FoldRate: A Web-Server for Predicting Protein Folding Rates from Primary Sequence

Kuo-Chen Chou[1,2,*] and Hong-Bin Shen[1,2,*]

[1]*Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, California 92130, USA*

[2]*Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong University, 800 Dongchuan Road, Shanghai, 200240, China*

Abstract: With the avalanche of gene products in the postgenomic age, the gap between newly found protein sequences and the knowledge of their 3D (three dimensional) structures is becoming increasingly wide. It is highly desired to develop a method by which one can predict the folding rates of proteins based on their amino acid sequence information alone. To address this problem, an ensemble predictor, called FoldRate, was developed by fusing the folding-correlated features that can be either directly obtained or easily derived from the sequences of proteins. It was demonstrated by the jackknife cross-validation on a benchmark dataset constructed recently that FoldRate is at least comparable with or even better than the existing methods that, however, need both the sequence and 3D structure information for predicting the folding rate. As a user-friendly web-server, FoldRate is freely accessible to the public at www.csbio.sjtu.edu.cn/bioinf/FoldRate/, by which one can get the desired result for a query protein sequence in around 30 seconds.

Keywords: Protein folding rate, Ensemble predictor, Fusion approach, Web-server, FoldRate.

I. INTRODUCTION

A protein can function properly only if it is folded into a very special and individual shape or conformation, i.e., has the correct secondary, tertiary and quaternary structure [1]. Failure to fold into the intended 3D (three-dimensional) structure usually produces inactive proteins or misfolded proteins [2] that may cause cell death and tissue damage [3] and be implicated in prion diseases such as bovine spongiform encephalopathy (BSE, also known as “mad cow disease”) in cattle and Creutzfeldt-Jakob disease (CJD) in humans. All prion diseases are currently untreatable and are always fatal [4].

Since each protein begins as a polypeptide translated from a sequence of mRNA as a linear chain of amino acids, it is interesting to study the folding rates of proteins from their primary sequences. Actually, protein chains can fold into the functional 3D structures with quite different rates, varying from several microseconds [5] to even an hour [6]. Since the 3D structure of a protein is determined by its primary sequence, we can assume the same is true for its folding rate. In view of this, we are challenged by an interesting question: Given a protein sequence, can we find its folding rate? Although the answer can be found by conducting various biochemical experiments, doing so is both time-consuming and expensive. Also, although a number of prediction methods were proposed [7-12], they need the input from the 3D structure of the protein concerned, and hence the prediction is feasible only after its 3D structure has been determined. Particularly, the newly-found protein sequences have been increasing explosively. For instance, in 1986 the Swiss-Prot databank (www.ebi.ac.uk/swissprot) contained merely 3,939 protein sequence entries, but the number has jumped to 428,650 according to version 57.0 of 24-March-2009, meaning that the number of protein sequence entries now is more than 108 times the number about 23 years ago. In contrast, as of 5-May-2009, the RCSB Protein Data Bank (http://www.rcsb.org/pdb) contains only 57,424 3D structure entries, meaning that the structure-known proteins is about 1.34% of sequence-known proteins. Facing the avalanche of protein sequences generated in the post-genomic age and also considering the huge gap between the numbers of known protein sequences and 3D structures, it is highly desired to develop an automated method that can rapidly and approximately predict the folding rates of proteins according to their sequence information alone.

The present study was initiated in an attempt to address this problem in hopes that our approach can play a complementary role to the existing methods [13, 14]. Below, let us first clarify the meaning of the protein folding rates as usually observed by experiments.

II. THE PROTEIN FOLDING RATE

Since the prediction object in the current study is the protein folding rate, a clear understanding of its implication is necessary. The folding rate of a protein chain observed by experiments is usually measured by the “apparent folding rate constant” [15], as denoted by K_f. It is instructive to unravel its relationship with the detailed rate constants, as given below.

