Computational Methods for Protein Folding: Advances in Chemical Physics, Volume 120.
Edited by Richard A. Friesner. Series Editors: I. Prigogine and Stuart A. Rice.

ISBNs: 0-471-20955-4 (Hardback); 0-471-22442-1 (Electronic)

STATISTICAL ANALYSIS OF PROTEIN FOLDING KINETICS

AARON R. DINNER

New Chemistry Laboratory, University of Oxford, Oxford, U.K.

SUNG-SAU SO

Hoffmann-La Roche Inc., Discovery Chemistry, Nutley, NJ, U.S.A.

MARTIN KARPLUS

New Chemistry Laboratory, University of Oxford, Oxford, U.K.; Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, U.S.A.; and Laboratoire de Chimie Biophysique, Institut le Bel, Université Louis Pasteur, Strasbourg, France

CONTENTS

I. INTRODUCTION

Experimental and theoretical studies have led to the emergence of a unified general mechanism for protein folding that serves as a framework for the design and interpretation of research in this area [1]. This is not to suggest that the details of the folding process are the same for all proteins. Indeed, one of the most striking computational results is that a single model can yield qualitatively different behavior depending on the choice of parameters [1–3]. Consequently, it remains to determine the behavior of individual sequences under given environmental conditions and to identify the specific factors that lead to the manifestation of one folding scenario rather than another. Although doing so requires investigation of the kinetics of particular proteins at the level of individual residues, for which protein engineering [4] and nuclear magnetic resonance (NMR) [5] experiments are very useful, complementary information about the roles played by the sequence and the structure can also be obtained by a statistical analysis of the folding rates of a series of proteins.

Statistical methods have been applied for many years in attempts to predict the structures of proteins (for a review of progress in this area, see the chapter by Meller and Elber, this volume), but their use in the analysis of folding kinetics is relatively recent. The first such investigations focused on ''toy'' protein models in which the polypeptide chain is represented by a string of beads restricted to sites on a lattice. It was found that the ability of a sequence to fold correlates strongly with measures of the stability of its native (ground) state (such as the Z-score or the gap between the ground and first excited compact states) [6–9], but the native structure also plays an important role for longer chains [10,11]. While lattice models are limited in their ability to capture the structural features of proteins, they have the important advantage that the results of statistical analyses can be compared with calculated folding trajectories to determine the physical bases of observed correlations. Consequently, studies based on such models are particularly useful for the quantitation of observed effects, the generalization from individual sequences, the identification of subtle relationships, and ultimately the design of additional sequences that fold at a given rate.

Analogous statistical analyses of experimentally measured folding kinetics of proteins were hindered by the fact that complex multiphasic behavior was exhibited by most of the proteins for which data were available (e.g., barnase and lysozyme). In recent years, an increasing number of proteins that lack

significantly populated folding intermediates and thus exhibit two-state folding kinetics have been identified, and a range of data have been tabulated for them [12–14]. The initial linear analyses of such proteins indicated that their folding rates are determined primarily by their native structures [12,14]. More recently, a nonlinear, multiple-descriptor approach revealed that there is a significant dependence on the stability as well [15]. These and related studies are discussed in Section IV.A, after an overview of the statistical methods employed in this area (Section II) and a review of the results from lattice models (Section III).

An in-depth analysis of a database of 33 proteins that fold with two- or weakly three-state kinetics is presented in Sections IV.B through V. We explore one-, two-, and three-descriptor nonlinear models. A structurally based cross-validation scheme is introduced. Its use in conjunction with tests of statistical significance is important, particularly for multiple-descriptor models, due to the limited size of the database. Consistent with the initial linear studies [12,14], it is found that the contact order and several other measures of the native structure are most strongly related to the folding rate. However, the analysis makes clear that the folding rate depends significantly on the size and stability as well. Due to the importance ascribed to the stability by analytic [16–18] and simulation [2,3,6–11] studies, as well as its recent use in one-dimensional models for fitting and interpreting experimental data [19,20], we examine its connection to the folding rate in more detail. The unfolding rate, which correlates more strongly with stability, is considered briefly. The relation of the statistical results to experiments and the model studies is discussed in Sections VI and VII.

II. STATISTICAL METHODS

Before reviewing the results for specific systems, we introduce the statistical methods that have been used to analyze folding kinetics. Perhaps the simplest such method is to group sequences; here, one categorizes each sequence in a database according to one or more of its native properties (''descriptors'') and its folding behavior. Visualization can be used to identify patterns, and averages and higher moments of the distributions of descriptors can be used to quantitate differences between groups. For properties on which the folding kinetics depend strongly, such as the energy gap in lattice models, this type of analysis has proven effective [6].

However, simple grouping is often insufficient to identify weaker but still significant trends and makes it difficult to determine the relative importance of relationships. Consequently, more quantitative methods are necessary. One statistic that is employed widely is the Pearson linear correlation coefficient ($r_{x,y}$):

$$r_{x,y} = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{1}$$

Typically, the x_i are a set of values of a particular descriptor, such as the sequence length, and the y_i are a set of values for a measure of the folding kinetics, such as the logarithm of the folding rate constant ($\log k_f$) [9,10,12]. The magnitude of $r_{x,y}$ determines its significance, and its sign indicates whether x_i and y_i vary in the same or opposite manner: $r_{x,y} = 1$ corresponds to a perfect correlation, $r_{x,y} = -1$ to a perfect anticorrelation, and $r_{x,y} = 0$ to no correlation. In spite of its popularity, this statistic has several shortcomings when used by itself. It is limited to the identification of linear relationships between pairs of properties; it is not straightforward to test or cross-validate those relationships, which is important, as discussed below; and it cannot be used directly to predict the behavior of additional sequences.

These limitations can be overcome by constructing models to predict folding behavior and then quantifying their accuracy. For the latter step, the Pearson linear correlation coefficient can be used with x_i as the observed values and y_i as the predicted ones (for which we introduce the shorthand notations r_{trn}, r_{jck}, and r_{cv}, described below). Alternatively, one can calculate the root-mean-square error or the closely related fraction of unexplained variance:

$$q^2 = 1 - \frac{\sum_i (y_i - x_i)^2}{\sum_i (x_i - \bar{x})^2} \tag{2}$$

Again, x_i (y_i) are the observed (predicted) values. Typically, r and q^2 behave consistently. The latter is useful for quantitating the improvement obtained upon extending a model with N descriptors to one with $N + 1$ with Wold's statistic: $E = (1 - q^2_{N+1})/(1 - q^2_N)$ [21,22]. A value of less than 1.0 for the latter shows that q^2 increases upon adding a descriptor. The statistical significance of a particular value of E depends on the specific data, but $E = 0.4$ has been suggested to correspond typically to the 95% confidence interval [23].

For constructing the models themselves, linear regression (on one or more descriptors) is attractive in that the best fit for a set of data can be determined analytically, but, as its name implies, it is limited to detecting linear relationships. While fits with higher-order polynomials are possible, a general and flexible alternative is to use neural networks (NNs). The latter are computational tools for model-free mapping that take their name from the fact that they are based on simple models of learning in biological systems [24,25]. Neural networks have been used extensively to derive quantitative structure–property relationships in medicinal chemistry (for a review, see Ref. 26) and were first used to analyze folding kinetics in Ref. 11. A schematic diagram of a neural network is shown in Fig. 1. In this example, there are three inputs (indicated by the rectangles on the left); in the present study these would each contain the value of a descriptor, such as the free energy of unfolding or the fraction of

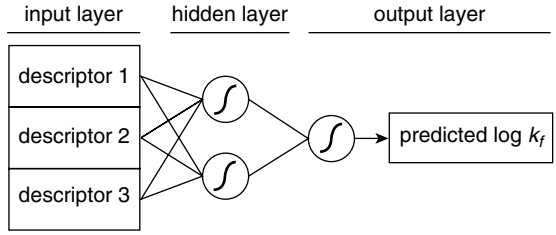


Figure 1. Schematic of a neural network.

helical contacts. The circles represent sigmoidal functions (nodes). There are many possible choices for the specific form of these functions; we use

$$f = \frac{1}{1 + \exp\left(-\theta - \sum_i w_i p_i\right)} \tag{3}$$

where the sum ranges over the previous layer (to the left in the diagram), p_i are the values of the elements of that layer, w_i are the weights for each of those elements (represented by the connecting lines in the diagram), θ is an arbitrary constant, and the data are assumed to be normalized for clarity. Thus, to ''fire'' the network in Fig. 1, a weighted sum over the three inputs to each hidden node is made, the resulting sums are used to calculate the values of the sigmoidal functions associated with those nodes, a weighted sum of those values is then made, and the final sigmoidal function of the output node is calculated. To fit data, the w_i are initialized to random values and adjusted with standard optimization techniques to maximize the accuracy of the output for the (training) set. In the present study, we varied the weights with the scaled conjugate gradient method [27].

When one wishes to test many different possible descriptors, the number of possible NN input combinations can be very large. One can avoid making an exhaustive search by using a genetic algorithm (GA) to select the descriptors to test. This tool is also biologically motivated—in this case, by evolution. A population is created in which each individual consists of a particular set of descriptors. Repeatedly, each such set (a ''parent'') is duplicated (''asexual reproduction''), the new copy (a ''child'') is changed by one descriptor (''mutated''), and then only the best (''fittest'') individuals in the combined pool of parents and children are kept. Here, ''best'' means that a linear regression or NN model employing those descriptors yields the greatest accuracy for the training set. Alternative schemes that involve combining features from different individuals (''sexual reproduction'') also exist but are not employed here; for a comprehensive review of the use of GAs in medicinal chemistry see Ref. 28. In the present study, we used 40 individuals with 20 genetic cycles; a few trials with 200 individuals and 50 cycles did not yield significantly different results.

An important point concerning neural networks, and indeed any multiple parameter model, is that it is possible to overfit the data. For small sample sizes (here, a small number of proteins), even relatively simple neural networks can memorize the examples in the training set at the expense of learning more general rules. Thus, it is important to test a model on novel data not used during the fitting process. One approach is cross-validation, in which one partitions the existing data into a series of training and test sets. In the special case of jackknife cross-validation, all possible combinations are formed in which a single protein is used to test the network and the remainder are used to train it. While jackknife cross-validation is straightforward to automate, it is not appropriate if any members of the database are significantly related (e.g., homologous proteins) because the inclusion of the similar data in the training set can bias the test. A structurally based partitioning scheme is presented in Section IV.B. Throughout, care is taken to distinguish statistics (r and q^2) for *fits* of the entire (training) set (denoted "trn") from those for *predictions* obtained with either jackknife or structurally based cross-validation (denoted "jck" and "cv," respectively).

III. LATTICE MODELS

The first study in which a large number of unrelated sequences were analyzed to identify the factors that determine their folding kinetics was based on a 27-residue chain of beads subject to Monte Carlo dynamics on a simple cubic lattice [6]. In this and the subsequent studies of 125-residue sequences [10,11], folding rate constants were calculated for only a few sequences due to the large number of trajectories required to obtain accurate results. Folding "ability" was measured by either (a) the fraction of Monte Carlo trials that reached the native state within the allotted simulation time or (b) the average fraction of native contacts in the lowest energy states sampled. When the results for the 27-residue sequences were grouped according to the former, it was found that the stability of the native (ground) state is the only feature that distinguishes those that folded repeatedly within the simulation time from those that did not. If the native state is maximally compact, the stability criterion can be simplified to a consideration of the difference in energy between the ground state and the first fully compact ($3 \times 3 \times 3$) excited state [6]. These criteria have been used in the design of fast folding sequences [29] and are consistent with similar studies which focus on exhaustive enumeration of folding paths for two-dimensional chains [7,30] or on the ratio of the folding and the "glass" transition temperatures for the (three-dimensional) 27-residue model [8].

