Chapter 1
Ab Initio Protein Structure Prediction

Jooyoung Lee, Sitao Wu, and Yang Zhang

Abstract Predicting protein 3D structures from the amino acid sequence still remains as an unsolved problem after five decades of efforts. If the target protein has a homologue already solved, the task is relatively easy and high-resolution models can be built by copying the framework of the solved structure. However, such a modelling procedure does not help answer the question of how and why a protein adopts its specific structure. If structure homologues (occasionally analogues) do not exist, or exist but cannot be identified, models have to be constructed from scratch. This procedure, called *ab initio* modelling, 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* modelling is low and the success is limited to small proteins (<100 residues). In this chapter, we give a review on the field of *ab initio* modelling. Focus will be put on three key factors of the modelling algorithms: energy function, conformational search, and model selection. Progresses and advances of a variety of algorithms will be discussed.

1.1 Introduction

With the tremendous success of the genome sequence projects, the number of available protein sequences is increasing exponentially. However, due to the technical difficulties and heavy labor and time costs of the experimental structure determination, the number of available protein structures lags far behind. By the end of 2007,

J. Lee
Center for Bioinformatics and Department of Molecular Bioscience,
University of Kansas, Lawrence, KS, 66047, USA
School of Computational Sciences, Korea Institute for Advanced Study,
Seoul, 130–722, Korea

S. Wand and Y. Zhang*
Centre for Bioinformatics and Department of Molecular Bioscience,
University of Kansas, Lawreance, KS, 66047, USA
*Corresponding author: e-mail: yzhang@ku.edu

D.J. Rigden (ed.) *From Protein Structure to Function with Bioinformatics*,

about 5.3 million protein sequences were deposited in the UniProtKB database (Bairoch et al. 2005) (http://www.ebi.ac.uk/swissprot). However, the corresponding number of protein structures in the Protein Data Bank (PDB) (Berman et al. 2000) (http://www.rcsb.org/pdb) is only about 44,000, less than 1% of the protein sequences. The gap is rapidly widening as indicated in Fig. 1.1. Thus, developing efficient computer-based algorithm to predicting 3D structures from sequences is 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 the framework of the solved proteins (templates). The procedure is called “template-based modelling (TBM)” (Karplus et al. 1998; Jones 1999; Shi et al. 2001; Ginalski et al. 2003b; Skolnick et al. 2004; Jaroszewski et al. 2005; Soding 2005; Zhou and Zhou 2005; Cheng and Baldi 2006; Pieper et al. 2006; Wu and Zhang 2008), which will be discussed in the subsequent chapters. Although high-resolution models can be often 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 called by several names, e.g. *ab initio* modelling (Klepeis et al. 2005; Liwo et al. 2005; Wu et al. 2007), *de novo* modelling (Bradley et al. 2005), physics-based modelling (Oldziej et al. 2005), or free modelling (Jauch et al. 2007). In this chapter, the term *ab initio* modelling is uniformly used to avoid confusion. Unlike the template-based modelling, successful *ab initio* modelling procedure could help answer the basic questions on how and why a protein adopts the specific structure out of many possibilities.

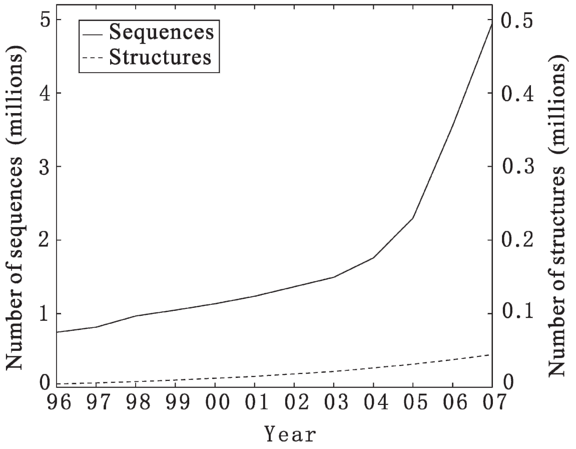

Fig. 1.1 The number of available protein sequences (left ordinate) and the solved protein structures (right ordinate) are shown for the last 12 years. The ratio of sequence/structure is rapidly increasing. Data are taken from UniProtKB (Bairoch et al. 2005) and PDB (Berman et al. 2000) databases

