PROTEINS: Structure, Function, and Genetics 53:720–730 (2003)

Elucidation of the Disulfide-Folding Pathway of Hirudin by a Topology-Based Approach

C. Micheletti,[1]* V. De Filippis,[2] A. Maritan,[1] and F. Seno[3]
[1]*International School for Advanced Studies, INFM and the Abdus Salam Centre for Theoretical Physics, Trieste, Italy*
[2]*Dipartimento di Scienze Farmaceutiche and CRIBI Biotechnology Centre, University of Padova, Padova, Italy*
[3]*INFM-Dipartimento di Fisica "G. Galilei," University of Padova, Padova, Italy*

ABSTRACT **A theoretical model for the folding of proteins containing disulfide bonds is introduced. The model exploits the knowledge of the native state to favor the progressive establishment of native interactions. At variance with traditional approaches based on native topology, not all native bonds are treated in the same way; in particular, a suitable energy term is introduced to account for the special strength of disulfide bonds, as well as their ability to undergo intramolecular reshuffling. The model thus possesses the minimal ingredients necessary to investigate the much debated issue of whether the refolding process occurs through partially structured intermediates with native or non-native disulfide bonds. This strategy is applied to a context of particular interest, the refolding process of hirudin, a thrombin-specific protease inhibitor, for which conflicting folding pathways have been proposed. We show that the only two parameters in the model (temperature and disulfide strength) can be tuned to reproduce well a set of experimental transitions between species with different number of formed disulfides. This model is then used to provide a characterization of the folding process and a detailed description of the species involved in the rate-limiting step of hirudin refolding. Proteins 2003;53:720–730.** © **2003 Wiley-Liss, Inc.**

Key words: protein folding; disulfide bonds; stochastic simulation; folding intermediates

INTRODUCTION

The characterization of the folding pathway of proteins is one of the fundamental problems in molecular biology and has received increasing scientific attention as a result of the continuous advancements in experimental and theoretical biochemistry. After the work of Anfinsen,[1] who demonstrated that ribonuclease unfolds and refolds reversibly into its native (active) three-dimensional (3D) structure, it has generally been accepted that the primary sequence usually contains sufficient information to direct the complete folding process. What typically remains elusive to experimental and theoretical investigations is the pathway of this spontaneous process and the mechanisms that govern it.

A considerable progress in this direction is possible for proteins containing native disulfide bonds. The formation of disulfide bonds during the folding process can be controlled experimentally through the use of an appropriate thiodisulfide redox couple and thiol quenching agent.[2–4] By these means, the regeneration process can be halted, and the intermediate species can be trapped, isolated, and characterized. Historically, one of the most investigated proteins containing disulfide bonds has been the bovine pancreatic trypsin inhibitor (BPTI). Starting with the work of Creighton,[5] a series of crucial studies have accumulated a wealth of evidence in favor of the existence of partially structured intermediates along the protein-folding pathway. Despite these efforts, the detailed characterization of the intermediates turned out to be a delicate experimental matter, and there is still not a universal agreement on whether intermediates contain native or non-native disulfide bonds,[3,4,6] and whether there exists more than one pathway.[4,7] In this context, the use of theoretical and computational tools[8–12] has been extremely useful in complementing the experimental findings with a more precise characterization of the folding pathway, albeit obtained for models that greatly simplify the complexity of the real system.

In this article, we propose a theoretical scheme to study the folding of proteins with disulfide bonds by suitably exploiting the (known) protein native structure. At a general level, our strategy falls in the class of approaches that builds on the importance of native-state topology in steering the folding process,[13] that is, in bringing into contact pairs of amino acids that are found in interaction in the native state. In the past few years, an increasing number of experimental and theoretical studies have confirmed the utility of these approaches in the characterization of various aspects of protein-folding processes.[14–26] In this work, we try to generalize this strategy by adding a suitable treatment of disulfide bonds accounting for both their strength and their capability to undergo intramolecular reshuffling.

In order to validate this approach, we investigated the folding process of the N-terminal core domain of hirudin HM2 from the blood-sucking leech *Hirudinaria manillen-*

Grant sponsor: INFM, FISR-MIUR 2001, and Murst Cofin 2001.

*Correspondence to: Cristian Micheletti, International School for Advanced Studies (SISSA/ISAS), Via Beirut 2–4, I-34014 Trieste, Italy. E-mail: michelet@sissa.it

Received 4 October 2002; Accepted 26 February 2003

sis.[27] Hirudins are the most potent and specific inhibitors of thrombin (a key enzyme in blood coagulation) identified so far, and they are currently used as effective anticoagulants. Hirudin HM2 is composed of a compact N-terminal domain (residues 1–47) stabilized by three disulfides (Cys6–Cys14, Cys16–Cys28, Cys22–Cys37) and a highly flexible, negatively charged C-terminal tail. Structural studies conducted on several leech-derived disulfide-rich small proteins (i.e., hirudin, decorsin, and antistasin) reveal that although these proteins display negligible sequence similarity and different function, they share a common disulfide topology and 3D fold,[28] suggesting that leeches use the same protein scaffold but different binding epitopes to affect hemostasis. Notably, it has been found that these leech antihemostatic proteins possess a T-knot scaffold similar to that observed for other unrelated proteins, including b-transforming growth factors, wheat germ agglutinins, and snake venom toxins.[29,30] In this view, the results of our study may have relevant and more general implications on the elucidation of the folding pathway(s) of other protein systems.

The first experimental attempts to identify the folding pathway of hirudin date back to the studies of Otto and Seckler, who argued that hirudin could be a viable alternative to BPTI, and showed that it is experimentally feasible to obtain and follow its unfolding/refolding processes.[31] As for the case of BPTI, hirudin has also been studied in different experiments[31–34] leading to alternative formulations of its folding pathway.[32,34] In particular, when dissolved atmospheric oxygen was used as an oxidizing agent, the folding process appeared to occur first through the establishment of three non-native (scrambled) disulfides, and later through their slow rearrangement into the native bonding pattern. On the contrary, no evidence for the importance of these fully oxidized disordered intermediates was found in the experiments of Thannhauser et al. using oxidized DL-dithiothreitol (DTT^{ox}).

These alternative views prompted the present investigation of the hirudin pathway within a topology-based model. To ascertain that, despite its simplified nature, the model was suitable to characterize the main aspects of the folding process, we have first tuned and validated it against the set of experimental measurements provided by Thannhauser et al.[34] Based on the success of this comparison we have undertaken a detailed characterization of the folding process by monitoring quantities inaccessible in previous experiments. Our study confirms the experimental evidence of Thannhauser et al.,[34] suggesting that, under a certain set of experimental conditions, the rate-limiting step of the refolding process involves a transition from species with two disulfides to ones with the three native disulfides.[34] A detailed analysis of the numeric dynamic trajectories further reveals the precise order of formation of the disulfide bonds.

THEORY AND RESULTS

The starting point of our analysis is the 1–47 fragment of hirudin[33] resolved by NMR,[35] shown in Figure 1 (see Methods and Materials section). The fragment under consideration contains three disulfide bonds between residues 6–14, 16–28, and 22–37, and differs from the whole protein because it lacks the 18-residue-long tail; this tail plays an important role in the biologic activity of the protein. However, it is highly mobile due to the virtual absence on any noncovalent contacts with the 1–47 globular fragment. For this reason we neglected the hirudin tail in our numeric characterization of the folding process.

