CHAPTER 19

CONFORMATIONAL SEARCH FOR THE PROTEIN NATIVE STATE

Amarda Shehu
Assistant Professor Dept. of Comp. Sci.
Affil. Appnt. Dept. of Comp. Biol. and Bioinf.
George Mason University
4400 University Blvd. MSN 4A5
Fairfax, Virginia, 22030, USA

Abstract

This chapter presents a survey of computational methods that obtain a structural description of the protein native state. This description is important to understand a protein's biological function. The chapter presents the problem of characterizing the native state in conformational detail in terms of the challenges that it raises in computation. Computing the conformations populated by a protein under native conditions is cast as a search problem. Methods such as Molecular Dynamics and Monte Carlo are treated first. Multiscaling, the combination of reduced and high complexity models of conformations, is briefly summarized as a powerful strategy to rapidly extract important features of the energy surface associated with the protein conformational space. Other strategies that narrow the search space through information obtained in the wet lab are also presented. The chapter then focuses on enhanced sam-

Protein Structure Prediction:Method and Algorithms. By H. Rangwala & G. Karypis

pling strategies employed to compute native-like conformations when given only amino-acid sequence. Fragment-based assembly methods are analyzed for their success and what they are revealing about the physical process of folding. The chapter concludes with a discussion of future research directions in the computational quest for the protein native state.

19.1 THE QUEST FOR THE PROTEIN NATIVE STATE

From the first formulation of the protein folding problem by Wu in 1931 to the experiments of Mirsky and Pauling in 1936, chemical and physical properties of protein molecules were attributed to the amino-acid composition and structural arrangement of the protein chain [86, 56]. Mirsky and Pauling hypothesized that denaturing conditions like heating abolished chemical and physical properties of a protein by "melting away" the protein structure. The relationship between sequence, structure, and function in protein molecules was under significant debate until the revolutionizing 1960's experiments by Christian Anfinsen and his co-workers at the National Institute of Health [3].

The Anfinsen experiments showed that ribonuclease would spontaneously reassume its structure and enzymatic activity after denaturation. This unique ability to regain both structure and function was confirmed on thousands of proteins. After a decade of experiments, Anfinsen concluded that the amino-acid sequence governed the folding of a protein chain into a "biologically-active conformation" under a "normal physiological milieu" [2]. Anfinsen used the terms "conformation" and "structure" interchangeably to describe a three-dimensional (3D) arrangement of the chain connecting amino acids in a protein. To this day, no distinction is drawn between structure and conformation.

The Anfinsen experiments posited that, if one were to understand how the amino-acid sequence determines the biologically-active conformation, one could find this conformation in silico. A computer algorithm would follow simple rules or instructions to compute the biologically-active conformation. Computing this conformation from knowledge of amino-acid sequence alone remains a grand challenge of computational biology [22, 89]. Nonetheless, significant computational progress has been made. Many methods now provide useful mechanistic insight into the biological function of a protein through a structural description of the functionally-relevant (native) state.

19.1.1 Native Structure versus Native Conformational Ensemble

The survey in this chapter focuses on methods that search the space of different conformations of a protein chain to find those associated with the native state. Significant computational research on describing the protein native state in conformational detail focuses on computing a single representative conformation, often referred to as the native structure [45, 8, 7, 16, 58, 28, 60, 88]. Such research, the focus of CASP [57], is justified in many proteins.

The single structure view of the native state does not fully take into account the inherent flexibility of a protein chain. Long gone are the days when protein molecules were considered rigid and solid-like [71]. In the words of Richard Feynmann [26]:

> "Everything that living things do can be understood in terms of the jiggling and wiggling of atom." [Lectures on Physics, 1964]

When flexibility under native conditions consists of mostly local insignificant deviations around an average structure, the single structure view is well warranted. For many proteins, however, an accurate description of flexibility may involve large-scale conformational rearrangements. This is often the case in proteins involved in biochemical processes like molecular recognition, enzymatic catalysis, and signal transduction [84, 61, 25]. Evidence of functionally-relevant flexibility, not necessarily around a unique structure, advocates a more general description of the native state through an ensemble of conformations, also referred to as the the native state ensemble [41, 43, 36].

19.1.2 Thermodynamic versus Kinetic Hypothesis

The current understanding of what drives the folding of a protein chain directly impacts the assumptions and strategies employed by computational methods to find the native state. Historically, two hypotheses have competed to explain the process of folding. The thermodynamic hypothesis states that the native state of a protein minimizes free energy, whereas the kinetic hypothesis attributes the native state to that which is kinetically accessible.

Anfinsen made the case that his experiments and those of other researchers established the generality of the thermodynamic hypothesis:

> "This hypothesis states that the three-dimensional structure of a native protein in its normal physiological milieu (solvent, pH, ionic strength, presence of other components such as metal ions or prosthetic groups, temperature, etc.) is the one in which the Gibbs free energy of the whole system is lowest; that is, that the native conformation is determined by the totality of interatomic interactions and hence by the amino acid sequence, in a given environment." [Nobel Lecture, December 11, 1972]

The thermodynamic hypothesis suggests that a naive computer algorithm can be written to systematically enumerate the distinct structures or conformations assumed by a protein chain. The algorithm can sum over the interatomic interactions to evaluate the energy of each computed conformation. Assuming the algorithm can properly identify the computed conformation(s) where the free energy reaches its global minimum value, the biologically-active state will have been captured in structural detail by simple enumeration.

Enumeration is possible when (i) the number of parameters employed to represent a protein conformation is small, and when (ii) these parameters take values from a finite set. In other words, the space of possible conformations, referred to as the conformational space, has to be low-dimensional and discretizable for enumeration to be a feasible strategy. Early computational re-

search on finding native conformations of a protein assumed a low-dimensional and discretizable space that was amenable to enumeration [38, 44, 42]. Such simplification allowed application of exhaustive search to the discovery of native conformations, albeit at the cost of a systematic inability to capture subtle structural details of the native state [63, 67].

Back-of-the-envelope calculations led Cyrus Levinthal to a supposed paradox. Even assuming a small number of configurations of the peptide bond connecting two consecutive amino acids (e.g., 3) in a short protein chain of 101 amino acids, the number of ensuing conformations is 3^{100}. Assuming a rate of 10^{13} conformations per second, it would still take 10^{27} years for a protein to sample all these conformations. This dramatic example showed that a protein could not possibly find its native structure by searching at random within a vast and high-dimensional conformational space [49].

