PROTEINS: Structure, Function, and Genetics 51:74–84 (2003)

Prediction of Folding Rates and Transition-State Placement From Native-State Geometry

Cristian Micheletti*
International School for Advanced Studies (SISSA) and INFM, Trieste, Italy

ABSTRACT **A variety of experimental and theoretical studies have established that the folding process of monomeric proteins is strongly influenced by the topology of the native state. In particular, folding times have been shown to correlate well with the contact order, a measure of contact locality. Our investigation focuses on identifying additional topologic properties that correlate with experimentally measurable quantities, such as folding rates and transition-state placement, for both two- and three-state folders. The validation against data from 40 experiments shows that a particular topological property that measures the interdependence of contacts, termed cliquishness or clustering coefficient, can account with statistically significant accuracy both for the transition state placement and especially for folding rates. The observed correlations can be further improved by optimally combining the distinct topological information captured by cliquishness and contact order. Proteins 2003;51:74–84.**

Key words: folding rates prediction; two-state folding kinetics; protein topology; stochastic simulations

INTRODUCTION

In the past three decades, there has been a growing effort of the scientific community for studying and understanding the principles that govern the folding process of a sequence of amino acids in the corresponding native structure.[1–3] In recent years, several proteins, in particular those folding via a two-state mechanism,[4] have provided an extraordinary benchmark for experimental and theoretical characterization of the folding pathways. The significant amount of experimental data available for several structurally unrelated proteins[4–24] has opened the possibility of identifying and isolating the factors that influence the folding rate. Besides considering detailed chemical interaction, such as those affecting free-energy barriers, an appealing and elegant line of investigation has focused on the effects of the native state structure on the folding process.[25–31]

From a qualitative point of view, the influences of structural effects was traditionally summarized in the tenet that proteins with high helical content fold faster than proteins with mixed α/β content, the slowest folding being for the all-β ones. This useful and intuitive rule of thumb fails to account for the very different rates observed between proteins in each of the α, α/β, or β-families.[5,32–34] A deep insight into this problem was provided by the work of Plaxco et al.[25] who introduced the concept of contact order, which captures, quantitatively, features beyond the mere secondary structure motifs. The highly significant correlation of contact order and experimental folding rates shows the extent to which the mere topology of native state can influence the folding process. However, the highly organized native structure of proteins is too rich to be captured by a single parameter such as the contact order. Indeed, the latter cannot account in the same satisfactory way for the transition-state placement, three-state folding rates, or the diversity of folding rates among structurally similar proteins.[35]

In the present study, we investigate how the topology of the native state can be further exploited to provide optimized predictions for protein folding rates and the transition-state placement. To do so we consider one particular topological descriptor that is crucial for characterizing the connection and interactions of native contacts: the clustering coefficient, or cliquishness. This parameter has been originally introduced in the context of graph theory.[36–38] As other graph properties, such as the distance and connectivity of nodes,[31] also the cliquishness is a useful descriptor of protein topology and is shown to have highly significant correlation with folding rates. The advantage of using this topological descriptor is that it allows one to capture the cooperative formation of native interactions, as proved by its statistically relevant correlations with the transition-state placement. Furthermore, we discuss how the different topological aspects captured by the cliquishness and contact order can be combined to yield optimal correlations higher than for the individual descriptors.

THEORY AND RESULTS

Customarily, at the heart of theoretical or numerical studies of topology-based folding models is the contact matrix (or map),[39] which will be used extensively also in the present context. The generic entry of the contact map, Δ_{ij}, takes on the value 1 if residues i and j are in contact and zero otherwise. Several criteria can be adopted to define a contact; in the present study, we consider two

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

Received 28 August 2002; Accepted 28 October 2002

amino acids in interaction if any pair of heavy atoms in the two amino acids are at a distance below a certain cutoff, d. All values of d between 3.5 and 6.5 Å have been considered and reported. The contact map provides a representation for the spatial distribution of contacts in the native structures that is both concise and often reversible (because native structures can be recovered when appropriate values of d are used). Plaxco and coworkers[25] have used the contact map to describe and characterize the presence and organization of secondary motifs in protein structures. The parameter that was introduced, the relative contact order, provides a measure of the average sequence separation of contacting residues and is defined as

$$\text{relative contact order} = \frac{1}{L}\frac{\sum_{i\neq j}\Delta_{ij}|i-j|w_{ij}}{\sum_{i\neq j}\Delta_{ij}w_{ij}}, \qquad (1)$$

where i and j run over the sequence indeces, w_{ij} is the contact degeneracy (i.e., the number of pairs of heavy atoms in interaction), and L is the protein length. Remarkably, the contact order was shown to have a highly significant linear correlation with experimental folding rates. The result of Plaxco and coworkers can be explained with intuitive arguments: a high contact order corresponds to few local interactions. One may thus expect that the route from the unfolded ensemble to the native state is slow, being hindered by the overcoming of several barriers[40–45] due to spacial restraints, as recently analyzed by Debe and Goddard[26] and previously by Chan and Dill[46] and also observed in topology-based numeric studies.[47] These considerations are based purely on geometric arguments and do not take into account the influence of specific interactions between the residues. In principle, the latter may well override the topological influence on the folding process, but surprisingly, as remarked in a recent review article,[48] this is often not the case.[47–58] A remarkable example is provided by structurally homologous proteins which, despite large variation of their primary sequences, share the same transition state.[48] Of course, although native topology can indicate an "ideal" reference folding rate, the ultimate real folding process is the result of the complex interplay of amino acid and solvent interactions, which may be sensitive even to single-point mutation of a few key sites.[4,59,60]

Our aim is to exploit as much as possible the topological information contained in the native state to improve both the accuracy of predictions for folding rates and gain more fundamental insight into the process. To this purpose, we have considered additional topological descriptors besides the contact order. The one that appeared most significant is a parameter termed cliquishness or clustering coefficient.[36–38] For a given site, i, the cliquishness is defined as:

$$\text{cliquishness}(i) = \frac{\sum_{j<l}\Delta_{ij}\Delta_{il}\Delta_{lj}}{\sum_{j<l}\Delta_{ij}\Delta_{il}} = \frac{\sum_{j<l}\Delta_{ij}\Delta_{il}\Delta_{lj}}{N_c(N_c-1)/2}, \qquad (2)$$

where N_c is the number of contacts to which site i takes part. As for the contact order, also the cliquishness has an intuitive meaning; in fact, it provides a measure of the extent to which different sites interacting with i are also interacting with each other. Of course, the cliquishness is properly defined only if site i is connected to, at least, two other sites. To ensure this, we included also the covalently bonded interactions $[i, i \pm 1]$ in Eq. 2. The importance of taking the cliquishness into account for discriminating fast/slow folders can be anticipated because a higher interdependency of contacts (large cliquishness) will likely result in a more cooperative folding process. In fact, the formation of a fraction of interactions will result in the establishment of a whole network of them. Consistently with this intuitive picture one should also expect that a large/small cliquishness will affect in different ways the amount of native-like content of the transition state.