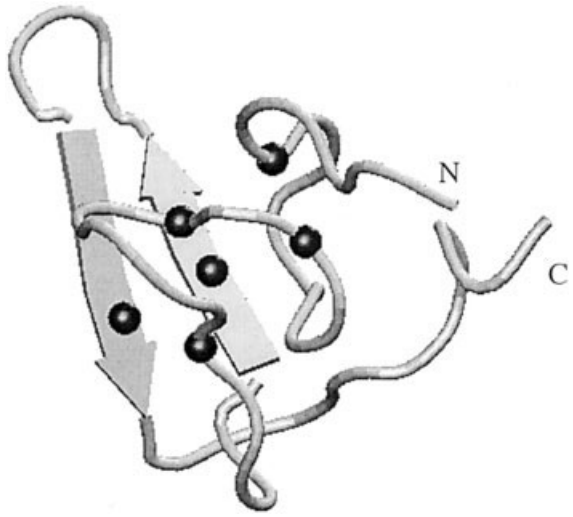

Fig. 1. Native structure of the 1–47 fragment of hirudin. The six cysteine residues, 6, 14, 16, 22, 28, 37 have been highlighted.

The effective energy scoring function that we adopted belongs to the class of topology-based Hamiltonians.[13,36] The knowledge of the above-mentioned native structure, Γ^N, is exploited to construct an effective energy function that admits Γ^N as the lowest energy state. The simplest form of such energy function is as follows:

$$E(\Gamma) = -\Sigma_{ij}\, \Delta(\Gamma^N) \cdot \Delta_{ij}(\Gamma), \qquad (1)$$

where $E(\Gamma)$ is the energy of a trial conformation Γ, and $\Delta(\Gamma)$ is its contact matrix, whose elements are 1 [0] if amino acids i and j are [not] in interaction. A standard criterion based on C_α or C_β distance of pairs of amino acids is used to decide whether two amino acids are interacting.

In the simple case of Eq. (1), the energy minimum is attained when all native bonds are established. Such native interactions are weighted equally. This simplification can be justified when the effective amino acid interactions in the protein are of comparable strength and has the advantage of keeping the model transparent by avoiding the use of imperfect energy parameterization.[37–40]

It is important to notice that the equal weighting of native contacts does not imply that, in a thermalized ensemble, they are formed with equal probability. In fact, it is the complex interplay of energy and structural entropy that dictates the most probable routes to the native state from an unfolded conformation, as well as the presence of rate-limiting steps due to the establishment of crucial sets of native contacts.[22]

In this study, the equal weighting of native contacts does not appear to be a good starting point, because one has to

account for the very energetic disulfide bridges that can occur between pairs of Cys. For this reason, in place of Eq. (1), we adopt a scoring function consisting of two terms (see also Methods and Materials section):

$$\mathcal{H} = V_{n-ss} + \mu V_{ss}. \quad (2)$$

V_{n-ss} enforces some general constraints on the peptide chain geometry and promotes the formation of native contacts between pairs of amino acids other than Cys–Cys. These contacts are weighted in the same manner. The second term, V_{ss} rewards the formation of disulfide bonds between pairs of cysteines. It is important to stress that the formation of disulfide bonds is allowed among any pair of Cys, not just the native ones. By doing so, we can investigate the extent to which species with one or more non-native disulfides are present and whether they influence the dynamics toward the native state, as proposed in recent experiments.

The strength of the disulfide bonds relative to other non-covalent interactions is controlled by the parameter μ. This parameter should not be regarded as a relative measure of the "bare" (i.e., in vacuum) disulfide strength. In fact, because our model does not treat the solvent explicitly, μ captures the effective strength of disulfide bonds in the presence of water and any other appropriate reducing/oxidizing agent. For this reason, the value of μ and also that of the heat bath temperature, T, have to be chosen in a suitable way to reproduce as accurately as possible the conditions of a given experiment.

In this study, we focus on the set of experiments carried out by Thannhauser et al.,[34] in which hirudin was refolded in the presence of various concentrations of DTT^{ox}. We first show that there exists a well-defined region in the $\mu - T$ parameter space in which the rates of conversion between species with different numbers of disulfides match well those observed in experiments. This is an important fact, because it proves that, despite its simple form, the energy scoring function of Eq. (2) can indeed be used to characterize the folding process obtaining the correct quantitative experimental picture. Based on such stringent validation of our strategy, we then monitor quantities that are inaccessible in current experiments and thus propose a vivid and detailed picture for the hirudin refolding process.

We want to stress the fact that the possibility of reproducing the results of Thannhauser et al.[34] for a suitable choice of the $\mu - T$ parameters supports the hypothesis that our model is general enough to study any folding process involving the formation of disulfide bonds.

Comparison With Experimental Rates

The experimental benchmark for our model is provided by a series of hirudin refolding experiments carried out by Thannhauser et al.[34] under various concentrations of DTT^{ox}. By using several combined techniques, they established that the refolding process occurs under the reaction shown in Figure 2, where R, $1S$, and $2S$ denote respectively the species with 0, 1, and 2 disulfides (either native or non-native). A certain species with three disulfides, denoted as $3S^*$, was seen to convert with extreme rapidity to native hirudin and was therefore identified with a native arrangement of the three disulfide contacts. The remainder of the ensemble of structures with three disulfides (i.e., with at least two non-native disulfides) is therefore denoted as $3S$. The experimental characterization of Thannhauser et al. provides the rates for the individual reactions of the above model for a few choices of the initial concentrations of DTT^{ox} and the reduced protein species. Our first goal was to see whether for suitably chosen values of T and μ it was possible to reproduce such rates.

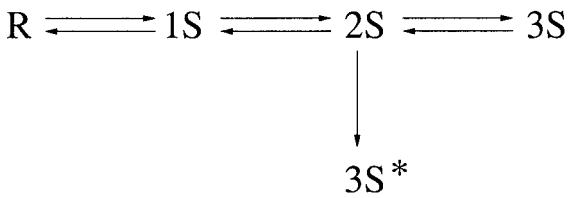

Fig. 2. Best fit model of Thannhauser et al.[34] to the hirudin refolding experiments. Species with 0, 1 and 2 disulfides are denoted as R, $1S$ and $2S$, respectively. The species denoted as $3S^*$ contains the three native disulfides. The remainder of the ensemble of structures with three disulfides is therefore denoted as $3S$.

Several values of T and μ were considered and, for each of them, we measured the rates of conversion between the R, $1S$, $2S$, $3S$, and $3S^*$ species by using a Monte Carlo technique. This was done by keeping track of how many transitions out of any given species occurred and how many of them ended up in each of the other species. To ensure that the obtained dynamics could be interpreted as a (coarse grained) viable dynamical trajectory,[41] only tiny, local distortions of the peptide chain were attempted in the Monte Carlo (MC) moves. During such coarse-grained trajectories, the formation, breaking, or reshuffling of disulfide bonds obeys a set of constraints dictated by the chemistry of the disulfide bonds and the thiol-disulfide coupling. In particular, no Cys residue is allowed to participate with more than one disulfide; the number of disulfide bonds can decrease or increase by, at most, one unit at each time step, respectively, when an existing disulfide is broken or through the establishment of a new disulfide between two previously unbonded Cys residues. Intramolecular rearrangements in which one bond is broken in favor of a new one (thus preserving the total number of disulfides) are also allowed under the requirement that the new bond involve one of the cysteines of the broken disulfide.