Levinthal was concerned with the time that it would take a protein to find its lowest free energy state, i.e., the actual kinetics of folding. Given that many proteins refold in a few microseconds after denaturation, random sampling of the conformational space does not explain the process of folding. Levinthal's paradox illustrates that (i) diffusion cannot be the only guiding force behind folding, and (ii) random searches are infeasible strategies for sampling the native conformation(s) of a protein chain.

Levinthal's calculations cast the process of finding the native structure as searching for a needle in a haystack. Early research showed in simulation that the problem of computing the lowest free energy state was indeed hard [47]. On the one hand, theoretical research in computer science proved that the problem, even when employing simple lattice models to represent conformations of a protein chain, is NP-hard [82, 32]. On the other hand, simulation studies showed that proteins could get trapped in structures that were energetically similar but topologically different from the native structure obtained in the wet lab [48]. An alternative hypothesis was offered to explain the discrepancy through the possibility of kinetic traps. The kinetic hypothesis suggested that proteins folded into structures that were kinetically accessible.

Despite the complexity associated with searching the protein conformational space and the seemingly competing views of protein folding, efficient algorithms exist today. These algorithms operate in a high-dimensional and continuous search space. This is made possible by a better understanding of protein physics and the process of folding accompanied by a steady increase in the number of calculations that can be performed in one CPU cycle. The predictive power of these algorithms and their ability to reproduce observations in the wet lab with high accuracy has significantly improved in the last decade. Significant progress in our understanding of protein folding came with the introduction of the energy landscape view, which unified the thermodynamic and kinetic hypotheses. This view is the focus of the next section.

19.1.3 The Energy Landscape View of Protein Folding

Levinthal's paradox ignored the energetic bias against unfavorable protein conformations. Seminal work that interpreted evidence emerging from folding experiments through the theory of statistical mechanics presented an energy landscape view of protein folding [21, 59] that reconciled the thermodynamic and kinetic hypotheses. The "New View" [21], offered a statistical description of the complex energy surface of a protein through an energy landscape. Despite the high-dimensionality of the conformational space and the intricate number of interatomic interactions in a protein, the energy surface associated with the conformational space can be projected onto a few coordinates to obtain an energy landscape view of how proteins fold.

Figure 19.1 The schematic diagram illustrates different folding scenarios. The vertical axis plots the internal free energy of a protein. The conformational space is projected on two coordinates (horizontal axes). The landscape on the left illustrates the classic scenario, where the native state labeled N is associated with the global energy minimum. The landscape on the right illustrates how a protein can be trapped in multiple deep minima. Reprinted from [21] with permission of Ken Dill.

Under this view, three main classes of energy surfaces emerge. They range from surfaces with a single global free energy minimum (Figure 20.1, left), corresponding to proteins with a very strong stability point, to surfaces with a few minima (Figure 20.1, right), and surfaces with a shallow native basin [5]. Actual energy surfaces of proteins may combine these three main cases.

The energy landscape view employs statistical mechanics to organize the multitude of protein conformations in terms of a minimal number of collective parameters [59, 30, 13]. This statistical formulation allows capturing essential features of the free energy surface of a protein with only a limited set of parameters. Various computational methods exploit this formulation to focus the search for native or near-native conformations of a protein chain to minima in the energy landscape that emerge when projecting the energy surface over the employed parameters. However, the general existence of underlying

collective parameters that guide a protein to quickly locate its lowest free energy state remains open to debate [13].

The energy landscape view provides a theoretical framework to explain how a protein may assume different low-energy conformations, for instance, upon binding [80]. The free energy minimum in the energy landscape could be populated by different low-energy conformations of a protein chain which map to the same region in the space of the underlying collective parameters. Since understanding protein function requires obtaining a comprehensive view of the conformational space associated with the free energy minimum (or minima) in the energy landscape [43], many computational methods describe the protein native state as an ensemble of conformations [37, 51, 78].

19.1.4 Computational Issues in the Search for Native Conformations

Figure 20.1 allows visualizing the thermodynamic versus the kinetic hypothesis [4, 29]. The energy landscape view brings into focus fundamental questions and computational issues that need to be addressed by algorithms designed to search a protein's conformational space. These issues and potential strategies to address them are summarized below. Methods that implement these strategies are then detailed in the rest of the chapter.

Does Search for the Native State Need the How? If one wants to design an algorithm to compute conformations of a protein chain under native conditions, should the algorithm consider the actual process of folding? Should physical timescales be associated with computed conformations? Considering how a protein chain tumbles down the energy landscape may help capture possible kinetic traps and find the actual native state. Many computational methods follow the physical process of folding to let a protein system "sample" the native state in simulation. Consideration of "the how" can result in very long simulations. Currently, methods that employ mainly the thermodynamic hypothesis exhibit faster sampling efficiency. Recent evidence emerging from the most successful methods suggests that incorporating features of the physical process of folding can actually improve both efficiency and accuracy [8, 20, 9].

Realms of Discretization The energy landscape view advocates that, if one were to know "the true" collective parameters (often referred to as reaction coordinates) that guide the folding reaction, the energy landscape obtained by projecting the protein energy surface over these parameters is not complex. The search for native conformations can be conducted over a discretization of the projection of the conformational space over the coordinates. At the very least, the discretization of the projected space can be employed to keep track and guide the search in the high-dimensional conformational space. Finding reaction coordinates, however, is challenging. A rich body of research beyond the scope of this survey pursues finding collective parameters that can serve as general reaction coordinates for protein systems [18].

Sampling Over Enumeration A fundamental problem in obtaining a structural description of the protein native state is that of efficiently computing conformations associated with the global minimum (or deepest minima) in a rugged energy surface constellated with local minima. Multiscale modeling, which combines coarse-grained and fine-grained detail when modeling a protein conformation, and a naturally-inspired discretization of the process through which conformations are assembled are currently the most successful strategies [78, 13, 16, 7]. The paradigm in some of the most successful methods is away from a systematic search and towards a probabilistic walk or probabilistic sampling of the conformational space [50, 1, 6, 58, 74].

