turn further enhance the performance of both the *ab initio* method and comparative modeling method.

In this paper, we survey a variety of protein structure reconstruction methods and introduce their underlying working principles. Besides, we also examine their prediction capabilities with the benchmark datasets ranging from CASP8 to CASP11. The results would give bioinformatics practitioners a preliminary guide of how they work, what their limitations are and how their performances are. Since the above tools are developed from different perspectives, they may demonstrate some complement in their capabilities. This would lay a good foundation to our future research in this field.

Many prediction servers adopted several algorithms or existing tools in their approaches. They more or less incorporate the sequence and structural information and potential energy functions, such as secondary structure prediction, distance homologies detection, multiple structure alignment, and folding simulation. However, the more information is considered in the approach, the more computational complexity it would have. Various improved alignment and modeling methods have been proposed to overcome this problem and yielded accurate and high-throughput prediction. Seeking a proper tool and inspecting their contribution and adequacy become an important issue. Because new target sequences with the different lengths, number of homologies, and so on, vary in a wide range, this makes it tough for the determination of the parameters of tools. For this reason, the modeling methods are updated constantly. Fortunately, a number of assessment methods provide us a perspective to select and evaluate among algorithms or tools, and therefore the tool developers can improve modeling methods accordingly.

As an important aspect of structural biology, structure-based drug design (SBDD) [67] facilitates the discovery and development of protein therapeutics. The SBDD relies heavily on the knowledge of atomic-level protein structures, and protein-ligand interactions. Benefited from the rapid growth of protein structure databases, which allows for the identification of ligand binding sites [28, 124], and thus provides an insight into the application of SBDD. However, existing structure modelling methods currently produce considerable errors, hence this would impede the progress of drug design. Besides, other alternatives, like X-ray crystallography, can hardly reveal the dynamic behavior of proteins. It also limits the progress of SBDD [24]. Many of those problems are related to the field of protein structure prediction and remain challenges for the structural researchers.

## References


[1] N. Alexandrov and I. Shindyalov, "PDP: protein domain parser," *Bioinformatics*, Vol. 19, pp. 429–430, 2003.

[2] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, Vol. 215, No. 3, pp. 403–410, 1990.

[3] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, W. M. Z. Zhang, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, Vol. 25, No. 17, pp. 3389–3402, 1997.

[4] C. B. Anfinsen, E. Haber, M. Sela, and F. H. W. Jr, "The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain," *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 47, pp. 1309–1314, 1961.

[5] C. B. Anfinsen, R. R. Redfield, W. L. Choate, J. Page, and W. R. Carroll, "Studies on the gross structure, cross-linkages, and terminal sequences in ribonuclease," *Journal of Biological Chemistry*, Vol. 207, pp. 201–210, 1954.

[6] A. Bairoch, "The ENZYME database in 2000," *Nucleic Acids Research*, Vol. 28, pp. 304–305, 2000.

[7] D. Baker and A. Sali, "Protein structure prediction and structural genomics," *Science*, Vol. 294, pp. 93–96, 2001.

[8] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic Acids Research*, Vol. 28, pp. 235–242, 2000.

[9] A. Biegert and J. Soding, "Sequence context-specific profiles for homology searching," *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 106, pp. 3770–3775, 2009.

[10] J. U. Bowie and D. Eisenberg, "An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function," *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 91, pp. 4436–4440, 1994.

[11] J. U. Bowie, R. Luthy, and D. Eisenberg, "A method to identify protein sequences that fold into a known three-dimensional structure," *Science*, Vol. 253, pp. 164–170, 1991.

[12] S. E. Brenner, C. Chothia, T. J. Hubbard, and A. G. Murzin, "Understanding protein

structure: Using scop for fold interpretation," *Methods in Enzymology*, Vol. 266, pp. 635–643, 1996.

[13] B. R. Brooks, C. L. Brooks III, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. B. A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, "CHARMM: The biomolecular simulation program," *Journal of Computational Chemistry*, Vol. 30, pp. 1545–1614, 2009.

[14] W. Browne, A. North, D. Phillips, K. Brew, T. C. Vanaman, and R. L. Hill, "A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme," *Journal of Molecular Biology*, Vol. 42, pp. 65–86, 1969.

