Theor Chem Acc (2006) 116: 75–86
DOI 10.1007/s00214-005-0026-8

REGULAR ARTICLE

Giorgio Colombo · Cristian Micheletti

Protein folding simulations: combining coarse-grained models and all-atom molecular dynamics

Received: 29 March 2005 / Accepted: 24 June 2005 / Published online: 21 December 2005

Abstract The investigation of protein folding and its ramifications in biological contexts is at the heart of molecular biology. Theoretical and computational studies provide a steadily growing contribution to the understanding of factors driving a given polypeptide sequence into the native state. Simplified coarse-grained protein models have proven very useful to gain insights into the general thermodynamic and kinetic features of the folding process. On the other hand, all-atom simulations allow to follow, with microscopic detail, the delicate interplay of the various chemical interactions leading to the formation of the native or intermediate states. In this paper we will discuss different computational strategies employed to tackle the protein folding problem, based on the use of either coarse-grained or all-atom protein descriptions. Finally we will discuss a recent approach that allows to extend the reach of ordinary folding simulations by using a simplified description of protein structures and energy functional in conjunction with all-atom molecular dynamics.

Keywords Protein folding · Molecular dynamics · Folding mechanisms

1 Introduction

Understanding the process through which proteins fold into their native structures lies at the heart of modern molecular and cellular biology. One of the paradigms for the folding process is constituted by the Anfinsenian principle which states that the folded state of an isolated protein corresponds to the *global* minimum of the system free energy at physiological temperature [1,2]. The existence of energetic and entropic forces for steering the folding process was argued by Levinthal more than thirty years ago [3]. His conclusion stemmed from the observation that in absence of such driving forces, the native state would be reached only through a random exploration of the available configurational space. Even for peptides of a few tens of residues the time required for this exploration would be incompatible with the time scales over which their folding is seen to occur (of the order of milliseconds). Although Levinthal did not rule out the possibility that, for naturally occurring proteins, the folding might occur through specific kinetic pathways, the reversibility of the folding process provides a strong experimental support for the Anfinsenian hypothesis. Indeed, unfolding/refolding experiments directly indicate that, for several proteins, their native state can be reached under a variety of initial conditions by virtue of mere thermodynamic forces, i.e. without the aid of any special cellular machinery. Several theoretical studies have focused on the characterization of the folding free energy landscape and in the way that it is explored by proteins en route to the native state [4–11]. In this context, one can hypothesize that, in principle, the knowledge of the mere primary sequence should contain all the physico-chemical information necessary to define the native structure of the protein and possibly the molecular mechanisms leading to it [1,2,12–14].

The Anfinsenian principle provides more than a framework to rationalize the phenomenology of protein folding. In fact, the notion that the native state of a protein is encoded entirely by its primary sequence provides the foundations for the rapidly-growing number of computational approaches to protein folding [4,6,8,15–19]. While at present, folding simulations are incapable of capturing kinetic processes occurring beyond the time scale of 1 μs (appropriate for the folding of peptides of 20–30 residues), in perspective they will play a paramount role in bridging the gap between the number of known primary sequences (extracted from genome-wide experiments) and of resolved protein structures with the atomic resolution detail necessary for in depth biochemical, medical and pharmaceutical applications.

G. Colombo (✉)
Istituto di Chimica del Riconoscimento Molecolare,
CNR, via Mario Bianco, 9, 20131 Milano, Italy
Tel.: +39-02-28500031,
Fax: +39-02-28500036
E-mail: giorgio.colombo@icrm.cnr.it

C. Micheletti
International School for Advanced Studies (S.I.S.S.A.) and INFM,
Via Beirut 2-4, 34014 Trieste, Italy

The scope of theoretical/computational approaches to protein folding also encompasses the investigations of the physico-chemical mechanisms which, by impairing the formation of the native state (mis-folding) lead to the onset of diseases. Molecular illnesses such as Alzheimer's and Parkinson's disease, Kreutzfeld–Jacob syndrome, bovine spongiform encephalopathy (BSE, or mad cow disease) share a common cause represented by the aggregation of mis-folded proteins or peptides [20]. Understanding the molecular basis of folding and mis-folding thus acquires considerable importance in the design of drugs for the treatment of these illnesses.

From a computational point of view, the folding problem has been approached from several perspectives and with different methodologies. In the context of "structure prediction", a commonly employed philosophy is to adopt a series of knowledge-based constraints and scoring functions to single out the putative native configuration among a pool of structures, possibly constructed from a representative repertoire of secondary and tertiary structures. These approaches identify the configurations of favorable free energy by importance sampling techniques that ensures that stochastically generated structures are drawn from the canonical distribution. In general, however, their temporal succession will not correspond to a legitimate kinetic trajectory. A realistic dynamical folding perspective is instead arguably captured in folding simulations where the dynamical evolution of a protein prepared in an unfolded state (typically a linear configuration) is followed by integrating the classical equations of motion for each atom in the protein and surrounding solvent. The large number of degrees of freedom of such systems limits the time-span that can be simulated in an ordinary molecular dynamics set-up. For this reason, a useful complementary approach is based on the use of simplified protein representations where each amino acid is represented by one or more centroids which interact through effective contact potentials. These simplified models, while lacking the chemical detail of all-atom representations, are amenable to extensive characterizations of both the folding kinetics and thermodynamics. It is apparent that these strategies have different advantages and shortcomings and thus, with varying demands on computer time, offer complementary insights on the folding process.

Being aware that it is impossible to present a comprehensive review of the various advancements made in this field, we have attempted to illustrate a selection of the ideas and methods that have turned protein folding simulations into a powerful tool for elucidating the pathways leading to the native state and for structure prediction. We will first start from methods employing simplified representations of polypeptide chains and then move to all-atom models of the protein in either implicit or explicit solvents. Finally, we will present our results on a combined use of coarse-grained Monte-Carlo (MC) searches and all-atom molecular dynamics (MD) which, through the appropriate selection of the MD starting configuration, allows to overcome some of the difficulties which hamper the scope and effectiveness of ordinary all-atom folding simulations.

2 Simplified protein models

Numerical simulations constitute a "virtual laboratory" where the influence on protein folding by the various physico-chemical determinants (hydrogen-bonding, hydrophobic effect and side-chain packing) can be dissected and characterized. Those aspects related to thermodynamic equilibrium can be captured only through an extensive canonical sampling of the accessible conformations. All-atom approaches are unsuitable for this task since, at present, they allow a satisfactory exploration of the free energy surface only of small peptides [21–24].

One natural route to make the task numerically feasible is to limit the phase space by reducing the conformational degrees of freedom [25–28]. Typically this is accomplished by replacing a given amino acid with one of more interaction centers. In some limiting cases several residues have been lumped together in a single interaction unit or, at the other extreme, the atoms of the main chain have been treated explicitly, while each side chain was represented with a single centroid. Finally, examples have appeared which use all-atom representation of the polypeptide and a simplified or biased force-field coupled to a specific conformational search scheme. It is not uncommon that the structural coarse-graining is accompanied by a discretization of the configurational space. This may be accomplished either by limiting the dihedral angles of the simplified backbone to take on a small number of possible values [29] or by forcing the centroids to lie on the nodes of a lattice.