*Address correspondence to these authors at the Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, California 92130, USA; Fax: 858-380-4623, 86-21-3420-5320; E-mail: kcchou@gordonlifescience.org; hbshen@sjtu.edu.cn

The apparent folding rate constant $K_{\rm f}$ for a protein chain is defined *via* the following differential equation:

$$\begin{cases} \dfrac{{\rm dP}_{\rm unfold}(t)}{{\rm d}t} = -K_{\rm f}\,{\rm P}_{\rm unfold}(t) \\ \dfrac{{\rm dP}_{\rm fold}(t)}{{\rm d}t} = K_{\rm f}\,{\rm P}_{\rm unfold}(t) \end{cases} \quad (1)$$

where ${\rm P}_{\rm unfold}(t)$ and ${\rm P}_{\rm fold}(t)$ represent the concentrations of its unfolded state and folded state, respectively. Suppose the total protein concentration is C_0, and initially only the unfolded protein is present; i.e., ${\rm P}_{\rm unfold}(t) = C_0$ and ${\rm P}_{\rm fold}(t) = 0$ when $t = 0$. Subsequently, the protein system is subjected to a sudden change in temperature, solvent, or any other factor that causes the protein to fold. Obviously, the solution for Eq. 1 is:

$$\begin{cases} {\rm P}_{\rm unfold}(t) = C_0 \exp(-K_{\rm f} t) \\ {\rm P}_{\rm fold}(t) = C_0\left[1-\exp(-K_{\rm f} t)\right] \end{cases} \quad (2)$$

It can be seen from the above equation that the larger the $K_{\rm f}$, the faster the folding rate will be. Given the value of $K_{\rm f}$, the half-life of an unfolded protein chain can be expressed by:

$$T_{1/2} = -\frac{\ln(1/2)}{K_{\rm f}} \simeq 0.693/K_{\rm f} \quad (3)$$

which can also be used to reflect the time that is needed for a protein chain to be half folded. However, the actual folding process is much more complicated than the one as described by Eq. 1 even if the reverse rate for the folding system concerned can be ignored. As an illustration, let us consider the following three-state folding mechanism:

$${\rm P}_{\rm unfold} \xrightarrow{k_{12}} {\rm P}_{\rm inter} \xrightarrow{k_{23}} {\rm P}_{\rm fold} \quad (4)$$

where ${\rm P}_{\rm inter}(t)$ represents the concentration of an intermediate state between the unfolded and folded states, k_{12} is the rate constant for ${\rm P}_{\rm unfold}$ converting to ${\rm P}_{\rm inter}$, and k_{23} the rate constant for ${\rm P}_{\rm inter}$ converting to ${\rm P}_{\rm fold}$. Thus we have the following kinetic equation:

$$\begin{cases} \dfrac{{\rm dP}_{\rm unfold}(t)}{{\rm d}t} = -k_{12}{\rm P}_{\rm unfold}(t) \\ \dfrac{{\rm dP}_{\rm inter}(t)}{{\rm d}t} = k_{12}{\rm P}_{\rm unfold}(t) - k_{23}{\rm P}_{\rm inter}(t) \\ \dfrac{{\rm dP}_{\rm fold}(t)}{{\rm d}t} = k_{23}{\rm P}_{\rm inter}(t) \end{cases} \quad (5)$$

Eqs.4 and 5 can be expressed *via* an intuitive diagram called “directed graph” or “digraph” $\mathbb{G}$ [15, 16] as shown in Fig. (**1a**). To reflect the variation of the concentrations of the three protein states with time, the digraph $\mathbb{G}$ is further transformed to the phase digraph $\tilde{\mathbb{G}}$ [15, 16] as shown in Fig. (**1b**), where s is an interim parameter associated with the following Laplace transform:

$$\begin{cases} \tilde{\rm P}_{\rm unfold}(s) = \int_0^\infty {\rm P}_{\rm unfold}(t)\exp(-ts){\rm d}t \\ \tilde{\rm P}_{\rm inter}(s) = \int_0^\infty {\rm P}_{\rm inter}(t)\exp(-ts){\rm d}t \\ \tilde{\rm P}_{\rm fold}(s) = \int_0^\infty {\rm P}_{\rm fold}(t)\exp(-ts){\rm d}t \end{cases} \quad (6)$$

where $\tilde{\rm P}_{\rm unfold}$, $\tilde{\rm P}_{\rm inter}$ and $\tilde{\rm P}_{\rm fold}$ are the phase concentrations of ${\rm P}_{\rm unfold}$, ${\rm P}_{\rm inter}$ and ${\rm P}_{\rm fold}$, respectively. Thus, according to the