We have tested and verified these expectations by calculating the average cliquishness for 38 distinct proteins for which folding rates and transition state placement, θ_m, have been measured. θ_m is deduced from the variation of folding/refolding rates on change of denaturant concentration and provides an indirect indication of how much the solvent-exposed surface of the transition state is similar to that of the native one. θ_m ranges between 0 and 1; higher values denote stronger similarity with the native state. It is worth pointing out that, although the model underlying the calculation of θ_m relies on a two-state analysis, an effective θ_m can be inferred for three-state folders as well.[5] Because reliable θ_m values are not available for all proteins, the number of entries used to correlate the cliquishness and θ_m (see Tables I and II) is slightly smaller than that used for the logarithm of estimated refolding rates in water, ln K_F. The set of proteins used, shown in Tables I and II, was built up from experimental data collected in previous studies and predictions (often topology-based) of folding rates.[5,25–30] The wide range of folding rates within each of the α, β, and αβ families leads to a substantial overlap between the folding rates of these classes and, hence, rules out the possibility that meaningful correlations against ln K_F can be obtained by a simple measure of the type of secondary content.

As indicated, the entries include both two-state and three-state folders. For a small number of entries (1hrc, 2rn2, 1hng) there exists more than one determination of the folding rates in water and/or θ_m. In this case, for the analysis of the correlation against topological descriptors, we used the average value of the corresponding experimental quantity as done in Ref. 28.

As discussed in detail below, when the comprehensive set of Tables I and II is used, one observes a highly significant correlation between cliquishness and folding rates. The existence of an intimate relationship between ln K_F and cliquishness is ascertained in two distinct ways. First, we calculate the standard linear correlation coefficient, r, as well as its statistical significance $P(r)$ by means of the Student's t-test; see Ref. 61. Such statistical test allows one to calculate the probability, $P(r)$, to observe a correlation higher than r (in modulus) by pure chance (i.e.,

TABLE I. List of Proteins Known to Fold via a Two-State Mechanism

Protein	Length	Family	$\ln K_F^{H_2O}$	θ_m
2pdd [27]	43	α	9.80	—
1lmb [6, 7]	80	α	8.50	0.46
2abd [8, 9]	86	α	6.55	0.61
1imq [26]	86	α	7.31	—
1ycc [10]	103	α	9.61	0.34
1hrc [78]	104	α	7.94	0.47
1hrc, horse, oxidized Fe^{III} [78]	104	α	5.99	0.40
2gb1 [21]	56	α/β	6.26	—
1div.n [26]	57	α/β	6.58	—
2ptl [16]	60	α/β	4.22	0.75
1coa [24]	64	α/β	3.87	0.61
1hdn [18]	85	α/β	2.70	0.64
1div.c [26]	93	α/β	1.15	—
1urn [17]	96	α/β	5.73	0.55
1aps [19]	98	α/β	−1.47	0.79
1fkb [5]	107	α/β	1.46	0.67
2vik [5]	126	α/β	6.80	0.73
1shg [79]	57	β	2.10	0.69
1srl [13]	56	β	4.04	0.69
1shf.a [80]	59	β	4.55	0.68
1tud [26]	60	β	3.45	—
1csp [81]	67	β	7.00	0.85
3mef [12]	69	β	5.30	0.94
2ait [11]	74	β	4.20	0.65
1pks [82]	76	β	−1.05	0.60
1ten [83]	89	β	−1.10	0.76
1fnf, 9FN-III [35]	90	β	−0.90	0.63
1wit [35]	93	β	0.41	0.70
1fnf, 10FN-III [35]	94	β	5.00	0.65

The experimental quantities K_F (s^{-1}) and θ_m are desumed from the cited references.

TABLE II. List of Proteins Known to Fold via a Three-State Mechanism

Protein	Length	Family	$\ln K_F^{H_2O}$	θ_m
1bta	89	α	3.40	0.87
1ubq	76	α/β	5.90	0.59
1bni	108	α/β	2.60	0.88
1hel	129	α/β	1.30	0.75
3chy	128	α/β	1.0	0.71
1dk7	146	α/β	0.80	0.78
2rn2, Urea, pH 5.5	155	α/β	−0.50	0.80
2rn2, GdnHCl, pH 5.5	155	α/β	1.40	0.63
1php.n	175	α/β	2.30	0.84
1php.c	208	α/β	−3.5	0.45
1hng, pH 7.0	97	β	1.80	—
1hng, pH 4.5	97	β	2.63	—

The structural families and experimental quantities K_F (s^{-1}) and θ_m are desumed from Ref. 5.

if the two quantities under study were statistically independent). Alternatively, $P(r)$ is the probability to observe a correlation higher than r if the quantities were statistically independent. Clearly $P(r)$ depends not only on r but also on the number of entries over which the correlation is measured. As a rule of thumb, the upper value of $P(r) = 0.05$ is taken as a threshold for statistically meaningful correlations.

Strictly speaking, the linear correlation analysis should be applied to cases where the joint distribution of the two variables of interest (e.g., $\ln K_F$ and contact order) is binormal. The limited number of data at disposal is not sufficient to fully corroborate this hypothesis. This problem, however, can be easily circumvented by using a non-parametric measure such as the Kendall coefficient, τ. In such analysis, it is entirely superfluous to know the original distribution from which the points (pairs of data) are taken; what matters is the ranking of the points according to each of the two variables. Kendall's parameter τ measures, in the data set, the degree of accord of the two ranking possibilities. Kendall's τ [which takes on the values of 1 (−1) for full (anti-)correlation] is "a more robust measure than linear correlation and resistant to unplanned defects in the data."[62] Also for this case there exists a definite way to evaluate the statistical significance of τ. We shall denote by $P(\tau)$ the probability to obtain, by pure chance, a value of τ higher, in modulus, than the observed one.

We have found, a posteriori, a good consistency among the the statistical significances of the linear and non-parametric correlations in the different cases under study. Indeed, the two types of analysis highlight the strong dependence of folding rates and cliquishness for the set comprising both two- and three-state folders. The probability that the observed correlation is fortuitus is as low as 10^{-5} (for both types of analysis), comparable to the statistical significance of the correlation between folding rates and a suitably defined contact order. The predicting power of the two quantities can be combined to achieve the optimal linear correlation coefficient of 0.73 over the combined set of 38 entries of Tables I and II.

The prediction of the transition-state placement turns out to be more difficult when either of the two topological parameters is used. In this context, the cliquishness appears to be better than absolute and relative contact order yielding correlations $r = 0.66$ and $\tau = 0.50$ (with a probability of <0.1% to have occurred by chance.) The significance of the correlation against θ_m is preserved on the addition of the three-state folders, although it diminishes to 5% because of the great difficulty of capturing the transition-state placement of three-state folders. This may possibly also reflect the difficulty to provide an effective θ_m for multiple-state folders. By comparison, the two-sided significance of the best non-parametric correlation between θ_m and a contact-order measure is only 31%.

Two-State Folders

Before considering the more general case of all entries in Tables I and II, we focus on two-state folders [i.e., proteins with a cooperative (all-or-none) transition between the unfolded and folded states]. The neatness of this process, due to the absence of any significantly populated intermediate state, makes them ideal candidates for identifying and isolating the factors that influence the folding rate.