Typically, *ab initio* modelling conducts a conformational search under the guidance of a designed energy function. This procedure usually generates a number of possible conformations (structure decoys), and final models are selected from them. Therefore, a successful *ab initio* modelling 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) selection of native-like models from a pool of decoy structures.

This chapter gives a review on the current state of the art in *ab initio* protein structure prediction. This review is neither complete to include all available *ab initio* methods nor in depth to provide all backgrounds/motivations behind them. For a comparative study of various *ab initio* modelling methods, readers are recommended to read a recent review by Helles (Helles 2008). The rest of the chapter is organized as follows. Three major issues of *ab initio* modelling, i.e. energy function, 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 discussed. Finally, current progresses and challenges of *ab initio* modelling are summarized.

1.2 Energy Functions

In this section, we will discuss energy functions used for *ab initio* modelling. 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 modelling procedure often loses its power/validity. We classify the energy into two groups: (a) physics-based energy functions and (b) knowledge-based energy functions, depending on the use of statistics from the existing protein 3D structures. A few promising methods from each group are selected to discuss according to their uniqueness and modelling accuracy. A list of *ab initio* modelling methods is provided in Table 1.1 along with their properties about energy functions, conformational search algorithms, model selection methods and typical running times.

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

Table 1.1 A list of *ab initio* modelling 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 & 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, 2005; Oldziej et al. 2005)	Physics-based	Conformational space annealing (CSA)	Clustering/free-energy	Hours
ASTRO-FOLD (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	Months
TASSER/Chunk-TASSER (Zhang and Skolnick 2004a; Zhou and Skolnick 2007) http://cssb.biology.gatech.edu/skolnick/webservice/MetaTASSER	Knowledge-based	MC	Clustering/free-energy	Hours
I-TASSER (Wu et al. 2007; Zhang 2007) http://zhang.bioinformatics.ku.edu/I-TASSER	Knowledge-based	MC	Clustering/free-energy	Hours

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 modelling is to use a compromised force field with a large number of selected atom types; in each atom type, the chemical and physical properties of the atoms are enough alike with the parameters calculated from crystal packing or quantum mechanical theory (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 Jr. 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.

For the study of protein folding, these classical force fields were often coupled with molecular dynamics (MD) simulations. However, the results, from the viewpoint of protein structure prediction, were not quite successful. (See Chapter 10 for the use of MD in elucidation of protein function from known structures). The first milestone in such MD-based *ab initio* protein folding is probably the 1997 work of Duan and Kollman who simulated the villin headpiece (a 36-mer) in explicit solvent for six months on parallel supercomputers. Although the authors did not fold the protein with high resolution, the best of their final model was within 4.5 Å to the native state (Duan and Kollman 1998). With Folding@Home, a worldwide-distributed computer system, this small protein was recently folded by Pande and coworkers (Zagrovic et al. 2002) to 1.7 Å with a total simulation time of 300 ms or approximately 1,000 CPU years. 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), not to mention the fact that the validity/accuracy has not yet been systematically tested even for a number of small proteins.

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 them 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 less 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) 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 root-mean-square deviation (RMSD) using CHARMM (Brooks et al. 1983) and the TIP3P water model (Jorgensen et al. 1983).

Later, using AMBER 5.0 (Case et al. 1997) and 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 is achieved (Lee et al. 2001). 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. Recently, Chen and Brooks (2007) used CHARMM22 (MacKerell Jr. 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 are used to guide the refinement procedure (Chen and Brooks 2007).

