Title

Trends in template/fragment-free protein structure prediction

Journal

Theoretical Chemistry Accounts: Theory, Computation, and Modeling, 128(1)

ISSN

1432-2234

Authors

Zhou, Yaoqi

Duan, Yong

Yang, Yuedong

et al.

Publication Date

2011

DOI

10.1007/s00214-010-0799-2

Peer reviewed

Theor Chem Acc (2011) 128:3–16
DOI 10.1007/s00214-010-0799-2

FEATURE ARTICLE

Trends in template/fragment-free protein structure prediction

Yaoqi Zhou · Yong Duan · Yuedong Yang · Eshel Faraggi · Hongxing Lei

Received: 17 May 2010 / Accepted: 15 August 2010 / Published online: 1 September 2010

Abstract Predicting the structure of a protein from its amino acid sequence is a long-standing unsolved problem in computational biology. Its solution would be of both fundamental and practical importance as the gap between the number of known sequences and the number of experimentally solved structures widens rapidly. Currently, the most successful approaches are based on fragment/template reassembly. Lacking progress in template-free structure prediction calls for novel ideas and approaches. This article reviews trends in the development of physical and specific knowledge-based energy functions as well as sampling techniques for fragment-free structure prediction. Recent physical- and knowledge-based studies demonstrated that it is possible to sample and predict highly accurate protein structures without borrowing native fragments from known protein structures. These emerging approaches with fully flexible sampling have the potential to move the field forward.

Keywords Protein structure prediction · Conformational sampling · Knowledge-based energy function · Protein folding · Molecular dynamics simulation · Molecular mechanics force field

Y. Zhou (✉) · Y. Yang · E. Faraggi
School of Informatics,
Indiana Center for Computational Biology and Bioinformatics,
Indiana University School of Medicine,
Indiana University Purdue University,
719 Indiana Ave #319, Walker Plaza Building,
Indianapolis, IN 46202, USA
e-mail: yqzhou@iupui.edu

Y. Duan · H. Lei
UC Davis Genome Center and Department of Applied Science,
University of California,
One Shields Avenue,
Davis, CA, USA

Y. Duan
College of Physics, Huazhong University of Science
and Technology, 1037 Luoyu Road,
430074 Wuhan, China

H. Lei
Beijing Institute of Genomics,
Chinese Academy of Sciences,
100029 Beijing, China

1 Introduction

One of the long-standing challenges in computational biology is to fold proteins of given amino acid sequences into native functional three-dimensional structures of experimental accuracy. Such reliable protein structure prediction method is in urgent need because it is far cheaper to sequence the entire genome of a species (<$10,000) [1] than to determine the structure of a single protein (~$100,000) [2]. As a result, the number of sequences generated from genome sequencing projects outpaces the growth of structures solved by experimental techniques by orders of magnitude. It is considered practically impossible to solve the structures of millions of proteins by experimental techniques, and the fact that not all protein structures can be solved by existing experimental techniques further exacerbates the challenge. For example, X-ray crystallography requires high-quality crystals that are not always possible to obtain while the Nuclear Magnetic Resonance (NMR) technique is currently limited to small-size proteins.

The most influential event in the structure prediction community is the biannual CASP meeting (Critical Assessment of Structure Prediction techniques) [3]. At two-year

intervals since 1994, sequences whose structures are soon to be solved are collected from structural biologists and distributed to computational biologists for prediction. Predicted structures are then compared to experimental solutions, and results from this comparison are reported in the bi-annual CASP meeting. The most effective structure prediction techniques highlighted by CASP include fragment-based assembly [4], profile and/or threading-based fold recognition [5–18], consensus and meta-server methods [12, 19–22], and template assembly [23]. While encouraging progress has been made, the overall pace of advancement since the first CASP remains slow [24]. The most successful techniques in structure prediction (e.g. ROSETTA [4] and TASSER [23]) appear to converge to a unified approach of mixing and matching known native structures either in whole (template-based modeling) or in part (fragment assembly) [24, 25]. The convergence of methods highlights the need for innovative techniques to break the impasse in protein structure prediction.

The CASP meeting has had a profound positive impact on the community by promoting the winners (the best predictors), regardless of the methods and databases employed. However, an unintended consequence of the performance-oriented evaluation is that it favors incremental changes from existing proven techniques that have been perfected over the years, rather than novel methods that are potentially game changing but not yet comparable in accuracy to the mature and proven techniques. It rewards the methods that employ the largest database and supercomputing powers and perform a relatively easier task of re-ranking models predicted by other methods, rather than the challenging task of structure prediction. The purpose of this review is to raise the attention to alternative approaches in protein structure prediction with the hope of preventing their premature termination. To limit our scope, we will focus on recent trends and several emerging "ab initio" approaches that are not fragment based. Focusing on fragment-free approaches in this review is not an attempt to reduce the historical or future importance of fragment-based approach but to stimulate new ideas to help solve this challenging problem.

2 Physics-based approaches

Most proteins fold into unique thermodynamically stable structures. The stability of the folded structures and the ability of proteins to perform a wide range of functional activities are determined by solvent-mediated physical interactions between the amino acid residues of the proteins. In principle, such physical interactions can be obtained by solving quantum mechanical equations. However, sufficiently accurate quantum–mechanical simulations of the large-scale motion of proteins are not yet possible because of the large number of complicated interactions in such systems (protein and water molecules). As a result, these interactions are usually approximated by empirical molecular mechanics force fields.

2.1 Molecular mechanics force fields

Molecular mechanics force fields are typically obtained by the combination of quantum mechanical calculations of small peptide fragments and empirical fitting of experimental data [26–28]. Earlier development of force fields focused on dynamics and free-energy simulations of proteins around their native conformations [29–31]. Direct ab initio folding simulations from random coils are hampered not only by the insufficient accuracy of molecular mechanics force fields but also by the astronomically large conformational space of polypeptide chains. Currently, typical molecular dynamics simulations last for a few hundred nanoseconds, compared to actual folding time from microseconds to seconds. Thus, most folding studies in explicit water molecules are limited to small peptides or very small proteins [32, 33]. One milestone study was a microsecond folding simulation of 36-residue villin headpiece starting from an unfolded conformation by Duan and Kollman [34]. Although the presence of water molecules can smooth the free-energy landscape [35], molecular dynamics simulations of low-resolution protein structures with explicit solvent models have mixed outcome: improving the structural accuracy for some but not other proteins [36–39]. In particular, a large-scale study of 75 proteins each with 729 near-native structures [40] indicates that molecular dynamics simulations with explicit solvent molecules started from near-native structures move further away from their respective native conformations. The results underscore the need for further improvement in the force fields and the approaches.

The performance with explicit water molecules described above does not justify the significant increase in computing time needed to include them. As a result, most studies in structure prediction employed simplified implicit solvation models (for reviews see e.g. [41–43]). While most studies are limited to short peptides and small proteins [32, 44–49], some successes for high-resolution ab initio predictions are noteworthy. Simmerling et al. [50], Pitera and Swope [51], and Duan et al. [52] all achieved high-resolution prediction of a 20-residue Trp-cage peptide with various versions of the AMBER force field and a generalized Born (GB) solvation model [53]. Duan et al. folded villin headpiece to less than 0.5 Å C_α-root-mean-squared distance (RMSD) from its native structure [54, 55]. Pande et al. folded villin headpiece to about 1.7 Å of the root-mean-squared of the inter-residue C_α–C_α distance

matrix (dRMS) from its native structure [56] and further developed a method for automatically constructing Markov state models to capture the thermodynamics and kinetics of folding [57]. Duan et al. also reached 2.0 Å RMSD for both three-helix bundles of 47-residue albumin binding domain and 60-residue B domain of protein A (BdpA) [58], and 1.3 Å for a 28-residue designed alpha/beta protein (FSD) [59]. Figure 1 shows the best folded structure achieved during folding simulation when compared to the native structure of BdpA. The lowest sampled conformations are less than 1.0 Å RMSD. It should point out that most of these are small helical proteins. Ab initio folding of proteins of mixed secondary structures and medium size remains a challenging endeavor. Nevertheless, the successful folding of small proteins to sub-angstrom C_α-RMSD by ab initio approach is encouraging. It suggests that, with improved force fields, folding proteins to their native states with experimental accuracy should be possible in the not-too-distant future.

