BIOCHEMISTRY
including biophysical chemistry & molecular biology

Current Topic

pubs.acs.org/biochemistry

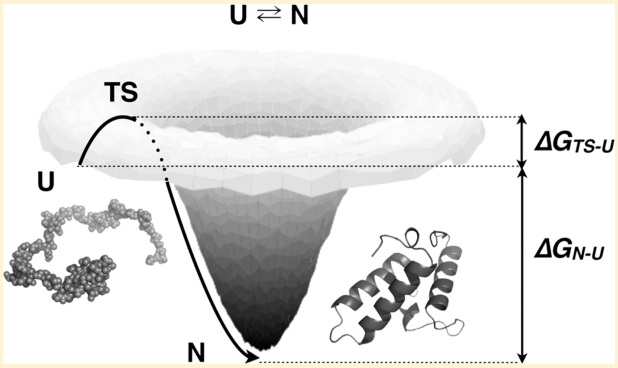
Computational and Theoretical Methods for Protein Folding

Mario Compiani*,†,§ and Emidio Capriotti*,‡,§

†School of Sciences and Technology, University of Camerino, Camerino, Macerata 62032, Italy

‡Division of Informatics, Department of Pathology, University of Alabama at Birmingham, Birmingham, Alabama 35233-7331, United States

ABSTRACT: A computational approach is essential whenever the complexity of the process under study is such that direct theoretical or experimental approaches are not viable. This is the case for protein folding, for which a significant amount of data are being collected. This paper reports on the essential role of *in silico* methods and the unprecedented interplay of computational and theoretical approaches, which is a defining point of the interdisciplinary investigations of the protein folding process. Besides giving an overview of the available computational methods and tools, we argue that computation plays not merely an ancillary role but has a more constructive function in that computational work may precede theory and experiments. More precisely, computation can provide the primary conceptual clues to inspire subsequent theoretical and experimental work even in a case where no preexisting evidence or theoretical frameworks are available. This is cogently manifested in the application of machine learning methods to come to grips with the folding dynamics. These close relationships suggested complementing the review of computational methods within the appropriate theoretical context to provide a self-contained outlook of the basic concepts that have converged into a unified description of folding and have grown in a synergic relationship with their computational counterpart. Finally, the advantages and limitations of current computational methodologies are discussed to show how the smart analysis of large amounts of data and the development of more effective algorithms can improve our understanding of protein folding.



Because proteins are only marginally stable under physiological conditions,[1−3] moderate fluctuations may cause deviations from the expected folding process (misfolding). In addition, it is becoming clear that prolonged exposure of hydrophobic surfaces during intermediate stages is responsible for pathogenic effects associated with protein aggregation,[4−6] whereas knowledge of the conformational dynamics of free proteins sheds light on the conformational changes of the molecule upon interaction with ligands.[7,8]

Thus, there is an urgent need to elucidate the dynamical mechanisms of folding and misfolding not only for clarifying the molecular basis of neurodegenerative diseases but also for boosting the constructive approach to protein science and polymer science. The new frontier in these areas includes modulating the folding dynamics of a natural protein,[9] the design of novel proteins,[10,11] and the synthesis of nonbiological foldamers.[12]

From a methodological point of view, the traditional modeling activity in terms of known physicochemical principles has been complemented by intensive processing of large amounts of raw data [crystallographic, thermodynamic, and kinetic or from nuclear magnetic resonance (NMR) and imaging techniques] because of the intrinsically statistical nature of the vast majority of the current investigations. This has brought about the massive development of computational methods, instrumental in processing the raw experimental data thus giving rise to further inferences, for shedding light on new correlations and possibly designing effective predictive methods. The resulting intermediate-level data, in their turn, provide the basic ingredients and insights for higher-level syntheses that result in general models, where established principles[13] and new heuristics[14] shape a deeper understanding of folding−misfolding processes. As a matter of fact, for systems with a complexity comparable to that of proteins, our comprehension is more realistically based on heuristics or qualitative "lessons" rather than on new quantitative laws.[15]

Computational and bioinformatics methodologies find themselves midway between experimental work and the more abstract activity of model building. This implies that besides changing our interpretation of the raw data,[16] they are expected to provide the basic elements that fuel more far-reaching modeling activity. However, one should not overlook the reverse influence. By way of example, hierarchical or modular models of the folding dynamics stimulate construction of new tools and call for new strategies for the organization of the existing data such as, for example, the introduction of searching strategies based on genetic algorithms and the construction of combinatorial libraries.[17]

Received: February 6, 2013
Revised: October 30, 2013
Published: November 4, 2013

 © 2013 American Chemical Society

To make the reader more familiar with the close interrelationships among theory, experiment, and computation, we first introduce the main physical and computational background notions in protein folding and summarize the recent evolution of the main theoretical models. We then focus on the main topic of this review, i.e., the large-scale organization of the data and the variety of computational tools designed to process them to address the threefold question posed by the protein folding problem: thermodynamic, kinetic, and computational.[18] In the last part of this paper, we describe the computational tools devised to cope with protein misfolding phenomena. In the concluding remarks, we comment on the noticeable changes induced in protein science in the era of big data.[19] The intermingled development of theory and computation is viewed as a direct consequence of the very fact that scientists have to manage, visualize, and interpret a deluge of experimental and computational data.

BASIC CONCEPTS OF PROTEIN FOLDING THEORY

In 1973, Anfinsen's paper[20] shifted interest away from the chemistry of disulfide bridges, and the conformational dynamics of proteins began to be examined through the eyes of the polymer physicist or the computer scientist.

Proving the uniqueness (stability) of the native structure entailed a further change in perspective in that one stopped thinking in terms of atomic coordinates. Actually it was argued that the relevant pieces of information for folding must reside in the sequence or, in logical terms, sequence → native structure. Since then, this has been known as the Anfinsen thermodynamic hypothesis. In the following sections that summarize the various models of protein folding, we shall see that an amended version of Anfinsen's hypothesis, a kind of extension to the realm of kinetics, is frequently invoked in the form sequence → folding kinetics. Clearly, this amounts to assuming that sequence also determines the properties of the basin of attraction associated with the native structure. This is at the core of the intriguing merger of two main lines of research concerning protein structure prediction and the investigation of protein folding dynamics.[21]

Going beyond the traditional chemical approach was also dictated by the excessive number of alternative conformations involved. The astronomical number of local minima in conformational space and the ensuing impossibility of exhaustive exploration are usually termed Levinthal's paradox.[22] Indeed, it has been shown that even a simple description of the protein folding based on the hydrophobic/hydrophilic (HP) model on a cubic lattice is NP-complete.[23] On the whole, the sequence-to-structure relationship and the elucidation of the folding events are such challenging problems that they are ranked among the most important scientific topics of the third millennium.[24]

Levinthal's paradox and Anfinsen's thermodynamic hypothesis, through their basic features, would seem to be in conflict. On the one hand, there is a need to constrain the folding process along a precise path (kinetic control), to ensure convergence toward the native state within a finite time. On the other hand, the dominant role of the final conformation makes the intervening path relatively unimportant (thermodynamic control) so that parallel alternative paths are admissible. These conflicting requirements become consistent within the context of the landscape theory of protein folding, in which both kinds of control are admitted.[25,26] Actually, the current view holds that parallelization is more sensible in the early stage of folding, whereas the process is more sequential-like in the later stages.[27,28]

The Physicist's View of Protein Folding. What matters within the context of this review is that the essential methodological contribution of physics is a rich tradition of methods for taming complexity and building up minimal (e.g., coarse grained) models.[29]

Protein physics introduced a mode of dealing with proteins forgetting the details of the biochemical machinery. This was paralleled by the so-called connectionist trend in cognitive sciences[30,31] where, to clarify the working mechanisms of the brain, very simple artificial models (artificial neural networks, in short ANNs) comprising simple units (neurons) and simple connections were devised and investigated.[32] In a similar vein, in protein science, the intricacies of biopolymers were condensed into the simple language of bead and string models borrowed from polymer physics[33] or in the even more skeletonized framework of Ising models,[34] the prototypical pictures of complex glassy systems,[35] which introduced the concept of frustration into the modern view of protein folding.

The wide applicability of notions such as frustration or topology reminds us of the basic intuition of synergetics, namely the existence of basic principles shared by many kinds of different systems (including proteins) irrespective of the specific nature of their elementary units.[36] More precisely, the slaving of variables, emergence of order parameters, and instability[36] have a precise counterpart in protein science in the concepts of hierarchy, reaction coordinate(s), and transition state, respectively. The comprehensive synthesis of these features is currently provided by statistical descriptions in terms of rugged energy landscapes (see Figure 1).[37]

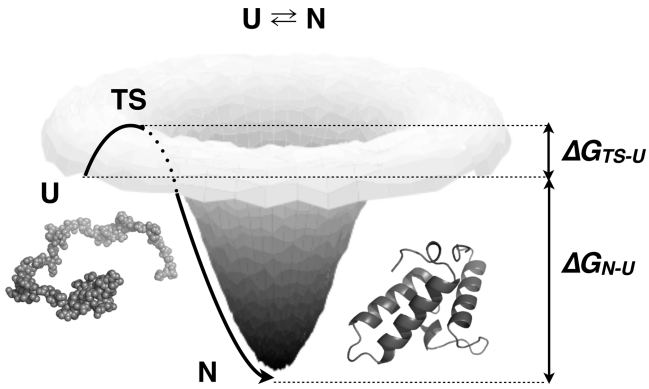


Figure 1. Typical funnel-like free energy profile for a two-state protein folding process from the unfolded (U) to the native (N) conformation.

The hierarchical nature of the energy landscape is reflected in the hierarchy of the time scales of the concurrent processes involved[38] and introduces the natural distinction between slaved (fast) and slaving (slow) variables.[39] Slaving of variables allows substantial simplification of the interdependent dynamics of coupled state parameters[36] in that the contribution of fast variables can be averaged out. As we shall see in Theoretical Models for Protein Folding, many folding models take advantage of this general property whenever structural elements at low levels in the hierarchy of protein structures are considered to evolve on shorter time scales than higher-level motifs (e.g., preformed helices, fragments of secondary structure elements corresponding to the so-called foldons,[40,41]

or in general the initiation sites of folding[42,43]) that drive the folding dynamics or feature in recognition processes in unstructured proteins.[44]

In protein folding studies, Kramers' kinetic theory for the barrier crossing of a Brownian particle[45] provides a useful paradigm for modeling a generic two-state process, i.e., the transition from the unfolded state (U) to the native state (N). Its major merit is to provide the simplest combination of the slaving principle, the notion of a reaction coordinate, and instability. In more refined theories, complete slaving of solvent (where relaxation of the solvent is assumed to be infinitely faster than the kinetics of the reaction coordinate) must be supplanted by more realistic pictures of solvent−protein coupling,[46] to account for the cooperative behavior of proteins[47] and environmental changes[48] and to provide a unified description of cold and warm denaturation.[49]

Deviations from Kramers kinetics and chevron plot rollovers reveal subtle effects of the multidimensionality of the energy landscape[50] and the attending multiple intramolecular interactions that set in during the folding of the protein (resulting in ruggedness, gating, and barrier fluctuations,[51] internal friction,[52] or multiple-barrier crossing[53]) or the coupling with the hydration shell that mediates the influence of the bulk solvent.

Kramers' theory and diffusion theory are also used within different contexts to estimate the rate of elementary folding processes such as the closing up of β-sheets,[54] the collision of elements of secondary structures,[55,56] or simply formation of an intramolecular contact.[57,58] This is in accord with a noteworthy suggestion from Ising-like descriptions (and also the tenets of synergetics) that both local folding (secondary structure formation) and global folding seem to obey the same basic principles.[59]

The Computer Scientist's Toolboox. A pondered assessment of the computer science methods that are currently applied in the protein folding literature would require a review of its own. Therefore, we only mention in passing that a frontal attack upon Levinthal's paradox aims to enhance brute-force computational strategies relying on parallel supercomputing architectures or using personal computing clusters,[60,61] although extensive number crunching cannot yet solve the time-demanding problem of folding a protein, though some interesting progress has been made recently.[62,63] Similar limitations also affect approaches based on low-cost worldwide parallel distributed computing (e.g., the Folding@Home project) involving thousands of contributors.[62]

The straightforward implementation of a mechanistic view of protein dynamics is represented by molecular dynamics (MD) simulations[61,64] and all its variants devised to cover a variety of time scales involved in folding.[65]

We prefer to turn our attention to a more accurate analysis of the impacting influence of Artificial Intelligence and in particular of some machine learning strategies. The story begins when the ANNs were applied to cope with the problem of predicting native protein structures directly from sequence (early 1990s). The efficacy of the machine learning approach is such that ANNs have definitely supplanted the earlier algorithms based on explicit statistics[21,66] and are currently used as standard tools.[67,68]

The essential novelty with respect to traditional computer programming is that the machine learning approaches are able to extract statistical information about the unknown sequence-to-structure mapping from a data set of examples in which sequences are associated with known, determined experimentally, native structures. From the description given above, it is clear that ANNs are equivalent to the implementation of a local version of Anfinsen's hypothesis (local sequence → local native secondary structure).

Although helical structures, as expected, can be well predicted,[66,69] helices can be ranked according to their sensitivity to long-range forces that makes them context-dependent. Such a distinction is crucial for the detection of the initiation sites (foldons) of the folding process that correspond to maximally context-independent helical stretches.[40]

These empirical findings have important implications for the general theory of folding. More precisely, two decades after their inception, structural studies based on machine learning are now playing a remarkable role in the process of unification of the theories of protein folding. To clarify this point, let us recall that, at the outset of protein science, structure predictions and the investigation of stability and folding pathways evolved into two separate areas of research.[21] However, in recent years, a fruitful dialogue between the two fields has been fostered by the finding that secondary structures can be quite successfully predicted by means of machine learning techniques, even though ANNs take into account only local interactions, but the most intriguing outcome is that progress in the application of the ANNs has facilitated the development of new folding models. A case in point is the foldon diffusion−collision model (see FDC model discussed in Theoretical Models for Protein Folding) in which ANNs and the Kramers approach merge in a single explanatory scheme.

In summary, the main implications of the successful results of machine learning approaches to the modeling of folding are as follows. (i) Simple structure prediction methods are able to circumvent Levinthal's paradox. (ii) They support the central role of a hierarchical mechanism in the formation of protein structures. (iii) The success of the FDC model corroborates the formulation of the "kinetic" Anfinsen hypothesis (sequence → folding kinetics). (iv) The fundamental role of secondary structures dictated by local interactions has been clearly recognized. (v) The existence of early embryos of native helical structures (foldons) has received convincing theoretical justification, which is coherent with the landscape view of protein folding, together with the related notion that the TS may be a deformed image of the native state. All these points will be more fully discussed in the next section.

■ THEORETICAL MODELS FOR PROTEIN FOLDING

In this section, we focus on simple folding models, which offer low-resolution pictures and depend on only a few parameters. These models are complementary to the more exact models that allow calculation of partition functions without resorting to free parameters.[70]

The rationale for building simple models is that the landscape structure of the folding scenario and the related remarkable robustness of proteins to mutations make quantitative predictions of the folding models rather insensitive to low-level details.[71] This property is closely related to the unexpected finding that generic pictures of the native state incorporated into topological parameters (contact order and all its variants) correlate well with the folding rates.[72] Hereafter, we follow a quasi chronological order to outline the main steps of the development of theoretical models of protein folding.

