Composition-based effective chain length for prediction of protein folding rates

Le Chang, Jun Wang,*,† and Wei Wang*,‡
National Laboratory of Solid State Microstructure and Department of Physics, Nanjing University, Nanjing 210093, China

(Received 13 September 2010; published 24 November 2010)

Folding rate prediction is a useful way to find the key factors affecting folding kinetics of proteins. Structural information is more or less required in the present prediction methods, which limits the application of these methods to various proteins. In this work, an "effective length" is defined solely based on the composition of a protein, namely, the number of specific types of amino acids in a protein. A physical theory based on a minimalist model is employed to describe the relation between the folding rates and the effective length of proteins. Based on the resultant relationship between folding rates and effective length, the optimal sets of amino acids are found through the enumeration over all possible combinations of amino acids. This optimal set achieves a high correlation (with the coefficient of 0.84) between the folding rates and the optimal effective length. The features of these amino acids are consistent with our model and landscape theory. Further comparisons between our effective length and other factors are carried out. The effective length is physically consistent with structure-based prediction methods and has the best predictability for folding rates. These results all suggest that both entropy and energetics contribute importantly to folding kinetics. The ability to accurately and efficiently predict folding rates from composition enables the analysis of the kinetics for various kinds of proteins. The underlying physics in our method may be helpful to stimulate further understanding on the effects of various amino acids in folding dynamics.

DOI: 10.1103/PhysRevE.82.051930 PACS number(s): 87.15.A−, 87.15.Cc, 87.15.Qt

I. INTRODUCTION

Natural proteins generally can fold to their native conformations rapidly and reliably under physiological environment [1]. The folding rate describes the efficiency of such kind of dynamic process on a complex energy landscape. Besides environmental conditions [2–6], various properties of a protein chain would affect the features of its energy landscape and consequently modulate its folding rate [7–10]. Therefore, finding the factors related to folding rate under the standard environmental condition would greatly promote the understanding on folding processes [11–15].

Physically, folding rate of a protein is determined by the competition between conformational entropy and energetic driving force [15–18]. The former controls the size of the conformational space to search and the latter outlines the protocols to search such a space. Considering the minimal frustration principle [11,18,19], the estimation of conformational entropy may be more important in determining folding rates of the proteins with optimal interactions. With this notion, a series of sequence-irrelevant factors were derived to predict the folding rate. For example, the chain length of a protein was expected to be the determinant of its folding rate in many physical models [20–24]. Consistent with these models, apparent correlations between the folding rates and the chain lengths were actually observed based on the bioinformatic statistics on proteins [25,26]. More recently, a series of structure-based factors related to prediction of folding rates were discovered after the pioneer works of Plaxco *et al.* [27,28]. These factors are believed to be correlated with conformational entropy more precisely than those based on the polymeric estimation; thus, it could produce better predictions for the folding rates [26,29–40]. These observations further support the notion that folding rate is mainly determined by conformational entropy, which is based on the consideration of the minimal frustration principle.

Accompanied with the above studies, there is another thread focusing on the sequence-specific factors in rate prediction (see the review [41] and references therein). These studies are motivated by the apparent variations of folding rates after mutations [42–45], which could be explained with neither length- nor topology-based factors. This indicates that the sequence-dependent energetic contributions to the folding rates should not be neglected. Interestingly, it is found in recent works that the factors based on the compositions of various amino acids could predict folding rates with high correlation coefficients for specific sets of proteins [46–51]. These observations suggest that the sequential information of proteins should not be neglected in rate predictions [41,52,53]. Therefore, the inclusion of sequential information could be a promising direction to enhance the predictability for folding rates.

To find the sequence-based factors related to folding rate, various properties of amino acids should be considered. The questions are the following: among a large number of amino-acid properties [54–56], which ones would be useful for rate predictions? What kind of weights should be assigned to these properties? How can we integrate the sequential information with the chain length? Presently, there is no general paradigm to answer these questions except some intuitions and empirical rules. Therefore, at the present stage, it is expected not only to find new factors for folding rate predictions but also to provide some insights related to the above questions. Such insights could be helpful to build up factors for folding rate predictions with higher accuracy.

*Corresponding author. FAX: 86-25-83595535.

†wangj@nju.edu.cn

‡wangwei@nju.edu.cn

In this work, a concept of "effective length" is created as the number of amino acids with specific types. The relation between the folding rate and this kind of factor (effective length) is derived based on a minimalist model. To determine what kinds of amino acids should be included in the definition of effective length, an optimization over all combinations of amino acids is carried out based on the kinetic information of 95 proteins. It is found that the optimal definition of effective length could be realized by introducing only the hydrophobic and the flexible amino acids. This is consistent with our modeling analysis and landscape theory. With this optimal definition of effective length, a nice correlation between the folding rates and the effective lengths is achieved. Such a correlation indicates that the effective length could be a viable predictor for the folding rates. The statistical significance of this correlation is systematically checked with the bootstrapping method for the protein set. The predictability of our effective length is also compared with those by other factors, including absolute contact order (ACO), Finkelstein's effective length, and other predictors based on sequential information. Our effective length not only is physically consistent with other methods but also has the best predictability. These results suggest that our effective length catches the basic essence of folding kinetics and may help to build comprehensive pictures for the factors affecting the folding rates.

II. MODEL AND THEORY

A. Effective length of proteins

Among various factors which correlate with the folding rates, the idea of effective length proposed by Finkelstein and his co-workers was derived from the assumption that the formation of the local and global structures may experience totally different time scales. Consequently, the folding kinetics (at least the rate) could be described with a simplified system with a set of effective units. The number of the effective units is defined as the effective length. Practically, the effective units are defined as the helices or other kinds of residues in Ref. [36]. Clearly, this kind of definition largely depends on the features of the amino acids in the proteins. It is natural to extend this kind of definition by considering the contributions of various kinds of amino acids. This kind of combination with chain length and the features of amino acids is a valuable direction to develop better factors to describe folding kinetics.

In this work, we propose a definition of effective length for proteins based on their amino-acid compositions. In our definition, the effective units are essential amino acids in the original sequence. The selected amino acids are believed to take important roles in folding processes, while other amino acids are less significant in either the thermodynamics or the kinetics. This is consistent with the observation that various kinds of amino acids contribute differently in folding kinetics. The selection of these effective amino acids is based on the types of amino acids. That is, only certain kinds of amino acids are kept in calculating the effective length of the proteins. Thus, the effective length L_{eff} is defined as $L_{\text{eff}} = \Sigma_{s \in \mathcal{S}} n_s$, where n_s represents the number of s-type amino acids and the set $\mathcal{S}$ is the collection of the amino-acid types used as the effective units. It is assumed that each effective unit contributes equally to the whole kinetics of the protein system. Operationally, the format of the set $\mathcal{S}$ is given in Sec. II C.

This kind of definition is based on the assumptions that the folding kinetics can be described with the combination of the features of various amino acids and that the features of amino acids are largely determined by its type. These assumptions have been widely applied in protein modeling [57,58], sequence alignments [59], structural features [60–62], as well as many practices of rate predictions [39,41]. The previous implementation of effective length also implicitly includes this kind of idea. The successes of those studies suggest that these assumptions are reasonable approximations for real protein systems.

B. Relation between folding rates and effective length

To determine the set $\mathcal{S}$ of amino-acid types for the effective length, there are several kinds of methods. The intuitive way is to pick up the amino acids based on prior knowledge about the relation between the features of amino acids and the folding behaviors. The amino acids with assumed important properties may be chosen to build up the set $\mathcal{S}$. Due to the complexity of folding processes, this kind of assignment for the set $\mathcal{S}$ may produce biases originated from the subjective selections for the properties of amino acids. Another more physical method is to derive the relationship between folding kinetics and various kinds of factors directly from physical models. This kind of method may be instructive to build up a thorough picture for the problems. In this work, the latter method is employed. The free-energy landscapes of proteins are build up based on a simplified model. This model gives out the qualitative relationships between folding rates and composition-based effective length.

