Chapter 1
Ab Initio Protein Structure Prediction

Jooyoung Lee, Peter L. Freddolino and Yang Zhang

Abstract Predicting a protein's structure from its amino acid sequence remains an unsolved problem after several decades of efforts. If the query protein has a homolog of known structure, the task is relatively easy and high-resolution models can often be built by copying and refining the framework of the solved structure. However, a template-based modeling procedure does not help answer the questions of how and why a protein adopts its specific structure. In particular, if structural homologs do not exist, or exist but cannot be identified, models have to be constructed from scratch. This procedure, called ab initio modeling, is essential for a complete solution to the protein structure prediction problem; it can also help us understand the physicochemical principle of how proteins fold in nature. Currently, the accuracy of ab initio modeling is low and the success is generally limited to small proteins (<120 residues). With the help of co-evolution based contact map predictions, success in folding larger-size proteins was recently witnessed in blind testing experiments. In this chapter, we give a review on the field of ab initio structure modeling. Our focus will be on three key components of the modeling algorithms: energy function design, conformational search, and model selection. Progress and advances of several representative algorithms will be discussed.

Keyword Protein structure prediction · Ab initio folding · Contact prediction · Force field

J. Lee
School of Computational Sciences,
Korea Institute for Advanced Study, Seoul 130-722, Korea

P.L. Freddolino · Y. Zhang
Department of Biological Chemistry, University of Michigan,
Ann Arbor, MI 48109, USA

P.L. Freddolino · Y. Zhang (✉)
Department of Computational Medicine and Bioinformatics,
University of Michigan, Ann Arbor, MI 48109, USA
e-mail: zhng@umich.edu

D.J. Rigden (ed.), *From Protein Structure to Function with Bioinformatics*,
DOI 10.1007/978-94-024-1069-3_1

1.1 Introduction

With the success of an expanding array of genome sequencing projects, the number of known protein sequences has been increasing exponentially. However, the sequences on their own cannot tell what each protein does in cell. Although protein structure information is essential for understanding the function, the speed of protein structure determination lags far behind the increase of sequences, due to the technical difficulties and laborious nature of structural biology experiments. By the end of 2015, about 90 million protein sequences were deposited in the UniProtKB database (Bairoch et al. 2005) (http://www.uniprot.org/). However, the corresponding number of protein structures in the Protein Data Bank (PDB) (Berman et al. 2000) (http://www.rcsb.org) is only about 100,000. The gap is rapidly widening as indicated in Fig. 1.1, where the ratio of sequences over structure increased from less than 1 magnitude to around 3 magnitudes in the last two decades. Thus, developing efficient computer-based algorithms that can generate high-resolution 3D structure predictions becomes probably the only avenue to fill up the gap.

Depending on whether similar proteins have been experimentally solved, protein structure prediction methods can be grouped into two categories. First, if proteins of a similar structure are identified from the PDB library, the target model can be constructed by copying and refining framework of the solved proteins (templates). The procedure is called “template-based modeling (TBM)” (Sali and Blundell 1993; Karplus et al. 1998; Jones 1999; Skolnick et al. 2004; Soding 2005; Wu and Zhang 2008a; b; Yang et al. 2011), and will be discussed in the subsequent chapters. Although high-resolution models can often be generated by TBM, the procedure cannot help us understand the physicochemical principle of protein folding.

If protein templates are not available, we have to build the 3D models from scratch. This procedure has been given different names, e.g. ab initio modeling (Klepeis et al. 2005; Liwo et al. 2005; Wu et al. 2007; Taylor et al. 2008; Xu and

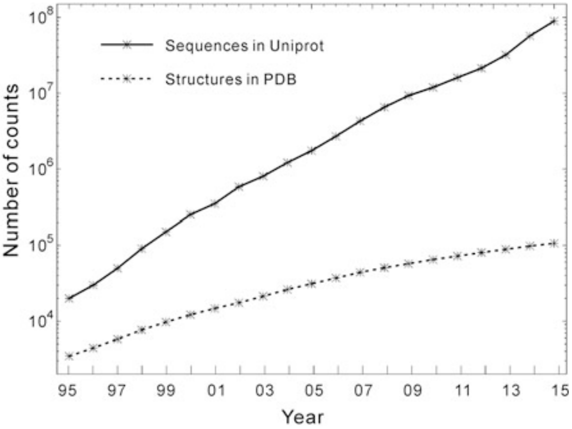

Fig. 1.1 The numbers of available protein sequences and solved protein structures are shown for the last 20 years. The ratio of sequences over structures increases from less than 10 in 1995 to three orders of magnitude in 2015. Data are taken from UniProtKB (Bairoch et al. 2005) and PDB (Berman et al. 2000) databases

Zhang 2012); de novo modeling (Bradley et al. 2005a, b), physics-based modeling (Oldziej et al. 2005), or free modeling (Jauch et al. 2007; Kinch et al. 2015). In this chapter, the term ab initio modeling is uniformly used to avoid confusion. Unlike the template-based modeling, a successful ab initio modeling procedure could help address the basic questions on how and why a protein adopts the specific structure out of many possibilities.

Typically, ab initio modeling conducts a conformational search under the guidance of a designed energy function. This procedure usually generates a number of possible conformations (also called structure decoys), and final models are selected from them. Therefore, a successful ab initio modeling depends on three factors: (1) an accurate energy function with which the native structure of a protein corresponds to the most thermodynamically stable state, compared to all possible decoy structures; (2) an efficient search method which can quickly identify the low-energy states through conformational search; (3) a strategy that can select near-native models from a pool of decoy structures.

This chapter gives a review on the most recent progress in ab initio protein structure prediction. This review is neither sufficiently complete to include all available ab initio methods nor sufficiently in depth to provide all backgrounds/motivations behind them. For a quantitative comparison of the state-of-the-art ab initio modeling methods, readers are suggested to read the assessment articles on template-free modeling in the recent CASP experiments (Kinch et al. 2011; Tai et al. 2014; Kinch et al. 2015). The rest of the chapter is organized as follows. First, the three major issues of ab initio modeling, i.e. energy function design, conformational search engine and model selection scheme, will be described in detail. New and promising ideas to improve the efficiency and effectiveness of the prediction are then discussed. Finally, current progress and challenges of ab initio modeling are summarized.

1.2 Energy Functions

In this section, we discuss energy functions used for ab initio modeling. It should be noted that in many cases energy functions and the search procedures are intricately coupled to each other, and as soon as they are decoupled, the modeling procedure often loses its power and/or validity. We classify the energy functions into two groups: (a) physics-based energy functions and (b) knowledge-based energy functions, depending on whether they make use of statistics from the existing protein 3D structures in the PDB. A few promising methods from each group are selected to discuss according to their uniqueness and modeling accuracy. A list of ab initio modeling methods is provided in Table 1.1 along with their properties about energy functions, conformational search algorithms, model selection methods and typical running times.

Table 1.1 A list of ab initio modeling algorithms reviewed in this chapter is shown along with their energy functions, conformational search methods, model selection schemes and typical CPU time per target

Algorithm and server address	Force-field type	Search method	Model selection	Time cost per CPU
AMBER/CHARMM/OPLS (Brooks et al. 1983; Weiner et al. 1984; Jorgensen and Tirado-Rives 1988; Duan and Kollman 1998; Zagrovic et al. 2002)	Physics-based	Molecular dynamics (MD)	Lowest energy	Years
UNRES (Liwo et al. 1999; Liwo et al. 2005, Oldziej et al. 2005)	Physics-based	Conformational space annealing (CSA)	Clustering/free-energy	Hours
ASTRO-FOLD (Klepeis et al. Klepeis and Floudas 2003; Klepeis et al. 2005)	Physics-based	αBB/CSA/MD	Lowest energy	Months
ROSETTA (Simons et al. 1997, Das et al. 2007) http://www.robetta.org	Physics- and knowledge-based	Monte Carlo (MC)	Clustering/free-energy	Days
TASSER/Chunk-TASSER (Zhang et al. 2004, Zhou and Skolnick 2007) http://cssb.biology.gatech.edu/skolnick/webservice/MetaTASSER	Knowledge-based	MC	Clustering/free-energy	Hours
I-TASSER (Roy et al. 2010; Yang et al. 2015a, b) http://zhanglab.ccmb.med.umich.edu/I-TASSER	Knowledge-based	MC	Clustering/free-energy	Hours
QUARK (Xu and Zhang 2012) http://zhanglab.ccmb.med.umich.edu/QUARK	Physics- and knowledge-based	MC	Clustering/free-energy	Hours

1.2.1 Physics-Based Energy Functions

In a strictly-defined physics-based ab initio method, interactions between atoms should be based on quantum mechanics and the Coulomb potential with only a few fundamental parameters such as the electron charge and the Planck constant; all atoms should be described by their atom types where only the number of electrons is relevant (Hagler et al. 1974; Weiner et al. 1984). However, there have not been serious attempts to start from quantum mechanics to predict structures of (even small) proteins, simply because the computational resources required for such calculations are far beyond what is available now. Without quantum mechanical treatments, a practical starting point for ab initio protein modeling is to use a force field treating atoms as point particles interacting through a defined potential form, with the parameters governing inter-atomic interactions obtained through the comparisons of the force field with a combination of experimental and quantum mechanical data (Hagler et al. 1974; Weiner et al. 1984). Well-known examples of such all-atom physics-based force fields include AMBER (Weiner et al. 1984; Cornell et al. 1995; Duan and Kollman 1998), CHARMM (Brooks et al. 1983; Neria et al. 1996; MacKerell et al. 1998), OPLS (Jorgensen and Tirado-Rives 1988; Jorgensen et al. 1996), and GROMOS96 (van Gunsteren et al. 1996). These potentials contain terms associated with bond lengths, angles, torsion angles, van der Waals, and electrostatics interactions. The major difference between them lies in the selection of atom types and the interaction parameters.

Coupling Physics-Based Potentials With Molecular Dynamics Simulations For the study of protein folding, these classical force fields were often coupled with molecular dynamics (MD) simulations. The obvious appeal of such an approach is that the prediction of protein folding via MD simulations provides not only information on the folded structure, but also the folding process itself, which must be fully simulated *en route*. However, the results, from the viewpoint of protein structure prediction, have until quite recently been disappointing. (See Chap. 12 for the use of MD in elucidation of protein function from known structures).

The first milestone in MD-based ab initio protein folding was probably the 1997 work of Duan and Kollman, who simulated the villin headpiece subdomain (a 36 amino acid protein) in explicit solvent for 6 months on parallel supercomputers. Although the authors did not fold the protein with high resolution, the best of their final models was within 4.5 Å RMS deviation of the native state (Duan and Kollman 1998). With Folding@Home, a worldwide-distributed computer system, this small protein was later folded by Pande and coworkers (Zagrovic et al. 2002) to 1.7 Å with a total simulation time of 300 μs or approximately 1000 CPU years. The years since then have seen an increasing number of successful ab initio folding simulations using molecular dynamics (Chowdhury et al. 2003; Ensign et al. 2007; Lei et al. 2007; Freddolino and Schulten 2009), although all have required heroic amounts of computing time either through supercomputing centers or distributed community projects. During the same period, ab initio folding simulations also revealed secondary structure biases in several physics-based force fields that hampered their general applicability to different folds (Best et al. 2008; Freddolino et al. 2008, Best and Hummer 2009; Freddolino and Schulten 2009; Lindorff-Larsen et al. 2012).