[15] S. H. Bryant and C. E. Lawrence, "An empirical energy function for threading protein sequence through the folding motif," *Proteins*, Vol. 16, pp. 92–1992, 1993.

[16] C. Bystroff and D. Baker, "Prediction of local structure in proteins using a library of sequence-structure motifs," *Journal of Molecular Biology*, Vol. 281, pp. 565–577, 1998.

[17] A. C. Camproux, R. Gautier, and P. Tuffery, "A hidden markov model derived structural alphabet for proteins," *Journal of Molecular Biology*, Vol. 339, pp. 591–605, 2004.

[18] G. Casari and M. J. Sippl, "Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds," *Journal of Molecular Biology*, Vol. 224, pp. 725–732, 1992.

[19] J. M. Chandonia, G. Hon, N. S. Walker, L. L. Conte, P. Koehl, M. Levitt, and S. E. Brenner, "The ASTRAL compendium in 2004," *Nucleic Acids Research*, Vol. 32, pp. 187–192, 2004.

[20] J. Cheng, J. Li, Z. Wang, J. Eickholt, and X. Deng, "The MULTICOM toolbox for protein structure prediction," *BMC Bioinformatics*, Vol. 13, p. 2, 2012.

[21] J. Cheng, Z. Wang, J. Eickholt, and X. Deng, "Recursive protein modeling: a divide and conquer strategy for protein structure prediction and its case study in CASP9," *Journal of Bioinformatics and Computational Biology*, Vol. 10, No. 3, 2012.

[22] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, Vol. 20, pp. 273–297, 1995.

[23] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, Vol. 39, pp. 1–38, 1977.

[24] M. A. DePristo, P. I. de Bakker, and T. L. Blundell, "Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography," *Structure*, Vol. 12, pp. 831–838, 2004.

[25] R. D. Finn, J. Clements, and S. R. Eddy, "HMMER web server: interactive sequence similarity searching," *Nucleic Acids Research*, Vol. 39, pp. W29–W37, 2011.

[26] R. D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. Sonnhammer, S. R. Eddy, and A. Bateman, "The Pfam protein families database," *Nucleic Acids Research*, Vol. 38, pp. D211–D222, 2010.

[27] A. Fiser and A. Sali, "Modeller: generation and refinement of homology-based protein structure models," *Methods in Enzymology*, Vol. 374, pp. 461–491, 2003.

[28] C. J. Francoijs, J. P. Klomp, and R. M. Knegtel, "Sequence annotation of nuclear receptor ligand-binding domains by automated homology modeling," *Protein Engineering*, Vol. 13, pp. 391–394, 2000.

[29] Gene Ontology Consortium, "The Gene Ontology: enhancements for 2011," *Nucleic Acids Research*, Vol. 40, pp. D559–D564, 2012.

[30] J. Greer, "Model for haptoglobin heavy chain based upon structural homology," *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 77, pp. 3393–3397, 1980.

[31] M. Gribskov, A. D. McLachlan, and D. Eisenberg, "Profile analysis: Detection of distantly related proteins," *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 84, pp. 4355–4358, 1987.

[32] D. Gront, S. Kmiecik, and A. Kolinski, "Backbone building from quadrilaterals: A fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates," *Journal of Computational Chemistry*, Vol. 28, pp. 1593–1597, 2007.

[33] D. Gusfield, "Efficient methods for multiple sequence alignment with guaranteed error bounds," *Bulletin of Mathematical Biology*, Vol. 55, pp. 141–154, 1993.

[34] D. Gusfield, *Algorithms on strings, trees, and sequences: computer science and computational biology.* Cambridge University Press, New York, 1997.

[35] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proceedings of the National Academy of Sciences of the United States of America*, pp. 10915–10919, 1992.

[36] D. A. Hinds and M. Levitt, "A lattice model for protein structure prediction at low resolution," *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 89, pp. 2536–2540, 1992.

[37] L. Holm and C. Sander, "Database algorithm for generating protein backbone and side-chain coordinates from a c alpha trace application to model building and detection of coordinate errors," *Journal of Molecular Biology*, Vol. 21, No. 1, pp. 183–194, 1991.

[38] L. Holm and C. Sander, "Parser for protein folding units," *Proteins*, Vol. 19, pp. 256–268, 1994.