Guiding the Search and Narrowing the Search Space The vast high-dimensional conformational space available to a protein chain raises practical computability problems. Many computational methods (detailed below) resort to employing additional information to narrow the conformational space relevant for their search. This information, either in the form of thermodynamic averages over conformations of the native state or in the form of an average native structure captured in experiment allows constraining the search for native conformations by what is observed in experiment.

Enhancing Sampling of a Vast High-dimensional Space Ab-initio methods that employ only knowledge of amino-acid sequence for a protein at hand have to enhance sampling of the vast conformational space. Enhanced sampling methods include simulated annealing, importance and umbrella sampling, replica exchange (also known as parallel tempering), local elevation, activation relaxation, local energy flattening, jump walking, multicanonical ensemble, conformational flooding, Markov state models, discrete timestep MD, and many more. Since a complete survey of these methods is not possible, the following summary focuses on a few successful representative methods.

Combining the Discrete and the Continuous This survey of conformational search methods for the protein native state concludes with a discussion of potential benefits to future research by a combination of discrete and continuous exploration. The discussion focuses on combining search in a discretized energy landscape with search on a continuous conformational space. The chapter concludes with an outlook on how knowledge of collective parameters that can serve as reaction coordinates can be employed to guide conformational search towards relevant energy minima in the underlying energy landscape.

19.2 EXHAUSTIVE SEARCH: DISCRETIZATION OF CONFORMATIONAL SPACE

Early simulations of protein chains showed that important physical properties could be obtained with considerably less than atomic detail. Coarse-grained

modeling of protein conformations opened up the possibility of exhaustively searching a simplified conformational space through explicit enumeration of possible conformations of a protein chain.

Some of the first coarse-grained models were based on lattices, explicitly modeling a representative (often C_α) atom of each amino acid in a protein chain and restricting atoms to lie on a lattice [81]. Lattice modeling not only allowed computing native conformations of very long protein chains, but incidentally exposed an interesting complexity result: finding the lowest-energy conformation on a 3D cubic lattice is NP-hard [82]. Despite this complexity result, lattice models offered both analytical and computational simplicity [87]. In addition to very fast integer-math evaluations of conformational energies on a lattice, lattice modeling made it feasible for exhaustive searches to explicitly enumerate conformations [38, 44, 42].

Despite the simplicity, exhaustive search methods that employ lattice modeling can reproduce the backbone with accuracy no greater than half the lattice spacing [67]. These methods cannot capture subtle structural details and may bias towards specific secondary structures [63]. Research on improving accuracy and getting the full computational benefits of searching in a discretized conformational space remains active. Indeed, some of the most successful enhanced sampling methods implicitly employ discretization by assembling conformations of a protein chain with naturally-occurring structures of short fragments defined over the chain [28, 7, 9, 16, 78, 20].

Coarse-grained models that capture realistic protein structures are predominantly off-lattice [39, 13]. Rather than discretize the conformational space, these models simplify and lower the dimensionality of this space. Sophisticated force fields designed for these models associate potential energies with computed conformations [62, 53, 17, 54, 39]. For instance, backbone-resolution models, where only heavy backbone atoms are explicitly modeled, allow obtaining highly-accurate native conformations (some of these models include C_β atoms to represent side chains) [58, 31, 16, 28, 78].

Recent methods combine coarse- and fine-grained modeling to enhance sampling of the conformational space [46, 52, 11]. For instance, methods that predict native conformations from the amino-acid sequence conduct most of their search in a coarse-grained space, adding atomic detail (as in [34]) only when it is imperative to refine conformations or determine which low-energy minima are relevant for the native state [7, 77, 78]. Multiscaling is one of the many computational strategies to enhance the exploration of the high-dimensional protein conformational space.

19.3 SYSTEMATIC SEARCH: MOLECULAR DYNAMICS

Computational methods that follow the physics of folding to let a protein "sample" its native state implement the Molecular Dynamics (MD) approach [83]. MD-based methods systematically search the conformational space by numer-

ically solving Newton's equations of motion. The solution accuracy dictates a small timestep in the order of femtoseconds between consecutive conformations in an MD trajectory. As a result, MD-based simulations may demand long trajectories before attaining native conformations. Moreover, when no knowledge of the global energy minimum is available, multiple trajectories may need to be computed in order to determine that no significantly lower energies can be obtained. The issue of convergence has practical implications for time demands. Most MD studies circumvent this issue by employing similarity between computed conformations and experimentally-available native structures of tested proteins as a termination criterion.

Since MD-based methods follow the process of folding, they offer more than just a set of the conformations relevant for the native state. They also reveal kinetic information; that is, how the unfolded protein chain reaches its native state and in what timescales. This added information increases the computational requirements of MD-based methods. These requirements are often alleviated by distributing the MD search of the conformational space on supercomputers [24] and grids of desktops [79]. Specific architectures like the IBM BLUE Gene and Anton and distributed MD implementations like Desmond are devoted to surpassing computational milestones and achieving high-resolution native conformations through MD simulations [64, 72].

19.4 BIASED RANDOM WALK: METROPOLIS MONTE CARLO

Rather than solving Newton's equations of motions, random search techniques such as Monte Carlo (MC) conduct biased random (probabilistic) walks in conformational space to obtain a sequence of conformations [69, 83]. The random walk ensures through the Metropolis criterion [55] that a conformation is obtained with frequency proportional to its Boltzmann probability. While exhibiting higher sampling efficiency than MD simulations, MC simulations also obtain conformations sequentially. Like MD, they also spend considerable time sampling rare events such as crossing maxima in the energy landscape.

The tendency of MD- and MC-based methods to converge to local minima in the energy surface that underlies the protein conformational space underscores the fact that MD and MC are local optimization techniques. This tendency is usually circumvented in two ways: (i) by narrowing the search to specific regions in the energy surface or conformational space through experimentally-available information; (ii) by enhancing the local optimization in the MD systematic search or the MC sampling through an array of enhanced sampling strategies beyond multiscaling. These two (not mutually exclusive) groups of methods are discussed next.