TABLE III. Linear (r) and Non-parametric (τ) Coefficients for the Correlation Found Among the Different Topological Descriptors and Folding Rates or Transition-State Placement

Protein set	Variables	r	$P(r)$	τ	$P(\tau)$
2-state	Cliquishness $-\ln K_F$	0.60	7×10^{-4}	0.42	2×10^{-3}
2-state	Relative contact order $-\ln K_F$	−0.75	4×10^{-6}	−0.58	1×10^{-5}
2-state	Absolute contact order $-\ln K_F$	−0.70	3×10^{-5}	−0.46	6×10^{-4}
2-state	Cliquishness $-\theta_m$	−0.66	9×10^{-4}	−0.50	1×10^{-3}
2-state	Relative contact order $-\theta_m$	0.44	0.04	0.23	0.14
2-state	Absolute contact order $-\theta_m$	0.20	0.38	0.11	0.46
3-state	Cliquishness $-\ln K_F$	0.64	0.04	0.56	0.02
3-state	Relative contact order $-\ln K_F$	0.56	0.09	0.38	0.13
3-state	Absolute contact order $-\ln K_F$	−0.36	0.31	−0.33	0.18
3-state	Cliquishness $-\theta_m$	−0.44	0.23	−0.39	0.14
3-state	Relative contact order $-\theta_m$	0.50	0.17	0.33	0.21
3-state	Absolute contact order $-\theta_m$	0.15	0.70	0.17	0.53
2- and 3-state	Cliquishness $-\ln K_F$	0.67	4×10^{-6}	0.49	2×10^{-6}
2- and 3-state	Relative contact order $-\ln K_F$	−0.26	0.12	−0.23	0.04
2- and 3-state	Absolute contact order $-\ln K_F$	−0.66	7×10^{-6}	−0.45	7×10^{-5}
2- and 3-state	Cliquishness $-\theta_m$	−0.43	0.02	−0.25	0.05
2- and 3-state	Relative contact order $-\theta_m$	0.20	0.28	0.06	0.61
2- and 3-state	Absolute contact order $-\theta_m$	0.21	0.27	0.13	0.31

The statistical significance of the correlations are given by $P(r)$ and $P(\tau)$. The data are reported for the two-state folders of Table I, the three-state folders of Table II, and the combined set.

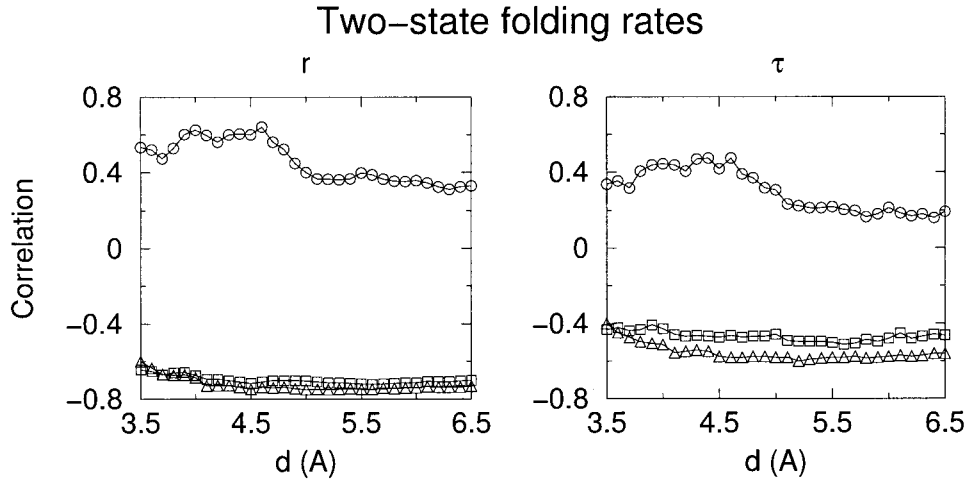

Fig. 1. Linear (**left**) and non-parametric (**right**) correlation of cliquishness, relative, and absolute contact order against folding rates of two-state folders (the entries of Table I were used). The values of the correlation coefficients are plotted as a function of the cutoff, d (Å), used in the definition of the contact map. The open circles, squares, and triangles are used for cliquishness, absolute, and relative contact order, respectively.

In the present context, this separate test is important because it appears that the relative contact order is a much stronger descriptor for two-state folders than for the general case. As a matter of fact, when both two- and three-state folders are considered, the influence of the average sequence separation of native contacts on folding properties is better captured by a different version of the contact order, which we shall term "absolute," obtained when the right-hand side of Eq. 1 is not divided by the protein length, L:

$$\text{absolute contact order} = \frac{\sum_{i\neq j} \Delta_{ij}|i-j|w_{ij}}{\sum_{i\neq j} \Delta_{ij} w_{ij}}. \qquad (3)$$

In the following, we report and compare the performance of both parameters; furthermore, we often consider the absolute value of the linear correlation coefficient $|r|$, or its non-parametric counterpart $|\tau|$, without regard to their sign, which can be easily inferred from the plots and Table III.

We have reported the correlation of two-state folding rates with contact order and cliquishness in Figure 1. For both families of results, the optimal correlation is found for cutoffs around 4.5–5.0 Å. In an unbiased estimate of the statistical significance of correlation, one should consider the value of r for a cutoff value, d, determined *a priori*. If one had at disposal a large number of entries, one could sacrifice a portion of them to determine the optimal value

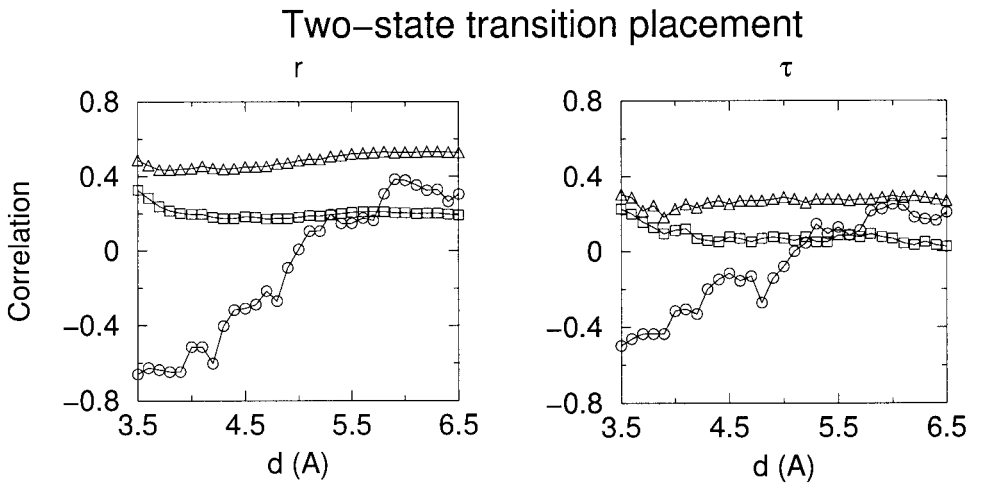

Fig. 2. Linear (**left**) and non-parametric (**right**) correlation of cliquishness, relative, and absolute contact order against transition-state placement of two-state folders (the entries of Table I were used). The values of the correlation coefficients are plotted as a function of the cutoff, *d* (Å), used in the definition of the contact map. The open circles, squares, and triangles are used for cliquishness, absolute, and relative contact order, respectively.

of d for processing the remaining data. However, the limited number of experimental entries in Tables I and II does not allow us to follow this route. Hence, to estimate the degree of correlation with a simple rule, we measured r and τ for $d = 4.5$ Å when dealing with cliquishness and $d = 5.0$ Å for contact order. The same values of d are used for the entries of Table II and the joint set of two- and three-state folders. The corresponding values of r, τ, and their significance are reported in Table III.