Particular care is necessary to ensure that the Monte Carlo dynamics subject to these constraints do not violate detailed balance. The difficulties arise from the fact that an elementary MC distortion of distinct starting configurations may result in structures that are compatible with a different number of allowed disulfide-bonding patterns.

To overcome this problem, we have proceeded as follows: To a newly generated conformation Γ we randomly associate one of the 15 possible pairing patterns of the three disulfides. If the proposed bonding pattern of the new structure violates the previous criteria, the structure is

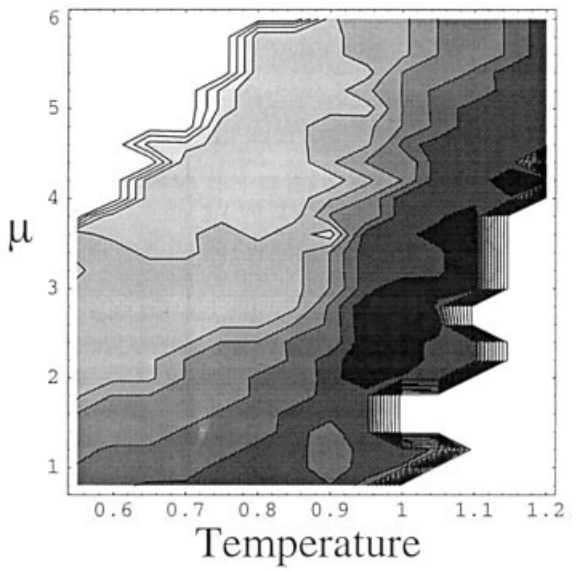

Fig 3. Contour and density plot of the Kendall correlation across the $T - \mu$ plane. The white regions correspond to the minimum values of the correlations were observed ($\tau \approx 0$). The highest correlation, $\tau = 0.81$, was observed in the dark area located around $\mu = 2.8$, $T = 1.0$.

rejected and time is incremented. Otherwise, the energy function is recalculated and the structure is rejected or accepted with the usual Metropolis criterion (see Methods section for further details). In this way, detailed balance is obviously satisfied, because the same number of bonding patterns is proposed for each structure. The downside is that one encounters frequent Metropolis rejections as a result of the "blind" proposal of bonding patterns.

For each of the explored values of T and μ, we have measured the correlation between the experimental rates and those obtained numerically. As a measure of the degree of correlation between these two quantities, we used the nonparametric Kendall analysis.[42] This statistical tool allows us to establish whether there is a relationship between two sets of data and how statistically significant it is. Being based on the comparative ranking of corresponding data in the two sets, the Kendall analysis does not rely on any preassigned parametric dependence (e.g., linear) between the two quantities. For this reason, it is regarded as a very robust measure of correlation and appears particularly appropriate in this context in which the measured rates (both experimental and numerical) span several orders of magnitude.[42]

Our findings are summarized in Figure 3, where we have shown the contour and density plot of the Kendall correlation coefficient, τ, against the rates pertaining to the experimental conditions of Figure 5(a) in Thannhauser et al.[34] These effective first order–like rates were obtained starting from the best-fit experimental rates of Table 2[34] multiplied by the asymptotic concentration of DTT^{ox} or DTT^{red} measured under the given experimental conditions (see also Fig. 6[34]). A darker/lighter shadow in Figure 3 denotes the presence of a higher/lower degree of correlation. It is apparent that the experimental data are well reproduced in the neighbourhood of $\mu = 2.8$, $T = 1.0$ for which the highest correlation $\tau = 0.81$ is observed. To ascertain the statistical significance of observing such correlation, we have computed the probability to observe by pure chance a correlation larger than the observed one. It turns out that this probability (double-sided) is equal to $p = 0.01$, which testifies to the statistical significance of the observed correlation. This establishes that there is a strong monotonic relation between the experimental and theoretical rates.

A scatter plot of the logarithm of experimental rates versus the numerically obtained ones is provided in Figure 4, where the good interdependence of the data can be visually inspected. If one were able to model the folding process with detailed and accurate energy functions, one would expect an equality of the theoretical and experimental rates, apart from a time-conversion factor. In our case, due to the simplicity of our model, we do not observe such dependence but encounter instead another simple, linear relationship between the logarithms of both sets of rates. This is visible in the linear fit of Figure 3; the corresponding correlation coefficient is $r = 0.88$. Its statistical significance (two-sided) over the set of seven data points is 4%, which is compatible with the significance of the more general (and robust) Kendall correlation. These values indicate that the correlation is highly significant from a statistical point of view.

A further validation can be carried out by comparing the asymptotic concentration of the various species observed in the experiment and in an equilibrated MC trajectory. In case of a perfect correlation between the experimental and numerical rates, this further check would be redundant. In this context, where the correlation is not perfect (a significant discrepancy is seen for the transition between the $3S$ and $2S$ species) this validation is useful to ascertain whether the differences in the rates result in significantly different equilibrium conditions. The plot of Figure 4(b) reveals that the asymptotic concentrations are in good accord; hence, by setting $T = 1$ and $\mu = 2.8$, one can be confident that the MC trajectory is compatible with both the dynamical and equilibrium properties of the experimental system.

Thermodynamics

It is interesting to analyze the thermodynamic behavior of the system at $\mu = 2.8$ as a function of the heat bath temperature, T. By using the standard yet powerful method of histogram reweighting (see Methods section), we have computed the average internal energy as a function of T. The data, shown in Figure 5(a), indicate the presence of a point of inflection for temperatures close to 1. The presence of a transition at this temperature is further corroborated by the behavior of the specific heat C_v, which displays a clear peak at $T = T_F \approx 0.91$, the folding temperature, and by the presence of two minima in the free-energy profile in Figure 5(b).

This peak is associated to the folding transition in the system, and its neatness suggests that the folding process has a two-state character.[12] This is indeed confirmed by an analysis of the free-energy landscape at