A flurry of force field development efforts spurred by these shortcomings have resulted in a new generation of parameter sets that are able to reliably fold a wide variety of protein structures (Lindorff-Larsen et al. 2010; Mittal and Best 2010; Piana et al. 2011; Lindorff-Larsen et al. 2012), leaving simulation timescales as the main barrier for MD ab initio folding simulations. Even this barrier has begun to crumble in the face of recent advances in computing hardware. The special purpose Anton machine, designed by Shaw and co-workers specifically for extreme-performance molecular dynamics simulations, has allowed complete, reversible folding simulations of proteins up to ~100 residues long in explicit solvent (Lindorff-Larsen et al. 2011; Piana et al. 2012, Piana et al. 2013a, b; Piana et al. 2014). Following a separate path, the use of GPU acceleration in most major molecular dynamics packages has enabled ab initio folding simulations on commodity hardware to reach performances of 1 microsecond per GPU-day for small proteins with implicit solvent (Nguyen et al. 2014), and allowed successful folding of 16 out of 17 test proteins (10–100 residues). Despite these remarkable efforts, the all-atom physics-based MD simulation is far from being routinely used for structure prediction of typical-size proteins (~100–300 residues), and it is instead primarily used to provide additional information on folding pathways or equilibriums.

Application to Atomic-Level Structure Refinement Another protein structure niche where physics-based MD simulation can contribute is structure refinement. Starting from low-resolution protein models, the goal is to draw the structure closer to the native by refining the local side-chain and peptide-backbone packing. When the starting models are not very far away from the native, the intended conformational change is relatively small and the simulation time would be much shorter than that required in ab initio folding. One of the early MD-based protein structure refinements was for the GCN4 leucine zipper (33-residue dimer) (Nilges and Brunger 1991; Vieth et al. 1994), where a low-resolution coiled-coil dimer structure (2–3 Å RMS deviation from native) was first assembled by Monte Carlo (MC) simulation before the subsequent MD refinement. With the help of helical dihedral-angle restraints, Skolnick and coworkers (Vieth et al. 1994) were able to generate a refined structure of GCN4 with below 1 Å backbone RMSD using CHARMM (Brooks et al. 1983) with the TIP3P water model (Jorgensen et al. 1983).

Later, using AMBER 5.0 (Case et al. 1997) and the TIP3P water model (Jorgensen et al. 1983; Lee et al. 2001) attempted to refine 360 low-resolution models generated by ROSETTA (Simons et al. 1997) for 12 small proteins (<75 residues); but they concluded that no systematic structure improvement was achieved (Lee et al. 2001). Fan and Mark (Fan and Mark 2004) tried to refine 60 ROSETTA models for 11 small proteins (<85 residues) using GROMACS 3.0 (Lindahl et al. 2001) with explicit water (Berendsen et al. 1981) and they reported that 11/60 models were improved by 10% in RMSD, but 18/60 got worse in RMSD after refinement. Similarly, Chen and Brooks (Chen and Brooks 2007) used CHARMM22 (MacKerell et al. 1998) to refine five CASP6 CM targets (70–144 residues). In four cases, refinements with up to 1 Å RMSD reduction were achieved. In this work, an implicit solvent model based on the generalized Born (GB) approximation (Im et al. 2003) was used, which significantly speeded up the computation. In addition, the spatial restraints extracted from the initial models were used to guide the refinement procedure (Chen and Brooks 2007).

More recently, Zhang et al. (2011) proposed to use analogous fragments from known structures to bias the physics-based force field and improve structure refinement. In this work, the initial structure model was split into segments of 2–4 secondary structure elements, which are structurally matched through the PDB library by TM-align (Zhang and Skolnick 2005a, b) to identify analogous fragments. The distance map from the analogous fragments is then used as restraints to reshape the MD energy funnel. The protocol was tested on 181 benchmarking and 26 CASP targets. It was found that structure models of correct folds with TM-score >0.5 can be often pulled closer to native with higher GDT-HA score, but improvement for the models of incorrect folds (TM-score <0.5) were much less pronounced. The previous experiments have shown that the physics-based force field can often recognize the native but lacks middle-range correlation to the RMSD in the high RMSD region (Bradley et al. 2005a, b; Jagielska et al. 2008), which leads to a golfcourse like energy landscapes with a deep basin around the native that cannot help for refining low-resolution models. The data by Zhang et al. seemed to indicate that template-based fragmental distance maps reshaped the MD energy landscape from golfcourse-like to funnel-like in the successfully refined targets with an approximate radius of TM-score ∼0.5. Similarly, Feig and coworkers used the Cα maps collected from initial structure models to guide the MD based structure refinement simulations (Mirjalili and Feig 2013). In the recent CASP experiment (Feig and Mirjalili 2015), the approach showed a small but consistent improvement on the structural models, with average RMSD improvement by 0.13 Å for the first submitted models and 0.52 Å for the best in top five models.

Molecular Mechanics Approaches A noteworthy observation was made by Summa and Levitt (2007) who exploited various molecular mechanics (MM) potentials (AMBER99 (Wang et al. 2000; Sorin and Pande 2005), OPLS-AA (Kaminski et al. 2001), GROMOS96 (van Gunsteren et al. 1996), and ENCAD (Levitt et al. 1995)) to refine 75 proteins by *in vacuo* energy minimization. They found that a knowledge-based atomic contact potential outperformed the MM potentials by moving almost all test proteins closer to their native states, while the MM potentials, except for AMBER99, essentially drove decoys further away from their native structures. The vacuum simulation without solvation may be partly the reason for the failure of the MM potentials. This observation demonstrates the possibility of combining knowledge-based potentials with physics-based force fields for more successful protein structure refinement.

While the physics-based potential driven by MD simulations was not particularly successful in structure prediction due to the immense computational cost of MD simulations on the timescales of folding processes, fast search methods (such as Monte Carlo simulations and genetic algorithms) combined with similar physics-based potentials have been shown to be promising in both structure prediction and structure refinement. One example is the effort by Scheraga and coworkers (Liwo et al. 1999; Liwo et al. 2005; Oldziej et al. 2005) who have been developing a physics-based protein structure prediction method solely based on the thermodynamic hypothesis. The method combines the coarse grained potential UNRES with a global optimization algorithm called conformational space annealing (Oldziej et al. 2005). In UNRES, each residue is described by two interacting off-lattice united atoms, C_α and the side-chain center. This effectively reduces the number of atoms by 10, enabling one to handle polypeptide chains of larger than 100 residues. The resulting prediction time for small proteins can be then reduced to 2–10 h. The UNRES energy function (Liwo et al. 1993) consists of pair-wise interactions between all interacting parties and additional terms such as local energy and correlation energy. The low energy UNRES models are then converted into all-atom representations based on ECEPP/3 (Nemethy et al. 1992). Although many of the parameters of the energy function are calculated by quantum-mechanical methods, some of them are derived from the distributions and correlation functions calculated from the PDB library. For this reason, one might question classifying it as a truly physics-based approach. Nevertheless, this method is one of the most faithful ab initio methods available (in terms of the application of a thorough global optimization to a physics-based energy function) and has been systematically applied to many CASP targets since 1998. The most notable prediction success by this approach was for T061 from CASP3, for which a model of 4.2 Å RMSD for a 95-residue α-helical protein was generated with an accuracy gap between it and the models of others. It was shown in a clear-cut fashion that the ab initio method can sometime provide better models for the targets where the template-based methods fail. In CASP6, a structure genomics target of TM0487 (T0230, 102 residues) was folded to 7.3 Å by this approach. However, it seems that the scarcity and the best-but-still-low accuracy of such models by a pure ab initio modeling failed to draw much attention from the protein science community, where accurate protein models are in great demand.

Another example of the physics-based modeling approaches is the multi-stage hierarchical algorithm ASTRO-FOLD, proposed by Floudas and coworkers (Klepeis and Floudas 2003; Klepeis et al. 2005). First, secondary structure elements (α-helices and β-strands) are predicted by calculating a free energy function of overlapping oligopeptides (typically pentapeptides) and all possible contacts between 2 hydrophobic residues. The free energy terms used include entropic, cavity formation, polarization, and ionization contributions for each oligopeptide. After transforming the calculated secondary structure propensity into the upper and lower bounds of backbone dihedral angles and the distant restraints between Cα atoms, the final tertiary structure of the full length protein is modeled by globally minimizing the energy using the ECEPP/3 all-atom force field. This approach was successfully applied to an α-helical protein of 102 residues in a double-blind fashion (but not in an open community-wide way for relative performance comparison to other methods). The RMSD of the predicted model was 4.94 Å away from the experimental structure. The global optimization method used in this approach is a combination of α branch and bound (αBB), conformational space annealing, and MD simulations (Klepeis and Floudas 2003; Klepeis et al. 2005). The relative performance of this method on larger number of proteins is yet to be examined.

Taylor and coworkers (Taylor et al. 2008) proposed a novel approach which constructs protein structural models by enumerating possible topologies in a coarse-grained form, given the secondary structure assignments and the physical connection constraints of the secondary structure elements. The top scoring conformations, based on the structural compactness and element exposure, are then selected for further refinement (Jonassen et al. 2006). The authors successfully folded a set of five αβ sandwich proteins with length up to 150 residues with the first model having 4–6 Å RMS deviation from the known experimental structure. Again, although appealing in methodology, the performance of the approach in open blind experiments and on proteins of various fold-types is yet to be seen.

1.2.2 Knowledge-Based Energy Function Combined with Fragments

The term knowledge-based potential refers to a set of empirical energy terms derived from the statistics and regularities of the solved structures in deposited PDB. Such potentials can be divided into two types as described by Skolnick (Skolnick 2006). The first covers generic and sequence-independent terms such as the hydrogen bonding and the local backbone stiffness of a polypeptide chain (Zhang et al. 2003). The second contains amino-acid or protein-sequence dependent terms, e.g. pair-wise residue contact potential (Skolnick et al. 1997), distance dependent atomic contact potential (Samudrala and Moult 1998; Lu and Skolnick 2001; Zhou and Zhou 2002; Shen and Sali2006; Zhang and Zhang 2010), and secondary structure propensities (Zhang et al. 2003, Zhang and Skolnick 2005a, b; Zhang et al. 2006).

Although most knowledge-based force fields contain secondary structure propensities, it may be that local protein structures are rather difficult to reproduce in the reduced modeling. That is, in nature a variety of protein sequences prefer either helical or extended structures depending on the subtle differences in their local and global sequence environment, yet we have not yet developed force fields that can reproduce this subtlety properly. One way to circumvent this problem is to use secondary structure fragments, obtained from sequence or profile alignments, directly into 3D model assembly. One additional advantage of the fragment-based approach is that the use of excised secondary structure fragment can significantly reduce the entropy of the conformational search.

Here, we introduce several representative methods utilizing knowledge-based energy functions, which have proven to be the most successful in ab initio protein structure prediction methods in recent community competitions (Simons et al. 1997; Zhang and Skolnick 2004a, b; Xu and Zhang 2012).