19.5 GUIDED SEARCH OF CONFORMATIONAL SPACE

A special class of conformational search methods employ experimental data to guide MD or MC trajectories to the relevant search space. The data help to focus computational resources to regions of the energy surface or the conformational space that are relevant for the native state and to quickly guide the exploration towards native conformations. These data come in the form of thermodynamic averages (over the underlying native state ensemble) obtained from Nuclear Magnetic Resonance (NMR) experiments, density maps obtained from X-ray crystallography or cryo-Electron Microscopy (cryoEM), or an average structure obtained from X-ray or NMR.

19.5.1 Guiding the Search with Thermodynamic Averages

Methods that employ NMR thermodynamic averages such as NOE distance constraints, S^2 order parameters, three-bond scalar couplings (^3J), residual dipolar couplings (RDCs), chemical shifts, ϕ or ψ values, or protection factors often incorporate these averages in an additional term in the potential energy function [6, 14, 51, 15, 68]. The resulting pseudo-energy function biases trajectories launched in conformational space away from conformational ensembles that, while low in energy, do not reproduce the NMR averages.

The NMR data are averages over an ensemble of molecules over time. While structures obtained in NMR is refined to agree with these averages, the refinement cannot capture the possibly non-local conformational heterogeneity present in solution. For this reason, methods that conduct a guided exploration of the conformational space are able to obtain a broader picture of the native state through an ensemble of conformations whose statistical averages reproduce the NMR observables better than a single structure. Figure 20.2 shows one such ensemble obtained for ubiquitin that reproduces the NMR data better than a single native structure.

The strategy of incorporating experimental data in the energy function is often described as a way to overcome possible non-physical biases in the current generation of (semi-empirical) molecular mechanics force fields [40]. The pseudo-energy function distorts the energy surface by deepening those low-energy regions that reproduce the experimental data. In this way, local optimization techniques have a better chance of converging to the funnel of the true energy surface of a protein.

Rather than modify the underlying energy surface by enforcing agreement with experimental data, recent methods use the experimental data to either build probabilistic models of relevant regions of the conformational space or explicitly disqualify regions from further exploration [23, 65]. In particular, the work in [65] presents a complete method that subdivides the search space into regions worthy of further exploration and regions corresponding to structures in direct violation of NMR NOE distance constraints. A branch-

Figure 19.2 144 conformations computed in [68] are superimposed in transparent over the first one (in opaque) of ubiquitin. The ensemble is obtained from the Protein Data Bank (PDB) under id 2nr2. Reproduced with permission of Michele Vendruscolo.

Figure 19.3 184 phospholamban conformations (under id 2hyn in the PDB) computed in [65] are shown superimposed over one another. The five monomers of the complex are shown in different colors. Courtesy of Chris Bailey-Kellogg.

and-bound search computes native structures of cyclic complexes such as the phospholamban protein shown in Figure 20.3.

19.5.2 Narrowing the Search with a Template Structure

Other methods elucidate structural details of the native state by searching with geometric or rigidity constraints. These constraints are often extracted from an average structure obtained in experiment [74, 75, 85, 12]. By constraining their search around an experimental structure, these methods capture the conformational heterogeneity in proteins where flexibility under native conditions consists of local fluctuations around a representative structure. The representative structure is essentially employed as a semi-rigid template.

Work in [85, 12] is inspired from the constraint theory in the context of mechanical engineering considerations in bar and joint frameworks. Rigidity analysis over the template structure reveals under-constrained degrees of freedom (angles) at room temperature. These angles define a search space which is explored to obtain conformations that obey the rigidity constraints and exhibit as much internal mobility as allowed by the template [12].

Other work, inspired by the treatment of inverse kinematics in robotics, conducts a geometrically-constrained probabilistic sampling of the conformational space around the template structure [74, 75, 76]. Local fluctuations are obtained around the representative structure by computing geometrically-constrained conformations of consecutive overlapping fixed-length fragments defined over the protein chain. Figure 20.4 shows fragment conformations superimposed over the reference structure of P. magnus albumin-binding second

Figure 19.4 Left: The lowest-energy conformations computed with the method described in [76] are drawn in transparent over the opaque X-ray structure of ALB8-GA. Right: Amide S^2_{calc} data (orange squares) calculated over the ensemble are compared to available NMR S^2_{exp} data (yellow squares). Methyl S^2_{calc} data are shown in colored circles (no NMR data are available for comparison). Horizontal bars on the $x-$axis show the position of the three α-helices (also annotated over the ensemble). The parts of the bars in lighter colors indicate amino acids found in unfolded configurations. Reprinted from [76] with permission.

GA module of PAB (ALB8-GA). Thermodynamic averages calculated over the conformations reproduce well data obtained in experiment.

19.6 ENHANCED SAMPLING OF CONFORMATIONAL SPACE

Stochastic search is one of the strategies employed to enhance the sampling of the high-dimensional protein conformational space. Stochastic search (or stochastic optimization) is a powerful strategy to solve global optimization problems on surfaces marked by abundance of local minima [19]. When the energy surface is complex and decisions are made locally about which conformations map to minima in the energy surface, stochastic search becomes a viable strategy to explore the protein conformational space. For instance, work in [74, 77] employs a robotics-inspired probabilistic sampling to compute geometrically-constrained conformations.

Other strategies enhance sampling in the context of a trajectory-based exploration by replicating trajectories, varying temperature (such as, simulated annealing), and exchanging conformations from which trajectories are launched in conformational space. An incomplete list of successful enhanced sampling strategies applied to searching the protein conformational space include importance sampling, simulated annealing, umbrella sampling, genetic algorithms, replica exchange (also known as parallel tempering), local elevation, activation relaxation, local energy flattening, jump walking, multicanonical ensemble, conformational flooding, Markov state models, discrete timestep MD, fragment-based assembly, and many more (cf. [83]).

19.6.1 Principles of Self-organization in Protein Chains

Analysis of conformational search methods identifies two main ingredients as essential for success: (i) a powerful sampling strategy to obtain a broad coverage of the conformational space and (ii) an accurate energy function that allows a near-native conformation to converge to the nearby native basin. The design of accurate energy functions remains an active area of research and is pursued vigorously by many computational groups [17, 7, 62, 16]. Energy functions provide a search method with a local view of the energy surface. This local view biases the search towards low-energy regions of the emerging energy surface. The energy function should not significantly distort the energy surface of an amino-acid sequence under consideration.