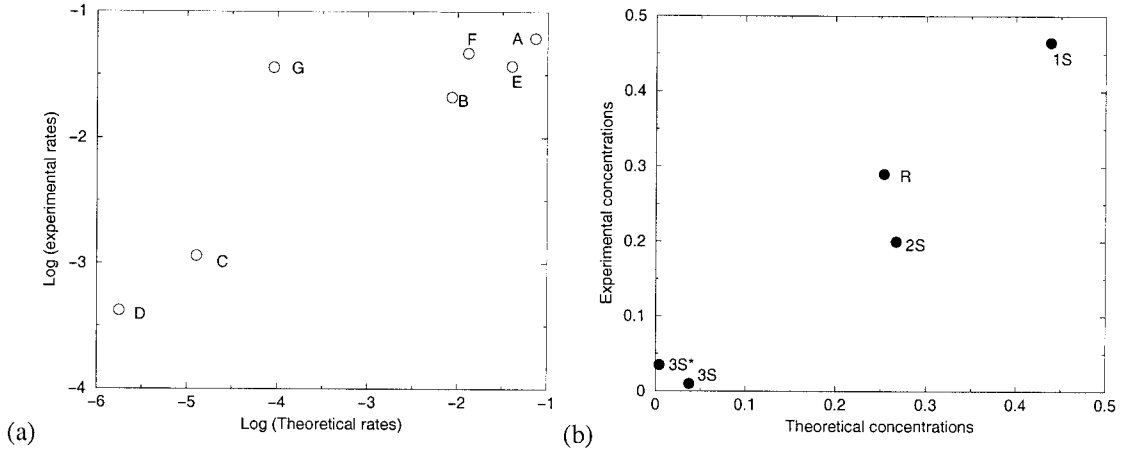

Fig 4. (a) Linear correlation between the log of experimental rates of Figs. 5a ref. [34] and those obtained through the present numerical calculation for $\mu = 2.8$, $T = 1.0$. The experimental rates are expressed in min^{-1}. The points in the plot correspond to the following transitions A: $R \rightarrow 1S$, B: $1S \rightarrow 2S$, C: $2S \rightarrow 3S$, D: $2S \rightarrow 3S^*$, E: $1S \rightarrow R$, F: $2S \rightarrow 1S$, G: $3S \rightarrow 2S$. (b) Scatter plot of the equilibrium fraction of the various species obtained in the Monte Carlo trajectory and in the experiment. The experimental concentrations were extracted from Fig. 5a of ref [34] for $t = 500$ s.

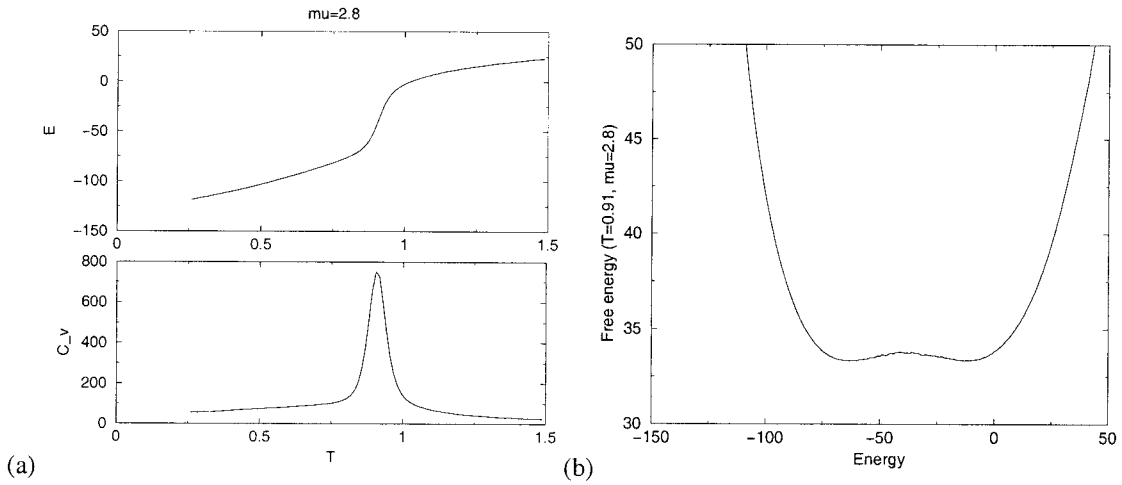

Fig 5. (a) Total energy and specific heat for $\mu = 2.8$. (b) Free energy profile, as a function of the internal energy, at the point $\mu = 2.8$, $T = 0.91$.

T_F, which exhibits two minima as a function of V_{n-ss} in correspondence of the unfolded and folded states (data not shown). The special point $\mu = 2.8$, $T = 1.0$, corresponding to the experimental conditions of Thannhauser et al.[34] appears therefore to be located slightly above the folding transition temperature (for $\mu = 2.8$). This is entirely consistent with the fact that the Thannhauser et al.[34] experimental conditions, Figure 5(a), do not particularly favor the formation of the native state that, indeed, involves only a small percentage of the total equilibrium population.

We have portrayed in Figure 6 the free-energy landscape for the different reduced species as a function of V_{n-ss}. It can be seen that the free-energy profiles have minima at different energy values according to the number of correctly formed disulfides. The R, $1S$, $2S$, and $3S$ species, in fact, display a minimum near $E \approx 0$, which denotes an unfolded ensemble [Fig. 5(b)]. On the contrary, the species with three native disulfides has a free-energy minimum for energies much closer to the native-state energy.

The relatively high values of free-energy associated with the $3S^*$ species reflect the particular experimental conditions reproduced here, in which the asymptotic fraction of species with native disulfides is low. Clearly, by lowering the temperature, one favors the formation of the native structure, which is accompanied by an increase of the concentration with native bonding patterns for the cysteines. This effect is visible in Figure 7(a), in which the average fraction of formed disulfide bonds is portrayed as a function of temperature. It can be seen that above T_F, one has a significant formation of non-native bonds that are superseded by native ones below T_F.

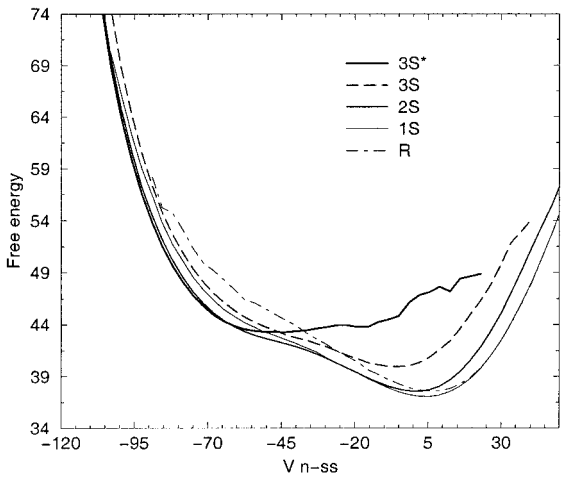

Fig 6. Free energy profile for $T = 1.0$ and $\mu = 2.8$ of the R, $1S$, $2S$, $3S$ and $3S^*$ ensembles as a function of V_{n-ss}.

An alternative picture was put forward by Chatrenet and Chang based on a hirudin refolding experiment.[32] Due to the much more oxidizing conditions than those considered by Thannhauser et al.,[34] Chatrenet and Chang[32] observed that the folding of hirudin occurred through the formation of intermediates with three (typically non-native) formed disulfides. Through a slow reshuffling process, the disulfides would then rearrange in the native pairing from which the native conformation could be easily reached.

Our findings indicate that this alternative scenario could probably be captured by our model with the use of larger values of the effective disulfide strength μ. This is consistent with the different oxidizing solvent conditions[34] adopted by Chatrenet and Chang.

To investigate this regime, we explored the value of $\mu = 10$. It is important to point out that the folding transition temperature (identified through the location of the specific heat peak) is almost insensitive to the disulfide strength in the range $0 \leq \mu \leq 10$. By comparing the plots in Figure 7, it is then possible to see that for the higher value of μ, the total number of formed disulfide bonds is higher than for $\mu = 2.8$. This result, accompanied by the fact that the total energy depends weakly on μ, confirms the intuition that, for large μ's, the disulfides are established at early stages of the folding process, when the rest of the protein is still unstructured. This "greedy" disulfide formation impacts not only on the total number of formed disulfides but also on the relative fraction of correct (native ones). As a result, a much higher fraction of wrong bonds is found at any temperature, as is noticeable in Figure 7(b). This picture suggests that the large μ regime of our model may be compatible with the scenario proposed by Chatrenet and Chang, in which the rate-limiting step corresponded to the disentanglement of non-native disulfides in intermediates states with three formed disulfide bonds. However, for this alternative case, the experimental rates are not available, and we cannot corroborate in a more quantitative way the parallel between the experimental conditions of Chatrenet and Chang and the "large" μ regime in our model.