Recently, Dill and his coworkers [60] made a blind prediction of six CASP 7 targets based on AMBER 96 with an implicit GB/SA (Solvent Accessible surface area) model of solvation with a sampling technique called the zipping and assembly [61]. They found that the accuracy of their method is about the average accuracy of other knowledge-based techniques. This is encouraging, considering that the method does not utilize any predicted secondary structures and fragments/templates from known protein structures. Their study will likely re-energize the physics-based approaches that were participants in early CASP experiments (e.g. [62–64]) and currently are overshadowed by knowledge-based or mixed approaches. However, in order to increase the competitive edge of physics-based approach over a knowledge-based one, it is clear that there is a need for further optimization of physics-based force fields and/or solvation models. For example, Jagielska et al. showed that protein models can be refined closer to their native structures using an AMBER force field with optimized relative weights [65]. Krieger et al. [66] re-tuned AMBER parameters by minimizing the deviations from 50 high-resolution protein crystal structures. Lin et al. found that hydrophobic potential of mean force is more useful than commonly used solvent accessible surface area for native structure discrimination [67]. Progress has been made in the development of efficient PB (Poisson-Boltzmann)/SA method that enabled MD simulations of proteins [68]. Because a force field–based approach relies on the continuum solvent models to treat the solvation effect, the overall accuracy and effectiveness of the approach thus requires the advancement in both. One area that may require additional effort is an efficient approach to treat the ionic effect including an accurate model of salt bridges in proteins.

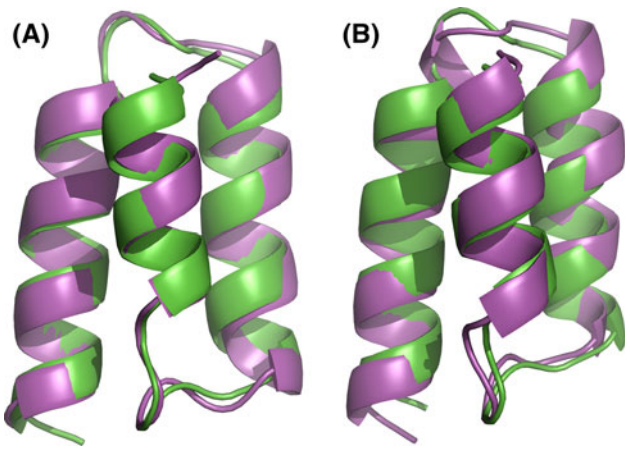

Fig. 1 Comparison between simulated structures (*magenta*) and NMR structure of BdpA (*green*). **a** The best folded structure with 0.8 Å RMSD (C_α only) from MD folding simulation of the truncated BdpA. **b** The best folded structure with 1.3 Å RMSD from the Replica Exchange MD of the full-length BdpA. Adopted from Fig. 2 of Ref. [58]

Most existing physics-based molecular mechanics force fields treat electrostatic interactions between atoms as a collection of fixed point charges. In reality, they are anisotropic and polarizable. As a result, there is a significant effort in the development of polarizable force fields [69–78]. Polarizability is handled by many different approaches including fluctuating charges, induced dipoles, Drude oscillator and distributed multipoles. Yet, despite the effort in development, applications of polarizable force fields are limited to validation of the developed polarizable force fields and a few dynamics simulations of proteins [79]. As the development of polarizable force fields continues [79], their application to structure prediction (structure refinement, in particular) will likely commence soon.

2.2 Quantum mechanics and mixed QM/MM

A more fundamental approach is to treat atomic interactions quantum mechanically. Most existing applications of quantum mechanics (QM) to proteins are a hybrid approach in which QM and molecular mechanics (MM) are applied to treat different portions of a system (QM/MM) [80, 81]. Typically, a small portion of a system (e.g. the active site of an enzyme [82]) is treated quantum mechanically and is coupled to the remaining portion that is treated classically for efficient conformational sampling. Applications of QM to entire proteins became possible with the development of linear scaling techniques [83, 84] and were found to be useful for refining experimental structures [85–87]. In 2001, Liu et al. [88] demonstrated

that it is possible to simulate a system where the entire protein crambin is represented on the semi-empirical quantum–mechanical level and water molecules are modeled at the MM level for 350 ps. The simulation of the protein crambin provides a more accurate description of structural detail than regular MM simulations, when compared to the high-resolution X-ray structure. Zhu et al. [89] further showed that the gas-phase and solution structures of non-natural beta- and mixed alpha/beta- peptides can be predicted by an approximate density functional method for peptides coupled with a MM model for the solvent. Renfrew also found that quantum mechanics allows a more accurate placement of side chains [90]. More recently, a new approach was proposed where valence and core electrons are treated at the QM and MM levels, respectively [91–93]. The resulting X-Pol model has been used to simulate the protein BPTI in water for 50 ps. These studies highlight the potential utility of QM/MM in protein structure prediction as computing power further improves. These ab initio physics-based approaches, however, are several orders of magnitude slower than molecular dynamics based on molecular mechanics force fields. They may prevail one day when GPU (Graphics processing unit) parallel processing [94–96] and specific hardware for molecular dynamics simulations [97] become mature techniques accessible to most researchers.

One of the most successful applications of quantum calculations to protein structures is their ability to make highly accurate structure prediction from NMR chemical shifts [98–100]. Several groups have achieved a 2.0 Å or better resolution for predicted protein structures by employing fragment-based, structure prediction techniques with NMR chemical shifts as the only experimental restraints [101–107].

3 Knowledge-based potentials

While purely physics-based approaches may have the potential to achieve accurate protein structure prediction in the future, it makes practical sense to take advantage of known sequence and structural information, as appropriate for aiding protein structure prediction. Knowledge-based information can be employed to derive restraints in order to achieve a significant reduction in the conformational sampling space; knowledge-based (free) energy functions have been applied rather successfully to discriminate the native conformations from other non-native ones. Here, we will limit our discussion on all-atom knowledge-based energy functions because they are required for high-resolution structure prediction and are usually more accurate than residue-level knowledge-based energy functions.

3.1 All-atom distance-dependent potentials

A knowledge-based or statistical energy function is obtained directly from statistical analysis of known experimental protein structures [108, 109]. Unlike physics-based energy functions, an all-atom statistical energy function is a potential of mean force and, thus, allows direct and efficient evaluation of the free energy involved in folding and binding of proteins. Developing distance-dependent statistical energy functions at the atomic level is a relatively new, under-explored approach, compared to distance-dependent all-atom physics-based force fields [26–28]. Although the residue-level distance-dependent potential was developed by Sippl in 1990 [110], the first all-atom distance-dependent statistical potential was not obtained until 1998 by Samudrala and Moult [111]. Only a few more have been developed since [112–120].

Different statistical energy functions differ in the reference states employed to estimate the expected number of atomic pairs at a given distance in the absence of any interaction. Samudrala and Moult used a conditional probability function [111], while Lu and Skolnick employed a quasi-chemical approximation [113]. The common approximation behind the two methods is the "uniform density" reference state [108] that statistically averages over the observed state for the distance dependence [110]. Zhou and Zhou proposed to employ uniformly distributed points in a finite-size sphere for the reference state (Distance-scaled Finite Ideal-gas REference state, DFIRE) [114] that led to an approximate analytical expression for the distance dependence. Shen et al. further refined the analytical expression to account for varied protein sizes and led to the DOPE (Discrete Optimized Potential Energy) energy function [115]. Cheng employed a free-rotating and self-avoiding chain model as the reference state to account for the effect of covalently bonded backbone [120]. The difference between these two new techniques and DFIRE is typically small [120–122].

The relatively slow development of all-atom knowledge-based energy functions is largely because a statistical energy function is not considered to be theoretically rigorous [123–125] and is thought to be useful for coarse-grained models only. Moreover, an all-atom statistical potential is often suspected to be less reliable than an all-atom physics-based energy function. However, all-atom statistical energy functions have been found to be comparable to, or more accurate than, physics-based energy functions in loop selections [126], restoring partially denatured segments with secondary structures [127], and refining near-native structures [128]. In restoring partially denatured segments [127], both explicit and implicit solvation physics-based force fields were less successful than the DFIRE energy function [114] together with a clustering

method. Moreover, specific interactions obtained from a statistical approach are directly comparable to quantum calculations. Morozov et al. [129] showed an excellent agreement between a statistical hydrogen-bonding potential and quantum mechanical calculations. Gillis et al. [130] illustrated that statistical descriptions of cation–π and amino–π interactions have a significant correlation with quantum calculations at the Hartree–Fock and the second-order Möller–Plesset perturbation theory levels. The correlation coefficient is 0.96. By comparison, the correlation coefficient between quantum calculations and the results from the physics-based energy function CHARMM [27] is 0.89. In addition, Zhou et al. showed that a DFIRE-based statistical potential has some characteristics of a physics-based energy function in terms of database independence and transferability [131–134]. These studies indicate that statistical energy functions are valuable counterpart to physics-based energy functions, even at the detailed atomic level. Thus, all-atom knowledge-based energy functions will likely play increasingly active roles in structure prediction beyond ranking decoy structures. For example, Yang and Zhou employed an improved version of DFIRE (DFIRE 2.0) based on finer grids to make an ab initio folding of terminal segments with secondary structures [122].

