Published in final edited form as:
Curr Opin Struct Biol. 2008 June ; 18(3): 342–348. doi:10.1016/j.sbi.2008.02.004.

Progress and challenges in protein structure prediction

Yang Zhang
Center for Bioinformatics and Department of Molecular Biosciences, University of Kansas, 2030 Becker Drive, Lawrence, KS 66047, United States

Abstract

Depending on whether similar structures are found in the PDB library, the protein structure prediction can be categorized into template-based modeling and free modeling. Although threading is an efficient tool to detect the structural analogs, the advancements in methodology development have come to a steady state. Encouraging progress is observed in structure refinement which aims at drawing template structures closer to the native; this has been mainly driven by the use of multiple structure templates and the development of hybrid knowledge-based and physics-based force fields. For free modeling, exciting examples have been witnessed in folding small proteins to atomic resolutions. However, predicting structures for proteins larger than 150 residues still remains a challenge, with bottlenecks from both force field and conformational search.

Introduction

In recent years, despite many debates, structure genomics is probably one of the most noteworthy efforts in protein structure determination, which aims to obtain 3D models of all proteins by an optimized combination of experimental structure solution and computer-based structure prediction [1,2•]. Two factors will dictate the success of the structure genomics: experimental structure determination of optimally selected proteins and efficient computer modeling algorithms. Based on about 40 000 structures in the PDB library (many are redundant) [3], 4 million models/fold-assignments can be obtained by a simple combination of the PSI-BLAST search and the comparative modeling technique [4•]. Development of more sophisticated and automated computer modeling approaches will dramatically enlarge the scope of modelable proteins in the structure genomics project.

The crucial problems/efforts in the field of protein structure prediction include: first, for the sequences of similar structures in PDB (especially those of weakly/distant homologous relation to the target), how to identify the correct templates and how to refine the template structure closer to the native; second, for the sequences without appropriate templates, how to build models of correct topology from scratch. The progress made along these directions was assessed in the recent CASP7 experiment [5] under the categories of template-based modeling (TBM) and free modeling (FM). Here, I will review the new progress and challenges in these directions.

Template-based modeling

The canonical procedure of the TBM consists of four steps: first, finding known structures (templates) related to the sequence to be modeled (target); second, aligning the target sequence to the template structure; third, building structural frameworks by copying the aligned regions or by satisfying the spatial restraints from templates; fourth, constructing the unaligned loop

Corresponding author: Zhang, Yang (E-mail: yzhang@ku.edu).

regions and adding side-chain atoms. The first two steps are actually done in a single procedure called threading (or fold recognition) [6,7] because the correct selection of templates relies on the accurate alignment. Similarly, the last two steps are performed simultaneously since the atoms of the core and loop regions are in close interaction.

The existence of similar structures in the PDB is a necessary precondition for the successful TBM. An important question is how complete the current PDB structure library is. Figure 1 shows a distribution of the best templates found by the structural alignment [8] for 1413 representative single-domain proteins between 80 and 200 residues. Remarkably, even excluding the homologous templates of sequence identity >20%, all the target 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% regions. The average rmsd and coverage are 2.96 Å and 86%, respectively. Zhang and Skolnick [9••] recently showed that high-quality full-length models could be built for all the protein targets with an average rmsd 2.25 Å when using the best templates in the PDB. These data demonstrate that the structural universe of the current PDB library is complete essentially for solving the protein structure problem for at least the single-domain proteins. However, most of the target–template pairs at this level of sequence identity (~15%) are difficult to identify by threading. In fact, after excluding the templates of sequence identity >30%, only two-third of the proteins could be assigned by the current threading techniques to the templates of a correct topology with some alignment errors (average rmsd ~ 4 Å) [10]. Thus, the role of the structure genomics initiative is to bridge the target–template gap for the remaining one-third proteins, as well as, to improve the alignment accuracy of the two-third proteins by providing evolutionarily closer template proteins.

Template structure identification

Since its invention in the early 1990s [6,7], threading has become one of the most active areas in proteins structure prediction. Numerous algorithms have been developed during the past 15 years for the purpose of identifying structure templates from the PDB, which use techniques including sequence profile–profile alignments (PPAs) [10–13], structural profile alignments [14], hidden Markov models (HMMs) [15,16••], machine learning [17,18], and others.

