Chapter 11
Protein Structure Prediction

Sitao Wu and Yang Zhang

11.1 Introduction

Owing to significant efforts in genome sequencing over nearly three decades (McPherson et al. 2001; Venter et al. 2001), gene sequences from many organisms have been deduced. Over 100 million nucleotide sequences from over 300 thousand different organisms have been deposited in the major DNA databases, DDBJ/EMBL/GenBank (Benson et al. 2003; Miyazaki et al. 2003; Kulikova et al. 2004), totaling almost 200 billion nucleotide bases (about the number of stars in the Milky Way). Over 5 million of these nucleotide sequences have been translated into amino acid sequences and deposited in the UniProtKB database (Release 12.8) (Bairoch et al. 2005). The protein sequences in UniParc triple this number. However, the protein sequences themselves are usually insufficient for determining protein function as the biological function of proteins is intrinsically linked to three dimensional protein structure (Skolnick et al. 2000).

The most accurate structural characterization of proteins is provided by X-ray crystallography and NMR spectroscopy. Owing to the technical difficulties and labor intensiveness of these methods, the number of protein structures solved by experimental methods lags far behind the accumulation of protein sequences. By the end of 2007, there were 44,272 protein structures deposited in the Protein Data Bank (PDB) (www.rcsb.org) (Berman et al. 2000) – accounting for just one percent of sequences in the UniProtKB database (http://www.ebi.ac.uk/swissprot). Moreover, the gap between the number of protein sequences and the number of structures has been increasing as indicated in Fig. 11.1.

One of the major efforts in protein structure determination in recent years is the structural genomics (SG) project initiated at the end of last century (Sali 1998; Terwilliger et al. 1998; Burley et al. 1999; Smaglik 2000; Stevens et al. 2001). The SG project aims to obtain 3D models of all proteins by an optimized combination of experimental

Y. Zhang (✉)
Center for Bioinformatics and Department of Molecular Bioscience, University of Kansas, Lawrence, KS, 66047
e-mail: yzhang@ku.edu

D. Edwards et al. (eds.), *Bioinformatics*: *Tools and Applications,*
DOI 10.1007/978-0-387-92738-1_11, © Springer Science+Business Media, LLC 2009

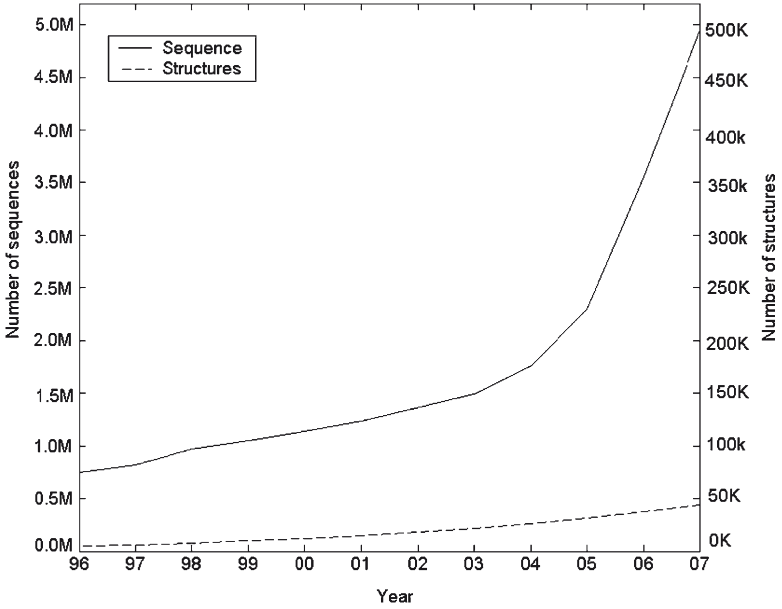

Fig. 11.1 Determination of amino acid sequences (left-hand scale) is outpacing that of 3D structures (right-hand scale) by a factor of 100. Data are taken from PDB (Berman et al. 2000) and UniProtKB (Bairoch et al. 2005)

structure determination and comparative model (CM) building (Pieper et al. 2006). One of the key aspects of the SG project is the selection of key target proteins for structure determination, so that the majority of sequences can be within a CM distance to solved structures. Using a sequence identity of 30% with 80% alignment coverage as the CM distance cutoff, Vitkup et al. (2001) estimated that at least 16,000 new structures need to be determined by experiments to ensure that the CM represents 90% of protein domain families. Without optimal coordination of target selection, as many as 50,000 structure determinations may be required.

Currently, 36% of Pfam families (Bateman et al. 2004) contain at least one member with the solved structure, allowing comparative modeling of other family members. According to Chandonia and Brenner (Chandonia and Brenner 2006), the SG project solved 1,887 protein structures between 2000 and 2005, 294 of which are the first solved structures in their respective Pfam families. During 2004, around half of the PDB structures with new Pfam family annotations were because of the efforts of the SG centers (Chandonia and Brenner 2006). Determination of these new Pfam structures has dramatically extended the range of computer-based predictions using Comparative Model (CM) techniques (Sali 1998; Pieper et al. 2006). For example, based on 53 newly solved proteins from SG projects, Sali and coworkers (Pieper et al. 2004) built reliable models for domains in 24,113 sequences from the UniProtKB database with their CM tool MODELLER (Sali and Blundell 1993). These models have been deposited in a comprehensive CM model database, MODBase (http://salilab.org/modbase). In February 2008, MODBase contained around 4.3 million models or fold assignments for domains from 1.34 million sequences. In this study, the structure assignments were based on an all-against-all search of the amino acid sequences in UniProtKB using the solved protein structures in PDB (Berman et al. 2000). Structural genomics can also benefit from improvements in high-resolution structure prediction algorithms. Vitkup et al. (2001) estimated that "a 10% decrease in the threshold needed for accurate modeling, from 30 to 20% sequence identity, would reduce the number of experimental structures required by more than a factor of two".