Naturally, proteins generally fold on funnel-shaped energy landscapes. On such kinds of landscapes, the two-state feature (i.e., the transition between the compact native state and loose denatured state cooperatively) is typical for the folding of single-domain proteins. In our model, the two-state feature is adopted for amino acids. Each amino acid would be either in the folded state (N) or in the unfolded state (U). These two states have distinct energetic and entropic features. This kind of assumption has been widely used in previous theoretical models [12,13,23,38,63]. To introduce compositional information, it is necessary to have the energetic and entropic terms specific for each type of amino acid. In detail, the ith amino acid of type τ_i would have the energy γ_{τ_i} and the entropy S_{τ_i} in the unfolded state. Here, $\gamma_{\tau_i} \geq 0$ describes the energy penalty when this amino acid is fully exposed to solvent. In the folded state, the amino acid would generally have structure-dependent energy $\gamma_{\tau_i} a_i$ and zero entropy. The parameter a_i is a factor measuring the degree of exposure of the amino acid, and the null value of the entropy reflects the rigid feature of native state.

With the above setups, the free energy of a protein chain could be calculated by summing up the energetic and entropic terms of all amino acids. For the unfolded state and the folded native state, the free energies of an entire protein chain take the forms as

$$F_U = \sum_{i=1}^{n} \gamma_{\tau_i} - TS_{\tau_i} = \sum_{\tau} n_\tau(\gamma_\tau - TS_\tau), \tag{1}$$

$$F_N = \sum_{i=1}^{n} \gamma_{\tau_i} a_i = \sum_{\tau} n_\tau \gamma_\tau \bar{a}_\tau = \sum_{\tau} \gamma_\tau A_\tau, \tag{2}$$

where T is the temperature, n is the length of the protein chain, n_τ is the number of τ-type amino acids in the concerned sequence, and $\bar{a}_\tau$ and A_τ are the averaged exposure and the total exposure of this kind of amino acid. As estimation for the exposure area of amino acids, a uniform distribution of various amino acids on the protein surface is assumed, that is, $A_\tau = (n_\tau/n)A_{\text{protein}}$ in which A_{protein} is the surface area of the whole protein. Consequently, the free energy in folded state could be further simplified as

$$F_N = \Gamma A_{\text{protein}}, \tag{3}$$

in which $\Gamma = \Sigma_\tau Q_\tau \gamma_\tau$, with $Q_\tau = n_\tau/n$ describes the strength of the interaction between protein and solvent. Considering the native states of proteins generally take the shape of compact globules, the exposed surface area after folding could be estimated from the length of protein chains with the power-law form of $A_{\text{protein}} \sim n^\alpha$, which is widely used in previous models [13,22,23]. The exponent α describes the shape of proteins. It is argued that typical value of α for native proteins ranges from 1/2 to 2/3 [64]. A smaller α suggests less exposed surface in the native state and vice versa. In this work, α is regarded as a constant for various proteins. Then, the free energy of the folded state could be replaced by

$$F_N = \mu \Gamma n^\alpha, \tag{4}$$

where μ is a constant parameter related to the surface area. Considering the two-state feature of amino acids, the free energy of a partially folded state could be expressed as

$$F = \mu \sum_{\tau} \gamma_\tau q_\tau m^\alpha + (n - m)\sum_{\tau} q_\tau(\gamma_\tau - TS_\tau), \tag{5}$$

where m is the number of the folded amino acids in this partially folded state and $q_\tau = m_\tau/m$ measures the ratio of τ-type folded amino acids correspondingly.

In our model, the amino acids with both small energy γ_τ and small entropy TS_τ would have little bias toward either unfolded or folded state. These amino acids would generally be the polar ones with strong preferences to certain local conformations. The invariance of free energies for these amino acids in folded and unfolded states suggests that these amino acids have no contributions to drive the folding. It is possible to assume that these amino acids have been preformed in unfolded state and are not necessary to be included when describing the folding kinetics. Therefore, an effective chain length, n_e, could be defined as the number of all other amino acids except those preformed ones. Based on this kind of effective length, the free energy of different states in folding process (including the unfolded, the folded, and the partially folded states) could be rewritten as

$$F_U = \sum_{\tau'} n_{\tau'}(\gamma_{\tau'} - TS_{\tau'}), \tag{6}$$

$$F_N = \mu' \Gamma' n_e^\alpha, \tag{7}$$

and

$$F = \mu' \sum_{\tau'} \gamma_{\tau'} q_{\tau'} m^\alpha + (n_e - m)\sum_{\tau'} q_{\tau'}(\gamma_{\tau'} - TS_{\tau'}). \tag{8}$$

Here, the primes indicate that these quantities are calculated based on the effective lengths. For simplicity, the primes are omitted in the following discussions.

Based on the free energy in Eq. (8), the transition state related to folding should satisfy the relationship

$$\frac{\partial F}{\partial m}(m = m^\ddagger, q_\tau = q_\tau^\ddagger) = 0, \tag{9}$$

where ‡ indicates that the values correspond to the transition state and τ represents the type of concerned amino acids. Since the derivative of F on q_τ is generally larger than 0, it is not considered.

To create kinetically preferable folding pathways, a maximum of route entropy should be achieved [65]. Since various kinds of amino acids may change their states (unfolded or folded) independently, the route entropy could be represented as

$$S_{\text{Route}} = \ln \prod_{\tau} C_{n_\tau}^{m_\tau}, \tag{10}$$

where the factor $C_n^m = n!/m!(n-m)!$ calculates the m combinations from the set of n elements. Maximizing the route entropy S_{Route} for a given set of $\{n_\tau\}$ would produce the relationship

$$\frac{m_\tau}{n_\tau} = \frac{m}{n_e}. \tag{11}$$

It is obvious that this relationship is rational for large n_τ. Thus, condition (11) will be more easy to satisfy when some coarse-grained groupings of amino acids are applied.

From Eqs. (9) and (11), the transition-state-related parameters $m^\ddagger$ and $q_\tau^\ddagger$ could be determined uniquely,

$$q_\tau^\ddagger = Q_\tau, \tag{12}$$

$$(m^\ddagger)^{\alpha-1} = \frac{1}{\alpha\mu\Gamma}\sum_{\tau} Q_\tau(\gamma_\tau - TS_\tau) = \frac{1}{\alpha}\frac{F_U}{F_N} n_e^{\alpha-1}. \tag{13}$$

Here, the location of the transition state $m^\ddagger$ depends on the temperature T. Therefore, the free-energy barrier related to the folding process $\Delta F = F^\ddagger - F_U$ could be derived as

$$\begin{aligned}\Delta F &= (1-\alpha)\mu \sum_{\tau} Q_\tau \gamma_\tau (m^\ddagger)^\alpha \\ &= (1-\alpha)\mu \sum_{\tau} Q_\tau \gamma_\tau n_e^\alpha \left(\frac{\alpha F_N}{F_U}\right)^{\alpha/(1-\alpha)} \sim n_e^\beta.\end{aligned} \tag{14}$$

Clearly, at the folding transition temperature, the relation $F_U = F_N$ would be satisfied. Thus, the relation $\Delta F \sim n_e^\alpha$

(namely, $\beta=\alpha$) would be established. While at the temperature with apparent native stability, the energy funnel is largely biased with the relation $\gamma_\tau \gg TS_\tau$. Therefore, the factor F_N/F_U would approach to $\sim n_e^{-\alpha}$ and the corresponding exponent $\beta \to 0$. Practically, for this case, a weaker dependence of the free-energy barrier on the n_e, such as $\ln n_e$, may be employed. This kind of effect of temperature on the folding barrier is also suggested previously [21,23]. For the standard experimental condition [66] to carry out the kinetics measurements, the temperature is generally apparently lower than the folding transition temperature. The logarithm dependence of the barriers would be employed in this work.