3.2 All-atom orientation-dependent potentials

Specific folding and binding of proteins rely on specific interactions. Evidence is abundant that many interactions are more specific and orientation dependent than what are described by existing statistical energy functions. The most well-studied specific interaction for protein folding is hydrogen-bonding interaction [135]. Hydrogen-bonding interaction is commonly described as an individual, physical or statistical term in many empirical functions for proteins (e.g. Refs. [23, 136–138]). However, hydrogen-bonding is only a special case of polar–polar interaction. The interaction between polar atoms that are not hydrogen-bonded should be orientation dependent as well. There is evidence that this orientation dependence plays an important role in the formation of α-helices and β-sheets [139–142]. Additionally, the interaction between polar and non-polar atoms is likely orientation dependent because the hydrophobic effect is caused by the re-orientation of water molecules (polar atoms) near a hydrophobic surface [143]. The orientation dependence described above is part of the physics-based approach through electrostatic interactions, but not yet accounted for by statistical energy functions. Recent advances in statistical orientation-dependent potentials focused on coarse-grained models [130, 144–147], rather than a systematic treatment of polar interactions on an atomic level.

Recently, Yang and Zhou introduced a dipolar DFIRE (dDFIRE) that treats polar atoms separately from non-polar atoms [148]. In this method, each polar atom is no longer approximated as a point but is a point with a direction. The directions of polar atoms are defined by the covalent bond vectors between heavy atoms. If a polar atom (e.g. main chain oxygen) is bonded with only one heavy atom, the direction of the polar atom is determined by the bond vector. If a polar atom (e.g. main chain nitrogen) is bonded with two heavy atoms, the direction of the polar atom is determined by the sum of two bond vectors. Polar atoms bonded with three heavy atoms (e.g. backbone nitrogen of residue proline) are approximated as non-polar atoms. Figure 2 displays all defined directions of polar atoms in 20 amino acid residues. Once the directions of polar atoms are defined, orientation-dependent polar interactions can be extracted from known protein structures based on distance and orientation angles of physical interactions of dipoles. Application of the DFIRE energy function to ab initio refolding of protein terminal segments with secondary structure elements indicates that hydrogen-bonded interactions alone are not enough to make high-resolution prediction of segment structures with secondary structure elements [148]. Specific interactions between polar atoms and between polar and non-polar atoms all contribute significantly to the prediction accuracy of the structure of a terminal segment. An all-atom orientation-dependent knowledge-based energy function has also been extracted with rigid block approximation in the absence of distance dependence and found to be useful for side chain modeling [149–151].

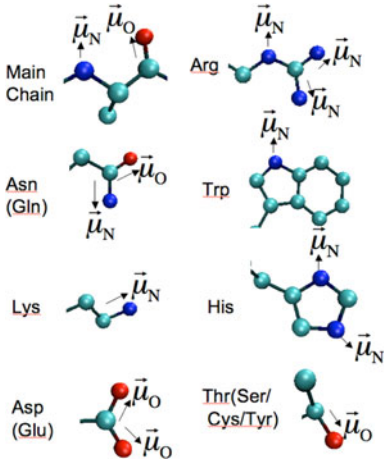

Fig. 2 Directions of all polar atoms for the main chain (*top left*) and the side chains of all amino acid residues. One diagram, sometimes, shows several residues with similar side chain structures for polar atoms (e.g. –OH/SH group in Thr, Ser, Cys and Tyr)

4 Conformational sampling

In addition to the lack of an accurate energy function, another bottleneck facing protein structure prediction is conformational space sampling [152]. This can be illustrated by the fact that from CASP 6 to CASP 8, some reasonable predictions were made for free-modeling targets with less than 100 residues but none for proteins with more than 100 residues [153]. Because several review articles provided an excellent overview on conformational sampling techniques [154–159] and a comprehensive review would require a separate article, we will only highlight a few newly developed sampling techniques that were implemented for protein folding and/or structure prediction. In particular, we will not discuss coarse-grained models [160–162] in this review as they have become a commonly used tool for speeding up sampling.

4.1 Barrier crossing/flattening techniques

Efficient sampling of protein conformational space is challenging because the energy landscape of proteins has numerous barriers that prevent proteins from moving freely from one conformational state to another. How to efficiently cross these energy barriers is the aim of many sampling techniques. They can be generally classified into methods modifying potential energy landscape such as umbrella sampling [163] and accelerated molecular dynamics [164, 165], methods employing a generalized ensemble of the system (multiple copies) such as replica exchange [166] and parallel tempering [167], and combinations of the two techniques such as simulated tempering [168, 169]. These three approaches have been substantially improved and/or implemented for protein structure prediction and folding in recent studies [154–157, 159]. A Grow-to-Fit method that reduces energy barriers due to side chain packing has been developed for the assignment of protein side chains using molecular mechanics force fields [170]. Among more recent examples, an improved accelerated molecular dynamics [171] demonstrated fast folding of Trp-CAGE and Trpzip2 [44, 172]. In this method, the energy surface is flattened to accelerate the barrier crossing process. Significantly faster convergence of thermodynamics properties of Trpzip2 [173] was observed by coupling replica exchange simulations to a non-Boltzmann structure reservoir generated from a high-temperature simulation [174, 175]. Replica exchange simulations were optimized by replica quenching [176] and reconstructing replica flow in the temperature ladder from first passage time [177]. Replica exchange simulations are also combined with specific biased potential such as hydrogen-bonding bias potential [178], repulsive and side chain interactions [179] and backbone-biased potential [180] for enhanced sampling. Enhanced sampling was also achieved by adaptive sampling of networks called Markov State Models [181]. Iteratively generating bias potentials targeting density of states has been shown to enhance the sampling of Go-type models [182, 183]. Similar to replica exchange, a forced random walk in temperature space allows a single simulation trajectory to traverse within a predetermined range of temperature to achieve accelerated sampling in MD simulations of small proteins with explicit solvent [184, 185]. A method has been proposed in which the simulation is initially performed at high temperature to sample the conformational space that is divided into smaller space within which subsequent room-temperature simulations are performed [186, 187]. Quick convergence was also demonstrated by coupling the replica exchange method with a general bias potential that does not correlate with the native protein structure [188–190] and by performing orthogonal space random walk [191]. Applications of these novel techniques are mostly limited to molecular mechanics force field simulations on peptides and/or a few small proteins, and a comprehensive comparison between different techniques is yet to be available. Their effectiveness on larger proteins of realistic size and knowledge-based energy functions is not known.

4.2 Local-guided/biased sampling

Another method to increase sampling efficiency is to restrict the conformational space to be sampled. The fragment-based approach was introduced as a technique to reduce the conformational space by focusing on sampling of known native local structures only. However, it has been found challenging to recognize structurally similar fragments or templates from a prebuilt structure/fragment library [25] because these structures are built using a preset threshold of structural or sequence similarity. As a result, these structures are similar but not identical to the structure of interest. Somewhat random imperfections in these fragments/templates make it difficult to design a universal energy function to recognize them and to make a correct assembly despite their imperfections. This adds more demands to the grand challenge of developing an accurate energy function for protein folding and structure prediction [192]. In addition, fragment rigidity may make it difficult to reach near-native structures for some proteins. Indeed, Hegler et al. found that under the same energy function, fragment-based sampling of larger proteins (>70 residues) encounters kinetic limitation that is not seen in unrestrictive molecular dynamics [193]. Kim et al. further showed that sampling is often limited by the inability to sample rarely occurring torsion angles of a few residues [194].

One approach to conformational sampling is to guide it by hierarchical folding pathways. Ozkan et al. predicted structures by zipping (local folding) and assembly [61].

This method involves independent folding of local structures and growth (zip) or coalescence (assemble) of these structures with other structures and achieved encouraging results in CASP [60]. DeBartolo et al. fixed secondary structure iteratively during Monte Carlo folding simulations [195] and further improved the technique with multiple sequence alignment for torsion angle sampling distribution with DOPE and other empirical energy functions including a collapse term [187]. For a benchmark of 12 small proteins, their method achieved higher accuracy for secondary structure prediction than sequence-based prediction, and the accuracy of their tertiary structure prediction is within 6Å for 8 of 12 proteins [196]. Brunette and Brock proposed a model-based search that guides the sampling with partially folded models during simulations with the Rosetta energy function [197]. The proposed method did sample lower energy conformations than the simple Monte Carlo technique in Rosetta. However, the test is quite limited because in the absence of homologous structural fragments, both the proposed method and Rosetta performed poorly for 29 out of 32 testing proteins, perhaps due to limited sampling in their experiments on homolog-free structure prediction.