[39] L. Holm and C. Sander, "Mapping the protein universe," *Science*, Vol. 273, pp. 595–603, 1996.

[40] R. Hughey and A. Krogh, "SAM: sequence alignment and modeling software system," tech. rep., University of California at Santa Cruz Santa Cruz, CA, USA, 1995.

[41] D. T. Jones, "GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences," *Journal of Molecular Biology*, Vol. 287, pp. 797–815, 1999.

[42] W. Kabsch, "A solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A*, Vol. 32, pp. 922–923, 1976.

[43] W. Kabsch, "A discussion of the solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A*, Vol. 34, pp. 827–828, 1978.

[44] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, Vol. 22, pp. 2577–2637, 1983.

[45] S. Kaczanowski and P. Zielenkiewicz, "Why similar protein sequences encode similar three-dimensional structures?," *Theoretical Chemistry Accounts*, Vol. 125, pp. 643–650, 2010.

[46] M. Kallberg, H. Wang, S. Wang, J. Peng, Z. Wang, H. Lu, and J. Xu, "Template-based protein structure modeling using the raptorx web server," *Nature Protocols*, Vol. 7, pp. 1511–1522, 2012.

[47] R. Kazmierkiewicz, A. Liwo, and H. A. Scheraga, "Energy-based reconstruction of a protein backbone from its $\alpha$-carbon trace by a Monte-Carlo method," *Journal of Computational Chemistry*, Vol. 23, pp. 715–723, 2002.

[48] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, Vol. 220, pp. 671–680, 1983.

[49] A. Kolinski, D. Gront, P. Pokarowski, and J. Skolnick, "A simple lattice model that exhibits a protein-like cooperative all-or-none folding transition," *Biopolymers*, Vol. 69, pp. 399–405, 2003.

[50] A. Krogh, M. Brown, I. S. Mian, K. Sjolander, and D. Haussler, "Hidden markov models in computational biology. Applications to protein modeling," *Journal of Molecular Biology*, Vol. 235, pp. 290–308, 1994.

[51] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," *Proceedings of the 18th International Conference*, San Francisco, CA, USA, pp. 282–289, 2001.

[52] K. F. Lau and K. A. Dill, "A lattice statistical mechanics model of the conformational and sequence spaces of proteins," *Macromolecules*, Vol. 22, pp. 3986–3997, 1989.

[53] A. Leaver-fay, M. Tyka, S. M. Lewis, F. Lange, J. Thompson, R. Jacak, K. Kaufman, P. D. Renfrew, C. A. Smith, W. Sheffler, I. W. Davis, S. Cooper, A. Treuille, D. J. M, F. Richter, Y. en Andrew Ban, S. J. Fleishman, E. Corn, D. E. Kim, S. Lyskov, M. Berrondo, S. Mentzer, K. Z. Popovic, K. J. J. Havranek, J. Karanicolas, R. Das, J. Meiler, T. Kortemme, J. J. Gray, B. Kuhlman, D. Baker, and P. Bradley, "ROSETTA3: An object-oriented software suite for the simulation and design of macromolecules," *Methods in Enzymology*, Vol. 487, pp. 545–547, 2011.

[54] C. Levinthal, "Are there pathways for protein folding?," *Journal de Chimie Physique*, Vol. 65, pp. 44–45, 1968.

[55] M. Levitt and A. Warshel, "Computer simulation of protein folding," *Nature*, Vol. 253, pp. 694–698, 1975.

[56] H. Li, R. Helling, C. Tang, and N. Wingreen, "Emergence of preferred structures in a simple model of protein folding," *Science*, Vol. 273, pp. 666–669, 1996.

[57] H. Li, C. Tang, and N. Wingreen, "Are protein folds atypical?," *Proceedings of the National Academy of Sciences of the United States of America*, pp. 4987–4990, 1998.

[58] H. Li, C. Tang, and N. Wingreen, "Designability of protein structures: a lattice-model study using the Miyazawa-Jernigan matrix," *Proteins*, Vol. 49, pp. 403–412, 2002.

[59] J. Li, X. Deng, J. Eickholt, and J. Cheng, "Designing and benchmarking the MULTICOM protein structure prediction system," *BMC Structural Biology*, Vol. 13, p. 2, 2013.