In a number of subsequent studies of the 27-residue model, it was argued that the kinetic folding behavior is determined by factors other than the energy gap

[31–33]. Unger and Moult [31] suggested that the dependence on the energy gap derived from the variation in the simulation temperature in Ref. 6 and identified the structure of the ground state as the primary determinant of the folding kinetics of this system. However, in a study of 15- and 27-residue three-dimensional chains that employed the Pearson linear correlation coefficient to quantitate the relationships between various sequence factors and the logarithm of the mean first passage time, the correlation with the *Z*-score was robust to use of a single temperature [9]. Examination of Ref. 31 showed that sequences were designed to have strong short-range contacts without mandating a certain fraction of long-range contacts, so that the resulting ground states were more appropriate for modeling a helix-coil transition than protein folding. Nevertheless, as will be discussed below, native structure does play a role for certain lattice models [10,11] as it does for proteins [12,14,15]. Klimov and Thirumalai [32,33] introduced the parameter $\sigma = 1 - T_f/T_\theta$, where T_f is the temperature at which the fluctuation of the order parameter is at its maximum and T_θ is the temperature at which the specific heat is at its maximum. They found that σ is positively correlated with the logarithm of the mean first passage time (i.e., small sigma gives fast folding). However, the interpretation of T_θ as the collapse transition temperature is not correct in general, and the correlation described above arises from the fact that σ is related to the energy gap [9]. These statistical studies of short chains are discussed in detail in Ref. 9.

The correlation of the folding time with the energy gap can be understood in terms of its effect on the energy surface. For random 27-residue sequences, folding proceeds by a fast collapse to a semicompact disordered globule, followed by a slow, nondirected search through the relatively small number of semicompact structures for one of the many transition states that lead rapidly to the native conformation [2]. A large energy gap results in a native-like transition state that is stable at a temperature high enough for the folding polypeptide chain to overcome barriers between random semicompact states. As the energy gap increases to the levels obtainable in designed sequences, the model exhibits Hammond behavior [34] in that there is a decrease in the fraction of native contacts required in the transition state from which the chain folds rapidly to the native state. Random sequences with relatively small gaps must form about 80% of the native contacts [2], whereas designed sequences with large gaps need form only about 20% [35]. This shift increases the ratio of the number of transition states to the number of semicompact states and results in a nucleation mechanism [35].

The first study to employ the Pearson linear correlation coefficients between various individual sequence properties and measures of folding ability concerned the analysis of 125-residue lattice model simulations [10]. It revealed that, in addition to the stability, the native structure plays an important role in determining

folding ability for chain lengths comparable to that typical of certain well-studied proteins (e.g., barnase and lysozyme); that is, a strong correlation was observed between the frequency of reaching the native state within the simulation time and the number of native contacts in tight turns or antiparallel sheets. On the lattice, these are the cooperative secondary structural elements that have the shortest sequential separations between contacts; lattice "helices," which typically consist only of $i, i + 3$ contacts, are noncooperative and thus do not accelerate folding. The physical basis of the relation between structure and kinetics in lattice models and in proteins is discussed in Section IV.E.

The initial linear analysis of the 125-residue model also made clear that one descriptor can compensate for others, so that it is necessary to consider more than one simultaneously [10]. Accordingly, the functional dependence of the folding ability on sets of sequence properties was derived with an artificial neural network, and a genetic algorithm was used to select the sets that maximize the accuracy of the predictions. Not only did the nonlinear, multiple-descriptor method increase the correlation coefficients between the observed folding abilities and the cross-validated predictions from about 0.5 to greater than 0.8, but it revealed (in addition to the strong dependences on the stability and structure of the native state) a role for the spatial distribution of strong and weak pairwise interactions within the native structure. Sequences with native structures that have more labile contacts between surface residues were found to fold faster in general because misfolded subdomains are less likely to form and lead to off-pathway traps [10,11,36]. This observation indicates that, as one goes to longer sequences, the relationship between the folding rate and the native state descriptors becomes more complex.

The genetic neural network (GNN) method was further validated by use of one of the resulting quantitative structure–property relationships (QSPRs) to design additional fast-folding 125-residue sequences [37]. The target native structure and the pairwise interaction energies were varied to maximize the output of a network trained on the original set of sequences to predict the average fraction of native contacts in the lowest energy structure sampled in each of 10 Monte Carlo simulations [10,11]. The specific descriptors employed were the number of contacts in antiparallel sheets, the estimated gap in energy between the native state and the lower limit of the quasi-continuous spectrum [38], and the total energy of the contacts between surface residues. On average, the designed sequences folded more rapidly than those for which only the stability of the native state was optimized [29,39]. The studies of the 125-residue lattice models thus make clear that simultaneous consideration of multiple descriptors can improve our understanding of protein folding and our ability to extrapolate from the analysis to predict the behavior of novel sequences. The utility of the statistical approach for obtaining a better understanding of the folding rates of proteins is described in the following section.

IV. FOLDING RATES OF PROTEINS

In this section we describe statistical analyses of measured rates of protein folding. Earlier studies are reviewed and an analysis of currently available experimental data is presented. The physical bases of the results are then discussed.

A. Review

As mentioned in the Introduction, statistical analyses of the folding kinetics of proteins were delayed until a sufficient number of proteins that fold with two-state kinetics overall were identified [12,13]. Plaxco et al. [12] carried out an analysis much like the initial 125-mer lattice model study mentioned above [10] for a set of 12 two-state proteins (extended to 24 proteins in Ref. 14); that is, they calculated linear correlation coefficients between several individual sequence properties and the logarithm of the measured folding rate constants ($\log k_f$). The only descriptor examined that exhibited a high correlation ($r_{c/n,\log k_f} = 0.81$) was the structure of the native state as measured by the normalized contact order (c/n), the average sequential residue separation of atoms in contact divided by the length of the sequence (see the footnote to Table III for the exact definition of c/n employed here). It is important to note that the contact order does not include any information about the energies of the interactions in the native state; it is only a measure of the structure (we use the term ''structure'' rather than ''topology'' [12,14] because, according to the standard mathematical meaning of the latter, all proteins that lack disulfide bonds have the same topology).

We used a neural network to carry out a nonlinear, two-descriptor analysis of the database of 33 proteins described in Section IV.B [15] and demonstrated that the stability contributes significantly to determining folding rates for a given contact order. Moreover, for 14 slow-folding proteins with high contact orders (mixed-α/β and β-sheet proteins), the free energy of unfolding can be used by itself to predict folding rates. By contrast, the folding rates of α-helical proteins show essentially no dependence on the stability. The variation in behavior observed for the structural classes suggests that, although there is a general mechanism of folding (see the Introduction), its expression for individual proteins can lead to very different behavior.

A number of simple physically motivated one-dimensional models have been introduced recently to fit and interpret data on peptide and protein folding [19, 20,40–42]. These models, which use only native state data, have elements in common with earlier theoretical treatments by Zwanzig, Wolynes, and their coworkers [16,17,43]. The conformation of a protein is represented by a series of binary variables (based on one or two residues), each of which can be either native or random coil. Pairwise interactions (which are assumed to be entirely favorable, as in a Gō model [44,45]) are counted if and only if both the sequence positions involved are native. Often, an additional approximation is made in

which the formation of the native structure is limited to one or two sequential segments [46]. Independent of this assumption, the one-dimensional character of these models and the choice of energy functions typically force the native structure to propagate in an essentially sequential manner. By adjusting parameters, one of these models was shown to fit log k_f with an accuracy of $0.83 \leq r_{trn} \leq 0.87$ for 18 proteins [20]. The fact that this correlation is somewhat higher than that obtained using only the contact order (Table I and Refs. 12,14, and 20) has been used as evidence for the physical basis of the model; that is, it provides an ''explanation'' of the empirical relationship between the folding rate and the contact order. However, the improvement appears to be due to the incorporation of the protein stabilities into the model. These were introduced by adjusting the pairwise interactions separately for each protein such that the model yielded free energies for folding that matched experimental ΔG values. Using the methods described in Section II and applied in Section IV.B, we were able to obtain $r_{trn} = 0.93$ with two descriptors (ΔG and q_a, described in Table I) and $r_{trn} = 0.98$ with three (ΔG, c, and b) for the same set of 18 proteins; for c/n, and $\Delta G/n$, $r_{trn} = 0.85$, which is very similar to the correlations reported in Ref. 20 ($0.83 \leq r_{trn} \leq 0.87$). Thus, further work is required to show that such simple phenomenological models can predict aspects of the folding reaction that go beyond the experimental data used in the fitting procedures. Although these model studies consider the prediction of ϕ values [4], it appears from the published results and statements in the text of Ref. 20 that the correlation is poor. This suggests that quantitative comparisons of predicted ϕ-values with the observed ones could serve as a meaningful test of such phenomenological models.

An alternative phenomenological model was developed by Debe and Goddard [47]. In essence, they assumed a sequence of events which is, in a certain sense, the reverse of the diffusion–collision model [48,49]: the correct overall (tertiary) structure is formed at low-resolution first by a random search and then local (secondary) refinement takes place within the manifold of states in that fold. Thus, the factor that determines the relative rate of folding for a series of proteins is the probability of randomly sampling a structure with the known native contacts (estimated by a Monte Carlo procedure); the distance at which a contact was counted was adjusted to optimize the fit. For mixed-α/β and β-sheet proteins, an accuracy of $r_{trn} = 0.78$ was obtained. This statistic is comparable to the correlation coefficients associated with the contact order (Table I and Refs. 12 and 14), which could suggest that this model is a rather complex procedure for reproducing the simple (essentially linear) dependence of log k_f on that descriptor. For α-helical proteins, the folding rates were considerably underestimated, which led Debe and Goddard to conclude that hose proteins must instead fold by a diffusion–collision mechanism [48,49]. The discussion in the present section shows that phenomenological models can be useful for

interpreting the observed statistical correlations. However, it is important to keep in mind that the ability to fit a particular set of data is not sufficient to demonstrate that the folding *mechanism* on which the model is based is correct.

B. Database

To illustrate the methods described in Section II and to show that simultaneous consideration of multiple descriptors improves prediction of protein folding kinetics, we describe a detailed analysis of the available data for the folding rates of two- and weakly three-state proteins. The descriptors tested are listed in Table I and can be divided into several categories: native state stability (0 and 1), size (2 to 5), native structure (8 to 15), and the propensity for a given structure (16 to 23). Definitions and sources for the descriptors as well as the data themselves are given in Tables II and III. Although certain descriptors are significantly

TABLE I
Descriptors Tested as Inputs to the GNN and Their Correlations[a]

Index	Symbol	Description	$r_{x,\log k_f}$	r_{trn}	r_{cv}	q^2_{cv}
0	ΔG	Stability	0.29	0.40	0.06	−0.16
1	$\Delta G/n$	Normalized stability	0.37	0.42	−0.00	−0.13
2	m	Buried surface area	−0.04	0.38	−0.16	−0.40
3	m/n	Normalized surface area	−0.04	0.24	−0.29	−0.21
4	n	Sequence length	−0.10	0.35	−0.52	−0.19
5	n_c	Number of atomic contacts	−0.08	0.34	−0.32	−0.18
6	c	Contact order	−0.73	0.74	0.67	0.45
7	c/n	Normalized contact order	−0.79	0.83	0.74	0.54
8	h	α-Helix content	0.63	0.64	0.39	0.11
9	e	β-Sheet content	−0.67	0.71	0.59	0.34
10	t	H-bonded turn content	0.04	0.34	−0.07	−0.21
11	s	Bend content	−0.11	0.31	−0.25	−0.26
12	g	3_{10}-Helix content	−0.01	0.35	−0.47	−0.28
13	b	β-Bridge content	−0.15	0.30	−0.36	−0.32
14	o	Other 2° structure	−0.05	0.27	−0.32	−0.44
15	a	Total helix content ($h+g$)	0.63	0.67	0.28	−0.04
16	P_h	Predicted α-helix	0.47	0.49	0.05	−0.10
17	P_e	Predicted β-sheet	−0.48	0.57	0.29	0.01
18	P_o	Predicted other 2°	−0.27	0.43	−0.39	−0.32
19	p_h	α-Helix propensity	0.51	0.55	0.21	−0.03
20	p_e	β-Sheet propensity	−0.47	0.64	0.42	0.14
21	p_o	Other 2° propensity	−0.40	0.50	−0.20	−0.16
22	q_e	Expected 2° prediction accuracy	0.21	0.42	0.07	−0.14
23	q_a	Actual 2° prediction accuracy	0.40	0.45	−0.14	−0.45

[a]Here r_{trn} and r_{cv} are correlation coefficients between observed and calculated values of log k_f for training set fits and cross-validated predictions, respectively. Correlations are the maximum ones observed for 10 independent trials, each with a different random number generator seed. Statistics for linear regression are available in Table V.