The sequence 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 and the log-odds of the amino acid in the template MSA, the profile [19]. There are alternatives in calculating the PPA scores [20]. The profile-alignment-based methods demonstrated advantages in several recent blind tests [21,22,23•]. In Live-Bench-8 [21], for example, all top four servers (BASD/MASP/MBAS, SFST/STMP, FFAS03, and ORF2/ORFS) were based on the sequence PPA. In CAFASP [22] and the recent CASP Server Section [23•], several sequence-profile-based methods were ranked at the top of single-threading servers. Wu and Zhang [24] recently showed that the accuracy of the sequence PPAs can be further improved by about 5–6% by incorporating a variety of additional structural information.

In CASP7, HHsearch [16••], a HMM–HMM alignment method, stands out to be the best single-threading server. The principle of the HMM–HMM alignments and the PPAs is similar in that both try to perform a pair-wise alignment of the target MSA with the template MSA. Instead of representing the MSAs by sequence profiles, HHsearch uses profile HMMs that can generate the sequences with certain probabilities, given by the product of amino-acid emission and insertion/deletion probabilities. HHsearch aligns the target and template HMMs by maximizing the probability that two models coemit the same amino-acid sequence. In this way, amino-acid frequencies and insertions and deletions of both HMMs are matched up together in an optimum way [16••].

Although the average performance differs among different algorithms, there is not a single-threading program that can outperform other methods for every target. This naturally leads to the prevalence of the so-called meta-server [25,26•,27], which collects and combines results from a set of different threading programs. There are two ways to generate predictions in meta-servers. One is to build a hybrid model by cut-and-paste of the selected structural fragments from multiple templates [27]. The combined model has on average larger coverage and better topology than the best single template. One draw-back is that often the hybrid models have nonphysical local clashes between atoms. The second way is to *select* the best model based on a variety of scoring functions or machine-learning techniques, which emerges as a new research topic called Model Quality Assessment Programs (MQAPs) [28]. Despite considerable efforts in developing various MQAP scores, the most robust score turns out to be the one based on the structure consensus [29•], that is, the best models are those simultaneously hit by various threading algorithms. The idea behind the consensus approach is simple because there are more ways for a threading program to select a wrong template than a right one. Therefore, the chances for multiple threading programs to make a common but wrong selection are much lower than the chances to make a common and correct selection.

The meta-server predictors have dominated the server predictions in previous experiments (e.g. CAFASP4 [28], LiveBench-8 [21], and CASP6 [30]). In the recent CASP7 experiment [23•], however, Zhang-Server (an automated server based on profile–profile threading and I-TASSER structure refinement [31••]) clearly outperforms others (including the meta-servers which include it as an input [29•]). A list of the top 10 automated servers in the CASP7 experiment is shown in Table 1. This data on the one hand highlight the challenge to the MQAP methods in correctly ranking and selecting the best models; on the other hand, the success of the composite threading plus refinement servers (as Zhang-Server, ROBETTA, and MetaTasser) demonstrates the advantage of structure refinement in the TBM prediction.

Template structure refinement

The goal of the protein structure refinement is to draw the templates closer to the native, which has proven to be an extremely nontrivial problem. Until only a few years ago, most of the TBM procedures either keep the templates unchanged or drive the templates away from the native structures [32,33].

Early efforts on template structure refinement have been focused on the molecular dynamics (MD)-based atomic simulations, which attempt to refine low-resolution models by running the classic software such as AMBER and CHARMM. Except for some isolated instances, however, no systematic improvement was achieved [34]. The failure of the MD-based structure refinements seems contrary to the reported successes of the MD potentials in discriminating the native from structural decoys. Wroblewska and Skolnick [35••] recently showed that the AMBER plus GB potential could only discriminate the native from roughly minimized TASSER structure decoys [36]. After a 2-ns MD simulation, none of the native structures have the lowest energy among decoys and the energy–rmsd correlation vanishes. A noteworthy observation was recently made by Summa and Levitt [37••] who exploited different molecular mechanics (MM) potentials (AMBER99, OPLS-AA, GROMOS96, and ENCAD) on the refinement of 75 proteins by *in vacuo* energy minimization. The authors found that a knowledge-based atomic contact potential based on the PDB statistics outperforms all the traditional MM potentials by moving almost all the test proteins closer to the native state, while the MM potentials, except for AMBER99, essentially drive the decoys away from the native. The vacuum simulation without solvation may be a part of the reason for the failure of the MM potentials. But this observation demonstrates the potential of the hybrid knowledge-based and physics-based potentials in the protein structure refinement.