In 1973, the framework model[73] introduced the idea that parallelization of folding may be the key to circumventing Levinthal's paradox. At the end of the 1970s, the dynamics of

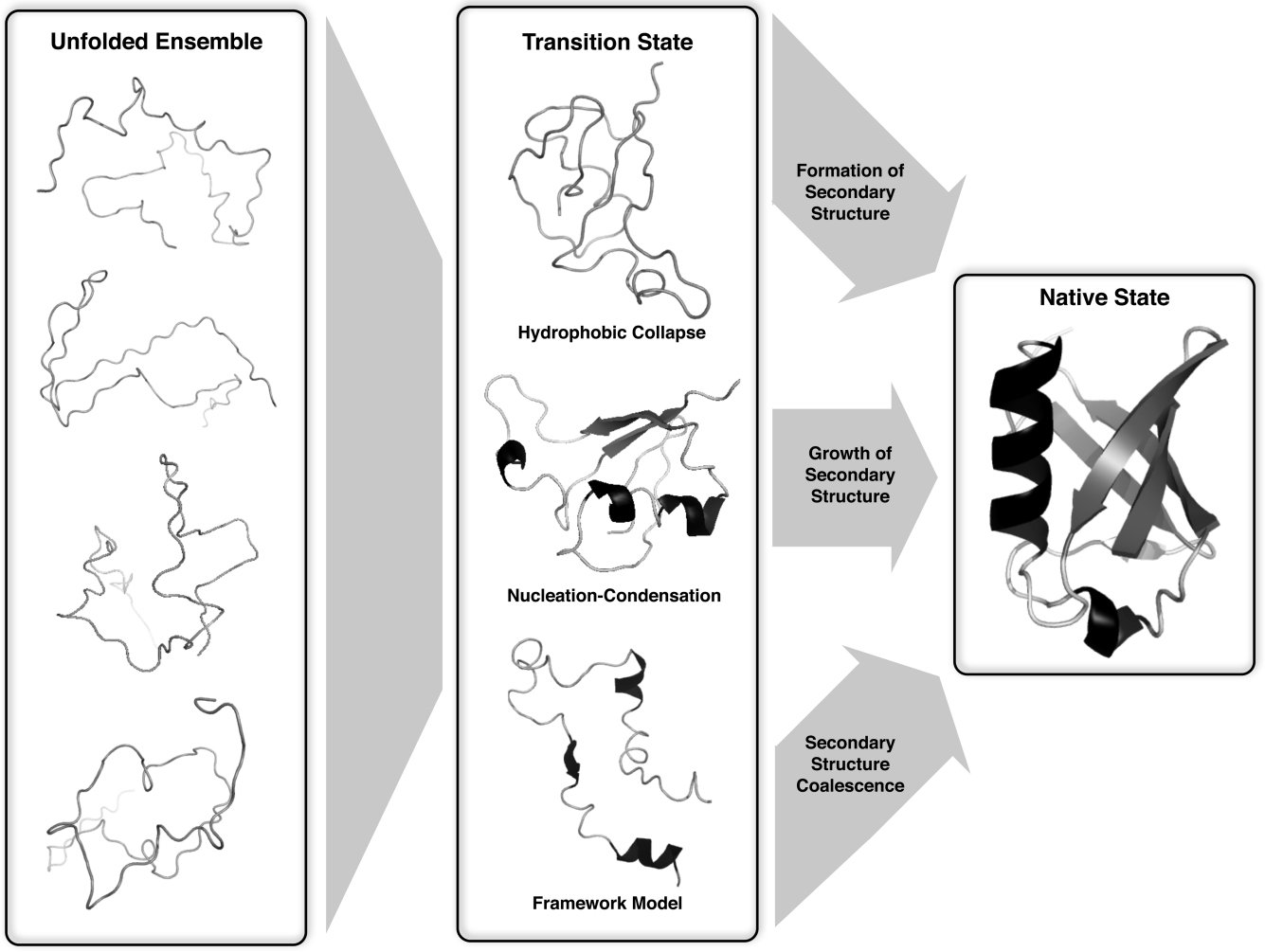


Figure 2. Three principal views of the protein folding dynamics according to the hydrophobic collapse (HC), nucleation−condensation (NC)/extended nucleus (EN) and framework models. The ordering (top-down in the middle column) follows the increasing hierarchical (or, equivalently, the decreasing cooperative) character of the three mechanisms. The essential differences among these scenarios are reflected in the properties of the transition state (TS), in which the role played by the secondary structure increases on passing from the HC to the framework model. The native state N of the protein, which is itself an ensemble of several conformations, is represented by a unique structure because, under physiological conditions, the fluctuations around state N are nearly negligible as compared with those observed around the unfolded state U.

protein folding was discussed in terms of a diffusion−collision (DC) model.[55] Since then, several different pictures have been proposed: the hydrophobic collapse (HC) mechanism[74] in 1977, the nucleation−condensation model[75] (NC) in 1995, and the foldon model[76] in 1996. The NC model was amended in 2000 and turned into the extended nucleus model (EN). Also, the DC model was subsequently modified into the foldon diffusion−collision (FDC) model[77] in 2004 for two-state proteins and the FDC3 version for three-state proteins in 2005.[78] The topomer search model (TSM) was introduced in 1999,[79] and the closely related hydrophobic zipper (HZ) model, the low-entropy-loss (LEL) routes model, and the zipping and assembly (ZA) model came to light between 1993 and 2007.[80−82] A visual representation of the main folding models is given in Figure 2.

Hydrophobic Collapse (HC). The HC mechanism predicts that hydrophobic forces and possibly backbone forces result in chain collapse prior to the formation of elements of secondary structure.[83] Long-range interactions precede or are concomitant with the establishment of local contacts; in the landscape perspective, collapse corresponds to the narrowing of the folding funnel and accounts for the essential part of the free energy balance (see Figure 1).[84,85] This view is in agreement with recent studies showing that burial information is conducive to effective reconstruction of the native structures of small globular proteins.[86,87] More generally, the paramount importance of the hydrophobic effect has motivated the investigation of minimalist models of folding based on hydrophobic interactions.[88,89] However, the nature main features of the collapse are still a matter of controversy. For example, it was argued that there are exceptions to the mandatory presence of collapse,[90] whereas in the case of collapse-mediated folding, other investigations have discussed the time scales of chain contraction or more general issues concerning kinetic versus thermodynamic control, such as the role of burial events in favoring a metastable active state with respect to the thermodynamically more stable native state.[91,92] A further critical topic is the relative importance of generic hydrophobic interactions versus specific local interactions in steering the unfolded protein toward a collapsed configuration.[91,92] Correlatively, there is an ongoing debate about the presence of residual structures under unfolding conditions and,

dx.doi.org/10.1021/bi4001529 | *Biochemistry* 2013, 52, 8601−8624

ultimately, about a more realistic characterization of the unfolded state.[91−94]

Conversely, the open question about the role of collapse in providing a suitable environment for the subsequent formation of secondary structure elements[91,92] may be of considerable importance for clarifying some aspects of current models of folding dynamics (summarized below in this section). It has been reported that metastable collapsed states [molten globules (MGs)] may be found on the way to the native state and have been assumed to play the role of general reaction intermediates.[95] Characteristic features of the MGs are a loose tertiary structure, a considerable amount of secondary structure, partial formation of a hydrophobic nucleus, and more or less tight packing of side chains according to whether they hinder (dry MG) or permit (wet MG) free passage of the solvent.[96,97]

The fold is already outlined in the MG state and turns out to be stabilized by hydrogen bonds and hydrophobic but not van der Waals interactions, so that at this stage, the fold is independent of the details of the side-chain interactions. The MG is the most compact of the unfolded conformations and is characterized by peculiar attractive or oscillatory hydrophobic interactions.[98] The typical magnitudes of the two major interactions stabilizing proteins (hydrophobic and hydrogen bonds) are nearly the same in the MG as in the native state, whereas van der Waals forces drive the later adjustments leading to the native state.

Hierarchical Mechanisms. The hierarchical view was initially put forward in the 1970s[73,99] and more recently proposed as a general mechanism.[100,101] Hierarchical organization of folding directly reflects the analogous organization of protein structures that can be dissected into separable domains endowed with marginal stability (e.g., the foldons in the FDC model or the microdomains in the DC model). Then the process of folding a protein can be considered a combinatorial mechanism of aggregation using the domains as building blocks that implements a bottom-up assembly mechanism.[100−102]

The archetypical example of a hierarchical mechanism is the framework model, a qualitative picture whose pivotal idea is that, because of the dominant role of local forces, formation of secondary structure precedes collapse of the backbone and native tertiary interactions. In general, three main steps are involved in the framework mechanism: the appearance of metastable segments of secondary structure in a manner independent of tertiary contacts, followed by a second step in which the local motifs undergo propagation or diffusion and collision to form a compact native-like scaffold, and finally a step in which all interactions are adjusted and the chain settles into the native structure. The framework scheme ensures that the compact state is essentially the collapse of elements of preexisting secondary structures and that the overall process is compatible with many alternative routes, i.e., different sequences of collisional events. Here we find the first clue about the idea that early choices must not be retracted, that is, location of the secondary structures must correspond quite precisely to the native location to avoid late reconfigurations that would substantially slow the process. This feature is the very essence of the requirement of minimal frustration (see Diffusion−Collision Models).

Nucleation−Condensation (NC) Mechanism. The HC mechanism and the framework model found a partial synthesis in the NC model. The NC picture, first put forward by Levinthal and Wetlaufer,[103,104] was corroborated by lattice simulations and experimental studies.[75] Ising-like models[105] for the folding of α-helices are the forerunner of the NC mechanism in that they introduce the same segmentation of the process into nucleation and propagation steps. It must be recalled that the NC mechanism lends itself to interpreting also the formation of β-hairpins.[106] The NC model envisages the formation of flickering embryos of secondary structure that are scarcely populated and are subsequently stabilized by long-range interactions. This results in the generation of a nucleus formed at the transition state (TS) and around which the remainder of the protein collapses. Distant contacts stabilize the nucleus, and the NC model is an example of a cooperative mechanism in which secondary and tertiary contacts are formed concomitantly. Furthermore, because multiple nucleation sites are conceivable, multiple paths are viable and are associated with the heterogeneity of the TS.

It has been noted that the NC scheme belongs to the category of models that seem more compatible with the large variability (6 orders of magnitude) of the experimental folding rates. The novelty is that the secondary structure becomes more important because it sheds light on some basic features of the TS.[107] Via examination of the TS, it turns out that the nucleus consists of residues that are more likely to belong to the developing native secondary structure. Thus, the coupling of secondary structure formation to the birth of the nucleus enhances the mechanism of substantial entropy reduction that accompanies the descent into the folding funnel (see Figure 1).

The NC model rationalizes the finding that the TS reproduces the native CO despite the high structural variability due to the fact that the majority of residues are accommodated in their native positions only in later stages of the folding path.[108]

Extended Nucleus. The extended nucleus scenario (EN) can be viewed as an amended version of the NC model.[109] The peculiarity of the EN mechanism is that it bridges the gap between seemingly irreconcilable theories, viz., the framework and NC mechanisms. As a consequence, the classical dichotomy, secondary structure first versus collapse first, is solved in terms of a variable synergy of stability and topology. This feature corresponds to recognizing the variable balance of long-range interactions and short-range interactions. The EN view envisages the possibility of having in the folding nucleus precursors of secondary structure elements of variable native character. As their native-like character decreases, the mechanism shifts from the framework picture to the NC scheme. This picture of the TS suggests that the stability of the stretches with native secondary structure and the loop closure entropic term (topology) are the essential determinants of the folding rate. In this manner, it has been noted that stability is very sequence-sensitive, whereas the folding dynamics is determined more by topology, i.e., gross properties of the fold.[72]

So far, we have sketched the path to a synthesis of folding schemes starting from the side of nucleation and collapse processes; we now examine other folding schemes that fall into the class of hierarchical processes and assign a more direct intervention to the elements of secondary structure.

Diffusion−Collision Models. In the late 1970s, Karplus and Weaver developed the DC model.[55,99] It was the first physics-based model of protein folding that provided a quantitative version of the framework model initially applied to helical proteins.[55,56,99,102] The DC model explores long-term protein evolution and allows reproduction of the large

dx.doi.org/10.1021/bi4001529 | Biochemistry 2013, 52, 8601−8624

amplitude changes that occur in the folding dynamics at the price of adopting a skeletonized description of the protein chain in the style of polymer physics. Pseudoparticles endowed with known structural properties (microdomains) are connected by structureless regions of the backbone. The DC model depicts the folding process as a sequence of collisional events involving the microdomains. The Brownian trajectories of the coupled microdomains result in the coalescence of these regions and eventually in the progressive formation of aggregates with an increasing degree of order. The maximally gross-grained choice was met in the DC scheme as the elementary microdomains were identified with the native helices. Considering a protein with N helices, each aggregate having order n ($1 < n < N$) corresponds to an intermediate state, whereas each native helix has $n = 1$ and $n = N$ indicates the aggregate with the highest possible order, which represents the end of the folding dynamics. Clearly, in the DC model, there is a natural mix of local and global forces that are assumed to be uncoupled and to operate on separated time scales in a clear-cut way. Helix formation is a very rapid process, which is then followed by slower diffusive motions that give rise to the progressive stabilization of larger aggregates. It should be noted that the assumption of preformed helices is a version of the slaving principle mentioned above. Concomitant effects of local forces (responsible for the formation and stability of the helices) and long-range forces (loop closure entropies and the hydrophobic effect) determine the coalescence probability of the microdomains upon collision and ultimately the rate of formation of all possible states (intermediates and native states).[55,56,99]

The various aggregates map the main states eligible as intermediates. The existence of such a web of interconnected states traversed by alternative folding routes leading to the native state was clearly demonstrated in the case of calmodulin.[110]

The DC model in its original formulation depends on several adjustable parameters and on the preliminary determination of structural features. For example, its implementation requires the determination of the native structure and the stability of each native helix. This difficulty was partially remedied by supplementing DC dynamics with algorithms designed to estimate the stability of helical domains.[55,56,99]

The FDC model is a more refined description of the folding that transforms the DC picture into a more self-contained tool for reconstructing the pathways and the kinetics of helical proteins. In the FDC model, an ANN is used to achieve the following goals: (i) to predict the secondary native structure, (ii) to identify the initiation sites of folding (foldons) of the protein[40] that replace the helical microdomains of the DC model, and (iii) to specify a measure of stability of the foldons themselves.[40,77] Helices without foldons do not participate in the rate-limiting step and belong to those structures that appear after the TS has been reached. Such a delayed formation of helices induced by foldon coalescence resembles the mechanisms in which binding and molecular recognition promote the stabilization of new helical structures.[96]

The FDC model uses sequence-specific features because the ANN reads the sequence of the protein. At this point, most of the free parameters of the DC model (location and stability of the microdomains) can be estimated directly from the protein sequence. Computing the coalescence probability requires the evaluation of the solvent-accessible surfaces of the colliding elements in the unfolded state and after the folding has been completed. This can be done by resorting to the tertiary structure of the protein. The FDC model has been successfully applied to two-state proteins,[77,111] to predict the kinetic effect of point mutations,[111] and to describe the folding dynamics of three-state proteins (FDC3 model).[78]

Topomer Search Model (TSM). The TSM[79,112] was motivated by the search for the physical factors underpinning the empirical correlation of the contact order (CO) with folding rate[113] and the cooperativity of folding dynamics. This view of folding originates from the finding that constrained simulations supplemented with simple burial criteria are sufficient to identify the native fold. The key concept is the topomer, which is the set of conformations sharing the same topology. The model assumes that there exists a clear-cut separation of the time scales between the topomer search, a large-scale rate-limiting event depicted as a random transition from the current topology to a drastically modified one and the intratopomer nonrandom local events, i.e., the growth or zipping of those local structures compatible with the current topology. The TSM works for β- and α/β-proteins but underestimates the rates of three-helix bundles. Moreover, the TSM cannot be extended to include prospective intermediates.

Like the FDC model and the TSM, a related approach[114] builds on the basic idea that bringing distant residues close to each other is the rate-limiting step of the folding process. Much as in the FDC model, in which the reaction coordinate is the number of coalesced foldons, here the coordinate is the number of native contacts formed. On the whole, the dynamics of contact formation provide a reasonable physical justification of the CO folding rate relationship essentially in terms of the entropy changes involved in the closure of the specific loops defining the topomer,[115] although excluded volume effects and local interactions are underestimated.[116]

Alternative folding models such as the hydrophobic zipping (HZ)[81] and low-entropy-loss (LEL) models[80,117] share some fundamental principles with the schemes mentioned above. The LEL model is based on the concept of effective contact order (ECO). ECO is a path-dependent CO providing information about the rates and routes. Minimizing the ECOs at each step of the folding amounts to minimizing the incremental entropy losses ensuing from loop closures and turns out to be a successful search strategy for identifying the dominant pathway on the energy landscape.

Determinants and Unification. As a main result of studies over the past decade, several widely debated dichotomies are being solved. Besides the dichotomy of secondary structure first versus collapse first (see Extended Nucleus), there is the distinction between two-state and three-state folding, closely related to counterposition of the NC scheme with the framework picture (exemplified by the DC model) or, alternatively, cooperativity with modularity. The turning point came with studies showing that proteins belonging to the same family exhibit either the former or the latter opposite folding mechanism[109,118,119] according to experimental conditions.[120]

This pointed to the existence of a common underlying mechanism expected to undergo a progressive shift from the NC to the DC extreme controlled by the most prominent determinants of folding, i.e., hydrophobicity (via TS rational redesign)[9] and helical propensity (in the N and TS states).[119,121,122] Gradual resolution of the dichotomies described above was favored by the improved temporal resolution of recent detection techniques stressing that intermediates are ubiquitous and their presence directly mirrors

the ruggedness of the energy landscape even in small single-domain proteins.[123]

The general convergence toward a universal picture is confirmed by comparing the same essential features captured by the EN, HZ, and FDC models.[82] Basically, the unifying mechanism starts with the formation of one or more nuclei, comprising a limited number of key residues that belong to nativelike marginally stable local structures; this step is followed by a stepwise global collapse eliciting formation of nonlocal contacts and accretion of the partially formed structures.

Not surprisingly, the EN and FDC pictures provide similar predictions for a set of proteins whose folding mechanisms are known to span the whole range between the NC and DC (fully cooperative and fully hierarchical) descriptions.[118] In addition, the flexibility of the FDC/FDC3 model (despite its limited applicability to all-α proteins) is also shown by its ability to reproduce the shift from two-state to three-state folding.[78] The alternative pathway view versus landscape view is also being reconsidered within a unified framework. This occurs in the FDC or the LEL route models in which the multiplicity of paths is compatible with the emergence of a bundle of similar paths (or a single path, as a limiting case) defining the dominant macro route.

DATABASES AND RESOURCES FOR PROTEIN FOLDING AND STABILITY

The growing opportunity to share information via the Internet has allowed the creation of curated databases in which experiments performed by single groups can be made available to the scientific community. This has alleviated the bottleneck effect plaguing the science of proteins due to the fact that the gathering of experimental data about the folding and structures of proteins is much more technically demanding than the recent sequencing technologies, so that structural studies lag far behind the more expedited characterization of new sequences.

Actually, a look at the current protein databases shows that the rate at which three-dimensional (3D) structures are determined is significantly lower than the rate of discovery of structural classes (see Figure 3). The reduced pace at which the amount of structural data grows is also due to the fact that, in general, different proteins belong to the same structural class, this finding being in agreement with the observation that protein structures are evolutionarily more conserved than sequences.[124] Not surprisingly, the question about the completeness of the known protein structural space still has to be answered. In this section, we first give a bird's-eye view of the main physical quantities that constitute the basic language in which most theoretical considerations of protein structures, thermodynamics, and kinetics are usually cast. We then review the online databases of experimental data on protein structures, folding thermodynamics, and kinetics.

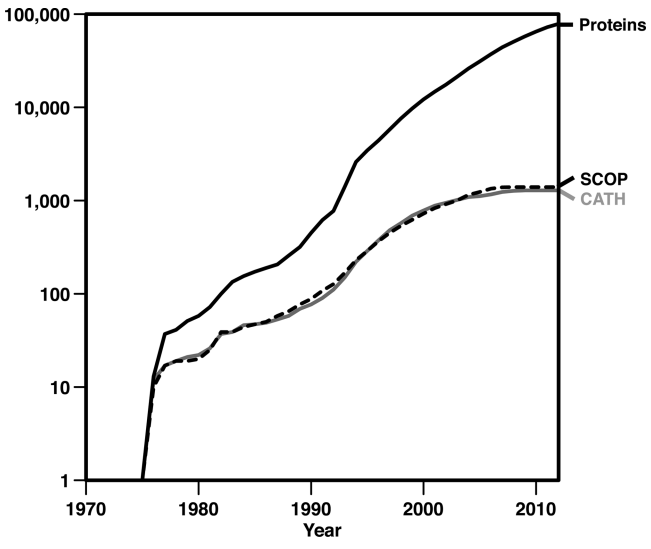


Figure 3. Growth of PDB, SCOP, and CATH data during the past several decades.