There are two critical problems in the field of protein structure prediction. The first problem is related to the template-based modeling: How to identify the most suitable templates from known protein structures in the PDB library? Furthermore, following template structure identification, how can the template structures be refined to better approximate the native structure? The second major problem is related to free modeling for the target sequences without appropriate templates: How can a correct topology for the target proteins be constructed from scratch? Progress made in these areas has been assessed in recent CASP7 experiments (Moult et al. 2007) under the categories of template based modeling (TBM) and free modeling (FM), respectively.

In the following sections, current protein structure prediction methods will be reviewed for both template-based modeling and free modeling. The basic ideas and advances of these directions will be discussed in detail.

11.2 Template-Based Predictions

For a given target sequence, template-based prediction methods build 3D structures based on a set of solved 3D protein structures, termed the template library. The canonical procedure of template-based modeling consists of four steps: (1) finding known structures (templates) related to the sequence to be modeled (target); (2) aligning the target sequence on the template structures; (3) building the structural framework by copying the aligned regions, or by satisfying spatial restraints from templates; (4) constructing the unaligned loop regions and adding side-chain atoms. The first two steps are usually performed as a single procedure because the correct selection of templates relies on their accurate alignment with the target. Similarly, the last two steps are also performed simultaneously since the atoms of the core and loop regions interact closely.

Historically, template-based methods can be categorized into two types: (1) comparative modeling (CM) and (2) threading. CM builds models based on evolutionary information between target and template sequences, while threading is designed to match target sequences directly onto 3D structures of templates with the goal to detect target-template pairs even without evolutionary relationships. The schematic overview of CM and threading is depicted in the upper part of Fig. 11.2. In recent years, as a general trend in the field, the borders between CM and threading are becoming increasingly blurred since both comparative modeling and threading methods rely on evolutionary relationships, e.g. both use sequence profile-based

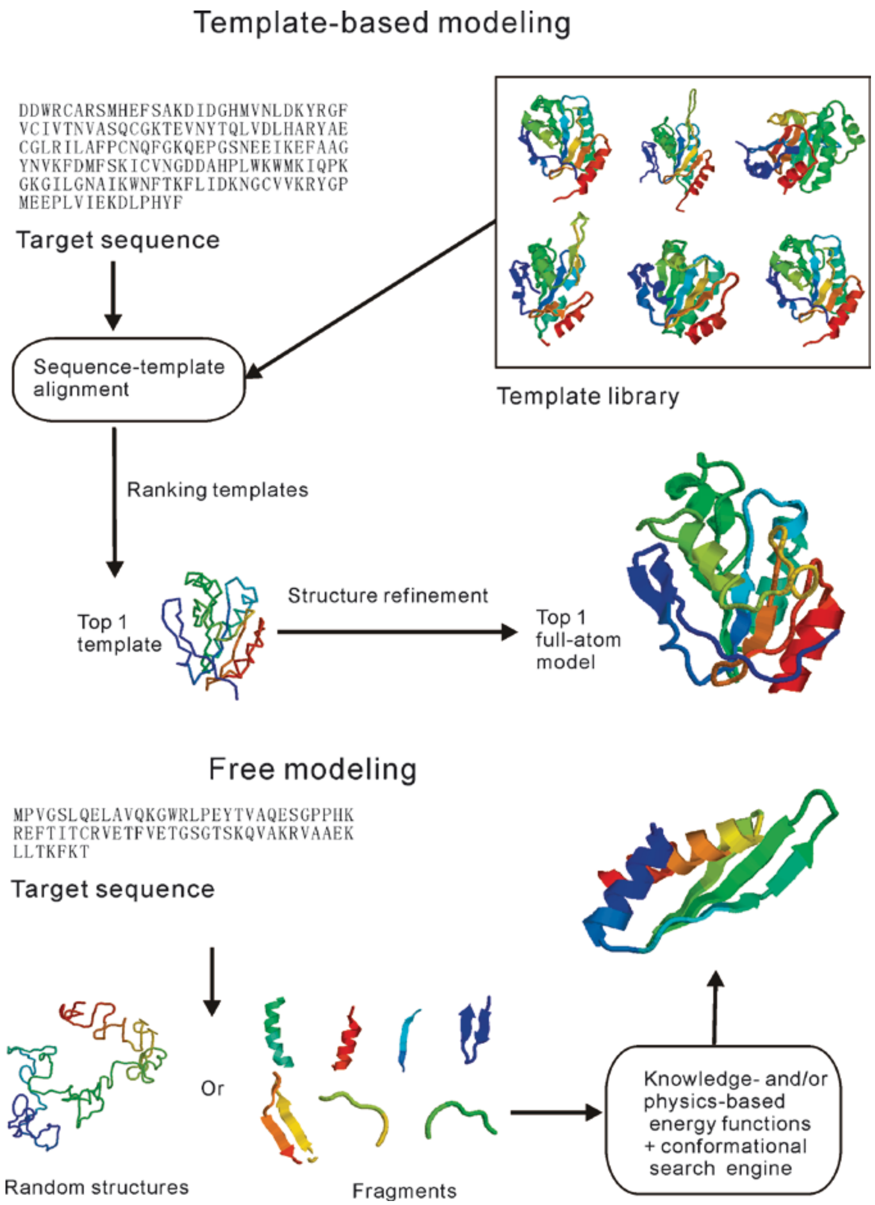
Template-based modeling
DDWRCARSMHEFSAKDIDGHMVNLDKYRGF
VCIVTNVASQCGKTEVNYTQLVDLHARYAE
CGLRILAFPCNQFGKQEPGSNEEIKEFAAG
YNVKFDMFSKICVNGDDAHPLWKWMKIQPK
GKGILGNAIKWNFTKFLIDKNGCVVKRYGP
MEEPLVIEKDLPHYF
Target sequence
Template library
Sequence-template alignment
Ranking templates
Top 1 template
Structure refinement
Top 1 full-atom model
Free modeling
MPVGSLQELAVQKGWRLPEYTVAQESGPPHK
REFTITCRVETFVETGSGTSKQVAKRVAAEK
LLTKFKT
Target sequence
Or
Random structures
Fragments
Knowledge- and/or physics-based energy functions + conformational search engine