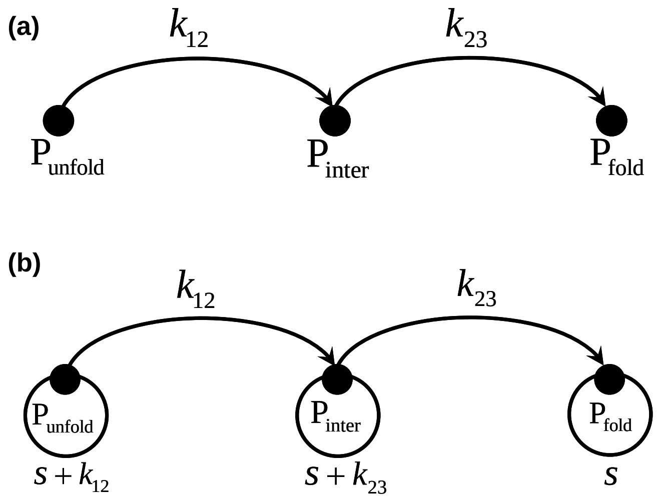

Fig. (1). (**a**) The directed graph or digraph $\mathbb{G}$ [15, 16] for the three-state protein folding mechanism as schematically expressed by Eq. 4 and formulated by Eq. 5. (**b**) The phase digraph $\tilde{\mathbb{G}}$ obtained from $\mathbb{G}$ of panel (**a**) according to graphic rule 4 for enzyme and protein folding kinetics [15, 16] that is also called “Chou’s graphic rule for non-steady-state kinetics” in literatures (see, e.g., [17]). The symbol s in the phase digraph $\tilde{\mathbb{G}}$ is an interim parameter (see the text for further explanation).

phase digraph $\tilde{\mathbb{G}}$ of Fig. (**1b**) and using the graphic rule 4 [15, 16], which is also called "Chou's graphic rule for non-steady-state kinetics" in literatures (see, e.g., [17]), we can directly write out the following phase concentrations:

$$\tilde{\mathrm{P}}_{\mathrm{unfold}}(s)=\frac{(s+k_{23})sC_0}{s\left[(s+k_{23})s+k_{12}s+k_{12}k_{23}\right]}=\frac{(s+k_{23})C_0}{(s+k_{12})(s+k_{23})}=\frac{C_0}{s+k_{12}} \quad (7.1)$$

$$\tilde{\mathrm{P}}_{\mathrm{inter}}(s)=\frac{k_{12}sC_0}{s\left[(s+k_{23})s+k_{12}s+k_{12}k_{23}\right]}=\frac{k_{12}C_0}{(s+k_{12})(s+k_{23})} \quad (7.2)$$

$$\tilde{\mathrm{P}}_{\mathrm{fold}}(s)=\frac{k_{12}k_{23}C_0}{s\left[(s+k_{23})s+k_{12}s+k_{12}k_{23}\right]}=\frac{k_{12}k_{23}C_0}{s(s+k_{12})(s+k_{23})} \quad (7.3)$$

Through the above phase concentrations and using Laplace transform table (see, e.g., [18] or any standard mathematical tables), we can immediately obtain the desired concentrations for $\mathrm{P}_{\mathrm{unfold}}$, $\mathrm{P}_{\mathrm{inter}}$ and $\mathrm{P}_{\mathrm{fold}}$ of Eq. 5, as given by:

$$\begin{cases} \mathrm{P}_{\mathrm{unfold}}(t)=C_0\,\mathrm{e}^{-k_{12}t} \\ \mathrm{P}_{\mathrm{inter}}(t)=\dfrac{k_{12}C_0}{k_{23}-k_{12}}\left(\mathrm{e}^{-k_{12}t}-\mathrm{e}^{-k_{23}t}\right) \\ \mathrm{P}_{\mathrm{fold}}(t)=\dfrac{C_0}{k_{23}-k_{12}}\left(k_{12}\,\mathrm{e}^{-k_{23}t}-k_{23}\,\mathrm{e}^{-k_{12}t}\right)+C_0 \end{cases} \quad (8)$$