Research shows that proteins have been designed by evolution to fold in spite of errors [66]. These findings advocate that a sampling strategy should not be highly sensitive to minor distortions of the energy surface and should be able to succeed as long as the energy function maps the explored conformational space on an energy surface that is funneled towards the native state. Significant computational efforts target the design of powerful sampling strategies to rapidly cover the conformational space [7, 22]. Work in [27] highlights that "the ultimate speed limit in protein folding is conformational search."

A good coverage of the conformational space should yield diverse conformations that are near energy minima relevant for the native state. Local optimization can then push near-native conformations to the native basin(s). It is worth mentioning that the notion of coverage is well-defined and employed in the AI and robotics community [10]. The complexity and high-dimensionality of the conformational space makes it very costly to estimate coverage [75]. Recent conformational search methods inspired in robotics are employing simple estimates of coverage to guide the search for the native state [73].

A powerful sampling strategy that does not incapacitate the predictive power of a method searching for the native state allows testing different hypothesis for how self-organization emerges in protein chains. Some of these hypotheses focus on determining the amount of detail that is necessary to capture the native state. For instance, work in [70] suggests a backbone-based theory of protein folding. Results emerging from multiscaling studies of the protein native state advocate employment of different scales [13].

Growing evidence points towards a hierarchical organization of the native structure, where local interactions bias the local structure that emerges in protein chains. This in turn limits the number of ways low-energy conformations can be put together. This realization is not new and continuous to emerge from various studies [8, 60, 39, 28, 16, 7, 33]. Fragment-based assembly (FA) methods, the focus of the next section, employs this realization to efficiently compute conformations by assembling them from smaller local structures.

19.6.2 Local Structure Limits Global Arrangements

FA methods are emerging as successful ab-initio methods in predicting the native state from knowledge of the amino-acid sequence [9]. The basic process in FA methods is to assemble conformations of a protein chain with local structures of fragments of the chain. The assembly can be implemented either in the context of an MC-based [35] or MD-based search [8, 16, 7, 28, 78, 73]. The sequence of the protein under consideration is divided into short fragments. Rarely, additional information is associated with the fragments from discretized Ramachandran maps of the backbone angles [28, 16].

The key feature is that conformations of a protein chain are assembled from local structures of short fragments. Candidate local structures for the fragments are compiled from a non-redundant database of protein structures, often extracted from the PDB. These local structures constitute a limited move set considered in an MC- or MD-framework to put together global tertiary structures (conformations) of a protein chain. The extent to which the limited move set reflects the naturally-occurring biases that the fragment sequence has on local structure depends on the length of the fragment and the richness of the PDB. Employed fragment lengths range from 3 to 9. Deciding on a suitable fragment length depends on the richness of the PDB to provide a comprehensive picture of the extent to which the fragment sequence determines the structures in which the fragment can be found in nature.

By considering a limited set of structures for a protein fragment, FA methods discretize the relevant conformational space to be explored. Yet, the success of these methods does not lie in this superficial observation. FA methods implement the experimental observation that the local sequence (implemented in the definition of a fragment) biases but does not uniquely decide the local structure of the fragment (implemented in the sampling of structures of a fragment from a database). These local structural biases limit the number of ways low-energy conformations can be assembled together.

Recent work in [20] improves upon the classic FA framework by iteratively fixing the secondary structure assignments of amino acids during the generation of conformations in an MC simulated annealing search. The available search space of local fragment structures is progressively narrowed by "locking in" predominant secondary structure assignments that emerge during the search. Besides improving the efficiency of the search and the accuracy of the resulting lowest-energy conformations (see Figure 20.5), the method outperforms homology-based secondary structure prediction methods while using only a coarse-grained modeling with no explicit side chains. The success of this method may shed insight into the actual process of folding not just as a hierarchical process but, a process that employs information on which secondary structures dominate a robust and efficient folding pathway.

Some studies suggest that biases on local structure, even when combined with nonspecific compaction forces (which promote compact conformations), are sufficient to result in a rapid sampling of native-like conformations of small

Figure 19.5 Alignments of the predictions generated in [20] (in red) have A) the lowest energy and B) the lowest least root-mean-squared-deviation (lRMSD) to the native structure (blue) for three proteins (PDB codes at top, lRMSD values indicated). Images are created using the PyMol visualization software. C) Scatter plots of lRMSD versus energy. Courtesy of Tobin Sosnick.

proteins [9]. The extent to which the energy function determines the success of FA methods is under some debate [35]. Application to larger proteins with multiple competing conformational ensembles under native conditions suggests a larger role for both a sophisticated energy function and an enhanced sampling strategy in order to sufficiently populate possibly multiple energy minima relevant for the native state [78, 35]. Figure 20.6 shows the energy landscape and competing conformational ensembles computed for calmodulin at room temperature with an FA-based MC simulated annealing [78].

19.7 DISCUSSION OF FUTURE RESEARCH DIRECTIONS

Enhanced sampling on a simplified search space and sophisticated energy functions are allowing FA methods to achieve high prediction accuracy of the native state and in the process shed light on the physical process of folding. While the prediction accuracy has improved both in proteins with high and low homology [7, 20] and even in proteins with multiple functional states [78], research on improving the efficiency of FA methods is active. Time demands remain a point to address through further research.

Figure 19.6 Left: The 2D energy pseudo-free energy landscape obtained for the calmodulin sequence with the method described in [78] is shown in a red-to-blue color scheme that denotes high-to-low energy values. The deepest minima are labeled A, B, and C. Right: Computed conformational ensembles that correspond to the minima are shown and labeled accordingly. Conformations are superimposed in transparent over lowest-energy ones drawn in opaque. Reproduced from [78] with permission.

One way to bring the computation of native-like conformations to a few hours on a single CPU is to discourage sampling of similar low-energy conformations. Currently, it seems difficult to ensure that computed conformations are geometrically-distinct and not representative of few regions of the conformational space [78]. Part of the difficulty lies in the inability to find a few meaningful parameters on which to define distance measures. The classic measures like lRMSD and radius of gyration (Rg) often mask away differences among conformations [78]. The robotics-inspired method in [73] proposes a way to address this problem.