With the relation between the free-energy barrier and the effective length, the logarithm of folding rate, $\ln k_{\text{fold}}$, could be computed based on the transition-state theory [67] as

$$\ln k_{\text{fold}} = \ln k_0 - \Delta F/k_B T_f = C_0 - C_1 \ln n_e \sum_\tau Q_\tau \gamma_\tau, \quad (15)$$

where the parameters $C_{0,1}$ are constants, $C_0 = \ln k_0$ and $C_1 = (1-\alpha)\alpha^{\alpha/(1-\alpha)}\mu/k_B T_f$, and k_B is the Boltzmann constant. It is clear that the folding rate depends on the compositional factor Q_τ and the effective length n_e. When the involved proteins have a large variation of sizes, the description with the effective length for rate prediction may be a reasonable simplification. Meanwhile, for proteins with similar sizes (such as middle-size proteins which are studied in Ref. [47]), the factor Γ may take an important role, and the predictor for folding rate may take a form of the linear combinations of compositions as used in many sequence-based studies [46–51]. Our formula provides an outline for the cooperation between amino-acid compositions and chain length effect in the determination of folding rates. As a remark, the definition of effective length does depend on the feature of amino-acid composition of proteins. Thus, the composition information is one of the important aspects to produce a proper estimation for folding rates.

Considering that the amino acids deleted from the definition of the effective length are those with both small γ_τ and small TS_τ as discussed above, the amino acids made up of the effective length should have either large γ_τ values or large TS_τ values. The former would generally be the hydrophobic amino acids, whereas the latter are possible flexible polar ones. The two categories of amino acids would be represented by h (hydrophobic) and s (soft). Based on such a kind of assignment, the free-energy barrier ΔF could be represented through the combination of Eqs. (13) and (14),

$$\Delta F \sim [Q_h(\gamma_h - T_f S_h) + Q_s(-T_f S_s)]^{\alpha/(\alpha-1)}. \quad (16)$$

Here, we assume $\gamma_s = 0$ for polar amino acids. Clearly, both the hydrophobic and the flexible amino acids contribute to folding barrier. Considering the fact that $\alpha < 1$, the strong hydrophobicity (i.e., with a large γ_h and/or a large Q_h) may reduce the barrier and may speed up the folding, and the high flexibility (i.e., with a large S_s and/or a large Q_s) would increase the barrier and retard the folding. This is consistent with the physical intuition. This form of barrier demonstrates that the folding of proteins is caused by the balance between the energetic and entropic terms.

With this kind of simplification, the factor $\Gamma = \Sigma_\tau Q_\tau \gamma_\tau$ in Eq. (15) could be simplified as $\gamma_h n_h / n_e$. For regular proteins, the composition of hydrophobic amino acids generally varies slightly around a certain amount. When considering a large set of proteins, the chain length may vary from tens to hundreds. In such a situation, the fluctuation of $\Gamma = \gamma_h n_h / n_e$ is relatively small, so that the quantity Γ could be regarded as a constant for the cases related to the proteins with a large range of lengths. This kind of opinion about the fluctuation of $Q_h = n_h / n_e$ is checked (as shown in Sec. III B). Based on this idea, the factors for folding rate prediction could be further simplified as

$$\ln k_{\text{fold}} = C_0 + C_1' \ln n_e, \quad (17)$$

where $C_1' = C_1 \gamma_h n_h / n_e$ is treated as a constant. Finally, the folding rate could be predicted solely from the effective length. Practically, the formulism based on the power-law dependence of n_e could also be evaluated as comparisons,

$$\ln k_{\text{fold}} = C_0 + C_1' n_e^\beta. \quad (18)$$

Here, the relation between the folding rates and the effective length is derived based on a simple two-state model. For the multistate folders, this relation is probably more complex. Yet, under the standard condition, the energy landscape of proteins would be clearly biased toward the native state. In such situation, the folding would generally have apparent similarity to that of the two-state folders. It is believed that the above relation may still work for the multistate proteins. This speculation is approved by further correlation analysis.

C. Determination of the optimal sets of amino-acid types

Based on the above theory, the first step to build a rational factor for rate prediction is to select certain kinds of amino acids to create reliable estimation for effective lengths. It is possible to pick up the hydrophobic and the flexible amino acids according to prior knowledge of routine classification. However, it is obscure to select the expected amino acids properly considering a large number of various indices for hydrophobicity and flexibility [54–56]. Besides, artificial assignment for types of amino acids cannot properly balance the contributions of two kinds of amino acids and would generally deteriorate the quality of predictions. Here in our work, a different approach is employed to optimize the set of amino-acid types to achieve the best predictability. This kind of derivations is also used in other statistical analysis [48,68,69].

In detail, given a set of amino-acid types (alphabets), $T = \{\tau_1, \tau_2, \ldots\}$, the effective length could be calculated as $n_e = \Sigma_{\tau \in T} n_\tau$, where n_τ is the number of τ-type amino acids in a concerned protein. A correlation between the experimental rates $\ln k_{\text{fold}}$ and the factor n_e^β thus could be determined based on the least-squares fitting method. The degree of correlation reflects the validity of the selected alphabet T for rate determination. A good correlation generally suggests a proper set of amino acids for the definition of effective length. Quantitatively, the degree of correlation is described with the correlation coefficient R or the averaged deviation D between $\ln k_{\text{fold}}$ and the corresponding predicted values $\ln k_{\text{pred}}$ [based

TABLE I. Proteins and their kinetic information. The table gives out the PDB codes, lengths, and folding rates for 95 proteins which are used in our work. The references for these kinetic data could be acquired from authors.

PDB code	Length	$\ln k_f$	PDB code	Length	$\ln k_f$	PDB code	Length	$\ln k_f$
1A6N	151	1.1	1ADW	123	0.6	1AEY(1SHG)	62	2.1
1AON(1DK7)	153	0.8	1APS	98	−1.5	1B9C	225	−2.8
1BA5	53	5.9	1BDC	58	11.7	1BEB	156	−2.2
1BNI	110	2.6	1BTA(1BRS)	89	3.4	1C8C	64	7.0
1C9O	66	7.2	1CBI	136	−3.2	1CEI	94	5.7
1COA	83	3.9	1CSP	67	6.5	1DIV-N	56	6.6
1DIV-C	93	3.3	1E0L	37	10.6	1E0M	37	8.8
1EAL	127	1.4	1ENH	61	10.5	1FEX	59	8.2
1FKB	107	1.4	1FMK	57	4.1	1FNF(ninth domain)	90	−0.9
1FNF(tenth domain)	94	5.0	1G6P	66	6.3	1GXT	91	5.4
1HCD	118	1.1	1HDN(1POH)	85	2.7	1HEL	129	1.2
1HMK	121	2.8	1HNG	98	1.8	1HRC	104	8.0
1I1B	151	−4.0	1IDY	54	8.7	1IFC	132	3.4
1IMQ	86	7.3	1JOO	149	0.3	1K8M	87	−0.7
1K9Q	40	8.4	1L2Y	20	12.4	1L8W	338	1.6
1LMB(N terminal)	80	8.5	1LOP	164	6.6	1OPA	134	1.4
1PBA	81	6.8	1PGB	56	6.0	1PGB(C terminal)	16	12.0
1PHP(N terminal)	175	2.2	1PHP(C terminal)	219	−3.5	1PIN	40	9.3
1PKS(1PNJ)	85	−1.1	1PRB	53	12.9	1PSE	69	3.2
1QOP-α	268	−2.5	1QOP-β	396	−6.9	1QTU	115	−0.4
1RA9	159	−3.2	1RFA	78	7.0	1RIS	101	5.9
1SCE	112	4.2	1SHF(1NYF)	59	4.5	1SRL	64	4.0
1TEN(third domain)	90	1.0	1TIT	98	3.5	1UBQ	76	7.3
1URN	97	5.8	1UZC	69	8.7	1VII	36	12.4
1WIT	93	0.4	2A3D	73	12.2	2A5E	156	3.5
2ABD	86	5.6	2ACY	98	0.8	2AIT	4.2	
2BLM	260	−1.2	2CRO	71	5.3	2HQI	72	0.2
2LZM(1L63)	164	4.3	2PDD	43	9.8	2PTL	78	4.1
2RN2	155	0.3	2VIK	126	4.2	3CHY	128	1.0
3MEF(1MJC)	69	5.3	1N88	96	2.0	1T8J	23	11.8
1JO8	58	2.5	1CUN(16th domain)	110	4.8	1CUN(17th domain)	103	3.4
Sho1 SH3 domain	76	2.1	Ubiquitin related modifier	101	2.6			

on Eq. (17)], $D=1/N_{\text{protein}}\Sigma_{i\in\text{protein}}|\ln k_{\text{fold}}^{(i)}-\ln k_{\text{pred}}^{(i)}|$. A better correlation generally has a larger correlation coefficient R and a smaller variance D. Based on the relation between correlation and the alphabet T, it is possible to find out the best alphabet T by optimizing the quantities R or D (namely, maximizing R or minimizing D) over all possible combinations of amino-acid types (totally $2^{20}-1$ combinations). With the resultant optimal set of amino acids T which has a best correlation with folding rates, the effective lengths of proteins for rate prediction could be self-consistently defined.