Accordingly, it follows from the above equation that:

$$\frac{\mathrm{dP}_{\mathrm{fold}}(t)}{\mathrm{d}t}=\frac{k_{12}k_{23}C_0}{k_{23}-k_{12}}\left(\mathrm{e}^{-k_{12}t}-\mathrm{e}^{-k_{23}t}\right)=\frac{k_{12}k_{23}}{k_{23}-k_{12}}\left[1-\mathrm{e}^{-(k_{23}-k_{12})t}\right]\mathrm{P}_{\mathrm{unfold}} \quad (9)$$

Comparing Eq. 9 with Eq. 1, we obtain the following equivalent relation:

$$K_{\mathrm{f}}\Leftrightarrow\frac{k_{12}k_{23}}{k_{23}-k_{12}}\left[1-\mathrm{e}^{-(k_{23}-k_{12})t}\right] \quad (10)$$

meaning that the apparent folding rate constant K_{f} is a function of not only the detailed rate constants, but also t. Accordingly, K_{f} is actually not a constant but will change with time. Only when $k_{23}\gg k_{12}$ and $k_{23}\gg 1$, can Eq. 10 be reduced to $K_{\mathrm{f}}\simeq k_{12}$ and Eq. 9 to:

$$\frac{\mathrm{dP}_{\mathrm{folded}}(t)}{\mathrm{d}t}\simeq k_{12}\mathrm{P}_{\mathrm{unfold}}(t)=K_f\mathrm{P}_{\mathrm{unfold}}(t) \quad (11)$$

and K_f be treated as a constant.

Even for a two-state protein folding system when the reverse effect needs to be considered, i.e., the system described by the following scheme and equation:

$$\mathrm{P}_{\mathrm{unfold}}\underset{k_{21}}{\overset{k_{12}}{\rightleftarrows}}\mathrm{P}_{\mathrm{fold}} \quad (12)$$

$$\begin{cases} \dfrac{\mathrm{dP}_{\mathrm{unfold}}(t)}{\mathrm{d}t}=-k_{12}\mathrm{P}_{\mathrm{unfold}}(t)+k_{21}\mathrm{P}_{\mathrm{fold}}(t) \\ \dfrac{\mathrm{dP}_{\mathrm{fold}}(t)}{\mathrm{d}t}=k_{12}\mathrm{P}_{\mathrm{unfold}}(t)-k_{21}\mathrm{P}_{\mathrm{fold}}(t) \end{cases} \quad (13)$$

where k_{21} represents the reverse rate constant converting $\mathrm{P}_{\mathrm{fold}}$ back to $\mathrm{P}_{\mathrm{unfold}}$. With the similar derivation by using the non-steady state graphic rule [15, 16] as described above, we have now the following equivalent relation:

$$K_{\mathrm{f}}\Leftrightarrow\left\{\frac{k_{12}\left(k_{12}+k_{21}\right)}{k_{21}+k_{12}\exp\left[-(k_{12}+k_{21})t\right]}\exp\left[-(k_{12}+k_{21})t\right]\right\} \quad (14)$$

indicating once again that, even for the two-state folding system of Eq. 12, the apparent folding rate constant K_{f} can be treated as a constant only when $k_{12}\gg k_{21}$ and $k_{12}\gg 1$.

It can be imagined that for a general multi-state folding system, K_{f} will be much more complicated. It is important to keep this in mind to avoid confusion of the apparent rate constants with the detailed rate constants.

We can also see from the above derivation that using the graphic analysis to deal with kinetic systems is quite efficient and intuitive, particularly in dealing with complicated kinetic systems. For more discussions about the graphic analysis and its applications to kinetic systems, see [19-25].

III. MATERIALS AND METHODS

To develop an effective statistical predictor, the following three things are indispensable: (1) a valid benchmark dataset; (2) a mathematical expression for the samples that can effectively reflect their intrinsic correlation with the object to be predicted; and (3) a powerful prediction algorithm or engine. The three necessities for establishing the current protein folding rate predictor were realized *via* the following procedures.