The use of simplified protein models in folding simulations has a relatively long tradition. One of the first notable examples is the seminal work of Warshel and Levitt who proposed, three decades ago, to consider the interaction among groups of atoms rather than detailed atom–atom interactions and torsion variation to drive conformational transitions. Their method was applied to simulate the folding of a small pancreatic trypsin inhibitor [30] and then extended to study the mainly α-helical protein carp myogen [31]. Since this first pioneering application, many others have followed taking also advantage of the rapidly-growing availability of computing power.

In the following subsections we will survey the main applications of lattice and off-lattice models of proteins and their implications for the study of polypeptide systems.

2.1 Discrete and continuous protein models

Arguably the most schematic description of a polypeptide chain is obtained by representing it as a self-avoiding walk on a cubic lattice: each amino acid is put in correspondence with a lattice site (consecutive amino acids occupying adjacent nodes of the lattice). Prior to their use in protein-related contexts, the characterization of self-avoiding walks on hypercubic lattices was extensively used to elucidate the physics of homopolymers in connection with the theory of critical phenomena [32,33]. Owing to the chemical equivalence of their constituent monomers, homopolymers are characterized by a huge degeneracy of the states having minimal energy. On the other hand, one of the distinctive features of naturally-occurring proteins is the uniqueness of the native state conformation corresponding to the (free) energy minimum. To reproduce this feature it is clearly necessary to go beyond the homopolymer model by introducing at least two types of amino acids and suitable parameters for their effective interaction. In the HP model of Chan and Dill [34], the two classes of amino acids are taken to correspond to polar and hydrophobic residues. To mimic the segregation of hydrophobic residues into a compact core avoiding interactions with the solvent, an attractive contact energy was introduced among hydrophobic residues. Despite its simplicity the HP model has proved invaluable in clarifying several issues associated to the "foldability" and designability of proteins.

In particular, concerning the issue of designability, both numerical and theoretical studies have shown that only a small fraction of the possible HP sequences admit a unique conformation of minimal energy. In turn, only a small number of the possible lattice conformations can be the unique ground state of some HP sequence. It therefore appears that, even in such simplified models, the introduction of selection criteria such as the non-degeneracy of the native state, introduces a drastic limitation on the number of viable sequences and structures. The systematic investigation of the set of uniquely-encodable structures through analytical and enumeration techniques has provided further insight into protein designability. In fact, a fraction of the viable structures are the ground state of a large number of distinct protein sequences, consistent with the many-to-one correspondence of primary sequences and structural families found in naturally-occurring proteins. Interestingly, the set of highly-designable structures is robust with respect to the changes of the interaction potentials among the HP classes and even to the number of classes used [35,36]. Indeed, the degree of designability of a model protein structure has often been associated to its geometrical properties, such as its symmetry or the number of turns and strands [35,37,38]. This has allowed a transparent investigation of the selection criteria that through natural evolution may have led to the surprisingly small number of distinct protein folds observed in nature.

The investigation of structural designability within the HP or related models typically relies on the characterization of the equilibrium thermodynamics [39–46]. This minimal models have, however, been also used in connection with the folding dynamics. The kinetic accessibility of both highly- and poorly-designable structures has been the object of several studies which aimed at clarifying not only how the ruggedness of the free energy landscape affects the folding kinetics but also the existence of a limited number of possible routes leading to the native state [8,10,47–56].

Lattice models of proteins have therefore been very useful in elucidating some general fundamental questions about proteins and heteropolymers [57–59]. Their computational simplicity has stimulated their applicability in more realistic and challenging contexts, such as the simulation of the folding process for naturally-occurring proteins. These approaches are typically confronted with two issues. First, the need to adopt a structural simplification yielding a satisfactory compromise between having a limited number of degrees of freedom and yet a faithful conformational representation. Secondly, the difficulty and ambiguity of choosing a suitable and transferable energy function for model. Several efforts have been made for identifying the best physico-chemical criteria for extracting reliable effective interaction potentials among the amino acids [60–72]. While the potentials extraction problem is sometimes approached without reference to any particular model, it is apparent that the determination of the effective potentials will depend both on the form of the employed energy function as well as the chosen structural representation for the model proteins. In such realistic cases, the cubic lattice representation mentioned before is rarely employed since it limits in a drastic and coarse manner the angles formed by consecutive virtual backbone bonds. A variety of discrete structural models have been introduced that allow to capture more faithfully the actual distribution of dihedral angles found in protein structures. In some instances, the coarse-graining of the configurational space is achieved through the discretization of the internal degrees of freedom of the model proteins (e.g. in the 4-state model) [29]. In other contexts, the amino acid centroids are constrained to occupy the modes of lattices with suitable unit spacings and geometry. Not infrequently, more than one centroid is used for the representation of amino acids. In these cases, the discretization of the degrees of freedom is typically applied only to one type of centroid, i.e. the C_α ones, while the positions of the remaining ones, e.g. the side chains, can take on continuous values (in particular they may be reconstructed through deterministic procedures). A good example of this approach is provided by the CABS (C_α, C_β, Side group) model [73] which adopts three interaction centers per amino acid: one for the side chain center of mass, one for the C_β and one for the C_α centroid. The C_α values are constrained to lie on a grid with lattice constant as fine as 0.61 Å. Though discrete lattices always introduce the computational advantage of enumerating a priori and storing all the possible configurations of segments up to a given number of amino acids, it is obvious that as the lattice spacing decreases, the distinction between discrete and off-lattice models is blurred. In particular, Kolinski et al. [74] have developed a finely-discretised lattice model that has been used with considerable success for homology modeling as well as for structure prediction.

Simplified models employing a continuous representation of the coordinates of internal degrees of freedom of a protein typically exploit the same ideas outlined above. In particular, one or more centroids can be used for representing amino acids, and a suitable interaction potential among them must be introduced. A well-known example is provided by the united residues (UNRES) model of Scheraga and Liwo where an amino acid is represented by two centers: one for the main chain and the other for the side chain [75,76]. The energy function governing the interaction of the centroids is based on a many-body expansion and is extensively used as a scoring function in structure prediction contexts. Within this scheme the simplified UNRES representation allows a good exploration of the conformational space and has proved effective in the ab initio prediction of proteins of up to 75 residues [75].

It is interesting to mention the introduction of models that combine the description of some parts of the protein at the atomic resolution level while others are captured with centroids. This concept is illustrated by the Large-C_β model of Irback and coworkers which retain a heavy-atoms description of the protein backbone while a C_β effective centroid is used for the side chain [77]. All the peptide bond lengths, angles and peptidic torsion angles are held fixed, leaving as the only two variables the ϕ and ψ – Ramachandran torsion angles. The interaction potential function is composed of four terms keeping into account local interactions, excluded volume effects, hydrogen bonding and hydrophobic interactions. Also this type of model is amenable to very efficient stochastic MC exploration of the phase space and appears to reproduce satisfactorily the main thermodynamic features of folding processes [78,79].

3 All-atom models of proteins

Though simplified models can satisfactorily elucidate several questions concerning protein folding, they are obviously incapable of capturing the rich variety of physical and chemical behavior of proteins. Within a classical perspective, the appropriate tool to capture the finer dynamical and thermodynamical aspects is constituted by simulations based on all-atom potentials. Though the time scale addressable by this method is limited by its large computational cost, it has proved useful in several important contexts. Examples include the detailed characterization of complete pathways, of ensembles of structures in the unfolded and/or native states, and, most importantly, the rational design of small molecules as possible drugs able to interact with partially folded or misfolded conformations or the design of specific mutants with particular (tailored) properties.