TABLE II
Rate, Stability, and Size Descriptors[a]

Group	Protein	Reference	log k_f	log k_u	ΔG	$\Delta G/n$	m	m/n	n	n_c
SH3	1NYF		1.973128	−3.00382	6.0	0.0895522	1.40	0.0208955	67	378
	1PKS		−0.455932	−3.17335	3.4	0.0404762	2.30	0.0273810	84	710
	1SHG		0.612784	−2.55238	2.9	0.0467742	0.80	0.0129032	62	406
	1SRL		1.755875	−0.99982	4.1	0.0640625	1.60	0.0250000	64	389
β-Sandwich	1FNF-9		−0.397940	−1.30080	1.2	0.0133333	3.00	0.0333333	90	686
	1FNF-10	61	2.380211	−3.63762	9.4	0.1000000	6.50	0.0691489	94	648
	1HNG	61	1.255273	−2.76905	6.8	0.0693878	1.10	0.0112245	98	560
	1TEN		0.462398	−2.55238	4.8	0.0533333	1.30	0.0144444	90	600
	1TIT	61	1.505150	−3.30921	7.5	0.0842697	2.50	0.0280899	89	472
	1WIT	61	0.176091	−3.55220	4.0	0.0430108	1.30	0.0139785	93	893
Acylphosphatase	1APS		−0.638272	−3.95789	5.4	0.0551020	1.25	0.0127551	98	833
	1HDN		1.173186	−2.67730	4.6	0.0541176	2.20	0.0258824	85	705
	1PBA		2.952792	−0.18705	4.1	0.0512500	1.00	0.0125000	80	589
	1URN		2.499687	−4.19990	9.3	0.0911765	2.30	0.0225490	102	749
	2HQI	63	0.079181		3.8	0.0527778	2.35	0.0326389	72	730
Cytochrome	1HRC		3.447158	−1.76923	6.9	0.0663462	2.40	0.0230769	104	828
	1HRC-oxidized		2.602060		17.7	0.1701923	3.30	0.0317308	104	828
	1YCC		4.176091		14.6	0.1417476	3.10	0.0300971	103	863
Cold shock	1CSP		3.029384	1.07899	3.0	0.0447761	0.76	0.0113433	67	346
	1MJC		2.274158	0.51842	2.9	0.0420290	0.57	0.0082609	69	400
λ-Repressor	1LMB		3.690196	1.47686	3.0	0.0375000	1.10	0.0137500	80	632
	1LMB-G46A/G48A	13,64	4.944483	1.55602	4.8	0.0600000	1.10	0.0137500	80	632
Ubiquitin	1UBQ		3.185259	−3.35891	7.1	0.0934211	1.90	0.0250000	76	510
	1UBQ-V26A		2.008600	−1.09671	3.9	0.0513158	2.00	0.0263158	76	510

Unique	1COA		1.681241	−3.74405	7.0	0.1093750	1.80	0.028250	64	376
	1DIV	65	2.857332	−0.12492	4.1	0.0725000	1.46	0.0260714	56	461
	1FKB		0.633468	−3.76887	5.5	0.0514019	1.40	0.0130841	107	758
	1IMQ	66	3.161368	−1.90623	6.3	0.0732558	1.10	0.0127907	86	936
	2ABD		2.445604	−3.99928	7.1	0.0825581	3.00	0.0348837	86	1118
	2AIT		1.826075	−4.34600	8.1	0.1094595	1.30	0.0175676	74	621
	2PDD	67,68	4.255272	−2.36130	3.1	0.0720930	0.80	0.0186047	43	199
	2PTL		1.778151	−1.69866	4.6	0.0741935	1.90	0.0306452	62	482
	2VIK		2.954243	−1.21445	6.2	0.0492063	1.60	0.0126984	126	1089

[a]Unless otherwise noted, the stabilities, m values, and rates were taken from Ref. 13. Concerning the measured values of $\log k_f$ and $\log k_u$, it should be noted that the available data were obtained at different temperatures, and no correction for this variation was made. For CI2, for which data are available, the folding rate varies by about one order of magnitude over the full temperature range in the database (5°C to 37°C), but it changes by only a factor of about 1.5 over a range that includes most of the database (20°C to 25°C).

TABLE III
Structural and Structural Propensity Descriptors[a]

Set	Protein	c	c/n	h	e	t	s	g	b	o	a	P_h	P_e	P_o	p_a	p_e	p_o	q_e	q_a
SH3	1NYF	18.11	27.03	0.00	35.82	7.46	11.94	4.48	4.48	35.82	4.48	0.00	26.87	73.13	8.84	31.59	60.66	73.36	79.10
	1PKS	24.75	29.46	4.71	30.59	14.12	4.71	10.59	3.53	31.76	15.29	17.65	16.47	65.88	18.93	23.95	57.97	72.14	57.65
	1SHG	18.05	29.11	0.00	41.94	9.68	8.06	4.84	4.84	30.65	4.84	19.35	29.03	51.61	23.51	29.68	48.03	75.48	58.06
	1SRL	20.49	32.02	0.00	25.00	3.12	20.31	4.69	1.56	45.31	4.69	0.00	32.81	67.19	6.80	37.56	56.79	73.91	82.81
β-Sandwich	1FNF-9	27.97	31.08	0.00	54.44	6.67	11.11	0.00	0.00	27.78	0.00	0.00	24.44	75.56	4.10	32.12	64.65	78.54	65.56
	1FNF-10	26.53	28.22	0.00	50.00	8.51	9.57	0.00	1.06	30.85	0.00	0.00	42.55	57.45	2.88	43.56	54.40	79.10	79.79
	1HNG	31.65	32.30	0.00	53.06	16.33	7.14	3.06	0.00	20.41	3.06	21.43	19.39	59.18	24.28	24.81	51.63	70.25	55.10
	1TEN	27.42	30.47	0.00	53.33	11.11	7.78	0.00	0.00	27.78	0.00	4.44	27.78	67.78	8.98	31.87	59.96	73.39	70.00
	1TIT	28.78	32.34	0.00	35.96	6.74	17.98	0.00	1.12	38.20	0.00	17.98	31.46	50.56	16.87	35.86	48.08	71.95	64.04
	1WIT	30.25	32.53	0.00	51.61	13.98	5.38	3.23	3.23	22.58	3.23	15.05	31.18	53.76	15.79	32.00	52.92	66.71	64.52
Acylphosphatase	1APS	34.38	35.08	18.37	36.73	9.18	14.29	0.00	0.00	21.43	18.37	2.04	37.76	60.20	11.47	34.69	54.53	67.42	52.04
	1HDN	24.24	28.52	37.65	30.59	8.24	11.76	0.00	0.00	11.76	37.65	36.47	21.18	42.35	37.00	23.48	40.46	80.03	77.65
	1PBA	20.67	25.84	25.93	11.11	4.94	18.52	3.70	0.00	35.80	29.63	18.52	23.46	58.02	22.00	26.30	52.56	69.92	70.37
	1URN	26.88	26.35	27.45	23.53	10.78	9.80	5.88	0.00	22.55	33.33	44.12	14.71	41.18	40.63	16.67	43.44	75.22	68.63
	2HQI	22.35	31.04	26.39	27.78	18.06	11.11	0.00	0.00	16.67	26.39	23.61	30.56	45.83	23.32	30.52	47.13	69.92	72.22
Cold shock	1CSP	19.51	29.12	0.00	53.73	14.93	7.46	4.48	1.49	17.91	4.48	0.00	38.81	61.19	5.09	44.23	51.83	77.16	70.15
	1MJC	18.57	26.91	0.00	47.83	15.94	5.80	4.35	1.45	24.64	4.35	0.00	34.78	65.22	4.00	37.38	59.76	78.64	75.36
Cytochrome	1HRC	17.00	16.35	41.35	0.00	17.31	7.69	0.00	1.92	31.73	41.35	20.19	14.42	65.38	26.93	19.10	54.69	75.50	64.42
	1HRC-oxidized	17.00	16.35	41.35	0.00	17.31	7.69	0.00	1.92	31.73	41.35	20.19	14.42	65.38	26.93	19.10	54.69	75.50	64.42
	1YCC	16.97	16.48	42.72	0.00	15.53	5.83	0.00	1.94	33.98	42.72	8.74	14.56	76.70	19.20	20.67	60.81	70.51	58.25
λ-Repressor	1LMB	9.77	12.21	70.00	0.00	7.50	5.00	3.75	0.00	13.75	73.75	67.50	0.00	32.50	61.23	5.65	34.08	76.44	77.50
	1LMB-G46A/G48A	9.77	12.21	68.75	0.00	7.50	5.00	3.75	0.00	15.00	72.50	70.00	0.00	30.00	65.07	4.69	31.23	78.19	81.25

Ubiquitin	1UBQ	17.44	22.95	15.79	31.58	15.79	5.26	7.89	2.63	21.05	23.68	11.84	28.95	59.21	19.22	31.34	50.35	68.92	84.21
	1UBQ-V26A	17.44	22.95	15.79	31.58	15.79	5.26	7.89	2.63	21.05	23.68	13.16	28.95	57.89	22.34	29.45	49.14	70.87	85.53
Unique	1COA	16.07	25.11	17.19	21.88	15.62	4.69	4.69	6.25	29.69	21.88	17.19	29.69	53.12	22.52	32.99	45.66	75.26	76.56
	1DIV	10.62	18.96	33.93	19.64	16.07	8.93	5.36	3.57	12.50	39.29	28.57	21.43	50.00	33.88	23.60	43.77	70.29	71.43
	1FKB	30.99	28.96	7.48	38.32	22.43	5.61	2.80	1.87	21.50	10.28	2.80	33.64	63.55	8.61	33.15	58.92	73.23	76.64
	1IMQ	16.14	18.77	52.33	0.00	11.63	11.63	0.00	0.00	24.42	52.33	40.70	5.81	53.49	38.60	10.73	51.59	79.38	82.56
	2ABD	17.03	19.80	56.98	0.00	10.47	8.14	3.49	0.00	20.93	60.47	54.65	0.00	45.35	50.22	8.33	42.34	76.71	81.40
	2AIT	27.62	37.32	0.00	40.54	8.11	16.22	0.00	2.70	32.43	0.00	12.16	36.49	51.35	15.25	36.12	49.67	77.62	67.57
	2PTL	20.53	33.11	19.35	38.71	9.68	14.52	0.00	1.61	16.13	19.35	24.19	41.94	33.87	24.46	38.45	38.30	75.24	85.48
	2PDD	6.45	15.00	44.19	0.00	16.28	9.30	0.00	0.00	30.23	44.19	34.88	16.28	48.84	32.75	23.04	45.88	71.79	74.42
	2VIK	25.87	20.53	19.05	23.81	15.08	16.67	2.38	0.00	23.02	21.43	24.60	23.02	52.38	24.16	27.98	48.43	71.23	77.78

[a]The secondary structure contents were obtained with the program (DSSP) [69], and the secondary structure predictions and propensities were obtained with the program PRED2ARY [70] (these descriptors are expressed as percentages of the total numbers of residues). Each of the mutations involved the substitution of an alanine into a helix; because such a change is likely to increase the propensity for forming a helix in that region, the contact orders and secondary structure content were taken to be the same as those of the wild types, and the secondary structure propensities and predictions were calculated with the modified sequences. Likewise, the structural data for the two forms of horse cytochrome *c* (1HRC) were taken to be the same. A contact was defined as two heavy atoms that are within 4 Å of each other and separated by at least two residues (i.e., $i, i+1$ and $i, i+2$ contacts are ignored). The (unnormalized) contact order is $c = \frac{1}{n_c}\sum_{i>j}\Delta(i,j)|s_i - s_j|$, where n_c is the total number of contacts, s_i is the sequence position of the residue containing atom i, and $\Delta(i,j)$ selects the atoms (i and j) that are in contact (as defined above). The normalized contact order (c/n) is multiplied by 100 for consistency with Refs. 12 and 13.

correlated with others (Table IV), consideration of all of them is useful because exhaustive enumeration or a genetic algorithm (GA) is employed to determine which to include for optimal fitting and prediction.