ROSETTA One of the best-known ideas for *ab initio*, pioneered by Bowie and Eisenberg, involves generating protein models by assembling small fragments (mainly 9-mers) taken from the PDB library (Bowie and Eisenberg 1994). Based on a similar idea, Baker and coworkers developed ROSETTA (Simons et al. 1997), which has been very successful for the free modeling (FM) targets in the CASP experiments, and which has greatly boosted the popularity of the fragment assembly approach in the field. In recent versions of ROSETTA (Bradley et al. 2005a, b; Das et al. 2007; Ovchinnikov et al. 2015), the authors first generated models in a reduced form with conformations specified with heavy backbone and Cβ atoms. In the second phase, a set of selected low-resolution models were subject to all-atom refinement procedure using an all-atom physics-based energy function, which includes van der Waals interactions, pair-wise solvation free energy, and an orientation-dependent hydrogen-bonding potential. The flowchart of the two-phase modeling is shown in Fig. 1.2 and details on the energy functions can be found in references (Bradley et al. 2005a, b; Das et al. 2007). For the conformational search, multiple rounds of Monte Carlo minimization (Li and Scheraga 1987) are carried out. One of the notable examples for this two-step protocol is the blind prediction of a FM target (T0281 from CASP6, 70 residues), whose Cα RMSD from its crystal structure is 1.6 Å (Bradley et al. 2005a, b), where a very extensive sampling was carried out using the distributed computing network of Rosetta@home allowing about 500,000 CPU hours for each target domain. Despite the significant success, the computational cost of the procedure is rather expensive for routine use.

Partially because of the notable success of the ROSETTA algorithm, as well as the limited availability of its energy functions to others, several groups initiated developments of their own energy functions following the idea of ROSETTA. Derivatives of ROSETTA include Simfold (Fujitsuka et al. 2006) and Profesy (Lee et al. 2004); their energy terms include van der Waals interactions, backbone dihedral angle potentials, hydrophobic interactions, backbone hydrogen-bonding potential, rotamer potential, pair-wise contact energies, beta-strand pairing, and a term controlling the protein radius of gyration. However, their predictions seems to be only partially successful in comparison to ROSETTA (Lee et al. 2004; Fujitsuka et al. 2006).

TASSER/I-TASSER Another successful free modeling approach, TASSER by Zhang and Skolnick (Zhang and Skolnick 2004a, b), constructs 3D models based on a purely knowledge-based approach. The target sequence is first threaded through a set of representative protein structures to search for possible folds. Contiguous fragments (>5 residues) are then excised from the threaded aligned regions and used to reassemble full-length models, while unaligned regions are built by a lattice-based ab initio modeling (Zhang et al. 2003). The protein conformation in TASSER is represented by a trace of Cα atoms and side-chain centers of mass,

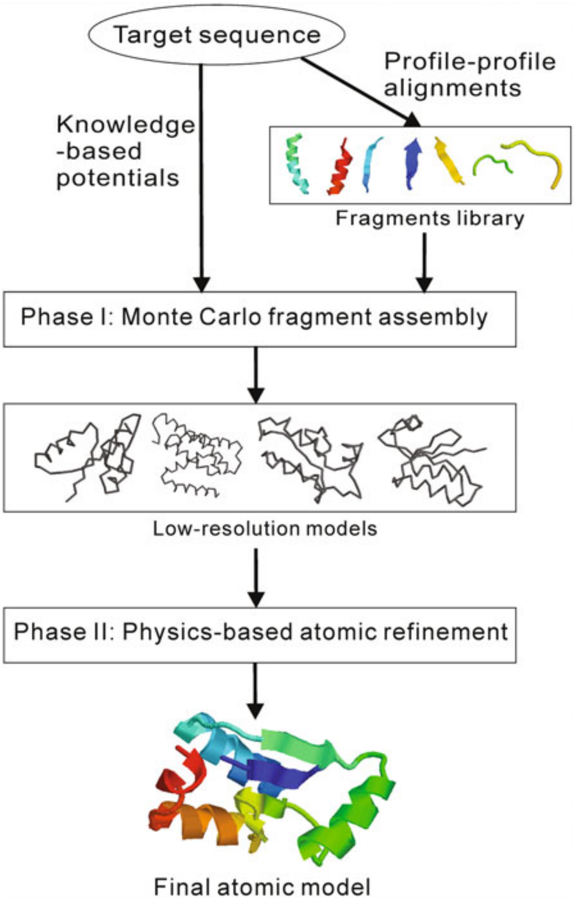
Target sequence
Profile-profile alignments
Knowledge -based potentials
Fragments library
Phase I: Monte Carlo fragment assembly
Low-resolution models
Phase II: Physics-based atomic refinement
Final atomic model

Fig. 1.2 Flowchart of the ROSETTA protocol (Simons et al. 1997). Fragments are first created from unrelated protein structures in the PDB, which are used to assemble full-length models by simulated annealing simulations guided by a knowledge-based force field. In the second phase, selected models are refined at atomic level using a physics-based potential

and the reassembly process is conducted by parallel-hyperbolic Monte Carlo simulations (Zhang et al. 2002). The energy terms of TASSER include information about predicted secondary structure propensities, backbone hydrogen bonds, a variety of short- and long-range correlations and hydrophobic energy based on the structural statistics from the PDB library. Weights of knowledge-based energy terms are optimized using a large-scale structure decoy set (Zhang et al. 2003) which coordinates the complicated correlations between various interaction terms.

Several derivatives of the TASSER approach have also found independent success. One is Chunk-TASSER (Zhou and Skolnick 2007), which first splits the target sequences into subunits (or “chunks”), each containing 3 consecutive regular secondary structure elements (helix and strand). These chunks are then folded separately. Finally, the spatial restraints are extracted from the chunk models and used for the subsequent TASSER simulations.

Another notable development is I-TASSER by Zhang and coworkers (Wu et al. 2007; Roy et al. 2010, Yang et al. 2015a, b), which refines TASSER cluster centroids by iterative fragment assembly simulations. The spatial restraints are extracted from the first round TASSER models and the template structures searched by TM-align (Zhang and Skolnick 2005a, b) from the PDB library, which are exploited in the second round simulations (Zhang and Skolnick 2013). The purpose is to remove the steric clashes from the first round models and refine the topology. Although the procedure uses structural fragments and spatial restraints from threading templates, it often constructs models of correct topology even when topologies of constituting templates are incorrect. From CASP7 to the latest CASP11 experiments, I-TASSER was consecutively ranked as one of the best methods for automated protein structure prediction (Battey et al. 2007; Cozzetto et al. 2009; Mariani et al. 2011; Montelione 2012; Kinch et al. 2015). As an independent test, Helles carried out a comparative study on 18 ab initio prediction algorithms and concluded that I-TASSER is about the best method in terms of the modeling accuracy and CPU cost per target (Helles 2008). Figure 1.3a shows an example of successful ab initio structure modeling by I-TASSER that constructed a correct model for the FM target T0604, which has a TM-score = 0.701 and RMSD = 2.66 Å from the X-ray structure.

Recently, many efforts have been made to improve the I-TASSER force field by the integration of sequence-based contact prediction (Wu and Zhang 2008a, b), short- and medium-range contact maps derived from segmental threading (Wu and Zhang 2010) and structure alignments (Zhang et al. 2011); these components have been proven particularly important for modeling distant-homology proteins in the CASP experiments (Zhang 2009; Xu et al. 2011; Zhang 2014; Zhang et al. 2015). The flowchart of current I-TASSER protocol is depicted in Fig. 1.4.

QUARK QUARK is a recently developed ab initio structural prediction method built on continuous fragment assembly using both knowledge and physics based energy terms (Xu and Zhang 2012). The flowchart of QUARK is shown in Fig. 1.5, which starts from position-specific fragment structure generation. At each residue position, 4000 (=200 × 20) structural fragments are generated, with lengths ranging from 1 to 20 residues, based on gapless threading of the fragment sequence through a non-redundant set of 6023 high-resolution PDB structures. The scoring function of the gapless threading consists of profile-profile, secondary structure, torsion angle and solvent accessibility matches (Wu et al. 2008a, b). Two types of information are derived from the fragments to assist next step of structure folding simulations. First, a torsion angle (φ, Ψ) distribution is collected from the 10-mer fragments at each residue position. Second, a residue-residue contact map is derived from the distance profiles between fragments. Here, a distance (d_{ij}) is recorded for each pair of fragments at two positions (i and j) if these two fragments come from the same PDB structure. A histogram is then generated for d_{ij} counting distances for all such fragment pairs. If the histogram of d_{ij} has a non-trivial peak below 9 Å, a contact between residue i and j will be predicted (Xu and Zhang 2013).

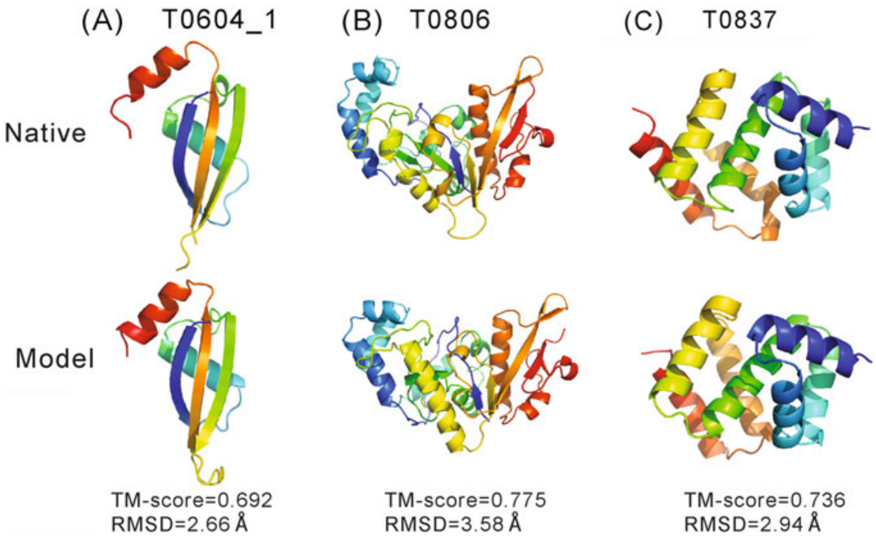

Fig. 1.3 Three examples of successful free modeling (FM) in recent CASP experiments. **a** T0604_1 is the first domain of the VP0956 protein from *Vibrio parahaemolyticus* in CASP9 that has 79 residues. The first model by the I-TASSER server has a TM-score = 0.692 and Cα-RMSD = 2.66 Å to the native. The success of this target was partially due to the sequence-based contact map prediction (Xu et al. 2011). **b** T0806 is the YaaA protein from *E. coli* K-12 in CASP11 that has 258 residues. The Rosetta human group (Ovchinnikov et al. 2015) constructed a correct model, using a co-evolution based contact prediction derived from >1100 homologous sequences, which has a TM-score = 0.775 and Cα-RMSD = 3.58 Å to the experimental structure. **c** T0837 is a hypothetical protein (YPO2654) from *Yersinia pestis* CO92 with 128 residues. The QUARK server generated a correct model with a TM-score = 0.736 and Cα-RMSD = 2.94 Å to the native, the success of which was attributed to the distance-profile based contact map prediction (Zhang et al. 2015). According to the assessors (Kinch et al. 2011; Kinch et al. 2015), there were no proteins in the PDB with a similar fold to any of these three targets at the time the predictions were made

In the next step, replica-exchange Monte Carlo (REMC) simulations are performed to assemble the fragments into full-length models under a composite physics- and knowledge-based potential, containing hydrogen bonding, van der Waals, solvation, Coulomb, backbone-torsion, bond-length and bond-angle, atomic distance, and strand pairing. The conformational changes are driven by 11 local and global movements shown in the top-right panel of Fig. 1.5. While the first feature, the torsion-angle distribution as collected from the fragments, is used to constrain local torsion movement selection, the second feature, the contact map derived from the fragment distance profiles, is used as a restraint to guide the simulations. The final models are selected by SPICKER (Zhang and Skolnick 2004a, b), which clusters all the decoys generated in the REMC simulations and ranks models by the size of the clusters.