Encouraging template refinements have been recently achieved by combining the hybrid potentials with spatial restraints from threading templates [9••,38••,39•]. Misura *et al.* [38••] first built low-resolution models by ROSETTA [40] using a fragment library enriched by the query-template alignment; the C_β-contact restraints were used to guide the assembly procedure. The low-resolution models were then refined by a physics-based atomic potential. As a result, in 22 of 39 test cases, at least 1 of the 10 lowest energy models was found closer to the native than the template.

A more comprehensive test of the template refinement procedure based on TASSER simulations, combined with consensus spatial restraints from multiple templates, was reported by Zhang and Skolnick [9••,36]. For 1489 test cases, TASSER reduces the rmsd of the templates in the majority of cases with an average rmsd reduction from 6.7 to 4.4 Å over the threading aligned regions. Even starting from the best templates as identified by the structural alignment, TASSER refines the models from 2.5 to 1.88 Å in the aligned regions. Here, TASSER has 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 the refinements is the use of multiple threading templates where the consensus spatial restraint is more accurate than that from the individual template. Second, the composite knowledge-based energy terms have been extensively optimized using large-scale structure decoys [41] which help coordinate the complicated correlations between different interaction terms.

The progress of threading template refinements has been assessed in the recent CASP7 experiment, where the assessors compared the predicted models with the best 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' [42•]. 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 and the latter requests the information of the native, which was not available when the predictions were made. On the contrary, a global GDT score comparison may favor the full-length models because the template alignment has a shorter length than the models. In a direct comparison of the rmsd over the same aligned regions, we find that the first I-TASSER model is closer to the native than the best initial template in 86 of 105 TBM cases while the other 13 (6) cases are worse than (equal to) the template. The average rmsd is 4.9 and 3.8 Å for the templates and models, respectively, over the same aligned regions [31••].

Free modeling

When structural analogs do not exist in the PDB library or could not be successfully identified by threading (which is more often the case as shown by Figure 1), the structure prediction has to be generated from scratch. This type of predictions has been termed as '*ab initio*' or '*de novo*' modeling, a term that may be easily understood as a modeling 'from first principle'. In CASP7, it is named as 'free modeling' which I think reflects more appropriately 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.

The best-known idea for free modeling is probably the one pioneered by Bowie and Eisenberg who assembled new tertiary structures using small fragments (mainly 9-mer) cut from other PDB proteins [43]. On the basis of similar idea, Baker and coworkers developed ROSETTA [40], which has worked extremely well for free modeling in the CASP experiments and made the fragment assembly approach popular in the field. In the new developments of ROSETTA [44••,45•], the authors first assemble structures in a reduced knowledge-based model with

conformations specified by the heavy backbone atoms and C_βs. 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 T0281 (70 residues) of 1.6 Å away from the crystal structure. In CASP7, ROSETTA built a model for T0283 (112 residues) with rmsd = 1.8 Å over 92 residues (Figure 2, left panel). Despite significant success, the computer cost of the procedure (~150 CPU days for a small protein <100 residues) is still too expensive for the routine use.

Another successful free modeling approach, called TASSER [36] by Zhang and Skolnick, constructs 3D models based on a purely knowledge-based approach. Continuous fragments of 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.* [46••], which refines the TASSER cluster centroids by iterative Monte Carlo simulations. Although the procedure uses structural fragments and spatial restraints from threading templates, it often constructs models of correct topology even when the topologies of individual templates are incorrect. In CASP7, 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. Figure 2 (right panel) shows one example of T0382 (123 residues) where all initial templates have a wrong topology (>9 Å) but the final model is 3.6 Å away from the X-ray structure.

Significant efforts have been made on the purely physics-based protein folding and structure prediction. The very first milestone of successful *ab initio* protein folding is probably the 1997 work of Duan and Kollman, who folded the villin headpiece (a 36-mer) by MD simulations in explicit solvent for two months on parallel supercomputers with models up to 4.5 Å [47]. With the help of the worldwide-distributed computers, this small protein was recently folded by Pande and coworkers [48] to 1.7 Å with a total simulation time of 300 μs or approximately 1000 CPU years. To reduce the computing cost, Scheraga and coworkers [49•] developed a reduced physics-based model, called UNRES, which represents protein conformations by C_α, side-chain center, and a virtual peptide group. The low-energy UNRES models are then converted to all-atom representations based on ECEPP/3. In CASP6, a structure genomic target of TM0487 (T0230, 102 residues) was folded to a structure within 7.3 Å by the approach. Using ASTRO-FOLD on the ECEPP/3 optimization, Floudas and coworkers [50] recently constructed a model of 5.2 Å for a four-helical bundle protein of 102 residues in a double-blind prediction.