The statistical significance of the selected optimal set of amino-acid types is also analyzed in our work. A Z-score-like quantity is defined for such kinds of analysis. The quantity Z is defined as $Z=(\bar{D}-D_{\text{opt}})/\sigma_D$, where $\bar{D}$ is the average of variance D for all concerned sets of amino-acid types and σ_D is the corresponding standard deviation respective to the average $\bar{D}$. This quantity of Z measures the specificity of the selected optimal set in the background of other sets. This quantity could also be defined for the cases focusing on the sets with n types of amino acids as $Z_n=(\bar{D}_n-D_n)/\sigma_{D_n}$, in which D_n, $\bar{D}_n$, and σ_{D_n} are calculated according to the sets with n types of amino acids.

D. Set of proteins for correlation analysis

In this work, 95 proteins are employed in our analysis. These proteins are collected from literatures including KineticsDB [70]. These proteins are listed in Table I. Some proteins with uncertainty in their rate measurements are omitted.

The dependence of our correlation analysis on the size of protein set is also checked with boot-strapping method. A series of subsets of proteins from 95 proteins are generated, and the same procedures are carried out to determine the optimal set of amino acids for these subsets. Generally, for a

TABLE II. Optimal sets of amino acids for various exponents. Optimal sets of amino acids with various β values. R is the correlation coefficient and Z is the Z score as described in main text.

β	R	Z	Optimal set of amino acids
0.0 (ln n_e)	0.838	3.27	CDGLPSTVWY
0.1	0.834	3.29	CDHMPSTVWY
0.2	0.830	3.22	CDHMPSTVWY
0.3	0.825	3.25	CDHMPSTVWY
0.4	0.819	3.49	CDHPSTVWY
0.5	0.812	3.60	CDHPSTVWY
0.6	0.805	3.42	CHPTVWY

certain size of the subsets N_s, about 1000 instances of the subsets are generated randomly. The probability P_{AA} of each amino acid to be in the optimal set for a certain size N_s is calculated through the statistics on the optimal sets for the 1000 instances of the subset. In a statistical sense, the optimal set of amino acids for the subset with N_s proteins could be determined as the set of amino acids with their P_{AA} larger than a threshold (0.5 in our analysis). Comparisons between these optimal sets could give us some useful information about the effect of size of protein set.

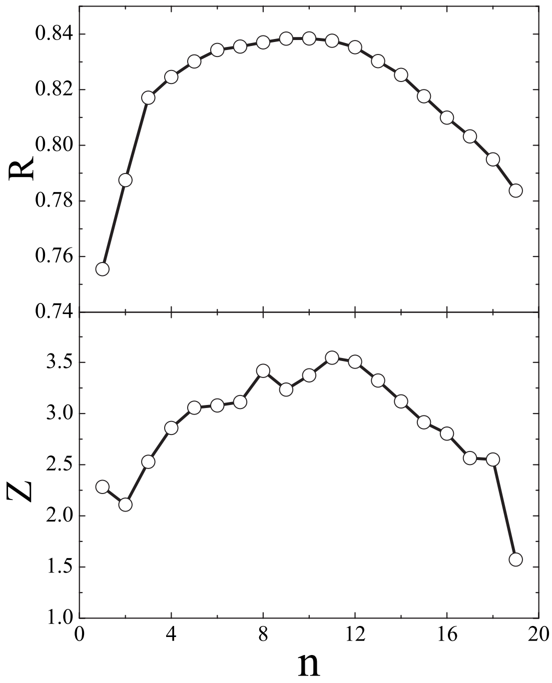

FIG. 1. Correlation coefficient and Z score. The correlation coefficient R_n and the corresponding Z score Z_n for the optimal set with n amino acids.

III. RESULTS AND DISCUSSIONS

A. Optimal set of amino-acid types

Based on Eq. (17), the optimal set of amino-acid types could be determined with the optimization procedure described in Sec. II C. As a comparison, the corresponding results with various exponent β from 0.6 to 0.1 based on Eq. (18) are also obtained. All these globally optimal sets are listed in Table II. The corresponding correlation coefficients and Z scores are also listed. All the cases have acceptable correlations. This illustrates that our formulas (17) and (18) catch the fundamentals of physics for folding processes. The optimal types of amino acids for different formula vary slightly, with the difference of only one or two kinds of amino acids generally. Such a small variation indicates that the optimal types of amino acids are not sensitive to the exponent β but are a kind of intrinsic feature of amino acids. Besides, it is also found that smaller values of β have larger correlation coefficients R. Especially, the logarithm format produces the best correlation. This kind of weak dependence on the lengths of proteins is consistent with previous observations [36] and may be related to the temperature in experimental measurement as discussed in Sec. II. As a result, the logarithm form $\ln k_{\text{fold}}=C_0+C_1' \ln n_e$ is generally used in the following discussion due to its best correlation with experimental data.

TABLE III. Optimal sets of various sizes. Optimal sets of amino acids for the relation $\ln k \sim \ln n_e$. n is the number of amino acids in optimal sets, R is the correlation coefficient, and Z is the Z score as described in main text.

n	R_n	Z_n	Optimal set of amino acids
01	0.755	2.28	V
02	0.788	2.11	VY
03	0.817	2.53	DSV
04	0.825	2.86	DSTV
05	0.830	3.06	DSTVW
06	0.834	3.08	DPSTVW
07	0.835	3.11	DGPSTVW
08	0.837	3.42	DGLSTVWY
09	0.838	3.23	CDGPSTVWY
10	0.838	3.37	CDGLPSTVWY
11	0.838	3.55	CDGHLPSTVWY
12	0.835	3.51	CDGHLMPSTVWY
13	0.830	3.32	CDFGHLMPSTVWY
14	0.825	3.12	CDFGHILMPSTVWY
15	0.818	2.91	CDFGHILMPQSTVWY
16	0.810	2.80	CDEFGHILMPRSTVWY
17	0.803	2.56	CDEFGHILMPQRSTVWY
18	0.795	2.55	CDEFGHILMNPQRSTVWY
19	0.784	1.57	CDEFGHIKLMNPQRSTVWY

Based on the logarithm format, the optimal sets with n types of amino acids ($1 \leq n \leq 20$) are obtained (as shown in Table III and Fig. 1). It is observed that the optimal set of amino-acid types is enlarged roughly in an accumulative manner following the increase of n (namely, adding new types of amino acids one by one). The order of the emergence of each type of amino acid following the increase of n outlines the influence of the corresponding type on folding rate. For example, hydrophobic amino acids V and Y appear in the first two steps. This demonstrates the importance of the hydrophobic interaction in folding kinetics. The amino-acid proline P also appears early when $n=6$. This is consistent with experimental observations that the amino-acid proline has important effect for protein kinetics [71,72]. Clearly, there are some competitions between the amino acids with different physical features, such as the hydrophobic amino acids and the flexible amino acids that appear in an interlacing order. This reflects that the folding is a kind of process affected by multiple kinds of physical properties rather than a single characteristic. At the same time, the correlation coefficient R_n has a nonmonotonic variation. A global maximum of 0.84 is reached at $n=10$, and the correlation coefficients from 5 to 13 are all larger than 0.83. Meanwhile, the variation of Z score Z_n is similar to that of R_n, peaking at $n=11$ and larger than 3.0 for n from 5 to 14. In this sense, the globally optimal set at $n=10$ not only has the largest predictability but also possesses a sufficient statistical significance. Therefore, the types of amino acids corresponding to $n=10$, *LVWYCGSTDP*, would be the most suitable to define the effective length. It is interesting that such optimal set is mainly composed of two kinds of amino acids: the hydrophobic ones *LVWYC* [73] and the flexible polar ones *GSTD* [54–56,74–76]. This matches the declaration in our theory automatically. Besides, the hydrophobic and the flexible polar amino acids appear alternatively as n increases. This indicates the balance of the hydrophobicity and the flexibility during folding as suggested in Eq. (16). It is worth noting that the amino-acid proline P (which is neither hydrophobic nor flexible) appears in the optimal set. Physically, the amino-acid proline may experience a slow isomerization between two predominant states of the pyrrolidine ring, so that the search for its native state would require a rather long time. This kind of behavior is similar to that of amino acids with large local entropy. The existence of amino-acid proline in optimal set also supports the physical view on the amino acids which are important for folding kinetics.