Interactions in Protein Folding. Structural, thermodynamic, and kinetic features are essential for defining the details of the folding process. In particular, the protein 3D structure gives important clues about important interactions occurring in protein folding (see Figure 4).

In proteins, the two main forces stabilizing the native structure are electrostatic interactions and the hydrophobic effect, whereas the main destabilizing force is the loss of conformational entropy.[125] Electrostatics is essential for elucidating the relationships between structure and function. Among electrostatic forces, the long-range interaction between two neighboring and oppositely charged residues (salt bridge) plays an important role in protein structure and function, including diverse processes such as oligomerization, molecular recognition, allosteric regulation, domain motions, and α-helix capping. Moreover, salt bridges are expected to stabilize the native state of proteins, but the current experimental and theoretical estimates of their free energy contributions vary significantly.[126] The local and nonlocal electrostatic interactions that determine the propensity of each residue for any secondary structure type have an only moderate influence on stability as shown by studies proving its weak dependence on the pH and salt concentration and the negligible evolutionary conservation of the charged residues. Other important electrostatic interactions include the hydrogen bonds whose contribution to the folding process is still controversial because it is strongly dependent on the polarity of the microenvironment.[127] Proper accounting for the wealth of electrostatic interactions in the interior of proteins is also complicated by recent studies pointing to the unexpected ability of proteins to tolerate the substitution of internal positions with charged residues, whether basic or acidic,[128] to the difficulty of assigning physically consistent values to the dielectric constant in the core of proteins,[129] and the controversy over the existence of stabilizing n$-\pi^*$ interactions[130] that have recently been proposed as dipole−dipole interactions.[131]

This experimental and theoretical evidence suggests that protein stability is to be ascribed to a delicate balance of many weak forces. By way of example, it has been argued that van der Waals interactions, associated with tight packing of the hydrophobic regions in the protein core, result in a contribution to stability comparable to that of the classic hydrophobic effect.[132]

In any case, the primary role in the early phases of folding is played by the aversion of nonpolar hydrophobic residues to water. The hydrophobic interaction ensues from a collective effect favoring segregation of nonpolar residues that tend to avoid direct contact with the polar aqueous environment. Although the prevailing opinion is that hydrophobicity is responsible for the compaction of the protein and the concomitant formation of a hydrophobic core in the inner region of the folded molecule,[96] several aspects of this

dx.doi.org/10.1021/bi4001529 | Biochemistry 2013, 52, 8601−8624

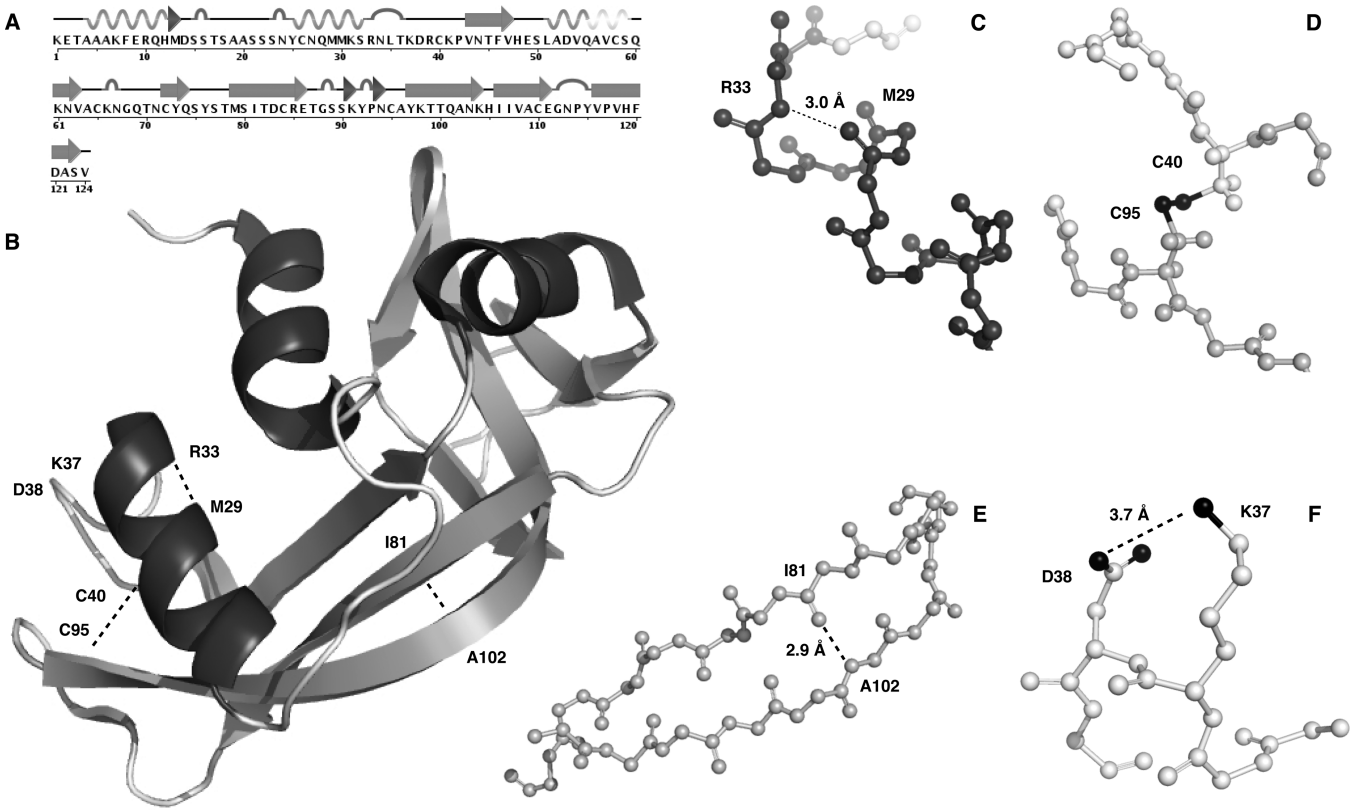


Figure 4. Different variety of interactions (hydrogen bonds, salt bridges, and disulfide bonds between cysteines) that define the secondary structure of proteins. (A) Protein's sequence in which the superimposed symbolism gives a one-dimensional visualization of the regions in α-helical (wavelet), β-sheet (arrow), and random coil (straight line) structure. The overall 3D conformation of the same protein is shown in panel B. Examples of hydrogen bonds stabilizing α-helices and β-sheets are reported in panels C and E, respectively. Examples of disulfide bond and a salt bridge are shown in panels D and F, respectively.

interaction need to be clarified,[133−135] including the operational definition of the hydrophobic free energies.

Thorough investigations of the final outcome of the hydrophobic effect (formation of an MG) have proposed to discriminate between conventional ("wet") and "dry" molten globules.[96] Both states precede the formation of the native structure; however, the wet MG corresponds to a dynamic core with nativelike secondary structure that can be accessed by the solvent and lacks close packing, whereas the dry MG is a tightly packed conformation from which water is segregated. An illustrative experiment has been conducted with a mutant of chorismate synthase in which a monomeric form of the enzyme, which folds through a wet MG, has been shown to retain the original activity of the dimeric enzyme.[136] More precisely, it turned out that the wet MG exhibits full enzymatic functionality and that after binding a substrate analogue the semifolded protein undergoes further compaction. The finding that a dry MG can occasionally give rise to detectable unfolding intermediates hints at the fact that protein−protein interactions may have a strong impact on the folding mechanism stabilizing or destabilizing intermediate conformations along the folding pathway.

The protein folding has many similarities with binding mechanism that can be described by funnel-shaped energy landscape.[137] Indeed, it has been frequently observed that folding and ligand binding may be closely coupled, although in most of those cases, the conformation of the protein is recognized by the ligand and the question of whether the ligand influences the folding kinetics is still poorly understood.[138]

These observations support the view that analysis of the protein structure is important for estimating the contributions of the various types of interactions to the stability of the native state. The recently proposed "fold approach"[139] involves extensive comparative analysis of the folding process of topologically, structurally, and/or evolutionarily related proteins to discern common patterns and trends in folding. This methodology helps in obtaining an improved understanding of the crucial steps of folding. Similarly, the distinction between the wet and dry MG that clarifies details of the later steps preceding the native state, the fold approach sheds light on the subtle dynamic arrangements leading to the TS (in terms of successive stabilization of the obligate and critical nuclei). This results in a more microscopic dissection of the folding process that allows the selective identification of the residues responsible for the definition of topology, function, folding, internal friction, and intermediates.[139]

Therefore, processing the raw structural data (atomic coordinates of the native structure) is important for generating a variety of useful intermediate data such as the solvent-accessible area per residue or group of residues that permits estimation of several structural and thermodynamic values, discriminates between buried and exposed amino acids, and allows evaluation of electrostatic forces and structural propensities for each residue.[140] Elaboration of the 3D structure is also important for characterizing the network properties of the native interactions;[141] it also allows visual inspection in the search for domains to be split into smaller building blocks that

dx.doi.org/10.1021/bi4001529 | Biochemistry 2013, 52, 8601−8624

may provide milestones for the most probable folding pathways (see Figure 1).[142]

In general, the development of computational methodologies tends to supplant these reductionistic approaches with more straightforward and phenomenological procedures. The price to pay is the lack of interpretation in microscopic terms; for example, structural preferences for each residue are no longer traced back to atomic interactions but are simply the outcome of some statistical treatment of the structural databases.

Since the beginning of the 1990s, statistical approaches have undergone significant evolution; today, sophisticated statistical algorithms predict protein folding features by implementing machine learning approaches. A related example is the use of structure-based methods for the prediction of protein thermodynamics that build on parametric equations whose free parameters are optimized by using structural and thermodynamic reference data on the native states of crystallized proteins.[143] It is to be stressed that although machine learning methods have attained a good level of accuracy in the prediction of several protein structural and thermodynamic features, their performance is still limited in the case of new folds that lack sufficient information about evolutionarily related proteins.

Protein Structure Data. As mentioned in the preceding sections, 3D protein structures are useful for the development of simplified models.[144] The most comprehensive repository of macromolecular structures is the Protein Data Bank (PDB).[144] Currently, the PDB contains more than 93000 macromolecular structures, ∼93% of which are proteins and ∼7% of which are nucleic acids (either isolated or in complexes with proteins). In the database, ∼88% of the structures were obtained by X-ray crystallography and the remainder using NMR and other techniques.

The complete definition of the space of protein structures is important for selecting key interactions for the stabilization of the native conformation and depends on the procedure used to classify the proteins included in the PDB. The current gold standards for the classification of protein structures are SCOP[145] and CATH.[146] The Structural Classification Of Proteins (SCOP) is a database composed by manually classified protein structure domains based on their similarities. It is a hierarchical classification comprised of the following levels: species, protein, family, superfamily, fold, and class. In the SCOP database, two domains that belong to the same fold have similar secondary structures in the same arrangement and with the same topological connections. CATH is a semiautomatic procedure for defining a hierarchical classification of the structures of protein domains. This classification is based on four levels: class, architecture, topology, and homologous superfamily. When two proteins have similar structural features and a high degree of sequence similarity in conjunction with similar functions, they are assumed to be evolutionarily related and, therefore, associated with the same CATH identifier.

To increase the rate of determination of structures and folds, the Structural Genomics (SG) project[147] has been started. The overall work of determination of new folds is distributed among the laboratories associated with the SG network in such a way that the activity of each of them focuses on selected target proteins.[148] However, despite the intensive work of the SG centers having determined more than 11500 structures, the number of known structural classes has not appreciably increased (see Figure 3). Therefore, we are still far from having a sufficient view of the universe of structural folds that, according to a recent estimation, amount to approximately 1700.[149] We are hopeful that an additional contribution from *in silico* protein design will allow the exploration of regions of the protein universe not yet observed in nature.[150]

Protein Thermodynamic Data. The stability of proteins can be quantified using various experimentally determined thermodynamic parameters. The most common are the free energy change (ΔG) in the protein, which measures the conformational stability estimated from a thermal denaturation curve, the free energy change of water (ΔG_{H_2O}), calculated from the unfolding curve as a function of the denaturant,[151] and the melting temperature (T_m) at which 50% of the protein is unfolded.[152]

Chemical Denaturation. In chemical unfolding experiments, the fractions of unfolded and folded protein are measured by adding denaturants such as urea or guanidine hydrochloride (GuHCl). In this case, the evaluation of the midpoint concentration (C_m) measures the concentration of denaturant at which 50% of the protein is unfolded. Three models are currently used for the interpretation of the unfolding data: the denaturant binding model, the solvent exchange model,[153] and the linear energy model.[154] They differ mainly in the description of the interactions responsible for unfolding. The first two models consider the protein as a collection of independent sites where the denaturant can bind reversibly. The last model assumes that the number of denaturant binding sites is proportional to the accessible surface area, and therefore, one expects a simple linear dependence between stability and the concentration of the denaturant. Because the fraction of folded and unfolded states cannot be measured directly, the relative population of folded molecules is estimated by various structural probes such as the absorbance and fluorescence at 287 nm, which quantifies the solvent exposure of tryptophan and tyrosine; far-ultraviolet circular dichroism (180−250 nm), which determines the secondary structure of the protein; dual polarization interferometry, which estimates the molecular size and density; and near-ultraviolet fluorescence, which detects changes in the environment of tryptophan and tyrosine.

Thermal Denaturation. Thermal denaturation is also generally modeled by assuming a two-state unfolding process. In these experiments, the free energy change (ΔG_u) is described as a function of the variations of the unfolding enthalpy (ΔH), entropy (ΔS), and heat capacity (ΔC_p) at constant pH and pressure.[143] Assuming that ΔC_p is temperature-independent, all the thermodynamic observables can be determined from a single differential scanning calorimetry thermogram of the system. More accurate estimates of ΔC_p can be obtained by subjecting the system to slight variations in pH or protein concentration. Alternatively, ΔC_p can be accurately measured from the change in the solvent-accessible surface area (SASA) of the protein upon thermal denaturation.[143] A recent paper provides full details about these experimental techniques for the determination of protein stability.[155]

The most comprehensive database of thermodynamic data of proteins and mutants is ProTherm,[156,157] a free online database that collects and documents experimental thermodynamic measurements taken from the literature. Its Web page includes an interface to facilitate searching in the database and sorting and visualizing the results as well as information about the protein sequence and protein structure (if available) for each thermodynamic experiment. The statistics of the entries in the

ProTherm database show that the amount of published thermodynamic data has increased significantly over the past few years (see Figure 5). In the current release (February 2013), ProTherm includes 25830 entries from 1902 bibliographic references regarding 740 individual proteins; for 311 wild types, supplementary data about mutants are also available (~81% of the entries are associated with single-point mutants). The largest part of the data was obtained by thermal (~61%) and chemical (~38%) denaturation using urea or GuHCl. Approximately 90% of the thermodynamic measurements use circular dichroism (43% of the total), differential scanning calorimetry (25%), and fluorescence (22%). The majority of ProTherm data concern two-state folding mechanisms, and only ~7% refer to multiple-state folding processes.

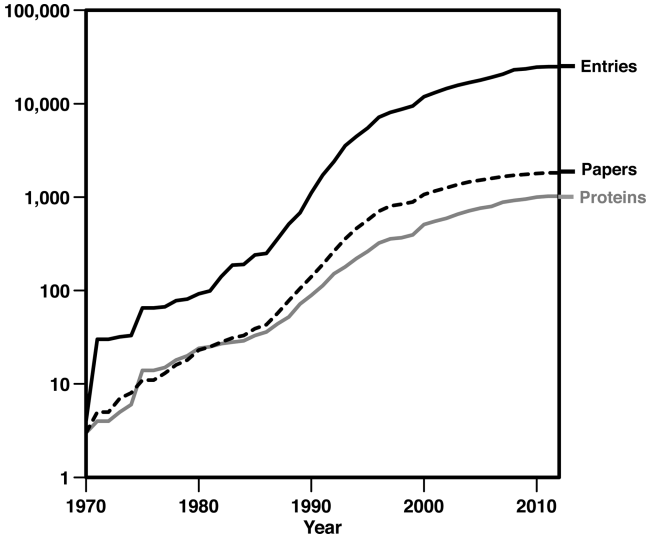


Figure 5. Growth of the amount of thermodynamic data vs time in the ProTherm database.

Protein Folding Kinetic Data. The kinetics of protein folding is another key aspect of the process. In the simple two-state folding mechanism, the equilibrium between unfolded and native states is established through a single TS; in this case, the rates of folding (k_f) and unfolding (k_u) are directly correlated with the equilibrium constant and consequently with the stability of the protein.

Because by definition the TS is not experimentally accessible, the folding kinetics are determined by studying the unfolding and refolding of the protein induced typically by a rapid jump in control parameters, such as denaturant concentration, temperature, pH, or pressure. The recovery of the equilibrium condition is monitored by spectroscopic techniques. Assuming that the folding and unfolding are single-exponential processes under a range of denaturant conditions, and that $\ln(k_{obs})$ (where $k_{obs} = k_f + k_u$) at different concentrations of denaturant gives rise to a "V-shaped" curve called a chevron plot, k_f and k_u can be evaluated by linear fitting the two sides of the curve and extrapolating to zero denaturant concentration. Many proteins show more complex folding mechanisms reflected in deviations from the standard chevron plot. From the experimental perspective, the presence of a populated intermediate state during the folding process is indicated by the nonlinear behavior of one of the chevron plot arms. Depending on whether the intermediate is on or off the pathway, different models have to be adopted to estimate the kinetic constants. A complete discussion of all possible scenarios was described in a recent review.[158]

Alternatively, the TS can be studied indirectly using protein engineering methods involving the introduction of single-point mutations into the protein sequence, which perturb the energies of the TS and the energy barriers between the native and unfolded states.