Fig. 11.2 Schematic overview of the methodologies employed in template-based and free modeling

alignments (Marti-Renom et al. 2000; Skolnick et al. 2004; Zhou and Zhou 2005; Wu and Zhang 2008). In this chapter, we put them in the same category of template-based modeling without explicitly distinguishing them unless necessary.

11.2.1 Completeness of the PDB Template Library

The existence of similar structures to the target in the PDB is a precondition for successful template-based modeling. An important concern is thus the completeness of the current PDB structure library. Figure 11.3 shows a distribution of the best templates found by the structural alignment (Zhang and Skolnick 2005b) for 1,413 representative single-domain proteins between 80 and 200 residues.

Remarkably, even excluding the homologous templates of sequence identity, >20%, all the proteins have at least one structural analog in the PDB with a C_α root-mean-squared deviation (RMSD) to the target <6 Å covering >70% of regions. The average RMSD and coverage are 2.96 Å and 86% respectively. Zhang and Skolnick (2005a,b) recently showed that high quality full-length models can be built for all the single-domain proteins with an average RMSD of 2.25 Å when using the best possible templates in the PDB. These data demonstrate that the structural universe of the current PDB library is likely to be complete for solving the protein structure for at least single-domain proteins. However, most of the target-template pairs have only around 15% sequence identity, which are difficult to

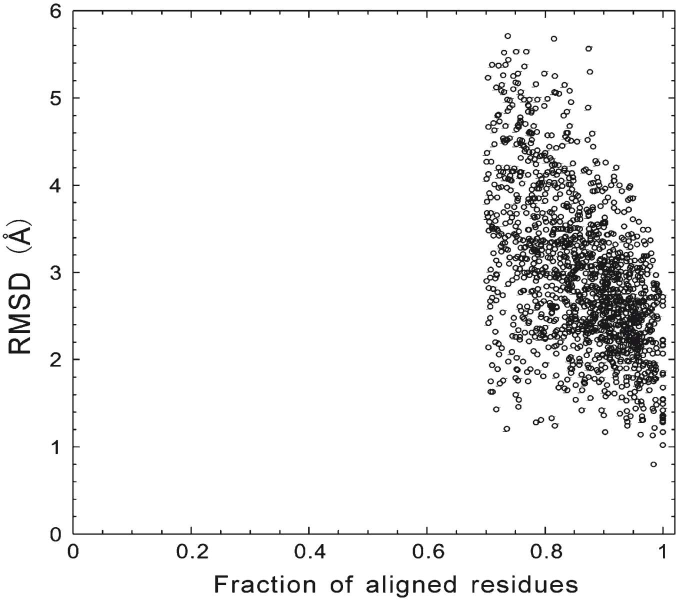

Fig. 11.3 Structural superposition results of 1,413 representative single-domain proteins on their structural analogs in the PDB library. The structural analogs are found using a sequence-independent structural-alignment tool, TM-align (Zhang and Skolnick 2005b), and ranked by a TM-score (a structural similarity measure balancing RMSD and coverage) (Zhang and Skolnick 2004b). All structural analogs with a sequence identity >20% to the target are excluded. If the analog with the highest TM-score has a coverage below 70%, the first structural analog with the coverage >70% is presented. As a result, all the structural analogs have a root-mean-squared deviation (RMSD) <6 Å; 80% have a RMSD <4 Å with >75% of regions covered

recover by current threading approaches. In fact, after excluding templates with a sequence identity >30%, current threading techniques could only assign templates with correct topologies (average RMSD ~4 Å) to 2/3 of the proteins (Skolnick et al. 2004). Here, the role of structural genomics is to bridge the target-template gap for the remaining 1/3 proteins as well as improve the alignment accuracy of the 2/3 proteins by providing evolutionarily closer template proteins.

11.2.2 Template Structure Identification Using Threading Programs

Since its first application in the early 1990s (Bowie et al. 1991; Jones et al. 1992), threading has become one of the most active areas in proteins structure prediction. Numerous algorithms have been developed during the previous 15 years for the purpose of identifying structure templates from the PDB. Threading techniques include sequence profile–profile alignments (Ginalski et al. 2003; Skolnick et al. 2004; Jaroszewski et al. 2005; Zhou and Zhou 2005), structural profile alignments (Shi et al. 2001), hidden Markov models (HMM) (Karplus et al. 1998; Soding 2005), and machine learning (Jones 1999; Cheng and Baldi 2006) among others.