The features of the amino acids in the optimal set are also consistent with the landscape theory of proteins [77]. Physically, the depth and the ruggedness of energy funnel are tightly related to the number of the hydrophobic amino acids, and the entropy of the whole conformational space could be estimated with the effective length. Therefore, the landscape feature of proteins could be estimated with the above compositional information. Our definition of the effective length based on the optimal set of amino acids really catches the essence of physics in protein systems.

As a comparison, the spectrum of optimal sets corresponding to the exponent $\beta=0.6$ is also derived (as listed in Table IV). Similar properties as that for logarithm case are observed, including the order and the classification of amino acids. Especially, the importance of the hydrophobic and the flexible in optimal set is also observed. These observations demonstrate that the optimal sets of amino acids reflect basic relationships between the composition of amino acids and folding kinetics.

B. Validity of the assumption for Q_h

TABLE IV. Optimal sets of various sizes for $\beta=0.6$. Optimal sets of amino acids for the relation $\ln k \sim n_e^{0.6}$. n is the number of amino acids in optimal sets, R is the correlation coefficient, and Z is the Z score as described in main text.

n	R_n	Z_n	Optimal set of amino acids
01	0.735	2.26	V
02	0.777	2.44	VY
03	0.791	2.69	VWY
04	0.794	2.72	IVWY
05	0.799	2.80	CIVWY
06	0.802	2.99	HPTVWY
07	0.805	3.12	CHPTVWY
08	0.804	3.33	CHIPTVWY
09	0.804	2.93	CDHPSTVWY
10	0.803	2.93	CDHMPSTVWY
11	0.799	2.86	CDHIMPSTVWY
12	0.795	3.23	CDEHIMPSTVWY
13	0.790	2.98	CDEHIMPQSTVWY
14	0.785	2.85	CDEFHILMPSTVWY
15	0.779	2.63	CDEFHILMPQSTVWY
16	0.774	2.78	CDEFGHILMPQSTVWY
17	0.769	2.10	CDEFGHILMPQRSTVWY
18	0.763	2.11	CDEFGHILMNPQRSTVWY
19	0.753	2.53	CDEFGHIKLMNPQRSTVWY

The composition of hydrophobic amino acids in an effective chain, $Q_h=n_h/n_e$, is a factor related to the folding rate determination [as shown in Eq. (15) in Sec. II]. Practically, the factor Q_h is expected to vary slightly around a certain value in our analysis. This assumption could be checked self-consistently based on various optimal sets of amino acids. First, for the global optimal set, the factor Q_h for various proteins has a small standard deviation of ~0.1. This kind of small variations of Q_h is also observed for the other cases with different α, as shown in Fig. 2(a). Compared with large variation of the factor $\ln n_e$ in Eq. (15), it is reasonable to regard the factor Q_h as a constant. Besides, for the optimal

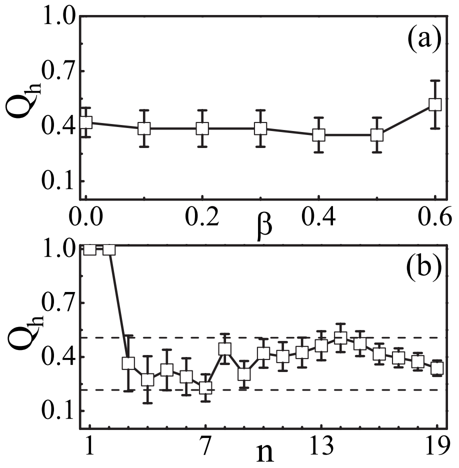

FIG. 2. *Dependence* of Q_h on the size of optimal sets. The average and the deviation of Q_h of various proteins for (a) the globally optimal sets with different exponents α and (b) the optimal sets with n amino acids.

TABLE V. The probabilities P_{AA} of amino acids in optimal sets for various subsets. This table gives the probabilities of amino acids in the optimal set corresponding to the subsets with various numbers of proteins. They are calculated through statistics over 1000 instances of subsets with a certain number of proteins.

N_s	5	15	25	35	45	55	65	75	85
A	0.175	0.084	0.048	0.026	0.011	0.002	0.002	0.000	0.000
C	0.378	0.416	0.416	0.450	0.477	0.489	0.545	0.613	0.736
D	0.295	0.326	0.383	0.474	0.544	0.623	0.698	0.797	0.918
E	0.200	0.095	0.100	0.071	0.052	0.034	0.029	0.012	0.003
F	0.290	0.176	0.142	0.111	0.102	0.082	0.067	0.052	0.023
G	0.261	0.214	0.220	0.249	0.276	0.312	0.380	0.472	0.643
H	0.350	0.252	0.285	0.310	0.331	0.358	0.386	0.388	0.384
I	0.229	0.140	0.159	0.170	0.180	0.178	0.145	0.144	0.089
K	0.198	0.086	0.061	0.051	0.041	0.022	0.013	0.002	0.000
L	0.191	0.167	0.192	0.228	0.231	0.265	0.290	0.337	0.395
M	0.289	0.208	0.182	0.175	0.176	0.214	0.227	0.218	0.195
N	0.263	0.088	0.057	0.038	0.025	0.014	0.005	0.001	0.000
P	0.357	0.408	0.482	0.545	0.594	0.630	0.667	0.721	0.813
Q	0.276	0.148	0.144	0.129	0.124	0.112	0.102	0.082	0.046
R	0.267	0.114	0.083	0.051	0.029	0.011	0.002	0.001	0.000
S	0.403	0.437	0.517	0.590	0.657	0.692	0.754	0.837	0.939
T	0.402	0.453	0.553	0.650	0.740	0.817	0.881	0.938	0.990
V	0.326	0.565	0.775	0.911	0.973	0.996	0.998	1.000	1.000
W	0.513	0.669	0.798	0.852	0.905	0.936	0.962	0.980	1.000
Y	0.464	0.470	0.546	0.641	0.656	0.716	0.772	0.818	0.892

sets with n amino acids, the average Q_h fluctuates in a limited range (from 0.4 to 0.6) when the optimal set is large enough ($n \geq 4$), as shown in Fig. 2(b). The standard deviation for each case is also as small as 0.1. Especially, for the optimal sets with large Z scores ($Z \geq 3.4$), namely, the cases with $n = 8-12$, the factor Q_h fluctuates much small. All these observations illustrate that the factor Q_h would vary in a rather limited range for our effective length. These results support the assumption for the factor Q_h. As a remark, in the above analysis, the hydrophobic amino acids are assigned according to the grouping scheme in Ref. [73] which was derived based on statistical potential [78].