Since its development, QUARK has been consistently ranked as one of the best methods in CASP for ab initio structure prediction (Kinch et al. 2011; Tai et al. 2014;

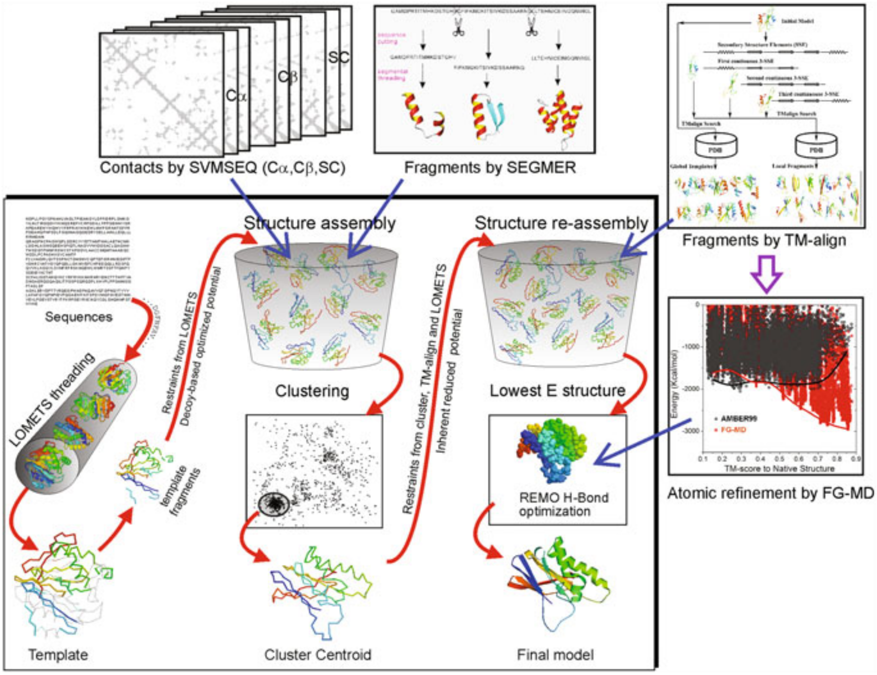

Fig. 1.4 Flowchart of I-TASSER protein structure modeling (Yang et al. 2015a, b). Multiple threading programs are used to identify templates and super-secondary structure fragments. Segments excised from the continuously aligned regions are used to reassemble the full-length models with the threading-unaligned regions built by lattice-based ab initio simulations. In the next step, templates structurally similar to the first-round models are identified from the PDB by structure alignments, with spatial restraints extracted from the templates to assist the second-round refinement simulations. In recent developments, sequence-based contact predictions and segmental threading were developed for improving results for distant homology modeling

Kinch et al. 2015). Figure 1.3c shows an example of the QUARK server modeling on T0837 in CASP11, where the distance profiles provided correct contacts for some of the critical medium-range contacts, which resulted in the first predicted models with a TM-score = 0.736 and RMSD = 2.94 Å to the experimental X-ray structure.

Coupling of Contact Prediction And Ab Initio Structure Prediction Sequence-based contact predictions have recently been investigated for improving ab initio modeling (Wu and Zhang 2008a, b; Wu et al. 2011; Marks et al. 2012; Kosciolek and Jones 2014). Unlike template-based protein structure prediction where high accuracy contacts can be derived from homologous structural templates, the CASP experiments for hard free-modeling (FM) protein targets show that purely sequence-based contact predictions can be more helpful than those collected from the best template-based models because the latter often have low quality for FM (Ezkurdia et al. 2009).

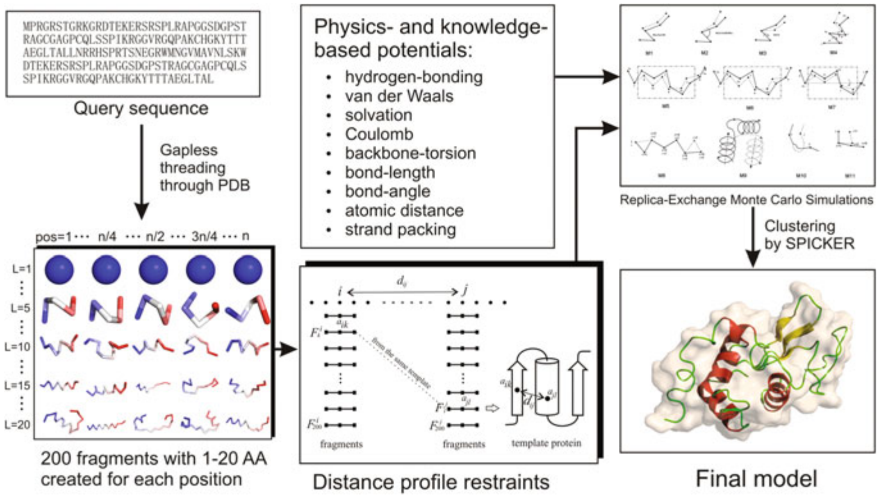
MPRGRSTGRKGRDTEKERSRSPLRAPGGSDGPST
RAGCGAGPCQLSSPIKRGGVRGQPAKCHGKYTTT
AEGLTALLNRRHSPRTSNEGRWMNGVMAVNLSKW
DTEKERSRSPLRAPGGSDGPSTRAGCGAGPCQLS
SPIKRGGVRGQPAKCHGKYTTTAEGLTAL
Query sequence
Gapless threading through PDB
pos=1 ··· n/4 ··· n/2 ··· 3n/4 ··· n
L=1
L=5
L=10
L=15
L=20
200 fragments with 1-20 AA created for each position
Physics- and knowledge-based potentials:
• hydrogen-bonding
• van der Waals
• solvation
• Coulomb
• backbone-torsion
• bond-length
• bond-angle
• atomic distance
• strand packing
fragments
fragments
template protein
Distance profile restraints
Replica-Exchange Monte Carlo Simulations
Clustering by SPICKER
Final model

Fig. 1.5 Flowchart of QUARK protein structure modeling (Xu and Zhang 2012). Multiple fragments with continuously distributed lengths are identified at each position from unrelated protein structures. Contact maps are then collected from distance profiles of the structural fragments, which are used to assist the fragment assembly simulations. Decoys are generated by replica-exchange Monte Carlo simulations under the guide of a composite physics and knowledge-based force field, with the final model selected by structure clustering

Some improvement of final models, with an average TM-score increase by 4.6%, was previously observed by Wu et al. after integrating nine SVM-based contact predictors (3 distance cutoffs multiplying 3 different contact atoms) into the I-TASSER force field (Wu et al. 2011). A handful of targets were converted from “nonfoldable” to “foldable” by several critical contacts when incorporated with the state-of-the-art structure assembly simulation methods. Similarly, Marks et al. (2011) showed that by integrating co-evolution based contact predictions with distance geometry programs, correct folds with RMSD values of 2.7–4.8 Å were generated for 15 test proteins with lengths between 50 and 260 residues. Later, Jones and coworker combined PSICOV (Jones et al. 2012), a co-evolution based contact predictor, with the fragment assembly program (Fragfold) and demonstrated the ability to fold 80% of cases with a TM-score above 0.5, when tested on a set of 150 proteins up to 266 amino acids in length (Kosciolek and Jones 2014).

One of the issues in applying co-evolution based contact predictions to ab initio structure prediction is that the accuracy of contact predictions depends on the number of homologous sequences that can be retrieved from the sequence databases, whereas hard FM targets often have few closely homologous sequences. Most recently, Baker and coworkers (Ovchinnikov et al. 2015) demonstrated an exciting achievement in the blind CASP11 experiment, where 4.6 *L* homologous sequences (with *L* being the protein length) were detected for a 256-residue FM target. The combination of the contact map with Rosetta simulations resulted in a first predicted model with the correct fold, with a TM-score = 0.775 and RMSD = 3.58 Å to the experimental structure (Fig. 1.3b). This probably represents the largest target that has been successfully folded in the CASP experiments, demonstrating the power of coupling contact map prediction and knowledge-based structure modeling.

1.3 Conformational Search Methods

Successful ab initio modeling of protein structures depends on the availability of a powerful conformation search method which can efficiently find the global minimum energy structure for a given energy function with a complicated energy landscape. Historically, Monte Carlo and molecular dynamics are two popular simulation methods to explore the conformational space of macromolecules such as proteins. For complicated systems like proteins, canonical MD/MC methods usually require a huge amount of computational resources for a complete exploration of the conformational space. The record for direct application of MD to obtain the protein native structure is not so impressive. One explanation for the failure could be that the simulation time required to fold a small protein takes as long as milliseconds, 10^{12} times longer than the usual incremental time step of femtoseconds (10^{-15} s). The technical difficulty of MC simulations mainly comes from that the energy landscape of protein conformational space is typically quite rugged containing many energy barriers, which may easily trap the Metropolis-based MC simulation procedures (Metropolis et al. 1953).

In this section we discuss recent development in conformational search methods to overcome these problems. We intend to illustrate the key ideas of conformational search methods used in various ab initio and related protein-modeling procedures. Unlike various energy functions used in ab initio modeling, the search methods should be, in principle, transferable between protein modeling methods, as well as other problems in science and technology. Currently, there exists no single omni-powerful search method that outperforms the others for all cases, and the investigation and systematic benchmarking on the performance of various search methods has yet to be carried out.

1.3.1 Monte Carlo Simulations

Simulated annealing (SA) (Kirkpatrick et al. 1983) is probably the most popular conformational search method. SA is general in that it is easy and straightforward to apply to any kind of optimization problem. In SA, one typically applies the Metropolis MC algorithm to generate a series of conformational states following the canonical Boltzmann energy distribution for a given temperature. SA initially executes high temperature MC simulation, followed by a series of simulations subject to a temperature-lowering schedule, hence the name simulated annealing. As much as SA is simple, its conformational search efficiency is not so impressive compared to other more sophisticated methods discussed below.

When the energy landscape of the system under investigation is rugged (due to numerous energy barriers), MC simulations are prone to get stuck in meta-stable states that will distort the distribution of sampled states by breaking the ergodicity of sampling. To avoid this malfunction, many simulation techniques have been developed, and one of the successful approaches is based on the generalized ensemble approach in contrast to the usual canonical ensemble. This kind of method was initially called by different names including multi-canonical ensemble (Berg and Neuhaus 1992) and entropic ensemble (Lee 1993). The underlying idea is to expedite the transition between states separated by energy barriers by modifying the transition probability so that the final energy distribution of sampling becomes more or less flat rather than bell-shaped. A popular method similar in this spirit is the replica exchange MC method (REM) (Swendsen and Wang 1986) where a set of many canonical MC simulations with temperatures distributed in a selected range are simultaneously carried out. From time to time one attempts to exchange structures (or equivalently temperatures) from neighboring simulations to sample states in a wide range of energy spectrum as the means to overcome energy barriers. Parallel hyperbolic sampling (PHS) (Zhang et al. 2002) further extends the REM by dynamically deforming energy using an inverse hyperbolic sine function to lower the energy barrier.