As visible in Figure 1, the original definition of contact order has an unrivaled performance in the prediction of folding rates for the two-state folders of Table I. The observed value of $r = 0.75, \tau = 0.58$ reported in Table III have a probability of 10^{-5} to have occurred by chance and are, therefore, extremely significant. On the other hand, the absolute contact order has a weaker performance, $r = 0.70, \tau = 0.46$. In this particular situation, the cliquishness yields a correlation $r = 0.60, \tau = 0.42$, which is the worst performance in the set, although still very significant.

By recoursing to the Fisher's z-transformation,[61] it is possible to ascertain how likely it is that the difference observed in two distinct measures of the linear correlation coefficient $|r|$ is a mere product of statistical fluctuations (because the signs of the two correlations are known, we shall use the one-sided significance). This probability is about 15% when comparing cliquishness and relative contact order, which supports the latter as as the best topological descriptor of two-state folding rates.

Consistent with previous results, we found that the transition-state placement is a much more elusive quantity to predict than folding rates. In fact, all topological descriptors yield a poorer correlation than $\ln K_F$ (see Fig. 2). It is apparent from Figure 2 that the values of d for which the optimal correlations are observed are lower than for the folding rates case. To reflect this, the average correlations against θ_m were taken for $d = 3.5$ Å for the case of cliquishness and $d = 4.0$ Å for the case of contact order.

For the relative contact order $r = 0.44, \tau = 0.23$ with a statistical significance $P(r) = 0.04$, $P(\tau) = 14\%$; the absolute contact order turns out to be a poorer descriptor of this further property of two-state folders. In this more challenging case, the cliquishness appears to capture the physical parameter, θ_m, much more faithfully. The values $r = 0.66$ and $\tau = 0.50$ both have a probability $<10^{-3}$ to occur by chance, with a significant improvement over the previous cases. In fact, the probability that the linear correlation improvement over the relative contact order is fortuitus is 16% and is even less (6%) against the absolute contact order. This helps to establish in a quantitative way how much the prediction of θ_m for two-state folders can be improved by introducing this additional topological descriptor. As discussed in the next subsection, the improvement persists when three-state folders are included in the test.

We conclude the analysis of two-state folders by pointing out that cliquishness-based correlations have a non-trivial dependence on the cutoff d. This reflects the fact that the measure of cliquishness involves third moments of the residue contact matrix Δ and is, therefore, more sensitive to cutoff changes than contact order, which involves only the first moment of Δ.

Two- and Three-State Folders

It is interesting to apply the same analysis to proteins known to fold via a three-state mechanism. For this class of proteins, the amount of precise experimental characterizations is more limited than two-state folders. We have analyzed the correlation of the topological descriptors against $\ln K_F$ for the 10 entries of Table II. In this context, the most reliable descriptor appears to be the cliquishness, which yields correlations equal to $r = 0.64$; $\tau = 0.56$, which have significances of $P(r) = 0.04$ and $P(\tau) = 0.02$, respectively. This result underscores that, although the

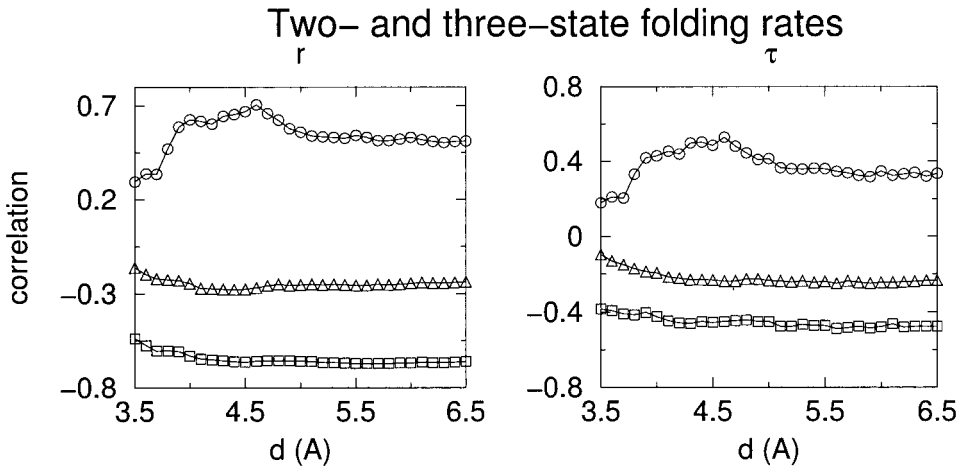

Fig. 3. Linear (**left**) and non-parametric (**right**) correlation of cliquishness, relative, and absolute contact order against folding rates of two- and three-state folders. The values of the correlation coefficients are plotted as a function of the cutoff, *d* (Å), used in the definition of the contact map. The open circles, squares, and triangles are used for cliquishness, absolute, and relative contact order, respectively.

performance has degraded with respect to two-state folders, where $P \approx 10^{-3}$, it is still very meaningful and reliable. The most significant change of performance is, however, observed for contact-order-based measurements. The case is particularly dramatic for the relative contact order, where the sign of the correlation has changed with respect to the two-state case. From this crucial fact one may anticipate that, when dealing with a set comprising both two- and three-state folders, the performance of the relative contact order will be poor because of the conflicting trends in the two subsets. The consistency of the correlation sign is preserved, on the other hand, by the absolute contact order. The associated absolute correlation values are $r = 0.36$, $\tau = 0.33$; the significance of the robust non-parametric dependence is 18%. This value suggests the absence of a meaningful correlation between $\ln K_F$ and contact order for three-state folders. However, it should be borne in mind that $P(r)$ and $P(\tau)$ capture the double-sided significance. If one knows *a priori* the *correct* sign of the correlation (e.g., by using arguments as in the beginning of Theory and Results), then it is appropriate to use the one-sided significance; accordingly, the values of P are halved. In this case $P(\tau) \approx 10\%$ for the absolute contact order, which may be regarded at the border or statistical significance given the limited number of entries in Table II. For relative contact order, the two-sided significance shows an improvement over the absolute case, $\tau = 0.38, P(\tau) = 13\%$. However, the sign of the correlation is the opposite to what expected *a priori* or, for example, from the two-state case. The one-sided analysis, therefore, rejects the hypothesis of any significant *anti*correlation between relative contact order and three-state folding rates. In summary, the linear and non-parametric analysis highlights the large change of performance of contact order from two- to three-state folders; the cliquishness appears to be more stable and reliable.

We further considered the case of θ_m. Given the fact that the prediction of the transition-state placement was elusive even for the "simpler" case of two-state folders, one may anticipate that, for three-state folders, the difficulty will be augmented. Indeed, this is the case: the cliquishness yields $r = 0.44$ and $\tau = 0.39$, whereas for the absolute contact order $r = 0.15$ and $\tau = 0.17$. The two-sided non-parametric statistical significance of the former is $P(\tau) = 0.14$. The one-sided significance appears to be marginally significant for cliquishness, $P \approx 7\%$, whereas it is at least twice as much for the contact order-based measurements. For this case, as for the relationship between absolute contact order and $\ln K_F$, it would be useful to take into account future additional experimental data to reach a definite conclusion about the existence/absence of a significant correlation.