[60] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical Programming*, Vol. 45, pp. 503–528, 1989.

[61] R. Luthy, J. U. Bowie, and D. Eisenberg, "Assessment of protein models with three-dimensional profiles," *Nature*, Vol. 356, pp. 83–85, 1992.

[62] J. Ma, J. Peng, S. Wang, and J. Xu, "A conditional neural fields model for protein threading," *Bioinformatics*, Vol. 28, pp. i59–i66, 2012.

[63] J. T. MacDonald, K. Maksimiak, M. I. Sadowski, and W. R. Taylor, "De novo backbone scaffolds for protein design," *Proteins*, Vol. 75, pp. 1311–1325, 2010.

[64] A. D. MacKerell, D. Bashford, Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wirkiewicz-Kuczera, D. Yin, and M. Karplus, "All-atom empirical potential for molecular modeling and dynamics studies of proteins," *Journal of Physical Chemistry B*, Vol. 102, pp. 3586–3616, 1998.

[65] M. Madera, "PRC, the profile comparer." http://supfam.org/PRC/.

[66] J. Maupetit, R. Gautier, and P. Tuffery, "SABBAC: online structural alphabet-based protein backbone reconstruction from alpha-carbon trace," *Nucleic Acids Research*, Vol. 34, pp. W147–W151, 2006.

[67] K. M. Merz, D. Ringe, and C. H. Reynolds, *Drug Design: Structure- and Ligand-Based Approaches*. Cambridge University Press, 2010.

[68] M. Milik, A. Kolinski, and J. Skolnick, "Algorithm for rapid reconstruction of protein backbone from alpha carbon coordinates," *Journal of Computational Chemistry*, Vol. 18, pp. 80–85, 1997.

[69] B. L. Moore, L. A. Kelley, J. W. Murray, and J. T. MacDonald, "High-quality

protein backbone reconstruction from alpha carbons using gaussian mixture models," *Journal of Computational Chemistry*, Vol. 34, pp. 1881–1889, 2013.

[70] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: A structural classification of proteins database for the investigation of sequences and structures," *Journal of Molecular Biology*, Vol. 247, pp. 536–540, 1995.

[71] S. Needleman and C. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, Vol. 48, pp. 443–453, 1970.

[72] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton, "CATH - a hierarchic classification of protein domain structures," *Structure*, Vol. 5, pp. 1093–1109, 1997.

[73] B. H. Park and M. Levitt, "The complexity and accuracy of discrete state models of protein structure," *Journal of Molecular Biology*, Vol. 249, pp. 493–507, 1995.

[74] J. Peng, L. Bo, and J. Xu, "Conditional neural fields," *Advances in Neural Information Processing Systems 22*, Vol. 22, Vancouver, British Columbia, Canada, pp. 1419–1427, 2009.

[75] J. Peng and J. Xu, "Boosting protein threading accuracy," *Proceedings of the 13th Annual International Conference on Research in Computational Molecular Biology (RECOMB), Lecture Notes in Computer Science*, Tucson, AZ, USA, pp. 31–45, 2009.

[76] J. Peng and J. Xu, "Low-homology protein threading," *Bioinformatics*, Vol. 26, pp. i294–i300, 2010.

[77] J. Peng and J. Xu, "A multiple-template approach to protein threading," *Proteins*, Vol. 79, pp. 1930–1939, 2011.

[78] J. Peng and J. Xu, "RaptorX: exploiting structure information for protein alignment by statistical inference," *Proteins*, Vol. 79, pp. 161–171, 2011.

[79] S. Pietrokovski, "Searching databases of conserved sequence regions by aligning protein multiple-alignments," *Nucleic Acids Research*, Vol. 24, pp. 3836–3845, 1996.

[80] E. O. Purisima and H. A. Scheraga, "Conversion from a virtual-bond chain to a complete polypeptide backbone chain," *Biopolymers*, Vol. 23, pp. 1207–1224, 1984.

[81] S. Raman, R. Vernon, J. Thompson, M. Tyka, R. Sadreyev, J. Pei, D. Kim, E. Kellogg, F. DiMaio, O. Lange, L. Kinch, W. Sheffler, B.-H. Kim, R. Das, N. V. Grishin, and D. Baker, "Structure prediction for CASP8 with all-atom refinement using Rosetta," *Proteins*, Vol. 77, pp. 89–99, 2009.