Conclusions

Since a detailed physicochemical description of protein folding principles does not yet exist, the protein structure prediction problem is largely defined by the evolutionary or structural distance between the target and the solved proteins 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 if 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 may be approaching complete for solving the protein structure prediction problem [9••]. However, most of the target–template pairs are evolutionarily too distant to be detected with the current threading approaches.

The development of efficient threading algorithms to detect weakly/distant homologous templates has been a central theme in the field and may persist as a principal direction, as the gap between threading and the best structural alignment is obvious and tempting. However, progress in reducing this gap is slow or incremental since the invention of the PPA techniques. There is no single-threading method that outperforms all others on every target; this results in

the prevalence of the meta-servers and MQAP which generate predictions by collecting and selecting models from a set of other threading programs. On the contrary, the template structure refinement has enjoyed promising progress. In the recent CASP7 experiment [23•], automated threading plus structure refinement servers outperforms by a margin the threading-only and the MQAP-based meta-servers. Nevertheless, the template refinement mainly occurs at the topology level. The demand for atomic-level structural refinements, which can generate models of use in drug screening and biochemical function inference, is keener than ever, especially when more and more template structures become available through the structure genomics and traditional structural biology.

Free modeling is certainly the 'Holy Grail' of the protein structure prediction because its success would mark the eventual solution to the problem. 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. Although there are consistent successes in building correct topology (3–6 Å) for small proteins, the more exciting high-resolution free modeling (<2 Å) is rarer and computationally 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 the insufficient conformational sampling [44••]. However, a golf-hole-like energy landscape without middle-range funnel should not be the one taken in nature, which can be a deeper reason for the failure of conformational search. Thus, the bottleneck for free modeling comes from the lack of both funnel-like force fields and efficient space searching, especially for proteins of larger sizes.

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 and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

•of special interest

••of outstanding interest

1. Burley SK, Almo SC, Bonanno JB, Capel M, Chance MR, Gaasterland T, Lin D, Sali A, Studier FW, Swaminathan S. Structural genomics: beyond the human genome project. Nat Genet 1999;23:151–157. [PubMed: 10508510]
2. Chandonia JM, Brenner SE. The impact of structural genomics: expectations and outcomes. Science 2006;311:347–351.351 [PubMed: 16424331] The authors review and assess the gain and loss of the structural genomics project in the past five years in contrast with traditional structural biology.
3. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242. [PubMed: 10592235]
4. Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS, Rossi A, Marti-Renom M, Karchin R, Webb BM, Eramian D, et al. MODBASE: a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res 2006;34:D291–D295.D295 [PubMed: 16381869] MODBASE is a database of 3D models built by the MODELLER pipeline for all protein sequences in SwissProt based on available structural templates in the PDB library.
5. Moult J, Fidelis K, Kryshtafovych A, Rost B, Hubbard T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP) — round VII. Proteins 2007;69:3–9. [PubMed: 17918729]

6. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170. [PubMed: 1853201]
7. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89. [PubMed: 1614539]
8. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302–2309. [PubMed: 15849316]
9. Zhang Y, Skolnick J. The protein structure prediction problem could be solved using the current PDB library. Proc Natl Acad Sci U S A 2005;102:1029–1034.1034 [PubMed: 15653774] Using the best available templates, TASSER could build high-quality models for all single-domain proteins. This shows that the current structure set in PDB is essentially complete for the protein structure prediction problem, though most of the templates are not detectable by current threading approaches.
10. Skolnick J, Kihara D, Zhang Y. Development and large scale benchmark testing of the PROSPECTOR 3.0 threading algorithm. Protein 2004;56:502–518.
11. Jaroszewski L, Rychlewski L, Li Z, Li W, Godzik A. FFAS03: a server for profile–profile sequence alignments. Nucleic Acids Res 2005;33:W284–W288. [PubMed: 15980471]
12. Zhou H, Zhou Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins 2005;58:321–328. [PubMed: 15523666]
13. Ginalski K, Pas J, Wyrwicz LS, von Grotthuss M, Bujnicki JM, Rychlewski L. ORFeus: detection of distant homology using sequence profiles and predicted secondary structure. Nucleic Acids Res 2003;31:3804–3807. [PubMed: 12824423]
14. Shi J, Blundell TL, Mizuguchi K. FUGUE: sequence–structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 2001;310:243–257. [PubMed: 11419950]
15. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics 1998;14:846–856. [PubMed: 9927713]
16. Soding J. Protein homology detection by HMM–HMM comparison. Bioinformatics 2005;21:951–960.960 [PubMed: 15531603] The sequence–HMM alignment is extended to the pair-wise profile HMM–HMM alignment for the remote homology detection. The HHsearch is one of the best single-threading servers in CASP7.
17. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287:797–815. [PubMed: 10191147]
18. Cheng J, Baldi P. A machine learning information retrieval approach to protein fold recognition. Bioinformatics 2006;22:1456–1463. [PubMed: 16547073]
19. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci U S A 1987;84:4355–4358. [PubMed: 3474607]
20. Sadreyev R, Grishin N. COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. J Mol Biol 2003;326:317–336. [PubMed: 12547212]
21. Rychlewski L, Fischer D. LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction. Protein Sci 2005;14:240–245. [PubMed: 15608124]
22. Fischer D, Rychlewski L, Dunbrack RL Jr, Ortiz AR, Elofsson A. CAFASP3: the third critical assessment of fully automated structure prediction methods. Proteins 2003;53:503–516. [PubMed: 14579340]
23. Battey JN, Kopp J, Bordoli L, Read RJ, Clarke ND, Schwede T. Automated server predictions in CASP7. Proteins 2007;69 Suppl 8:68–82.82 [PubMed: 17894354] It is an official assessment paper for the structure prediction servers in CASP7, which is especially helpful for the users who want to find appropriate servers for generating their own structure prediction.
24. Wu ST, Zhang Y. MUSTER: improving protein sequence profile–profile alignments by using multiple sources of structure information. Proteins. 2008
25. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 2003;19:1015–1018. [PubMed: 12761065]
26. Wu ST, Zhang Y. LOMETS: a local meta-threading-server for protein structure prediction. Nucleic Acids Res 2007;35:3375–3382.3382 [PubMed: 17478507] LOMETS is a new meta-server with all individual threading programs installed locally, which ensures a quick collection and selection of multiple threading results.

27. Fischer D. 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. Proteins 2003;51:434–441. [PubMed: 12696054]
28. Fischer D. Servers for protein structure prediction. Curr Opin Struct Biol 2006;16:178–182. [PubMed: 16546376]
29. Wallner B, Elofsson A. Prediction of global and local model quality in CASP7 using Pcons and ProQ. Proteins 2007;69 Suppl 8:184–193.193 [PubMed: 17894353] The Pcons-server generates structure predictions by ranking and selecting models generated by other programs. It shows that the structural consensus is the most robust score for protein model selection.
30. Moult J, Fidelis K, Rost B, Hubbard T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP) — round 6. Proteins 2005;61:3–7. [PubMed: 16187341]
31. Zhang Y. Template-based modeling and free modeling by I-TASSER in CASP7. Proteins 2007;69 Suppl 8:108–117.117 [PubMed: 17894355] Template structures can be refined significantly closer to the native by a purely knowledge-based I-TASSER modeling. I-TASSER also generated the correct topology for 7 of 19 free modeling targets in CASP7.
32. Tress M, Ezkurdia I, Grana O, Lopez G, Valencia A. Assessment of predictions submitted for the CASP6 comparative modeling category. Proteins 2005;61:27–45. [PubMed: 16187345]
33. Tramontano A, Morea V. Assessment of homology-based predictions in CASP5. Proteins 2003;53:352–368. [PubMed: 14579324]
34. Lee MR, Tsai J, Baker D, Kollman PA. Molecular dynamics in the endgame of protein structure prediction. J Mol Biol 2001;313:417–430. [PubMed: 11800566]
35. Wroblewska L, Skolnick J. Can a physics-based, all-atom potential find a protein's native structure among misfolded structures? I. Large scale AMBER benchmarking. J Comput Chem 2007;28:2059–2066.2066 [PubMed: 17407093] AMBER plus GB solvation potential can discriminate the native from the roughly minimized structural decoys. After a longer MD simulation, however, the energy–rmsd correlation vanishes. This finding partially explains the discrepancy between the discrimination ability and some unsuccessful folding/refinement results of the physics-based potentials.
36. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci U S A 2004;101:7594–7599. [PubMed: 15126668]
37. Summa CM, Levitt M. Near-native structure refinement using in vacuo energy minimization. Proc Natl Acad Sci U S A 2007;104:3177–3182.3182 [PubMed: 17360625] The *in vacuo* energy minimization experiments show that a knowledge-based atomic contact potential from the PDB statistics outperforms all traditional molecular mechanics potentials in driving the protein structure decoys toward the native state.
38. Misura KM, Chivian D, Rohl CA, Kim DE, Baker D. Physically realistic homology models built with ROSETTA can be more accurate than their templates. Proc Natl Acad Sci U S A 2006;103:5361–5366.5366 [PubMed: 16567638] The hybrid approaches of the ROSETTA structure assembly combined with atomic refinements guided by spatial restraints are shown to be able to draw 22 of 39 template models closer to the native.
39. Chen J, Brooks CL III. Can molecular dynamics simulations provide high-resolution refinement of protein structure? Proteins 2007;67:922–930.930 [PubMed: 17373704] CHARMM22/GBSW with spatial restraints are able to refine four of five CASP6 CM targets with up to 1 Å rmsd reduction, a new progress of the MD-based structure refinements.
40. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225. [PubMed: 9149153]
41. Zhang Y, Kolinski A, Skolnick J. TOUCHSTONE II: a new approach to ab initio protein structure prediction. Biophys J 2003;85:1145–1164. [PubMed: 12885659]
42. Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T. Assessment of CASP7 predictions for template-based modeling targets. Proteins 2007;69 Suppl 8:38–56.56 [PubMed: 17894352] The paper assesses the template-based modeling category, which includes 108 out of a total of 123 targets/domains in CASP7. Progress in the template refinement is highlighted.
43. Bowie JU, Eisenberg D. An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function. Proc Natl Acad Sci U S A 1994;91:4436–4440. [PubMed: 8183927]