C. Effect of size of protein sets

Based on the method in Sec. II D, the probabilities P_{AA} of amino acids for a series of sizes of subsets are calculated (as shown in Table V). It is easy to find that the distribution of the probability P_{AA} changes from a uniform distribution to a polarized distribution in which some amino acids have strong tendency to be in the optimal set. For these polarized distributions, the statistical optimal set is thus not sensitive to the threshold. This indicates that a large set of proteins would be helpful to determine the optimal set. Besides, the optimal sets are compared with the optimal set T determined with 95 proteins. The ratio R_T of the correctly identified amino acids (namely, the ratio of common amino acids to the size of the set T) is given in Fig. 3(a). This ratio is averaged over 1000 instances of subset. The ratio R_T grows monotonically as the size of subset increases. The growth of the ratio R_T slows down about after $n \geq 60$. When n is large enough, such as $n \geq 60$, large ratio $R_T \geq 0.8$ could be observed. The Z score for the optimal sets corresponding to the subsets is also calculated. It is given in Fig. 3(b). The average Z scores are generally larger than 3.0. The statistical errors also become smaller for larger N_s. These results indicate that the optimal

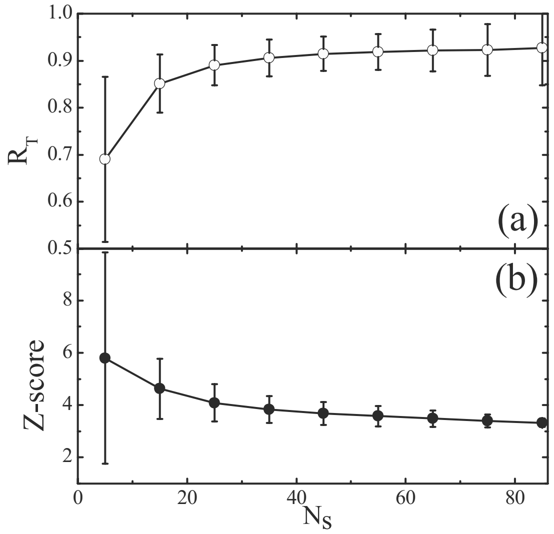

FIG. 3. Dependence of optimal sets on the size of protein set. (a) The ratio P_{AA} and (b) Z score for the optimal sets of amino acids corresponding to the subsets with n proteins. The standard deviations for various generations of subsets are also given.

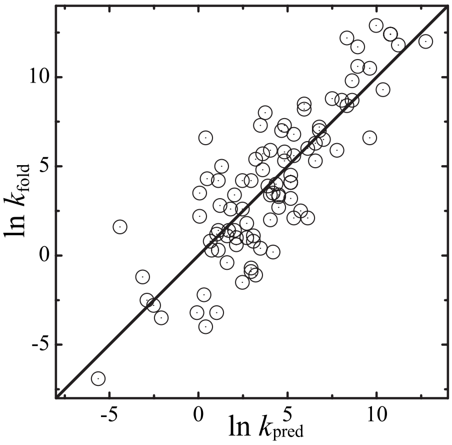

FIG. 4. Correlation of folding rates. The correlation between logarithm of experimental folding rates (ln k_{fold}) and the predicted rates with ln $k_{\mathrm{pred}}=C_0+C_1$ ln n_e. Here, effective length n_e is defined with the globally optimal set, and $C_0=28.09\pm1.64$ and $C_1=-6.39\pm0.43$. The correlation coefficient is 0.838. The diagonal line is plotted to guide the eyes.

sets become almost invariant when the set of proteins is large enough, and our choice is sufficient to extract reliable results.

D. Prediction for folding rates

Based on the globally optimal set of amino acids, the effective-length-related factor ln n_e acts as a nice predictor for the logarithm of folding rates with a high correlation ($R=0.84$), as shown in Fig. 4. The regression function obtained through least-squares fitting is ln $k_{\mathrm{fold}}=C_0+C_1$ ln n_e, where $C_0=28.09\pm1.64$ and $C_1=-6.39\pm0.43$. Compared with previous structure-based protocols (such as those in Refs. [36,39]), this function works for more proteins and has a slightly higher correlation coefficient. Clearly, this high correlation between folding rates and the effective length is achieved for both two-state and multistate folders. The independence of various kinds of proteins is also observed in previous studies [36]. Since our method merely needs sequence information, it would be easy to operate and be suitable for various cases without structural information. Such a feature enables our method to be a practical way for rate predictions.

This kind of prediction is clearly correlated with the length dependence of the folding rates. The correlation between the logarithm of the effective length ln L_{eff} and of the full length ln L is given in Fig. 5(a), with a correlation coefficient of 0.94. There is a large fluctuation in this correlation for shorter chains which mainly correspond to the two-state folders. This reflects the intrinsic dependence of the folding rates on the lengths of proteins. It is worth noting that there are clear improvements in prediction with our method compared with the naive length. The correlation coefficient is enhanced from 0.75 to 0.84. In more detail, the deviations of the predicted rates, $\Delta k=|\ln k_{\mathrm{pred}}-\ln k_{\mathrm{exp}}|$, based on the factors ln L_{eff} and ln L are calculated. The differences of the deviations, $\Delta\Delta k=\Delta k(L)-\Delta k(L_{\mathrm{eff}})$, for these two kinds of factors could be used to evaluate which kind of factor is better in prediction. For our data, the histograms of $\Delta\Delta k$ for the two-state folders and multistate folders are shown in Figs. 5(b) and 5(c). It is found that the factor ln L_{eff} works better for two-state folders with a majority of $\Delta\Delta k$ being positive [as shown in Fig. 5(b)], while two kinds of factors behave similarly for multistate folders. Combined with the fact that the optimal set of essential amino acids is physically meaningful and is insensitive to the choices for the func-

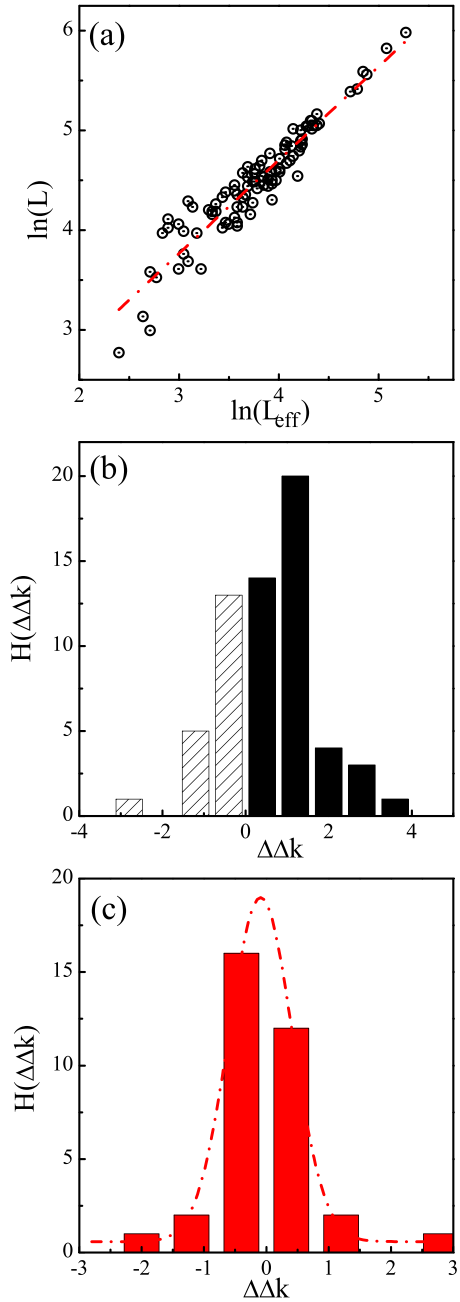

FIG. 5. (Color online) Comparison with the factor of full length of proteins. (a) The correlation between the logarithm of the effective length ln L_{eff} and that of the full length ln L. (b) and (c) give the histogram of the difference of the deviations of predicted rates $\Delta\Delta k$ with the factor L_{eff} and L for (b) two-state folders and (c) multistate folders. The solid and shadowed bars in (b) indicate the part with positive and negative $\Delta\Delta k$, respectively. The dashed-dotted line int (c) is used to guide the eyes.

tional format and the set of proteins, our method is a kind of improvement compared to the previous prediction solely with length information.