Monte Carlo with minimization (MCM), proposed by Li and Scheraga (1987), was successfully applied for the conformational search by several structure prediction programs (Simons et al. 1997). In MCM, one performs MC moves between local energy minima after local energy minimization of each perturbed protein structure. For a given local energy minimum structure A, a trial structure B is generated by random perturbation of A and is subsequently subject to local energy minimization. The usual Metropolis algorithm is used to determine the acceptance of B over A by calculating the energy difference between the two.

1.3.2 Molecular Dynamics

MD simulation (discussed in detail in Chap. 12) propagates physically realistic trajectories by applying Newton's equations of motion iteratively to allow atom movement, and is thus the most faithful method depicting atomistically what is occurring in proteins. The method is therefore often used for the study of protein folding pathways (Duan and Kollman 1998; Freddolino et al. 2010). The massive computational cost of long simulations is a major challenge with this method, since the incremental time scale is usually in the order of femtoseconds (10^{-15} s) while the fastest folding time of small proteins are on timescales of several microseconds (for folding model systems) or in the millisecond range (more typically). From the standpoint of search efficiency, molecular dynamics simulations are guaranteed to propagate some motion after each energy/force evaluation, but the steps that are taken are very small; in contrast, as described in the preceding section, Monte Carlo simulations may make larger steps, but not all steps will be accepted after energy evaluation. The relative sampling efficiency of the methods is thus dependent on the acceptance rate of Monte Carlo moves; with modern move sets (see, e.g., Fig. 1.5) Monte Carlo sampling of protein conformational space tends to be much more efficient. Thus, the application of molecular dynamics simulations using atomistic models is reserved for cases where the topic of interest is the folding process, rather than the folded structure per se. One unusual strength of MD sampling compared with Monte Carlo is that MD can accommodate the presence of explicit water much more readily, which might prove useful in the rare cases where implicit solvent models are directly responsible for failed structure predictions (Zhou 2003).

In addition, molecular dynamics simulations have been successfully applied in protein structure prediction using a variety of coarse-grained models, in which the computational complexity is substantially reduced and the folding accelerated due to the simulation of a smaller system with a less rugged energetic landscape, but of course with reduced resolution (Tozzini 2005; Hills and Brooks 2009). In addition, when a low-resolution model is available, MD simulations are often carried out for structure refinement since the conformational changes are assumed to be small (Zhang et al. 2011; Mirjalili and Feig 2013). Sampling in molecular dynamics simulations of protein folding may be enhanced using similar methods to those in Monte Carlo simulations, *e.g.* through the use of replica exchange simulations (Sugita and Okamoto 1999), but at the price of complicating the interpretation of folding kinetics and pathways. One particularly promising enhanced sampling method for future protein folding simulations and structure prediction is accelerated molecular dynamics (aMD) (Hamelberg et al. 2004), which applies a bias to lower the relative height of barriers on the potential energy surface. In a recent application, aMD allowed the prediction of the folded structures and folding free energy landscapes of a set of four commonly used model proteins with 10–100 fold less computational effort than unbiased simulations (Miao et al. 2015), providing promise for future applications to study folding pathways and equilibriums.

1.3.3 Genetic Algorithm

A genetic algorithm (GA) is a heuristic approach to the optimization problems based on a natural selection process mimicking the biological evolution. GA is designed to repeatedly modify a population of individual solutions. At each step, the algorithm randomly selects individuals from the current population, which are used as parents to produce the children for the next generation. Over successive generations, the population “evolves” toward the optimal solutions (Mitchell 1996).

Conformational space annealing (CSA) (Lee et al. 1998) is one of the most successful genetic algorithms developed for protein conformational search. By utilizing a local energy minimizer as in MCM and the concept of annealing in conformational space, it searches the whole conformational space of local minima in its early stages and then narrows the search to smaller regions with low energy as the distance cutoff is reduced. Here the distance cutoff is defined as the similarity between two conformations, and it controls the diversity of the conformational population. The distance cutoff plays the role of temperature in the usual SA, and initially its value is set to a large number in order to force conformational diversity. The value is gradually reduced as the search progresses. CSA has been successfully applied to various global optimization problems including protein structure prediction separately combined with ab initio modeling in UNRES (Oldziej et al. 2005) and ASTRO-FOLD (Klepeis and Floudas 2003; Klepeis et al. 2005), and with fragment assembly in Profesy (Lee et al. 2004).

1.3.4 Mathematical Optimization

The conformational searching approach by Floudas and coworkers, α branch and bound (αBB) (Klepeis and Floudas 2003; Klepeis et al. 2005), is unique in the sense that the method is mathematically rigorous, while all the others discussed here are stochastic and heuristic methods. The search space is successively cut into two halves while the lower and upper bounds of the global minimum (LB and UB) for each branched phase space are estimated. The estimate for the UB is simply the best currently obtained local minimum energy, and the estimate for the LB comes from the modified energy function augmented by a quadratic term of the dissecting variables with the coefficient α (hence the name αBB). With a sufficiently large value of α, the modified energy contains only one energy minimum, whose value serves as the lower bound. While performing successive dissection of the phase space accompanied by estimates of LB and UB for each dissected phase space, phase spaces with LB higher than the global UB can be eliminated from the search. The procedure continues until one identifies the global minimum by locating a dissected phase space where LB becomes identical to the global UB. Once the solution is found, the result is mathematically rigorous, but large proteins with many degrees of freedom are yet to be addressed by this method.

1.4 Model Selection

Ab initio modeling methods typically generate many non-native structure conformations (also called decoys) during the simulation. How to select appropriate models structurally close to the native state is an important issue. The development of algorithms for selection of protein models has been emerged as a new field called Model Quality Assessment Programs (MQAP) (Fischer 2006). In general, modeling selection approaches can be classified into two types, the energy based and the free energy based. In the energy-based methods, one designs a variety of specific potentials and identifies the lowest-energy state as the final prediction. In the free-energy based approaches, the free energy of a given conformation R can be written as

$$F(R) = -k_B T \ \ln Z(R) = -k_B T \ \ln \int\limits_{\Omega \in R} \mathrm{e}^{\frac{-E(R)}{k_B T}} \mathrm{d}\Omega \tag{1}$$

where $Z(R)$ is the restricted partition function which is proportional to the number of occurrences of the structures in the neighborhood of R during the simulation. This can be estimated by a clustering procedure at a given RMSD cutoff (Zhang and Skolnick 2004a, b).

For the energy-based model selection methods, we will discuss three energy/scoring functions: (1) physics-based energy function; (2) knowledge-based energy function; (3) scoring function describing the compatibility between the target sequence and model structures. In MQAP, there is another popular method which takes the consensus conformation from the predictions generated by different algorithms (Wallner and Elofsson 2007), also known as the meta-server approach (Ginalski et al. 2003; Wu et al. 2007). The essence of this method is similar to the clustering approach since both assume the most frequently occurring states to be the near-native structures. This approach has been mainly used for selecting models generated by threading-servers (Ginalski et al. 2003; Wu et al. 2007); but it has recently become popular for full-length model selection in the CASP experiments (Larsson et al. 2009; Kryshtafovych et al. 2015).

1.4.1 Physics-Based Energy Function

For the development of all-atom physics-based energy functions, Lazaridis and Karplus (1999a, b) exploited CHARMM19 (Neria et al. 1996) and EEF1 (Lazaridis and Karplus (1999a, b)) solvation potential to discriminate the native structure from decoys that are generated by threading on other protein structures. They found the energy of the native state is lower than those of decoys in most cases. Later, Petrey and Honig (Petrey and Honig 2000) used CHARMM and a continuum treatment of the solvent, Brooks and coworkers (Dominy and Brooks 2002; Feig and Brooks 2002) used CHARMM plus GB solvation, Felts et al. (2002) used OPLS plus GB, Lee and Duan (Lee et al. 2004) used AMBER plus GB, and Hsieh and Luo (2004) used AMBER plus Poisson-Boltzmann solvation potential on a number of structure decoy sets (including the Park-Levitt decoy set (Park and Levitt 1996), Baker decoy set (Tsai et al. 2003), Skolnick decoy set (Kihara et al. 2001; Skolnick et al. 2003), I-TASSER decoy set (Wu et al. 2007; Zhang and Zhang 2010), and CASP decoys set (Moult et al. 2001)). All these authors obtained similar results: the native structures have lower energy than decoys in their potentials. The claimed success of model discrimination of the physics-based potentials seems contradicted by other less successful physics-based structure prediction results. Wroblewska and Skolnick (Wroblewska and Skolnick 2007) showed that the AMBER plus GB potential can only discriminate the native structure from roughly minimized TASSER decoys (Zhang and Skolnick 2004a, b). After a 2-ns MD simulation on the decoys, none of the native structures were lower in energy than the lowest energy decoy, and the energy-RMSD correlation was close to zero. This result partially explains the discrepancy between the widely reported decoy discrimination ability of physics-based potentials and the less successful folding/refinement results.

Another related issue is that many of the decoy selection approaches are focused on the discrimination of the native structures from the decoy pools. However, such ability is of no practical usefulness in real cases of structure prediction because no structure prediction simulation could generate decoys exactly matching the native structure. Furthermore, the native structure has usually a nearly perfect local secondary structure packing, in addition to the fitness of global topology arrangement, whereas the computer generated decoys often have various flaws in the local structure packing and steric clashes. This makes it much more challenging to recognize the near-native structure decoys that are structurally closest to the native, compared to the task of discriminating the native structure from a set of computer-generated, flawed structure decoys (Deng et al. 2016).

1.4.2 Knowledge-Based Energy Function

Sippl proposed a pair-wise residue-distance based potential (Sippl 1990) using the statistics of known PDB structures in 1990 (its newest version is PROSA II (Sippl 1993; Wiederstein and Sippl 2007)). Since then, a variety of knowledge-based potentials have been developed, which include atomic interaction potential, solvation potential, hydrogen bond potential, torsion angle potential, etc. In the coarse-grained potentials, each residue is represented either by a single atom or by a few atoms, e.g., Cα-based potentials (Melo et al. 2002), Cβ-based potentials (Hendlich et al. 1990), side-chain-center-based potentials (Bryant and Lawrence 1993; Kocher et al. 1994; Thomas and Dill 1996; Skolnick et al. 1997; Zhang and Kim 2000; Zhang and Skolnick 2004a, b), side-chain and Cα-based potentials (Berrera et al. 2003).