Kinetic data are less abundant than thermodynamic data. The databases of protein kinetic data on the Internet are manually curated by extracting the relevant information from the literature. In 1998, the first seminal review reported folding kinetics data of all the proteins studied up to that point.[159] Over the past few years, as more and more data have been generated, updating of the data set has become a time-consuming job of questionable value, because of the absence of a standard protocol. Such a protocol was suggested only in 2005 when kinetic data for 30 two-state proteins under standard conditions were collected.[160] A few years later, a more comprehensive collection of experimental data about the kinetics of folding was organized in the PFD and KineticDB[161] databases.

The main goal of KineticDB is to provide users with the diverse set of protein folding rates known from experimental work. For each record, KineticDB reports a single protein folding kinetics measurement extracted from the literature, the details of the protein under study, its best known 3D structure, the experimental conditions, and reference to the original paper. In the event that the exact protein structure is unknown, the structure of the closest homologue is provided. The experimental data include the natural logarithms of k_f and k_u extrapolated to zero denaturant, the natural logarithm of the midtransition rate of folding, the TS coordinate, ΔG_u (unfolding free-energy change) in water, and the type of folding mechanism (two-state or multistate). KineticDB contains kinetic data from 87 single-domain proteins and

Table 1. Databases and Resources for Protein Folding

name	URL	description	ref
		Protein Structure Databases	
PDB	http://www.pdb.org	macromolecular tertiary structure database	144
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop	structure classification of proteins	145
CATH	http://www.cathdb.info	HMM and domain protein 3D structure classification	146
		Thermodynamic and Kinetic Databases	
ProTherm	http://www.abren.net/protherm/	protein thermodynamic database	156
KineticDB	http://kineticdb.protres.ru/db/index.pl	manually curated databases of kinetic data	161
Muñoz lab	http://tmg.cib.csic.es/servers/data-tables	database of kinetic rates for proteins and their mutants	162, 163
PFD	http://pfd.med.monash.edu	kinetic database collecting folding rates and energies	247
REFOLD	http://refold.med.monash.edu.au	database of optimized refolding protocols	165

dx.doi.org/10.1021/bi4001529 | Biochemistry 2013, 52, 8601−8624

~100 mutants. Protein domains and short polypeptides without disulfide bonds in the native structure are treated separately. The protein lengths range from 16 to 396 residues. Currently, the database comprises 172 experimental kinetic data relative to 124 two-state and 45 multistate folding mechanisms. The $\ln k_f$ ranges from approximately −7 for slow folding proteins through multistate mechanisms to 16 for two-state fast folding proteins. Another database of kinetic data worth mentioning is PFD; in the latest version, PFD adopts standards for data acquisition, analysis, and reports that facilitate the comparison of folding rates, energies, and structures across various sets of proteins.

Alternatives to KineticDB and PFD are available on the Muñoz lab Web site (see Table 1). This Web site hosts two manually curated sets of data: the database of k_f and k_u values[162] for wild-type proteins, including kinetic data from 52 proteins, and the single-point mutant protein folding database,[163] collecting experimental kinetic data of ~1866 mutants from 26 proteins. The latter data set is particularly useful for the direct calculation of the ϕ values because the data are organized in different tables where, for each protein, the experimental kinetic data of the wild type and its mutants are grouped together. The tables in the Muñoz lab Web site report experimental values of parameters such as k_f and k_u and also the variations in ΔG_{TS-U} and ΔG_{N-U} between the wild type and mutant ($\Delta\Delta G_{TS-U}$ and $\Delta\Delta G_{N-U}$, respectively).

Manual curation of data extracted from the literature has proven to be useful for a number of new theoretical, empirical, and computational studies. In parallel, there has been growing interest in the sources of experimental error and variability, and recent work has shown that the experimental precision of $\Delta\Delta G_{N-U}$ across various laboratories is ~1.3 kJ/mol.[164] The accuracy of experiments is a significant issue for the development of predictive algorithms. The error associated with the experimental measurements is particularly important when selecting a reliable set of experimental data and for the statistical assessment of the results obtained from their analysis.

From the experimental point of view, another critical issue concerns the methodologies used to synthesize sufficient amounts of any protein. Overexpression of proteins in bacteria is a routinely used method. Unfortunately, a considerable proportion of the proteins expressed in bacterial hosts aggregate, forming inclusion bodies. Moreover, before the expressed protein can be subjected to any kind of biophysical or functional assays, it must refold to its native state. Therefore, it is important to define efficient protocols for protein refolding, which minimize the undesired competition of misfolding and aggregation reactions. These requirements are met by the REFOLD database.[165] It is a Web-accessible database describing published methods employed in the refolding of proteins. Currently, REFOLD contains 1156 annotated refolding records from 735 proteins in 288 organisms. Possibly, with an increase in the number of refolding protocols, REFOLD will become a reference resource for the optimization of protein renaturation.

A selection of Web available databases and resources is reported in Table 1.

METHODS AND TOOLS FOR PROTEIN FOLDING AND STABILITY

Over the past several years, the amount of content in the databases of experimental protein folding measurements has substantially increased. This has resulted in the development of several computational methods for predicting important aspects of protein folding using sequence and structure information. In this section, we describe a selected set of available Web tools for the prediction of protein structures, folding thermodynamics, and kinetics.

Algorithms for Protein Structure Prediction. Because of the growing appreciation of the importance of the native structures as a guide in the search for alternative conformations during folding, much work has been done in structural bioinformatics to develop computational tools for the prediction and assessment of native protein structures.[166,167] As a major consequence, this paved the way for the study of the folding process of proteins of unknown structure. Here we describe the tools for predicting protein tertiary structure providing a short list of algorithms representing the state of the art for this task.

Following the observation that protein structure tends to be conserved within families of proteins performing the same functions in different organisms, several approaches have been developed to address the problem of protein structure prediction. In general, current methods for structure prediction fall into two main categories: template-based modeling and free modeling. The methods belonging to the first class (e.g., threading and comparative modeling) rely on detectable similarities between the modeled sequence (target) and at least one known structure (template). The second class of methods, also termed *de novo* or *ab initio* methods, predict the structure directly from sequence, without relying on fold similarities. Today, it is commonly accepted that an empirical threshold of ~30% sequence similarity separates the region with a high degree of homology, where the target proteins can be predicted by homology modeling from the "twilight zone" (with a low degree of homology) where more sophisticated algorithms are needed. In general, the reliability of the resulting models is related to the level of sequence identity between target and template proteins. High-accuracy models require more than 50% sequence identity and have an average root-mean-square deviation (rmsd) of ~1 Å for the main-chain atoms. Medium-quality structure predictions (with 30−50% sequence identity) have an average rmsd of 1.5 Å on ~90% of the main-chain atoms.

For low-quality predictions belonging to the twilight zone (<30% sequence identity), the alignment errors may be considerable so that the model may predict a substantially incorrect fold.[168] On the other hand, for proteins whose structures have not been experimentally characterized, high-resolution models predicted by template-based methods can be reliably used to determine key interactions in the native structure and to interpret experimental thermodynamic and kinetic data.

Structure Prediction Tools. Where homology modeling applies, MODELER[169] represents one of the best choices for automatically performing all the steps required to build an accurate model.[170] MODELER predicts the tertiary structure of the target protein fulfilling the spatial restraints consistent with the sequence alignment between the target and template and the 3D structure of the template. Conservation is the key criterion for extrapolating sensible restraints from the template to the equivalent residues of the target, in that the more evolutionarily conserved the residues, the more stringent the structural constraints they introduce into the final putative structure. In addition, servers like ModBase[171] and Protein Model Portal[172] give access to large repositories of predicted

dx.doi.org/10.1021/bi4001529 | *Biochemistry* 2013, 52, 8601−8624

structures of millions of target proteins. If there is no clear similarity between the target protein and at least one template in the PDB, I-TASSER[173] and Robetta[174] can predict the structure using more complex template-free approaches. The I-TASSER algorithm implements a threading procedure in which the query sequence is matched against a nonredundant database of sequences to identify possible evolutionary relatives. The sequence profile resulting from the multiple-sequence alignment of homologue proteins is then used to predict the secondary structure of the protein. In the next step, all the templates are threaded through a representative PDB structure library combining seven threading programs to find the top-scoring templates. Finally, continuous fragments from the templates are used to assemble the predicted structure, including an *ab initio* approach to model the unaligned regions. A standard *de novo* structure prediction tool to cope with the latter problem is Robetta, which is an implementation of Rosetta.[175] The Rosetta algorithm predicts protein structures using a *de novo* approach based on a library of three- and nine-residue fragments. The fragments are selected according to their sequence similarity with the target and assembled using a Monte Carlo simulated annealing procedure.

Prediction Assessment. The evaluation of the predicted models is a key issue that must be addressed prior to their application in research studies. In 1994, the series of critical assessment of techniques for protein structure prediction (CASP) experiments was initiated to objectively test protein structure prediction methods and to serve as an official forum for the independent assessment of the state of the art in protein structure modeling.[176] The biannual CASP meeting has celebrated its 10th edition and is the reference for the international community of scientists working on protein structure prediction as well as for software developers and users. In parallel to the CASP experiments, several automatic methods have been developed to assess the quality of the predicted protein structures.[177] These algorithms use scoring functions that rely on physics-based energies, knowledge-based potentials, combined scoring functions, and clustering approaches. In particular, the first two methods calculate pseudoenergies that allow one to detect putatively destabilizing interactions in the protein structure and/or simulate the folding process. Methods implementing physics-based scoring functions compute the interaction energies via a force field that combines experimental observations with quantum mechanics. The most commonly used force fields are AMBER,[178] CHARMM,[179] and GROMOS,[180] which are also routinely applied in MD simulations.

In general, statistical potentials, also termed knowledge-based potentials, encode the statistical preferences of residues or atom types to be exposed to the solvent, or to interact with each other in a pairwise or higher-order fashion. Such preferences are extracted from subsets of protein structures, which describe the known structural space of globular proteins. The basic hypothesis underlying this approach is that protein structures contain clues about the stabilizing forces of protein folding, which can be reconstructed under the following assumptions. (i) The folding can be described in terms of a free energy function. (ii) The conformational energy can be approximated by two-body interactions. (iii) Conformations occurring frequently are expected to correspond to low-free energy structures. If such assumptions are true, it is likely that the minimization of the scoring function mirrors the minimization of the protein's free energy and corresponds to the observed native structure. Several knowledge-based potentials are currently in use,[181] including ANOLEA,[182] DFIRE,[183] and PROSA-web.[184]

Algorithms for the Prediction of Protein Stability. The characterization of the chemicophysical rules governing protein stability is one of the long-term goals of protein structure analysis[185] that is necessary for improving the effectiveness of rational protein design. The most favored method for pursuing this end is to study the effects of mutations on protein stability. This has led to the development of several methods for predicting stability changes induced by residue substitution.[186] These algorithms are mainly based on energy functions designed to assess the stability free energy of the protein and its mutants and/or machine learning-based methods trained to predict the stability changes upon mutation.

Energy Function-Based Approaches. Methods based on energy functions implement an algorithm for sampling alternative 3D conformations and ranking them according to an appropriate scoring function. The reason why such methods have been devised is that although, in principle, quantum mechanics is the proper framework for calculating rigorous solutions for these problems,[187] its application to proteins is not feasible because of the huge number of degrees of freedom involved. In general, the energy functions can be grouped into three major categories: (i) physical effective energy functions,[188,189] (ii) statistical potential energy functions,[190−192] and (iii) empirically defined energies.[193−196] The physically effective energy functions generally make use of MD simulations to estimate the difference in ΔG_{N-U} between the wild type and mutant ($\Delta\Delta G_{N-U}$). Although with distributed computing techniques it has been possible to perform up to 0.5 ms of MD simulation,[197] the estimation methods of physicochemical energy functions are computationally demanding, and their application to the analysis of large sets of mutations is still not viable.[195]

To reduce the search space, EGAD describes the unfolded state, aggregates, and alternative conformers explicitly with empirical models featuring the Optimized Potentials for Liquid Simulations All-Atom (OPLS-AA) force field approach to represent the atom−atom interactions, the generalized Born continuum model to describe the electrostatics, and solvent-accessible surface area-dependent terms to account for the hydrophobic effect. EGAD has been optimized using the binding free energy changes upon mutations for protein−protein interactions and independently tested in the prediction of $\Delta\Delta G_{N-U}$ upon mutations.

Recently, alternative approaches based on statistical potentials and empirical energy functions have been developed. PopMusic calculates the changes in stability of a given protein using different combinations of database-derived potentials that, in turn, depend on the change in solvent accessibility of the mutated residues. The first version of PopMusic implemented torsional and distance potentials derived from high-quality structures from the PDB. This combined energy function uses distance potentials to represent the main hydrophobic interactions that stabilize the core of the protein and torsion potentials to describe local interactions that stabilize the protein surface. In a recent version, the program includes an ANN to improve the accuracy of the resulting estimates.[198] A different statistical potential based on DFIRE[196] is used in DMUTANT. DFIRE is a potential that implements the distance-scaled finite ideal gas reference state approach to optimize a residue-specific all-atom statistical potential where

the background distributions of atom pairs is likened to an ideal mixture of uniformly distributed atoms. DFIRE was derived from a data set of 1011 high-resolution structures of nonhomologous proteins and has been tested in the prediction of the stability of 895 mutants.

Fold-X is one of the most popular algorithms for the prediction of $\Delta\Delta G_{N-U}$ based on empirical energy functions.[195] It allows fast and quantitative estimation of ΔG_{N-U} for proteins and protein complexes. The energy function implemented in Fold-X is a weighted sum of various free energy contributions, among them van der Waals interactions, solvation, electrostatic interactions, hydrogen bonds, and entropy terms. The weighted terms have been optimized over a set of 339 mutants from nine different proteins, whereas the method was tested using a blind test set of 667 mutants, and a set of mutants of 82 protein−protein complexes. After 5% of the outliers had been removed, the correlation between experimental and predicted $\Delta\Delta G_{N-U}$ values is 0.83 with a standard deviation of 0.81 kcal/mol. More recently, new energy functions have been proposed,[199,200] and although accurate comparisons are difficult to conduct, one can tentatively conclude that Fold-X and DMUTANT are the best downloadable tools in this category.[201]

Machine Learning Methods. Over the past decade, a large amount of thermodynamic data has become available online through the ProTherm database, and this has fostered the design of a new generation of machine learning algorithms. In particular, machine learning methods using ANNs,[202] SVMs,[203−205] random forest,[206,207] and decision trees[208] have proliferated once the prediction of $\Delta\Delta G_{N-U}$ upon single-point mutation was simplified to the mere binary prediction of the sign of $\Delta\Delta G_{N-U}$. All of these methods rely on preliminary information about protein sequences and/or structures. In 2004, I-Mutant was the first machine learning approach for the prediction of the sign of $\Delta\Delta G_{N-U}$ that implemented an ANN cross-validated over ∼1600 single-point mutations.[202] I-Mutant takes as input a 41-dimension vector encoding the mutation under study and the structure of the protein around the mutated position to predict the sign of $\Delta\Delta G_{N-U}$. A new release based on SVMs (I-Mutant2.0) introduces two improvements: it predicts the thermodynamic effect of the mutation using only sequence information[209] and predicts the actual value of $\Delta\Delta G_{N-U}$.[203] More recent machine learning methods[186] have a broader scope and are able to predict the effect of double- and multiple-point mutations.[207,210]

Recently, two independent studies assessed and compared the performances of the predictors based on machine learning techniques, using data sets of experimentally characterized mutants with no overlap with the training sets of any of the individual methods being tested.[201,211] The results show that sequence-based approaches are less accurate than structure-based methods, and although DMUTANT, FOLD-X, and I-Mutant are among the best methods for the prediction of the sign of $\Delta\Delta G_{N-U}$, the value prediction of $\Delta\Delta G_{N-U}$ remains a critical task. Interesting progress in this direction has been made with AUTO-MUTE[212] and the new version of PopMusic[198] that implements a hybrid approach combining statistical potentials and machine learning methods. As a general caveat, let us recall that machine learning approaches can be biased toward destabilizing mutations[204] and may be affected by overtraining effects. Thus, special attention must be paid to the cross-validation procedure, the critical process that ultimately determines the real efficiency of this class of predictors.

Methods for the Prediction of Protein Folding Kinetics. The growing number of experimental investigations of protein folding have led to the development of novel methods for the prediction of protein folding kinetics and mechanisms. These methods are mainly based on empirical models that take into account the topology of the protein and/or its residue composition. More recently, machine learning methods have been implemented to predict k_f and the folding mechanism.