A similar approach employs locally biased sampling. Hegler et al. showed improved sampling by a local energy term that is derived from local fragment sequence alignment and tested their technique in CASP 8 [193]. Chen et al. developed a move set for protein folding based on statistical knowledge of torsion angles [198]. Their test is limited to a native-contact biased model. Yang and Liu improved protein sampling by genetic algorithm in discrete backbone dihedral angle space [199]. Zhao et al. sampled the backbone via local biases from a probabilistic, conditional random/neutral fields model on the relation between protein sequences and backbone structures [200–202]. Their application to CASP 8 targets is on a par with other best predictors. Similarly, Boomsma et al. [203, 204] proposed a generative, probabilistic model for local structure sampling. Testing of the technique was limited to the ability to sample near-native conformations.

To summarize, the above studies on local-guided/biased sampling suggested significant potential. However, large-scale benchmark tests and optimized integration with a suitable energy function with an all-atom model for final packing are needed to further improve or confirm the accuracy of protein structure prediction.

4.3 Secondary structure and torsion angle restraints

Another approach for reducing conformational space is to employ predicted secondary structures (e.g. [4, 205–209]). However, predicted secondary structure is often represented by coarse-grained three states of helices, coils and strands because the accuracy of predicting more than three states is too low to be useful [210]. Restraints based on predicted secondary structures are limited to ideal shapes of helical and strand residues only because coil residues do not have a well-defined structure.

One way to avoid the limitation of predicted secondary structures is to predict backbone torsion angles. However, multistate torsion angles are as difficult as secondary structure to predict [211–215]. For example, Zimmermann and Hansmann [216] obtained a three-state prediction accuracy of 79%, the same level of accuracy for secondary structure prediction [217]. Recently, Zhou et al. demonstrated that real-value backbone torsion angles could be predicted with reasonable accuracy [218–220]. One limitation of direct real-value angle prediction is that many predicted angles are located in sterically prohibited regions. This limitation was remedied by mixing the advantage of multistate prediction (avoiding prohibited regions) and that of real-value prediction (continuous representation) [221]. This was done by making a two-state peak prediction first and followed by predicting the deviation from the predicted peak. The final method (SPINE XI) further refines the prediction by a conditional random field model and leads to an accurate prediction of real-value torsion angles that is close to the accuracy of angles derived from NMR chemical shifts with the methods TALOS [222] and TOPOS [107]. Multistate prediction derived from predicted real values by SPINE XI is even more accurate than predicted states from those methods dedicated to multistate prediction. For example, a three-state prediction accuracy based on a five-residue block of 8 consecutive torsion angles defined by multistate predictor LOCUSTRA is 81% by SPINE XI and 79% by LOCUSTRA [216].

Predicted real values of torsion angles serve as significantly more powerful restraints for fragment-free protein structure prediction than predicted secondary structure. Using a benchmark of 16 proteins and defining success as the ability to sample a structure with less than 6 Å RMSD from the native structure within top 15 predicted structures, Faraggi et al. [221] showed that the success rate increases from 6 with predicted secondary structure as restraints, 10 with predicted real-value torsion angles for helical and strand residues only as restraints, to 12 with predicted real-value torsion angles as restraints for all residues. The median RMSD value for these three cases decreased to 6.3, 5.4 and 4.3 Å RMSD, respectively. Here, torsion angles are not restrained if they are within the predicted ranges of error, restrained harmonically if greater than predicted ranges but within twice the predicted ranges, and subjected to a constant penalty if above twice the predicted ranges. This result demonstrates the importance of real-value prediction (67% increase in success rate and 14% reduction in

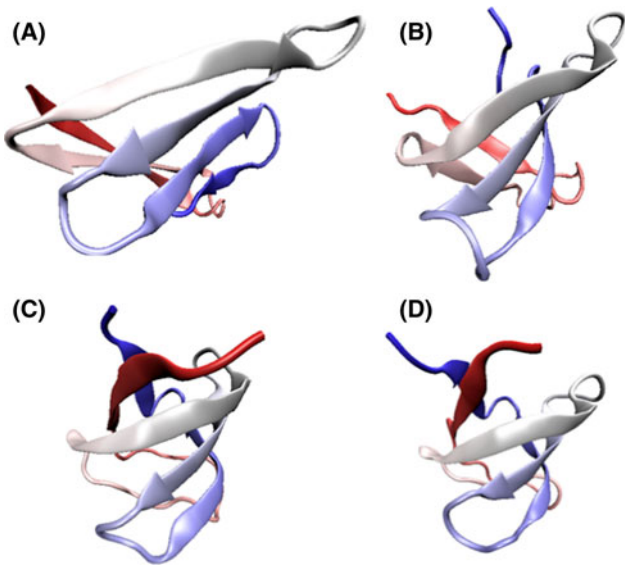

Fig. 3 The best structures in top 15 predicted structures obtained by predicted secondary structure as restraints (8.5 Å RMSD, **a**), predicted real-value torsion angles for strand residues as restraints (6.9 Å RMSD, **b**), predicted real-value torsion angles for all residues as restraints (3.1 Å RMSD, **c**) are compared to the native structure (**d**) for the SH3 domain protein (pdb id 1shf). It is clear that only real-value prediction allows the sampling of bended strand conformation

the median RMSD value), and of coil residue restraints (another 20% increase in success rate and 20% reduction in the median RMSD value) in structure prediction. In Fig. 3, the case of the SH3 domain protein (PDB ID: 1shf) is given to illustrate the importance of real-value torsion angles for sampling of non-ideal beta strands.

5 Summary and outlook

Some progress has been made towards ab initio prediction of protein structure by physics-based force fields. The progress, however, is limited to a few small helical or mixed helical and strand proteins. With intensive development in next generation force fields and advances in computing power, there is hope that physics-based methods may emerge as a powerful tool for structure prediction. Meanwhile, lack of progress in knowledge-based approaches for template-free modeling calls for fresh ideas. This review describes several trends in recent literature: development of physical, polarizable force fields and specific orientation-dependent all-atom, statistical energy functions, and smoothing or reduction of sampling space via improved sampling techniques and local bias or restraints.

One noticeable trend is the increased use of molecular force fields coupled with solvation free energy for scoring or ranking near-native conformations generated from conformational sampling. This approach, however, neglects the contribution of entropy (dynamic motions) in stabilizing native conformations of proteins because typical molecular force fields characterize the energy rather than free-energy surface of proteins. A more effective scoring function would require re-training all force field parameters (van der Waals parameters and partial charges) to mimic the free-energy surface and allow a more accurate account of the effect of atomic movement on solvent dielectric [223, 224].

In summary, recent studies suggest that it is possible to reach near-native structures without borrowing native fragments or templates from other proteins. Although fully flexible conformational search is one or more orders of magnitude slower than rigid fragment–based search, it has the potential to reach more accurate, high-resolution structure needed for function prediction and analysis. This fully flexible, continuous sampling approach coupled with more specific, accurate energy functions will likely lead to the next generation methods in structure prediction.

Acknowledgments We thank with Professors K. A. Dill, Y. Q. Gao, J. Ma, V. S. Pande, J. Skolnick, J. Xu, W. Yang for very helpful discussions. This work was supported by the National Institutes of Health grants GM R01 085003 to Y. Z., GM 067168 to D. Y. and Y. Z. and R01GM079383 to D.Y. and by the National Science Foundation of China grant 30870474 to HL.

References

1. Pettersson E, Lundeberg J, Ahmadian A (2009) Generations of sequencing technologies. Genomics 93:105–111
2. Terwilliger TC, Stuart D, Yokoyama S (2009) Lessons from structural genomics. Ann Rev Biophys 38:371–383
3. Moult J, Fidelis K, Kryshtafovych A, Rost B, Tramontano A (2009) Critical assessment of methods of protein structure prediction—round VIII. Proteins 77(9):1–4
4. 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 268:209–225
5. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402
6. Yona G, Levitt M (2002) Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. J Mol Biol 315:1257–1275
7. 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 310:243–257
8. Shan Y, Wang G, Zhou HX (2001) Fold recognition and accurate query-template alignment by a combination of PSI-BLAST and threading. Proteins 42:23–37