44. Bradley P, Misura KM, Baker D. Toward high-resolution de novo structure prediction for small proteins. Science 2005;309:1868–1871.1871 [PubMed: 16166519] This is the first work to report successful high-resolution modeling casesby free modeling. It states that atomic potentials have the lowest energy near the native state and the bottleneck for high-resolution free modeling is the insufficient conformation search.
45. Das R, Qian B, Raman S, Vernon R, Thompson J, Bradley P, Khare S, Tyka MD, Bhat D, Chivian D, et al. Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. Proteins 2007;69 Suppl 8:118–128.128 [PubMed: 17894356] The paper summarizes the recent progress of ROSSETA using distributed computing resource and the performance of ROSETTA@home on the CASP7 targets.
46. Wu ST, Skolnick J, Zhang Y. Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol 2007;5:17. [PubMed: 17488521] By iterative TASSER assembly, I-TASSER is able to generate medium-resolution to high-resolution models for small proteins without using homologous templates. The computing cost is significantly lower than the corresponding atomic-based structure predictions.
47. Duan Y, Kollman PA. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science 1998;282:740–744. [PubMed: 9784131]
48. Zagrovic B, Snow CD, Shirts MR, Pande VS. Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. J Mol Biol 2002;323:927–937. [PubMed: 12417204]
49. Oldziej S, Czaplewski C, Liwo A, Chinchio M, Nanias M, Vila JA, Khalili M, Arnautova YA, Jagielska A, Makowski M, et al. Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: assessment in two blind tests. Proc Natl Acad Sci U S A 2005;102:7547–7552.7552 [PubMed: 15894609] By using a reduced physics-based approach, UNRES is able to generate correct topologies for proteins up to 102 residues.
50. Klepeis JL, Wei Y, Hecht MH, Floudas CA. Ab initio prediction of the three-dimensional structure of a de novo designed protein: a double-blind case study. Proteins 2005;58:560–570. [PubMed: 15609306]
51. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004;57:702–710. [PubMed: 15476259]