Empirical Model-Based Methods. Methods implementing empirical models are based on the observed correlation between the logarithm of the in-water k_f and some topological parameters computed from protein 3D structure or from closely related proteins, such as single-point mutants or homologues with high levels of sequence identity.[113,213,214] In particular, several studies have shown that k_f can be predicted using topological properties of the protein structure such as the contact order (CO), the long-range order (LRO), the total contact distance (TCD), the cliquishness, and the multiple-contact index (MCI).[186] At the end of the 1990s, the correlation between ln k_f and CO in the native state of two-state proteins (correlation coefficient of −0.81) for the first time hinted at the importance of topology-dependent properties.[186] In that work, a distance threshold of 6 Å between heavy atoms was considered to discriminate contacting from distant residues. Similar papers focused on the fraction of nonlocal residue contacts with a sequence separation of >12 positions and a distance threshold of 8 Å.[215] This investigation, performed with 23 two-state proteins, resulted in a linear correlation coefficient between ln k_f and LRO of −0.78. In a more recent work,[214] the TCD (taking into account the CO and LRO) was the eligible topological parameter. Analysis of 28 two-state proteins has shown that the TCD is linearly correlated with k_f with a correlation coefficient of −0.88. Testing the correlations of alternative topological properties with experimental kinetic data revealed that parameters measuring contact interdependence, namely cliquishness or the clustering coefficient, can be used to predict k_f values of both two- and three-state proteins.[216] Indeed, it was reported that a combination of cliquishness and absolute CO (contact order not normalized to protein length) correlates with ln k_f with a correlation coefficient of 0.73 over a set of 40 proteins. In a recent work,[217] it was shown that two alternative definitions of the MCI for two- and three-state proteins correlate with the experimental k_f. This analysis, conducted over a data set of 75 proteins, resulted in correlation coefficients of −0.80 and −0.83 for the sets of 50 two-state and 25 three-state proteins, respectively. New algorithms based only on protein sequence information[218,219] have been developed to render the prediction methods independent of features relating to 3D structures. One such method[218] uses the predicted secondary structure to calculate the effective chain length as a function of the number of helices and number of residues in a helical conformation. When the experimental k_f was fit to the effective length according to a power law function, the comparison between experimental and predicted k_f values resulted in a negative correlation larger than 0.8 over a data set of 64 two-state and multistate proteins. A further approach to the prediction of k_f capitalizes on the nonlocal residue contacts (via the computed values of LRO normalized to the square of the protein length) and protein sequence information.[219] Quite interestingly, the results obtained by analyzing a set of 37 two-state proteins demonstrate that rate predictions based on

dx.doi.org/10.1021/bi4001529 | Biochemistry 2013, 52, 8601−8624

Table 2. Methods and Tools for Protein Folding

name	URL	description	ref
	Protein Structure Prediction Tools and Resources		
I-TASSER	http://zhanglab.ccmb.med.umich.edu/I-TASSER	structure prediction by threading	173
ModBase	https://modbase.compbio.ucsf.edu/scgi/modweb.cgi	repository of models predicted by homology	171
MODELER	http://www.salilab.org/modeller	standard homology modeling tool	169
Protein Model Portal	http://www.proteinmodelportal.org	resources and services for protein structure prediction	172
ROBETTA	http://robetta.bakerlab.org	*de novo* and homology modeling algorithm	174
	Physics-Based Energy Functions		
AMBER	http://amber.scripps.edu	molecular mechanics force field and package for simulations	178
CHARMM	http://www.charmm.org	empirical atomic force fields for molecular dynamics	179
GROMOS	http://www.igc.ethz.ch/gromos	energy function included in GROMACS	180
	Knowledge-Based Potentials		
ANOLEA	http://protein.bio.puc.cl/cardex/servers/anolea	atomic statistical potential scoring nonlocal interactions	182
DFIRE	http://sparks.informatics.iupui.edu/yueyang/DFIRE	residue-specific and distance-scaled mean force potential	183
PROSA-web	https://prosa.services.came.sbg.ac.at	knowledge-based potential for scoring protein structures	184
	Prediction of Protein Stability		
AUTO-MUTE	http://proteins.gmu.edu	machine learning and statistical potential for $\Delta\Delta G$ predictions	212
CUPSAT	http://cupsat.tu-bs.de	statistical potentials for the prediction of $\Delta\Delta G$	193
DMUTANT	http://sparks.informatics.iupui.edu/hzhou/mutation.html	prediction of $\Delta\Delta G$ using DFIRE statistical potential	196
Fold-X	http://foldx.crg.es	empirical scoring function for the prediction of protein stability	195
I-Mutant	http://folding.biofold.org/i-mutant	sequence and structure SVM-based method	203
PopMusic	http://babylone.ulb.ac.be/popmusic	neural network and statistical potential for $\Delta\Delta G$ predictions	248
PreThermut	http://www.mobioinfor.cn/prethermut	random forest for single- and multiple-mutation predictions	207
ProMaya	http://bental.tau.ac.il/ProMaya	random forest and filtering model for $\Delta\Delta G$ predictions	206
MuPro	http://mupro.proteomics.ics.uci.edu	structure-based SVM for $\Delta\Delta G$ predictions	205
SDM	http://mordred.bioc.cam.ac.uk/sdm/sdm.php	statistical potential for the prediction of $\Delta\Delta G$	192
sMMGB	http://compbio.clemson.edu/downloadDir/mentaldisorders/sMMGB_pdb.rar	generalized Born method for $\Delta\Delta G$ predictions	200
	Prediction of Folding Kinetics		
FREEDOM	http://bioinformatics.myweb.hinet.net/freedom.htm	prediction of protein folding rate change upon mutation	222
FOLD-RATE	http://psfs.cbrc.jp/fold-rate	folding rate predictor based on amino acid properties	221
K-FOLD	http://folding.biofold.org/k-fold	SVM-based methods for folding kinetics predictions	220
SeqRate	http://casp.rnet.missouri.edu/fold_rate/index.html	sequence-based SVM method for folding kinetics predictions	249
TCD	http://sparks.informatics.iupui.edu/Softwares-Services_files/tcd.htm	folding rate predictions based on total contact distance	214

estimated LRO values from noisy contact predictions are almost as accurate as those deduced from the known contacts.

Machine Learning Methods. In 2006, publication of the experimental kinetic data for the folding of more than 60 proteins made possible the first applications of machine learning methods for the prediction of ln k_f.[220,221]

The physicochemical, energetic, and conformational properties of amino acid residues and the structural classification of proteins lie at the foundation of FOLD-RATE,[221] which implements a multiple regression for predicting k_f. Testing with a jackknife procedure on 77 two- and three-state proteins resulted in a correlation coefficient between experimental and predicted k_f values of 0.97. Soon after that K-Fold, the first SVM algorithm for the prediction of k_f and the folding mechanism, was proposed.[220] K-Fold has been tested using a 5-fold cross-validation procedure on a set of 64 two-state and multistate proteins. To avoid overestimation of the performance, in the cross-validation procedure homologous proteins were grouped into the same subset. The method takes as input a vector encoding the chain length and CO and discriminates between two-state and multistate protein with an accuracy of 81% and a correlation coefficient of 0.60. When trained and tested for the prediction of lnk_f, K-Fold gives a correlation coefficient of 0.74.

At present, a significant number of experimental data about the variation of k_f upon mutation are known. These data have been incorporated into FREEDOM,[222] which implements a quadratic regression model to predict accelerating or decelerating mutations using 12 selected amino acid features. Tested by a 10-fold cross-validation procedure over 467 mutants, FREEDOM has an accuracy of 74% and a correlation coefficient of 0.31.

A general comment on the reliability of these methods is in order. Because the available number of experimental kinetic data is still relatively limited, a definitive assessment of the current methods is difficult. Nonetheless, in the presence of homologous proteins in the data set, the jackknife procedure is likely to overestimate the performance in the prediction of k_f. Thus, we suggest that a fair evaluation of the algorithms should be made using a 5- and/or 10-fold cross-validation procedure where homologous proteins are kept in the same subset.

Table 3. Tools for Protein Aggregation Available on the Web

name	URL	description	ref
		Empirical Methods	
AGGRESCAN	http://bioinf.uab.es/aggrescan	aggregation propensity scale from *in vivo* experiments	232
Tango	http://tango.crg.es	physicochemical properties of β-sheet formation in core regions	230
Zyggregator	http://www-vendruscolo.ch.cam.ac.uk/zyggregator.php	physicochemical propensities of residues	233
		Structure-Based Tools	
BETASCAN	http://groups.csail.mit.edu/cb/betascan	β-strands and strand pair scores from parallel β-sheet	250
FoldAmyloid	http://bioinfo.protres.ru/fold-amyloid/oga.cgi	hydrogen bond probability and residue packing density	251
Net-CSSP	http://cssp2.sookmyung.ac.kr	β-strand propensity from buried and highly interacting regions	252
PASTA	http://protein.cribi.unipd.it/pasta	β-parallel and antiparallel scores for amyloid formations	231
Waltz	http://waltz.switchlab.org	position-specific scoring matrix to predict aggregating regions	253

A summary of available Web tools discussed in this section is reported in Table 2, while a comprehensive and updated reference can be found in a recent review.[186]

PROTEIN FOLDING AND DISEASE

Our understanding of the protein folding mechanism becomes even more challenging when its implications for human morbidity are considered. In general, the formation of the native three-dimensional structure is an obligatory step to render the protein functional, though in certain cases the unfolded or partially unfolded states are known to be equally important in events such as translocation across membranes, protein trafficking and transport, the degradation of protein molecules,[5] and the normal working conditions of enzymes.[92]

Several recent publications reported a broad range of misfolding diseases,[223] related to the failure of proteins to adopt their native functional conformations. Although these pathological conditions can operate through various mechanisms, their general end result is a reduction in the quantity of the functional protein available. The low protein concentration may be associated with the increased probability of degradation by the quality control system of the endoplasmic reticulum,[224] improper trafficking,[225] or the decreasing solubility and formation of amyloid fibrils.[226] These alternatives to the standard folding mechanism are activated when the protein adopts a misfolded conformation, which in general is transiently stable. A protein can be trapped in a misfolded conformation when a partially folded intermediate state is formed or when interactions with other molecules are established. This scenario is in agreement with the idea that, under particular conditions, changes in the folding energy landscape result in the formation of non-native protein states. Thus, biological evolution seems to exert a kinetic control over the folding process to avoid the formation of alternative incorrect states that may end up with the molecular degradation or the formation of amyloids.[227]

In particular, understanding the aggregation process is even more important if we consider that inside the cell, proteins are expressed at very high concentrations and interact with a heterogeneous environment. Contrary to *in vitro* experiments, where the protein in dilute solution tends to follow a relatively small set of rules, *in vivo* experiments show that proteins exhibit a more complex behavior.[228] The growing interest in misfolding diseases led to the development of several algorithms for the prediction of peptide and protein aggregation. The prediction tools can be divided into two main categories: empirical and structure-based. The empirical algorithms use individual or combined amino acid properties such as hydrophobicity, β-propensity, and solubility to evaluate aggregation propensities. Alternatively, the methods based on structural information predict amyloid aggregation analyzing available 3D structures of peptides that are known to adopt fibrillar structures. The first method for the prediction of *in vitro* experimental aggregation data was published in 2003.[229] In that study, it was found that the change in the rate of aggregation of AcP mutants correlates with changes in amino acid hydrophobicity, the propensity to form α-helical and β-sheet structures, and overall charge. The same applies to the Tango algorithm[230] that predicts protein aggregation using the physicochemical principles underlying β-sheet formation. Tango accurately accounts for the aggregation propensities of 179 peptides collected from the literature as well as 71 peptides derived from human disease-related proteins, including prion protein, lysozyme, and β2-microglobulin. A sequence-based method for the prediction of protein aggregation is PASTA,[231] which uses two different propensity scores, depending on the orientation (parallel or antiparallel) of the neighboring strands; the scores are calculated from data sets of known native structures of globular proteins. PASTA associates the more likely conformation adopted by the fibril with the minimum of the two scores. In 2007, AGGRESCAN[232] was developed to detect short and specific sequences prone to aggregation and is based on an amino acid aggregation propensity scale derived from *in vivo* experiments. AGGRESCAN has been validated by comparing experimental data for regions that promote aggregation, experiments with fragments, and short synthetic peptides that notoriously aggregate *in vivo*. The more recent Zyggregator algorithm[233] relies on a position-dependent score calculated over a seven-residue window. The score combines amino acid scales for α-helix and β-sheet formation, hydrophobicity, and charge. Zyggregator, which has been optimized through fitting aggregation data from *in vivo* experiments, has been successfully tested in the prediction of the toxicity of aggregates produced by genetic variants in *Drosophila* models of Alzheimer's disease. Further examples of tools available on the Web implementing both empirical and structure-based algorithms are listed in Table 3. A recent review[234] reports the exhaustive assessment of 12 methods for predicting data generated from *in vitro* studies. The selected algorithms were tested on the task of the indirect prediction of the aggregation propensity change upon mutation that is known to inversely correlate with the experimental solubility. Analysis of the predictions shows that AGGRESCAN and Tango are among the best methods for the prediction of aggregation propensity changes measured *in vivo* in experiments with the Aβ_{42} mutants of *Escherichia coli*. Similar tests conducted with experimental data from other mutated proteins showed that the accuracy of some methods is protein-dependent. In the same paper, the accuracy of Zyggregator has been tested in the prediction of *in vivo* phenotypic effects of the

dx.doi.org/10.1021/bi4001529 | *Biochemistry* 2013, 52, 8601−8624

aggregating variants. The results show a strong negative correlation between the predicted changes in aggregation propensity upon mutation and the longevity and locomotor ability of the model organism.

Although available methods for the prediction of peptide and protein aggregation work with satisfactory accuracy even in the case of *in vivo* experiments, more thorough testing is needed to understand their robustness in view of their application to large-scale protein sequence analysis.

CONCLUDING REMARKS AND FUTURE PROSPECTS

The impressive proliferation of databases and computational tools described in this review is the inescapable manifestation of the recent exponential development of the science of proteins. Dealing with large data sets is now a trend that is a common feature of a wide variety of scientific fields,[235] and this implies a change in the way of doing science, requiring a suitable cultural shift involving personal attitudes, professional expertise, and institutional organization.[236]

This suggests the need for some general comments to put the computational approach to protein folding in an overall perspective. The novel character of this stage of science is the rapid dissemination of large amounts of experimental and computational results in such a way that their further elaboration can be distributed worldwide. In a sense, this is the materialization of the vision of H. G. Wells who more than a century ago was dreaming of a utopia with "reports of scientific experiments, as full, as prompt as telegraphic reports of cricket".[237] The @Home projects (e.g., SETI@home, Einstein@home, Rosetta@home, and Folding@home) testify to the last frontier of such a parallelization process now underway in many research areas.

Also the community of protein scientists experiences the fact that experimental data are so complex or are being produced at such a rate that exceeds by far the working capabilities of any single laboratory or research center. Thus, the challenging task of processing unmanageable amounts of data or conducting time-demanding minimizations or simulations is solved in the context of the parallel distributed processing paradigm. Relevant implementations of such a paradigm are provided by the PSI (protein structure initiative) and its sequel, PSI:Biology. These are large-scale initiatives pursuing global parallelization of research on structural and functional problems in medicine and biology through a highly collaborative worldwide network of laboratories and individual investigators.

Quite frequently, the computational tools are expected to remedy the scarcity of experimental data. In these cases (e.g., MD simulations), according to the traditional view, theory formulates in rigorous terms the task to be accomplished whereas computation is ascribed a merely executive role. However, it may happen that the computational tool is not preliminarily programmed following a previously defined theoretical framework but, instead, can stand on its own two feet by virtue of some learning algorithm and the ensuing automatic learning capability.

This fact is related to the predominantly statistical and heuristic character of most of the computational methodologies discussed in the previous sections. Clearly, the statistics, be they computed in a direct fashion or automatically through machine learning algorithms, require a set of experimental reference data (training set) whose general properties influence the reliability of the final results or predictions.

The general consideration that heterogeneous data generate hardly comparable results suggests that special care should be taken to ensure that the databases include data that satisfy as uniformly as possible reliability requirements. On this point, it is worth mentioning that the Critical Assessment of protein Structure Prediction (CASP) and the Critical Assessment of Genome Interpretation (CAGI) initiatives represent a valuable step toward the establishment of standard assessment criteria and the creation of a collaborative environment.

Furthermore, once the standardization requirement is met, one should not forget that the sets of data used to implement statistical and machine learning approaches are to be accurately balanced to minimize the risk of generating biased results.

From all these considerations, it turns out that increasing levels of coordination between the computational and theoretical communities and the experimental teams are essential for guaranteeing sensible scientific results. In the near future, the role of scientific curator (biocurator) and scientific animator will become extremely important. They will serve as interfaces between humans and computers as well as between data suppliers and data users.[236,238] The basic point is that existing data are useful only if they can be easily accessed and logically related to each other. On top of that, we recall that complex data are informative provided they are managed through appropriate visualization techniques providing powerful means of extracting new pieces of information from torrents of data.[239]

Actually, one of the noteworthy consequences of handling large data sets is the rising awareness that data visualization[239] and smart searching are the compelling complement of smart science. Coming back to the main subject of this review, the very fact that searching, visualizing, and researching are becoming so strictly connected is one of the basic reasons for our emphasis on close links between the computational and theoretical modeling activities.

In this regard, there is an increasing number of examples that show that dwelling on alternative ways of representing data may be conducive to rethinking some theoretical aspects, and this eventually opens up new paths toward understanding science as well as communicating science.[238,239] Briefly, proper (improper) visualization may promote (hinder) theoretical progress as well as viable applications of standard theories. One striking example is the Foldit multiplayer online game that translates the refinement of protein structures into a puzzle-solving problem.[240]

The novel feature emerging from these recent developments in protein folding research is that the visualization techniques may happen to exceed their traditional auxiliary role and can, instead, play a key role in unorthodox approaches to tackling the minimization problems of protein structures. Remarkably, a similar change has revolutionized our way of looking at the relative weights of theory and computation. More precisely, the machine learning approach has prompted a methodological twist that has entailed a radical revision of the former hierarchical paradigm envisaging the complete dependence of computation on theory. In other words, in the traditional approach, computation is devoted to the mere elaboration of the theoretical premises, whereas in the scenario presented here, computation may play a more autonomous role and even act as a substitute for theory. For example, machine learning algorithms can solve a variety of problems without building on a predefined theoretical framework. In particular, ANNs accomplish the task independently of any theoretical

description whatsoever of the folding dynamics or the native structures of proteins. What is essential to this discussion is the fact that machine learning can succeed where the theoretical approach fails operating as if they had incorporated a theory of protein folding during the learning stage.

However, the even more intriguing feature is that the computational techniques may suggest new perspectives ultimately culminating in new advancements of the preexisting theory.