9. Pei J, Sadreyev R, Grishin NV (2003) PCMA: fast and accurate multiple sequence alignment based on profile consistency. Bioinformatics 19:427–428
10. Panchenko AR, Marchler-Bauer A, Bryant SH (2000) Combination of threading potentials and sequence profiles improves fold recognition. J Mol Biol 296:1319–1331
11. Kim D, Xu D, Guo JT, Ellrott K, Xu Y (2003) PROSPECT II: protein structure prediction program for genome-scale applications. Protein Eng 16:641–650
12. Kelley LA, MacCallum RM, Sternberg MJ (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 299:499–520
13. Karplus K, Barrett C, Hughey R (1998) Hidden Markov models for detecting remote protein homologies. Bioinformatics 14:846–856
14. Jones DT (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 287:797–815
15. Elofsson A, Fischer D, Rice DW, Le Grand SM, Eisenberg D (1996) A study of combined structure/sequence profiles. Fold Des 1:451–461
16. Zhou H, Zhou Y (2004) Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins 55:1005–1013
17. Zhou H, Zhou Y (2005) Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins 58:321–328
18. Xu J, Li M, Kim D, Xu Y (2003) RAPTOR: optimal protein threading by linear programming. J Bioinform Comput Biol 1:95–117
19. Fischer D (2000) Hybrid fold recognition: combining sequence derived properties with evolutionary information. Pac Symp Biocomput 119–130
20. Lundstrom J, Rychlewski L, Bujnicki J, Elofsson A (2001) Pcons: a neural-network-based consensus predictor that improves fold recognition. Protein Sci 10:2354–2362
21. Bujnicki JM, Elofsson A, Fischer D, Rychlewski L (2001) Structure prediction meta server. Bioinformatics 17:750–751
22. Chivian D, Kim DE, Malmstrom L, Bradley P, Robertson T, Murphy P, Strauss CEM, Bonneau R, Rohl CA, Baker D (2003) Automated prediction of CASP-5 structures using the Robetta server. Proteins 53:524–533
23. Zhang Y, Skolnick J (2004) Automated structure prediction of weakly homologous proteins on a genomic scale'. Proc Natl Acad Sci USA 101:7594–7599
24. Zhang Y (2009) Protein structure prediction: when is it useful? Curr Opin Struct Biol 19:145–155
25. Bujnicki JM (2006) Protein-structure prediction by recombination of fragments. Chembiochem 7:19–27
26. Weiner SJ, Kollman P, Nguyen D, Case D (1986) An all atom force field for simulations of proteins and nucleic acids. J Comput Chem 7:230–252
27. Brooks BR, Brooks CL, Mackerell AD, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M (2009) CHARMM: the biomolecular simulation program. J Comput Chem 30:1545–1614
28. Ponder JW, Case DA (2003) Force fields for protein simulations. Adv Protein Chem 66:27
29. McCammon JA, Gelin BR, Karplus M (1977) Dynamics of folded proteins. Nature 267:585–590
30. Bash PA, Singh UC, Langridge R, Kollman PA (1987) Free-energy calculations by computer-simulation. Science 236:564–568
31. McCammon JA (1991) Free energy from simulations. Curr Opin Struc Biol 1:196–200
32. Brooks CL (2002) Protein and peptide folding explored with molecular simulations. Accounts Chem Res 35:447–454
33. Seibert MM, Patriksson A, Hess B, van der Spoel D (2005) Reproducible polypeptide folding and structure prediction using molecular dynamics simulations. J Mol Biol 354:173–183
34. Duan Y, Kollman PA (1998) Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science 282:740–744
35. Papoian GA, Ulander J, Eastwood MP, Luthey-Schulten Z, Wolynes PG (2004) Water in protein structure prediction. Proc Natl Acad Sci USA 101:3352–3357
36. Lee MR, Tsai J, Baker D, Kollman PA (2001) Molecular dynamics in the endgame of protein structure prediction. J Mol Biol 313:417–430
37. Fan H, Mark AE (2004) Refinement of homology-based protein structures by molecular dynamics simulation techniques. Protein Sci 13:211–220
38. Vieth M, Kolinski A, Brooks CL, Skolnick J (1994) Prediction of the folding pathways and structure of the Gcn4 leucine-zipper. J Mol Biol 237:361–367
39. Simmerling C, Lee MR, Ortiz AR, Kolinski A, Skolnick J, Kollman PA (2000) Combining MONSSTER and LES/PME to predict protein structure from amino acid sequence: application to the small protein CMTI-1. J Am Chem Soc 122:8392–8402
40. Chopra G, Summa CM, Levitt M (2008) Solvent dramatically affects protein structure refinement. Proc Natl Acad Sci USA 105:20239–20244
41. Wagner F, Simonson T (1999) Implicit solvent models: combining an analytical formulation of continuum electrostatics with simple models of the hydrophobic effect. J Comput Chem 20:322–335
42. Lazaridis T, Karplus M (2000) Effective energy functions for protein structure prediction. Curr Opin Struct Biol 10:139–145
43. Roux B, Simonson T (1999) Implicit solvent models. Biophys Chem 78:1–20
44. Yang LJ, Shao Q, Gao YQ (2009) Thermodynamics and folding pathways of Trpzip2: an accelerated molecular dynamics simulation study. J Phys Chem B 113:803–808
45. Roy S, Goedecker S, Field MJ, Penev E (2009) A minima hopping study of all-atom protein folding and structure prediction. J Phys Chem B 113:7315–7321
46. Zhu J, Alexov E, Honig B (2005) Comparative study of generalized Born models: Born radii and peptide folding. J Phys Chem B 109:3008–3022
47. Liu YX, Beveridge DL (2002) Exploratory studies of ab initio protein structure prediction: multiple copy simulated annealing, AMBER energy functions, and a Generalized Born/Solvent Accessibility solvation model. Proteins 46:128–146
48. Vila JA, Ripoll DR, Scheraga HA (2003) Atomically detailed folding simulation of the B domain of staphylococcal protein A from random structures. Proc Natl Acad Sci USA 100:14812–14816
49. Katagiri D, Fuji H, Neya S, Hoshino T (2008) Ab initio protein structure prediction with force field parameters derived from water-phase quantum chemical calculation. J Comput Chem 29:1930–1944
50. Simmerling C, Strockbine B, Roitberg AE (2002) All-atom structure prediction and folding simulations of a stable protein. J Am Chem Soc 124:11258–11259
51. Pitera JW, Swope W (2003) Understanding folding and design: replica-exchange simulations of "Trp-cage" fly miniproteins. Proc Natl Acad Sci USA 100:7587–7592