A noteworthy observation is recently 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 outperforms 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, fast search methods (such as Monte Carlo simulations and genetic algorithms) based on physics-based potentials have shown to be promising in both structure prediction and structure refinement. One example is the ongoing project by Scheraga and coworkers (Liwo et al. 1999, 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 of UNRES with the 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 centre. 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 2–10h. 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 the authenticity of the true *ab initio* nature of their approach. Nevertheless, this method is probably the most faithful *ab initio* method available (in terms of the application of a thorough global optimization to a physics-based energy function) and it has been systematically applied to many CASP targets since 1998. The most notable prediction success by this approach is for T061 from CASP3, for which a model of 4.2Å RMSD to the native for a 95-residue α-helical protein was generated with an accuracy gap from the rest of models by others. It is shown, for the first time in a clear-cut fashion that the *ab initio* method can 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 7.3Å by this approach. However, it seems that the scarcity and the best-but-still-low accuracy of such models by a pure *ab initio* modelling failed to draw much attention from the protein science community, where accurate protein models are in great demand.

Another example of the physics-based modelling 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 two 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 C_α atoms, the final tertiary structure of the full length protein is modeled by globally minimizing 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 C_α 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 for a number of proteins is yet to be seen in the future.

Taylor and coworkers (2008) recently 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 fold a set of five αβ sandwich proteins with length up to 150 residues with the first model within 4–6 Å RMSD of the native structure. Again, although appealing in methodology, the performance of the approach in the open blind experiments and on the proteins of various fold-types is yet to be seen.

In the recent development of ROSETTA (Bradley et al. 2005; Das et al. 2007), a physics-based atomic potential is used in the second stage of Monte Carlo structure refinement following the low-resolution fragment assembly (Simons et al. 1997), which we will discuss in the next section.

1.2.2 Knowledge-Based Energy Function Combined with Fragments

Knowledge-based potential refers to the empirical energy terms derived from the statistics of the solved structures in deposited PDB, which can be divided into two types as described by Skolnick (2006). The first one 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 Sali 2006), and secondary structure propensities (Zhang et al. 2003, 2006; Zhang and Skolnick 2005a).

Although most knowledge-based force fields contain secondary structure propensity propensities, it may be that local protein structures are rather difficult to reproduce in the reduced modelling. 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 found 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. Another advantage of this approach is that the use of excised secondary structure fragment can significantly reduce the entropy of the conformational search.

Here, we introduce two prediction methods utilizing knowledge-based energy functions, which are proved to be the most successful in *ab initio* protein structure prediction (Simons et al. 1997; Zhang and Skolnick 2004a).

One of the best-known ideas for *ab initio* modelling is probably the one pioneered by Bowie and Eisenberg, who generated 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 was extremely successful for the free modelling (FM) targets in CASP experiments and made the fragment assembly approach popular in the field. In the recent developments of ROSETTA (Bradley et al. 2005; Das et al. 2007), the authors first generated models in a reduced form with conformations specified with heavy backbone and C_β 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 modelling is shown in Fig. 1.2 and details on the energy functions can be found in references (Bradley et al. 2005; Das et al. 2007). For the conformational search, multiple rounds of Monte Carlo minimization (Li and Scheraga 1987) are carried out. The most notable example for this two-step protocol is the blind prediction of an *ab initio* target (T0281 from CASP6, 70 residues), whose C_α RMSD from its crystal structure is 1.6 Å (Bradley et al. 2005). In CASP7, 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. There was one target, T0283, which was a template-based modelling (TBM) target but was modeled by the ROSETTA *ab initio* protocol. It generated a model of RMSD = 1.8 Å over 92 residues out of the 112 residues (Fig. 1.3, left panel). 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