1. Benchmark Dataset

The dataset recently constructed by Ouyang and Liang [12] was used in the current study. It contains 80 proteins whose apparent folding rate constants (K_{f}) have been experimentally determined. However, it is instructive to point out that, when the experimentally measured K_{f} is a constant independent on time t, the conditions as mentioned in Section II (see Eqs.10 and 14 and the relevant texts) must be satisfied. Accordingly, the folding kinetic mechanisms for all these 80 proteins can be approximately described by Eq. 1, and hence there is no need here to specify which proteins belong to the two-state folding and which ones to the three-state or other multiple-state as done in [12]. Furthermore, although the experimental 3D structures of the 80 proteins are known, none of this kind of information will be used here because we are intending to develop a statistical predictor purely based on the experimental K_{f} values of proteins and their sequence information alone. If the success rates thus

Table 1. The Apparent Folding Rate Constant K_f (sec^{-1}) of the 80 Proteins in the Benchmark Dataset $\mathbb{S}_\mathrm{bench}$ and their Half-Folding Time $T_{1/2}$ (sec) (cf. Eq. 3)

Number	PDB Code	$\ln K_\mathrm{f}$	K_f (sec^{-1})	$T_{1/2}$ (sec)
1	1APS	-1.47	2.299×10^{-1}	3.015
2	1BA5	5.91	3.687×10^{2}	1.88×10^{-3}
3	1BDD	11.69	1.194×10^{5}	6.0×10^{-6}
4	1C8C	6.95	1.043×10^{3}	6.64×10^{-4}
5	1C9O	7.20	1.339×10^{3}	5.17×10^{-4}
6	1CSP	6.54	6.92×10^{2}	1.001×10^{-3}
7	1DIV_c	0.0	1.000	6.932×10^{-1}
8	1DIV_n	6.61	7.425×10^{2}	9.34×10^{-4}
9	1E0L	10.37	3.1888×10^{4}	2.2×10^{-5}
10	1E0M	8.85	6.974×10^{3}	9.9×10^{-5}
11	1ENH	10.53	3.742×10^{4}	1.9×10^{-5}
12	1FEX	8.19	3.604×10^{3}	1.92×10^{-4}
13	1FKB	1.45	4.263	1.626×10^{-1}
14	1FMK	4.05	5.7440×10^{1}	1.208×10^{-2}
15	1FNF_9	-0.92	3.985×10^{-1}	1.739
16	1G6P	6.30	5.446×10^{2}	1.273×10^{-3}
17	1HDN	2.69	1.473×10^{1}	4.705×10^{-2}
18	1IDY	8.73	6.186×10^{3}	1.12×10^{-4}
19	1IMQ	7.28	1.451×10^{3}	4.78×10^{-4}
20	1K8M	-0.71	4.916×10^{-1}	1.410
21	1K9Q	8.37	4.316×10^{3}	1.61×10^{-4}
22	1L2Y	12.40	2.428×10^{5}	3.0×10^{-6}
23	1LMB	8.50	4.915×10^{3}	1.41×10^{-4}
24	1MJC	5.23	1.868×10^{2}	3.711×10^{-3}
25	1N88	3.0	2.009×10^{1}	3.451×10^{-2}
26	1NYF	4.54	9.369×10^{1}	7.398×10^{-3}
27	1PGB_b	12.0	1.628×10^{5}	4.0×10^{-6}
28	1PIN	9.37	1.173×10^{4}	5.9×10^{-5}
29	1PKS	-1.06	3.465×10^{-1}	2.001
30	1PRB	12.90	4.003×10^{5}	2.0×10^{-6}
31	1PSE	1.17	3.222	2.151×10^{-1}
32	1QTU	-0.36	6.977×10^{-1}	9.935×10^{-1}
33	1RFA	7.0	1.097×10^{3}	6.32×10^{-4}
34	1SHG	2.10	8.166	8.488×10^{-2}
35	1TEN	1.06	2.886	2.402×10^{-1}
36	1URN	5.76	3.173×10^{2}	2.184×10^{-3}
37	1VII	11.51	9.971×10^{4}	7.0×10^{-6}
38	1WIT	0.41	1.507	4.6×10^{-1}
39	2A3D	12.7	3.277×10^{5}	2.0×10^{-6}
40	2ACY	0.84	2.317	2.992×10^{-1}
41	2AIT	4.21	6.736×10^{1}	1.029×10^{-2}
42	2CI2	3.87	4.794×10^{1}	1.446×10^{-2}
43	2HQI	0.18	1.197	5.790×10^{-1}
44	2PDD	9.69	1.616×10^{4}	4.3×10^{-5}
45	2PTL	4.10	6.034×10^{1}	1.149×10^{-2}
46	2ABD	6.48	6.520×10^{2}	1.063×10^{-3}
47	2CRO	5.35	2.106×10^{2}	3.291×10^{-3}
48	1UZC	8.68	5.884×10^{3}	1.18×10^{-4}
49	1CEI	5.8	3.303×10^{2}	2.099×10^{-3}
50	1BRS	3.37	2.908×10^{1}	2.384×10^{-2}

(Table 1). Contd…..