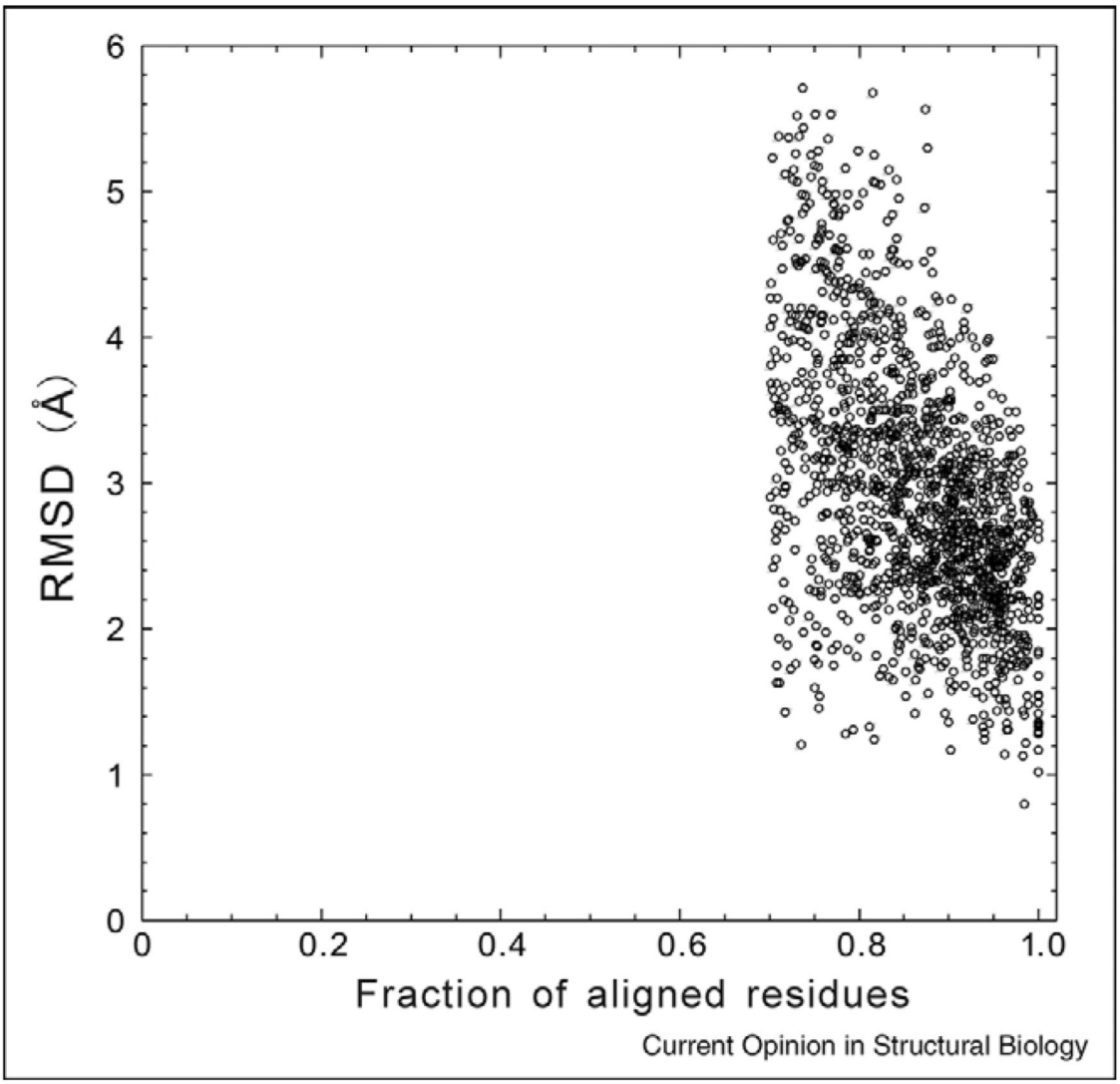

Figure 1.
Structural superimposition results of 1413 representative single-domain proteins on their analogs in the PDB library. The structural analogs are searched by a sequence-independent structural-alignment tool, TM-align [8], and ranked by TM-score (a structural similarity measure balancing rmsd and coverage) [51]. All structural analogs with a sequence identity >20% to the target are excluded. If the analog of 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 rmsd < 6 Å; 80% have a rmsd < 4 Å with >75% regions covered.

Figure 2.
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-based and physics-based approach) [45•] 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) [31••] with a TM-score 0.66 and a rmsd 3.6 Å. Blue and red represent the model and the crystal structure, respectively.

Table 1

Top 10 servers in CASP7 as ranked by the accumulative GDT-TS score

Servers	Number of targets	GDT-TS score	Server type; 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); server not avaliable

Multiple servers from the same lab are represented by the highest rank one.