One of the most widely-used knowledge-based potentials is a residue-specific, all-atom, distance-dependent potential, which was first formulated by Samudrala and Moult (RAPDF) (Samudrala and Moult 1998); it counts the distances between 167 amino acid specific pseudo-atoms. Following this, several atomic potentials with various reference states have been proposed, including those by Lu and Skolnick (KBP) (Lu and Skolnick 2001), Zhou and Zhou (DFIRE) (Zhou et al. 2002), Wang et al. (self-RAPDF) (Wang et al. 2004), Tostto (victor/FRST) (Tosatto 2005), Shen and Sali (DOPE) (Shen and Sali 2006), Zhang and Zhang (RW) (Zhang and Zhang 2010), and Zhou and Skolinck (GOAP) (Zhou and Skolnick 2011). All these potentials claimed that native structures could be distinguished from decoy structures in their tests. Deng et al. (2012) recently conducted a comparative investigation on all these potentials. To eliminate biases from the datasets and computing environments, they re-derived the potentials from a unified PDB structure dataset but based on the same original reference states. It was found that the performance varies with the tested decoy datasets and no potential could clearly outperform the others for all decoy sets.

The task of selecting the near-native models out of many decoys remains a challenge for these potentials (Skolnick 2006). Based on the CAFASP4-MQAP experiment in 2004 (Fischer 2006), the best-performing energy functions were Victor/FRST (Tosatto 2005) which incorporates an all-atom pair-wise interaction potential, solvation potential and hydrogen bond potential, and MODCHECK (Pettitt et al. 2005) which includes Cβ atom interaction potential and solvation potential. From CASP7-MQAP in 2006, the consensus-based method, Pcons developed by Elofsson group, showed the best performance (Wallner and Elofsson 2007). In the most recent CASP experiments, the consensus-based model selection scheme has kept ranking higher than any of the physics or knowledge-based scoring functions (Kryshtafovych et al. 2011; Kryshtafovych et al. 2014; Kryshtafovych et al. 2015). Several of the advanced structure modeling approaches in the CASP experiment have exploited a combined consensus and statistics scoring system to select models in the recent CASP (Cao et al. 2015; Yang et al. Yang et al. 2015a, b; Zhang et al. 2015).

1.4.3 Sequence-Structure Compatibility Function

In the third type of MQAPs, selection of the best models is not purely based on energy functions. Instead, they are selected based on the compatibility of target sequences to model structures. The earliest and still successful example is that by Luthy et al. (1992), who used threading scores to evaluate structures. Colovos and Yeates (1993) later used a quadratic error function to describe the non-covalently bonded interactions among atom pairs CC, CN, CO, NN, NO and OO, showing that near-native structures have fewer errors than other decoys. Verify3D (Eisenberg et al. 1997) improves the method of Luthy et al. (Luthy et al. 1992) by considering local threading scores in a 21-residue window. Jones developed GenThreader (Jones 1999) and used neural networks to classify native and non-native structures. The inputs of GenThreader include pairwise contact energy, solvation energy, alignment score, alignment length, and sequence and structure lengths. Similarly, based on neural networks, Wallner and Ellofsson built ProQ (Wallner and Elofsson 2003) for quality prediction of decoy structures. The inputs of ProQ include contacts, solvent accessible area, protein shape, secondary structure, structural alignment score between decoys and templates, and the fraction of protein regions to be modeled from templates. Later, McGuffin developed a consensus MQAP

(McGuffin 2007) called ModFold that includes ProQ (Wallner and Elofsson 2003), MODCHECK (Pettitt et al. 2005) and ModSSEA. The author showed that ModFold outperforms its component MQAP programs.

1.4.4 Clustering of Decoy Structures

For the purpose of identifying the lowest free-energy state, structure clustering techniques were adopted by many ab initio modeling approaches. In the work by Shortle et al. (1998), for all 12 cases tested, the cluster-center conformation of the largest cluster was closer to native structures than the majority of decoys. Cluster-center structures were ranked as the top 1–5% closest to their native structures.

Zhang and Skolnick developed an iterative structure clustering method, called SPICKER (Zhang and Skolnick 2004a, b). Based on 1489 representative benchmark proteins each with up to 280,000 structure decoys, the best of the top 5 models was ranked in the top 1.4% of all decoys. For 78% of the 1489 proteins, the RMSD difference between the best of the top 5 models and the most native-like decoy structure was less than 1 Å.

In ROSETTA ab initio modeling (Bradley et al. 2005a, b), structure decoys are clustered to select low-resolution models and these models are further refined by all-atom simulations to obtain final models. In the case of TASSER/I-TASSER (Zhang and Skolnick 2004a, b; Yang et al. 2015a, b) and QUARK (Xu and Zhang 2012), thousands of decoy models from MC simulations are clustered by SPICKER (Zhang and Skolnick 2004a, b) to generate cluster centroids as final models. In the approach by Scheraga and coworkers (Oldziej et al. 2005), decoys are clustered and the lowest-energy structures among the clustered structures are selected.

1.5 Remarks and Discussions

Successful ab initio modeling from amino acid sequence alone is considered the “Holy Grail” of protein structure prediction (Zhang 2008), since this will mark an eventual and complete solution to the problem. In addition to the generation of 3D structures, ab initio modeling can also help us understand the underlying principles of how proteins fold in nature; this could not be done by the template-based modeling approaches which build 3D models by copying and refining the framework of other solved structures.

An ideal approach to ab initio modeling would be to treat atoms in a protein as interacting particles according to an accurate physics-based potential, and fold the protein by solving Newton’s equations of motion in each step of movements. A number of molecular dynamics simulations were carried out along this line of approach by using the classic CHARMM and AMBER force fields. Although the

MD based simulation is very important for the study of protein folding, the success in the viewpoint of structure prediction is quite limited. One reason is the prohibitive computing demand for a normal size protein. On the other hand, knowledge-based (or hybrid knowledge- and physics-based) approaches making use of Monte Carlo sampling schemes appear to be progressing rapidly, producing many examples of successful low-to-medium accuracy models often with correct topology for small and medium size proteins. Although very rare, successful higher resolution models (<2–3 Å in Cα-RMSD) have been witnessed in blind experiments (Bradley et al. 2005a, b; Xu et al. 2011; Zhang et al. 2015).

The current state-of-the-art ab initio protein structure prediction methods often utilize as much information as possible from known structures, in several different ways. First, the use of local structure fragments directly excised from the PDB structures helps reduce the degrees of freedom and the entropy of the conformational search and yet keep the fidelity of the native protein structures. Second, the knowledge-based potential derived from the statistics of a large number of solved structures can appropriately grasp the subtle balance of the complicated correlations between different sources of energy terms (Summa and Levitt 2007). With the carefully parameterized knowledge-based potential terms aided by various advances in the conformational search methods, the accuracy of ab initio modeling for proteins up to 100–150 residues has been significantly improved in the last decade. With the help of co-evolution based contact map predictions, an exciting examples has been recently reported on a free-modeling target (T0806) up to 258 residues in the most recent CASP experiment (Ovchinnikov et al. 2015). However, such performance is only possible when sufficient number of homologous sequences can be obtained to ensure the accuracy of contact predictions: this situation is rare for ab initio modeling target proteins that have no homologues in the PDB.

For further improvement, parallel developments of accurate potential energy functions and efficient optimization methods are both necessary. That is, separate examination/development of potential energy functions is important; meanwhile, systematic benchmarking of various conformational search methods should be performed, so that the advantages as well as the limitations of available search methods can be explored separately. Currently, the ab initio modeling methods solely based on the physicochemical principles of interaction are still far behind, in terms of their modeling speed and accuracy, compared with the methods utilizing bioinformatics and knowledge-based information. However, the physics-based atomic potentials have recently demonstrated their potential in refining the detailed packing of side-chain atoms and peptide backbones (Zhang et al. 2011; Mirjalili and Feig 2013). Development of composite methods using both knowledge-based and physics-based energy terms should represent a promising approach to the problem of ab initio modeling.

It is important to acknowledge that with the progress in structure genomics and structural biology, the number of experimental structures in the PDB has been rapidly increasing, significantly extending the scope of the template-based protein structure predictions. Nevertheless, the traditional comparative modeling approaches can only yield model predictions with the accuracy of the templates, whereas the efficiency of template structure refinements is highly correlated with our ability in ab initio protein folding, because structure refinements often involve reconstruction of part of the side-chain and local backbone structures, and sometime the global topology for the low-resolution templates. Meanwhile, for most templates available in the PDB, a considerable portion of the sequence is either disordered or unaligned in the query-template alignments; the structures of these portions must be constructed using ab initio modeling. Finally, a very important bottleneck drawback in template-based modeling is that the alignment accuracy dramatically decreases with the sequence identity between query and template becomes low (e.g. <30%). Most recently, it has been demonstrated that the structural models built by free modeling can be used to help identify analogous templates that are of low sequence similarity but high structural similarity to the native, by matching the low-resolution ab initio models to experimentally solved structures in the PDB and thereby improve the success rate of distant-homologous structure predictions (Zhang 2014). Thus, the development of efficient ab initio folding algorithms will remain a major theme in the field and should have important impacts on all aspects of protein structure prediction.

Acknowledgements Authors want to thank Drs. Sitao Wu and Haiyou Deng for their contribution to the article. The project is supported in part by the National Institute of General Medical Sciences (GM083107, GM116960, and GM097033).

References

Bairoch A, Apweiler R, Wu CH et al (2005). The universal protein resource (UniProt). Nucleic Acids Res 33(Database issue): D154–159

Battey JN, Kopp J, Bordoli L et al (2007) Automated server predictions in CASP7. Proteins 69 (S8):68–82

Berendsen HJC, Postma JPM, van Gunsteren WF et al (1981) Interaction models for water in relation to protein hydration. Intermolecular forces, Reidel, The Netherlands

Berg BA, Neuhaus T (1992) Multicanonical ensemble: a new approach to simulate first-order phase transitions. Physical Review Letters 68(1):9–12

Berman HM, Westbrook J, Feng Z et al (2000) The protein data bank. Nucleic Acids Research 28(1):235–242

Berrera M, Molinari H, Fogolari F (2003) Amino acid empirical contact energy definitions for fold recognition in the space of contact maps. BMC Bioinform 4:8

Best RB, Buchete NV, Hummer G (2008) Are current molecular dynamics force fields too helical? Biophysical Journal 95(1):L07–09

Best RB, Hummer G (2009) Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. J Phys Chem B 113(26):9004–9015

Bowie JU, Eisenberg D (1994) An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function. Proc Natl Acad Sci U S A 91(10):4436–4440

Bradley P, Malmstrom L, Qian B et al (2005a) Free modeling with Rosetta in CASP6. Proteins 61(Suppl 7):128–134

Bradley P, Misura KM, Baker D (2005b) Toward high-resolution de novo structure prediction for small proteins. Science 309(5742):1868–1871

Brooks BR, Bruccoleri RE, Olafson BD et al (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. Journal of Computational Chemistry 4(2): 187–217

Bryant SH, Lawrence CE (1993) An empirical energy function for threading protein sequence through the folding motif. Proteins 16(1):92–112

Cao R, Bhattacharya D, Adhikari B et al (2015). Massive integration of diverse protein quality assessment methods to improve template based modeling in CASP11. Proteins 84:247–259

Case DA, Pearlman DA, Caldwell JA et al (1997). AMBER 5.0, University of California, San Francisco

Chen J, Brooks CL 3rd (2007) Can molecular dynamics simulations provide high-resolution refinement of protein structure? Proteins 67(4):922–930