The tree-based method in [73] advocates that conformational search be guided through low-dimensional projections of the conformational space and the energy surface. The projections afford a discretized view of the explored conformational space and its corresponding energy surface and allow defining a two-layered probability distribution by which to guide the search towards conformations that are both low-energy and geometrically-distinct. While the projection coordinates employed in [73] are not proposed as general reaction coordinates, the two-layered search may be an interesting framework through which to maximize the sampling of low-energy conformations that populate a desired conformational subspace.

The survey in this chapter has highlighted the complexity of computing conformations that populate the native state from minimal information such as amino-acid sequence. Given that the prediction of native conformations is a stringent test of the ability of computers to fold protein sequences, research on effective and accurate conformational search for the protein native state will be active. Considering the interdisciplinary challenges that arise in the

context of this problem, contributions will likely emerge from collaborations that reach across exact and life science communities of researchers.

The future holds promises for both communities: computer (and computational) scientists will be challenged and will learn how to mimic in silico the efficient steps that apparently proteins employ to fold within a few microseconds; physicists, biologists, and chemists will complement their experimental and theoretical understanding of the process of folding by testing diverse hypotheses in silico. Continued scientific progress is expected to result from discoveries of efficient search algorithms in computer science and further improvements in our understanding of protein physics.

References

1. R. Abagyan and M. Totrov. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.*, 235(3):983–1002, 1994.
2. C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.
3. C. B. Anfinsen, E. Haber, M. Sela, and F. H. Jr. White. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci. USA*, 47(9):1309–1314, 1961.
4. D. Baker and D. A. Agard. Kinetics versus thermodynamics in protein folding. *Biochemistry*, 33(24):7505–7509, 1994.
5. O. M. Becker and M. Karplus. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *J. Chem. Phys.*, 106(4):1495–1517, 1997.
6. R. B. Best and M. Vendruscolo. Determination of ensembles of structures consistent with NMR order parameters. *J. Am. Chem. Soc.*, 126(26):8090–8091, 2004.
7. P. Bradley, K. M. S. Misura, and D. Baker. Toward high-resolution de novo structure prediction for small proteins. *Science*, 309(5742):1868–1871, 2005.
8. G. Chikenji, Y. Fujitsuka, and S. Takada. A reversible fragment assembly method for de novo protein structure prediction. *J. Chem. Phys.*, 119(13):6895–6903, 2003.

9. G. Chikenji, Y. Fujitsuka, and S. Takada. Shaping up the protein folding funnel by local interaction: lesson from a structure prediction study. *Proc. Natl. Acad. Sci. USA*, 103(9):3141–3146, 2006.

10. H. Choset, K. M. Lynch, S. Hutchinson, G. Kantor, W. Burgard, L. E. Kavraki, and S. Thrun. *Principles of Robot Motion: Theory, Algorithms, and Implementations*. MIT Press, Cambridge, MA, 1st edition, 2005.

11. S. Christakos, C. Gabrielides, and W. B. Rhoten. Multigraining: an algorithm for simultaneous fine-grained and coarse-grained simulation of molecular systems. *J. Chem. Phys.*, 125(15):154106, 2006.

12. M. Chubunsky, B. Hespenheide, D. J. Jacobs, L. A. Kuhn, M. Lei, S. Menor, A. J. Rader, M. F. Thorpe, W. Whiteley, and M. I. Zadovsky. Constraint theory applied to proteins. *Nanotech. Res. J.*, 2(1):61–72, 2008.

13. C. Clementi. Coarse-grained models of protein folding: Toy-models or predictive tools? *Curr. Opinion Struct. Biol.*, 18(1):10–15, 2008.

14. G. M. Clore and C. D. Schwieters. How much backbone motion in ubiquitin is required to account for dipolar coupling data measured in multiple alignment media as assessed by independent cross-validation? *J. Am. Chem. Soc.*, 126(9):2923–2938, 2004.

15. G. M. Clore and C. D. Schwieters. Concordance of residual dipolar couplings, backbone order parameters and crystallographic B-factors for a small α/β protein: A unified picture of high probability, fast atomic motions in proteins. *J. Mol. Biol.*, 355(5):879–886, 2006.

16. A. Colubri, A. K. Jha, M.-Y. Shen, A. Sali, R. S. Berry, T. R. Sosnick, and K. F. Freed. Minimalist representations and the importance of nearest neighbor effects in protein folding simulations. *J. Mol. Biol.*, 363(4):835–857, 2006.

17. P. Das, S. Matysiak, and C. Clementi. Balancing energy and entropy: A minimalist model for the characterization of protein folding landscapes. *Proc. Natl. Acad. Sci. USA*, 102(29):10141–10146, 2005.

18. P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi. Low-dimensional free energy landscapes of protein folding reactions by nonlinear dimensionality reduction. *Proc. Natl. Acad. Sci. USA*, 103(26):9885–9890, 2006.

19. K. A. De Jong. *Evolutionary Computation: A Unified Approach*. MIT Press, Cambridge, MA, 2006.

20. J. DeBartolo, A. Colubri, A. Jha, J. E. Fitzgerald, K. F. Freed, and T. R. Sosnick. Mimicking the folding pathway to improve homology-free protein structure prediction. *Proc. Natl. Acad. Sci. USA*, 106(10):3734–3739, 2009.

21. K. A. Dill and H. S. Chan. From Levinthal to pathways to funnels. *Nat. Struct. Biol.*, 4(1):10–19, 1997.

22. K. A. Dill, B. Ozkan, M. S. Shell, and T. R. Weikl. The protein folding problem. *Annu. Rev. Biophys.*, 37:289–316, 2008.

23. F. DiMao, D. Kondrashov, E. Bitto, A. Soni, G. Bingman, G. Phillips, and J. Shavlik. Creating protein models from electron-density maps using particle-filtering methods. *Bioinformatics*, 23(21):2851–2858, 2007.

24. Y. Duan and P. A. Kollman. Pathways to a protein folding intermediate observed in a 1-μs simulation in aqueous solution. *Science*, 282(5389):740–744, 1998.

25. E. Z. Eisenmesser, O. Millet, W. Labeikovsky, D. M. Korzhnev, M. Wolf-Watz, D. A. Bosco, J. J. Skalicky, L. E. Kay, and D. Kern. Intrinsic dynamics of an enzyme underlies catalysis. *Nature*, 438(7064):117–121, 2005.

26. R. P. Feynman, R. B. Leighton, and M. Sands. *The Feynman Lectures on Physics: Volume I.* Addison Wesley Longman, 1964.