Number	PDB Code	$\ln K_f$	K_f (sec^{-1})	$T_{1/2}$ (sec)
51	2A5E	3.50	3.312×10^{1}	2.093×10^{-2}
52	1TIT	3.6	3.660×10^{1}	1.894×10^{-2}
53	1FNF_1	5.48	2.399×10^{2}	2.890×10^{-3}
54	1HNG	1.8	6.050	1.146×10^{-1}
55	1ADW	0.64	1.897	3.654×10^{-1}
56	1EAL	1.3	3.669	1.889×10^{-1}
57	1IFC	3.4	2.996×10^{1}	2.313×10^{-2}
58	1OPA	1.4	4.055	1.709×10^{-1}
59	1HCD	1.1	3.004	2.307×10^{-1}
60	1BEB	-2.20	1.108×10^{-1}	6.256
61	1B9C	-2.76	6.329×10^{-2}	1.095×10^{1}
62	1I1B	-4.01	1.813×10^{-2}	3.822×10^{1}
63	1PGB_a	6.40	6.018×10^{2}	1.152×10^{-3}
64	1UBQ	5.90	3.650×10^{2}	1.899×10^{-3}
65	1GXT	4.39	8.064×10^{1}	8.596×10^{-3}
66	1SCE	4.17	6.472×10^{1}	1.071×10^{-2}
67	1HMK	2.79	1.628×10^{1}	4.257×10^{-2}
68	3CHY	1.0	2.718	2.550×10^{-1}
69	1HEL	1.25	3.490	1.986×10^{-1}
70	1DK7	0.83	2.293	3.022×10^{-1}
71	1JOO	0.30	1.350	5.135×10^{-1}
72	2RN2	1.41	4.096	1.692×10^{-1}
73	1RA9	-2.46	8.543×10^{-2}	8.113
74	1PHP_c	-3.44	3.207×10^{-2}	2.162×10^{1}
75	1PHP_n	2.30	9.974	6.949×10^{-2}
76	2BLM	-1.24	2.894×10^{-1}	2.395
77	1QOP_a	-2.5	8.209×10^{-2}	8.444
78	1QOP_b	-6.9	1.008×10^{-3}	6.878×10^{2}
79	1BTA	1.11	3.034	2.284×10^{-1}
80	1L63	4.10	6.034×10^{1}	1.1487×10^{-2}

obtained can be comparable or about the same as those by the method of Ouyang and Liang where the 3D structure information was needed as an input [12], the new predictor will have the advantage of being able to also cover those proteins whose 3D structures are unknown yet. This is particularly useful due to the huge gap between the number of known protein sequences and the number of known protein 3D structures, as mentioned in Section I.

For readers' convenience, the benchmark dataset, denoted as $\mathbb{S}_{\text{bench}}$, is given in Appendix **A** which can also be downloaded from the web-site at www.csbio.sjtu.edu.cn/bioinf/FoldRate/. As we can see there, $\ln K_{\text{f}}$ (where ln means taking the natural logarithm for the number right after it) ranges from -6.9 to 12.9; i.e., K_{f} ranges from $\mathrm{e}^{-6.9} \simeq 1.01\times10^{-3}$ to $\mathrm{e}^{12.9} \simeq 4.00\times10^{6}$ (where $\mathrm{e} \simeq 2.718$ is the natural number, sometimes called Euler's number), meaning that the apparent folding rate constants of the 80 proteins span more than eight orders of magnitude (cf. Table **1**).