Folding simulations based on both MD and MC approaches adopting all-atom force fields are being routinely used. MD simulations are particularly appealing in this context, as they represent the only computational method that can provide a time-dependent analysis of a system in molecular biology and, consequently, can be used to gain a complete description of the folding mechanism of a protein. The study of complete pathways by these means is however still computationally very demanding and intensive for proteins of more than 50–60 residues. In any case, folding simulations have the potential to yield the native structure, the folding pathways and the structures of the intermediates and transition states. From these data, important kinetic and thermodynamic information can be calculated. In contrast, most structure prediction methods aim "only" at finding the native structure, sometimes accompanied by a free energy estimate of this state.

3.1 Extended MD simulations

The first major effort to simulate from first principles – the folding process of a reasonably-sized protein – was undertaken by Duan and Kollman [16]. In their seminal study of the Villin Headpiece subdomain (36 residues) they started from a completely extended conformation and followed the dynamical evolution of the system for about 1 μs using classical molecular dynamics on parallel computers. The simulated time-span was about two orders of magnitude longer than the longest simulation reported at that time and still among the longest MD simulations on real systems. The inspection of the dynamical trajectory revealed a very rich behavior. In particular, it was observed that the extended conformation underwent a rapid hydrophobic collapse accompanied by helix formation. Notably, a marginally stable state was detected, with a lifetime of 150 ns, and having a favorable solvation free energy and significant resemblance to the native state structure. During the whole trajectory the protein tended to populate mainly compact states, as revealed by the near-native values of the radius of gyration. The main-chain root mean square deviation of all residues from the native state varied between 0.45 and 1.2 nm while that of the core (residues 9–32) fluctuated between 0.3 and 0.88 nm. Up to 80% of the native helical content and up to 62% of the native contacts were observed. Interestingly, the solvation component of the free energy (SFE) also reached levels comparable to those of the native structure. The folding process appeared to begin with a burst phase, characterized by a steady rise in native helical content and in native contacts and the decrease of the SFE, which lasted from the beginning of the simulation to about 60 ns. The analysis of solvation energy terms and the solvent accessible area indicated that the initial phase was driven by the burial of hydrophobic surface. Therefore, the initial phase closely represents the so called hydrophobic collapse occurring on the same time scale as formation of secondary structure.

The fact that the dynamical evolution of the initially-extended Villin headpiece was not seen to proceed through a steady build-up of native structure emphasizes the long time scales required to capture the salient aspects of the folding process. The computational requirements would be even heavier if a reliable thermodynamic characterization is sought. If this were to be accomplished with a single trajectory, it would be necessary to consider a time-span long enough to record several unfolding/refolding events [80]. By these means one would achieve a reliable sampling of the accessible phase space and thus address issues that are important from both the conceptual and practical point of view. One of these important aspects concerns the spatial characterization of conformers representing the unfolded ensemble.

In this respect, it is interesting to mention a series of studies undertaken by Daura et al. who focused on simple short peptides capable to form secondary structures in isolation [81–84]. The small size of the systems allowed a reliable characterization of the properties in thermodynamic equilibrium which, in turn compared well with NMR-based measurements. The typical peptides used for the studies consisted of peptides of 3–10 residues designed to fold into either helical or β-hairpin conformations (see [80] and references therein). The results of MD simulations of each short peptide at different temperatures and in various solvents showed that the peptides could actually fold and unfold multiple times, also starting from completely extended conformations. The use of multiple trajectories with different temperatures, and the application of a suitable conformational clustering technique allowed to identify the family of most populated native-like conformations and to identify the transitions leading to this cluster from other non-native families. Thanks to the observation of multiple folding–unfolding events, the free energy differences between the folded and unfolded ensembles could be computed and the melting temperature for the peptides in the force field could be estimated. The validation of these MD simulations was performed by comparing calculated interatomic distances with experimental NOE restraints. This approach proved that even at 340 K, in highly denaturing conditions, most of the NOE restraints were not violated in the simulations, a fact that demonstrates on the one hand the high conformational variability of short peptide segments, and, on the other hand, that the denatured states for those peptides is significantly less heterogeneous than previously hypothesized. For example, for peptides with about 20 rotatable bonds, the denatured state can be represented by only 100 conformers.

These conclusions are believed to apply also to the case of proteins consisting of tens of residues, consistently with recent experimental findings[85]. The existence of a limited number of unfolded conformers may have profound implications for the folding process since it may imply that the native state can be reached only through a limited number of routes[80]. From a computational point of view it would be most efficient to start the simulation not from the out-of-equilibrium extended state but rather in correspondence of one of the unfolded conformers populated in thermal equilibrium. This is consistent with that observed in recent MD simulations on β-hairpin structures, such as the B1 hairpin of protein G and Betanova. The analysis of the simulated folding trajectories revealed that the turn sequence was fundamental in determining the initial collapse of the strands from a completely extended conformation towards a structure with a high native-like character [86,87]. The establishment of the correct juxtaposition of hydrophobic residues and in-register hydrogen bonds between facing strands appeared to follow the fast hydrophobic collapse leading towards the native β-hairpin.

3.2 Distributed computing

We conclude this section by mentioning a novel approach that has been adopted to characterize some overall features of the folding/unfolding processes. Instead of using a small number of long MD trajectories, Pande and coworkers [88, 89] have adopted the opposite perspective of dealing with a very large number of weakly-coupled short simulations. The approach lent itself very naturally to a distributed computing implementation which was indeed successfully used on thousands of computers distributed worldwide. At present this approach has been used to estimate the folding rate of several fast-folding proteins, including α-helices, a β-hairpin, and a three-helix bundle protein from the Villin headpiece [88,89].

Though the excessively short duration of the individual simulated trajectories may introduce some difficulties in the effective use of the distributed computing approach [90], the latter has been used to characterize putative unfolded ensembles and their role in steering a correct rapid folding. Using a supercluster of over 10,000 processors, almost 800 μs of MD simulation were performed with atomistic detail of the folded and unfolded states of three polypeptides from a range of structural classes. A comparison between the folded and the unfolded ensembles revealed that, even though virtually none of the individual members of the unfolded ensemble was found to exhibit native-like features, the mean unfolded structure (averaged over the entire unfolded ensemble) had a native-like geometry [91]. This finding is consistent with the above-mentioned investigations of Daura et al. [81–84] and Shortle [85]. The latter group, in fact, observed through NMR measurements of residual dipolar coupling that a native-like spatial positioning and orientation of chain segments persisted to concentrations of at least 8 M urea [85]. These data were used to demonstrate that long-range ordering can occur well before a folding protein attains a compact conformation.