27. K. Ghosh, S. B. Ozkan, and K. A. Dill. The ultimate speed limit to protein folding is conformational searching. *J. Am. Chem. Soc.*, 129(39):11920–11927, 2007.

28. H. Gong, P. J. Fleming, and G. D. Rose. Building native protein conformations from highly approximate backbone torsion angles. *Proc. Natl. Acad. Sci. USA*, 102(45):16227–16232, 2005.

29. S. Govindarajan and R. A. Goldstein. On the thermodynamic hypothesis of protein folding. *Proc. Natl. Acad. Sci. USA*, 95(10):5545–5549, 1997.

30. M. Gruebele. Protein folding: the free energy surface. *Curr. Opinion Struct. Biol.*, 12(2):161–168, 2002.

31. C. Hardin, Z. Luthey-Schulten, and P. G. Wolynes. Backbone dynamics, fast folding, and secondary structure formation in helical proteins and peptides. *Proteins: Struct. Funct. Genet.*, 34(3):281–294, 1999.

32. W. E. Hart and S. Istrail. Robust proofs of NP-hardness for protein folding: General lattices and energy potentials. *J. Comp. Biol.*, 4(1):1–22, 1997.

33. N. Haspel, C. J. Tsai, H. Wolfson, and R. Nussinov. Hierarchical protein folding pathways: A computational study of protein fragments. *Proteins: Struct. Funct. Bioinf.*, 51(2):203–215, 2003.

34. A. P. Heath, L. E. Kavraki, and C. Clementi. From coarse-grain to all-atom: Towards multiscale analysis of protein landscapes. *Proteins: Struct. Funct. Bioinf.*, 68(3):646–661, 2007.

35. J. Hegler, J. Laetzer, A. Shehu, C. Clementi, and P. G. Wolynes. Restriction vs. guidance: Fragment assembly and associative memory hamiltonians for protein structure prediction. *Proc. Natl. Acad. Sci. USA*, 106(36):15302–15307, 2009.

36. V. J. Hilser, B. Garcia-Moreno, G. T. Oas, G. Kapp, and S. T. Whitten. A statistical thermodynamic model of the protein ensemble. *Chem. Rev.*, 106(5):1545–1558, 2006.

37. V. J. Hilser, T. Oas, D. Dowdy, and E. Freire. The structural distribution of cooperative interactions in proteins: Analysis of the native state ensemble. *Proc. Natl. Acad. Sci. USA*, 95(17):9903–9908, 1998.

38. D. A. Hinds and M. Levitt. Exploring conformational space with a simple lattice model for protein structure. *J. Mol. Biol.*, 243(4):668–682, 1994.

39. T. H. Hoang, A. Trovato, F. Seno, J. R. Banavar, and A. Maritan. Geometry and symmetry presculpt the free-energy landscape of proteins. *Proc. Natl. Acad. Sci. USA*, 101(21):7960–7964, 2007.

40. V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins: Struct. Funct. Bioinf.*, 65(3):712–725, 2006.

41. Y.P. J. Huang and G. T. Montellione. Structural biology: Proteins flex to function. *Nature*, 438(7064):36–37, 2005.

42. K. Ishikawa, K. Yue, and K. A. Dill. Predicting the structures of 18 peptides using Geocore. *Protein Sci.*, 8(4):716–721, 1999.

43. M. Karplus and J. Kuriyan. Molecular dynamics and protein function. *Proc. Natl. Acad. Sci. USA*, 102(19):6679–6685, 2005.

44. A. Kolinski and J. Skolnick. Monte carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins: Struct. Funct. Genet.*, 18(4):338–352, 1994.

45. B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker. Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302(5649):1364–1368, 2003.

46. W. Kwak and U. H. Hansmann. Efficient sampling of protein structures by model hopping. *Phys. Rev. Lett.*, 95(13):138102, 2005.

47. K. F. Lau and A. K. Dill. A lattice statistical mechanics model of the conformational and sequence spaces of of proteins. *Macromolecules*, 22(10):3986–3997, 1989.

48. T. Lazaridis and M. Karplus. New view of protein folding reconciled with the old through multiple unfolding simulations. *Science*, 278(5345):1928–1931, 1997.

49. C. Levinthal. Are there pathways for protein folding? *J. Chem. Phys.*, 65(1):44–45, 1968.

50. Z. Li and H. A. Scheraga. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc. Natl. Acad. Sci. USA*, 84(19):6611–6615, 1987.

51. K. Lindorff-Larsen, R. B. Best, M. A. DePristo, C. M. Dobson, and M. Vendruscolo. Simultaneous determination of protein structure and dynamics. *Nature*, 433(7022):128–132, 2005.

52. E. Lyman, F. M. Ytreberg, and D. M. Zuckermann. Resolution exchange simulations. *Phys. Rev. Lett.*, 96(2):028105, 2006.

53. S. Matysiak and C. Clementi. Optimal combination of theory and experiment for the characterization of the protein folding landscape of S6: How far can a minimalist model go? *J. Mol. Biol.*, 343(8):235–248, 2004.

54. S. Matysiak and C. Clementi. Minimalist protein model as a diagnostic tool for misfolding and aggregation. *J. Mol. Biol.*, 363(1):297–308, 2006.

55. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21(6):1087–1092, 1953.

56. A. E. Mirsky and L. Pauling. On the structure of native, denatured and coagulated proteins. *Proc. Natl. Acad. Sci. USA*, 22:439–447, 1936.

57. J. Moult, K. Fidelis, A. Kryshtafovych, B. Rost, T. Hubbard, and A. Tramontano. Critical assessment of methods of protein structure prediction (CASP) round VII. *Proteins: Struct. Funct. Bioinf.*, 69(S8):3–9, 2007.

58. S. Oldziej, C. Czaplewski, A. Liwo, M. Chinchio, M. Nanias, J. A. Vila, M. Khalili, Y. A. Arnautova, A. Jagielska, M. Makowski, H. D. Schafroth,

R. Kazmierkiewicz, D. R. Ripoll, J. Pillardy, J. A Saunders, Y. K. Kang, K. D. Gibson, and H.A. Scheraga. Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: Assessment in two blind tests. *Proc. Natl. Acad. Sci. USA*, 102(21):7547–7552, 2005.