The sequence profile–profile alignment (PPA) is probably the most often-used and robust threading approach. Instead of matching the single sequences of target and template, PPA aligns a target multiple sequence alignment (MSA) with a template MSA. The alignment score in the PPA is usually calculated as a product of the amino-acid frequency at each position of the target MSA with the log-odds of the matching amino-acid in the template MSA, though there are also alternative methods for calculating the profile–profile alignment scores (Sadreyev and Grishin 2003). Profile–profile alignment based methods demonstrated advantages in several recent blind tests (Fischer et al. 2003; Rychlewski and Fischer 2005; Battey et al. 2007). In LiveBench-8 (Rychlewski and Fischer 2005), for example, the top four servers (BASD/MASP/MBAS, SFST/STMP, FFAS03, and ORF2/ORFS) are all based on sequence profile–profile alignment. In CAFASP (Fischer et al. 2003) and the recent CASP Server Section (Battey et al. 2007), several sequence profile based methods were ranked at the top of single threading servers. Wu and Zhang (2008) recently showed that the accuracy of the sequence profile–profile alignments can be further improved by about 5–6% by incorporating a variety of additional structural information.

In CASP7, HHsearch (Soding 2005), a HMM–HMM alignment method, was distinguished as the best single threading server. The principles of the HMM–HMM alignments and the profile–profile alignments are similar in that both attempt pair-wise alignments of the target MSA with the template MSA. Instead of representing the MSAs by sequence profiles, HHsearch uses profile HMMs which can generate the sequences with certain probabilities determined by the product of the amino acid emission and insertion/deletion probabilities. HHsearch aligns the target and template HMMs by maximizing the probability that two models co-emit the same amino acid sequence. In this way, amino acid frequencies and insertions/deletions of both HMMs are matched in an optimum way (Soding 2005).

11.2.3 Consensus of Various Threading Programs: Meta-Servers

Although average performance differs among threading algorithms, there is no single threading program which outperforms all others on every target. This naturally leads to the popularity of the meta-server (Fischer 2003; Wu and Zhang 2007), which collects and combines results from a set of existing threading programs. There are two ways to generate predictions in meta-servers. One is to build a hybrid model by cutting and pasting the selected structure fragments from the templates identified by threading programs (Fischer 2003). The combined model has on average larger coverage and better topology than any single template. One defect is that the hybrid models often have non-physical local clashes. The second approach is to select the best models based on a variety of scoring functions or machine-learning techniques. This approach has emerged as a new research area called Model Quality Assessment Programs (MQAP) (Fischer 2006). Despite considerable efforts in developing various MQAP scores, the most robust score turns out to be the one based on the structure consensus, i.e. the best models are those simultaneously hit by different threading algorithms. The idea behind the consensus approach is simple: there are more ways for a threading program to select a wrong template than a right one. Therefore, the chance for multiple threading programs working collectively to make a commonly wrong selection is lower than the chance to make a commonly correct selection.

The meta-server predictors have dominated the server predictions in previous experiments (e.g. CAFASP4, Livebench8, and CASP6). However, in the recent CASP7 experiment (Battey et al. 2007) Zhang-Server (an automated server based on profile–profile threading and I-TASSER structure refinement (Wu et al. 2007; Zhang 2007)) clearly outperforms others (including the meta-servers which included it as an input (Wallner and Elofsson 2007)). A list of the top ten automated servers in the CASP7 experiment is shown in Table 11.1. This data, highlights the challenge of the MQAP methods in correctly ranking and selecting the best models; while the success of the composite threading plus refinement servers (as Zhang-Server, ROBETTA, and MetaTasser) demonstrates the advantage of the structure refinement in the TBM prediction, which is discussed in the next section.

11.2.4 Template Structure Assembly/Refinement

The goal of protein structure assembly/refinement is to draw the templates closer to the native structure. This has proven to be a non-trivial task. Until only a few years ago, most of the TBM procedures either kept the templates unchanged or drove the templates away from the native structures.

Early efforts on template structure refinement have relied on molecular dynamics (MD) based atomic-level simulations; these attempt to refine low-resolution models using classic MD programs such as AMBER and CHARMM. However, with the

Table 11.1 Top 10 servers in CASP7 as ranked by the accumulative GDT-TS score. Multiple servers from the same lab are represented by the highest rank one

Servers	# of targets	GDT-TS	Server type and URL address
Zhang-Server	124	76.04	Threading, refinement and free modeling http://zhang.bioinformatics.ku.edu/I-TASSER
HHpred2	124	71.94	HMM–HMM alignment (single threading server) http://toolkit.tuebingen.mpg.de/hhpred
Pmodeller6	124	71.69	Meta threading server http://pcons.net
CIRCLE	124	71.09	Meta threading server http://www.pharm.kitasato-u.ac.jp/fams/fams.html
ROBETTA	123	70.87	Threading, refinement and free modeling http://robetta.org/submit.jsp
MetaTasser	124	70.77	Threading, refinement and free modeling http://cssb.biology.gatech.edu/skolnick/webservice/MetaTASSER
RAPTOR-ACE	124	69.70	Meta threading server http://ttic.uchicago.edu/~jinbo/RAPTOR_form.htm
SP3	124	69.38	Profile–profile alignment (single threading server) http://sparks.informatics.iupui.edu/hzhou/anonymous-fold-sp3.html
beautshot	124	69.26	Meta threading server http://inub.cse.buffalo.edu/form.html
UNI-EID-expm	121	69.13	Profile–profile alignment (single threading server) (not available)

exception of some isolated instances, this approach has not achieved systematic improvements.