The so-called "fold approach" and the FDC model are two clarifying examples of how the computational treatment of data has opened new avenues of research. In the former example, extensive use of computational methods has made possible the comparative study of homologous proteins. These investigations, in their turn, have given rise to novel theoretical perspectives that never would have emerged from the analysis of individual proteins.[139,241]

The ensuing notions such as the malleability of the folding mechanism and the detection of specific residues responsible for nucleation, dynamics, and intermediates bear directly on the theoretical aspects of folding. These findings lend support to the novel view that a continuum of folding mechanisms exists between the extreme schemes represented by the fully cooperative and fully hierarchical descriptions. Thus, the long-standing dichotomy of the NC model versus the DC model can eventually be solved. In this respect, the implications of the fold approach are in agreement with the general unified view consistent with the EN, HZ, and FDC models.

Another reason for furthering the synergy of theory and computation is that notoriously blind computational treatment of the data allows the exploration of expected and unexpected correlations between the relevant parameters but by no means guarantees the reduction of the processes being investigated to first principles or to known theoretical frameworks. Nonetheless, this may be useful for conducting a kind of preliminary analysis looking for putative causal relationships between different variables, with the caveat that a mere correlation analysis may lead to statistical fallacies because correlation is by no means equivalent to causation. This implies that the computational effort is to be substantiated by physics-based modeling that permits easier identification of causal relationships. A case in point mentioned in the previous sections is the serendipitous discovery of the CO folding rate correlation with the related follow-up of physical interpretations that clearly exemplifies how the synergy between theory and computation is highly desirable for the advancement of both.

A final comment is in order with regard to future directions of protein folding studies, in light of the explosive growth of initiatives and projects, a new generation of high-throughput technologies, and the challenges in personal genomics[242] and personalized medicine.[243,244]

Within such a context, the primary tasks to be tackled are (i) the definition of standard and unified protocols describing protein folding mechanisms and experimental setups, (ii) the design of large-scale high-throughput biophysics experiments[245] coherent with task i, which should systematically relate the folding process to controllable physical parameters, (iii) the curation of publicly available databases ensuring methodologically homogeneous criteria and standardized nomenclature, and (iv) the development of holistic approaches to the studies of protein folding according to the multilevel perspective of systems biology and proteomics.[246]

One fundamental problem still awaiting a solution is the prediction of protein stability.[245] In agreement with the dominant trend, the most promising strategies take advantage of large numbers either in experiments (via high-throughput techniques) or in statistical sequence and structure analysis. Finally, when the single protein is plunged into a metabolic network, or in the broader scenario of a whole cell or organ, protein scientists are confronted with the task of bridging the divide between the outcomes of the folding models and the multiple levels of analysis regarding the proteome, the metabolome, or the connectome. Clearly, the relevant data expected from investigation of the folding process should allow one to predict the response of a protein to physical interactions with its environment.

In this perspective, the unified picture emerging from the converging models of the folding process provides useful information. First, it clarifies the determinants of folding, discriminating the hot residues seen as the accelerator pedals of the folding mechanism from the other residues that might be essential for function or stability. This may be beneficial for predicting the effects of mutations. Second, it provides a mesoscopic description of the most significant intermediates as well as estimates of their lifetimes. Gaining knowledge of the essential properties of these states might be useful to indicate the possible branching points where different processes competing with the normal folding may start. This could lead to significant progress in the elucidation of etiology at the molecular level of a whole class of neurodegenerative diseases. Also, modeling the kinetics of the exposure to solvent of selected regions of the protein may be profitable in inferring conditions under which interactions with other proteins, complexes, or other metabolites set in. This shifts the emphasis to the still largely unexplored properties of the unfolded conformations.

These improvements are conducive to a more satisfactory description of the interactions caused by cellular crowding that make the folding *in vivo* markedly different from the folding *in vitro* and, in the last analysis, will prompt the development and benchmarking of new and more comprehensive methods for the prediction of the key features of protein folding processes under more realistic conditions. A step toward achieving this goal is a hybrid model that uses a combination of simulations, coarse grained models, and experimentally determined parameters for reproducing the effects of osmolytes and denaturants on the protein molecule.[227]

AUTHOR INFORMATION

Corresponding Authors

*(M.C.) School of Sciences and Technology, University of Camerino, Via S. Agostino 1, Camerino, MC 62032, Italy. Phone: +39 0737 402275. Fax: + 39 0737 404508. E-mail: mario.compiani@unicam.it.

*(E.C.) Division of Informatics, Department of Pathology, University of Alabama at Birmingham, 619 19th St., South, WP Building, Suite P220, Birmingham, AL 35249. E-mail: emidio@uab.edu. Phone: (205) 975-2928. Fax: (205) 934-5499.

Author Contributions

§The authors contributed equally to this work.

Funding

M.C. is supported by the Italian Ministry for University and Research. E.C. is supported by start-up funds from the

Department of Pathology at the University of Alabama at Birmingham.

Notes

The authors declare no competing financial interest.

Abbreviations

ANN, artificial neural network; CO, contact order; DC, diffusion−collision; ECO, effective contact order; EN, extended nucleus; FDC, foldon diffusion−collision; HC, hydrophobic collapse; HP, hydrophobic/polar; HZ, hydrophobic zipping; LEL, low-entropy-loss; LRO, long-range order; MCI, multiple-contact index; MD, molecular dynamics; MG, molten globule; NC, nucleation−condensation/collapse; NP, nonpolynomial; rmsd, root-mean-square deviation; SASA, solvent-accessible surface area; TSM, topomer search model; TCD, total contact distance; TS, transition state; TSE, transition state ensemble; SVM, support vector machine; ZA, zipping and assembly.

■ ACKNOWLEDGMENTS

We acknowledge Dr. Derek Jones (CNR Bologna, Italy), for the careful revision of the manuscript.

■ REFERENCES

(1) Pace, C. N. (1975) The stability of globular proteins. *CRC Crit. Rev. Biochem. 3*, 1−43.
(2) Privalov, P. L. (1979) Stability of proteins: Small globular proteins. *Adv. Protein Chem. 33*, 167−241.
(3) Pace, C. N., Trevino, S., Prabhakaran, E., and Scholtz, J. M. (2004) Protein structure, stability and solubility in water and other solvents. *Philos. Trans. R. Soc., B 359*, 1225−1234.
(4) Chien, P., Weissman, J. S., and DePace, A. H. (2004) Emerging principles of conformation-based prion inheritance. *Annu. Rev. Biochem. 73*, 617−656.
(5) Dobson, C. M. (2003) Protein folding and misfolding. *Nature 426*, 884−890.
(6) Rochet, J. C., and Lansbury, P. T., Jr. (2000) Amyloid fibrillogenesis: Themes and variations. *Curr. Opin. Struct. Biol. 10*, 60−68.
(7) Tobi, D., and Bahar, I. (2005) Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. *Proc. Natl. Acad. Sci. U.S.A. 102*, 18908−18913.
(8) Bakan, A., and Bahar, I. (2009) The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. *Proc. Natl. Acad. Sci. U.S.A. 106*, 14349−14354.
(9) Friel, C. T., Beddard, G. S., and Radford, S. E. (2004) Switching two-state to three-state kinetics in the helical protein Im9 via the optimization of stabilising non-native interactions by design. *J. Mol. Biol. 342*, 261−273.
(10) Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., and Baker, D. (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science 302*, 1364−1368.
(11) Hecht, M. H., Das, A., Go, A., Bradley, L. H., and Wei, Y. (2004) De novo proteins from designed combinatorial libraries. *Protein Sci. 13*, 1711−1723.
(12) Goodman, C. M., Choi, S., Shandler, S., and DeGrado, W. F. (2007) Foldamers as versatile frameworks for the design and evolution of function. *Nat. Chem. Biol. 3*, 252−262.
(13) Dill, K. A., and MacCallum, J. L. (2012) The protein-folding problem, 50 years on. *Science 338*, 1042−1046.
(14) Kadanoff, L. P. (2001) Turbulent heat flow: Structures and scaling. *Phys. Today 54*, 34−39.
(15) Amaral, L. A. N., and Ottino, J. M. (2004) Complex networks. Augmenting the framework for the study of complex systems. *Eur. Phys. J. B 38*, 147−162.
(16) Hasnain, S. S., and Wakatsuki, S. (2012) Frontiers and challenges of biophysical methods: From computational biology to X-ray free electron laser. *Curr. Opin. Struct. Biol. 22*, 591−593.
(17) Voigt, C. A., Martinez, C., Wang, Z. G., Mayo, S. L., and Arnold, F. H. (2002) Protein building blocks preserved by recombination. *Nat. Struct. Biol. 9*, 553−558.
(18) Dill, K. A., Ozkan, S. B., Weikl, T. R., Chodera, J. D., and Voelz, V. A. (2007) The protein folding problem: When will it be solved? *Curr. Opin. Struct. Biol. 17*, 342−346.
(19) Marx, V. (2013) Biology: The big challenges of big data. *Nature 498*, 255−260.
(20) Anfinsen, C. B. (1973) Principles that govern the folding of protein chains. *Science 181*, 223−230.
(21) Honig, B. (1999) Protein folding: From the Levinthal paradox to structure prediction. *J. Mol. Biol. 293*, 283−293.
(22) Levinthal, C. (1969) How to Fold Graciously. In *Mössbauer Spectroscopy in Biological Systems*, pp 22−24, Allerton House, Monticello, IL.
(23) Berger, B., and Leighton, T. (1998) Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. *J. Comput. Biol. 5*, 27−40.
(24) (2005) So much more to know. *Science 309*, 78−102.
(25) Bryngelson, J. D., Onuchic, J. N., Socci, N. D., and Wolynes, P. G. (1995) Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins 21*, 167−195.
(26) Dill, K. A., and Chan, H. S. (1997) From Levinthal to pathways to funnels. *Nat. Struct. Biol. 4*, 10−19.
(27) Schonbrun, J., and Dill, K. A. (2003) Fast protein folding kinetics. *Proc. Natl. Acad. Sci. U.S.A. 100*, 12678−12682.
(28) Kaya, H., and Chan, H. S. (2005) Explicit-chain model of native-state hydrogen exchange: Implications for event ordering and cooperativity in protein folding. *Proteins 58*, 31−44.
(29) Hyeon, C., and Thirumalai, D. (2011) Capturing the essence of folding and functions of biomolecules using coarse-grained models. *Nat. Commun. 2*, 487.
(30) Compiani, M. (1996) Remarks on the paradigms of connectionism. In *Connectionism, Concepts, and Folk Psychology. The Legacy of Alan Turing* (Clark, A., and Millican, P. J. R., Eds.) Clarendon Press, Oxford, U.K.
(31) Cliff, D. (2003) Biologically-Inspired Computing Approaches to Cognitive Systems: A partial tour of the literature. In Technical Report HPL-2003-11, Hewlett-Packard Laboratories, Bristol, England.
(32) Hopfield, J. J. (1994) Neurons, dynamics and computation. *Phys. Today 47*, 40−46.
(33) Dill, K. A. (1999) Polymer principles and protein folding. *Protein Sci. 8*, 1166−1180.
(34) Levantino, M., Cupane, A., Zimanyi, L., and Ormos, P. (2004) Different relaxations in myoglobin after photolysis. *Proc. Natl. Acad. Sci. U.S.A. 101*, 14402−14407.
(35) Bryngelson, J. D., and Wolynes, P. G. (1987) Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. U.S.A. 84*, 7524−7528.
(36) Haken, H. (1983) *Synergetics, an Introduction: Nonequilibrium Phase Transitions and Self-Organization in Physics, Chemistry, and Biology*, Springer-Verlag, New York.
(37) Laughlin, R. B., Pines, D., Schmalian, J., Stojkovic, B. P., and Wolynes, P. G. (2000) The middle way. *Proc. Natl. Acad. Sci. U.S.A. 97*, 32−37.
(38) Frauenfelder, H., and McMahon, B. (1998) Dynamics and function of proteins: The search for general concepts. *Proc. Natl. Acad. Sci. U.S.A. 95*, 4795−4797.
(39) Frauenfelder, H., Fenimore, P. W., Chen, G., and McMahon, B. H. (2006) Protein folding is slaved to solvent motions. *Proc. Natl. Acad. Sci. U.S.A. 103*, 15469−15472.
(40) Compiani, M., Fariselli, P., Martelli, P. L., and Casadio, R. (1998) An entropy criterion to detect minimally frustrated intermediates in native proteins. *Proc. Natl. Acad. Sci. U.S.A. 95*, 9290−9294.

(41) Maity, H., Maity, M., Krishna, M. M., Mayne, L., and Englander, S. W. (2005) Protein folding: The stepwise assembly of foldon units. *Proc. Natl. Acad. Sci. U.S.A. 102*, 4741−4746.
(42) Freund, S. M., Wong, K. B., and Fersht, A. R. (1996) Initiation sites of protein folding by NMR analysis. *Proc. Natl. Acad. Sci. U.S.A. 93*, 10600−10603.
(43) Galzitskaya, O. V., and Finkelstein, A. V. (1999) A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl. Acad. Sci. U.S.A. 96*, 11299−11304.
(44) Fuxreiter, M., Simon, I., Friedrich, P., and Tompa, P. (2004) Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J. Mol. Biol. 338*, 1015−1026.
(45) Kramers, H. A. (1940) Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica 7*, 284−304.
(46) Halle, B., and Davidovic, M. (2003) Biomolecular hydration: From water dynamics to hydrodynamics. *Proc. Natl. Acad. Sci. U.S.A. 100*, 12135−12140.
(47) Kaya, H., and Chan, H. S. (2003) Solvation effects and driving forces for protein thermodynamic and kinetic cooperativity: How adequate is native-centric topological modeling? *J. Mol. Biol. 326*, 911−931.
(48) Thirumalai, D., Liu, Z., O'Brien, E. P., and Reddy, G. (2013) Protein folding: From theory to practice. *Curr. Opin. Struct. Biol. 23*, 22−29.
(49) Caldarelli, G., and De los Rios, P. (2001) Cold and warm denaturation of proteins. *J. Biol. Phys. 27*, 229−241.
(50) Compiani, M. (1993) Escape rates in bistable systems with position-dependent friction coefficients. *J. Chem. Phys. 98*, 602−606.
(51) Ye, X., Ionascu, D., Gruia, F., Yu, A., Benabbas, A., and Champion, P. M. (2007) Temperature-dependent heme kinetics with nonexponential binding and barrier relaxation in the absence of protein conformational substates. *Proc. Natl. Acad. Sci. U.S.A. 104*, 14682−14687.
(52) Kaya, H., and Chan, H. S. (2002) Towards a consistent modeling of protein thermodynamic and kinetic cooperativity: How applicable is the transition state picture to folding and unfolding? *J. Mol. Biol. 315*, 899−909.
(53) Sanchez, I. E., and Kiefhaber, T. (2003) Evidence for sequential barriers and obligatory intermediates in apparent two-state protein folding. *J. Mol. Biol. 325*, 367−376.
(54) Yapa, K., Weaver, D. L., and Karplus, M. (1992) β-Sheet coil transitions in a simple polypeptide model. *Proteins 12*, 237−265.
(55) Karplus, M., and Weaver, D. L. (1979) Diffusion−collision model for protein folding. *Biopolymers 18*, 1421−1437.
(56) Islam, S. A., Karplus, M., and Weaver, D. L. (2002) Application of the diffusion-collision model to the folding of three-helix bundle proteins. *J. Mol. Biol. 318*, 199−215.
(57) Moglich, A., Joder, K., and Kiefhaber, T. (2006) End-to-end distance distributions and intrachain diffusion constants in unfolded polypeptide chains indicate intramolecular hydrogen bond formation. *Proc. Natl. Acad. Sci. U.S.A. 103*, 12394−12399.
(58) Schellman, J. A. (2002) Fifty years of solvent denaturation. *Biophys. Chem. 96*, 91−101.
(59) Munoz, V. (2001) What can we learn about protein folding from Ising-like models? *Curr. Opin. Struct. Biol. 11*, 212−216.
(60) Jayachandran, G., Vishal, V., Garcia, A. E., and Pande, V. S. (2007) Local structure formation in simulations of two small proteins. *J. Struct. Biol. 157*, 491−499.
(61) Adcock, S. A., and McCammon, J. A. (2006) Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem. Rev. 106*, 1589−1615.
(62) Zagrovic, B., Snow, C. D., Shirts, M. R., and Pande, V. S. (2002) Simulation of folding of a small α-helical protein in atomistic detail using worldwide-distributed computing. *J. Mol. Biol. 323*, 927−937.
(63) Shaw, D. E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R. O., Eastwood, M. P., Bank, J. A., Jumper, J. M., Salmon, J. K., Shan, Y., and Wriggers, W. (2010) Atomic-level characterization of the structural dynamics of proteins. *Science 330*, 341−346.
(64) Rizzuti, B., and Daggett, V. (2013) Using simulations to provide the framework for experimental protein folding studies. *Arch. Biochem. Biophys. 531*, 128−135.
(65) Daggett, V. (2006) Protein folding-simulation. *Chem. Rev. 106*, 1898−1916.
(66) Fariselli, P., Compiani, M., and Casadio, R. (1993) Predicting secondary structures of membrane proteins with neural networks. *Eur. Biophys. J. 22*, 41−51.
(67) Rost, B. (2001) Review: Protein secondary structure prediction continues to rise. *J. Struct. Biol. 134*, 204−218.
(68) Pavlopoulou, A., and Michalopoulos, I. (2011) State-of-the-art bioinformatics protein structure prediction tools. *Int. J. Mol. Med. 28*, 295−310.
(69) Casadio, R., Fariselli, P., Taroni, C., and Compiani, M. (1996) A predictor of transmembrane α-helix domains of proteins based on neural networks. *Eur. Biophys. J. 24*, 165−178.
(70) Dill, K. A., Bromberg, S., Yue, K., Fiebig, K. M., Yee, D. P., Thomas, P. D., and Chan, H. S. (1995) Principles of protein folding: A perspective from simple exact models. *Protein Sci. 4*, 561−602.
(71) Itzhaki, L., and Wolynes, P. G. (2008) The quest to understand protein folding. *Curr. Opin. Struct. Biol. 18*, 1−3.
(72) Baker, D. (2000) A surprising simplicity to protein folding. *Nature 405*, 39−42.
(73) Ptitsyn, O. B. (1973) [Stages in the mechanism of self-organization of protein molecules]. *Dokl. Akad. Nauk SSSR 210*, 1213−1215.
(74) Dill, K. A. (1985) Theory for the folding and stability of globular proteins. *Biochemistry 24*, 1501−1509.
(75) Itzhaki, L. S., Otzen, D. E., and Fersht, A. R. (1995) The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: Evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol. 254*, 260−288.
(76) Panchenko, A. R., Luthey-Schulten, Z., and Wolynes, P. G. (1996) Foldons, protein structural modules, and exons. *Proc. Natl. Acad. Sci. U.S.A. 93*, 2008−2013.
(77) Compiani, M., Capriotti, E., and Casadio, R. (2004) Dynamics of the minimally frustrated helices determine the hierarchical folding of small helical proteins. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys. 69*, 051905.
(78) Stizza, A., Capriotti, E., and Compiani, M. (2005) A minimal model of three-state folding dynamics of helical proteins. *J. Phys. Chem. B 109*, 4215−4226.
(79) Debe, D. A., and Goddard, W. A., III (1999) First principles prediction of protein folding rates. *J. Mol. Biol. 294*, 619−625.
(80) Weikl, T. R., and Dill, K. A. (2003) Folding rates and low-entropy-loss routes of two-state proteins. *J. Mol. Biol. 329*, 585−598.
(81) Dill, K. A., Fiebig, K. M., and Chan, H. S. (1993) Cooperativity in protein-folding kinetics. *Proc. Natl. Acad. Sci. U.S.A. 90*, 1942−1946.
(82) Ozkan, S. B., Wu, G. A., Chodera, J. D., and Dill, K. A. (2007) Protein folding by zipping and assembly. *Proc. Natl. Acad. Sci. U.S.A. 104*, 11987−11992.
(83) Haran, G. (2012) How, when and why proteins collapse: The relation to folding. *Curr. Opin. Struct. Biol. 22*, 14−20.
(84) Ziv, G., and Haran, G. (2009) Protein folding, protein collapse, and Tanford's transfer model: Lessons from single-molecule FRET. *J. Am. Chem. Soc. 131*, 2942−2947.
(85) Ziv, G., Thirumalai, D., and Haran, G. (2009) Collapse transition in proteins. *Phys. Chem. Chem. Phys. 11*, 83−93.
(86) Pereira de Araujo, A. F., Gomes, A. L., Bursztyn, A. A., and Shakhnovich, E. I. (2008) Native atomic burials, supplemented by physically motivated hydrogen bond constraints, contain sufficient information to determine the tertiary structure of small globular proteins. *Proteins 70*, 971−983.
(87) Pereira de Araujo, A. F., and Onuchic, J. N. (2009) A sequence-compatible amount of native burial information is sufficient for determining the structure of small globular proteins. *Proc. Natl. Acad. Sci. U.S.A. 106*, 19001−19004.