E. Comparisons with other prediction protocols

Presently, the predictions of folding rates with structural information are widely accepted in protein studies. It would be somehow surprising to have a such a nice predictor solely based on the amino-acid compositions of proteins. Considering the fact that the structural features of proteins are generally determined by the composition and sequential arrangement of amino acids [1], the concurrence of these two kinds of predictors would be physically relevant. Here, we carry out a phenomenological demonstration on the consistence between effective length and a contact-order-based factor. The absolute contact order (ACO) is a good predictor based on the structural information [25], which is generally described as

$$\mathrm{ACO} = \frac{1}{N_c}\sum_{\mathrm{contact}\ i} \mathrm{CO}_i = \frac{1}{N_c}\sum_{j<k} \Gamma_{jk}(k-j), \qquad (19)$$

where N_c is the number of native contacts, CO_i is the contact order of the contact i, and Γ_{jk} gives the contact map and takes the value 1 when the residues j and k form a contact and 0 otherwise. This quantity could rewritten as

$$\mathrm{ACO} = \frac{1}{N_c}\sum_{j<k}\sum_{r} \Gamma_{jk}\theta_{r,jk} = \frac{1}{N}\sum_{r} n(r). \qquad (20)$$

Here, $\theta_{r,jk}$ describes if the residue r is in the related loop of the contact formed between the residues j and k, which takes the value 1 when $j \le r < k$ and 0 otherwise. The quantity $n(r) = \Sigma_{j<k}\Gamma_{jk}\theta_{r,jk}$ records the number of contacts which the residue r is involved in. Clearly, various amino acids may have different values of $n(r)$, thus have the different contributions to the quantity ACO. Physically, the hydrophobic residues and the flexible residues may have larger $n(r)$. Strong interactions between hydrophobic amino acids may effectively shorten the loop lengths for long-range contacts, and the flexible residues could reduce the Kuhn length of protein chains and help the formations of many local contacts. As a binary approximation [namely, assuming hydrophobic residues and flexible residues have the same $n(r)$ and $n(r)=0$ for other amino acids], the quantity ACO could be represented with the number of hydrophobic and flexible amino acids (namely, our effective length). This demonstrates the intrinsic connection between the structure-based methods and the composition-related protocols. The kind of consistence is also observed in recent studies [39]. It is found that the folding rates could be predicted from both the structure-related factor (the geometric contact number) and the composition of proteins [39]. Especially, the definition of geometric contact number is related to the well-packed nonlocal contacts. This is consistent with our assumption for effective length. All these agreements indicate that the success of various protocols generally comes from the same physical principles, and our theory catches the key factors of such considerations.

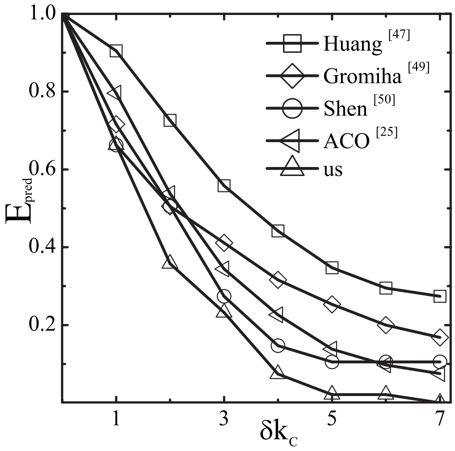

FIG. 6. Comparison between various prediction methods. The percentage of failed predictions E_{pred} for various threshold δk_c with five kinds of prediction methods.

It is worth noting that our method shares the same philosophy as that of Ivankov and Finkelstein [36] except for the different definitions of the effective length. It is found out that the amino acids favoring helical structure are suppressed in the definition for the effective length. The strong helix formers E, A, and L generally appear late in the optimal sets ($n \ge 16$ for E, $n=20$ for A, and $n \ge 8$ for L). Especially, although the amino acid L has high hydrophobicity, it enters into the optimal set rather late and even later for the cases with larger β, such as $n \ge 14$ for the case with $\beta=0.6$. Similar trends are also observed for those weak helix formers such as K, Q, and M. These observations is physically consistent with the structure-based effective length by Ivankov and Finkelstein [36]. This kind of consistency reveals that the physics behind both effective lengths is the same.

Compared with other sequence-based protocols, our method has its simplicity. By integrating sequence information with chain length, our method works for a larger number of proteins with various sizes. As a comparison, the folding rates of all the 95 proteins are predicted with our method and three other methods published recently [47,49,50]. For a protein, a prediction could be practically regarded as a false one when the difference between the predicted and experimental rates, Δk, is larger than a certain threshold δk_c. Therefore, for an assigned threshold δk_c, the percentage of false predictions with the concerned method, E_{pred}, could act as a measure for the quality of the concerned method. The variations of E_{pred} for four kinds of methods with different thresholds δk_c are given in Fig. 6. It is found that our method generally has the smallest value of E_{pred} for various thresholds. This indicates that our method has the best predictability though we have fewer parameters. More detailed analysis points out that the failed predictions by other methods are often related to the small peptides or large proteins. This kind of failure may be ascribed to the neglect of length effect in those studies. These comparisons reflect the necessity to consider the length effect for the prediction of folding rates and reveal the physical reason for the success with our simple factor in rate predictions. The similar quantitative comparison between our method and the prediction with ACO is also given in Fig. 6.

The result also demonstrates that our effective length has a more accurate predictability.

IV. CONCLUSIONS

The folding of proteins is a process controlled by complex free-energy landscapes. In this work, the relationship between amino-acid compositions and folding rates is discussed through a model including both the entropic and energetic terms. An effective length is defined to be a predictor for the folding rates. Although derived from a coarse-grained model, this predictor could make prediction for folding rates within a reasonable precision. Our simplifications of the amino-acid alphabet into hydrophobic, flexible polar, and rigid polar ones grasp the key factors related to the folding rate. In fact, there are other factors besides hydrophobicity and flexibility, such as long-range electrostatic interactions. The influence of these factors on the folding rate should not be neglected. We believe that it is easy to extend our method with more detailed considerations of amino-acid compositions, which would probably enhance the predictability further.

ACKNOWLEDGMENTS

This work was supported by National Basic Research Program (Grants No. 2006CB910302 and No. 2007CB814806), National Natural Science Foundation (Grants No. 10834002 and No. 10774069), and National Natural Science Foundation of Jiangsu Province (Grant No. BK2009008). J.W. thanks the support of FANEDD.