Chowdhury S, Lee MC, Xiong GM et al (2003) Ab initio folding simulation of the Trp-cage mini-protein approaches NMR resolution. Journal of Molecular Biology 327(3):711–717

Colovos C, Yeates TO (1993) Verification of protein structures: patterns of nonbonded atomic interactions. Protein Science 2(9):1511–1519

Cornell WD, Cieplak P, Bayly CI et al (1995) A Second generation force field for the simulation of proteins, nucleic acids, and organic molecules. Journal of the American Chemical Society 117:5179–5197

Cozzetto D, Kryshtafovych A, Fidelis K et al (2009) Evaluation of template-based models in CASP8 with standard measures. Proteins 77(Suppl 9):18–28

Das R, Qian B, Raman S et al (2007) Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. Proteins 69(S8):118–128

Deng H, Jia Y, Zhang Y (2016) 3DRobot: automated generation of diverse and well-packed protein structure decoys. Bioinformatics 32(3):378–387

Deng HY, Jia Y, Wei YY et al (2012) What is the best reference state for designing statistical atomic potentials in protein structure prediction? Proteins-Structure Function and Bioinformatics 80(9):2311–2322

Dominy BN, Brooks CL (2002) Identifying native-like protein structures using physics-based potentials. Journal of Computational Chemistry 23(1):147–160

Duan Y, Kollman PA (1998) Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science 282(5389):740–744

Eisenberg D, Luthy R, Bowie JU (1997) VERIFY3D: assessment of protein models with three-dimensional profiles. Methods in Enzymology 277:396–404

Ensign DL, Kasson PM, Pande VS (2007) Heterogeneity even at the speed limit of folding: large-scale molecular dynamics study of a fast-folding variant of the villin headpiece. Journal of Molecular Biology 374(3):806–816

Ezkurdia I, Grana O, Izarzugaza JM et al (2009) Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8. Proteins 77(Suppl 9):196–209

Fan H, Mark AE (2004) Refinement of homology-based protein structures by molecular dynamics simulation techniques. Protein Science 13(1):211–220

Feig M, Brooks CL 3rd (2002) Evaluating CASP4 predictions with physical energy functions. Proteins 49(2):232–245

Feig M, Mirjalili V (2015). Protein structure refinement via molecular-dynamics simulations: what works and what does not? Proteins 84:282–292

Felts AK, Gallicchio E, Wallqvist A et al (2002) Distinguishing native conformations of proteins from decoys with an effective free energy estimator based on the OPLS all-atom force field and the Surface Generalized Born solvent model. Proteins 48(2):404–422

Fischer D (2006) Servers for protein structure prediction. Current Opinion in Structural Biology 16(2):178–182

Freddolino PL, Harrison CB, Liu Y et al (2010) Challenges in protein folding simulations: timescale, representation, and analysis. Nature Physics 6(10):751–758

Freddolino PL, Liu F, Gruebele M et al (2008) Ten-microsecond molecular dynamics simulation of a fast-folding WW domain. Biophysical Journal 94(10):L75–77

Freddolino PL, Park S, Roux B et al (2009) Force field bias in protein folding simulations. Biophysical Journal 96(9):3772–3780

Freddolino PL, Schulten K (2009) Common structural transitions in explicit-solvent simulations of villin headpiece folding. Biophysical Journal 97(8):2338–2347

Fujitsuka Y, Chikenji G, Takada S (2006) SimFold energy function for de novo protein structure prediction: consensus with Rosetta. Proteins 62(2):381–398

Ginalski K, Elofsson A, Fischer D et al (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 19(8):1015–1018

Hagler A, Euler E, Lifson S (1974) Energy functions for peptides and proteins i. derivation of a consistent force field including the hydrogen bond from amide crystals. Journal of the American Chemical Society 96:5319–5327

Hamelberg D, Mongan J, McCammon JA (2004) Enhanced sampling of conformational transitions in proteins using full atomistic accelerated molecular dynamics simulations. Protein Science 13:76

Helles G (2008) A comparative study of the reported performance of ab initio protein structure prediction algorithms. Journal of the Royal Society, Interface 5(21):387–396

Hendlich M, Lackner P, Weitckus S et al (1990) Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. Journal of Molecular Biology 216(1):167–180

Hills RD Jr, Brooks CL 3rd (2009) Insights from coarse-grained go models for protein folding and dynamics. International Journal of Molecular Sciences 10(3):889–905

Hsieh MJ, Luo R (2004) Physical scoring function based on AMBER force field and poisson-boltzmann implicit solvent for protein structure prediction. Proteins 56(3):475–486

Im W, Lee MS, Brooks CL 3rd (2003) Generalized born model with a simple smoothing function. Journal of Computational Chemistry 24(14):1691–1702

Jagielska A, Wroblewska L, Skolnick J (2008) Protein model refinement using an optimized physics-based all-atom force field. Proceedings of the National Academy of Sciences of the United States of America 105(24):8268–8273

Jauch R, Yeo HC, Kolatkar PR et al (2007) Assessment of CASP7 structure predictions for template free targets. Proteins 69(Suppl 8):57–67

Jonassen I, Klose D, Taylor WR (2006) Protein model refinement using structural fragment tessellation. Computational Biology and Chemistry 30(5):360–366

Jones DT (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. Journal of Molecular Biology 287(4):797–815

Jones DT, Buchan DW, Cozzetto D et al (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28(2):184–190

Jorgensen WL, Chandrasekhar J, Madura JD et al (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926–935

Jorgensen WL, Maxwell DS, Tirado-Rives J (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. Journal of the American Chemical Society 118:11225–11236

Jorgensen WL, Tirado-Rives J (1988) The OPLS potential functions for proteins. energy minimizations for crystals of cyclic peptides and crambin. Journal of the American Chemical Society 110:1657–1666

Kaminski GA, Friesner RA, Tirado-Rives J et al (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. J Phys Chem B 105:6474–6487

Karplus K, Barrett C, Hughey R (1998) Hidden markov models for detecting remote protein homologies. Bioinformatics 14:846–856

Kihara D, Lu H, Kolinski A et al (2001) TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. Proc Natl Acad Sci U S A 98(18): 10125–10130

Kinch L, Yong Shi S, Cong Q et al (2011) CASP9 assessment of free modeling target predictions. Proteins 79(Suppl 10):59–73

Kinch LN, Li W, Monastyrskyy B, et al. (2015). Evaluation of free modeling targets in CASP11 and ROLL. Proteins 84: 51–66

Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220 (4598):671–680

Klepeis JL, Floudas CA (2003) ASTRO-FOLD: a combinatorial and global optimization framework for Ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. Biophysical Journal 85(4):2119–2146

Klepeis JL, Wei Y, Hecht MH et al (2005) Ab initio prediction of the three-dimensional structure of a de novo designed protein: a double-blind case study. Proteins 58(3):560–570

Kocher JP, Rooman MJ, Wodak SJ (1994) Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. Journal of Molecular Biology 235(5): 1598–1613

Kosciolek T, Jones DT (2014) De novo structure prediction of globular proteins aided by sequence variation-derived contacts. PLoS ONE 9(3):e92197

Kryshtafovych A, Barbato A, Fidelis K et al (2014) Assessment of the assessment: evaluation of the model quality estimates in CASP10. Proteins 82(Suppl 2):112–126

Kryshtafovych A, Barbato A, Monastyrskyy B, et al (2015) Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in CASP11. Proteins 84: 349–369

Kryshtafovych A, Fidelis K, Tramontano A (2011) Evaluation of model quality predictions in CASP9. Proteins 79(Suppl 10):91–106

Larsson P, Skwark MJ, Wallner B et al (2009) Assessment of global and local model quality in CASP8 using Pcons and ProQ. Proteins 77(Suppl 9):167–172

Lazaridis T, Karplus M (1999a) Discrimination of the native from misfolded protein models with an energy function including implicit solvation. Journal of Molecular Biology 288(3):477–487

Lazaridis T, Karplus M (1999b) Effective energy function for proteins in solution. Proteins 35(2): 133–152

Lee J (1993) New monte carlo algorithm: entropic sampling. Physical Review Letters 71(2): 211–214

Lee J, Kim SY, Joo K et al (2004) Prediction of protein tertiary structure using PROFESY, a novel method based on fragment assembly and conformational space annealing. Proteins 56 4):704–714

Lee J, Scheraga HA, Rackovsky S (1998) Conformational analysis of the 20-residue membrane-bound portion of melittin by conformational space annealing. Biopolymers 46 (2):103–116

Lee MC, Duan Y (2004) Distinguish protein decoys by using a scoring function based on a new AMBER force field, short molecular dynamics simulations, and the generalized born solvent model. Proteins 55(3):620–634

Lee MR, Tsai J, Baker D et al (2001) Molecular dynamics in the endgame of protein structure prediction. Journal of Molecular Biology 313(2):417–430

Lei HX, Wu C, Liu HG et al (2007) Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. Proceedings of the National Academy of Sciences of the United States of America 104(12):4925–4930

Levitt M, Hirshberg M, Sharon R et al (1995) Potential-energy function and parameters for simulations of the molecular-dynamics of proteins and nucleic-acids in solution. Computer Physics Communications 91(1–3):215–231

Li Z, Scheraga HA (1987) Monte carlo-minimization approach to the multiple-minima problem in protein folding. Proc Natl Acad Sci U S A 84(19):6611–6615

Lindahl E, Hess B, van der Spoel D (2001) GROMACS 3.0: a package for molecular simulation and trajectory analysis. J Mol Modeling 7:306–317

Lindorff-Larsen K, Maragakis P, Piana S et al (2012) Systematic validation of protein force fields against experimental data. PLoS ONE 7(2):e32131

Lindorff-Larsen K, Piana S, Dror RO et al (2011) How fast-folding proteins fold. Science 334(6055):517–520

Lindorff-Larsen K, Piana S, Palmo K et al (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins 78(8):1950–1958

Liwo A, Khalili M, Scheraga HA (2005) Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. Proc Natl Acad Sci U S A 102(7):2362–2367

Liwo A, Lee J, Ripoll DR et al (1999) Protein structure prediction by global optimization of a potential energy function. Proc Natl Acad Sci U S A 96(10):5482–5485

Liwo A, Pincus MR, Wawak RJ et al (1993) Calculation of protein backbone geometry from alpha-carbon coordinates based on peptide-group dipole alignment. Protein Science 2(10): 1697–1714

Lu H, Skolnick J (2001) A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins 44(3):223–232

Luthy R, Bowie JU, Eisenberg D (1992) Assessment of protein models with three-dimensional profiles. Nature 356(6364):83–85

MacKerell AD Jr, Bashford D, Bellott M et al (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 102(18):3586–3616

Mariani V, Kiefer F, Schmidt T et al (2011) Assessment of template based protein structure predictions in CASP9. Proteins 79(Suppl 10):37–58

Marks DS, Colwell LJ, Sheridan R et al (2011) Protein 3D structure computed from evolutionary sequence variation. PLoS ONE 6(12):e28766

Marks DS, Hopf TA, Sander C (2012) Protein structure prediction from sequence variation. Nature Biotechnology 30(11):1072–1080

McGuffin LJ (2007) Benchmarking consensus model quality assessment for protein fold recognition. BMC Bioinformatics 8:345

Melo F, Sanchez R, Sali A (2002) Statistical potentials for fold assessment. Protein Science 11(2): 430–448