(88) Barbosa, M. A., and de Araujo, A. F. (2003) Relevance of structural segregation and chain compaction for the thermodynamics of folding of a hydrophobic protein model. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys. 67*, 051919.
(89) Barbosa, M. A., Garcia, L. G., and Pereira de Araujo, A. F. (2005) Entropy reduction effect imposed by hydrogen bond formation on protein folding cooperativity: Evidence from a hydrophobic minimalist model. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys. 72*, 051903.
(90) Jacob, J., Krantz, B., Dothager, R. S., Thiyagarajan, P., and Sosnick, T. R. (2004) Early collapse is not an obligate step in protein folding. *J. Mol. Biol. 338*, 369−382.
(91) Sadqi, M., Lapidus, L. J., and Munoz, V. (2003) How fast is protein hydrophobic collapse? *Proc. Natl. Acad. Sci. U.S.A. 100*, 12117−12122.
(92) Stocks, B. B., Sarkar, A., Wintrode, P. L., and Konermann, L. (2012) Early hydrophobic collapse of α_1-antitrypsin facilitates formation of a metastable state: insights from oxidative labeling and mass spectrometry. *J. Mol. Biol. 423*, 789−799.
(93) Jha, A. K., Colubri, A., Freed, K. F., and Sosnick, T. R. (2005) Statistical coil model of the unfolded state: Resolving the reconciliation problem. *Proc. Natl. Acad. Sci. U.S.A. 102*, 13099−13104.
(94) Shortle, D., and Ackerman, M. S. (2001) Persistence of native-like topology in a denatured protein in 8 M urea. *Science 293*, 487−489.
(95) Ptitsyn, O. B., and Uversky, V. N. (1994) The molten globule is a third thermodynamical state of protein molecules. *FEBS Lett. 341*, 15−18.
(96) Baldwin, R. L., and Rose, G. D. (2013) Molten globules, entropy-driven conformational change and protein folding. *Curr. Opin. Struct. Biol. 23*, 4−10.
(97) Bhattacharyya, S., and Varadarajan, R. (2013) Packing in molten globules and native states. *Curr. Opin. Struct. Biol. 23*, 11−21.
(98) Israelachvili, J., and Wennerström, H. (1996) Role of hydration and water structure in biological and colloidal interactions. *Nature 379*, 219−225.
(99) Karplus, M., and Weaver, D. L. (1976) Protein-folding dynamics. *Nature 260*, 404−406.
(100) Baldwin, R. L., and Rose, G. D. (1999) Is protein folding hierarchic? I. Local structure and peptide folding. *Trends Biochem. Sci. 24*, 26−33.
(101) Baldwin, R. L., and Rose, G. D. (1999) Is protein folding hierarchic? II. Folding intermediates and transition states. *Trends Biochem. Sci. 24*, 77−83.
(102) Myers, J. K., and Oas, T. G. (2001) Preorganized secondary structure as an important determinant of fast protein folding. *Nat. Struct. Biol. 8*, 552−558.
(103) Wetlaufer, D. B. (1973) Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl. Acad. Sci. U.S.A. 70*, 697−701.
(104) Levinthal, C. (1968) Are there pathways for protein folding? *J. Chem. Phys. 65*, 44−45.
(105) Zimm, B. H., and Bragg, J. K. (1959) Theory of the phase transition between helix and random coil in polypeptide chains. *J. Chem. Phys. 31*, 526−535.
(106) Yoda, T., Sugita, Y., and Okamoto, Y. (2007) Cooperative folding mechanism of a β-hairpin peptide studied by a multicanonical replica-exchange molecular dynamics simulation. *Proteins 66*, 846−859.
(107) Nolting, B., and Andert, K. (2000) Mechanism of protein folding. *Proteins 41*, 288−298.
(108) Paci, E., Lindorff-Larsen, K., Dobson, C. M., Karplus, M., and Vendruscolo, M. (2005) Transition state contact orders correlate with protein folding rates. *J. Mol. Biol. 352*, 495−500.
(109) Fersht, A. R. (2000) Transition-state structure as a unifying basis in protein-folding mechanisms: Contact order, chain topology, stability, and the extended nucleus mechanism. *Proc. Natl. Acad. Sci. U.S.A. 97*, 1525−1529.
(110) Stigler, J., Ziegler, F., Gieseke, A., Gebhardt, J. C., and Rief, M. (2011) The complex folding network of single calmodulin molecules. *Science 334*, 512−516.
(111) Capriotti, E., and Compiani, M. (2006) Diffusion-collision of foldons elucidates the kinetic effects of point mutations and suggests control strategies of the folding process of helical proteins. *Proteins 64*, 198−209.
(112) Debe, D. A., Carlson, M. J., and Goddard, W. A., III (1999) The topomer-sampling model of protein folding. *Proc. Natl. Acad. Sci. U.S.A. 96*, 2596−2601.
(113) Plaxco, K. W., Simons, K. T., and Baker, D. (1998) Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol. 277*, 985−994.
(114) Makarov, D. E., Keller, C. A., Plaxco, K. W., and Metiu, H. (2002) How the folding rate constant of simple, single-domain proteins depends on the number of native contacts. *Proc. Natl. Acad. Sci. U.S.A. 99*, 3535−3539.
(115) Makarov, D. E., and Plaxco, K. W. (2003) The topomer search model: A simple, quantitative theory of two-state protein folding kinetics. *Protein Sci. 12*, 17−26.
(116) Wallin, S., and Chan, H. S. (2005) A critical assessment of the topomer search model of protein folding using a continuum explicit-chain model with extensive conformational sampling. *Protein Sci. 14*, 1643−1660.
(117) Weikl, T. R., Palassini, M., and Dill, K. A. (2004) Cooperativity in two-state protein folding kinetics. *Protein Sci. 13*, 822−829.
(118) Gianni, S., Guydosh, N. R., Khan, F., Caldas, T. D., Mayor, U., White, G. W., DeMarco, M. L., Daggett, V., and Fersht, A. R. (2003) Unifying features in protein-folding mechanisms. *Proc. Natl. Acad. Sci. U.S.A. 100*, 13286−13291.
(119) White, G. W., Gianni, S., Grossmann, J. G., Jemth, P., Fersht, A. R., and Daggett, V. (2005) Simulation and experiment conspire to reveal cryptic intermediates and a slide from the nucleation-condensation to framework mechanism of folding. *J. Mol. Biol. 350*, 757−775.
(120) Gianni, S., Geierhaas, C. D., Calosci, N., Jemth, P., Vuister, G. W., Travaglini-Allocatelli, C., Vendruscolo, M., and Brunori, M. (2007) A PDZ domain recapitulates a unifying mechanism for protein folding. *Proc. Natl. Acad. Sci. U.S.A. 104*, 128−133.
(121) Baldwin, R. L. (2002) Making a network of hydrophobic clusters. *Science 295*, 1657−1658.
(122) Nishimura, C., Lietzow, M. A., Dyson, H. J., and Wright, P. E. (2005) Sequence determinants of a protein folding pathway. *J. Mol. Biol. 351*, 383−392.
(123) Brockwell, D. J., and Radford, S. E. (2007) Intermediates: Ubiquitous species on folding energy landscapes? *Curr. Opin. Struct. Biol. 17*, 30−37.
(124) Chothia, C., and Lesk, A. M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J. 5*, 823−826.
(125) Pace, C. N. (2009) Energetics of protein hydrogen bonds. *Nat. Struct. Mol. Biol. 16*, 681−682.
(126) Kumar, S., and Nussinov, R. (1999) Salt bridge stability in monomeric proteins. *J. Mol. Biol. 293*, 1241−1255.
(127) Gao, J., Bosco, D. A., Powers, E. T., and Kelly, J. W. (2009) Localized thermodynamic coupling between hydrogen bonding and microenvironment polarity substantially stabilizes proteins. *Nat. Struct. Mol. Biol. 16*, 684−690.
(128) Isom, D. G., Cannon, B. R., Castaneda, C. A., Robinson, A., and Garcia-Moreno, B. (2008) High tolerance for ionizable residues in the hydrophobic interior of proteins. *Proc. Natl. Acad. Sci. U.S.A. 105*, 17784−17788.
(129) Schutz, C. N., and Warshel, A. (2001) What are the dielectric "constants" of proteins and how to validate electrostatic models? *Proteins 44*, 400−417.
(130) Bartlett, G. J., Choudhary, A., Raines, R. T., and Woolfson, D. N. (2010) n→π* interactions in proteins. *Nat. Chem. Biol. 6*, 615−620.

(131) Worley, B., Richard, G., Harbison, G. S., and Powers, R. (2012) ^{13}C NMR reveals no evidence of n-π^* interactions in proteins. *PLoS One 7*, e42075.
(132) Chen, J., and Stites, W. E. (2001) Packing is a key selection factor in the evolution of protein hydrophobic cores. *Biochemistry 40*, 15280−15289.
(133) Baldwin, R. L. (2013) Properties of hydrophobic free energy found by gas-liquid transfer. *Proc. Natl. Acad. Sci. U.S.A. 110*, 1670−1673.
(134) Chandler, D. (2005) Interfaces and the driving force of hydrophobic assembly. *Nature 437*, 640−647.
(135) Southall, N. T., Dill, K. A., and Haymet, A. D. J. (2002) A view of the hydrophobic effect. *J. Phys. Chem. B 106*, 521−533.
(136) Vamvaca, K., Vogeli, B., Kast, P., Pervushin, K., and Hilvert, D. (2004) An enzymatic molten globule: Efficient coupling of folding and catalysis. *Proc. Natl. Acad. Sci. U.S.A. 101*, 12860−12864.
(137) Tsai, C. J., Kumar, S., Ma, B., and Nussinov, R. (1999) Folding funnels, binding funnels, and protein function. *Protein Sci. 8*, 1181−1190.
(138) Kiefhaber, T., Bachmann, A., and Jensen, K. S. (2012) Dynamics and mechanisms of coupled protein folding and binding reactions. *Curr. Opin. Struct. Biol. 22*, 21−29.
(139) Nickson, A. A., Wensley, B. G., and Clarke, J. (2013) Take home lessons from studies of related proteins. *Curr. Opin. Struct. Biol. 23*, 66−74.
(140) Avbelj, F., and Moult, J. (1995) Role of electrostatic screening in determining protein main chain conformational preferences. *Biochemistry 34*, 755−764.
(141) Greene, L. H., and Higman, V. A. (2003) Uncovering network systems within protein structures. *J. Mol. Biol. 334*, 781−791.
(142) Tsai, C. J., Maizel, J. V., Jr., and Nussinov, R. (2000) Anatomy of protein structures: Visualizing how a one-dimensional protein chain folds into a three-dimensional shape. *Proc. Natl. Acad. Sci. U.S.A. 97*, 12038−12043.
(143) Freire, E. (2001) The thermodynamic linkage between protein structure, stability, and function. *Methods Mol. Biol. 168*, 37−68.
(144) Rose, P. W., Beran, B., Bi, C., Bluhm, W. F., Dimitropoulos, D., Goodsell, D. S., Prlic, A., Quesada, M., Quinn, G. B., Westbrook, J. D., Young, J., Yukich, B., Zardecki, C., Berman, H. M., and Bourne, P. E. (2011) The RCSB Protein Data Bank: Redesigned web site and web services. *Nucleic Acids Res. 39*, D392−D401.
(145) Andreeva, A., Howorth, D., Chandonia, J. M., Brenner, S. E., Hubbard, T. J., Chothia, C., and Murzin, A. G. (2008) Data growth and its impact on the SCOP database: New developments. *Nucleic Acids Res. 36*, D419−D425.
(146) Sillitoe, I., Cuff, A. L., Dessailly, B. H., Dawson, N. L., Furnham, N., Lee, D., Lees, J. G., Lewis, T. E., Studer, R. A., Rentzsch, R., Yeats, C., Thornton, J. M., and Orengo, C. A. (2013) New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res. 41*, D490−D498.
(147) Brenner, S. E. (2001) A tour of structural genomics. *Nat. Rev. Genet. 2*, 801−809.
(148) Chandonia, J. M., and Brenner, S. E. (2006) The impact of structural genomics: Expectations and outcomes. *Science 311*, 347−351.
(149) Sadreyev, R. I., and Grishin, N. V. (2006) Exploring dynamics of protein structure determination and homology-based prediction to estimate the number of superfamilies and folds. *BMC Struct. Biol. 6*, 6.
(150) Koga, N., Tatsumi-Koga, R., Liu, G., Xiao, R., Acton, T. B., Montelione, G. T., and Baker, D. (2012) Principles for designing ideal protein structures. *Nature 491*, 222−227.
(151) Pace, C. N. (1990) Measuring and increasing protein stability. *Trends Biotechnol. 8*, 93−98.
(152) Baldwin, R. L., and Eisenberg, D. E. (1987) Protein stability. In *Protein engineering* (Oxender, D. L., and Fox, C. F., Eds.) pp 127−148, Alan R. Liss, New York.
(153) Schellman, J. A. (1994) The thermodynamics of solvent exchange. *Biopolymers 34*, 1015−1026.
(154) Myers, J. K., Pace, C. N., and Scholtz, J. M. (1995) Denaturant m values and heat capacity changes: Relation to changes in accessible surface areas of protein unfolding. *Protein Sci. 4*, 2138−2148.
(155) Grimsley, G. R., Trevino, S. R., Thurlkill, R. L., and Scholtz, J. M. (2013) Determining the conformational stability of a protein from urea and thermal unfolding curves. *Current Protocols in Protein Science*, Chapter 28, Unit 28.24, Wiley, New York.
(156) Kumar, M. D., Bava, K. A., Gromiha, M. M., Prabakaran, P., Kitajima, K., Uedaira, H., and Sarai, A. (2006) ProTherm and ProNIT: Thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res. 34*, D204−D206.
(157) Bava, K. A., Gromiha, M. M., Uedaira, H., Kitajima, K., and Sarai, A. (2004) ProTherm, version 4.0: Thermodynamic database for proteins and mutants. *Nucleic Acids Res. 32*, D120−D121.
(158) Morris, E. R., and Searle, M. S. (2012) Overview of protein folding mechanisms: Experimental and theoretical approaches to probing energy landscapes. *Current Protocols in Protein Science*, Chapter 28, Unit 28.22, pp 21−22, Wiley, New York.
(159) Jackson, S. E. (1998) How do small single-domain proteins fold? *Folding Des. 3*, R81−R91.
(160) Maxwell, K. L., Wildes, D., Zarrine-Afsar, A., De Los Rios, M. A., Brown, A. G., Friel, C. T., Hedberg, L., Horng, J. C., Bona, D., Miller, E. J., Vallee-Belisle, A., Main, E. R., Bemporad, F., Qiu, L., Teilum, K., Vu, N. D., Edwards, A. M., Ruczinski, I., Poulsen, F. M., Kragelund, B. B., Michnick, S. W., Chiti, F., Bai, Y., Hagen, S. J., Serrano, L., Oliveberg, M., Raleigh, D. P., Wittung-Stafshede, P., Radford, S. E., Jackson, S. E., Sosnick, T. R., Marqusee, S., Davidson, A. R., and Plaxco, K. W. (2005) Protein folding: Defining a "standard" set of experimental conditions and a preliminary kinetic data set of two-state proteins. *Protein Sci. 14*, 602−616.
(161) Bogatyreva, N. S., Osypov, A. A., and Ivankov, D. N. (2009) KineticDB: A database of protein folding kinetics. *Nucleic Acids Res. 37*, D342−D346.
(162) De Sancho, D., and Munoz, V. (2011) Integrated prediction of protein folding and unfolding rates from only size and structural class. *Phys. Chem. Chem. Phys. 13*, 17030−17043.
(163) Naganathan, A. N., and Munoz, V. (2010) Insights into protein folding mechanisms from large scale analysis of mutational effects. *Proc. Natl. Acad. Sci. U.S.A. 107*, 8611−8616.
(164) de los Rios, M. A., Muralidhara, B. K., Wildes, D., Sosnick, T. R., Marqusee, S., Wittung-Stafshede, P., Plaxco, K. W., and Ruczinski, I. (2006) On the precision of experimentally determined protein folding rates and ϕ-values. *Protein Sci. 15*, 553−563.
(165) Chow, M. K., Amin, A. A., Fulton, K. F., Fernando, T., Kamau, L., Batty, C., Louca, M., Ho, S., Whisstock, J. C., Bottomley, S. P., and Buckle, A. M. (2006) The REFOLD database: A tool for the optimization of protein expression and refolding. *Nucleic Acids Res. 34*, D207−D212.
(166) Fariselli, P., Rossi, I., Capriotti, E., and Casadio, R. (2007) The WWWH of remote homolog detection: The state of the art. *Briefings Bioinf. 8*, 78−87.
(167) Zhang, Y. (2008) Progress and challenges in protein structure prediction. *Curr. Opin. Struct. Biol. 18*, 342−348.
(168) Pantazes, R. J., Grisewood, M. J., and Maranas, C. D. (2011) Recent advances in computational protein design. *Curr. Opin. Struct. Biol. 21*, 467−472.
(169) Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M. Y., Pieper, U., and Sali, A. (2007) Comparative protein structure modeling using MODELLER. *Current Protocols in Protein Science*, Chapter 2, Unit 2.9, Wiley, New York.
(170) Liu, T., Tang, G. W., and Capriotti, E. (2011) Comparative Modeling: The state of the art and protein drug target structure prediction. *Comb. Chem. High Throughput Screening 14*, 532−537.
(171) Pieper, U., Webb, B. M., Barkan, D. T., Schneidman-Duhovny, D., Schlessinger, A., Braberg, H., Yang, Z., Meng, E. C., Pettersen, E. F., Huang, C. C., Datta, R. S., Sampathkumar, P., Madhusudhan, M. S., Sjolander, K., Ferrin, T. E., Burley, S. K., and Sali, A. (2011) ModBase, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res. 39*, D465−D474.