We now turn to the more interesting case of the combined set of two- and three-state folders. Despite the addition of the 10 entries corresponding to three-state folders, the performance of cliquishness-based predictions for folding rates and θ_m retain their significance. In fact, as shown in Figures 3 and 4, the associated non-parametric correlations for $\ln K_F$ and θ_m are $\tau = 0.49$ and $\tau = 0.25$, respectively, again observed for the same contact order ranges reported for the two-state case. The corresponding statistical significances are now $P(\tau) = 2 \times 10^{-6}$ and $P(\tau) = 5\%$. The linear analysis yields almost exactly the same values.

From Figures 3 and 4, it can be noticed that the performance of the relative contact order is noticeably poorer than the absolute one, which, therefore, becomes the focus of our analysis. The corresponding non-parametric correlation is $\tau = 0.45$ with a significance of 7×10^{-5}.

Although folding rates appear to be predicted by cliquishness and absolute contact order, with comparable accuracy the difference concerning θ_m has become more dramatic because the latter yields $\tau = 0.13$ and $P(\tau = 0.38)$. Fisher's z transformation shows that there is $<$ a 16% chance that improvement in the linear correlation coefficient obtained with cliquishness over the best contact order case is

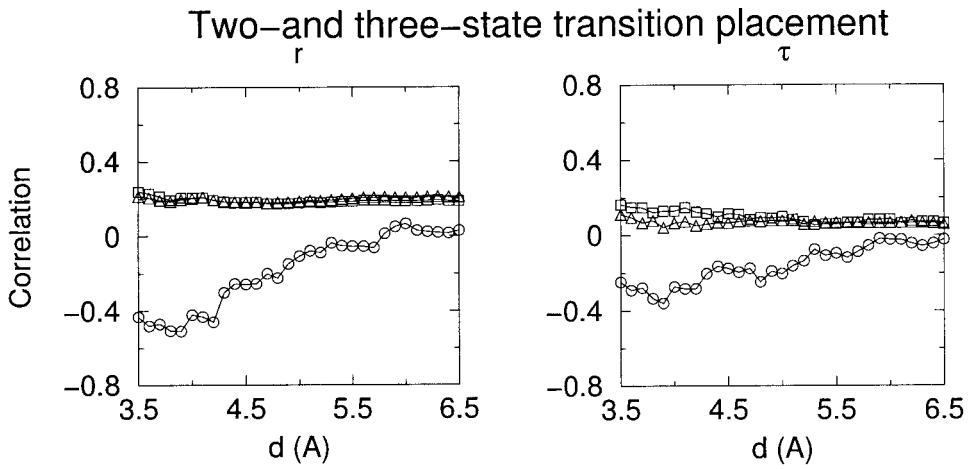

Fig. 4. Linear (**left**) and non-parametric (**right**) correlation of cliquishness, relative, and absolute contact order against transition-state placement of two- and three-state folders. The values of the correlation coefficients are plotted as a function of the cutoff, *d* (Å), used in the definition of the contact map. The open circles, squares, and triangles are used for cliquishness, absolute, and relative contact order, respectively.

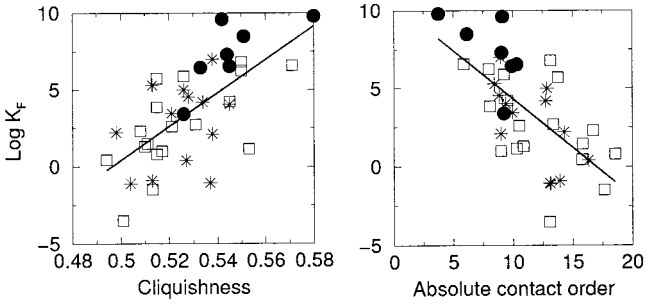

Fig. 5. Scatter plot of cliquishness (**left**) and absolute contact order (**right**) versus folding rates of the 38 entries of Tables I and II. The cutoff values used for cliquishness and absolute contact order were $d = 4.5$ Å and $d = 5.0$ Å, respectively. Filled circles, open squares, and starred points denote proteins belonging to the α, α/β, and β families, respectively.

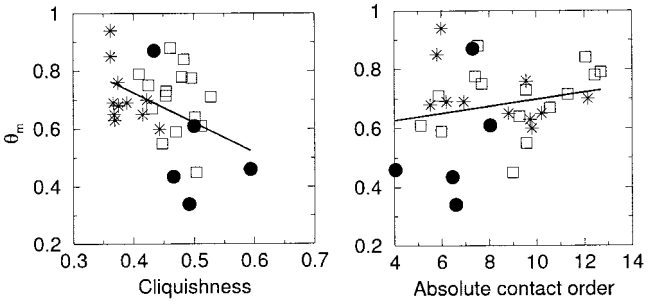

Fig. 6. Scatter plot of cliquishness (**left**) and absolute contact order (**right**) versus θ_m of the entries of Tables I and II. The cutoff value was chosen in the neighborhood maximum performance of cliquishness ($d = 3.5$ Å) and contact order $d = 4.5$ Å. Filled circles, open squares, and starred points denote proteins belonging to the α, α/β, and β families, respectively.

accidental (analogous to the significance of the improvement of relative contact order against cliquishness for two-state folders).

A direct comparison of how the clustering coefficient and the absolute contact order correlate with $\ln K_F$ and θ_m can be made by inspecting the plots of Figures 5 and 6. Although for both folding parameters the cliquishness gives a more significant correlation than contact order, the difference is particularly dramatic for the transition-state placement, which is notoriously difficult to capture with topology-based predictions.[25]

An important conclusion stemming out of this observation is that the transition-state structure (and hence θ_m) is more influenced by the degree of interdependency of native contacts than their average sequence separation. This is in accord with the intuition that highly interdependent contacts may mutually enhance their probability of formation, thus facilitating the progress toward the native state during the folding process. This is, indeed, consistent with the negative correlation observed between cliquishness and native content, θ_m, at the transition state. It is important to stress that the presence and effects of the cooperative formation of native interactions are captured with considerable difficulty by parameters based on measures of contact locality. This highlights the importance of considering all viable topological descriptors to characterize the folding process, because they do not impact in the same way on various folding properties.

Optimal Combined Correlation

A natural question that arises is whether it is possible to combine the predicting power of cliquishness and contact order to achieve correlations with experimental folding rates and transition-state placements that are better than the individual cases.

Indeed, as shown in the Methods section, it is straightforward to combine in an optimal linear way the two quantities to improve the prediction accuracy. The quantitative increment in the correlation is clearly related to the amount of independent information contained in the two topological descriptors. Hence, an important issue is to what extent cliquishness and contact order are mutually correlated.