FOLDING PATHWAYS

So far, we have focused on the overall thermodynamic characterization of the refolding of hirudin, paying particular attention to the validation of our model against experimental data. Having established that the detailed experimental reaction rates of Thannhauser et al.[34] can be well reproduced, we can confidently use our model to investigate finer aspects of the refolding process.

We begin by examining the details of formation of the native arrangement of disulfides. Our interest is to find whether the formation of the $3S^*$ species is aided by the establishment of some non-native disulfides that are later broken in favor of native bondings. To do so, instead of subdividing the structures only according to the overall number of disulfides, we indexed them with a pair of numbers (n_c, n_w) denoting the number of correct (native) disulfides, n_c, and wrong (non-native) ones, n_w. In this way, the fully reduced state, R, is indicated as (0, 0), whereas the $3S^*$ state corresponds to (3, 0). Altogether, there are nine possible states, as indicated in Figure 8.

We have generated several MC trajectories, each spanning about 20 million time steps (see Methods and Materials sections) at ($T = 1.0$, $\mu = 2.8$), and then recorded the probability of occurrence of all allowed transitions between the states of Figure 8. The typical dispersion on the measured rates over 10 MC trajectories was typically less than 1% but augmented to about 10% for the few very improbable transitions. Thus, we have considered all possible productive routes taking from R to $3S^*$ in a preassigned number of transitions. The probability of occurrence for any such routes is obtained by multiplying the individual probabilities of any of the elementary steps. By considering productive routes involving an increasing number of transitions, it can be established that the most probable route (see Fig. 8) is $R \equiv (0, 0) \rightarrow (1, 0) \rightarrow (2, 0) \rightarrow (3, 0) \equiv 3S^*$, which is more than 10 times as likely than the second ranking path, $R \rightarrow (0, 1) \rightarrow (1, 0) \rightarrow (2, 0) \rightarrow \equiv 3S^*$. These findings seem robust compared to variations of the parameters in our model. For example, even changing the temperature to 0.8 (keeping $\mu = 2.8$), so that the formation of native species is strongly favored [see Fig. 7(a)], it is found that the top-ranking routes from the reduced state to the native one are the same as above. Interestingly, the relative weight ratio of the top productive routes is analogous to the one encountered for $T = 1.0$; on the other hand, the fact that the native formation is highly favored at $T = 0.8$ is reflected in an increase, by several orders of magnitude, of the weight of the productive routes.

We complete this section by identifying the rate-limiting steps of the folding process. In a sequential reaction, the rate-limiting step is straightforwardly identified as the slowest one. In this scheme, such simple analysis cannot be carried out, because of the presence of several alternative routes to the native state. An objective and convenient way to identify the rate-limiting step in such a situation is to identify the reaction whose rate change affects most the production of the species of interest, in this case, $3S^*$. Therefore, by using the measured rates for the transitions

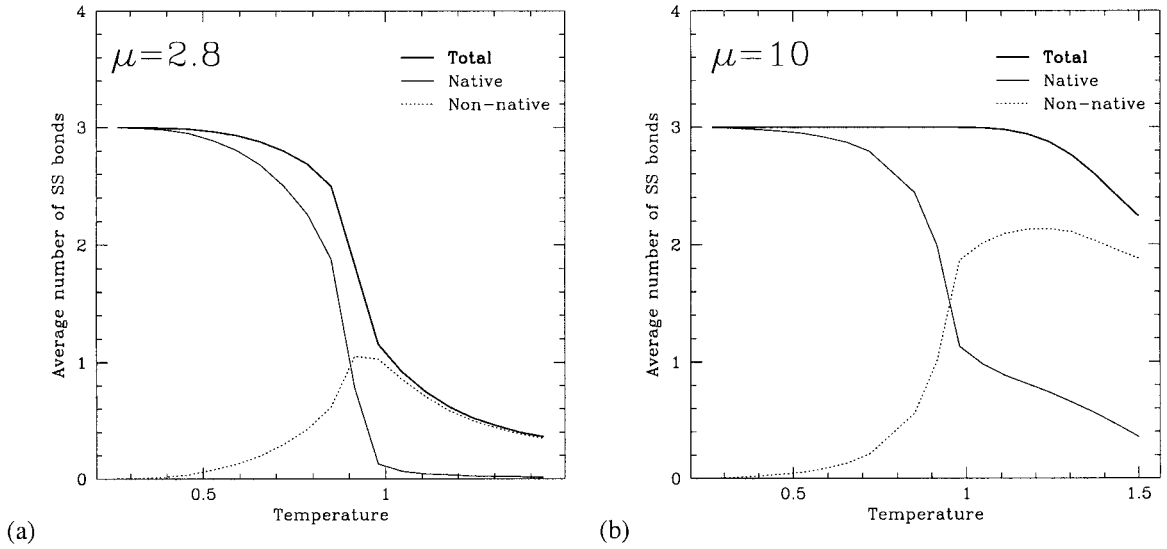

Fig 7. Average number of total (thick continuous line), native (continuous line) and non-native (dotted line) disulfide bonds for (a) $\mu = 2.8$ and (b) $\mu = 10$.

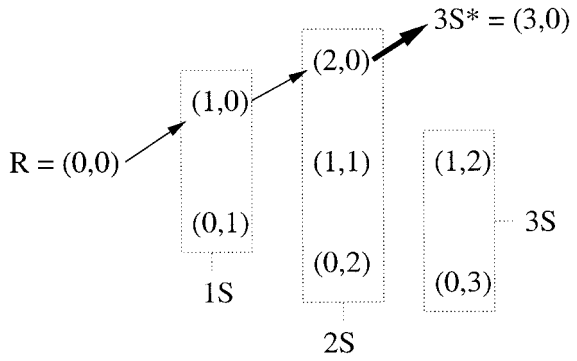

Fig 8. The different states considered in our trajectory analysis. The arrows indicate the most probable route from the reduced ensemble toward the native arrangement of disulfides. The thickest arrow denotes the rate-limiting step of the reaction.

between the states in Figure 8, we have integrated the associated master equations starting from a fully reduced population. We have then changed by 10% each of the rates and identified the time at which the concentration of $3S^*$ crosses a preassigned threshold value. We found that, almost independently of the threshold value, the most sensitive step among all the allowed ones was $(2, 0) \rightarrow 3S^*$ which is the last step of the most probable route leading to $3S^*$. In principle, this may have not been the case, especially in the presence of equally important paths leading to $3S^*$. The fact that the most probable routes include the rate-limiting steps confirms the existence of a well-defined succession of events to the native state.