4 Combining coarse-grained and all-atom methods

Several experimental and theoretical observations on the conformational properties of peptides and proteins suggest that, despite a seemingly random organization [92], the unfolded ensemble contains structures with a residual native organization [85,93,94]. It would be physically appealing to use these structures as starting points for the MD folding simulations. This would also affect the computational efficiency of the simulation. In fact, when starting from a fully-extended configuration a significant time is spent in relaxing the system from this out-of-equilibrium situation. The resulting slow-down is particularly severe for simulations where the solvent is treated explicitly, since a very large number of solvent molecules needs to be considered due to the large size of the simulation box. The computational advantage over extended initial configurations would be even more conspicuous if one could start from structures picked from the transition state ensemble. In this case all-atom dynamics would progress towards the native ensemble in a time scale much shorter than the typical protein folding time. The difficulty in pursuing this strategy lies in the determination of the transition state ensemble, which ought to be done in an unbiased way, i.e. with the sole input of the primary sequence. In the following we shall discuss how the accomplishment of this

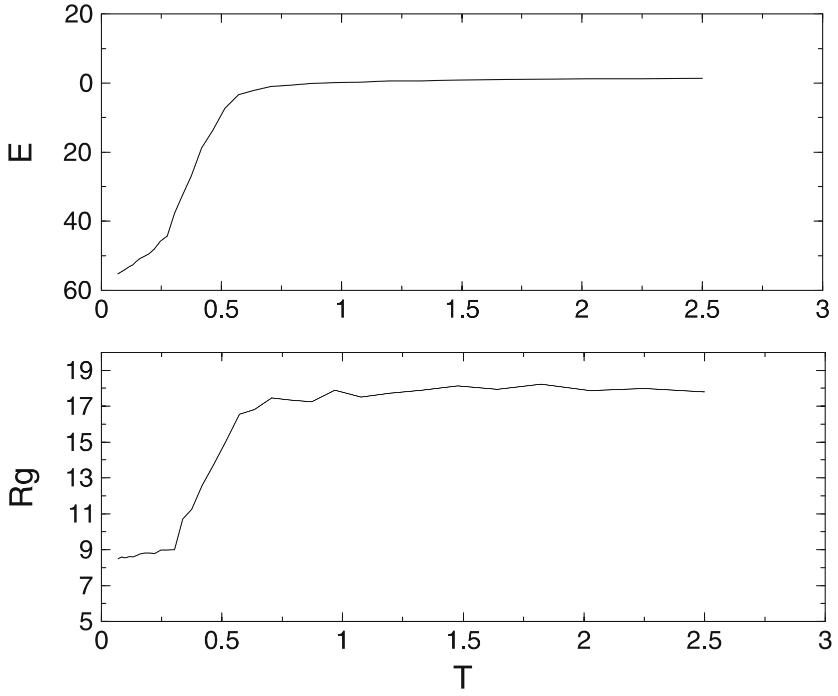


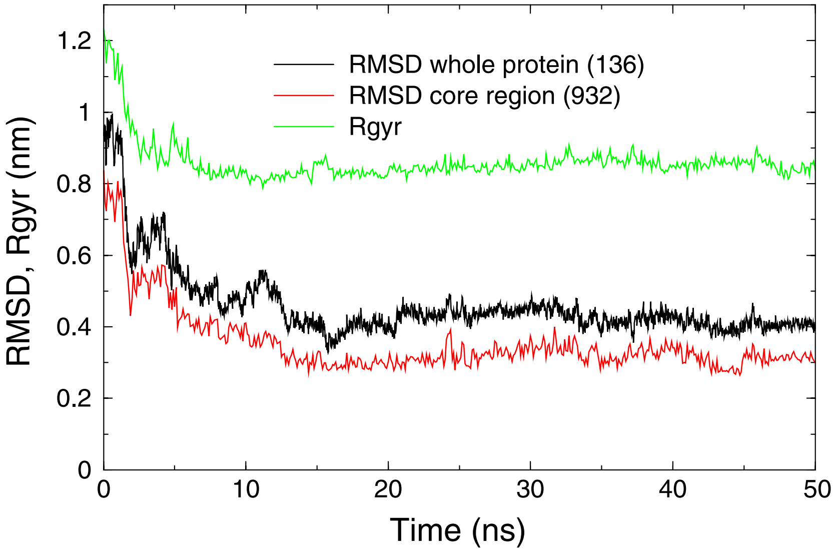
Fig. 1 Average energy and radius of gyration as a function of temperature for the coarse-grained model evolution of the 36 residue Villin headpiece domain. The "cooperative" like transition corresponding to the formation of secondary structures is evident in the two graphs [95]

difficult task may be attempted with the aid of simplified protein models.

The coarse-grained models reviewed in the previous sections have been shown to allow a vast and efficient exploration of the conformational space, vastly larger than in all-atom MD. However, the limitations of both the structural representations and of the energy functions prevent reproducing the finer features of the folding process. At the opposite extreme of resolution, all-atom simulations, based e.g. on MD, are unable to cross high energy barriers and allow an exhaustive exploration of the phase space. However they can bring into the system much of the atomic detail necessary for the high resolution structural and dynamical studies of primary interest to the biochemical community.

These observations led us to suggest a new approach which combines a coarse-grained MC search of viable starting configurations for a subsequent all–atom MD simulations in explicit solvent [95,96]. The effect of the coarse-grained part is to simplify the energy landscape of the protein to identify efficiently physically meaningful starting conformations for the subsequent MD. Explicit solvent MD is then used to reintroduce the fine chemical details which are ultimately responsible for driving the evolution towards the native state. The link between the two structural representations is a fine-graining algorithm which allows to reconstruct reliably the full atomic detail of the protein using a library of previously generated protein fragments. The details of the reconstruction procedure are provided in [95,96].

The test system chosen for this study was the same as considered by Duan and Kollman and other investigators [16, 89,97–99], namely the Villin headpiece, HP36 (1VII.pdb). The preliminary coarse-grained MC exploration of the free energy landscape is achieved by describing the protein in terms of its C_α trace and of effective C_β centroids, building upon the lessons learned from the papers cited in paragraph 2. This is accompanied by a simplification of the energy function which incorporates effective pairwise interactions among amino acids (KGS potentials) [63], knowledge-based constraints for backbone chirality, local propensities to form secondary motifs, and a term favoring their tertiary packing. [95,96]. The relative weight of the potential energy terms was chosen so that, on a set of short proteins (unrelated to the Villin headpiece), the build up of secondary structure occurred at the collapse temperature, Tc. Within this simplified framework, the thermodynamics of HP36 was characterized by several MC evolutions at distinct temperatures. The MC dynamics entailed the use of pivot and crankshaft moves which preserve the length of the bonds, initially set to 3.8 Å, joining consecutive C_α centroids. As temperature is decreased, the model protein undergoes a collapse, as signaled by the rapid decrease of both the radius of gyration, Rg, and the average system energy. The peak in the specific heat in correspondence of the collapse temperature, Tc, is further associated with significant fluctuations in energy reflecting the coexistence of rather swollen and globular conformations (see Fig. 1). Consistently with the relative strength of the potential terms in the model Hamiltonian, the latter ones typically possess local secondary elements and are further compactified at lower temperatures. The protein at Tc is thus poised to collapse into compact conformations with