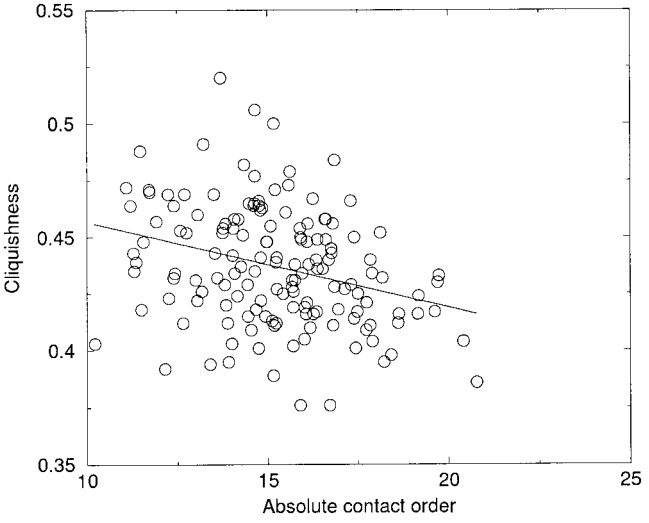

Fig. 7. Scatter plot of average cliquishness versus absolute contact order, for randomly collapsed structures generated by stochastic numerical methods.

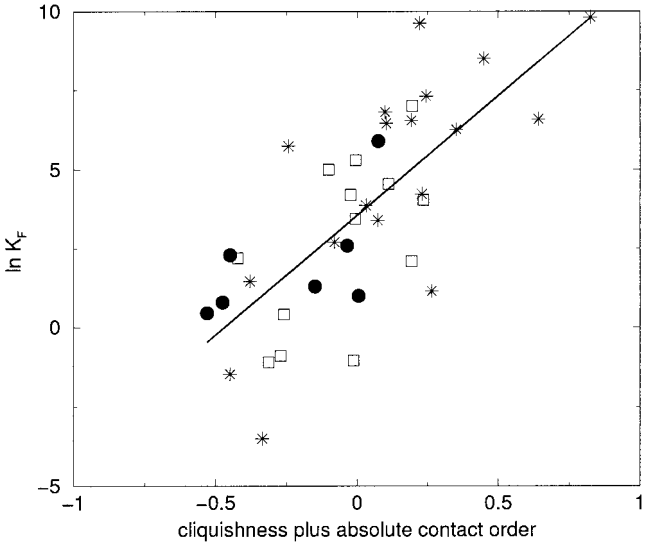

Fig. 8. Scatter plot of the logarithm of folding rates for the entries of Tables I and II, against data from optimally combined cliquishness and contact order. The optimal linear superposition, see Methods, is obtained for $b = 0.7$, $\{x\}$ and $\{y\}$ being the cliquishness and contact order data, respectively. Filled circles, open squares, and starred points denote proteins belonging to the α, α/β, and β families, respectively.

If, in place of a physical contact map, Δ_{ij}, one uses a random symmetric matrix, no meaningful correlation will be found. The contact maps of real proteins, however, display features that are highly non-random and reflect both (i) the physical constraints to which a compact three-dimensional chain is subject and (ii) the presence and organization of secondary motifs.[63–66]

With the aid of numeric simulations it is possible to assess the degree of interdependency of clustering coefficient of native contacts and their average sequence separation resulting from the first of the mentioned effects. This is accomplished by considering, in place of the proteins of Tables I and II, 150 computer-generated compact structures respecting basic steric constraints found in real proteins (details can be found in the Methods section). As visible in the plot of Figure 7 the level of mutual contact order-cliquishness correlation observed in these artificial structures is $r = 0.25$, which is significantly smaller than the actual correlation of the two quantities found in real proteins. In fact, the typical correlation for cliquishness and contact order (either relative or absolute) is around 0.65. Such non-trivial correlation can been ascribed to the special topological properties of naturally occurring proteins whose ramifications have been investigated in a variety of contexts[33,49,67–72] also with the aid of concepts developed in graph theory, as done in the present study.[31,73] In particular, the picture presented here, where folding rates are significantly affected by the native spatial organization of secondary elements is connected with the possible existence of a hierarchical protein assembly during the folding process.[72,74,75] Thus, the very presence and organization of secondary motifs in proteins makes it possible, on one hand, to exploit the native topology to predict, for example, folding rates, but on the other hand, it limits the amount of independent information contained in different topological descriptors.

Nevertheless, because the mutual correlation is not perfect, it is still possible to achieve, by definition, better predictions by combining cliquishness and contact order. The degree of enhancement depends also on the statistical significance of the individual starting correlations. For these reasons, the improvement is noticeable for folding rates, whereas it is not significant for transition-state placement. For the case of two-state folders, the optimal combination of cliquishness and relative contact order (see Methods section) yields a correlation of $r = 0.79$. For the more general case of two and three-state folding rates, by using cliquishness and absolute contact order, one has $r = 0.73$, which leads to a discernible improvement over previous cases, as visible in Figure 8. Because the optimal combined correlations are found *a posteriori,* the associated values of $P(r)$ are no more meaningful indicators of statistical significance.

CONCLUSIONS

We have analyzed important topological descriptors of organized networks (in our case, the spatial network of native contacts) that could be used, individually, or in mutual conjunction, to describe and predict experimental parameters used to characterize the folding process. It is found that, besides the previously introduced contact order, a topological parameter, termed cliquishness or clustering coefficient, is a powerful indicator of both the folding velocity and the transition-state placement for two- and three-state folders. The predicting power of the cliquishness is that it takes into account the presence and organization of clusters of interdependent contacts that are putatively responsible for the cooperative formation of native-like regions. This property appears well suited to reproduce important features in the transition state that are otherwise elusive to other topological analysis. The high statistical significance of the observed correlations

testifies to the strong influence of geometrical structural issues on the folding process. The maximum predicting power is obtained when the topological information contained in the cliquishness is used in combination with the contact order; this allows one to reach a linear correlation as high as 0.73 with experimental folding rates recorded in about 40 experimental measurements.

METHODS

Cross-Correlations

The linear correlation between two sets of data, $\{x\}$ and $\{y\}$, is obtained from the normalised scalar products of the covariations:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_j (y_j - \bar{y})^2}} \quad (4)$$

Without loss of generality, in the following we consider the sets of data to be with zero average and with unit norm, so that the expression of the correlation simplifies

$$r = \vec{x} \cdot \vec{y} \quad (5)$$

We now formulate the following problem. Two sets of data, $\{x\}$ and $\{y\}$, have linear correlation r_x and r_y, respectively with a third (reference set), $\{z\}$. What is the maximum and minimum correlations we can expect between sets $\{x\}$ and $\{y\}$? We assume that r_x and r_y are positive because this condition can always be met by changing sign, if necessary, to the vector components.