As mentioned before, based on experimental dynamic plots, Thannhauser et al.[34] had determined that the rate limiting step was the $2S \rightarrow 3S^*$ one. In their study, it was not possible to characterize by direct means whether the transition to the $3S^*$ state occurred from a $2S$ state comprising only native cysteine bonds, although this was reputed to be the most likely scenario. Our picture thus fully supports the experimental results concerning the $2S \rightarrow 3S^*$ rate-limiting step but also adds novel insight into the process by indicating explicitly that the most crucial step involves a particular $2S$ species, namely, one with two native disulfides. In the remainder of this section, we further characterize the folding process by identifying which of the native disulfides are formed in the most probable routes (or rate-limiting steps). This is done by following a series of individual dynamic trajectories in which the native conformation is reached starting from a reduced state.

Folding Trajectories

The Monte Carlo procedure allows for a detailed study of the folding pathways and a systematic analysis of the specific order of formation of disulfide bonds. By setting $\mu = 2.8$, the protein is initially thermalized at a high temperature ($T \sim 10$) where, in our model, the disulfide bonds are completely reduced, then suddenly quenched to $T = 1.0$. For each kinetic trajectory, the order of formation of the disulfide bonds was stored and a statistical analysis accomplished by comparing several runs. In particular, we have recorded the exact type of disulfide bridges forming the different species ($1S$, $2S$), and from their relative concentration we have inferred the pathway.

The six Cys residues of the fully reduced protein (6, 14, 16, 22, 28, and 37) are equally likely to participate in forming the initial disulfide bond. Although, in the early folding stages, the contact (14–16) appears quite frequently due to the short sequence distance between the amino acids, after the molecule has equilibrated through a series of rapid internal disulfide interchange reactions, only 5 of the 15 possible $1S$ states exist in significant quantities. They are (6–14), (6–16), (16–22), and (28–37), which, respectively, are present in the following equilibrium concentrations: 7%, 6%, 12%, and 26%. It is interesting to notice that the most common bond among the $1S$

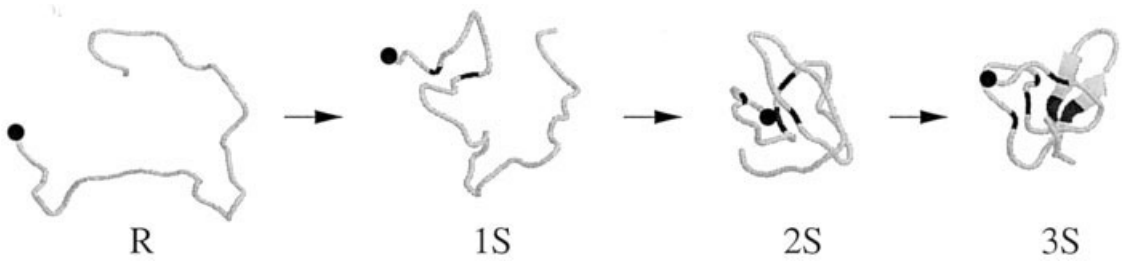

Fig 9. Snapshots of one of the possible folding trajectories. The 1*S* state involves the native contact (6–14), while the 2*S* species involves the native bonds (6–14) and (16–28). A black sphere is used to distinguish the N-terminus while cysteines involved in disulfide bonds are highlighted as dark segments.

ensemble is a non-native one (28–37), whereas the only native bond present in significant quantities is (6–14).

Of the 45 possible 2*S* states, seven occur in significant quantities. Two of them (6–14, 16–28) present in 1% of the cases and (6–14, 22–37) present in 0.7% of the cases, are formed by native bonds. The other two (6–14, 16–22) and (6–14, 28–37) present with a concentration of 2% and 5%, respectively, have a native and a non-native contact, whereas the last three, (6–16, 28–37) with a frequency of 9%, (14–22, 28–37) with a frequency of 7%, and (16–22, 28–37) with a frequency of 7%, are formed by non-native disulfide bonds. The relatively high concentration of species with non-native states, which are not present in the most probable productive routes expected for the folding (see the previous section), reflect the particular experimental conditions reproduced here (induced by the choice of T and μ) in which the formation of the native state is not particularly favored [Fig. 4(b)].

The analysis of the folding pathway reported in the previous section and the statistical analysis of the dynamic productive pathway can be summarized in the following refolding picture: The reduced proteins forms first the native contact (6–14) and, consequently, either the states (6–14, 16–28) or (6–14, 28–37) almost with the same probability. The folding proceeds with the formation of the last disulfide bond. This conversion is relatively slow, in agreement with the finding of the previous section on the rate limiting step of the reaction. The presence of conformation with three scrambled disulfide bonds also turns out not to be statistically significant in this kinetic analysis. A graphic representation of the most probable trajectory is shown in Figure 9.

The fact that the majority of structures sampled in the quenching process involve non-native bonding patterns is not in contradiction with the findings of the previous section, where the most probable productive route was shown to be free of non-native bonds. In fact, the high concentration of structures with non-native bonds does not automatically imply that they contribute significantly to the refolding "flux" toward the native state. On the contrary, species involving native bonds, even if they do not accumulate to high equilibrium concentration, appear to take part in the most efficient routes leading to the native state.

METHODS AND MATERIALS

Structure of Hirudin

The structure of the synthetic analogues of fragment 1–47 of hirudin HM2 in the free state was modeled on the NMR solution structure of natural fragment 1–47,[35] almost super-imposable on that of the corresponding segment (1–49) in intact hirudin HV1 variant[43] (PDB code: 5HIR), showing 75% sequence identity with hirudin HM2, or to that of HV1 fragment 1–51[44] (PDB code: 1HIC). The best representative model in the NMR ensemble was selected with the program OLDERADO,[45] available online at http://neon.chem.le.ac.uk/olderado/.

The Model

In our model, a generic conformation $\Gamma\{\vec{r}_i\}$ of the protein is modeled as a self-avoiding chain of connected C_α atoms located at position $\vec{r}_i^{\,\alpha}$, where i is the amino acid chain index (ranging from 1 to 47). Starting from the C_α coordinates, the peptide dihedral angles are calculated; hence, following standard geometrical rules,[39,46] we construct an effective C_β centroid for all residues, with the exception of the two end residues and the seven glycines (indices: 10, 18, 23, 25, 34, 40, 42).

The energy-scoring function consists of two terms. The first incorporates a standard bias toward the formation of native nondisulfide bonds but also disfavors the formation of non-native bonds (except for disulfide ones) and penalizes significant deviations from the native dihedral angles. These details are known to improve the cooperativity of the modeled folding process.[47–50] The explicit expression of this term, evaluate on a trial structure Γ is given by

$$V_{n-ss}(\Gamma) = V_0 \,\Sigma'_{i,j>i+2}\Delta_{ij}(\Gamma^N)\left[5\left(\frac{r_{ij}^N}{r_{ij}}\right)^{12} - 6\left(\frac{r_{ij}^N}{r_{ij}}\right)^{10}\right] \quad (3)$$

$$+V_1\,\Sigma'_{i,j>i+2}[1-\Delta_{ij}(\Gamma^N)]\left(\frac{r_{ij}^N}{r_{ij}}\right)^{12} + V_2\,[\Sigma_{i=2,46}(\theta_i-\theta_i^N)^2 + \Sigma_{i=3,46}(\phi_i-\phi_i^N)^2] + V_{constraints}, \quad (4)$$

where r_{ij} denotes the distance of the ith and jth C_α atoms in the trial structure Γ; a superscript N is used to denote analogous quantities pertaining to the native structure. The contact map Δ is computed by considering as threshold a distance of 8 Å between the C_α atoms. The prime in

the summation indicates that no contribution is considered if both the amino acids are Cys. Finally, we have constructed the angular term using the standard dihedral angles, θ and ϕ.[39,46] The coefficients V_0, V_1, and V_2 are used to control the strength of interactions and are set equal to 1, 5, and 1 energy units, respectively. The last term in the expression, $V_{constraints}$, is used to enforce a series of knowledge-based constraints whose violation is penalized through an "infinite" energy penalty (that is, through a rejection of the violating conformation). The constraints are as follows: (1) the distance between two consecutive C_α atoms must remain in the interval 3.7–3.9 Å; (2) the distance between two nonconsecutive C_α atoms must be greater than 4 Å; (3) the distance between any two C_β must be greater than 2 Å; and (4) the distance between any two C_α and C_β centroids must be greater than 2 Å.