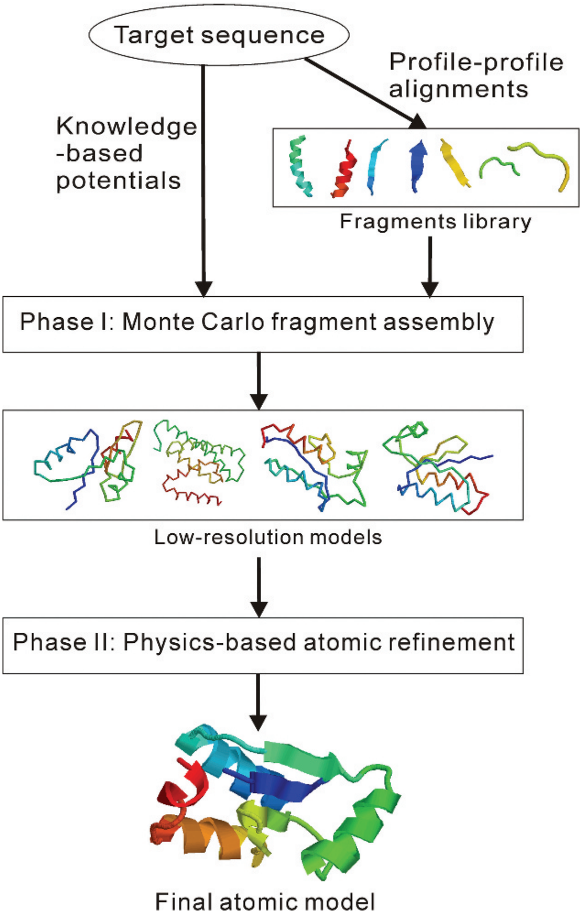
Target sequence
Profile-profile alignments
Knowledge -based potentials
Fragments library
Phase I: Monte Carlo fragment assembly
Low-resolution models
Phase II: Physics-based atomic refinement
Final atomic model

Fig. 1.2 Flowchart of the ROSETTA protocol

controlling the protein radius of gyration. However, their prediction seems to be only partially successful in comparison to ROSETTA.

Another successful free modelling approach, TASSER by Zhang and Skolnick (2004a), 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 *ab initio* modelling (Zhang et al. 2003). The protein conformation in TASSER is represented by a trace of C_α atoms and side chain centres of mass, and the reassembly process is conducted by parallel 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

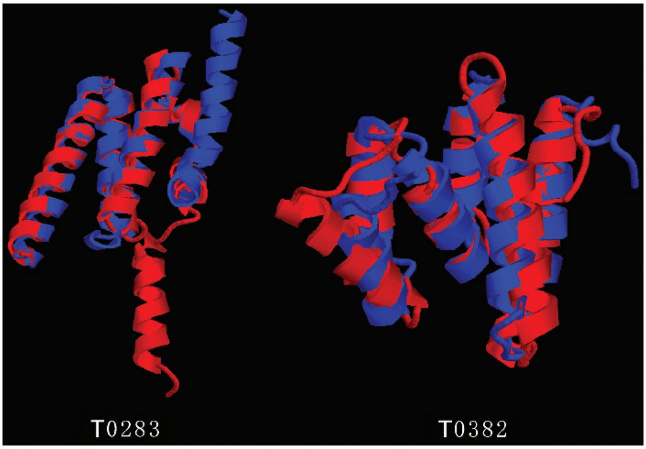

Fig. 1.3 Two examples of successful free modelling from CASP7 are shown. T0283 (left panel) is a TBM target (from *Bacillus halodurans*) of 112 residues; the model was generated by all-atom ROSETTA (a hybrid knowledge- and physics-based approach) (Das et al. 2007) based on free modelling, which gives a TM-score 0.74 (Zhang and Skolnick 2004b) and a RMSD 1.8 Å over the first 92 residues (13.8 Å overall RMSD is due to the wrong orientation of the C-terminal helix). T0382 (right panel) is a FM/TBM target (from *Rhodopseudomonas palustris* CGA009) of 123 residues; the model was generated by I-TASSER (a purely knowledge-based approach) (Zhang 2007) with a TM-score 0.66 and a RMSD 3.6 Å. Blue and red represent the model and the crystal structures, respectively

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.