52. Chowdhury S, Lee MC, Xiong GM, Duan Y (2003) Ab initio folding simulation of the Trp-cage mini-protein approaches NMR resolution. J Mol Biol 327:711–717
53. Tsui V, Case DA (2000) Theory and applications of the generalized Born solvation model in macromolecular simulations. Biopolymers 56:275–291
54. Lei HX, Wu C, Liu HG, Duan Y (2007) Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. Proc Natl Acad Sci USA 104:4925–4930
55. Lei HX, Duan Y (2007) Two-stage folding of HP-35 from ab initio simulations. J Mol Biol 370:196–206
56. 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 323:927–937
57. Bowman GR, Beauchamp KA, Boxer G, Pande VS (2009) Progress and challenges in the automated construction of Markov state models for full protein systems. J Chem Phys 131
58. Lei HX, Wu C, Wang ZX, Zhou YQ, Duan Y (2008) Folding processes of the B domain of protein A to the native state observed in all-atom ab initio folding simulations. J Chem Phys 128:235105
59. Lei HX, Wang ZX, Wu C, Duan Y (2009) Dual folding pathways of an alpha/beta protein from all-atom ab initio folding simulations. J Chem Phys 131:165105
60. Shell MS, Ozkan SB, Voelz V, Wu GHA, Dill KA (2009) Blind test of physics-based prediction of protein structures. Biophys J 96:917–924
61. Ozkan SB, Wu GA, Chodera JD, Dill KA (2007) Protein folding by zipping and assembly. Proc Natl Acad Sci USA 104:11987–11992
62. Liwo A, Arlukowicz P, Czaplewski C, Oldziej S, Pillardy J, Scheraga HA (2002) A method for optimizing potential-energy functions by a hierarchical design of the potential-energy landscape: application to the UNRES force field. Proc Natl Acad Sci USA 99:1937–1942
63. Srinivasan R, Rose GD (2002) Ab initio prediction of protein structure using LINUS. Proteins 47:489–495
64. Oldziej S, Czaplewski C, Liwo A, Chinchio M, Nanias M, Vila JA, Khalili M, Arnautova YA, Jagielska A, Makowski M, Schafroth HD, Kazmierkiewicz R, Ripoll DR, Pillardy J, Saunders JA, Kang YK, Gibson KD, Scheraga HA (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 102:7547–7552
65. Jagielska A, Wroblewska L, Skolnick J (2008) Protein model refinement using an optimized physics-based all-atom force field. Proc Natl Acad Sci USA 105:8268–8273
66. Krieger E, Darden T, Nabuurs SB, Finkelstein A, Vriend G (2004) Making optimal use of empirical energy functions: force-field parameterization in crystal space. Proteins 57:678–683
67. Lin MS, Fawzi NL, Head-Gordon T (2007) Hydrophobic potential of mean force as a solvation function for protein structure prediction. Structure 15:727–740
68. Wang J, Luo R (2010) Assessment of linear finite-difference Poisson-Boltzmann solvers. J Comput Chem 31:1689–1698
69. Wang ZX, Zhang W, Wu C, Lei HX, Cieplak P, Duan Y (2006) Strike a balance: optimization of backbone torsion parameters of AMBER polarizable force field for simulations of proteins and peptides. J Comput Chem 27:994
70. Tan YH, Tan CH, Wang J, Luo R (2008) Continuum polarizable force field within the Poisson-Boltzmann framework. J Phys Chem B 112:7675–7688
71. Stork M, Tavan P (2007) Electrostatics of proteins in dielectric solvent continua II first applications in molecular dynamics simulations. J Chem Phys 126:166106
72. Patel S, Mackerell AD, Brooks CL (2004) CHARMM fluctuating charge force field for proteins: II—protein/solvent properties from molecular dynamics simulations using a nonadditive electrostatic model. J Comput Chem 25:1504–1514
73. Masella M, Borgis D, Cuniasse P (2008) Combining a polarizable force-field and a coarse-grained polarizable solvent model: application to long dynamics simulations of bovine pancreatic trypsin inhibitor. J Comput Chem 29:1707–1724
74. Kaminski GA, Stern HA, Berne BJ, Friesner RA, Cao YXX, Murphy RB, Zhou RH, Halgren TA (2002) Development of a polarizable force field for proteins via ab initio quantum chemistry: first generation model and gas phase tests. J Comput Chem 23:1515–1531
75. Grossfield A, Ren PY, Ponder JW (2003) Ion solvation thermodynamics from simulation with a polarizable force field. J Am Chem Soc 125:15671–15682
76. Warshel A, Levitt M (1976) Theoretical studies of enzymatic reactions: dielectric electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. J Mol Biol 103:227–249
77. Jiao D, Golubkov PA, Darden TA, Ren P (2008) Calculation of protein-ligand binding free energy by using a polarizable potential. Proc Natl Acad Sci USA 105:6290–6295
78. Piquemal JP, Cisneros GA, Reinhardt P, Gresh N, Darden TA (2006) Towards a force field based on density fitting. J Chem Phys 124:104101
79. Lopes PEM, Roux B, MacKerell AD (2009) Molecular modeling and dynamics studies with explicit inclusion of electronic polarizability: theory and applications. Theor Chem Acc 124:11–28
80. Warshel A, Bromberg A (1970) Oxidation of 4a, 4b-dihydrophenanthrenes. III. A theoretical study of the large kinetic isotope effect of deuterium in the initiation step of the thermal reaction with oxygen. J Chem Phys 52:1262–1269
81. Warshel A, Levitt M (1976) Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. J Mol Biol 103:227–249
82. Gao JL, Truhlar DG (2002) Quantum mechanical methods for enzyme kinetics. Annu Rev Phys Chem 53:467–505
83. Challacombe M, Schwegler E (1997) Linear scaling computation of the Fock matrix. J Chem Phys 106:5526–5536
84. Van Alsenoy C, Yu CH, Peeters A, Martin JML, Schafer L (1998) Ab initio geometry determinations of proteins. 1. Crambin. J Phys Chem A 102:2246–2251
85. Ryde U, Olsen L, Nilsson K (2002) Quantum chemical geometry optimizations in proteins using crystallographic raw data. J Comput Chem 23:1058–1070
86. Yu N, Li X, Cui GL, Hayik SA, Merz KM (2006) Critical assessment of quantum mechanics based energy restraints in protein crystal structure refinement. Protein Sci 15:2773–2784
87. Yu N, Yennawar HP, Merz KM (2005) Refinement of protein crystal structures using energy restraints derived from linear-scaling quantum mechanics. Acta Crystallogr D 61:322–332
88. Liu HY, Elstner M, Kaxiras E, Frauenheim T, Hermans J, Yang WT (2001) Quantum mechanics simulation of protein dynamics on long timescale. Proteins 44:484–489
89. Zhu X, Yethiraj A, Cui Q (2007) Establishing effective simulation protocols for beta- and alpha/beta-mixed peptides I. QM and QM/MM models. J Chem Theory Comput 3:1538–1549
90. Renfrew PD, Butterfoss GL, Kuhlman B (2008) Using quantum mechanics to improve estimates of amino acid side chain rotamer energies. Proteins 71:1637–1646
91. Song LC, Han JB, Lin YL, Xie WS, Gao JL (2009) Explicit polarization (X-Pol) potential using ab initio molecular orbital

theory and density functional theory. J Phys Chem A 113:11656–11664
92. Xie W, Orozco M, Truhlar DG, Gao J (2009) X-Pol potential: an electronic structure-based force field for molecular dynamics simulation of a solvated protein in water. J Chem Theory Comput 5:459–467
93. Xie WS, Gao JL (2007) Design of a next generation force field: the X-POL potential. J Chem Theory Comput 3:1890–1900
94. Stone JE, Phillips JC, Freddolino PL, Hardy DJ, Trabuco LG, Schulten K (2007) Accelerating molecular modeling applications with graphics processors. J Comput Chem 28:2618–2640
95. Friedrichs MS, Eastman P, Vaidyanathan V, Houston M, Legrand S, Beberg AL, Ensign DL, Bruns CM, Pande VS (2009) Accelerating molecular dynamic simulation on graphics processing units. J Comput Chem 30:864–872
96. Voelz VA, Bowman GR, Beauchamp K, Pande VS (2010) Molecular Simulation of ab initio protein folding for a millisecond folder NTL9(1-39). J Am Chem Soc 132:1526
97. Shaw DE, Deneroff MM, Dror RO, Kuskin JS, Larson RH, Salmon JK, Young C, Batson B, Bowers KJ, Chao JC, Eastwood MP, Gagliardo J, Grossman JP, Ho CR, Ierardi DJ, Kolossvary I, Klepeis JL, Layman T, Mcleavey C, Moraes MA, Mueller R, Priest EC, Shan YB, Spengler J, Theobald M, Towles B, Wang SC (2008) Anton, a special-purpose machine for molecular dynamics simulation. Commun ACM 51:91–97
98. Xu XP, Case DA (2001) Automated prediction of N-15, C-13(alpha), C-13(beta) and C-13 'chemical shifts in proteins using a density functional database. J Biomol NMR 21:321–333
99. Neal S, Nip AM, Zhang HY, Wishart DS (2003) Rapid and accurate calculation of protein H-1, C-13 and N-15 chemical shifts. J Biomol NMR 26:215–240
100. Meiler J (2003) PROSHIFT: protein chemical shift prediction using artificial neural networks. J Biomol NMR 26:25–37
101. Wishart DS, Arndt D, Berjanskii M, Tang P, Zhou J, Lin G (2008) CS23D: a web server for rapid protein structure generation using NMR chemical shifts and sequence data. Nucleic Acids Res 36:W496–W502
102. Shen Y, Vernon R, Baker D, Bax A (2009) De novo protein structure generation from incomplete chemical shift assignments. J Biomol NMR 43:63–78
103. Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu GH, Eletsky A, Wu YB, Singarapu KK, Lemak A, Ignatchenko A, Arrowsmith CH, Szyperski T, Montelione GT, Baker D, Bax A (2008) Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci USA 105:4685–4690
104. Robustelli P, Cavalli A, Vendruscolo M (2008) Determination of protein structures in the solid state from NMR chemical shifts. Structure 16:1764–1769
105. Montalvao RW, Cavalli A, Salvatella X, Blundell TL, Vendruscolo M (2008) Structure determination of protein-protein complexes using NMR chemical shifts: case of an endonuclease colicin-immunity protein complex. J Am Chem Soc 130:15990–15996
106. Gong HP, Shen Y, Rose GD (2007) Building native protein conformation from NMR backbone chemical shifts using Monte Carlo fragment assembly. Protein Sci 16:1515–1521
107. Cavalli A, Salvatella X, Dobson CM, Vendruscolo M (2007) Protein structure determination from NMR chemical shifts. Proc Natl Acad Sci USA 104:9615–9620
108. Tanaka S, Scheraga HA (1976) Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. Macromolecules 9:945–950
109. Miyazawa S, Jernigan RL (1985) Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 18:534–552
110. 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 213:859–883
111. Samudrala R, Moult J (1998) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 275:895–916
112. Mirzaie M, Eslahchi C, Pezeshk H, Sadeghi M (2009) A distance-dependent atomic knowledge-based potential and force for discrimination of native structures from decoys. Proteins 77:454–463
113. Lu H, Skolnick J (2001) A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins 44:223–232
114. Zhou HY, Zhou YQ (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 11:2714–2726
115. Shen MY, Sali A (2006) Statistical potential for assessment and prediction of protein structures. Protein Sci 15:2507–2524
116. Yoshidome T, Oda K, Harano Y, Roth R, Sugita Y, Ikeguchi M, Kinoshita M (2009) Free-energy function based on an all-atom model for proteins. Proteins 77:950–961
117. Ferrada E, Melo F (2009) Effective knowledge-based potentials. Protein Sci 18:1469–1485
118. Kamisetty H, Xing EP, Langmead CJ (2008) Free energy estimates of all-atom protein structures using generalized belief propagation. J Comput Biol 15:755–766
119. Ferrada E, Vergara IA, Melo F (2007) A knowledge-based potential with an accurate description of local interactions improves discrimination between native and near-native protein conformations. Cell Biochem Biophys 49:111–124
120. Cheng J, Pei JF, Lai LH (2007) A free-rotating and self-avoiding chain model for deriving statistical potentials based on protein structures. Biophys J 92:3868–3877
121. Eramian D, Shen MY, Devos D, Melo F, Sali A, Marti-Renom MA (2006) A composite score for predicting errors in protein structure models. Protein Sci 15:1653–1666
122. Yang YD, Zhou Y (2008) Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. Protein Sci 17:1212–1219
123. Thomas PD, Dill KA (1996) Statistical potentials extracted from protein structures: how accurate are they? J Mol Biol 257:457–469
124. BenNaim A (1997) Statistical potentials extracted from protein structures: are these meaningful potentials? J Chem Phys 107:3698–3706
125. Betancourt MR, Thirumalai D (1999) Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Protein Sci 8:361–369
126. Zhang C, Liu S, Zhou YQ (2004) Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential. Protein Sci 13:391–399
127. Zhu J, Xie L, Honig B (2006) Structural refinement of protein segments containing secondary structure elements: local sampling, knowledge-based potentials, and clustering. Proteins 65:463–479
128. Summa CM, Levitt M (2007) Near-native structure refinement using in vacuo energy minimization. Proc Natl Acad Sci USA 104:3177–3182
129. Morozov AV, Kortemme T, Tsemekhman K, Baker D (2004) Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum

mechanical calculations. Proc Natl Acad Sci USA 101:6946–6951
130. Gilis D, Biot C, Buisine E, Dehouck Y, Rooman M (2006) Development of novel statistical potentials describing cation-pi interactions in proteins and comparison with semiempirical and quantum chemistry approaches. J Chem Inf Model 46:884–893
131. Zhang C, Liu S, Zhu QQ, Zhou YQ (2005) A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. J Med Chem 48:2325–2335
132. Zhang C, Liu S, Zhou HY, Zhou Y (2004) The dependence of all-atom statistical potentials on structural training database. Biophys J 86:3349–3358
133. Zhou Y, Zhou HY, Zhang C, Liu S (2006) What is a desirable statistical energy function for proteins and how can it be obtained? Cell Biochem Biophys 46:165–174
134. Liu S, Zhang C, Zhou HY, Zhou Y (2004) A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. Proteins 56:93–101
135. Haber E, Anfinsen CB (1961) Regeneration of enzyme activity by air oxidation of reduced subtilisin-modified ribonuclease. J Biol Chem 236:422–424
136. Kortemme T, Morozov AV, Baker D (2003) An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. J Mol Biol 326:1239–1259
137. Pillardy A, Czaplewski C, Liwo A, Lee J, Ripoll DR, Kazmierkiewicz R, Oldziej S, Wedemeyer WJ, Gibson KD, Arnautova YA, Saunders J, Ye YJ, Scheraga HA (2001) Recent improvements in prediction of protein structure by global optimization of a potential energy function. Proc Natl Acad Sci USA 98:2329–2333
138. 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 USA 98:10125–10130
139. Maccallum PH, Poet R, Milnerwhite EJ (1995) Coulombic interactions between partially charged main-chain atoms not hydrogen-bonded to each other influence the conformations of alpha-helices and antiparallel beta-sheet—a new method for analyzing the forces between hydrogen-bonding groups in proteins includes all the coulombic interactions. J Mol Biol 248:361–373
140. Maccallum PH, Poet R, Milnerwhite EJ (1995) Coulombic attractions between partially charged main-chain atoms stabilize the right-handed twist found in most beta-strands. J Mol Biol 248:374–384
141. Deane CM, Allen FH, Taylor R, Blundell TL (1999) Carbonyl-carbonyl interactions stabilize the partially allowed Ramachandran conformations of asparagine and aspartic acid. Protein Eng 12:1025–1028
142. Paulini R, Muller K, Diederich F (2005) Orthogonal multipolar interactions in structural chemistry and biology. Angew Chem Int Edit 44:1788–1805
143. Blokzijl W, Engberts JBFN (1993) Hydrophobic effects—opinions and facts. Angew Chem Int Edit 32:1545–1579
144. Wu YH, Lu MY, Chen MZ, Li JL, Ma JP (2007) OPUS-Ca: a knowledge-based potential function requiring only C alpha positions. Protein Sci 16:1449–1463
145. Miyazawa S, Jernigan RL (2005) How effective for fold recognition is a potential of mean force that includes relative orientations between contacting residues in proteins? J Chem Phys 122:024901
146. Hoppe C, Schomburg D (2005) Prediction of protein thermostability with a direction- and distance-dependent knowledge-based potential. Protein Sci 14:2682–2692
147. Buchete NV, Straub JE, Thirumalai D (2004) Development of novel statistical potentials for protein fold recognition. Curr Opin Struc Biol 14:225–232
148. Yang YD, Zhou Y (2008) Specific interactions for ab initio folding of protein terminal regions with secondary structures. Proteins 72:793–803
149. Lu M, Dousis AD, Ma J (2008) OPUS-Rota: a fast and accurate method for side-chain modeling. Protein Sci 17:1576–1585
150. Ma JP (2009) Explicit orientation dependence in empirical potentials and its significance to side-chain modeling. Accounts Chem Res 42:1087–1096
151. Lu M, Dousis AD, Ma J (2008) OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. J Mol Biol 376:288–301
152. Bradley P, Misura KMS, Baker D (2005) Toward high-resolution de novo structure prediction for small proteins. Science 309:1868–1871
153. Ben-David M, Noivirt-Brik O, Paz A, Prilusky J, Sussman JL, Levy Y (2009) Assessment of CASP8 structure predictions for template free targets. Proteins 77:50–65
154. Liwo A, Czaplewski C, Oldziej S, Scheraga HA (2008) Computational techniques for efficient conformational sampling of proteins. Curr Opin Struc Biol 18:134–139
155. Lei HX, Duan Y (2007) Improved sampling methods for molecular simulation. Curr Opin Struc Biol 17:187–191
156. Christen M, Van Gunsteren WF (2008) On searching in, sampling of, and dynamically moving through conformational space of biomolecular systems: a review. J Comput Chem 29:157–166
157. Knight JL, Brooks CL (2009) Lambda-dynamics free energy simulation methods. J Comput Chem 30:1692–1700
158. de Bakker PI, Furnham N, Blundell TL, DePristo MA (2006) Conformer generation under restraints. Curr Opin Struct Biol 16:160–165
159. Leone V, Marinelli F, Carloni P, Parrinello M (2010) Targeting biomolecular flexibility with metadynamics. Curr Opin Struct Biol 20:148–154
160. Tozzini V (2005) Coarse-grained models for proteins. Curr Opin Struct Biol 15:144–150
161. Tozzini V (2010) Multiscale modeling of proteins. Accounts Chem Res 43:220–230
162. Sherwood P, Brooks BR, Sansom MSP (2008) Multiscale methods for macromolecular simulations. Curr Opin Struct Biol 18:630–640
163. Torrie GM, Valleau JP (1977) Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling. J Chem Phys 23:187–199
164. Voter AF (1997) Hyperdynamics: accelerated molecular dynamics of infrequent events. Phys Rev Lett 78:3908–3911
165. Hamelberg D, Mongan J, McCammon JA (2004) Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. J Chem Phys 120:11919–11929
166. Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. Chem Phys Lett 314:141–151
167. Hansmann UHE (1997) Parallel tempering algorithm for conformational studies of biological molecules. Chem Phys Lett 281:140–150
168. Lyubartsev AP, Martsinovski AA, Shevkunov SV, Vorontsovvelyaminov PN (1992) New approach to monte-carlo calculation of the free-energy—method of expanded ensembles. J Chem Phys 96:1776–1783
169. Marinari E, Parisi G (1992) Simulated tempering—a new Monte-Carlo scheme. Europhys Lett 19:451–458
170. Zhang W, Duan Y (2006) Grow to fit molecular dynamics (G2FMD): an ab initio method for protein side-chain assignment and refinement. Protein Eng Des Sel 19:55–65