The database consists of 33 proteins. Twenty-four of these fall into six structurally related groups, and nine are structurally unique. The former are SH3 domains [1NYF (82 to 148), 1PKS, 1SHG, and 1SRL], Ig-like β-sandwiches [1FNF (1326 to 1415), 1FNF (1416 to 1509), 1HNG, 1TEN (802 to 891), 1TIT, and 1WIT], members of the acylphosphatase family (1APS, 1HDN, 1PBA, 1URN, and 2HQI), cytochromes (1HRC, 1HRC-oxidized, 1YCC), cold shock proteins [1CSP and 1MJC (2 to 70)], λ-repressor variants (1LMB wild type and G46A/G48A), and ubiquitin variants (1UBQ wild type and V26A). The remainder of the proteins are 1COA (20 to 83), 1DIV (1 to 56), 1FKB, 1IMQ, 2ABD, 2AIT, 2PDD, 2PTL (94 to 155), and 2VIK. Numbers in parentheses indicate the residue numbers of the domain or fragment studied.

To cross-validate the results, each group of structurally related proteins is left out of the training set in turn and used to test the network. Such a partitioning scheme (in contrast to a jackknife one, for example) minimizes the likelihood of biasing the results in favor of structural descriptors (see Section II). Its use yields true predictions (denoted "cv") in contrast to fits of the data, in which all the proteins are included during the training (denoted "trn"). The latter tend to yield inflated accuracy statistics, but we describe them here as well for comparison with earlier studies [12,13,20,47], which failed to cross-validate their results [however, it should be noted that the relationship in Ref. 12 has been used successfully for blind predictions (K. W. Plaxco and D. Baker, personal communication)].

C. Single-Descriptor Models

We begin by examining the relationship between $\log k_f$ and each individual descriptor.

1. *Linear Correlations*

The first column of statistics given in Table I contains the Pearson linear correlation coefficients between the descriptor values (x) and $\log k_f$ ($r_{x,\log k_f}$). This is the statistical measure used by Plaxco et al. in their analysis of a subset of the descriptors considered here [12,14]. Consistent with their results, the two coefficients with the largest magnitudes are associated with the contact order (c and c/n). Several descriptors not examined by Plaxco et al. [12,14] exhibit $|r_{x,\log k_f}| > 0.5$ as well: the α-helix content and propensity (h and p_h), total helix content (a), and β-sheet content (e). Additional linear statistics are provided in Table V. Physical interpretations of the results are given in Section IV.E.

2. *Neural Network Predictions*

The second and third columns of statistics in Table I measure the ability of a single-input neural network to predict the folding rate. They contain Pearson

TABLE IV
Descriptor–Descriptor Pearson Linear Correlation Coefficients[a]

		ΔG	$\Delta G/n$	m	m/n	n	n_c	c	c/n	h	e	t	s	g	b	o	a	P_h	P_e	P_o	p_h	p_e	p_o	q_e	q_a
0	ΔG		0.94	0.50	0.38	0.50	0.37	−0.02	−0.29	0.24	−0.41	0.26	−0.05	−0.34	0.02	0.28	0.20	0.01	−0.17	0.16	0.06	−0.18	0.10	−0.04	−0.17
1	$\Delta G/n$	0.93		0.43	0.40	0.20	0.16	−0.16	−0.25	0.20	−0.39	0.19	0.03	−0.25	0.22	0.30	0.16	−0.01	−0.07	0.09	0.05	−0.08	0.00	−0.02	−0.02
2	m	0.28	0.27		0.95	0.39	0.31	0.12	−0.09	0.02	−0.06	−0.05	−0.01	−0.27	−0.08	0.22	−0.01	−0.11	0.04	0.14	−0.13	0.08	0.14	0.20	0.06
3	m/n	0.21	0.26	0.98		0.12	0.13	0.01	−0.04	−0.03	−0.02	−0.12	0.06	−0.16	0.06	0.20	−0.05	−0.12	0.13	0.06	−0.13	0.16	0.04	0.21	0.21
4	n	0.21	−0.14	−0.10	−0.27		0.73	0.48	−0.10	0.17	−0.15	0.27	−0.17	−0.40	−0.47	0.07	0.12	0.02	−0.26	0.23	0.02	−0.27	0.28	−0.12	−0.41
5	n_c	−0.19	−0.38	−0.20	−0.28	0.60		0.17	−0.28	0.50	−0.44	0.14	−0.15	−0.34	−0.42	−0.12	0.46	0.34	−0.52	−0.03	0.34	−0.55	0.02	0.00	−0.19
6	c	0.11	0.00	−0.34	−0.43	−0.34	−0.02		0.81	−0.63	0.66	−0.03	0.29	−0.28	−0.22	0.18	−0.67	−0.49	0.51	0.28	−0.56	0.50	0.44	−0.26	−0.46
7	c/n	−0.08	0.18	−0.23	−0.12	−0.68	−0.54	0.43		−0.84	0.83	−0.22	0.47	−0.10	0.08	0.24	−0.86	−0.61	0.77	0.20	−0.68	0.77	0.35	−0.18	−0.26
8	h	−0.01	−0.06	−0.10	−0.08	0.08	0.39	−0.39	−0.39		−0.89	−0.01	−0.22	−0.11	−0.34	−0.41	0.99	0.83	−0.82	−0.51	0.88	−0.88	−0.60	0.23	0.24
9	e	−0.18	−0.12	0.13	0.12	−0.15	−0.39	0.44	0.45	−0.75		0.00	0.07	0.02	0.12	0.00	−0.89	−0.64	0.77	0.26	−0.71	0.80	0.38	−0.11	−0.18
10	t	−0.15	−0.26	−0.35	−0.36	0.29	0.34	0.02	−0.27	0.08	−0.17		−0.70	0.20	0.32	−0.27	0.01	−0.17	0.01	0.26	−0.11	0.00	0.21	−0.23	−0.14
11	s	0.30	0.34	0.01	0.01	0.04	0.02	−0.05	0.04	0.20	−0.35	−0.52		−0.38	−0.25	0.45	−0.27	−0.20	0.32	0.00	−0.23	0.31	0.06	0.01	0.02
12	g	−0.05	−0.15	−0.14	−0.16	0.17	0.19	−0.11	−0.28	−0.01	−0.31	0.36	−0.55		0.42	−0.09	0.02	0.07	−0.09	−0.03	0.09	−0.06	−0.10	−0.29	0.15
13	b	−0.04	0.08	−0.05	−0.02	−0.26	0.00	0.03	0.30	−0.47	0.09	0.15	−0.34	0.49		0.26	−0.29	−0.32	0.31	0.20	−0.27	0.32	0.13	−0.22	−0.03
14	o	0.26	0.33	0.29	0.28	−0.13	−0.42	0.08	0.23	−0.73	0.24	−0.41	0.23	0.08	0.43		−0.42	−0.50	0.22	0.59	−0.50	0.31	0.59	−0.10	−0.23
15	a	−0.02	−0.10	−0.13	−0.12	0.11	0.43	−0.41	−0.45	0.97	−0.81	0.16	0.07	0.22	−0.34	−0.70		0.84	−0.84	−0.51	0.89	−0.89	−0.62	0.19	0.27
16	P_h	0.23	0.19	−0.20	−0.19	0.04	0.13	−0.45	−0.38	0.68	−0.66	0.05	0.11	0.33	−0.19	−0.41	0.74		−0.80	−0.80	0.99	−0.86	−0.83	0.27	0.23
17	P_e	0.22	0.32	0.31	0.33	−0.17	−0.08	0.31	0.43	−0.33	0.33	−0.13	0.25	−0.62	0.16	0.26	−0.46	−0.69		0.29	−0.83	0.98	0.38	−0.18	−0.03
18	P_o	−0.51	−0.53	0.01	−0.01	0.08	−0.12	0.35	0.16	−0.66	0.63	0.03	−0.36	0.07	0.13	0.35	−0.62	−0.79	0.10		−0.76	0.41	0.96	−0.25	−0.34
19	p_h	0.21	0.16	−0.29	−0.28	0.06	0.15	−0.36	−0.33	0.74	−0.67	0.07	0.10	0.32	−0.24	−0.51	0.79	0.98	−0.68	−0.77		−0.89	−0.82	0.21	0.22
20	p_e	0.12	0.22	0.42	0.43	−0.16	−0.19	0.24	0.37	−0.55	0.48	−0.22	0.25	−0.60	0.19	0.47	−0.67	−0.80	0.93	0.31	−0.83		0.48	−0.15	−0.06
21	p_o	−0.47	−0.48	0.06	0.04	0.05	−0.08	0.35	0.19	−0.69	0.63	0.09	−0.41	0.06	0.21	0.38	−0.65	−0.84	0.21	0.97	−0.84	0.39		−0.21	−0.36
22	q_e	0.18	0.26	0.53	0.53	−0.21	−0.34	−0.42	−0.14	0.08	0.06	−0.48	0.14	−0.22	−0.13	0.12	0.03	0.03	−0.02	−0.02	0.00	0.06	−0.06		0.37
23	q_a	0.20	0.15	0.42	0.39	0.18	0.21	−0.55	−0.58	0.29	−0.20	0.12	0.09	−0.30	−0.15	−0.19	0.21	0.11	0.17	−0.30	0.05	0.16	−0.24	0.54	

[a]Data for all 33 proteins are above the diagonal, and data for the 14 proteins with $c > 21$ are below the diagonal.

TABLE V
Linear Regression Statistics for $\log k_f$

Index	Symbol	r_{trn}	r_{cv}	q^2_{cv}
0	ΔG	0.29	−0.02	−0.09
1	$\Delta G/n$	0.37	0.13	−0.05
2	m	0.04	−0.65	−0.19
3	m/n	0.04	−0.52	−0.20
4	n	0.10	−0.53	−0.27
5	n_c	0.08	−0.60	−0.24
6	c	0.73	0.70	0.48
7	c/n	0.79	0.77	0.59
8	h	0.63	0.55	0.30
9	e	0.67	0.59	0.34
10	t	0.04	−0.76	−0.23
11	s	0.11	−0.52	−0.19
12	g	0.01	−0.75	−0.41
13	b	0.15	−0.43	−0.26
14	o	0.05	−0.74	−0.31
15	a	0.63	0.57	0.32
16	P_h	0.47	0.29	0.06
17	P_e	0.48	0.31	0.08
18	P_o	0.27	−0.27	−0.28
19	p_h	0.51	0.37	0.13
20	p_e	0.47	0.28	0.05
21	p_o	0.40	0.07	−0.09
22	q_e	0.21	−0.21	−0.14
23	q_a	0.40	0.12	−0.07

linear correlation coefficients (r_{trn} and r_{cv}) between observed and calculated values of $\log k_f$; thus, only positive values of r are significant. Because there are only 24 different input possibilities, it is feasible to consider each one in turn, so that use of a genetic algorithm is not necessary at this stage. However, the NN weights depend on the random number generator seed through the training procedure. Consequently, for each descriptor, the network was trained independently with ten different seeds. The maximum correlation coefficient for each set of 10 networks corresponding to a particular descriptor is listed in Table I; the average standard deviation for a given descriptor was 0.03 for r_{trn} and 0.06 for r_{cv}.