[1] C. B. Anfinsen, Science **181**, 223 (1973).
[2] B. van den Berg, R. J. Ellis, and C. M. Dobson, EMBO J. **18**, 6927 (1999).
[3] W. Wang, W. X. Xu, Y. Levy, E. Trizac, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **106**, 5517 (2009).
[4] T. E. Creighton, Curr. Biol. **7**, R380 (1997).
[5] Y. Cao and H. B. Li, J. Mol. Biol. **375**, 316 (2008).
[6] C. J. Camacho and D. Thirumalai, Protein Sci. **5**, 1826 (1996).
[7] J. S. Merkel and L. Regan, J. Biol. Chem. **275**, 29200 (2000).
[8] D. Baker, Nature (London) **405**, 39 (2000).
[9] B. Oztop, M. R. Ejtehadi, and S. S. Plotkin, Phys. Rev. Lett. **93**, 208105 (2004).
[10] C. Clementi and S. S. Plotkin, Protein Sci. **13**, 1750 (2004).
[11] S. Takada, Proc. Natl. Acad. Sci. U.S.A. **96**, 11698 (1999).
[12] E. Alm and D. Baker, Proc. Natl. Acad. Sci. U.S.A. **96**, 11305 (1999).
[13] O. V. Galzitskaya and A. V. Finkelstein, Proc. Natl. Acad. Sci. U.S.A. **96**, 11299 (1999).
[14] C. J. Tsai and R. Nussinov, Protein Eng. **14**, 723 (2001).
[15] V. Muñoz and W. A. Eaton, Proc. Natl. Acad. Sci. U.S.A. **96**, 11311 (1999).
[16] M. H. Hao and H. A. Scheraga, J. Mol. Biol. **277**, 973 (1998).
[17] P. Das, S. Matysiak, and C. Clementi, Proc. Natl. Acad. Sci. U.S.A. **102**, 10141 (2005).
[18] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, Science **267**, 1619 (1995).
[19] P. G. Wolynes, Q. Rev. Biophys. **38**, 405 (2005).
[20] D. Thirumalai, J. Phys. I **5**, 1457 (1995).
[21] A. M. Gutin, V. I. Abkevich, and E. I. Shakhnovich, Phys. Rev. Lett. **77**, 5433 (1996).
[22] A. V. Finkelstein and A. Y. Badretdinov, Folding Des. **2**, 115 (1997).
[23] P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **94**, 6170 (1997).
[24] A. V. Finkelstein, D. N. Ivankov, S. O. Garbuzynskiy, and O. V. Galzitskaya, Curr. Protein Pept. Sci. **8**, 521 (2007).
[25] O. V. Galzitskaya, S. O. Garbuzynskiy, D. N. Ivankov, and A. V. Finkelstein, Proteins: Struct., Funct., Bioinf. **51**, 162 (2003).
[26] A. Y. Istomin, D. J. Jacobs, and D. R. Livesay, Protein Sci. **16**, 2564 (2007).
[27] K. W. Plaxco, K. T. Simons, and D. Baker, J. Mol. Biol. **277**, 985 (1998).
[28] K. W. Plaxco, K. T. Simons, I. Ruczinski, and D. Baker, Biochemistry **39**, 11177 (2000).
[29] N. Koga and S. Takada, J. Mol. Biol. **313**, 171 (2001).
[30] M. M. Gromiha and S. Selvaraj, J. Mol. Biol. **310**, 27 (2001).
[31] H. Zhou and Y. Zhou, Biophys. J. **82**, 458 (2002).
[32] D. N. Ivankov, S. O. Garbuzynskiy, E. Alm, K. W. Plaxco, D. Baker, and A. V. Finkelstein, Protein Sci. **12**, 2057 (2003).
[33] T. R. Weikl and K. A. Dill, J. Mol. Biol. **329**, 585 (2003).
[34] H. Gong, D. G. Isom, R. Srinivasan, and G. D. Rose, J. Mol. Biol. **327**, 1149 (2003).
[35] B. Nölting, W. Schälike, P. Hampel, F. Grundig, S. Gantert, N. Sips, W. Bandlow, and P. X. Qi, J. Theor. Biol. **223**, 299 (2003).
[36] D. N. Ivankov and A. V. Finkelstein, Proc. Natl. Acad. Sci. U.S.A. **101**, 8942 (2004).
[37] I. B. Kuznetsov and S. Rackovsky, Proteins: Struct., Funct., Bioinf. **54**, 333 (2003).
[38] O. V. Galzitskaya and S. O. Garbuzynskiy, Proteins: Struct., Funct., Bioinf. **63**, 144 (2006).
[39] Z. Ouyang and J. Liang, Protein Sci. **17**, 1256 (2008).
[40] D. V. Ivankov, N. S. Bogatyreva, Y. L. M, and O. V. Galzitskaya, PLoS ONE **4**, e6476 (2009).
[41] M. M. Gromiha and S. Selvaraj, Curr Bioinf **3**, 1 (2008).
[42] D. E. Kim, H. D. Gu, and D. Baker, Proc. Natl. Acad. Sci. U.S.A. **95**, 4982 (1998).
[43] H. Gu, N. Doshi, D. E. Kim, K. T. Simons, J. V. Santiago, S. Nauli, and D. Baker, Protein Sci. **8**, 2734 (1999).
[44] H. Kono and J. G. Saven, J. Mol. Biol. **306**, 607 (2001).
[45] A. Zarrine-Afsar, S. Dahesh, and A. R. Davidson, J. Mol. Biol. **373**, 764 (2007).
[46] G. Calloni, N. Taddei, K. W. Plaxco, G. Ramponi, M. Stefani, and F. Chiti, J. Mol. Biol. **330**, 577 (2003).
[47] J. Huang and J. Tian, Proteins: Struct., Funct., Bioinf. **63**, 551 (2006).
[48] B. Ma, J. Guo, and H. Zang, Proteins: Struct., Funct., Bioinf. **65**, 362 (2006).

[49] M. M. Gromiha, A. M. Thangakani, and S. Selvaraj, Nucleic Acids Res. **34**, W70 (2006).

[50] H. B. Shen, J. N. Song, and K. C. Chou, J Biomed Sci Eng **2**, 136 (2009).

[51] L. T. Huang and M. M. Gromiha, J. Comput. Chem. **29**, 1675 (2008).

[52] M. M. Gromiha, J. Chem. Inf. Comput. Sci. **43**, 1481 (2003).

[53] M. M. Gromiha, J. Chem. Inf. Comput. Sci. **45**, 494 (2005).

[54] S. Kawashima, H. Ogata, and M. Kanehisa, Nucleic Acids Res. **27**, 368 (1999).

[55] S. Kawashima and M. Kanehisa, Nucleic Acids Res. **28**, 374 (2000).

[56] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa, Nucleic Acids Res. **36**, D202 (2008).

[57] G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, J. Mol. Biol. **7**, 95 (1963).

[58] A. Irback, F. Sjunnesson, and S. Wallin, Proc. Natl. Acad. Sci. U.S.A. **97**, 13614 (2000).

[59] *Bioinformatics: Sequence and Genome Analysis*, edited by D. W. Mount (Cold Spring Harbor Laboratory Press, New York, 2001).

[60] K. C. Chou, FEBS Lett. **363**, 127 (1995).

[61] B. Mao, K. C. Chou, and C. T. Zhang, Protein Eng. **7**, 319 (1994).

[62] C. T. Zhang, K. C. Chou, and G. M. Maggiora, Protein Eng. **8**, 425 (1995).

[63] X. Qi and J. J. Portman, Proc. Natl. Acad. Sci. U.S.A. **105**, 11164 (2008).

[64] S. Reuveni, R. Granek, and J. Klafter, Phys. Rev. Lett. **100**, 208101 (2008).

[65] S. S. Plotkin and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **97**, 6509 (2000).

[66] K. L. Maxwell, D. Wildes, A. Zarrine-Afsar, M. A. De Los Rios, A. G. Brown, C. T. Friel, L. Hedberg, J. C. Horng, D. Bona, E. J. Miller, A. Vallée-Bélisle, E. R. G. Main, F. Bemporad, L. L. Qiu, K. Teilum, N. D. Vu, A. M. Edwards, I. Ruczinski, F. M. Poulsen, B. B. Kragelund, S. W. Michnick, F. Chiti, Y. W. Bai, S. J. Hagen, L. Serrano, M. Oliveberg, D. P. Raleigh, P. Wittung-Stafshede, S. E. Radford, S. E. Jackson, T. R. Sosnick, S. Marqusee, A. R. Davidson, and K. W. Plaxco, Protein Sci. **14**, 602 (2005).

[67] *IUPAC Compendium of Chemical Terminology*, 2nd ed., edited by A. D. McNaught and A. Wilkinson (Blackwell Scientific Publications, Oxford, 1997).

[68] L. A. Mirny and E. I. Shakhnovich, J. Mol. Biol. **264**, 1164 (1996).

[69] M. Vendruscolo, L. A. Mirny, E. I. Shakhnovich, and E. Domany, Proteins Struct Genet Funct **41**, 192 (2000).

[70] N. S. Bogatyreva, A. A. Osypov, and D. N. Ivankov, Nucleic Acids Res. **37**, D342 (2009).

[71] F. Rousseau, J. W. Schymkowitz, H. R. Wikinson, and L. S. Itzhaki, Proc. Natl. Acad. Sci. U.S.A. **98**, 5596 (2001).

[72] W. J. Wedemeyer, E. Welker, and H. A. Scheraga, Biochemistry **41**, 14637 (2002).

[73] J. Wang and W. Wang, Nat. Struct. Biol. **6**, 1033 (1999).

[74] F. Huang and W. M. Nau, Angew. Chem., Int. Ed. **42**, 2269 (2003).

[75] M. Vihinen, E. Torkkila, and P. Riikonen, Proteins **19**, 141 (1994).

[76] P. A. Karplus and G. E. Schulz, Naturwiss. **72**, 212 (1985).

[77] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997).

[78] S. Miyazawa and R. L. Jernigan, J. Mol. Biol. **256**, 623 (1996).