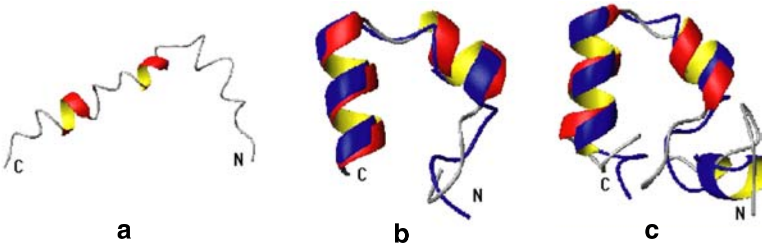


Fig. 2 *Top panel* time evolution during simulation F1 of the radius of gyration (*green*), of the RMSD over the HP36 whole protein (*black*) and over the HP36 core region (*red*). *Lower panel* **a** the starting structure of simulation F1; best native structural alignment over the **b** core region and **c** entire protein of the representative structure of trajectory F1. The representative is colored in red while the native reference NMR conformation in *blue* [95]

non-trivial secondary content. Thus, the structures encountered in the MC trajectory at Tc represent attractive candidates for all-atom MD evolution for several reasons: (1) secondary elements are typically formed, (2) the structures are not unnaturally compact and (3) the conformational variability is such that significantly different structures can be picked (the average RMSD between any pair of structures sampled at Tc being 5.9±1.2 Å). If the free energy landscape of the coarse-grained model at T_c retained the relevant features of the "true" one at the folding temperature, one would expect that representatives of the transition ensemble would be found among the sampled model structures. The dynamical evolution of these configurations should, on average, progress towards the native basin much more rapidly than starting from unfolded (or extended) conformations. Simplified structural models are manifestly unable to capture the delicate chemical interplay which sculpts the free energy landscape governing a folding process (for otherwise they would be routinely used for determining the native state of any given sequence!). However, it has been argued that the bottleneck of the folding process is constituted by the formation of a few crucial contacts which, by establishing the correct overall native topology, predispose the harmonious formation of the remaining native interactions [17, 18]. The generation of structures capturing the main traits of native topology, is a much simpler task than predicting the native state and is within reach of simplified protein models [74]. This fact provided the motivation for the present attempt to extend the scope of ordinary MD simulations by using simple physicochemical criteria to identify the starting conformations.

Seven different uncorrelated coarse-grained conformations were thus chosen at Tc. Each of them was used, after the fine-graining, as an input for a subsequent all-atom MD evolution of 50 ns at 300 K and in explicit water. The general Gromos96 force field, with the SPC water model and PME treatment of electrostatics were used. For the sake of brevity we shall discuss the main features that emerged out of the seven simulations, (indicated as F1, F2, ..., F7); a more detailed account can be found in [95]. Trajectory F1 starts from a rather open configuration, the gyration radius being $R_g \approx 12$ Å, and with partially formed helices in regions H2 and H3 (cf. Fig. 2). After a short equilibration time, it undergoes a rapid compaction as shown by the fast decrease in R_g values. Not only is this collapse rapid (10 ns), but the initial helical segments grow to the full native extension of H2 and H3 while also achieving their correct tertiary packing. The core RMSD with respect to the minimized average NMR structure often attains values as low as 2.8 Å and stabilizes around a mean value of 3.0 Å (cf. Fig. 2). This result constitutes a significant advancement over the pioneering study of Duan and Kollman where the representative structure of their 1 μs-long simulation was found to be at 4 and 5.7 Å

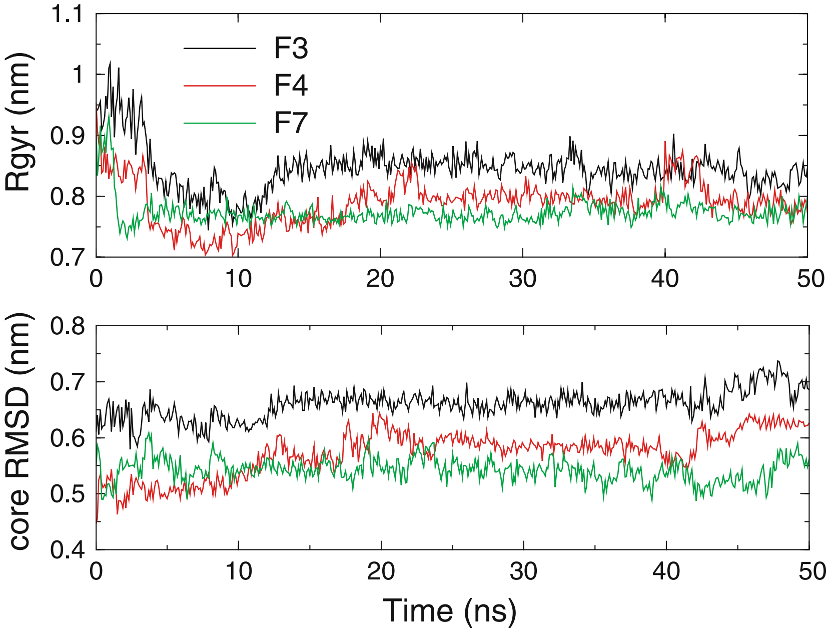


Fig. 3 Time evolution of radius of gyration (*top*) and core-RMSD (*bottom*) of HP36 in runs F3, F4 and F7. The RMSD was calculated against the average minimized NMR structure of HP36

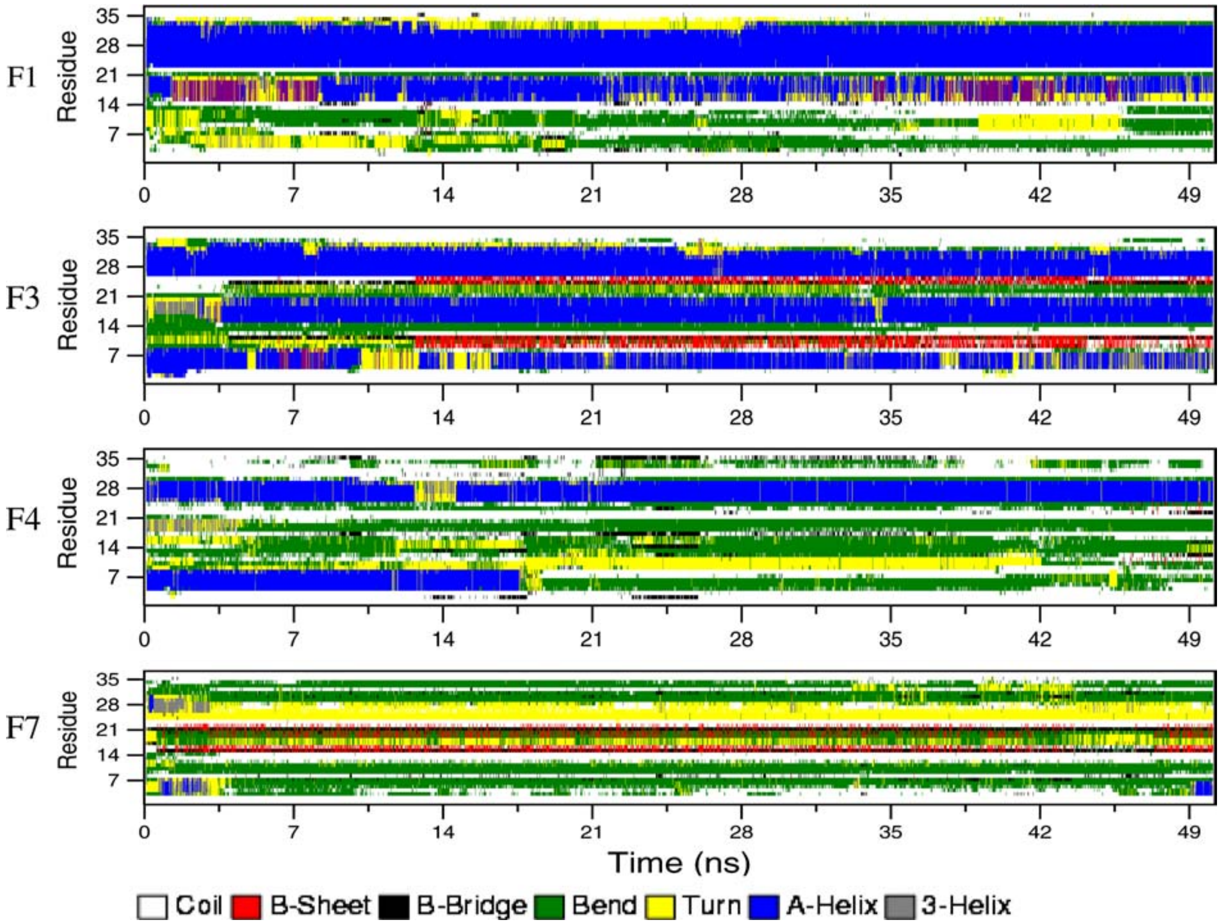


Fig. 4 Secondary structure time evolution for simulations F1, F3, F4 and F7. The DSSP criterion is used to define secondary structure motifs