The second term of the Hamiltonian, V_{ss} rewards the formation of disulfides. As explained in the next subsection, each proposed configuration, described in terms of C_α's and C_β's, comes with a set of three putative bonds among the six cysteines. Of these putative bonds, only those among residues whose C_β's are a separation smaller than 5 Å are considered to be effectively present and hence give a contribution equal to $-\mu$ to the total energy.

Monte Carlo Method

As mentioned, we used Monte Carlo dynamics for studying the folding process. At each Monte Carlo step, the current structure is distorted through local deformations[51] based on two, equally probable moves: (1) *single-bead move:* a random *alpha*-carbon is chosen and displaced randomly by, at most, 1 Å along each Cartesian direction; (2) *crankshaft move:* two sites, i and j, with sequence separation that is, at most, 6 are chosen, and all the sites between them are rotated around the axis joining i and j by an angle chosen randomly in the interval $-\pi/10 \leq \Omega \leq \pi/10$. As mentioned before, besides these structural rearrangements, at each Monte Carlo step, we also associate a randomly chosen pairing pattern for the six cysteines, so that each of them is involved in a *putative* disulfide bond. These bonds are termed *putative* because, to satisfy detailed balance, the pairing assignment is done blindly, that is, without inspecting whether a given pair of cysteines is at a distance compatible with the existence of a disulfide bond. One then checks whether (irrespective of the pairing distances) the proposed bonding pattern is compatible with the undistorted configuration (i.e., if it respects the rules of the section on comparison with experimental rates about disulfide formation and thiol/disulfide coupling). If not, the configuration is rejected, the Monte Carlo clock is advanced, and a new distortion and bonding pattern is considered. Otherwise, one proceeds as in ordinary Metropolis schemes, after having calculated the energy of the proposed configuration. It is important to notice that this latter step involves the inspection of the distance of the putatively bonded cysteines to reward only those bonds that are geometrically feasible.

The efficiency of the Monte Carlo algorithm to study the thermodynamics was enhanced by the multiple-Markov chain-sampling scheme,[52] a method that has proved quite effective in exploring the low-temperature phase diagrams of proteins and interacting polymers. All the runs have been performed by covering the temperature range of interest, $T = [0.2, 1.5]$, with 20 Markov chains uniformly spaced in temperature.

The data obtained in the multiple Markov chain runs at different values of T and μ were further processed through a generalized multiple-histogram technique inspired by the work of Ferrenberg and Swensen.[53] This strategy allowed us to reconstruct faithfully the density of states (number of configurations) in the multidimensional space of reaction coordinates constituted by V_{n-ss}, and the number of correct (native) and wrong (non-native) disulfides: (n_c, n_w). The data from the different runs were combined so as to minimize the error propagation on the density of states. The typical uncertainty of the reconstructed free energy at the values of T and μ considered here is about 0.5 energy unit (this estimate follows from the analysis of free-energy dispersion when half of the collected data is used). By these means, it was possible to obtain the total energy and specific heat curves in Figure 5 and the free-energy profiles in Figure 6.

CONCLUSIONS

We have proposed a theoretical framework to model and study the folding of proteins containing disulfide bonds. The approach is based on the knowledge of the native state of a protein but contains an appropriate term to account for the possibility that native or non-native disulfide bonds can form. The main advantage of the model proposed here is its simplicity, which allows for a detailed description of all the folding pathways through the monitoring of the correct–incorrect contact formation.

We validated the model by investigating the debated refolding pathways of hirudin, which has been the object of several experimental studies. It was shown that there exists a region in our 2D-parameter space where the rates of conversions between different oxidized species are in good agreement with experimental measurements.[34] Starting from this successful comparison, we have then attempted a detailed characterization of the whole folding process.

At a coarse-grained level, our results are consistent with the scenario described by Thannhauser et al.,[34] suggesting that the rate-limiting step turns out to be $2S \rightarrow 3S^*$. Our approach, which allow for a precise identification of the formed contacts, shows clearly that the $2S$ state appears to involve native intermediates, possibility reputed to be the most probable situation in Thannhauser et al.,[34] although, experimentally, it was impossible to show it directly. The analysis of several folding trajectories also allowed the identification of the most probable folding route and the typical associated succession of formation of disulfide bridges. Furthermore, we elucidated the thermodynamics of the system by using statistical mechanical techniques to recon-

struct the free-energy profiles for the whole system and also for the different oxidized species.

Due to its simplicity, the proposed model cannot capture those aspects of the folding process that result from the delicate interplay of amino acid specific interactions. The successful comparison of the theoretical predictions for hirudin with the experimental findings suggests that, also, for disulfide-containing proteins, suitable topology-based models can be profitably used to elucidate the folding pathways, even in the presence of non-native intermediates. Thus, this approach, which is general and not specifically tailored for hirudin, ought to be straightforwardly applicable in other contexts, providing a useful complement of experimental techniques in the characterization of the folding process in the presence of disulfide bonds.

ACKNOWLEDGMENTS

We thank F. Cecconi, G.L. Lattanzi and D. Marenduzzo for useful discussions.