As stated above, the coefficients denoted "trn" are for results obtained with networks trained on all 33 proteins; in other words, they are not true predictions since all the data are included in the training set. For descriptors that are linearly related to $\log k_f$, r_{trn} is expected to be comparable in magnitude to $r_{x,\log k_f}$ (in fact, for linear regression, $r_{trn} = |r_{x,\log k_f}|$), whereas, for ones that are non-linearly related, it should be higher. Thus, r_{trn} can be viewed as essentially a nonlinear version of the statistic employed in Ref. 12. Accordingly, most of the descriptors that exhibit high r_{trn} were included in the analysis of $r_{x,\log k_f}$.

The coefficients denoted "cv" are for the predictions obtained with the structurally based cross-validation scheme. Negative values of r_{cv} indicate that the accuracy of the network is lower than that which would be obtained from random guesses. If a network fails in this way when confronted with novel test data, it has derived a spurious relationship by memorizing the information in the training set at the expense of learning more general rules. The highest r_{cv} do correspond to the highest r_{trn}, but overall the cross-validated coefficients are much lower. The large differences between r_{trn} and r_{cv} in many cases (Table I) make clear that the former is a relatively indiscriminate statistic for such a small database. If linear regression is used, r_{trn} and r_{cv} are often closer due to the decreased flexibility of the fitting method (Table V). However, such an approach fails to identify nonlinear relationships and can hide complexities in the results.

In summary, the contact order yields relatively good prediction of log k_f but is not alone in doing so. Several measures of the propensity of the sequence for a given structure also exhibit significant relationships with the folding rate. Although r_{cv} values for the various descriptors obtained from the secondary structure prediction program (indices 16 to 21 in Table I) are lower than those for measures of the known native structure (indices 6 to 15), the former correlations may be sufficiently high that the calculated descriptors could be used to identify particularly fast or slow proteins without the need for high-resolution structures. The stability, which has been suggested to be of importance based on model studies, exhibits no clear relation to the folding rate. An essential additional point of the single-descriptor analysis is that large differences are observed between most of the values obtained with and without cross-validation. This highlights the need for care in assessing the significance of correlations when working with small numbers of sequences.

D. Multiple-Descriptor Models

We present results for two- and three-descriptor models; addition of a fourth descriptor yielded no significant improvement in predictive accuracy. In the two-descriptor case there are only 276 possible input combinations, so we examine each explicitly, whereas, in the three-descriptor case there are 2024, so we use the genetic algorithm (GA) to optimize the descriptor selection. Use of the GA in the two-descriptor case gives models of comparable quality to the exhaustive search, but this test of the algorithm is not very stringent because the space of input combinations is small. Because both the GA and the NN depend on the random number generator seed, several trials were performed in each case (as detailed in Section IV.D.2).

1. Two Descriptors

The best five two-descriptor models are shown in Table VI, and selected examples to illustrate the types of behavior that are observed are shown in Fig. 2.

There is a significant increase in fitting ability (training) and, more importantly, in predictive accuracy (cross-validation) upon adding a second descriptor. In Figure 2, we see that the squares (□) tend to be closer to the ideal line than the circles (○), particularly for lower log k_f (slower-folding proteins). To quantitate the improvement, we calculated Wold's E statistic from the q^2_{cv} values (Table VI). While these figures suggested to us that the additional descriptors significantly improve the accuracies of the cross-validated predictions, general confidence limits are not straightforward to calculate. Consequently, we did the following. We shuffled the values of each secondary descriptor (other than c/n) 10 times and then trained neural networks to predict log k_f as for the actual data. Averages and standard deviations of the correlation coefficients are reported in Table VII. We see that, even though the r_{trn} values are comparable to those in Table VI, the

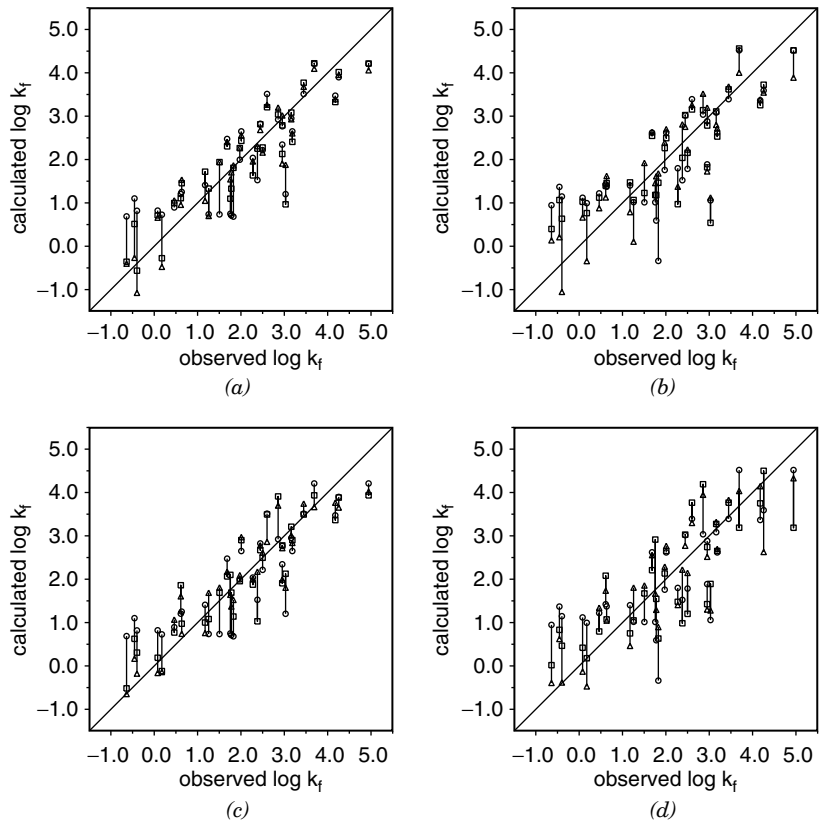


Figure 2. Comparison of observed and calculated values of log k_f for selected models. (a and b) c/n(○); c/n and $\Delta G/n$ (□); and c/n, $\Delta G/n$ and p_e(△). (*c* and *d*) c/n(○); c/n and n_c (□); and $c/n, n_c,$ and ΔG(△). (a and c) Training set fits. (b and d) Cross-validated predictions.

TABLE VI
The Best (as Measured by r_{cv}) Five Two-Descriptor Models Obtained by Examining All Possible Combinations for Ten Different Random Number Generator Seeds[a]

Descriptors		r_{trn}	r_{cv}	q^2_{cv}	E
c/n	$\Delta G/n$	0.89	0.81	0.66	0.74
c/n	P_h	0.87	0.80	0.63	0.81
c/n	n_c	0.89	0.79	0.62	0.82
c/n	p_h	0.86	0.77	0.57	0.93
c/n	q_a	0.84	0.77	0.59	0.89

[a]For the calculation of E, q^2_{cv} was compared with that for c/n. Statistics for linear regression and additional measures of the predictive accuracy are available in Tables VII and VIII.

r_{cv} values are close to that for c/n by itself (Table I); the NN ignores the randomized descriptor. The fact that the r_{cv} values for the actual data are two to four standard deviations above the average r_{cv} values for the randomized data demonstrates that the improvement is significant and is not due to the increase in the number of fitting parameters.

The best predictions are obtained with $\Delta G/n$ paired with c/n (ΔG with c is the sixth best set of inputs with $r_{cv} = 0.77$ and $E = 0.76$) This combination of input descriptors was investigated previously [15], but it is of interest that it ranks first in the exhaustive search performed here. To better understand the physical basis for the correlations, we show the dependence of log k_f on c/n and $\Delta G/n$ in Fig. 3a. When c/n is small ($c/n \leq 19$; mainly α-helical proteins), folding is always fast ($k_f > 400\,\mathrm{s}^{-1}$), whereas when c/n is large ($c/n \geq 25$; either mixed-α/β or β-sheet proteins), the rate spans over three orders of magnitude. Thus, proteins with lower contact orders fold fast regardless of their stabilities, whereas for those with higher contact orders, the rate increases with $\Delta G/n$. As described in Ref. 15, a single-input neural network can be trained to predict log k_f from ΔG for the 14 proteins with $c > 21$ (Fig. 4); $r_{trn} = 0.81$, and $r_{cv} = 0.64$, which confirms that stability plays a significant role in determining the folding rates of mixed-α/β and β-sheet proteins. For these 14

TABLE VII
Randomization Tests for the Models in Table VI[a]

Descriptors		r_{trn}	r_{cv}	q^2_{cv}
c/n	$\Delta G/n$	0.83 ± 0.01	0.71 ± 0.03	0.49 ± 0.04
c/n	P_h	0.84 ± 0.03	0.68 ± 0.07	0.43 ± 0.12
c/n	n_c	0.87 ± 0.02	0.69 ± 0.04	0.46 ± 0.05
c/n	p_h	0.84 ± 0.02	0.68 ± 0.06	0.42 ± 0.10
c/n	q_a	0.84 ± 0.00	0.68 ± 0.07	0.44 ± 0.11

[a]In each case, the second descriptor listed was shuffled 10 times, and the networks were trained as for the original data. Values shown are averages for the 10 trials; ranges indicate standard deviations.

TABLE VIII
Linear Regression Statistics for the Models in Table VI

Descriptors		r_{trn}	r_{cv}	q^2_{cv}	E
c/n	$\Delta G/n$	0.81	0.72	0.47	1.27
c/n	P_h	0.79	0.75	0.57	1.04
c/n	n_c	0.82	0.79	0.62	0.92
c/n	p_h	0.79	0.75	0.56	1.05
c/n	q_a	0.80	0.77	0.60	0.97

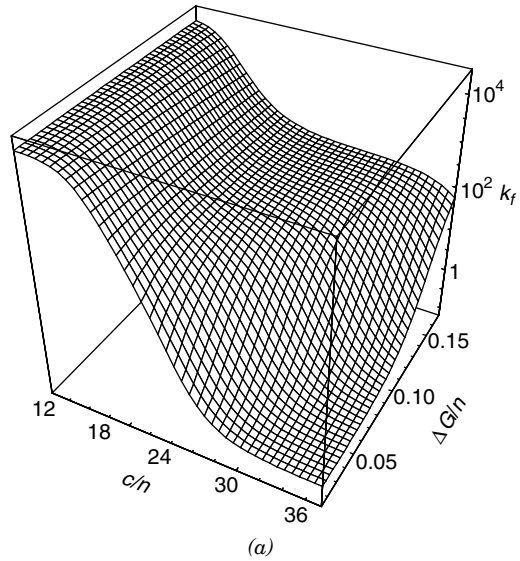
proteins, $r_{\Delta G,\log k_f} = 0.80$ while $r_{c,\log k_f} = -0.22; E = (1 - q^2_{c,\Delta G})/(1 - q^2_c) = 0.23$.

Two of the other models in Table VI combine the contact order with a measure of the α-helical propensity: c/n with either P_h or p_h. These pairings essentially reflect the results of the previous section. The remaining model couples c/n with n_c, which reveals a secondary dependence on protein size. Consistent with the sign of $r_{n_c,\log k_f}$ (Table I), the functional dependences of $\log k_f$ on these descriptors for the models in Table VI indicate that shorter proteins fold faster than longer ones (Fig. 3b).