The answer is easily found by decomposing $\{x\}$ and $\{y\}$ into their components parallel and orthogonal to $\{z\}$:

$$\vec{x} \cdot \vec{y} = b_{\|} c_{\|} + \vec{b}_{\perp} \cdot \vec{c}_{\perp} \quad (6)$$

Because $b_{\|} c_{\|}$ is equal to $r_x r_y$, and hence is fixed, the maximum [minimum] correlation is found when $\vec{b}_{\perp}$ and $\vec{c}_{\perp}$ are [anti]parallel. Thus,

$$r_x r_y - \sqrt{(1 - r_x^2)(1 - r_y^2)} \le r \le r_x r_y + \sqrt{(1 - r_x^2)(1 - r_y^2)} \quad (7)$$

Now we turn to a different, but related problem. How can we combine linearly $\{x\}$ and $\{y\}$ to have the maximum correlation with $\{z\}$. The generic linear combination,

$$\vec{k} = \frac{\vec{x} + b\vec{y}}{\sqrt{1 + b^2 + 2b\vec{x} \cdot \vec{y}}} \quad (8)$$

leads to the following correlations

$$\vec{k} \cdot \vec{z} = \frac{r_x + b r_y}{\sqrt{1 + b^2 + 2b\vec{x} \cdot \vec{y}}} \quad (9)$$

The maximal is achieved for

$$b = \frac{r_y - \vec{x} \cdot \vec{y} r_x}{r_x - \vec{x} \cdot \vec{y} r_y} \quad (10)$$

which yields

$$\mathrm{Max}(\vec{k} \cdot \vec{z}) = \sqrt{\frac{r_x^2 + r_y^2 - 2\vec{x} \cdot \vec{y} r_x r_y}{1 - (\vec{x} \cdot \vec{y})^2}} \quad (11)$$

Generation of Alternative Compact Structures

To generate the 30 randomly collapsed structures used in the comparison of Figure 8, we adopted a Monte Carlo technique. The length of the artificial proteins ranged uniformly in the interval 80–110. Starting from an open conformation, each structure was modified under the action of typical MC moves (single-bead, crankshaft, pivot).[76] A newly generated modified configuration is accepted according to the ordinary Metropolis rule. The energy-scoring function is composed of two terms. The first one contains a homopolymeric part that rewards the establishment of attractive interactions (cutoff of 6.0 Å) between any pair of non-consecutive residues. The second term is introduced to penalize structure realizations with radii of gyration larger than that found in naturally occurring proteins with the same length. The Monte Carlo evaluation if embedded in a simulated annealing scheme,[77] which allows one to minimize efficiently the scoring function by slowly decreasing a temperature-like control parameter.

ACKNOWLEDGMENTS

We are indebted with F. Cecconi, A. Flammini, D. Marenduzzo, and A. Maritan for several illuminating discussions and for a careful reading of the manuscript. Support from INFM and MURST Cofin 2001 is acknowledged.

REFERENCE

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 ed. New York: W.H. Freeman and Company; 1993.
3. Branden C, Tooze J. Introduction to protein structure. New York: Garland Publishing; 1991.
4. Jackson SE, Fersht AR. Folding of chymotrypsin inhibitor-2: evidence for a two-state transition. Biochemistry 1991;30:10428–10435.
5. Jackson SE. How do small single-domain proteins fold? Fold Design 1998;3:R81–R91.
6. Huang GS, Oas TG. Structural and stability of monomeric lambda repressor: NMR evidence for two-state folding. Biochemistry 1995;34:3884–3892.
7. Burton RE, Huang GS, Daugherty MA, Fullbright PW, Oas TG. Microsecond protein folding through a compact transition state. J Mol Biol 1996;163:311–322.
8. Kragelund BB, Robinson CV, Knudsen J, Dobson CM, Poulsen FM. Folding of a four-helix bundle: studies of acyl-coenzyme A binding protein. Biochemistry 1995;34:7217–7224.
9. Kragelund BB, Hojrup P, Jensen MS, Schjerling JE, Knudsen J, Poulsen FM. Fast and one-step folding of closely and distantly related homologous proteins of a four-helix bundle family. J Mol Biol 1996;256:187–200.
10. Mines GA, Pascher T, Lee SC, Winkler JR, Gray HB. Cytochrome-c folding triggered by electron-transfer. Chem Biol 1996;3:491–497.
11. Schonbrunner N, Kofler KP, Kiefhaber T. Folding of the disulfide-bonded 3-sheet protein tendamistat: rapid two-state folding without hydrophobic collapse. J Mol Biol 1997;268:526–538.
12. Reid KL, Rodriguez HM, Hillier BJ, Gregoret LM. Stability and folding properties of a model beta-sheet protein and escherichia coli cspA. Protein Sci 1998;7:470–479.