Metropolis N, Rosenbluth AW, Rosenbluth MN et al (1953) Equation of state calculations by fast computing machines. J Chem Phys 21:1087–1092

Miao YL, Feixas F, Eun CS et al (2015) Accelerated molecular dynamics simulations of protein folding. Journal of Computational Chemistry 36(20):1536–1549

Mirjalili V, Feig M (2013) Protein structure refinement through structure selection and averaging from molecular dynamics ensembles. Journal of Chemical Theory and Computation 9(2): 1294–1303

Mitchell M (1996). An Introduction to Genetic Algorithms. Cambridge, MIT Press

Mittal J, Best RB (2010) Tackling force-field bias in protein folding simulations: folding of Villin HP35 and Pin WW domains in explicit water. Biophysical Journal 99(3):L26–28

Montelione GT (2012). Template based modeling assessment in CASP10. 10th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction. Gaeta, Italy

Moult J, Fidelis K, Zemla A et al (2001) Critical assessment of methods of protein structure prediction (CASP): round IV. Proteins Suppl 5:2–7

Nemethy G, Gibson KD, Palmer KA et al (1992) Energy parameters in polypeptides. 10. Improved geometric parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. J Phys Chem B 96:6472–6484

Neria E, Fischer S, Karplus M (1996) Simulation of activation free energies in molecular systems. J Chem Phys 105(5):1902–1921

Nguyen H, Maier J, Huang H et al (2014) Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent. Journal of the American Chemical Society 136(40):13959–13962

Nilges M, Brunger AT (1991) Automated modeling of coiled coils: application to the GCN4 dimerization region. Protein Engineering 4(6):649–659

Oldziej S, Czaplewski C, Liwo A et al (2005) Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: assessment in two blind tests. Proceedings of the National Academy of Sciences of the United States of America 102(21): 7547–7552

Ovchinnikov S, Kim DE, Wang RY, et al (2015) Improved de novo structure prediction in CASP11 by incorporating Co-evolution information into rosetta. Proteins 84:67–75

Park B, Levitt M (1996) Energy functions that discriminate X-ray and near native folds from well-constructed decoys. Journal of Molecular Biology 258(2):367–392

Petrey D, Honig B (2000) Free energy determinants of tertiary structure and the evaluation of protein models. Protein Science 9(11):2181–2191

Pettitt CS, McGuffin LJ, Jones DT (2005) Improving sequence-based fold recognition by using 3D model quality assessment. Bioinformatics 21(17):3509–3515

Piana S, Klepeis JL, Shaw DE (2014) Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. Current Opinion in Structural Biology 24:98–105

Piana S, Lindorff-Larsen K, Shaw DE (2011) How robust are protein folding simulations with respect to force field parameterization? Biophysical Journal 100(9):L47–49

Piana S, Lindorff-Larsen K, Shaw DE (2012) Protein folding kinetics and thermodynamics from atomistic simulation. Proceedings of the National Academy of Sciences of the United States of America 109(44):17845–17850

Piana S, Lindorff-Larsen K, Shaw DE (2013a) Atomic-level description of ubiquitin folding. Proc Natl Acad Sci U S A 110(15):5915–5920

Piana S, Lindorff-Larsen K, Shaw DE (2013b) Atomistic description of the folding of a dimeric protein. J Phys Chem B 117(42):12935–12942

Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nature Protocols 5(4):725–738

Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. Journal of Molecular Biology 234(3):779–815

Samudrala R, Moult J (1998) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. Journal of Molecular Biology 275(5):895–916

Shen MY, Sali A (2006) Statistical potential for assessment and prediction of protein structures. Protein Science 15(11):2507–2524

Shortle D, Simons KT, Baker D (1998) Clustering of low-energy conformations near the native structures of small proteins. Proc Natl Acad Sci U S A 95(19):11158–11162

Simons KT, Kooperberg C, Huang E et al (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. Journal of Molecular Biology 268(1):209–225

Sippl MJ (1990) 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 213(4):859–883

Sippl MJ (1993) Recognition of errors in three-dimensional structures of proteins. Proteins 17(4): 355–362

Skolnick J (2006) In quest of an empirical potential for protein structure prediction. Current Opinion in Structural Biology 16(2):166–171

Skolnick J, Jaroszewski L, Kolinski A et al (1997) Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? Protein Science 6:676–688

Skolnick J, Kihara D, Zhang Y (2004) Development and large scale benchmark testing of the PROSPECTOR 3.0 threading algorithm. Protein 56:502–518

Skolnick J, Zhang Y, Arakaki AK et al (2003) TOUCHSTONE: A unified approach to protein structure prediction. Proteins 53(Suppl 6):469–479

Soding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21(7): 951–960

Sorin EJ, Pande VS (2005) Exploring the helix-coil transition via all-atom equilibrium ensemble simulations. Biophysical Journal 88(4):2472–2493

Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. Chemical Physics Letters 314(1–2):141–151

Summa CM, Levitt M (2007) Near-native structure refinement using in vacuo energy minimization. Proc Natl Acad Sci U S A 104(9):3177–3182

Swendsen RH, Wang JS (1986) Replica Monte Carlo simulation of spin glasses. Physical Review Letters 57(21):2607–2609

Tai CH, Bai H, Taylor TJ et al (2014) Assessment of template-free modeling in CASP10 and ROLL. Proteins 82(Suppl 2):57–83

Taylor WR, Bartlett GJ, Chelliah V et al (2008) Prediction of protein structure from ideal forms. Proteins 70(4):1610–1619

Thomas PD, Dill KA (1996) Statistical potentials extracted from protein structures: how accurate are they? Journal of Molecular Biology 257(2):457–469

Tosatto SC (2005) The victor/FRST function for model quality estimation. Journal of Computational Biology 12(10):1316–1327

Tozzini V (2005) Coarse-grained models for proteins. Current Opinion in Structural Biology 15(2):144–150

Tsai J, Bonneau R, Morozov AV et al (2003) An improved protein decoy set for testing energy functions for protein structure prediction. Proteins 53(1):76–87

van Gunsteren WF, Billeter SR, Eising AA et al (1996). Biomolecular simulation: The GROMOS96 Manual and User Guide Univ Publ House, Zurich

Vieth M, Kolinski A, Brooks CL et al (1994) Prediction of the folding pathways and structure of the GCN4 leucine zipper. Journal of Molecular Biology 237(4):361–367

Wallner B, Elofsson A (2003) Can correct protein models be identified? Protein Science 12(5): 1073–1086

Wallner B, Elofsson A (2007) Prediction of global and local model quality in CASP7 using Pcons and ProQ. Proteins 69(S8):184–193

Wang JM, Cieplak P, Kollman PA (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? Journal of Computational Chemistry 21(12):1049–1074

Wang K, Fain B, Levit M et al (2004). Improved protein structure selection using decoy-dependent discriminatory functions. BMC Structural Biology 4(8)

Weiner SJ, Kollman PA, Case DA et al (1984) A new force field for molecular mechanical simulation of nucleic acids and proteins. Journal of the American Chemical Society 106:765–784

Wiederstein M, Sippl MJ (2007). ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res 35(Web Server issue): W407–410

Wroblewska L, Skolnick J (2007) Can a physics-based, all-atom potential find a protein's native structure among misfolded structures? i. large scale AMBER benchmarking. Journal of Computational Chemistry 28(12):2059–2066

Wu S, Skolnick J, Zhang Y (2007) Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biology 5:17

Wu S, Szilagyi A, Zhang Y (2011) Improving protein structure prediction using multiple sequence-based contact predictions. Structure 19(8):1182–1191

Wu S, Zhang Y (2007) LOMETS: a local meta-threading-server for protein structure prediction. Nucleic Acids Research 35(10):3375–3382

Wu S, Zhang Y (2008a) A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. Bioinformatics 24(7):924–931

Wu S, Zhang Y (2008b) MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins 72(2):547–556

Wu S, Zhang Y (2010) Recognizing protein substructure similarity using segmental threading. Structure 18(7):858–867

Xu D, Zhang J, Roy A et al (2011) Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. Proteins 79(Suppl 10):147–160

Xu D, Zhang Y (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins 80(7):1715–1735

Xu D, Zhang Y (2013) Toward optimal fragment generations for ab initio protein structure assembly. Proteins 81(2):229–239

Yang J, Yan R, Roy A et al (2015a) The I-TASSER Suite: protein structure and function prediction. Nature Methods 12(1):7–8

Yang J, Zhang W, He B, et al (2015) Template-based protein structure prediction in CASP11 and retrospect of I-TASSER in the last decade. Proteins 84: 233–246

Yang Y, Faraggi E, Zhao H et al (2011) 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 27(15):2076–2082

Zagrovic B, Snow CD, Shirts MR et al (2002) Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. Journal of Molecular Biology 323(5):927–937

Zhang C, Kim SH (2000) Environment-dependent residue contact energies for proteins. Proc Natl Acad Sci U S A 97(6):2550–2555

Zhang C, Liu S, Zhou H et al (2004) An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. Protein Science 13(2):400–411

Zhang J, Liang Y, Zhang Y (2011) Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. Structure 19(12):1784–1795

Zhang J, Zhang Y (2010) A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. PLoS ONE 5 (10):e15386

Zhang W, Yang J, He B et al (2015). Integration of QUARK and I-TASSER for Ab initio protein structure prediction in CASP11. Proteins 84: 76–86

Zhang Y (2008). Progress and Challenges in protein structure prediction. Curr Opin Struct Biol: In press

Zhang Y (2009) I-TASSER: Fully automated protein structure prediction in CASP8. Proteins 77(S9):100–113

Zhang Y (2014) Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. Proteins 82(Suppl 2):175–187

Zhang Y, Hubner I, Arakaki A et al (2006) On the origin and completeness of highly likely single domain protein structures. Proc Natl Acad Sci U S A 103:2605–2610

Zhang Y, Kihara D, Skolnick J (2002) Local energy landscape flattening: parallel hyperbolic monte carlo sampling of protein folding. Proteins-Struct Func Genet 48(2):192–201

Zhang Y, Kolinski A, Skolnick J (2003) TOUCHSTONE II: A new approach to ab initio protein structure prediction. Biophysical Journal 85(2):1145–1164

Zhang Y, Skolnick J (2004a) Automated structure prediction of weakly homologous proteins on a genomic scale. Proceedings of the National Academy of Sciences of the United States of America 101:7594–7599

Zhang Y, Skolnick J (2004b) SPICKER: a clustering approach to identify near-native protein folds. Journal of Computational Chemistry 25(6):865–871

Zhang Y, Skolnick J (2005a) The protein structure prediction problem could be solved using the current PDB library. Proc Natl Acad Sci U S A 102:1029–1034

Zhang Y, Skolnick J (2005b) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Research 33(7):2302–2309

Zhang Y, Skolnick J (2013) Segment assembly, structure alignment and iterative simulation in protein structure prediction. BMC Biology 11:44

Zhou H, Skolnick J (2007) Ab initio protein structure prediction using chunk-TASSER. Biophysical Journal 93(5):1510–1518

Zhou H, Skolnick J (2011) GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophysical Journal 101(8):2043–2052

Zhou H, Zhou Y (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Science 11(11):2714–2726

Zhou R (2003) Free energy landscape of protein folding in water: explicit vs. implicit solvent. Proteins 53(2):148–161