59. J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes. Theory of protein folding: The energy landscape perspective. *Annual Review of Physical Chemistry*, 48:545–600, 1997.

60. S. B. Ozkan, G. H. A. Wu, J. D. Chodera, and K. A. Dill. Protein folding by zipping and assembly. *Proc. Natl. Acad. Sci. USA*, 104(29):11987–11992, 2007.

61. G. A. III Palmer. Nmr characterization of the dynamics of biomacromolecules. *Annu. Rev. Biophys. and Biomolec. Struct.*, 104(8):3623–3640, 2004.

62. G. A. Papoian, J. Ulander, M. P. Eastwood, Z. Luthey-Schulten, and P. G. Wolynes. Water in protein structure prediction. *Proc. Natl. Acad. Sci. USA*, 101(10):3352–3357, 2004.

63. B. H Park and M. Levitt. The complexity and accuracy of discrete state models of protein structure. *J. Mol. Biol.*, 249(2):493–507, 1995.

64. J. W. Pitera and W. Swope. Understanding folding and design: Replica-exchange simulations of Trp-cage miniproteins. *Proc. Natl. Acad. Sci. USA*, 100(13):7587–7592, 2003.

65. S. Potluri, A. K. Yan, J. J. Chou, B. R. Donald, and C. Bailey-Kellogg. Structure determination of symmetric homo-oligomers by a complete search of symmetry configuration space, using NMR restraints and van der Waals packing. *Proteins: Struct. Funct. Bioinf.*, 65(1):203–219, 2006.

66. M. C. Prentiss, C. Hardin, M. P. Eastwood, C. Zong, and P. G. Wolynes. Protein structure prediction: The next generation. *J. Chem. Theory Comput.*, 2(3):705–716, 2006.

67. B. A. Reva, A. V. Finkelstein, M. F. Sanner, and A. J. Olson. Adjusting potential energy functions for lattice models of chain molecules. *Proteins: Struct. Funct. Genet.*, 25(3):379–388, 1996.

68. B. Richter, J. Gsponer, P. Várnai, X. Salvatella, and M. Vendruscolo. The MUMO (minimal under-restraining minimal over-restraining) method for the determination of native state ensembles of proteins. *J. Biomol. NMR*, 37(2):117–135, 2007.

69. D. R. Ripoll, J. A. Vila, and H. A. Scheraga. Folding of the villin headpiece subdomain from random structures. Analysis of the charge distribution as a function of pH. *J. Mol. Biol.*, 339(4):915–925, 2004.

70. G. D. Rose, P. J. Fleming, J. R. Banavar, and A. Maritan. A backbone-based theory of protein folding. *Proc. Natl. Acad. Sci. USA*, 103(45):16623–16633, 2006.

71. E. Schroedinger. *What is life?* Cambridge University Press, 1944.

72. D. E. Shaw and et al. Anton, a special-purpose machine for molecular dynamics simulation. *Comm. of ACM*, 51(7):91–97, 2008.

73. A. Shehu. Guiding a tree-based search for protein conformations in a projection space. In *Robotics: Science and Systems*, pages 31–39, 2009.

74. A. Shehu, C. Clementi, and L. E. Kavraki. Modeling protein conformational ensembles: From missing loops to equilibrium fluctuations. *Proteins: Struct. Funct. Bioinf.*, 65(1):164–179, 2006.

75. A. Shehu, C. Clementi, and L. E. Kavraki. Sampling conformation space to model equilibrium fluctuations in proteins. *Algorithmica*, 48(4):303–327, 2007.

76. A. Shehu, L. E. Kavraki, and C. Clementi. On the characterization of protein native state ensembles. *Biophys. J.*, 92(5):1503–1511, 2007.

77. A. Shehu, L. E. Kavraki, and C. Clementi. Unfolding the fold of cyclic cysteine-rich peptides. *Protein Sci.*, 17(3):482–493, 2008.

78. A. Shehu, L. E. Kavraki, and C. Clementi. Multiscale characterization of protein conformational ensembles. *Proteins: Struct. Funct. Bioinf.*, 76(4):837–851, 2009.

79. M. Shirts and V. J. Pande. COMPUTING: Screen savers of the world unite! *Science*, 290(5498):1903–1904, 2000.

80. G. R. Smith, M. J. E. Sternberng, and P. A. Bates. The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking. *J. Mol. Biol.*, 347(5):1077–1101, 2005.

81. H. Taketomi, Y. Ueda, and N Go. Studies on protein folding, unfolding and fluctuations by computer simulation: The effect of specific amino acid sequence represented by specific inter-unit interactions. *Int. J. Peptide Prot. Res.*, 7(6):445–459, 1975.

82. R. Unger and J. Moult. Finding lowest free energy conformation of a protein is an NP-hard problem: Proof and implications. *Bull. Math. Biol.*, 55(6):1183–1198, 1993.

83. W. F. van Gunsteren, D. Bakowies, R. Baron, I. Chandrasekhar, M. Christen, X. Daura, P. Gee, D. P. Geerke, A. Glättli, Hünenberger P. H., M. A. Kastenholz, C. Oostenbrink, M. Schenk, D. Trzesniak, N. F. van der Vegt, and H. B. Yu. Biomolecular modeling: Goals, problems, perspectives. *Angew. Chem. Int. Ed. Engl.*, 45(25):4064–4092, 2006.

84. A. J. Wand. Dynamic activation of protein function: A view emerging from NMR spetroscopy. *Nat. Struct. Biol.*, 8(11):926–931, 2001.

85. S. Wells, S. Menor, B. Hespenheide, and M. F. Thorpe. Constrained geometric simulation of diffusive motion in proteins. *J. Phys. Biol.*, 2(4):127–136, 2005.

86. H. Wu. Studies on denaturation of proteins XIII. A theory of denaturation. *Chinese J. Physiol.*, 5(4):321–344, 1931.

87. K. Yue, K. M. Fiebig, P. D. Thomas, H. S. Chan, E. I. Shakhnovich, and K. A. Dill. A test of lattice protein folding algorithms. *Proc. Natl. Acad. Sci. USA*, 92(1):325–329, 1995.

88. Y. Zhang. Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins: Struct. Funct. Bioinf.*, 8(S1):108–117, 2007.

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