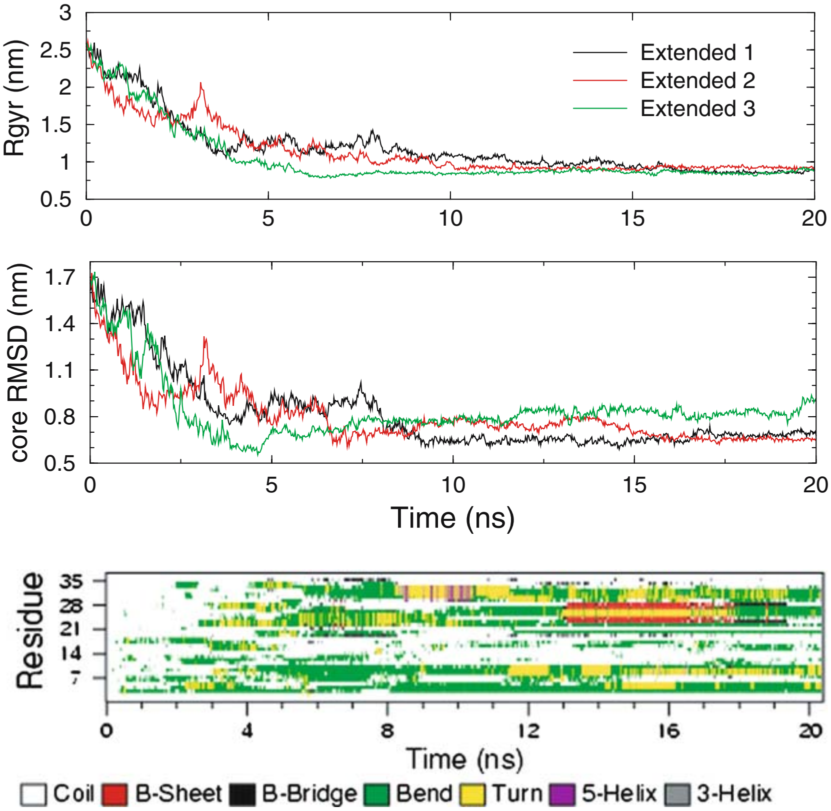


Fig. 5 Time evolution of radius of gyration (*top*) and core-RMSD (*middle*) of HP36 in three different runs starting from the fully-extended state. *Bottom* typical time evolution of the secondary content during simulations starting from the extended conformation

RMSD from the NMR reference for the core and the whole protein respectively. The advantage of our approach, which overall involved a fraction of the time-span simulated is that valuable insight about the folding dynamics can be obtained by the analysis and comparison of the evolution of the various starting structures. The significant secondary content and organization found in all starting structures resulted in interesting dynamical evolutions that, even when not progressing towards the folded state, convey valuable information on the folding process, as e.g. the trapping mechanism associated with the formation of contacting strands. In one remarkable case, run F3, analysis of the secondary structure content reveals that helices H1, H2 and H3 are correctly formed (see Figs. 3, 4). However, within the simulated time span the trajectory does not approach the native conformation since the RMSD from the native state stabilizes around 6 Å. The secondary elements, while formed in the correct native helical regions, assemble in a non-native geometry mainly due to the formation of a stable hydrophobic core involving Phe11, Leu21, Trp24 and Leu29. Run F4 also presents an interesting behavior since the starting structure possesses an acceptable native similarity (the core RMSD being about 4.5 Å) and a good helical content in regions H1 and H3 (see Fig. 4). This initially promising similarity is gradually eroded in a few nanoseconds of dynamical evolution, eventually leading to negligible native content. Interestingly, the loss of native content is paralleled by the formation of a turn conformation involving residues 8–10 and the pairing of β-like structures involving residues 2–7 and 10–15. The tendency to form β-sheets is also observed in the other four trajectories, where no native progress is recorded (the average RMSD being 6 Å) and that persistently display contacting extended segments organized in an overall compact structure. Clearly these insights into the events that impair the progress towards the native state may give important clues about the presence of intermediates that may twarth trajectories away from the native basin. It is important to point out that among the 7 runs, the one having the lowest internal energy is the one reaching the native basin, F1, $E_{\text{int}} = -1109$ KJ/mol. All other runs have energies in the range $[-1078, -1021]$ KJ/mol, while the native state has energy -1094 KJ/mol. Besides the internal energy, other free energy estimators can be used to discriminate the trajectory approaching the native basin from the others [100].

As a term of comparison, we have carried out three MD runs starting from extended conformations. Due to the much larger number of water molecules present in the simulation cells that accommodate the starting configurations, the total simulation time needed to evolve the three structures over 20 ns was about equal to the one used for runs F1–F7. The behavior of these runs is summarized in Fig. 5. As seen from the graphs, between 5 and 10 ns a chain collapse occurs. The resulting structures, despite possessing native-like values of R_g have very poor secondary content (and hence native simi-

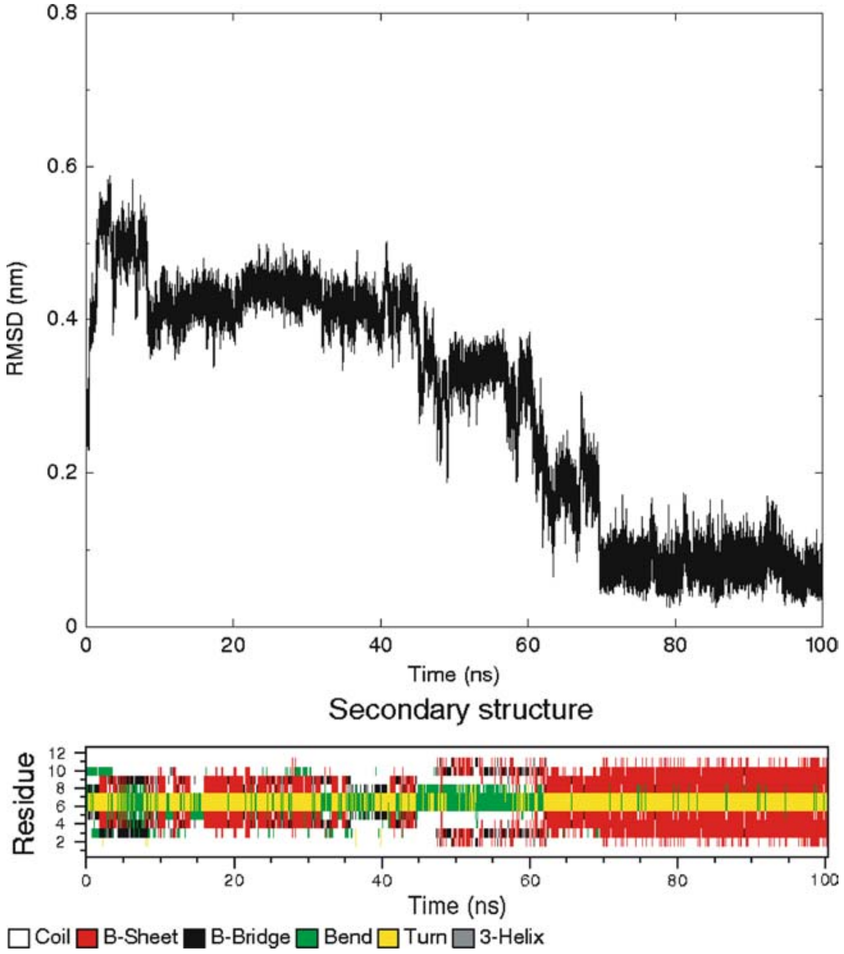


Fig. 6 Time evolution of RMSD (*top*) from the average NMR structure and of the secondary structure content (*Bottom*) for the Trp-Zipper β-hairpin