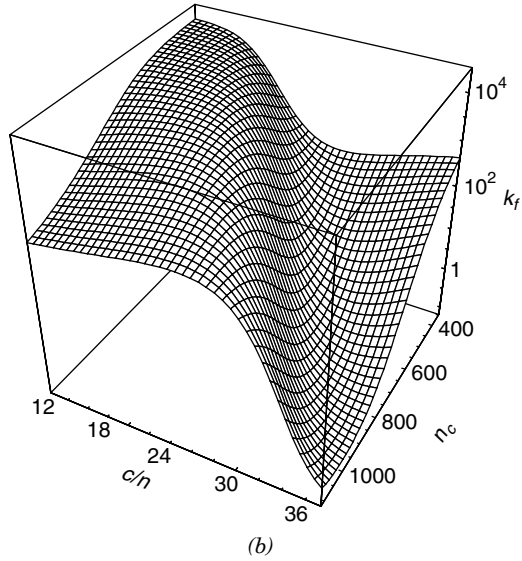
2. *Three Descriptors*

As mentioned above, there are 2024 possible combinations of three descriptors, so we use a GA to identify the inputs that are likely to yield the greatest predictive accuracy. Use of the GA requires selection of a particular measure of predictive accuracy to decide which models to keep at each cycle. Because we are interested primarily in cross-validated predictions, r_{cv} is a natural choice. However, the structurally based partitioning scheme is less straightforward to automate than a jackknife one. Consequently, for the GNN, we used the Pearson linear correlation coefficient for the jackknife cross-validated outputs (r_{jck}) and subsequently tested each selected combination of descriptors with the structurally based cross-validation scheme (r_{cv}). We performed five GNN trials, from each of which we saved the best 20 models. Of these 100 models, 46 were unique, and each of these was subjected to 10 trials with the structurally based cross-validation scheme.

In general, the GA combines the descriptors that were identified above by the two-dimensional exhaustive search ($c, c/n, \Delta G, \Delta G/n$, and n_c) to further refine the predictions (Tables IX to XI and Fig. 2). The propensity for sheet structure (p_e) appears in two of the five models; not surprisingly, it is strongly anti-correlated with the propensity for helical structure, which appeared in Table VI ($r_{p_e,p_h} = -0.89$). In considering these results, it is necessary to keep in mind that the database is small, so that there is a danger of overfitting (but see Table X). Nevertheless, given this disclaimer, we see that simultaneous consideration of multiple descriptors improves prediction of the folding rate and that both the



(a)



(b)

Figure 3. Functional dependence of calculated folding rate (k_f, in s^{-1}) on the normalized contact order (c/n) and either (a) the normalized stability ($\Delta G/n$ in kcal/mol) or (b) the total number of atomic contacts (n_c).

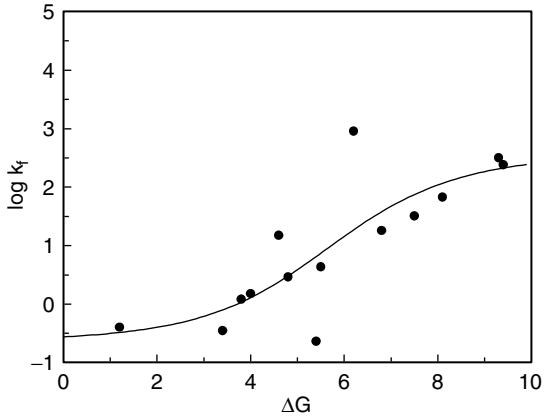


Figure 4. Observed (points) and calculated (line) log k_f as a function of the stability in kcal/mol for the 14 proteins in the database with $c > 21$.

size and the stability play significant secondary roles that could not have been anticipated from the single-descriptor analyses.

E. Physical Bases of the Observed Correlations

Consistent with earlier, single-descriptor linear analyses of protein folding [12,13,50], the primary determinants of the folding rate are measures that characterize the native structure; that is, proteins with more sequentially local interactions tend to fold faster. As discussed below, the equilibrium structure and the kinetics are connected by the fact that the structure of the transition state resembles that of the native state in many small proteins [50]. Thus, the kinetics and the underlying thermodynamics of the reaction are affected in a similar way, in accord with linear free energy relations.

The microscopic origin for the statistical dependence of the folding kinetics on the structure is the stochastic diffusive search that is required to find the

TABLE IX
The Best (as Measured by r_{cv}) Five Unique Three-Descriptor Models Obtained from the GNN Protocol for Ten Different Random Number Generator Seeds[a]

Descriptors			r_{trn}	r_{jck}	r_{cv}	q^2_{cv}	E
c/n	$\Delta G/n$	p_e	0.92	0.84	0.86	0.74	0.76
c/n	ΔG	n_c	0.93	0.84	0.84	0.70	0.80
c/n	$\Delta G/n$	n_c	0.92	0.81	0.83	0.67	0.97
c/n	ΔG	c	0.90	0.83	0.83	0.66	0.81
c/n	ΔG	p_e	0.91	0.80	0.83	0.67	0.72

[a]For the calculation of E, q^2_{cv} was compared with the highest observed q^2_{cv} of the six possible two-descriptor models that could be formed from the three selected inputs (corresponding to the unshuffled pair in Table X). Statistics for linear regression and additional measures of the predictive accuracy are available in Table X and XI.

TABLE X
Randomization Tests for the Models in Table IX

Descriptors			Randomized	r_{trn}	r_{cv}	q^2_{cv}
c/n	$\Delta G/n$	p_e	p_e	0.89 ± 0.02	0.80 ± 0.03	0.61 ± 0.07
c/n	ΔG	n_c	ΔG	0.88 ± 0.02	0.72 ± 0.05	0.48 ± 0.10
c/n	$\Delta G/n$	n_c	n_c	0.89 ± 0.01	0.74 ± 0.04	0.49 ± 0.09
c/n	ΔG	c	c/n	0.89 ± 0.01	0.71 ± 0.04	0.46 ± 0.08
c/n	ΔG	p_e	ΔG	0.88 ± 0.01	0.69 ± 0.06	0.41 ± 0.10

transition state. As described in the formulation of the ''hydrophobic zipper hypothesis'' [51,52] and in the statistical analyses of 125-residue lattice models [10,11], having sequentially short-range contacts in the transition state should increase the folding rate for two reasons. First, such contacts are found more readily because there are fewer conformations to search (the number grows exponentially with loop length). Second, making sequentially long-range contacts costs more entropy because they constrain the chain to a greater degree. These advantages correspond to different components of the macroscopic rate law [$k_f = A(T)\exp(-\Delta G/k_B T)$]. In this regard, it is necessary to keep in mind that the preexponential factor can be nontrivial for protein folding [53,54]. If $A(T)$ is sufficiently large, there is a separation of time scales; the protein reaches an effective equilibrium within the unfolded state rapidly, and the rate is dominated by the time required to surmount the barrier [55]. In this case, the observed statistical dependence on the structure implies that the barrier is entropic (as in Fig. 3a of Ref. 1 and Figs. 6 and 7 of Ref. 36). Based on these ideas, Fersht recently derived a simple relationship to show that changes in contact order are directly proportional to changes in log k_f [50]. On the other hand, if $A(T)$ is small, there is no separation of time scales. Because a dependence on the structure enters through the preexponential factor in this case, the barrier, if there is one, could be either entropic or energetic (as in Fig. 3b of Ref. 1).

Free energy surfaces for folding have now been determined for high-resolution (all-atom) models of several peptides and proteins [72–77]. For both α-helical and β-hairpin peptides, decomposition of the surfaces into contributions from the effective energies (which include the full solvent free

TABLE XI
Linear Regression Statistics for the Models in Table IX

Descriptors			r_{trn}	r_{cv}	q^2_{cv}	E
c/n	$\Delta G/n$	p_e	0.83	0.71	0.46	1.57
c/n	ΔG	n_c	0.84	0.73	0.46	1.42
c/n	$\Delta G/n$	n_c	0.84	0.76	0.55	1.29
c/n	ΔG	c	0.83	0.71	0.41	1.40
c/n	ΔG	p_e	0.82	0.69	0.38	1.34

energies) and configurational entropies indicated that the free energy barriers derive primarily from the fact that the entropy decreases more rapidly than the energy [75–77], as in Ref. 36 discussed above. However, consistent with the statistical analyses of proteins, differences in secondary structure content correspond to differences in the general shapes of the free energy surfaces. For α-helical sequences, the transition states tend to be less folded, and secondary and tertiary structure form concurrently [72,77]. For peptides and proteins which contain β-hairpins and β-sheets, a collapse to a native-like radius of gyration occurs first, and rearrangement to the native state follows wihout significant expansion [73–75]. At least for peptides at elevated temperatures [76,77], determination of the rate of diffusion on the free energy surfaces, which relates directly to the pre-exponential factor in the rate law [53], should now be possible but has not been done and would be of interest.

In connecting these ideas with earlier phenomenological models, it is not obvious how to reconcile the dependence of the rate on the structure with a nucleation mechanism, as in Ref. 50. The statistical relationship suggests that the transition state contains a considerable amount of native structure, while a nucleus, in the classic sense of the word, is a small part of the structure. However, it could be that a limited number of native contacts (i.e., those in the nucleus) are sufficient to confine the transition state ensemble to a native-like fold. This idea is supported by a recent analysis of the folding transition state of acylphosphatase in which key residues, as determined by a ϕ value analysis, play a critical role [56].

V. UNFOLDING RATES OF PROTEINS

To function, a protein must not only fold (kinetic criterion) but populate its native state for a significant fraction of the time (thermodynamic criterion). The unfolding rate (k_u) as well as k_f contribute to the equilibrium constant, which determines to what degree the latter condition is satisfied. To find the factors that affect the unfolding rate, we carried out an analysis for $\log k_u$. Rate data for unfolding in water were not available for three of the proteins (2HQI, 1YCC, and 1HRC-oxidized), so these were excluded from the analysis; the choice of descriptors was the same.

For single-descriptor models, the best cross-validated predictions are obtained with the contact order (c and c/n), the free energy of unfolding (ΔG and $\Delta G/n$), and the buried surface area (m) (Table XII). The strong dependence of the unfolding rate on the contact order for these proteins is somewhat surprising because no significant correlation was observed in a previous study of a database of 24 proteins [14], 19 of which are included here. For those 19 proteins we have $r_{\Delta G,\log k_u} = -0.61$, $r_{c,\log k_u} = -0.56$ and $r_{c/n,\log k_u} = -0.45$, whereas for the 11 additional proteins included in the present analysis of the unfolding rate we have $r_{\Delta G,\log k_u} = -0.64$, $r_{c,\log k_u} = -0.85$, and $r_{c/n,\log k_u} = -0.83$. The proteins

TABLE XII
Single-Input Correlations for Unfolding Rates

Index	Symbol	$r_{x,\log k_u}$	r_{trn}	r_{cv}	q^2_{cv}
0	ΔG	−0.64	0.69	0.53	0.21
1	$\Delta G/n$	−0.45	0.55	0.40	0.12
2	m	−0.41	0.61	0.45	0.14
3	m/n	−0.31	0.36	0.08	−0.11
4	n	−0.43	0.58	0.09	−0.09
5	n_c	−0.40	0.53	0.09	−0.05
6	c	−0.68	0.77	0.67	0.44
7	c/n	−0.58	0.69	0.52	0.20
8	h	0.40	0.49	−0.57	−0.86
9	e	−0.34	0.53	0.16	−0.06
10	t	−0.01	0.39	−0.25	−0.12
11	s	−0.08	0.26	−0.19	−0.24
12	g	0.03	0.36	−0.16	−0.32
13	b	−0.27	0.27	−0.19	−0.23
14	o	−0.20	0.55	0.15	−0.08
15	a	0.40	0.50	−0.27	−0.27
16	P_h	0.29	0.53	−0.64	−0.32
17	P_e	−0.28	0.30	−0.38	−0.47
18	P_o	−0.20	0.52	−0.22	−0.20
19	p_h	0.29	0.50	−0.31	−0.42
20	p_e	−0.23	0.50	−0.38	−0.40
21	p_o	−0.27	0.49	−0.56	−0.11
22	q_e	0.14	0.35	−0.11	−0.14
23	q_a	0.24	0.48	0.19	−0.06

that appear to be primarily responsible for decreasing the correlation with the free energy of unfolding and increasing the correlation with the contact order are the helical proteins—in particular, 2PDD and 1LMB. Because for the 30 proteins considered in this section there is no significant correlation between the contact order and either the free energy of unfolding ($r_{\Delta G,c} = 0.28$) or the amount of buried surface area ($r_{m,c} = 0.23$), higher predictive accuracy is obtained by combining these descriptors (Table XIII). Only a slight improvement was observed upon adding a third descriptor.