REFERENCES

1. Anfinsen C. Principles that govern the folding of protein chains. Science 1973;181:223–230.
2. Creighton T. Proteins, structure and molecular properties. 2nd edition. New York: W.H. Freeman; 1993.
3. Creighton T. The disulfide folding pathway of BPTI. Science 1992;256:111–114.
4. Weismann J, Kim P. Reexamination of the folding of BPTI: Predominance of native intermediates. Science 1991;253:1386–1393.
5. Creighton T. The two-disulphide intermediates and the folding pathway of reduced pancreatic trypsin inhibitor. J Mol Biol 1975;95:167–169.
6. Weismann J, Kim P. Kinetic role of nonnative species in the folding of bovine pancreatic trypsin inhibitor. PNAS 1992;89:9900–9904.
7. Scheraga H, Konishi Y, Rothwarf D, Mui P. Toward an understanding of the folding of ribonuclease a. PNAS 1987;84:5740–5744.
8. Camacho C, Thirumalai D. Modeling the role of disulfide bonds in protein folding: Entropic barriers and pathways. Proteins 1995;22:27–40.
9. Camacho C, Thirumalai D. Theoretical predictions of folding pathways using the proximity rule with application to BTPI. PNAS 1995;92:1277–1281.
10. Abkevich V, Shakhnovich E. What can disulphide bond tell us about protein energetics, function and folding: Simulations and bioinformatics analysis. J Mol Biol 2000;300:975–985.
11. Thirumalai D, Klimov D, Dima R. Insights into specific problems in protein folding using simple concepts. Adv Chem Phys 2002;120:35–77.
12. Chan H, Kaya H, Shimizu S. In: Current topics in computational molecular biology. Cambridge, MA: MIT Press; 2002. p. 403–447.
13. Go N, Scheraga HA. On the use of classical statistical mechanics in the treatment of polymer chain conformations. Macromolecules 1976;9:535–542.
14. Micheletti C, Banavar JR, Maritan A, Seno F. Protein structures and optimal folding from a geometrical variational principle. Phys Rev Lett 1999;82:3372–3375.
15. Munoz V, Eaton W. A simple model for calculating the kinetics of protein folding from three-dimensional structures. Proc Natl Acad Sci USA 1999;96:11305–11310.
16. Alm E, Baker D. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. Proc Natl Acad Sci USA 1999;96:11305–11310.
17. Galzitskaya O, Finkelstein A. A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. Proc Natl Acad Sci USA 1999;96:11299–11304.
18. Maritan A, Micheletti C, Banavar JR. Role of secondary motifs in fast folding polymers: A dynamical variational principle. Phys Rev Lett 2000;84:3009–3012.
19. Baker D. A surprising simplicity to protein folding. Nature 2000;405:39–42.
20. Clementi C, Nymeyer H, Onuchic J. Topological and energetic factors: What determines the structural details of the transition state ensemble and en-route intermediates for protein folding? An investigation for small globular proteins. J Mol Biol 2000;298:937–953.
21. Orlandini E, Seno F, Banavar J, Laio A, Maritan A. Deciphering the folding kinetics of transmembrane helical proteins. PNAS 2000;97:14229–14234.
22. Cecconi F, Micheletti C, Carloni P, Maritan A. Molecular dynamics studies of HIV-1 protease: Drug resistance and folding pathways. Proteins 2000;43:365–372.
23. Hoang TX, Cieplak M. Sequencing of folding events in go-type proteins. J Chem Phys 2000;113:8319–8328.
24. Zhou YQ, Karplus M. Interpreting the folding kinetics of helical proteins. Nature 1999;401:400–403.
25. Shea J, Nochomovitz YD, Guo Z.Y., Brooks C.L. Exploring the space of protein folding Hamiltonians: The balance of forces in a minimalist beta-barrel model. J Chem Phys 1998;109:2895–2903.
26. Plotkin S. Speeding protein folding beyond the go model: How a little frustration sometimes helps. Proteins 2001;45:337–345.
27. Scaccheri E, Nitti G, Valsasina B, Orsini G, Visco C, Ferreia M, Sawyer RT, Sarmientos P. Eur J Biochem 1993;214:295–304.
28. Krezel AM, Wagner G, Seymourulmer J, Lazarus RA. Structure of the rgd protein dicorsin-conserved motifs and distinct function in leech proteins that affect blood-clotting. Science 1994;264:1944–1947.
29. Lin S, Nussinov R. A disulfide-reinforced structural scaffold shared by small proteins with diverse functions. Nat Struct Biol 1995;2:835–837.
30. Ascenzi P, Bolognesi M, Catalucci D, Pascarella S, Ruoppolo M, Rizzi M. Leech antihemostatic proteins share the T-knot scaffold, a disulfide-reinforced motif. Biol Chem 1998;379:1387–1389.
31. Otto A, Seckler R. Characterization, stability and refolding of recombinant hirudin. Eur J Biochem 1991;202:67–73.
32. Chatrenet B, Chang J. The disulfide folding pathway of hirudin elucidated by stop/go folding experiments. J Biol Chem 1993;268:20988–20996.
33. Filippis VD, Vindigni A, Altichieri L, Fontana A. Core domain of hirudin from the leec *Hirudinaria manillensis:* Chemical synthesis, purification and characterization of a trp^3 analog of fragment 1–47. Biochemistry 1995;34:9552–9564.
34. Thannhauser T, Rothwarf D, Scheraga H. Kinetic studies of the regeneration of recombinant hirudin variant 1 with oxidized and reduced dithiothreitol. Biochemistry 1997;36:2154–2165.
35. Nicastro G, Baumer L, Bolis G, Tato M. NMR solution structure of a novel hirudin variant hm2, n-terminal 1–47 and n64 → v + g mutant. Biopolymers 1997;41:731–749.
36. Go N, Abe H. Non interacting local structures: Model of folding and unfolding transition in globular protein: I. Formalism. Biopolymers 1981;20:991–1011.
37. Maiorov VN, Crippen GM. Contact potential that recognizes the correct folding of globular proteins. J Mol Biol 1992;227:876–888.
38. Thomas PD, Dill KA. An iterative method for extracting energy-like quantities from protein structures. Proc Natl Acad Sci U. S. A. 1996;93:11628–11633.
39. Micheletti C, Seno F, Banavar JR, Maritan A. Learning effective amino acid interactions through iterative stochastic techniques. Proteins 2001;42:422–431.
40. Chang I, Cieplak M, Dima R, Banavar J, Maritan A. Protein threading by learning. Proc Natl Acad Sci U. S. A. 2001;98:14350–14355.
41. Sokal AD. Monte Carlo methods for the self-avoiding walk. Nuclear Phys 1996;B47:172–179.
42. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical recipes. Cambridge, UK: Cambridge University Press; 1999.
43. Folkers P, Clore G, Driscoll P, Dodt J, Koheler S, Gronenborn A. Solution structure of recombinant hirudin and the lys-47→glu mutant: A nuclear magnetic resonance and hybrid distance geometry–dynamical simulated annealing study. Biochemistry 1989;28:2601–2617.

44. Szyperski T, Guntert P. Impact of protein–protein contacts on the conformation of thrombin-bound hirudin studied by comparison with the nuclear magnetic resonance solution structure of hirudin (1–51). J Mol Biol 1992;228:1206–1211.
45. Kelley L, Sutcliffe M. Olderado: On-line database of ensemble representatives and domains. Protein Sci 1997;12:2628–2630.
46. Park B, Levitt M. Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. J Mol Biol 1996; 258:367–392.
47. Chan HS, Dill KA. The effects of internal constraints on the configurations of chain molecules. J Chem Phys 1990;92:3118–3135.
48. Kaya H, Chan HS. Energetic components of cooperative protein folding. Phys Rev Lett 2000;85:4823–4826.
49. Kaya H, Chan HS. Polymer principles of protein calorimetric two-state cooperativity. Proteins 2000;40:637–661.
50. Settanni G, Cattaneo A, Maritan A. Role of native-state topology in the stabilization of intracellular antibodies. Biophys J 2001;81: 2935–2945.
51. Micheletti C, Seno F, Maritan A. Recurrent oligomers in proteins—an optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies. Proteins 2000;40:662–674.
52. Tesi M, van Rensburg EJ, Orlandini E, Whittington S. Monte-Carlo study of the interacting self-avoiding walk in three dimensions. J Stat Phys 1996;82:155–181.
53. Ferrenberg AM, Swendsen RH. Optimized Monte Carlo data analysis. Phys Rev Lett 1989;63:1195–1198.