There are several new developments of TASSER. One is Chunk-TASSER (Zhou and Skolnick 2007) in Skolnick's group, which first splits the target sequences into subunits (or "chunks"), each containing three 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 development is I-TASSER by Wu et al. (2007), which refines TASSER cluster centroids by iterative Monte Carlo simulations. The spatial restraints are extracted from the first round TASSER models and the template structures searched by TM-align (Zhang and Skolnick 2005b) from the PDB library, which are exploited in the second round simulations. The purpose is to remove the steric clashes from the first round models and refine the topology. The flowchart of I-TASSER is shown in Fig. 1.4. 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. In CASP7, out of 19 FM and FM/TBM targets, I-TASSER built models with correct topology (~3–5 Å) for seven cases with sequences up to 155 residues long. Figure 1.3 (right panel) shows the example of T0382 (123 residues) where all initial templates were of

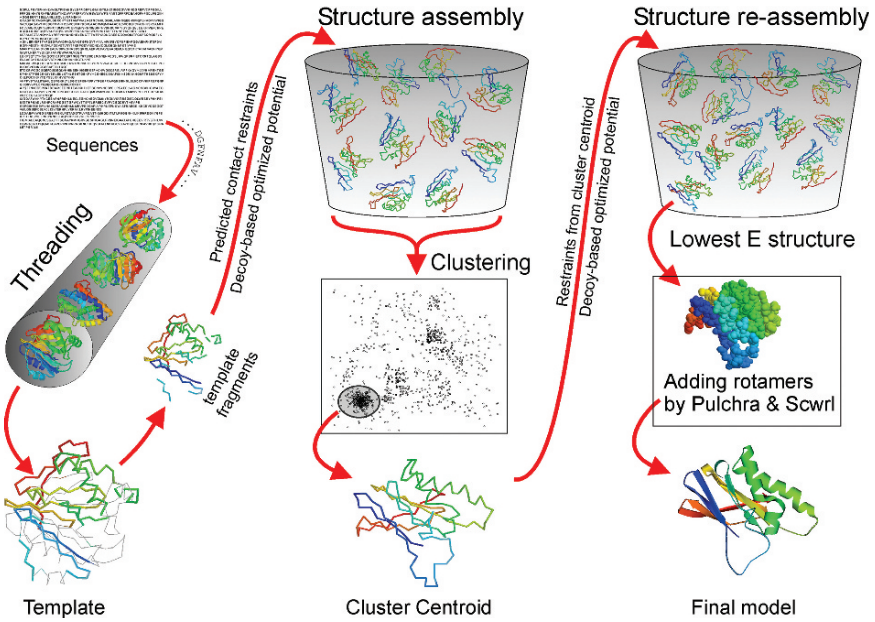
Sequences
...DGFNPAV...
Threading
Template
template fragments
Predicted contact restraints
Decoy-based optimized potential
Structure assembly
Clustering
Cluster Centroid
Restraints from cluster centroid
Decoy-based optimized potential
Structure re-assembly
Lowest E structure
Adding rotamers by Pulchra & Scwrl
Final model

Fig. 1.4 Flowchart of I-TASSER protein structure modelling

wrong topology (>9 Å) but the final model is 3.6 Å away from the X-ray structure. Recently, Helles carried out a comparative study on 18 *ab initio* prediction algorithms and concluded that I-TASSER is about the best method in term of the modelling accuracy and CPU cost per target (Helles 2008).

1.3 Conformational Search Methods

Successful *ab initio* modelling 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 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 MC simulation procedures.