Encouraging template refinements have been achieved by combining the knowledge- and physics-based potentials with spatial restraints from templates (Zhang and Skolnick 2005a; Misura et al. 2006; Chen and Brooks 2007). Misura et al. (2006) first built low-resolution models with ROSETTA (Simons et al. 1997) using a fragment library enriched by the query-template alignment. The C_β-contact restraints are used to guide the assembly procedure, and the low-resolution models are then refined by a physics-based atomic potential. As a result, in 22 out of 39 test cases, the ten lowest-energy models were found closer to the native structure than the template.

A more comprehensive test of the template refinement procedure, based on TASSER simulation combined with consensus spatial restraints from multiple templates, was reported by Zhang and Skolnick (2004a,b, 2005a,b). For 1,489 test cases, TASSER reduced the RMSD of the templates in the majority of cases, with an average RMSD reduction from 6.7 Å to 4.4 Å over the threading-aligned regions. Even starting from the best templates identified by the structural alignment, TASSER refines the models from 2.5 Å to 1.88 Å in the aligned regions. Here, TASSER built the structures based on a reduced model (specified by C_α and side-chain center of mass) with a purely knowledge-based force field. One of the major contributions to these refinements is the use of multiple threading templates, where the consensus restraint is more accurate than that from the individual template.

In addition, the composite knowledge-based energy terms have been extensively optimized using large-scale structure decoys (Zhang et al. 2003) which helps coordinate the complicated correlations of different interaction terms.

The recent CASP7 experiment assessed the progress of threading template refinements. The assessment team compared the predicted models with the best possible structural template (or "virtual predictor group") and commented that "The best group in this respect (24, Zhang) managed to achieve a higher GDT-TS score than the virtual group in more than half the assessment units and a higher GDT-HA score in approximately one-third of cases." (Kopp et al. 2007) This comparison may not entirely reflect the template refinement ability of the algorithms because the predictors actually start from threading templates rather than the best structural alignments; the latter requests the information of the native structures, which were not available when the predictions were made. On the other hand, a global GDT score comparison may favor the full-length model because the template alignment has a shorter length than the model. In a direct comparison of the RMSD over the same aligned regions, we found that the first I-TASSER model is closer to the native than the best initial template in 86 out of 105 TBM cases, while the other 13 (6) cases are worse than (or equal to) the template. The average RMSD is 4.9 Å and 3.8 Å for the templates and models, respectively, over the same aligned regions (Zhang 2007).

11.3 Free Modeling

When structural analogs do not exist in the PDB library or could not be detected by threading (which is more often the case as demonstrated by the data shown in Fig. 11.3), the structure prediction has to be generated from scratch. This type of prediction has been termed ab initio or de novo modeling, a term that may be easily understood as modeling "from first principle". Since CASP7, it is termed *free modeling*, which more appropriately reflects the status of the field, since the most efficient methods in this category still consider hybrid approaches including both knowledge-based and physics-based potentials. Evolutionary information is often used in generating sparse spatial restraints or identifying local structural building blocks.

11.3.1 Physics-Based Free Modeling

Compared to template-based approaches, the purely physics-based ab initio methods – all-atom potential functions, like AMBER (Weiner et al. 1984), CHARMM (Brooks et al. 1983) and OPLS (Jorgensen and Tirado-Rives 1988), combined with molecular dynamics (MD) conformational sampling – have been less successful in protein structure prediction. Significant efforts have been made on the purely physics-based protein folding. The first widely recognized milestone of successful ab initio protein folding is the 1997 work of Duan and Kollman, who folded the *villin headpiece* (a 36-mer). This work used MD simulations in explicit solvent for 2 months on parallel supercomputers with models up to 4.5 Å (Duan and Kollman 1998). With the help of the worldwide-distributed computers, 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 μs or approximately 1,000 CPU years. Despite this remarkable effort, physics-based folding is far from routine for general protein structure prediction of normal size proteins, mainly because of the prohibitive computing demand.

Another niche for physics-based simulation is protein-structure refinement. This approach starts from low-resolution structures with the goal to draw the initial models closer to the native structure. Because the starting models are usually not far away from the native, the conformational change is relatively small and the simulation time is much less than in ab initio folding. One of the earliest MD-based protein structure refinements was for the GCN4 leucine zipper (a 33 residue dimer) (Nilges and Brunger 1991; Vieth et al. 1994). In that work, a low resolution coiled-coil dimer structure (2 ~ 3 Å) was first assembled using Monte Carlo simulation. With the help of the helical dihedral-angle restraints, Skolnick and coworkers (Vieth et al. 1994) refined the GCN4 structure with a backbone RMSD below 1 Å using CHARMM (Brooks et al. 1983) with the TIP3P water model (Jorgensen et al. 1983). Using AMBER 5.0 (Case et al. 1997) and the same explicit water model (Jorgensen et al. 1983), Lee et al. (2001) attempted to refine 360 low-resolution models generated using ROSETTA (Simons et al. 1997) for 12 small proteins (<75 residues), but concluded that there was no systematic structure improvement (Lee et al. 2001). Later, 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 reported that 11/60 models had 10% RMSD reduction and 18/60 had increased RMSD after refinement. Recently, Chen and Brooks (2007) used CHARMM22 (MacKerell et al. 1998) to refine five CASP6 CM targets with lengths in the 70–144 residue range. In four cases, considerable refinements with up to 1 Å RMSD reduction were achieved. One of the major differences of this work is that an implicit solvent force field based on the generalized Born (GB) approximation (Im et al. 2003) was exploited, which significantly speeds up the MD simulations, while the spatial restraints extracted from the initial models are used to guide the refinement procedure (Chen and Brooks 2007). A particularly noteworthy observation was recently made by Summa and Levitt (Summa and Levitt 2007) who exploited different 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)) on the refinement of 75 proteins by in vacuo energy minimization. The authors found that a knowledge-based atomic contact potential outperforms all the traditional MM potentials in moving almost all the test proteins closer to the native state, while all the MM potentials, except for AMBER99, essentially drive the decoys away from the native. The vacuum simulation without solvation may be part of the reason for the failure of the MM potentials. But this observation demonstrates the potential of combining knowledge-based potentials with physics-based force field in protein structure refinement.