We end this section by noting that, for these 30 proteins, there is a significant correlation between the folding and unfolding rates ($r_{\log k_f,\log k_u} = 0.59$). At least in the case that k_f and k_u are determined by an entropic barrier (Section IV.E), this relationship can be understood in the following way. Because all the proteins are roughly the same size, the stability of the native state does not depend on contact order (for the overall reaction, $\Delta S \propto n$). Changes to c that raise or lower the free energy of the transition state (TS) relative to the fixed endpoints (U and F) will change Δ_{U-TS} and ΔG_{F-TS} in the same manner. This dependence of the activation free energies is the basis not only for the correlation of $\log k_u$ with $\log k_f$ but also that with c.

TABLE XIII
The Best (as Measured by r_{cv}) Five Two-Descriptor Models for the Unfolding Rates

Descriptors		r_{trn}	r_{cv}	q^2_{cv}	E
c	$\Delta G/n$	0.90	0.85	0.71	0.53
c	ΔG	0.88	0.81	0.66	0.62
c/n	ΔG	0.89	0.80	0.61	0.49
c	m	0.83	0.73	0.53	0.85
c	m/n	0.90	0.71	0.49	0.92

VI. HOMOLOGOUS PROTEINS

Information about the transition state of a protein can be obtained from protein engineering experiments in which one compares the effects of mutations on the folding rate to their effects on the overall stability (ϕ values). Several proteins have been mutated extensively, and their kinetics have been measured. The fact that proteins with related structures but low sequence homologies are found to have similar transition states has been taken to support the relation between native structure and folding behavior; this is the case for the transition states of the src [57] and α-spectrin [58] SH3 domains, which have 36% sequence homology. A particularly interesting transition state comparison involves acylphosphatase (AcP) [59] and procarboxypeptidase A2 [60]. These two proteins fold to sandwich structures with two α-helices packed against a five- or four-stranded antiparallel sheet, respectively. Although their sequences have only 13% identity, the average ϕ values for all elements of secondary structure (except one, β-strand 4) are almost the same. Moreover, it has been suggested that the reason that procarboxypeptidase A2 folds about 4000 times faster than AcP is that the transition state of the latter involves longer loops and secondary structure elements; consistent with this observation, there is a strong correlation between $\log k_f$ and the contact order for proteins with this fold [59].

The dependence of the folding rate on the stability can be evaluated by measuring the kinetics of a family of proteins with native states that have similar structures but different ΔG values. Such an analysis was made recently for a set of six immunoglobulin-like β-sandwich domains [61]. They have stabilities that are distributed relatively uniformly over the range $1.2 \leq \Delta G \leq 9.4$ kcal/mol (in contrast to the AcP family discussed above, for which four of the five members have $3.8 \leq \Delta G \leq 5.4$ kcal/mol). Although there is some variation in the detailed structures of these six proteins, using the definition of the contact order given in Section II, all of them have $c/n > 28$ (for these six, $28.22 \leq c/n \leq 32.53$; for the five members of the AcP family, $25.83 \leq c/n \leq 35.08$; for all 33 proteins, $12.21 \leq c/n \leq 37.32$). In accord with the functional dependence on ΔG shown in Figs. 3a and 4, a strong positive correlation between $\log k_f$ and ΔG was observed for this family ($r_{\Delta G, \log k_f} = 0.99$). The data

TABLE XIV
Relation Between Stability and Folding Rate for Six Two-State Proteins That Have Been Mutated Extensively[a]

Protein	Reference	c	Number of Mutants	$r_{\Delta G,\log k_f}$	r_{trn}	r_{jck}
Acylphosphatase	59	34.4	25	0.614	0.667	0.386
Procarboxypeptidase A2	60	20.7	19	0.531	0.712	0.464
src SH3	57	20.5	58	0.552	0.556	0.408
α-Spectrin SH3	58	18.0	18	0.481	0.476	0.099
CI2	71	16.1	86	0.554	0.606	0.519
λ-Repressor	64	9.8	9	0.720	0.760	0.307

[a]The coefficients r_{trn} and r_{jck} are for single-input (ΔG) neural networks. The α-spectrin SH3 domain values are those for pH 7; the src SH3 domain values are for pH 6. The λ-repressor values are for 2 M urea.

suggest that for a given structural family with significant variation in ΔG, the folding rates of individual sequences are determined by their stabilities.

This conclusion is consistent with the fact that both $\log k_f$ and $\log k_u$ typically vary linearly with the stability of the native state as a protein is mutated. Such Brønsted behavior has been used in protein engineering studies to argue that fractional ϕ values derive from partial structure formation rather than multiple parallel folding pathways [62]. Correlation coefficients for published folding rates of mutants of six two-state proteins are given in Table XIV. For the most part, there is a strong, essentially linear relation that is reasonably robust to jackknife cross-validation. For all the sequences, increases in stability tend to accelerate folding. Similar behavior is obtained simply by varying the conditions to affect the stability of a protein (for example, see Fig. 2a of Ref. 14). This analysis thus confirms that the stability is an important secondary factor in determining folding rate. As described in Ref. 9, in accord with the Hammond postulate [34], stabilizing the native state of a protein in most cases also lowers the energy of its transition state relative to the unfolded state and thus increases the folding rate.

VII. RELATING PROTEIN AND LATTICE MODEL STUDIES

The fact that the folding (and unfolding) kinetics of relatively small, two-state proteins can be predicted with reasonable accuracy from global features of the native state like the contact order, stability, and number of contacts supports the idea that the details of protein structure are not required to capture the key features of protein folding, so that reduced representations should be adequate. However, the most widely used simple heteropolymer models, those restricted to a simple cubic lattice, predict that stability is more important than native structure, in contrast to the experimental data for proteins. In this section we seek to understand why lattice models differ from proteins in this regard. Doing so is of importance because complete details of the folding trajectories of such models

can be obtained and used to test phenomenological models like those described in Section IV.A.

In the case of the 27-residue model described in Section III [6,9], it is likely that the chain length is too short for there to be contacts that are sufficiently long range to slow-folding. In the case of the 125-residue model, which is larger than all but one (2VIK) of the proteins considered in the present study, significant correlations between various measures that characterize the native structure and the folding behavior were observed [10,11] (it should be mentioned that, in contrast to the number of antiparallel sheet contacts discussed in Section III, the contact order is a poor measure for characterizing lattice model structure; $18.7 \leq c \leq 31.0$ for the 100 helical proteins in Refs. 10 and 11, whereas $17.2 \leq c \leq 32.0$ for the 100 sheet proteins). However, in the lattice model, the functional dependence of the folding stability is essentially the same regardless of the native structure; at a particular threshold value of the stability (which varies only slightly with the number of antiparallel sheet contacts), the folding ability rises rapidly and then levels off [11,37]. There are two likely reasons that the functional dependence is much simpler than that for proteins (Fig. 3a). First, the 125-residue sequences were energetically optimized to observe folding on the time scale of feasible simulations and are thus expected to correspond to the more stable region in Fig. 3a. Second, due to the highly restricted conformational space of the lattice and the choice of move set, helices that form in isolation cannot diffuse as semirigid units [49]; as a result, lattice models cannot correctly capture the lower contact order region of Fig. 3a. Once one restricts oneself to the remaining part of Fig. 3a, the behaviors observed in the lattice models and proteins are consistent; in both, the folding ability increases sigmoidally with the stability [compare Fig. 4 with Fig. 16 of Ref. 11 and Fig. 1 of Ref. 37]. It should be noted, however, that an exact correspondence is not expected because, in the lattice model [2,6–11] and related analytic [16–18] studies, the stability descriptors are calculated from effective energies that include solvent effects implicitly rather than from full free energies, while the experimental ΔG values include the protein configurational entropy as well. It would be useful in this regard to have experimental enthalpies of folding for the proteins considered.

VIII. CONCLUSIONS

In the present study a nonlinear, multiple-descriptor method was applied to the prediction of the logarithm of the folding rate constant for a set of 33 two- and weakly three-state proteins. With two (three) descriptors, the Pearson linear correlation coefficient between the observed values and the training set and cross-validated predictions reach 0.89 (0.93) and 0.81 (0.86), respectively. These results are to be compared with those obtained by using the contact order by itself: $r_{trn} = 0.83$ and $r_{cv} = 0.74$. In addition to the contact order, some measures

of the propensity of the sequence for a given structure also exhibited significant relationships with the folding rate; for example, $r_{cv} = 0.42$ for p_e. Although the propensity correlations are somewhat lower than those for measures obtained from the observed native structure, the sequence-based predictions may be sufficient to identify fast- or slow-folding proteins without the need for high-resolution structures. For example, using n and p_e, the folding rates for all 33 proteins, which range over almost six orders of magnitude, are predicted within a factor of 200; these (cross-validated) predictions are to be compared with those based on n_c and c/n, which are accurate within a factor of 60. In addition to the contact order, the size and stability play significant roles and are selected frequently for two- and three-descriptor models. Of particular interest is the finding that, for mixed-α/β and β-sheet proteins with higher contact orders ($c > 21$), the stability not only significantly improves the accuracy of multiple-descriptor models but gives excellent predictions by itself. The explicit or implicit inclusion of the stability in phenomenological models accounts for recent improvements in fitting experimental kinetic data [19,20,42]. Given the high quality of predictions that are obtained with the present analysis, further investigation of such correlations and their physical origins appear worthwhile, as has been suggested elsewhere [50].

Acknowledgments

A.R.D. is a Burroughs Wellcome Fund Hitchings-Elion Postdoctoral Fellow, and M.K. is the Eastman Visiting Professor at the Oxford Centre for Molecular Sciences. They would like to thank W. Graham Richards and Christopher M. Dobson for the hospitality that has been extended to them during their time in Oxford, John-Marc Chandonia for helpful discussions concerning the PRED2ARY program, and Jane Clarke for providing unpublished data for the proteins in Ref. 61. This work was supported in part by a grant from the National Science Foundation.

References

1. A. R. Dinner, A. Šali, L. J. Smith, C. M. Dobson, and M. Karplus, Understanding protein folding via free energy surfaces from theory and experiment. *Trends Biochem. Sci.* **25**, 331–339 (2000).
2. A. Šali, E. Shakhnovich, and M. Karplus, How does a protein fold? *Nature* **369**, 248–251 (1994).
3. Y. Zhou and M. Karplus, Interpreting the folding kinetics of helical proteins. *Nature* **401**, 400–403 (1999).
4. A. Fersht, *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*, W. H. Freeman, New York, 1999.
5. C. M. Dobson, P. A. Evans, and S. E. Radford, Understanding how proteins fold: The lysozyme story so far. *Trends Biochem. Sci.* **19**, 31–37 (1994).
6. A. Šali, E. Shakhnovich, and M. Karplus, Kinetics of protein folding: A lattice model study of the requirements for folding to the native state. *J. Mol. Biol.* **235**, 1614–1636 (1994).
7. H. S. Chan and K. A. Dill, Transition states and folding dynamics of proteins and heteropolymers. *J. Chem. Phys.* **100**, 9238–9257 (1994).
8. N. D. Socci and J. N. Onuchic, Folding kinetics of proteinlike heteropolymers. *J. Chem. Phys.* **101**, 1519–1528 (1994).
9. A. R. Dinner, V. Abkevich, E. Shakhnovich, and M. Karplus, Factors that affect the folding ability of proteins. *Proteins* **35**, 34–40 (1999).