[82] R. J. Read and G. Chavali, "Assessment of CASP7 predictions in the high accuracy template-based modeling category," *Proteins*, Vol. S8, pp. 27–37, 2007.

[83] M. Remmert, A. Bieger, A. Hauser, and J. Soding, "HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment," *Nature Methods*, Vol. 9, pp. 173–175, 2011.

[84] J. S. Richardson, "The anatomy and taxonomy of protein structure," *Advances in Protein Chemistry*, Vol. 34, pp. 167–339, 1981.

[85] C. Rohl, C. Strauss, K. Misura, and D. Baker, "Protein structure prediction using rosetta," *Methods in Enzymology*, Vol. 383, pp. 66–93, 2004.

[86] P. Rotkiewicz and J. Skolnick, "Fast procedure for reconstruction of full-atom protein models from reduced representations," *Journal of Computational Chemistry*, Vol. 29, pp. 1460–1465, 2008.

[87] A. Roy, A. Kucukural, and Y. Zhang, "I-TASSER: a unified platform for automated protein structure and function prediction," *Nature Protocols*, Vol. 5, pp. 725–738, 2010.

[88] R. Sadreyev and N. Grishin, "COMPASS: A tool for comparison of multiple protein alignments with assessment of statistical significance," *Journal of Molecular Biology*, Vol. 326, pp. 317–336, 2003.

[89] A. Sali and T. L. Blundell, "Comparative protein modelling by satisfaction of spatial restraints," *Journal of Molecular Biology*, Vol. 234, pp. 779–815, 1993.

[90] A. Sali, L. Potterton, F. Yuan, H. van Vlijmen, and M. Karplus, "Evaluation of comparative protein modeling by modeller," *Proteins*, Vol. 23, pp. 318–326, 1995.

[91] F. Sanger, E. O. P. Thompson, and R. Kitai, "The amide groups of insulin," *Biochemical Journal*, Vol. 59, pp. 509–518, 1955.

[92] S. Santini, G. Wei, N. Mousseau, and P. Derreumaux, "Exploring the folding pathways of proteins through energy landscape sampling: Application to alzheimer's $\beta$-amyloid peptide," *Internet Electronic Journal of Molecular Design*, Vol. 2, No. 9, pp. 564–577, 2003.

[93] M. J. Sippl, "Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins," *Journal of Molecular Biology*, Vol. 213, pp. 859–883, 1990.

[94] J. Soding, A. Biegert, and A. Lupas, "The HHpred interactive server for protein homology detection and structure prediction," *Nucleic Acids Research*, Vol. 33, pp. W244–W248, 2005.

[95] J. Soding, "Protein homology detection by HMM-HMM comparison," *Bioinformatics*, Vol. 21, pp. 951–960, 2005.

[96] Y. Song, F. DiMaio, R. Wang, D. Kim, C. Miles, T. Brunette, J. Thompson, and D. Baker, "High-resolution comparative modeling with RosettaCM," *Structure*, Vol. 21, pp. 1735–1742, 2013.

[97] J. tao Guo, D. Xu, D. Kim, and Y. Xu, "Improving the performance of DomainParser for structural domain partition using neural network," *Nucleic Acids Research*, Vol. 31, pp. 944–952, 2003.

[98] P. Tuffery and P. Derreumaux, "Dependency between consecutive local conformations helps assemble protein structures from secondary structures using Go potential and greedy algorithm," *Proteins*, Vol. 61, p. 732740, 2005.

[99] P. Tuffery, F. Guyon, and P. Derreumaux, "Improved greedy algorithm for protein structure reconstruction," *Journal of Computational Chemistry*, Vol. 26, pp. 506–513, 2005.

[100] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, Vol. 13, pp. 260–269, 1967.

[101] D. Vitkup, E. Melamud, J. Moult, and C. Sander, "Completeness in structural genomics," *Nature Structural & Molecular Biology*, Vol. 8, pp. 559–566, 2001.