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 modelling procedures. Readers are recommended to read appropriate references for details. Unlike various energy functions used in *ab initio* modelling, the search methods should be, in principle, transferable between protein modelling methods, as well as other problems in science and technology. Currently, there exist 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 performs 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 most 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) (Kihara et al. 2001) 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), originally developed by Li and Scheraga (Li and Scheraga 1987), was successfully applied to the conformational search of ROSETTA's high-resolution energy function. 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 Chapter 10) solves Newton's equations of motion at each step of atom movement, which is probably the most faithful method depicting atomistically what is occurring in proteins. The method is therefore most-often used for the study of protein folding pathways (Duan and Kollman 1998). The long simulation time is one of the major issues of this method, since the incremental time scale is usually in the order of femtoseconds (10^{-15} s) while the fastest folding time of a small protein (less than 100 residues) is in the millisecond range in nature. Currently no serious all-atom MD simulations are attempted for protein structure prediction starting from either an extended or a random initial structure. 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. One notable approach is the recent work of Scheraga and his coworkers, who have implemented torsion space MD simulation with the coarse-grained energy function UNRES (see the discussion above).

1.3.3 Genetic Algorithm

Conformational space annealing (CSA) (Lee et al. 1998) is one of the most successful genetic algorithms. 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* modelling 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 search 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 modelling methods typically generate lots of decoy structures during the simulation. How to select appropriate models structurally close to the native state is an important issue. The selection of protein models has been emerged as a new field called Model Quality Assessment Programs (MQAP) (Fischer 2006). In general, modelling selection approaches can be classified into two types, i.e. 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

$$F(R) = -k_B T \quad \ln Z(R) = -k_B T \ \ln \int e^{-\beta E(R)} d\Omega, \tag{1}$$

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 the clustering procedure at a given RMSD cutoff (Zhang and Skolnick 2004c).

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), which has also called meta-server approaches (Ginalski et al. 2003a; Wu and Zhang 2007). The essence of this method is similar to the clustering approach since both assume the most frequently occurring state as the near-native ones. This approach has been mainly used for selecting models generated by threading-servers (Ginalski et al. 2003; Wallner and Elofsson 2007; Wu and Zhang 2007).

1.4.1 Physics-Based Energy Function

For the development of all-atom physics-based energy functions, Lazaridis and Karplus (1999a) exploited CHARMM19 (Neria et al. 1996) and EEF1 (Lazaridis and Karplus 1999b) 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 (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 (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), and CASP decoys set (Moult et al. 2001)). All these authors obtained similar results, i.e. 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. Recently, 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). 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.

1.4.2 Knowledge-Based Energy Function

Sippl developed 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 proposed, which include atomic interaction potential, solvation potential, hydrogen bond potential, torsion angle potential, etc. In coarse-grained potentials, each residue is represented either by a single atom or by a few atoms, e.g., C_α-based potentials (Melo et al. 2002), C_β-based potentials (Hendlich et al. 1990), side chain centre-based potentials (Bryant and Lawrence 1993; Kocher et al. 1994; Thomas and Dill 1996; Skolnick et al. 1997; Zhang and Kim 2000; Zhang et al. 2004), side chain and C_α-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 and Zhou 2002), Wang et al. (self-RAPDF) (Wang et al. 2004), Tostto (victor/FRST) (Tosatto 2005), and Shen and Sali (DOPE) (Shen and Sali 2006). All these potentials claimed that native structures can be distinguished from decoy structures in their tests. However, the task of selecting the near native models out of many decoys remains as a challenge for these potentials (Skolnick 2006); this is actually more important than native structure recognition because in reality there are no native structures available from computer simulations. Based on the CAFASP4-MQAP experiment in 2004 (Fischer 2006), the best-performing energy functions are 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 C_β atom interaction potential and solvation potential. From CASP7-MQAP in 2006, Pcons developed by Elofsson group based on structure consensus performed best (Wallner and Elofsson 2007).