13. Grantcharova VP, Baker D. Folding dynamics of the src SH3 domain. Biochemistry 1997;36:15685–15692.
14. Villegas V, Azuaga A, Catasus L, Reverter D, Mateo PL, Aviles FX, Serrano L. Evidence for a two-state transition in the folding process of the activation domain of human procarboxypeptidase A2. Biochemistry 1995;34:15105–15110.
15. Khorasanizadeh S, Peters ID, Butt TR, Roder H. Evidence for a three-state model of protein folding from kinetic analysis of ubiquitin variants with altered core residues. Nat Struct Biol 1993;3:193–205.
16. Scalley ML, Yi Q, Gu HD, McCormack A, Yates JR, Baker D. Kinetics of folding of the ig binding domain of peptastreptococcal protein L. Biochemistry 1997;36:3373–3382.
17. Silow M, Oliveberg M. High-energy channeling in protein folding. Biochemistry 1997;36:7633–7636.
18. Van Nuland NAJ, Meijberg W, Warner J, Forge V, Schee RM, Robillard GT, Dobson CM. Slow co-operative folding of a small globular protein HPr. Biochemistry 1998;37:622–637.
19. Taddei N, Chiti F, Paoli P, Fiaschi T, Bucciantini M, Stefani M, Dobson CM, Ramponi G. Thermodynamics and kinetics of folding of common-type acylophosphatase: comparison to the highly homologous muscle isoenzyme. Biochemistry 1999;38:2135–2141.
20. Ferguson N, Capaldi AP, James R, Kleanthous C, Radford SE. Rapid folding with and without populated intermediates in the homologous four-helix proteins im7 and IM9. J Mol Biol 1999;286: 1597–1608.
21. Smith CK, Bu ZM, Anderson JMSKS, Engelman DM, Regan L. Surface point mutations that significantly alter the structure and stability of a protein's denatured state. Protein Sci 1996;5:2009–2019.
22. Kuhlman B, Luisi DL, Evans PA, Raleigh DP. Global analysis of the effects of temperature and denaturant on the folding and unfolding kinetics of the N-terminal domain of the protein L9. J Mol Biol 1998;284:1661–1670.
23. Hamill SJ, Meekhof AE, Clarke J. The effect of boundary selection on the stability and folding of the third fibronectin type III domain from human tenascin. Biochemistry 1998;37:8071–8079.
24. Tan Y-J, Oliveberg M, Fersht AR. Titration properties and thermodynamics of the transition state for folding: comparison of two-state and multistate folding pathways. J Mol Biol 1996;264: 377–389.
25. Plaxco KW, Simons KT, Baker D. Contact order and transition state placement and the refolding rates of single domain proteins. J Mol Biol 1998;277:985–994.
26. Debe DA, Goddard WA III. First principles prediction of protein folding rates. J Mol Biol 1999;294:619–625.
27. Dinner AR, Karplus M. The roles of stability and contact order in determining protein folding rates. Nat Struct Biol 2001;8:21–22.
28. Ivankov DN, Finkelstein A. Theoretical study of a landscape of protein folding-unfolding pathways. folding rates and midtransition. Biochemistry 2001;40:9957–9961.
29. Grantcharova V, Alm EJ, Baker D, Horwich AL. Mechanisms of protein folding. Curr Opin Struct Biol 2001;11:70–82.
30. Zhou H, Zhou Y. Folding rate prediction using total contact distance. Biophys J 2002;82:458–463.
31. Dokholyan N, Li L, Ding F, Shakhnovich E. Topological determinants of protein folding. Proc Natl Acad Sci USA 2002;99:8637–8641.
32. Aurora R, Creamer TP, Srinivasan R, Rose GD. Local interactions in protein folding: lessons from alpha-helix. J Mol Biol 1997;272: 1413–1416.
33. 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.
34. Capaldi AP, Radford SE. Kinetic studies of beta-sheet protein folding. Curr Opin Struct Biol 1998;8:86–92.
35. Clarke J, Cota E, Fowler SB, Hamill SJ. Folding studies of immunoglobulin-like beta-sandwich proteins suggest that they share a common folding pathway. Structure 1999;7:1145–1153.
36. Bollobas B. Random graphs. London: Academic; 1985.
37. Watts DJ, Strogatz SH. Collective dynamics of "small-world" network. Nature 1998;393:440–442.
38. Strogatz SH. Exploring complex networks. Nature 2001;410:268–276.
39. Go N, Scheraga HA. On the use of classical statistical mechanics in the treatment of polymer chain conformations. Macromolecules 1976;9:535–542.
40. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Funnels and pathways and the energy landscape of protein folding: a synthesis. Proteins 1995;21:167–195.
41. Wolynes PG, Onuchic JN, Thirumalai D. Navigating the folding routes. Science 1995;267:1619–1620.
42. Dill KA, Chan HS. From levinthal to pathways to funnels. Nat Struct Biol 1997;4:10–19.
43. Dobson MC, Sali A, Karplus M. Protein folding: a perspective from theory and experiment. Angew Chem Int Edit 1998;37:868–893.
44. Sali A, Shakhnovich E, Karplus M. How dows a protein fold. Nature 1994;369:28–251.
45. Gutin AM, Abkevich VI, Shakhnovich EI. Chain length scaling of protein folding time. Phys Rev Lett 1996;77:5433–5436.
46. Chan HS, Dill KA. The effects of internal constraints on the configurations of chain molecules. J Chem Phys 1990;92:3118–3135.
47. Cecconi F, Micheletti C, Carloni P, Maritan A. Molecular dynamics studies of HIV-1 protease: drug resistence and folding pathways. Proteins 2001;43:365–372.
48. Baker DA. Surprising simplicity to protein folding. Nature 2000; 45:39–42.
49. 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.
50. Clementi C, Nymeyer H, Onuchic JN. 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.
51. Galzitskaya OV, Finkelstein AV. A theoretical search for folding/unfolding nuclei in 3D protein structure. Proc Natl Acad Sci USA 1999;96:11299–11304.
52. Munoz V, Henry ER, Hofrichter J, Eaton WA. A simple model for calculating the kinetics of protein folding from three-dimensional structures. Proc Natl Acad Sci USA 1999;95:5872.
53. 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.
54. Klimov DK, Thirumalai D. Native topology determines force-induced unfolding pathways in globular proteins. Proc Natl Acad Sci USA 2000;97:7254–7259.
55. Hoang TX, Cieplak M. Sequencing of folding events in go-type proteins. J Chem Phys 2000;113:8319–8328.
56. Cieplak M, Hoang TX. Scaling of folding properties in go models of proteins. J Biol Phys 2000;26:273–294.
57. Shimada J, Shakhnovich E. The ensemble folding kinetics of protein g from an all-atom monte carlo simulation. Proc Natl Acad Sci USA 2002;99:11175–11180.
58. Dokholyan NV, Buldyrev SV, Stanley HE, Shaknovich EI. Identifying the protein folding nucleus using molecular dynamics. J Mol Bill 2000;296:1183–1188.
59. Kragelund B, Poulsen K, Andersen K, Baldursson T, Kroll JB, Neergard TB, Jepsen J, Roepstorff P, Kristiansen K, Poulsen F, Knudsen J. Conserved residues and their role in the structure, function, and stability of acyl-coenzyme a binding protein. Biochemistry 1999;38:2386–2394.
60. Tiana G, Broglia RA. Statistical analysis of native contact formation in the folding of designed model proteins. J Chem Phys 2001;114:2503–2510.
61. Dunn OJ, Clark VA. Applied statistics. New York: Wiley; 1974.
62. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical recipes. Cambridge: CUP; 1999.
63. Chothia C. Principles that determine the structure of proteins. Annu Rev Biochem 1984;53:537–572.
64. Chothia C. One thousand families for the molecular biologist. Nature 1992;357:543–544.
65. Levitt M, Chothia C. Structural patterns in globular proteins. Nature 1976;261:552–558.
66. Rose GD, Seltzer JP. A new algorithm for finding the peptide chain turns in a globular proteins. J Mol Biol 1977;113:153–164.
67. Maritan A, Micheletti C, Trovato A, Banavar JR. Optimal shapes of compact strings. Nature 2000;406:287.
68. Denton M, Marshall C. Laws of form revisited. Nature 2001;410: 417.
69. Hunt NG, Gregoret LM, Cohen FE. The origins of protein secondary structures. J Mol Biol 1994;241:214–225.

70. Socci ND, Bialek WS, Onuchic JN. Properties and origins of protein secondary structures. Phys Rev E 1994;49:3440–3443.
71. Dokholyan NV, Shakhnovich B, Shakhnovich E. Expanding protein universe and its origin from the biological big bang. Proc Natl Acad Sci USA 2002;99:14132–14136.
72. Micheletti C, Lattanzi G, Maritan A. Elastic properties of proteins: insight on the folding process and evolutionary selection of native structures. J Mol Biol 2002;321:909–921.
73. Vendruscolo M, Dokholyan NV, Paci E, Karplus M. A small-world view of the amino acids that play a key role in protein folding. Phys Rev E 2002;65:DOI 061910.
74. Baldwin RL, Rose GD. Is protein folding hierarchic? I. local structure and peptide folding. Trends Biochem Sci 1999;24:26–33.
75. Broglia R, Tiana G. Hierarchy of events in the folding of model proteins. J Chem Phys 2001;114:7267–7273.
76. Sokal AD. Monte carlo methods for the self-avoiding walk. Nuclear Phys 1996;B47:172–179.
77. Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealin. Science 1983;220:671–680.
78. Chan C, Hu Y, Takahashi S, Rousseau DL, Eaton WA, Hofrichter J. Submillisecond protein folding kinetics studied by ultrarapid mixing. Proc Natl Acad Sci USA 1997;94:1779–1784.
79. Viguera A, Martinez J, Filimonov V, Mateo P, Serrano L. Thermodynamic and kinetic analysis of the sh3 domain of spectrin shows a 2-state folding transition. Biochemistry 1994;33:2142–2150.
80. Plaxco K, Guijarro J, Morton C, Pitkeathly M, Campbell I, Dobson C. The folding kinetics and thermodynamics of the fyn-sh3 domain. Biochemistry 1998;37:2529–2537.
81. Perl D, Welker C, Schindler T, Schroeder K, Marahiel MA, Jaenicke R, Schmid F. Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins. Nat Struct Biol 1998;5:229–235.
82. Guijarro I, Morton C, Plaxco KW, Campbell ID, Dobson C. Folding kinetics of the sh3 domain of pi3 kinase by real-time nmr combined with optical spectroscopy. J Mol Biol 1998;276:657–667.
83. Clarke J, Hamill SJ, Johnson CM. Folding and stability of a fibronectin type iii domain of human tenascin. J Mol Biol 1997;270:771–778.