Another use of the physics-based potential is in the discrimination of the native/near-native structures from structure decoys. For example, Lazaridis and Karplus (1999) exploited CHARMM19 (Neria et al. 1996) and EEF1 (Lazaridis and Karplus 1999) solvation potential to discriminate the native structure from the decoys generated by threading the native sequences on other protein structures. They found the energy of the native states is lower than that of the decoys in most cases. Later, Dominy and Brooks (2002), and Feig and Brooks (2002) used CHARMM plus GB, 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 the Park–Levitt decoy set (Park and Levitt 1996), Baker decoy set (Tsai et al. 2003), Skolnick decoy set (Kihara et al. 2001; Zhang et al. 2003), and CASP decoys set (Moult et al. 2001). Similar results were obtained by all the authors, i.e. the native structure can be distinguished from non-native decoys by the physics-based potentials. Recently, however, 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, none of the native structures have lower energy than decoys, and the energy-RMSD correlation was close to zero. This result partially explains the widely-reported discrepancy between the decoy-discrimination ability of the physics-based potentials and less-successful folding/refinement results (Wroblewska and Skolnick 2007).

In contrast, fast Monte Carlo simulations on the physics-based potentials have enjoyed considerable success in both protein structure prediction and refinement. For example, Scheraga and coworkers (Liwo et al. 1999) successfully built models of 4.2 Å for a fragment of 61 residues based on the MC optimization of a physics-based united-residue force field (Liwo et al. 1993) combined with the atomic ECEPP potential (Nemethy et al. 1992). Using ASTRO-FOLD (Klepeis and Floudas 2003) on the ECEPP optimization, Floudas and coworkers (Klepeis et al. 2005) constructed a model of 5.2 Å for a four-helical bundle protein of 102 residues. In the recent development of ROSSETA (Bradley et al. 2005; Das et al. 2007), the authors also cooperated the physics-based atomic potential in the final stage of Monte Carlo structure refinement, which is discussed in the next section.

11.3.2 Knowledge-Based Free Modeling

Probably the most well-known approach for efficient free-modeling was pioneered by Bowie and Eisenberg, who assembled new tertiary structures using small fragments (mainly 9-mers) cut from other PDB proteins (Bowie and Eisenberg 1994). Based on this idea, Baker and coworkers later developed ROSETTA (Simons et al. 1997), which works extremely well for free modeling in the CASP experiments, and popularized the fragment assembly approach in the field. In new developments with ROSETTA (Das et al. 2007), the authors first assemble structures in a reduced knowledge-based model with conformations specified by the heavy backbone atoms and C_β. In the second stage, Monte Carlo simulations with an all-atom physics-based potential are performed to refine the details of the low-resolution models. An exciting achievement was demonstrated in CASP6 by generating a model for a small hard target T0281 (70 residues) that is 1.6 Å away from the crystal structure. In CASP7, the atomic ROSETTA built a model for T0283 (112 residues) with RMSD = 1.8 Å over 92 residues (see Fig. 11.4). Despite significant success, the computer cost of the procedure (~150 CPU days for a small protein <100 residues) is still too expensive for routine use.

Another successful free modeling approach, called TASSER by Zhang and Skolnick (2004a,b), constructs 3D models based on a purely knowledge-based approach. Continuous fragments with various sizes are excised from threading alignments and used to reassemble protein structures in an on-and-off lattice system. A newer version of I-TASSER was recently developed by Wu et al. (2007), which refines the TASSER cluster centroids by iterative Monte Carlo simulations. Although the procedure uses structural fragments and spatial restraints from threading