(172) Arnold, K., Kiefer, F., Kopp, J., Battey, J. N., Podvinec, M., Westbrook, J. D., Berman, H. M., Bordoli, L., and Schwede, T. (2009) The Protein Model Portal. *J. Struct. Funct. Genomics 10*, 1−8.
(173) Roy, A., Kucukural, A., and Zhang, Y. (2010) I-TASSER: A unified platform for automated protein structure and function prediction. *Nat. Protoc. 5*, 725−738.
(174) Kim, D. E., Chivian, D., and Baker, D. (2004) Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res. 32*, W526−W531.
(175) Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y. E., Fleishman, S. J., Corn, J. E., Kim, D. E., Lyskov, S., Berrondo, M., Mentzer, S., Popovic, Z., Havranek, J. J., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, J. J., Kuhlman, B., Baker, D., and Bradley, P. (2011) ROSETTA3: An object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol. 487*, 545−574.
(176) Moult, J., Fidelis, K., Kryshtafovych, A., and Tramontano, A. (2011) Critical assessment of methods of protein structure prediction (CASP): Round IX. *Proteins 79* (Suppl. 10), 1−5.
(177) Capriotti, E., and Marti-Renom, M. A. (2008) Assessment of protein structure predictions. In *Computational Structural Biology: Methods and Applications* (Schwede, T., and Peitsch, M. C., Eds.) pp 89−109, World Scientific Publishing Co., Singapore.
(178) Yang, L., Tan, C. H., Hsieh, M. J., Wang, J., Duan, Y., Cieplak, P., Caldwell, J., Kollman, P. A., and Luo, R. (2006) New-generation amber united-atom force field. *J. Phys. Chem. B 110*, 13166−13176.
(179) Brooks, B. R., Brooks, C. L., III, Mackerell, A. D., Jr., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post, C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W., York, D. M., and Karplus, M. (2009) CHARMM: The biomolecular simulation program. *J. Comput. Chem. 30*, 1545−1614.
(180) Riniker, S., Christ, C. D., Hansen, H. S., Hunenberger, P. H., Oostenbrink, C., Steiner, D., and van Gunsteren, W. F. (2011) Calculation of relative free energies for ligand-protein binding, solvation, and conformational transitions using the GROMOS software. *J. Phys. Chem. B 115*, 13570−13577.
(181) Poole, A. M., and Ranganathan, R. (2006) Knowledge-based potentials in protein design. *Curr. Opin. Struct. Biol. 16*, 508−513.
(182) Melo, F., and Feytmans, E. (1998) Assessing protein structures with a non-local atomic interaction energy. *J. Mol. Biol. 277*, 1141−1152.
(183) Yang, Y., and 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.
(184) Wiederstein, M., and Sippl, M. J. (2007) ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res. 35*, W407−W410.
(185) Daggett, V., and Fersht, A. R. (2003) Is there a unifying mechanism for protein folding? *Trends Biochem. Sci. 28*, 18−25.
(186) Gromiha, M. M., and Huang, L. T. (2011) Machine learning algorithms for predicting protein folding rates and stability of mutant proteins: Comparison with statistical methods. *Curr. Protein Pept. Sci. 12*, 490−502.
(187) Morozov, A. V., Kortemme, T., Tsemekhman, K., and Baker, D. (2004) Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proc. Natl. Acad. Sci. U.S.A. 101*, 6946−6951.
(188) Benedix, A., Becker, C. M., de Groot, B. L., Caflisch, A., and Bockmann, R. A. (2009) Predicting free energy changes using structural ensembles. *Nat. Methods 6*, 3−4.
(189) Pokala, N., and Handel, T. M. (2005) Energy functions for protein design: Adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. *J. Mol. Biol. 347*, 203−227.
(190) Cohen, M., Potapov, V., and Schreiber, G. (2009) Four distances between pairs of amino acids provide a precise description of their interaction. *PLoS Comput. Biol. 5*, e1000470.
(191) Kwasigroch, J. M., Gilis, D., Dehouck, Y., and Rooman, M. (2002) PoPMuSiC, rationally designing point mutations in protein structures. *Bioinformatics 18*, 1701−1702.
(192) Worth, C. L., Preissner, R., and Blundell, T. L. (2011) SDM: A server for predicting effects of mutations on protein stability and malfunction. *Nucleic Acids Res. 39*, W215−W222.
(193) Bordner, A. J., and Abagyan, R. A. (2004) Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations. *Proteins 57*, 400−413.
(194) Funahashi, J., Takano, K., and Yutani, K. (2001) Are the parameters of various stabilization factors estimated from mutant human lysozymes compatible with other proteins? *Protein Eng. 14*, 127−134.
(195) Guerois, R., Nielsen, J. E., and Serrano, L. (2002) Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *J. Mol. Biol. 320*, 369−387.
(196) Zhou, H., and Zhou, Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci. 11*, 2714−2726.
(197) Jayachandran, G., Vishal, V., and Pande, V. S. (2006) Using massively parallel simulation and Markovian models to study protein folding: Examining the dynamics of the villin headpiece. *J. Chem. Phys. 124*, 164902.
(198) Dehouck, Y., Grosfils, A., Folch, B., Gilis, D., Bogaerts, P., and Rooman, M. (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics 25*, 2537−2543.
(199) Wickstrom, L., Gallicchio, E., and Levy, R. M. (2012) The linear interaction energy method for the prediction of protein stability changes upon mutation. *Proteins 80*, 111−125.
(200) Zhang, Z., Wang, L., Gao, Y., Zhang, J., Zhenirovskyy, M., and Alexov, E. (2012) Predicting folding free energy changes upon single point mutations. *Bioinformatics 28*, 664−671.
(201) Khan, S., and Vihinen, M. (2010) Performance of protein stability predictors. *Hum. Mutat. 31*, 675−684.
(202) Capriotti, E., Fariselli, P., and Casadio, R. (2004) A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics 20* (Suppl. 1), i63−i68.
(203) Capriotti, E., Fariselli, P., and Casadio, R. (2005) I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res. 33*, W306−W310.
(204) Capriotti, E., Fariselli, P., Rossi, I., and Casadio, R. (2008) A three-state prediction of single point mutations on protein stability changes. *BMC Bioinf. 9* (Suppl. 2), S6.
(205) Cheng, J., Randall, A., and Baldi, P. (2006) Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins 62*, 1125−1132.
(206) Wainreb, G., Wolf, L., Ashkenazy, H., Dehouck, Y., and Ben-Tal, N. (2011) Protein stability: A single recorded mutation aids in predicting the effects of other mutations in the same amino acid site. *Bioinformatics 27*, 3286−3292.
(207) Tian, J., Wu, N., Chu, X., and Fan, Y. (2010) Predicting changes in protein thermostability brought about by single- or multi-site mutations. *BMC Bioinf. 11*, 370.
(208) Huang, L. T., Gromiha, M. M., and Ho, S. Y. (2007) iPTREE-STAB: Interpretable decision tree based method for predicting protein stability changes upon mutations. *Bioinformatics 23*, 1292−1293.
(209) Capriotti, E., Fariselli, P., Calabrese, R., and Casadio, R. (2005) Predicting protein stability changes from sequences using support vector machines. *Bioinformatics 21* (Suppl. 2), ii54−ii58.
(210) Huang, L. T., and Gromiha, M. M. (2009) Reliable prediction of protein thermostability change upon double mutation from amino acid sequence. *Bioinformatics 25*, 2181−2187.

(211) Potapov, V., Cohen, M., and Schreiber, G. (2009) Assessing computational methods for predicting protein stability upon mutation: Good on average but not in the details. *Protein Eng. Des. Sel. 22*, 553−560.
(212) Masso, M., and Vaisman, I. I. (2008) Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. *Bioinformatics 24*, 2002−2009.
(213) Ivankov, D. N., Garbuzynskiy, S. O., Alm, E., Plaxco, K. W., Baker, D., and Finkelstein, A. V. (2003) Contact order revisited: Influence of protein size on the folding rate. *Protein Sci. 12*, 2057−2062.
(214) Zhou, H., and Zhou, Y. (2002) Folding rate prediction using total contact distance. *Biophys. J. 82*, 458−463.
(215) Gromiha, M. M., and Selvaraj, S. (2001) Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: Application of long-range order to folding rate prediction. *J. Mol. Biol. 310*, 27−32.
(216) Micheletti, C. (2003) Prediction of folding rates and transition-state placement from native-state geometry. *Proteins 51*, 74−84.
(217) Gromiha, M. M. (2009) Multiple contact network is a key determinant to protein folding rates. *J. Chem. Inf. Model. 49*, 1130−1135.
(218) Ivankov, D. N., and Finkelstein, A. V. (2004) Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. *Proc. Natl. Acad. Sci. U.S.A. 101*, 8942−8944.
(219) Punta, M., and Rost, B. (2005) Protein folding rates estimated from contact predictions. *J. Mol. Biol. 348*, 507−512.
(220) Capriotti, E., and Casadio, R. (2007) K-Fold: A tool for the prediction of the protein folding kinetic order and rate. *Bioinformatics 23*, 385−386.
(221) Gromiha, M. M., Thangakani, A. M., and Selvaraj, S. (2006) FOLD-RATE: Prediction of protein folding rates from amino acid sequence. *Nucleic Acids Res. 34*, W70−W74.
(222) Huang, L. T., and Gromiha, M. M. (2010) First insight into the prediction of protein folding rate change upon point mutation. *Bioinformatics 26*, 2121−2127.
(223) Chiti, F., and Dobson, C. M. (2006) Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem. 75*, 333−366.
(224) Amaral, M. D. (2004) CFTR and chaperones: Processing and degradation. *J. Mol. Neurosci. 23*, 41−48.
(225) Lomas, D. A., and Carrell, R. W. (2002) Serpinopathies and the conformational dementias. *Nat. Rev. Genet. 3*, 759−768.
(226) Westermark, P., Benson, M. D., Buxbaum, J. N., Cohen, A. S., Frangione, B., Ikeda, S., Masters, C. L., Merlini, G., Saraiva, M. J., and Sipe, J. D. (2005) Amyloid: Toward terminology clarification. Report from the Nomenclature Committee of the International Society of Amyloidosis. *Amyloid 12*, 1−4.
(227) Thirumalai, D., and Reddy, G. (2011) Protein thermodynamics: Are native proteins metastable? *Nat. Chem. 3*, 910−911.
(228) Vendruscolo, M. (2012) Proteome folding and aggregation. *Curr. Opin. Struct. Biol. 22*, 138−143.
(229) Chiti, F., Stefani, M., Taddei, N., Ramponi, G., and Dobson, C. M. (2003) Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature 424*, 805−808.
(230) Fernandez-Escamilla, A. M., Rousseau, F., Schymkowitz, J., and Serrano, L. (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol. 22*, 1302−1306.
(231) Trovato, A., Chiti, F., Maritan, A., and Seno, F. (2006) Insight into the structure of amyloid fibrils from the analysis of globular proteins. *PLoS Comput. Biol. 2*, e170.
(232) Conchillo-Sole, O., de Groot, N. S., Aviles, F. X., Vendrell, J., Daura, X., and Ventura, S. (2007) AGGRESCAN: A server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinf. 8*, 65.
(233) Tartaglia, G. G., and Vendruscolo, M. (2008) The Zyggregator method for predicting protein aggregation propensities. *Chem. Soc. Rev. 37*, 1395−1401.
(234) Belli, M., Ramazzotti, M., and Chiti, F. (2011) Prediction of amyloid aggregation in vivo. *EMBO Rep. 12*, 657−663.
(235) (2008) Community cleverness required. *Nature 455*, 1.
(236) Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D. P., Kania, R., Schaeffer, M., St Pierre, S., Twigger, S., White, O., and Rhee, S. Y. (2008) Big data: The future of biocuration. *Nature 455*, 47−50.
(237) Boyd Rayward, W. (1999) H. G. Wells's idea of a world brain: A critical re-assessment. *J. Am. Soc. Inf. Sci. 50*, 557−579.
(238) Lok, C. (2011) Biomedical illustration: From monsters to molecules. *Nature 477*, 359−361.
(239) Frankel, F., and Reid, R. (2008) Distilling meaning from data. *Nature 455*, 30.
(240) Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popovic, Z., and Players, F. (2010) Predicting protein structures with a multiplayer online game. *Nature 466*, 756−760.
(241) Nickson, A. A., and Clarke, J. (2010) What lessons can be learned from studying the folding of homologous proteins? *Methods 52*, 38−50.
(242) Capriotti, E., Nehrt, N. L., Kann, M. G., and Bromberg, Y. (2012) Bioinformatics for personal genome interpretation. *Briefings Bioinf. 13*, 495−512.
(243) Fernald, G. H., Capriotti, E., Daneshjou, R., Karczewski, K. J., and Altman, R. B. (2011) Bioinformatics challenges for personalized medicine. *Bioinformatics 27*, 1741−1748.
(244) Lahti, J. L., Tang, G. W., Capriotti, E., Liu, T., and Altman, R. B. (2012) Bioinformatics and variability in drug response: A protein structural perspective. *J. R. Soc., Interface 9*, 1409−1437.
(245) Magliery, T. J., Lavinder, J. J., and Sullivan, B. J. (2011) Protein stability by number: High-throughput and statistical approaches to one of protein science's most difficult problems. *Curr. Opin. Chem. Biol. 15*, 443−451.
(246) Ideker, T., Galitski, T., and Hood, L. (2001) A new approach to decoding life: Systems biology. *Annu. Rev. Genomics Hum. Genet. 2*, 343−372.
(247) Fulton, K. F., Bate, M. A., Faux, N. G., Mahmood, K., Betts, C., and Buckle, A. M. (2007) Protein Folding Database (PFD 2.0): An online environment for the International Foldeomics Consortium. *Nucleic Acids Res. 35*, D304−D307.
(248) Dehouck, Y., Kwasigroch, J. M., Gilis, D., and Rooman, M. (2011) PoPMuSiC 2.1: A web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinf. 12*, 151.
(249) Lin, G. N., Wang, Z., Xu, D., and Cheng, J. (2010) SeqRate: Sequence-based protein folding type classification and rates prediction. *BMC Bioinf. 11* (Suppl. 3), S1.
(250) Bryan, A. W., Jr., Menke, M., Cowen, L. J., Lindquist, S. L., and Berger, B. (2009) BETASCAN: Probable β-amyloids identified by pairwise probabilistic analysis. *PLoS Comput. Biol. 5*, e1000333.
(251) Galzitskaya, O. V., Garbuzynskiy, S. O., and Lobanov, M. Y. (2006) Prediction of amyloidogenic and disordered regions in protein chains. *PLoS Comput. Biol. 2*, e177.
(252) Yoon, S., and Welsh, W. J. (2004) Detecting hidden sequence propensity for amyloid fibril formation. *Protein Sci. 13*, 2149−2160.
(253) Maurer-Stroh, S., Debulpaep, M., Kuemmerer, N., Lopez de la Paz, M., Martins, I. C., Reumers, J., Morris, K. L., Copland, A., Serpell, L., Serrano, L., Schymkowitz, J. W., and Rousseau, F. (2010) Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat. Methods 7*, 237−242.

NOTE ADDED IN PROOF

The reader is referred to the October Festschrift issue of the *Journal of Physical Chemistry B*, in honor of P.G. Wolynes, which was published while the present paper was in proof. It

provides a useful source of recent contributions on several theoretical topics discussed in this review.

dx.doi.org/10.1021/bi4001529 | *Biochemistry* 2013, 52, 8601−8624