10. A. R. Dinner, A. Šali, and M. Karplus, The folding mechanism of larger model proteins: Role of native structure. *Proc. Natl. Acad. Sci. USA* **93**, 8356–8361 (1996).
11. A. R. Dinner, S.-S. So, and M. Karplus, Use of quantitative structure–property relationships to predict the folding ability of model proteins. *Proteins* **33**, 177–203 (1998).
12. K. W. Plaxco, K. T. Simons, and D. Baker, Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985–994 (1998).
13. S. E. Jackson, How do small single-domain proteins fold? *Folding & Design* **3**, R81–R91 (1998).
14. K. W. Plaxco, K. T. Simons, I. Ruczinski, and D. Baker, Topology, stability, sequence, and length: Defining the determinants of two-state protein folding kinetics. *Biochemistry* **39**, 11177–11183 (2000).
15. A. R. Dinner and M. Karplus, The roles of stability and contact order in determining protein folding rates. *Nature Struct. Biol.* **8**, 21–22 (2001).
16. J. D. Bryngelson and P. G. Wolynes, Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. USA* **84**, 7524–7528 (1987).
17. J. D. Bryngelson and P. G. Wolynes, Intermediates and barrier crossing in a random energy model (with applications to protein folding). *J. Phys. Chem.* **93**, 6902–6915 (1989).
18. E. I. Shakhnovich and A. M. Gutin, Formation of unique structure in polypeptide chains: Theoretical investigation with the aid of a replica approach. *Biophys. Chem.* **34**, 187–199 (1989).
19. E. Alm and D. Baker, Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl. Acad. Sci. USA* **96**, 11305–11310 (1999).
20. V. Muñoz and W. A. Eaton, A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl. Acad. Sci. USA* **96**, 11311–11316 (1999).
21. S. Wold, Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics* **20**, 397–405 (1978).
22. S. Wold, Validation of QSARs. *Quant. Struct.–Act. Relat.* **10**, 191–193 (1991).
23. J. A. Malpass, D. W. Salt, M. G. Ford, E. W. Wynn, and J. Livingstone, Continuum regression: A new algorithm for the prediction of biological activity, in *Methods and Principles of Medicinal Chemistry*, Vol. 3, H. van de Waterbeemd, ed., VCH Publishers, New York, 1994, pp. 163–189.
24. J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA* **79**, 2554–2558 (1982).
25. J. Hertz, A. Krogh, and R. G. Palmer. *Introduction to the Theory of Neural Computation*, Addison-Wesley, Redwood City, CA, City, 1991.
26. D. T. Manallack and D. J. Livingstone, Neural networks in drug discovery: Have they lived up to their promise? *Eur. J. Med. Chem.* **34**, 195–208 (1999).
27. M. F. Møller, A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks* **6**, 525–533 (1993).
28. D. E. Clark, ed., *Evolutionary Algorithms in Molecular Design*, Wiley-VCH, Cambridge, 2000.
29. E. I. Shakhnovich and A. M. Gutin, Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl. Acad. Sci. USA* **90**, 7195–7199 (1993).
30. H. S. Chan and K. A. Dill, Protein folding in the landscape perspective: Chevron plots and non-Arrhenius kinetics. *Proteins* **30**, 2–33 (1998).
31. R. Unger and J. Moult, Local interactions dominate folding in a simple protein model. *J. Mol. Biol.* **259**, 988–994 (1996).
32. D. K. Klimov and D. Thirumalai, Criterion that determines the foldability of proteins. *Phys. Rev. Lett.* **76**, 4070–4073 (1996).
33. D. K. Klimov and D. Thirumalai, Factors governing the foldability of proteins. *Proteins* **26**, 411–441 (1996).

34. G. S. Hammond, A correlation of reaction rates. *J. Am. Chem. Soc.* **77**, 334–338 (1955).
35. V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, Specific nucleus as the transition state for protein folding: Evidence from the lattice model. *Biochemistry* **33**, 10026–10036 (1994).
36. A. R. Dinner and M. Karplus, The thermodynamics and kinetics of protein folding: A lattice model analysis of multiple pathways with intermediates. *J. Phys. Chem. B* **103**, 7976–7994 (1999).
37. A. R. Dinner, E. Verosub, and M. Karplus, Use of a quantitative structure–property relationship to design larger model proteins that fold rapidly. *Prot. Eng.* **12**, 909–917 (1999).
38. E. I. Shakhnovich and A. M. Gutin, Implications of thermodynamics of protein folding for evolution of primary sequences. *Nature* **346**, 773–775 (1990).
39. E. I. Shakhnovich and A. M. Gutin, A new approach to the design of stable proteins. *Prot. Eng.* **8**, 793–800 (1993).
40. V. Muñoz, P. A. Thompson, J. Hofrichter, and W. A. Eaton, Folding dynamics and mechanism of β-hairpin formation. *Nature* **390**, 196–199 (1997).
41. V. Muñoz, E. R. Henry, J. Hofrichter, and W. A. Eaton, A statistical mechanical model for β-hairpin kinetics. *Proc. Natl. Acad. Sci. USA* **95**, 5872–5879 (1998).
42. O. V. Galzitskaya and A. V. Finkelstein, A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl. Acad. Sci. USA* **96**, 11299–11304 (1999).
43. R. Zwanzig, A. Szabo, and B. Bagchi, Levinthal's paradox. *Proc. Natl. Acad. Sci. USA* **89**, 20–22 (1992).
44. S. Takada, Gō-ing for the prediction of protein folding mechanisms. *Proc. Natl. Acad. Sci. USA* **96**, 11698–11700 (1999).
45. N. Gō, Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* **12**, 183–210 (1983).
46. J. A. Schellman, The factors affecting the stability of hydrogen-bonded polypeptide structures in solution. *J. Phys. Chem.* **62**, 1485–1494 (1958).
47. D. A. Debe and W. A. Goddard III, First principles prediction of protein folding rates. *J. Mol. Biol.* **294**, 619–625 (1999).
48. M. Karplus and D. L. Weaver, Protein-folding dynamics. *Nature* **260**, 404–406 (1976).
49. M. Karplus and D. L. Weaver, Protein folding dynamics: The diffusion-collision model and experimental data. *Prot. Sci.* **3**, 650–668 (1994).
50. A. R. Fersht, Transition-state structure as a unifying basis in protein-folding mechanisms: Contact order, chain topology, stability, and the extended nucleus mechanism. *Proc. Natl. Acad. Sci. USA* **97**, 1525–1529 (2000).
51. K. M. Fiebig and K. A. Dill, Protein core assembly processes. *J. Chem. Phys.* **98**, 3475–3487 (1993).
52. K. A. Dill, K. M. Fiebig, and H. S. Chan, Cooperativity in protein-folding kinetics. *Proc. Natl. Acad. Sci. USA* **90**, 1942–1946 (1993).
53. N. D. Socci, J. N. Onuchic, and P. G. Wolynes, Diffusive dynamics of the reaction coordinate for protein folding funnels. *J. Chem. Phys.* **104**, 5860–5868 (1996).
54. M. Karplus, Aspects of protein reaction dynamics: Deviations from simple behavior. *J. Phys. Chem. B* **104**, 11–27 (2000).
55. R. Zwanzig, Two-state models of protein folding kinetics. *Proc. Natl. Acad. Sci. USA* **94**, 148–150 (1997).
56. M. Vendruscolo, E. Paci, C. M. Dobson, and M. Karplus, Three key residuees form a critical contact network in a transition state for protein folding. *Nature* **409**, 641–645 (2001).
57. D. S. Riddle, V. P. Grantcharova, J. V. Santiago, E. Alm, I. Ruczinski, and D. Baker, Experiment and theory highlight role of native state topology in SH3 folding. *Nature Struct. Biol.* **6**, 1016–1024 (1999).

58. J. C. Martínez and L. Serrano, The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nature Struct. Biol.* **6**, 1010–1016 (1999).
59. F. Chiti, N. Taddei, P. M. White, M. Bucciantini, F. Magherini, M. Stefani, and C. M. Dobson, Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nature Struct. Biol.* **6**, 1005–1009 (1999).
60. V. Villegas, J. C. Martínez, F. X. Avilés, and L. Serrano, Structure of the transition state in the folding process of human procarboxypeptidase A2 activation domain. *J. Mol. Biol.* **283**, 1027–1036 (1998).
61. J. Clarke, E. Cota, S. B. Fowler, and S. J. Hamill, Folding studies of immunolobulin-like β-sandwich proteins suggest that they share a common folding pathway. *Structure Fold. Des.* **7**, 1145–1153 (1999).
62. A. R. Fersht, L. S. Itzhaki, N. F. El Masry, J. M. Matthews, and D. E. Otzen, Single versus parallel pathways of protein folding and fractional formation of structure in the transition state. *Proc. Natl. Acad. Sci. USA* **91**, 10426–10429 (1994).
63. G. Aronsson, A.-C. Brorsson, L. Sahlman, and B.-H. Jonsson, Remarkably slow folding of a small protein. *FEBS Lett.* **411**, 359–364 (1997).
64. R. E. Burton, B. S. Huang, M. A. Daugherty, T. L. Calderone, and T. G. Oas, The energy landscape of a fast-folding protein mapped by ala→gly substitutions. *Nature Struct. Biol.* **4**, 305–310 (1997).
65. S. Sato, B. Kuhlman, W.-J. Wu, and D. P. Raleigh, Folding of the multidomain ribosomal protein L9: The two domains fold independently with remarkably different rates. *Biochemistry* **38**, 5643–5650 (1999).
66. N. Ferguson, A. P. Capaldi, R. James, C. Kleanthous, and S. E. Radford, Rapid folding with and without populated intermediates in the homologous four-helix proteins Im7 and Im9. *J. Mol. Biol.* **286**, 1597–1608 (1999).
67. S. Spector, B. Kuhlman, R. Fairman, E. Wong, J. A. Boice, and D. P. Raleigh, Cooperative folding of a protein mini domain: The peripheral subunit-binding domain of the pyruvate dehydrogenase multienzyme complex. *J. Mol. Biol.* **276**, 479–489 (1998).
68. S. Spector and D. P. Raleigh, Submillisecond folding of the peripheral subunit-binding domain. *J. Mol. Biol.* **293**, 763–768 (1999).
69. W. Kabsch and C. Sander, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
70. J. M. Chandonia and M. Karplus, New methods for accurate prediction of protein secondary structure. *Proteins* **35**, 293–306 (1999).
71. L. S. Itzhaki, D. E. Otzen, and A. R. Fersht, The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: Evidence for a nucleation–condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260–288 (1995).
72. Z. Guo, C. L. Brooks III, and E. M. Boczko, Exploring the folding free energy surface of a three-helix bundle protein. *Proc. Natl. Acad Sci. USA* **94**, 10161–10166 (1997).
73. F. B. Sheinerman and C. L. Brooks III, Molecular picture of folding of a small α/β protein. *Proc. Natl. Acad. Sci. USA* **95**, 1562–1567 (1998).
74. B. D. Bursulaya and C. L. Brooks III, Folding free energy surface of a three-stranded β-sheet protein. *J. Am. Chem. Soc.* **121**, 9947–9951 (1999).
75. A. R. Dinner, T. Lazaridis, and M. Karplus, Understanding β-hairpin formation. *Proc. Natl. Acad. Sci. USA* **96**, 9068–9073 (1999).
76. P. Ferrara and A. Caflisch, Folding simulations of a three-stranded antiparallel β-sheet. *Proc. Natl. Acad. Sci. USA* **97**, 10780–10785 (2000).
77. A. Hiltpold, P. Ferrara, J. Gsponer, and A. Caflisch, Free energy surface of the helical peptide $Y(MEARA)_6$, *J. Phys. Chem. B* **104**, 10080–10086 (2000).