1.4.3 Sequence-Structure Compatibility Function

In the third type of MQAPs, best models are selected not purely based on energy functions. 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 CC, CN, CO, NN, NO and OO, where near-native structures have fewer errors than other decoys. Verify3D (Eisenberg et al. 1997) improves the method of 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. Recently, 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* modelling approaches. In the work by Shortle et al. (1998), for all 12 cases tested, the cluster-centre conformation of the largest cluster was closer to native structures than the majority of decoys. Cluster-centre 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 2004c). Based on the 1,489 representative benchmark proteins each with up to 280,000 structure decoys, the best of the top five models was ranked as top 1.4% among all decoys. For 78% of the 1,489 proteins, the RMSD difference between the best of the top five models and the most native-like decoy structure was less than 1 Å.

In ROSETTA *ab initio* modelling (Bradley et al. 2005), 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; Wu et al. 2007), thousands of decoy models from MC simulations are clustered by SPICKER (Zhang and Skolnick 2004c) 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* modelling from amino acid sequence alone is considered as the “Holy Grail” of protein structure prediction (Zhang 2008), since this will mark an eventual and complete solution to the problem. Except for the generation of 3D structures, *ab initio* modelling can also help us understand the underlying principles on how proteins fold in nature; this could not be done by the template-based modelling approaches which build 3D models by copying the framework of other solved structures.

An ideal approach to *ab initio* modelling 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 exploiting the classic CHARMM and AMBER force fields. Although the MD based simulation is extremely 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 appear to be progressing rapidly, producing many examples of successful low-to-medium accuracy models often with correct topology for proteins of up to 100 residues. Although very rare, successful higher resolution models (<2 Å of C_α atoms) have also been reported (Bradley et al. 2005).

The current state-of-the-art *ab initio* protein structure prediction methods often utilize as much as possible knowledge-based information from known structures, which is multi-purpose. First, the employment of local structure fragments directly excised from the PDB structures helps reduce the degrees of freedom and the entropy of 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* modelling for proteins up to 100–120 residues has been significantly improved in the last decade.

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 limitations of available search methods can be explored separately.

It is important to acknowledge that *ab initio* prediction methods solely based on the physicochemical principles of interaction are currently far behind, in terms of their modelling speed and accuracy, compared with the methods utilizing bioinformatics and knowledge-based information. However, the physics-based atomic potentials have proven to be useful in refining the detailed packing of the side chain atoms and the peptide backbones. Thus, developing composite methods using both knowledge-based and physics-based energy terms may represent a promising approach to the problem of *ab initio* modelling.

Acknowledgements The project is supported in part by KU Start-up Fund 06194, the Alfred P. Sloan Foundation, and Grant Number R01GM083107 of the National Institute of General Medical Sciences.

References

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

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

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

Berman HM, Westbrook J, Feng Z, et al. (2000) The protein data bank. Nucleic Acids Res 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 Bioinformatics 4:8

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 USA 91(10):4436–4440

Bradley P, Misura KM, Baker D (2005) 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. J Comput Chem 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

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

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

Cheng J, Baldi P (2006) A machine learning information retrieval approach to protein fold recognition. Bioinformatics 22(12):1456–1463

Colovos C, Yeates TO (1993) Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci 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. J Am Chem Soc 117:5179–5197

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

Dominy BN, Brooks CL (2002) Identifying native-like protein structures using physics-based potentials. J Comput Chem 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. Method Enzymol 277:396–404

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

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

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. Curr Opin Struct Biol 16(2):178–182

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. (2003a) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 19(8):1015–1018

Ginalski K, Pas J, Wyrwicz LS, et al. (2003b) ORFeus: detection of distant homology using sequence profiles and predicted secondary structure. Nucleic Acids Res 31(13):3804–3807

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. J Am Chem Soc 96:5319–5327

Helles G (2008) A comparative study of the reported performance of ab initio protein structure prediction algorithms. J R Soc 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. J Mol Biol 216(1):167–180