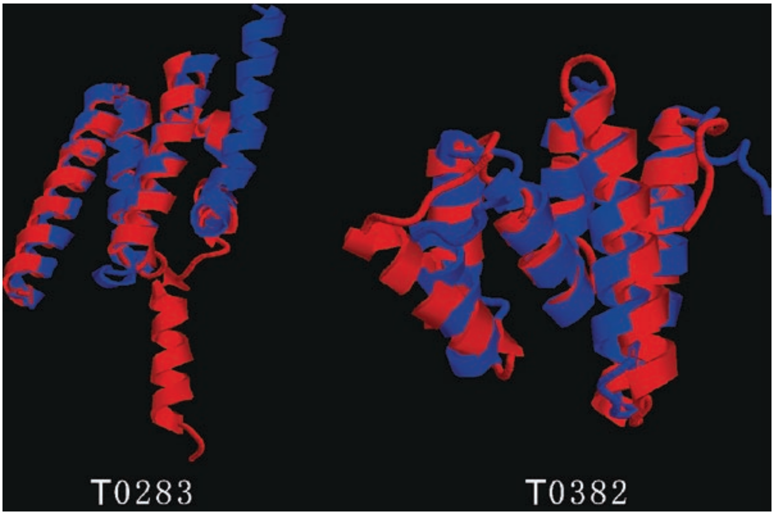

Fig. 11.4 Representative examples of free modeling in CASP7 generated by two different approaches. T0283 (left panel) is a TBM target (from *Bacillus halodurans*) of 112 residues; but the model is generated by all-atom ROSETTA (a hybrid knowledge- and physics-based approach) (Das et al. 2007) based on free modeling, which gives a TM-score 0.74 and a RMSD 1.8 Å over the first 92 residues (the overall RMSD is 13.8 Å mainly because of the misorientation of C-terminal). T0382 (right panel) is a FM/TBM target (from *Rhodopseudomonas palustris* CGA009) of 123 residues; the model is 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 structure representatively

templates, it often constructs models of correct topology even when the topologies of individual templates are incorrect. In CASP7 (Zhang 2007), among 19 FM and FM/TBM targets, I-TASSER builds correct topology (~3–5 Å) for 7 cases with sequences up to 155 residues long. In the right panel of Fig. 11.4, we show an example of T0382 (123 residues), where all initial templates have incorrect topologies (>9 Å); but the final model by I-TASSER is 3.6 Å away from the X-ray structure. Recently, Helles made a comparative study of 18 different ab initio prediction algorithms in the literature and concluded that I-TASSER is currently the best method in the balance of modeling accuracy and CPU cost (Helles 2008). However, as indicated by the fact that no high-resolution model has been predicted in the CASP7 New Fold category (Jauch et al. 2007), I-TASSER modeling has a resolution limit by the inherent reduced potential. One of the on-going efforts is to extend the reduced I-TASSER modeling to the atomic representation with the goal to improve the modeling accuracy in the atomic-level (Zhang 2007).

11.4 Conclusion

Since a detailed physicochemical description of protein folding principles does not yet exist, the most accurate structure predictions are generated based on evolutionary relationships between the target and solved structures in the PDB library. For the proteins with close templates, full-length models can be constructed by copying the template framework. Recent studies show that using the best possible template structures in PDB, the state-of-the-art modeling algorithms could build high-quality full-length models for almost all single-domain proteins with an average RMSD ~2.3 Å. This suggests that the current PDB structure universe is essentially complete for solving protein structure prediction problem (Zhang and Skolnick 2005a). However, most of the target-template pairs are evolutionarily too distant to be detected with current threading approaches.

The development of efficient threading algorithms to detect weakly/distant structure templates has been a central theme in the field and may persist as a principal direction; the gap between threading and the best structural alignment is obvious and tempting. However, progress in reducing this gap progresses slowly. As mentioned above, there is no single threading method that outperforms all others on every target. Consequently, meta-servers and MQAP have been used to generate predictions by collecting and selecting models from a set of different threading programs. In contrast, the template structure refinement has enjoyed promising progress. In the recent CASP7 experiment (Battey et al. 2007), automated threading plus structure refinement servers outperform the threading and MQAP based meta-servers by a noticeable margin. Nevertheless, template refinement mainly occurs at the topology level. The demand for atomic-level models, which can generate models of real use for new drug screening and biochemical function inference, is keener than ever as more template structures become available through the structure genomics and traditional structural biology.

Free modeling is the 'Holy Grail' of protein structure prediction because its success would mark the eventual solution to a problem manifested at genome scales. Although a purely physics-based ab initio simulation has the advantage in revealing the pathway of protein folding, the best current free-modeling results come from those which combine both knowledge-based and physics-based approaches. While there are consistent successes in building correct topologies (3 ~ 6 Å) for small proteins, the more exciting high-resolution free modeling (<2 Å) is much rarer and computationally more expensive. There is evidence that the current atomic potentials have the lowest energy near the native state, and the bottleneck of high-resolution folding seems to be insufficient conformational sampling (Bradley et al. 2005). However, a golf-hole-like energy landscape without middle range funnel is far from the one taken in nature and this can be a deeper reason for failures in conformational searches. Thus, the bottleneck for free modeling comes from the lack of both funnel-like force fields and efficient space searching methods, especially for proteins of larger sizes.

Acknowledgment We are grateful to Dr. Terry Clark for reading the manuscript and giving us helpful comments. The project is partly supported by KU Start-up Fund 06194. Y.Z. is supported by the Alfred P. Sloan Foundation.

References

Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S et al (2005) The Universal Protein Resource (UniProt). Nucleic Acids Res 33(Database issue):D154–D159

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S et al (2004) The Pfam protein families database. Nucleic Acids Res 32(Database issue):D138–D141

Battey JN, Kopp J, Bordoli L, Read RJ, Clarke ND, Schwede T (2007) Automated server predictions in CASP7. Proteins 69(S8):68–82

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2003) GenBank. Nucleic Acids Res 31(1):23–27

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

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H et al (2000) The Protein Data Bank. Nucleic Acids Res 28(1):235–242

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

Bowie JU, Luthy R, Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253:164–170

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, States DJ, Swaminathan S, Karplus M (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 4(2):187–217

Burley SK, Almo SC, Bonanno JB, Capel M, Chance MR, Gaasterland T et al (1999) Structural genomics: beyond the human genome project. Nat Genet 23(2):151–157

Case DA, Pearlman DA, Caldwell JA, Cheatham TE, Ross WS (1997) AMBER 5.0. University of California, San Francisco, CA