171. Gao YQ, Yang LJ (2006) On the enhanced sampling over energy barriers in molecular dynamics simulations. J Chem Phys 125:114103
172. Yang LJ, Grubb MP, Gao YQ (2007) Application of the accelerated molecular dynamics simulations to the folding of a small protein. J Chem Phys 126:125102
173. Roitberg AE, Okur A, Simmerling C (2007) Coupling of replica exchange simulations to a non-Boltzmann structure reservoir. J Phys Chem B 111:2415–2418
174. Brown S, Head-Gordon T (2003) Cool walking: a new Markov chain Monte Carlo sampling method. J Comput Chem 24:68–76
175. Li HZ, Li GH, Berg BA, Yang W (2006) Finite reservoir replica exchange to enhance canonical sampling in rugged energy surfaces. J Chem Phys 125:144902
176. Li XF, Latour RA, Stuart SJ (2009) TIGER2: an improved algorithm for temperature intervals with global exchange of replicas. J Chem Phys 130:174106
177. Nadler W, Meinke JH, Hansmann UHE (2008) Folding proteins by first-passage-times-optimized replica exchange. Phys Rev E 78
178. Vreede J, Wolf MG, de Leeuw SW, Bolhuis PG (2009) Reordering hydrogen bonds using Hamiltonian replica exchange enhances sampling of conformational changes in biomolecular systems. J Phys Chem B 113:6484–6494
179. Mu YG (2009) Dissociation aided and side chain sampling enhanced Hamiltonian replica exchange. J Chem Phys 130:164107
180. Kannan S, Zacharias M (2007) Enhanced sampling of peptide and protein conformations using replica exchange simulations with a peptide backbone biasing-potential. Proteins 66:697–706
181. Bowman GR, Ensign DL, Pande VS (2010) Enhanced modeling via network theory: adaptive sampling of markov state models. J Chem Theory Comput 6:787–794
182. Kamberaj H, van der Vaart A (2009) An optimized replica exchange molecular dynamics method. J Chem Phys 130:074906
183. Wang FG, Landau DP (2001) Efficient, multiple-range random walk algorithm to calculate the density of states. Phys Rev Lett 86:2050–2053
184. Zhang C, Ma JP (2009) Enhanced sampling in generalized ensemble with large gap of sampling parameter: case study in temperature space random walk. J Chem Phys 130:194112
185. Zhang C, Ma J (2010) Enhanced sampling and applications in protein folding in explicit solvent. J Chem Phys 132:244101
186. Gao YQ (2008) An integrate-over-temperature approach for enhanced sampling. J Chem Phys 128:064105
187. Yang LJ, Shao Q, Gao YQ (2009) Comparison between integrated and parallel tempering methods in enhanced sampling simulations. J Chem Phys 130:124111
188. Piana S, Laio A (2007) A bias-exchange approach to protein folding. J Phys Chem B 111:4553–4559
189. Piana S, Laio A, Marinelli F, Van Troys M, Bourry D, Ampe C, Martins JC (2008) Predicting the effect of a point mutation on a protein fold: the villin and advillin headpieces and their Pro62Ala mutants. J Mol Biol 375:460–470
190. Todorova N, Marinelli F, Piana S, Yarovsky I (2009) Exploring the folding free energy landscape of insulin using bias exchange metadynamics. J Phys Chem B 113:3556–3564
191. Zheng LQ, Chen MG, Yang W (2009) Simultaneous escaping of explicit and hidden free energy barriers: application of the orthogonal space random walk strategy in generalized ensemble based conformational sampling. J Chem Phys 130:234105
192. Skolnick J (2006) In quest of an empirical potential for protein structure prediction. Curr Opin Struc Biol 16:166–171
193. Hegler JA, Latzer J, Shehu A, Clementi C, Wolynes PG (2009) Restriction versus guidance in protein structure prediction. Proc Natl Acad Sci USA 106:15302–15307
194. Kim DE, Blum B, Bradley P, Baker D (2009) Sampling bottlenecks in De novo protein structure prediction. J Mol Biol 393:249–260
195. DeBartolo J, Colubri A, Jha AK, Fitzgerald JE, Freed KF, Sosnick TR (2009) Mimicking the folding pathway to improve homology-free protein structure prediction. Proc Natl Acad Sci USA 106:3734–3739
196. DeBartolo J, Hocky G, Wilde M, Xu JB, Freed KF, Sosnick TR (2010) Protein structure prediction enhanced with evolutionary diversity: SPEED. Protein Sci 19:520–534
197. Brunette TJ, Brock O (2008) Guiding conformation space search with an all-atom energy potential. Proteins 73:958–972
198. Chen WW, Yang JS, Shakhnovich EI (2007) A knowledge-based move set for protein folding. Proteins 66:682–688
199. Yang YD, Liu HY (2006) Genetic algorithms for protein conformation sampling and optimization in a discrete backbone dihedral angle space. J Comput Chem 27:1593–1602
200. Zhao F, Li SC, Sterner BW, Xu JB (2008) Discriminative learning for protein conformation sampling. Proteins 73:228–240
201. Zhao F, Peng JA, Xu JB (2010) Fragment-free approach to protein folding using conditional neural fields. Bioinformatics 26:i310–i317
202. Zhao F, Peng J, Debartolo J, Freed KF, Sosnick TR, Xu J (2009) A probabilistic graphical model for Ab initio folding. Lect Notes Comput Sci 5541:59–73
203. Boomsma W, Mardia KV, Taylor CC, Ferkinghoff-Borg J, Krogh A, Hamelryck T (2008) A generative, probabilistic model of local protein structure. Proc Natl Acad Sci USA 105:8932–8937
204. Hamelryck T, Kent JT, Krogh A (2006) Sampling realistic protein conformations using local structural bias. PLoS Comput Biol 2:1121–1133
205. Ortiz AR, Kolinski A, Skolnick J (1998) Nativelike topology assembly of small proteins using predicted restraints in Monte Carlo folding simulations. Proc Natl Acad Sci USA 95:1020–1025
206. Eyrich VA, Standley DM, Friesner RA (1999) Prediction of protein tertiary structure to low resolution: performance for a large and structurally diverse test set. J Mol Biol 288:725–742
207. Hardin C, Eastwood MP, Luthey-Schulten Z, Wolynes PG (2000) Associative memory Hamiltonians for structure prediction without homology: alpha-helical proteins. Proc Natl Acad Sci USA 97:14235–14240
208. Fain B, Levitt M (2003) Funnel sculpting for in silico assembly of secondary structure elements of proteins. Proc Natl Acad Sci USA 100:10700–10705
209. Nanias M, Chinchio M, Pillardy J, Ripoll DR, Scheraga HA (2003) Packing helices in proteins by global optimization of a potential energy function. Proc Natl Acad Sci USA 100:1706–1710
210. Pollastri G, Przybylski D, Rost B, Baldi P (2002) Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins 47:228–235
211. Kang HS, Kurochkina NA, Lee B (1993) Estimation and use of protein backbone angle probabilities. J Mol Biol 229:448–460
212. Rooman MJ, Kocher JP, Wodak SJ (1991) Prediction of protein backbone conformation based on seven structure assignments. Influence of local interactions. J Mol Biol 221:961–979
213. Gibrat JF, Robson B, Garnier J (1991) Influence of the local amino-acid-sequence upon the zones of the torsional angles-phi and angle-psi adopted by residues in proteins. Biochemistry 30:1578–1586
214. Bystroff C, Baker D (1998) Prediction of local structure in proteins using a library of sequence-structure motifs. J Mol Biol 281:565–577

215. Kuang R, Leslie CS, Yang AS (2004) Protein backbone angle prediction with machine learning approaches. Bioinformatics 20:1612–1621
216. Zimmermann O, Hansmann UHE (2008) LOCUSTRA: accurate prediction of local protein structure using a two-layer support vector machine approach. J Chem Inf Model 48:1903–1908
217. Dor O, Zhou Y (2007) Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. Proteins 66:838–845
218. Faraggi E, Xue B, Zhou Y (2009) Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. Proteins 74:847–856
219. Xue B, Dor O, Faraggi E, Zhou Y (2008) Real-value prediction of backbone torsion angles. Proteins 72:427–433
220. Dor O, Zhou Y (2007) Real-SPINE: an integrated system of neural networks for real-value prediction of protein structural properties. Proteins 68:76–81
221. Faraggi E, Yang YD, Zhang SS, Zhou Y (2009) Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction. Structure 17:1515–1527
222. Cornilescu G, Delaglio F, Bax A (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J Biomol NMR 13:289–302
223. Duan Y, Chowdhury S, Xiong G, Wu C, Zhang W, Lee T, Cieplak P, Caldwell J, Luo R, Wang J, Kollman PA (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase QM calculations. J Comput Chem 24:1999–2012
224. Wang Z, Duan Y (2004) Solvent effects on alanine dipeptide: a MP2/cc-pVTZ//MP2/6–31G** study on its (Φ, Ψ) energy maps and conformers in the gas phase, ether and water. J Comput Chem 25:1699–1716