larity too). The further evolution of these compact and disordered structures is very slow at 300 K. These features contrast with the build up of secondary and tertiary structure observed with the same computational investment, among the trajectories F1–F7. This comparison provides a clear illustration of the advantages of the hybrid MC–MD approach. The fact that the energy-function of the coarse-grained model is simple and not tuned for the Villin headpiece testifies to the general feasibility of employing simplified models to identify viable starting configurations for the MD evolution. Though it is tempting to speculate that some of the structures sampled at T_c are related to the transition state ensemble, our limited number of runs does not allow to establish this with the desired statistical confidence. We have however, applied the same hybrid MC–MD methodology to study β-hairpins and all-β-proteins with promising results: RMSD of 0.6 Å of the best predicted with respect to the native structure for a 13 residue long β-hairpin forming peptide as shown in Fig. 6 (Trp-zipper, 1leo). In this case, thanks to the limited dimensions of the hairpin, simulations could be run up to 100 ns starting from seven MC structures selected at the Tc. Starting all-atom MD from completely extended conformations yields a minimum RMSD value of 3 Å, but with a computational expense which is about three times bigger than in the case of the mixed MC–MD approach.

The results obtained from these studies show that the combined approach, MC–MD approach, can provide valuable insight into the details of folding and mis-folding mechanisms and, particularly about the delicate influence of local and non-local interactions in steering the folding process.

5 Conclusions and perspectives

In this account, we presented the basics of protein simulation methodology, and we highlighted different approaches that have been used to tackle the protein folding problem. Several examples were given highlighting the effectiveness of simplified models in capturing the basic thermodynamics, kinetics and structural features of different protein systems. We also discussed the great potential held by all-atom MD simulations in describing the details of folding pathways. With continuing advances in the methodologies and in computer power, all of these studies will be steadily extended to more complex and larger systems.

In this respect, the combination of simplified models and all-atom methodologies can provide many fundamental insights concerning several mechanisms of protein systems. Coarse-grained methods can be used to overcome the slower and time-consuming steps in the process creating suitable starting points for all-atom studies. The "time advancement" can be exploited to bring the real fine chemical detail only into the most significant selected structures, resulting in a net saving of computer time (less brute force) and allowing a higher degree of transparency and control in the analysis of the process.

We think that this philosophy can be successfully extended to the study of other important topics of modern molecular biology. The detailed structural characterization of the ensemble of conformations available to "intrinsically disordered proteins" can shed light on the presence of secondary structure content of particular regions, on the presence of possible long range contacts, on the degree of presence (or absence) and flexibility of tertiary structural organization. This information can be linked to NMR derived data in the effort to characterize the structure–activity relationships of this emerging important class of proteins. Another fundamental aspect of post-genomics biology is the understanding and characterization of protein–protein interactions and interaction networks. The nature of the *intermolecular* forces involved in the formation of complexes is the same as the nature of *intramolecular* forces involved in the folding of single molecules. Based on this, one could try to model a simplified intermolecular recognition pathway, treating the binding partners at a coarse-grained level with the same potentials as we have described above. Selected complexes can subsequently be projected into the all-atom world to allow the characterization of their structures and of possible molecular motions involved in the recognition process.

The involvement of complex formation and of intrinsically disordered proteins in regulatory and signaling pathways represent a cornerstone of modern biology and even nanotechnology. The structural and dynamical characterization is the main step towards the deep understanding of these biological mechanisms. For these aspects, the insight that can be offered by NMR technology, X-ray crystallography and other techniques is, despite the continuous advancements, still limited. In these contexts there appears to be ample scope for theoretical and numerical contributions. Methods combining simplified and all-atom protein descriptions appear to possess many of the characteristics necessary to tackle these issues with a contained computational investment and yet retaining the relevant chemical details.

Acknowledgements We gratefully acknowledge G. Carrea, F. Seno and A. van der Vaart for their critical comments and suggestions on the manuscript. Majo and Ile are also acknowledged for their critical advice.

References

1. Anfinsen CB (1973) Science 181:223
2. Anfinsen C, Scheraga HA (1975) Adv Protein Chem 29:205
3. Levinthal C (1969) Mossbauer spectroscopy in biological systems, University of Illinois Press, Urbana
4. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG (1995) Proteins 21:167
5. Wolynes PG, Onuchic JN, Thirumalai D (1995) Science 267:1619
6. Dill KA, Chan HS (1997) Nat Struct Biol 4:10
7. Dobson CM, Sali A, Karplus M (1998) Angew Chem Int Ed 37:868
8. Sali A, Shakhnovich E, Karplus M (1994) Nature 369:248
9. Micheletti C, Banavar JR, Maritan A, Seno F (1999) Phys Rev Lett 82:3372
10. Mirny LA, Abkevich VI, Shakhnovich EI (1998) Proc Natl Acad Sci USA 95:4976
11. Hoang TX, Trovato A, Seno F, Banavar JR, Maritan A (2004) Proc Natl Acad Sci USA 101:7960
12. Branden C, Tooze J (1991) Introduction to protein structure. Garland Publishing, New York
13. Creighton T (1993) Proteins, structure and molecular properties. 2nd ed. W.H. Freeman and Company, New York
14. Fersht AR (1999) Structure and mechanism in protein science. V.H. Freeman, New York
15. Pande VS, Grosberg AY, Tanake T, Rokhsar DS (1998) Curr Opin Struct Biol 8:68
16. Duan Y, Kollman PA (1998) Science 282:740
17. Debe DA, Carlson MJ, Goddard WA (1999) Proc Natl Acad Sci USA 96:2596
18. Makarov DE, Plaxco KW (2003) Prot Sci 12:17
19. Rohl CA, Strauss CE, Misura KM, Baker D (2004) Methods Enzymol 383:66
20. Dobson CM (2003) Nature 426:884
21. Gnanakaran S, Nymeyer H, Portman J, Sanbonmatsu KY, Garcia AE (2003) Curr Opin Struct Biol 13:168
22. Zhou RH (2003) Proteins 53:148
23. Ishikuza T, Terada T, Ans K, Shimizu SN (2004) Chem Phys Lett 393:546
24. Pascheck D, Garcia AE (2004) Phys Rev Lett 93:238105
25. Dill KA, Bromberg S, Yue K, Fiebig KM, Yee DP, Thomas PD, Chan HS (1995) Prot Sci 4:561
26. Sali A, Shakhnovich E, Karplus M (1994) Nature 369:248
27. Kolinski A, Skolnick J (1996) Lattice models of protein folding, dynamics and thermodynamics. Landes, Austin
28. Pokarowski P, Kolinski A, Skolnick J (2003) Biophys J 84:1518
29. Park B, Levitt M (1996) Proteins 258:367
30. Levitt M, Warshel A (1975) Nature 253:694
31. Warshel A, Levitt M (1976) J Mol Biol 106:421
32. de Gennes PG (1979) Scaling concepts in polymer physics. Cornell University Press, Ithaca
33. Sokal AD (1997) Nucl Phys B Suppl 47:172
34. Chan HS, Dill KA (1991) Annu Rev Biophys Biophys Chem 20:447
35. Micheletti C, Maritan A, Banavar JR (1999) J Chem Phys 110:9730
36. Ejtehadi MR, Hamedani N, Shahrezaei V (1999) Phys Rev Lett 82:4723
37. Li H, Tang C, Wingreen N (1996) Science 273:666
38. England JL, Shakhnovich EI (2003) Phys Rev Lett 90:218101
39. Shakhnovich EI, Gutin AM (1993) Protein Eng 6:793
40. Sun S, Brem R, Chan HS, Dill KA (1995) Protein Eng 8:1205
41. Seno F, Vendruscolo M, Maritan A, Banavar JR (1996) Phys Rev Lett 77:1901
42. Morrisey MP, Shakhnovich EI (1996) Fold Des 1:391
43. Micheletti C, Seno F, Maritan A, Banavar JR (1998) Phys Rev Lett 80:2237
44. Micheletti C, Seno F, Maritan A, Banavar JR (1998) Proteins 32:80
45. Tiana G, Broglia RA (2001) J Chem Phys 114:2503
46. Rossi A, Micheletti C, Seno F, Maritan A (2001) Biophys J 80:480
47. Mirny L, Shakhnovich EI (2001) Annu Rev Biophys Biomol Struct 30:361
48. Abkevich VI, Gutin AM, Shakhnovich EI (1994) Biochemistry 22:10026