Chandonia JM, Brenner SE (2006) The impact of structural genomics: expectations and outcomes. Science 311(5759):347–351

Chen J, Brooks CL III (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

Das R, Qian B, Raman S, Vernon R, Thompson J, Bradley P et al (200) 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

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

Felts AK, Gallicchio E, Wallqvist A, Levy RM (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 (2003) 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. Proteins 51(3):434–441

Fischer D (2006) Servers for protein structure prediction. Curr Opin Struct Biol 16(2):178–182

Fischer D, Rychlewski L, Dunbrack RL Jr, Ortiz AR, Elofsson A (2003) CAFASP3: the third critical assessment of fully automated structure prediction methods. Proteins 53(Suppl 6):503–516

Ginalski K, Pas J, Wyrwicz LS, von Grotthuss M, Bujnicki JM, Rychlewski L (2003) ORFeus: Detection of distant homology using sequence profiles and predicted secondary structure. Nucleic Acids Res 31(13):3804–3807

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

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

Jaroszewski L, Rychlewski L, Li Z, Li W, Godzik A (2005) FFAS03: a server for profile–profile sequence alignments. Nucleic Acids Res 33(Web Server issue):W284–W288

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

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

Jones DT, Taylor WR, Thornton JM (1992) A new approach to protein fold recognition. Nature 358(6381):86–89

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

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

Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL (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, Skolnick J (2001) TOUCHSTONE: An ab initio protein structure prediction method that uses threading-based tertiary restraints. Proc Natl Acad Sci U S A 98:10125–10130

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, Floudas CA (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

Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T (2007) Assessment of CASP7 predictions for template-based modeling targets. Proteins 6(S8):38–56

Kulikova T, Aldebert P, Althorpe N, Baker W, Bates K, Browne P et al (2004) The EMBL nucleotide sequence database. Nucleic Acids Res 32(Database issue):D27–D30

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

Lee MR, Tsai J, Baker D, Kollman PA (2001) Molecular dynamics in the endgame of protein structure prediction. J Mol Biol 313(2):417–430

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

Levitt M, Hirshberg M, Sharon R, Daggett V (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

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

Liwo A, Lee J, Ripoll DR, Pillardy J, Scheraga HA (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, Rackovsky S, Scheraga HA (1993) Calculation of protein backbone geometry from alpha-carbon coordinates based on peptide-group dipole alignment. Protein Sci 2(10):1697–1714

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

Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A (2000) Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 29:291–325

McPherson JD, Marra M, Hillier L, Waterston RH, Chinwalla A, Wallis J et al (2001) A physical map of the human genome. Nature 409(6822):934–941

Misura KM, Chivian D, Rohl CA, Kim DE, Baker D (2006) Physically realistic homology models built with ROSETTA can be more accurate than their templates. Proc Natl Acad Sci U S A 103(14):5361–5366

Miyazaki S, Sugawara H, Gojobori T, Tateno Y (2003) DNA Data Bank of Japan (DDBJ) in XML. Nucleic Acids Res 31(1):13–16

Moult J, Fidelis K, Kryshtafovych A, Rost B, Hubbard T, Tramontano A (2007) Critical assessment of methods of protein structure prediction-Round VII. Proteins 69(Suppl 8):3–9

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

Nemethy G, Gibson KD, Palmer KA, Yoon CN, Paterlini G, Zagari A 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

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

Pieper U, Eswar N, Braberg H, Madhusudhan MS, Davis FP, Stuart AC et al (2004) MODBASE, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res 32(Database issue):D217–D222

Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS, Rossi A et al (2006) MODBASE: a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res 34(Database issue):D291–D295

Rychlewski L, Fischer D (2005) LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction. Protein Sci 14(1):240–245

Sadreyev R, Grishin N (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. J Mol Biol 326(1):317–336

Sali A (1998) 100, 000 protein structures for the biologist. Nat Struct Biol 5(12):1029–1032

Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234(3):779–815

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

Simons KT, Kooperberg C, Huang E, Baker D (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

Skolnick J, Fetrow JS, Kolinski A (2000) Structural genomics and its importance for gene function analysis. Nat Biotechnol 18(3):283–287

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

Smaglik P (2000) Protein structure groups seek to draft common ground rules. Nature 403(6771):691

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

Stevens RC, Yokoyama S, Wilson IA (2001) Global efforts in structural genomics. Science 294(5540):89–92

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

Terwilliger TC, Waldo G, Peat TS, Newman JM, Chu K, Berendzen J (1998) Class-directed structure determination: foundation for a protein structure initiative. Protein Sci 7(9):1851–1856

Tsai J, Bonneau R, Morozov AV, Kuhlman B, Rohl CA, Baker D (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, Hunenberger PH, Kruger P, Mark AE et al (1996) Biomolecular Simulation: The GROMOS96 Manual and User Guide. Vdf Hochschulverlag AG an der ETH Zürich, Zürich

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG et al (2001) The sequence of the human genome. Science 291(5507):1304–1351

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

Vitkup D, Melamud E, Moult J, Sander C (2001) Completeness in structural genomics. Nat Struct Biol 8(6):559–566

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? J Comput Chem 21(12):1049–1074

Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G et al (1984) A new force field for molecular mechanical simulation of nucleic acids and proteins. J Am Chem Soc 106:765–784

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

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

Zagrovic B, Snow CD, Shirts MR, Pande VS (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 Y (2007) Template-based modeling and free modeling by I-TASSER in CASP7. Proteins 69(Suppl 8):108–117

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

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(4):702–710

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 Res 33(7):2302–2309

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