[102] Z. Wang, J. Eickholt, and J. Cheng, "MULTICOM: a multi-level combination approach to protein structure prediction and its assessments in CASP8," *Bioinformatics*, Vol. 26, pp. 882–888, 2010.

[103] Z. Wang, J. Eickholt, and J. Cheng, "APOLLO: a quality assessment service for single and multiple protein models," *Bioinformatics*, Vol. 27, pp. 1715–1716, 2011.

[104] Z. Wang, A. N. Tegge, and J. Cheng, "Evaluating the absolute quality of a single protein model using structural features and support vector machines," *Proteins*, Vol. 75, pp. 638–647, 2009.

[105] Z.-X. Wang, "How many fold types of protein are there in nature?," *Proteins*, Vol. 26, pp. 186–191, 1996.

[106] D. B. Wetlaufer, "Nucleation, rapid folding, and globular intrachain regions in proteins," *Proceedings of the National Academy of Sciences of the United States of America*, pp. 697–701, 1973.

[107] S. Wu and Y. Zhang, "LOMETS: a local meta-threading-server for protein structure prediction," *Nucleic Acids Research*, Vol. 35, pp. 3375–3382, 2007.

[108] D. Xu and Y. Zhang, "Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization," *Biophysical Journal*, Vol. 101, pp. 2525–2534, 2011.

[109] J. Xu, M. Li, D. Kim, and Y. Xu, "RAPTOR: optimal protein threading by linear programming," *Journal of Bioinformatics and Computational Biology*, Vol. 1, pp. 95–117, 2003.

[110] Y. Xu and D. Xu, "Protein threading using PROSPECT: design and evaluation," *Proteins*, Vol. 40, pp. 343–354, 2000.

[111] J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, and Y. Zhang, "The I-TASSER Suite:

Protein structure and function prediction," *Nature Methods*, Vol. 12, pp. 7–8, 2015.

[112] Y. Yang, E. Faraggi, H. Zhao, and Y. Zhou, "Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates," *Bioinformatics*, Vol. 27, pp. 2076–2082, 2011.

[113] A. Zemla, C. Venclovas, J. Moult, and K. Fidelis, "Processing and analysis of CASP3 protein structure predictions," *Proteins*, Vol. 3, pp. 22–29, 1999.

[114] C. Zhang and C. DeLisi, "Estimating the total number of protein folds," *Journal of Molecular Biology*, Vol. 284, pp. 1301–1305, 1998.

[115] Y. Zhang, "Automated assessment of protein structure prediction in CASP10." http://zhanglab.ccmb.med.umich.edu/casp10/.

[116] Y. Zhang, "Automated assessment of protein structure prediction in CASP11." http://zhanglab.ccmb.med.umich.edu/casp11/.

[117] Y. Zhang, "Automated assessment of protein structure prediction in CASP8." http://zhanglab.ccmb.med.umich.edu/casp8/.

[118] Y. Zhang, "Automated assessment of protein structure prediction in CASP9." http://zhanglab.ccmb.med.umich.edu/casp9/.

[119] Y. Zhang, "I-TASSER server for protein 3D structure prediction," *BMC Bioinformatics*, Vol. 9, p. 40, 2008.

[120] Y. Zhang and J. Skolnick, "Scoring function for automated assessment of protein structure template quality," *Proteins*, Vol. 57, pp. 702–710, 2004.

[121] Y. Zhang and J. Skolnick, "SPICKER: a clustering approach to identify near-native protein folds," *Journal of Computational Chemistry*, Vol. 25, pp. 865–871, 2004.

[122] Y. Zhang and J. Skolnick, "TM-align: a protein structure alignment algorithm based on the TM-score," *Nucleic Acids Research*, Vol. 33, pp. 2302–2309, 2005.

[123] H. Zhou and Y. Zhou, "SPEM: improving multiple sequence alignment with sequence profiles and predicted secondary structures," *Bioinformatics*, Vol. 21, pp. 3615–3621, 2005.

[124] Y. Zhou and M. E. Johnson, "Comparative molecular modeling analysis of-5-amidinoindole and benzamidine binding to thrombin and trypsin: specific H-bond formation contributes to high 5-amidinoindole potency and selectivity for thrombin and factor Xa," *Journal of Molecular Recognition*, Vol. 12, pp. 235–241, 1999.