49. Tiana G, Broglia RA (2001) J Chem Phys 114:7267
50. Chan HS, Dill KA (1997) Proteins Struct Funct Genet 8:2
51. Kolinsky A, Skolnick J (1994) Proteins 18:338
52. Socci ND, Onuchic JN (1994) J Chem Phys 101:1519
53. Yue K, Fiebig KM, Thomas PD, Chan HS, Shackhnovich EI, Dill KA (1995) Proc Natl Acad Sci USA 92:325
54. Gutin AM, Abkevich VI, Shakhnovich EI (1996) Phys Rev Lett 77:5433
55. Cieplak M, Hoang TX (1998) Phys Rev E 58:3589
56. Cieplak M, Henkel M, Karbowski J, Banavar JR (1998) Phys Rev Lett 80:3654
57. Kaya H, Chan HS (2000) Proteins 40:637
58. Kaya H, Chan HS (2000) Phys Rev Lett 85:4823
59. Chan HS, Shimizu S, Kaya H (2004) Methods Enzymol 380:350
60. Miyazawa S, Jernigan RL (1985) Macromolecules 18:543
61. Casari G, Sippl MJ (1992) J Mol Biol 224:725
62. Maiorov VN, Crippen GM (1992) J Mol Biol 227:876
63. Kolinsky A, Godzik A, Skolnick J (1993) J Chem Phys 98:7420
64. Maiorov VN, Crippen GM (1994) Proteins 20:173
65. Sippl MJ (1995) Curr Opin Struct Biol 5:229
66. Thomas PD, Dill KA (1996) Proc Natl Acad Sci USA 93:11628
67. Thomas PD, Dill KA (1996) J Mol Biol 257:457
68. Vendruscolo M, Domany E (1999) J Chem Phys 109:11101
69. van Mourik J, Clementi C, Maritan A, Seno F, Banavar JR (1999) J Chem Phys 110:10123
70. Micheletti C, Seno F, Banavar JR, Maritan A (2001) Proteins 42:422
71. Scala A, Dokholyan NV, Buldyrev SV, Stanley HE (2001) Phys Rev E 63:032901
72. Khatun J, Khare SD, Dokholyan NV (2004) J Mol Biol 336:1223
73. Samudrala R, Huang ES, Koehl P, Levitt M (2000) Protein Eng 13:453
74. Kolinski A (2004) Acta Biochim Pol 51:349
75. Liwo A, Khalili M, Scheraga HA (2005) Proc Natl Acad Sci USA 102:2362
76. Pillardy J, Czaplewsky C, Liwo A, Wedemeyer WJ, Ripoll DR, Arlukowics P, Oldziej S, Arnautova YA, Scheraga HA (2001) J Phys Chem B 105:7299
77. Irbäck A (2003) J Phys Condens Matter S1797–S1807
78. Favrin G, Irbäck A, Sjunnesson F (2001) J Chem Phys 114:8154
79. Irbäck A, Potthast F (1995) J Chem Phys 103:10298
80. van Gunsteren WF, Burgi R, Peter C, Daura X (2001) Angew Chem Int Ed 40:352
81. Daura X, Jaun B, Seebach D, van Gunsteren WF, Mark AE (1998) J Mol Biol 280:925
82. Daura X, Antes I, van Gunsteren WF, Thiel W, Mark AE (1999) Proteins Struct Funct Genet 36:542
83. Daura X, Gademann K, Jaun B, Seebach D, van Gunsteren WF, Mark AE (1999) Angew Chem Intl Ed 38:236
84. Daura X, van Gunsteren WF, Mark AE (1999) Proteins 34:269
85. Shortle D, Ackerman MS (2001) Science 293:487
86. Colombo G, De Mori GMS, Roccatano D (2003) Prot Sci 12:538
87. Colombo G, Roccatano D, Mark AE (2002) Proteins Struct Funct Genet 46:380
88. Shirts M, Pande VS (2000) Science 290(5498):1903
89. Pande VS, Baker I, Chapman J, Elmer SP, Khaliq S, Larson SM, Rhee YM, Shirts MR, Snow CD, Sorin EJ, Zagrovic B (2003) Biopolymers 68:91
90. Fersht AR (2002) Proc Natl Acad Sci USA 99(22):14122
91. Zagrovic B, Snow CD, Khaliq S, Shirts MR, Pande VS (2002) J Mol Biol 323:153
92. Kohn JE, Millett IS, Jacob J, Zagrovic B, Dillon TM, Cingel N, Dothager RS, Seifert S, Thiyagarajan P, Sosnick TR, Hasan MZ, Pande VS, Ruczinski I, Doniach S, Plaxco KW (2004) Proc Nat1 Acad Sci USA 101:12491
93. Fitzkee NC, Rose GD (2004) Proc Nat1 Acad Sci USA 101:12497
94. Fitzkee NC, Fleming PJ, Gong H, Panasik N Jr, Street TO, Rose GD (2005) Trends Biochem Sci 30:73
95. De Mori GMS, Colombo G, Micheletti C (2005) Proteins Struct Funct Bioinform 58:459
96. De Mori GMS, Micheletti C, Colombo G (2004) J Phys Chem B 108:12267
97. Min-yi Shen KFF (2002) Proteins Struct Funct Genet 49:439
98. Sullivan DC, Kuntz ID (2002) J Phys Chem 106:3255
99. Hansmann UHE (2002) Int J Quantum Chem 90:1515
100. Berrera M, Molinari H, Fogolari F (2004) BMC Bioinformatics 4:8