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 (2003) Generalized born model with a simple smoothing function. J Comput Chem 24(14):1691–1702

Jaroszewski L, Rychlewski L, Li Z, et al. (2005) FFAS03: a server for profile–profile sequence alignments. Nucleic Acids Res 33(Web Server issue):W284–288

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. Comput Biol Chem 30(5):360–366

Jones DT (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 287(4):797–815

Jorgensen WL, Tirado-Rives J (1988) The OPLS potential functions for proteins. Energy minimizations for crystals of cyclic peptides and crambin. J Am Chem Soc (110):1657–1666

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. J Am Chem Soc 118:11225–11236

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 USA 98(18):10125–10130

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. Biophys J 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. J Mol Biol 235(5):1598–1613

Lazaridis T, Karplus M (1999a) Discrimination of the native from misfolded protein models with an energy function including implicit solvation. J Mol Biol 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. Phys Rev Lett 71(2):211–214

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 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 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. J Mol Biol 313(2):417–430

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. Comput Phys Commun 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 USA 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 Model 7:306–317

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 Sci 2(10):1697–1714

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

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 USA 102(7):2362–2367

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 Jr. AD, 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

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 Sci 11(2):430–448

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

Nilges M, Brunger AT (1991) Automated modeling of coiled coils: application to the GCN4 dimerization region. Protein Eng 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. Proc Natl Acad Sci USA 102(21):7547–7552

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

Petrey D, Honig B (2000) Free energy determinants of tertiary structure and the evaluation of protein models. Protein Sci 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

Pieper U, Eswar N, Davis FP, et al. (2006) MODBASE: a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res 34(Database issue):D291–295

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

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

Shi J, Blundell TL, Mizuguchi K (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 310(1):243–257

Shortle D, Simons KT, Baker D (1998) Clustering of low-energy conformations near the native structures of small proteins. Proc Natl Acad Sci USA 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. J Mol Biol 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. J Mol Biol 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. Curr Opin Struct Biol 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, Zhang Y, Arakaki AK, et al. (2003) TOUCHSTONE: a unified approach to protein structure prediction. Proteins 53(Suppl 6):469–479

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

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. Biophys J 88(4):2472–2493

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

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? J Mol Biol 257(2):457–469

Tosatto SC (2005) The victor/FRST function for model quality estimation. J Comput Biol 12(10):1316–1327

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. VDF Hochschulverlag AG an der ETH, Zurich.

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

Wallner B, Elofsson A (2003) Can correct protein models be identified? Protein Sci 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? JComput Chem 21(12):1049–1074

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

Weiner SJ, Kollman PA, Case DA, et al. (1984) A new force field for molecular mechanical simulation of nucleic acids and proteins. J Am Chem Soc 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. J Comput Chem 28(12):2059–2066

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

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

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

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. J Mol Biol 323(5):927–937

Zhang C, Kim SH (2000) Environment-dependent residue contact energies for proteins. Proc Natl Acad Sci USA 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 Sci 13(2):400–411

Zhang Y (2007) Template-based modeling and free modeling by I-TASSER in CASP7. Proteins 69(Suppl 8):108–117

Zhang Y (2008) Progress and challenges in protein structure prediction. Curr Opin Struct Biol 18(3):342–348

Zhang Y, Skolnick J (2004a) Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci U S A 101:7594–7599

Zhang Y, Skolnick J (2004b) Scoring function for automated assessment of protein structure template quality. Proteins 57:702–710

Zhang Y, Skolnick J (2004c) SPICKER: a clustering approach to identify near-native protein folds. J Comput Chem 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 USA 102:1029–1034

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

Zhang Y, Kihara D, Skolnick J (2002) Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding. Proteins 48(2):192–201

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

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 USA 103:2605–2610

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

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 Sci 11(11):2714–2726

Zhou H, Zhou Y (2005) Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins 58(2):321–328