2. Sample Expression or Feature Extraction

As shown in [12], the features extracted from the 3D structures of proteins are very useful for predicting their folding rates. However, for the majority of proteins, their 3D structures are unknown yet. To enable the prediction model to cover as many proteins as possible, here let us focus on those features that can be derived from the amino acid sequential information alone, either directly or indirectly. Owing to the fact that smaller proteins usually (although far from always) fold faster than larger ones [26], and that α-helix and β-sheet are the two most major structural elements [27], our attention should be particularly focused on the size of proteins as well as the effects of α-helices and β-strands.

(a) Protein Size or Length Effect

In protein science, the length of a protein chain is usually measured by L, the number of amino acids it contains. Many lines of evidences (see, e.g., [12, 13]) have indicated that the length of a protein chain is correlated with its folding rate, suggesting that L, as well as its various functions, could be useful for representing protein samples in predicting their folding rates. Our preliminary studies showed that $\ln(L)$ was particularly remarkable in this regard and hence will be used in the current study.

(b) Predicted α-Helix Effect and the Effective Folding Chain Length

Driven by the short-range interaction, α-helices can be formed independently in a much faster pace than the entire structural frame. These helices can be treated as rigid blocks so as to reduce the original chain length L counted according to the number of amino acids. The effective folding chain length L_{eff}^{α} thus considered is given by [13]:

$$L_{\text{eff}}^{\alpha} = L - L_{\text{h}} + \lambda N_{\text{h-block}} \tag{15}$$

where L_{h} is the total number of amino acids in the helix blocks that can be easily predicted by using PSIPRED [28] for a given protein sequence; $N_{\text{h-block}}$ the number of predicted helix blocks; and λ the pseudo length of a helix block that was set at 3 in the current study, meaning that each helix block is equivalent to 3 amino acid units in length. Again, our preliminary studies showed that among various functions of L_{eff}^{α}, $\ln(L_{\text{eff}}^{\alpha})$ was particularly remarkable in correlation with the protein folding rates, and hence will be used in the current study.

(c) Effect of β-Sheet Propensity

It was hinted in some previous studies (see, e.g., [29, 30]) that the folding of a protein is strongly correlated with those amino acids that have a high propensity to form β-strands [31, 32]. To reflect the overall β-sheet propensity of a protein chain, let us take the following consideration. Suppose a protein chain is formulated by:

$$\mathbf{P} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\mathrm{R}_5\mathrm{R}_6\mathrm{R}_7\cdots\mathrm{R}_L \tag{16}$$

where the i-th residue R_i $(i = 1, 2, \cdots, L)$ can be one of the 20 different types of amino acids each having its own propensity to form β-strand [31]. The overall β-sheet propensity of the protein concerned is defined by:

$$\Phi^{\beta} = \frac{\sum_{i=1}^{L}\Psi_{\beta,i}}{L} \tag{17}$$

where $\Psi_{\beta,i}$ is the β-strand propensity for the i-th $(i = 1, 2, \cdots, L)$ amino acid in the protein $\mathbf{P}$. Note that before substituting the values of β-strand propensity into Eq. 17, they are subject to a Max-Min normalization as given by:

$$\Psi_{\beta,i} = \frac{\Psi_{\beta,i}^{0}}{\mathbf{Max}\{\Psi_{\beta}^{0}\} - \mathbf{Min}\{\Psi_{\beta}^{0}\}} \tag{18}$$

where $\Psi_{\beta,i}^{0}$ represent the original β-strand propensity value for R_i in Eq. 16 and can be obtained from [31] because it must be one of the 20 native amino acids, $\mathbf{Max}\{\Psi_{\beta}^{0}\}$ means taking the maximum value among the 20 original β-strand propensities, and $\mathbf{Min}\{\Psi_{\beta}^{0}\}$ the corresponding minimum one. For reader's convenience, the converted β-strand propensity value obtained through the Max-Min normalization procedure (cf. Eq. 18) for each of the 20 native amino acids is given in Table **2**, from which one can easily derive its overall β-sheet propensity, Φ^{β}, for any given protein sequence.

The values of $\ln(L), \ln(L_{\text{eff}}^{\alpha})$, and Φ^{β} for the 80 proteins in the benchmark dataset $\mathbb{S}_{\text{bench}}$ are given in Appendix **B**.

3. Prediction Algorithm

According to the above discussion, we have the following three quantitative features extracted from a protein sequence: $\ln(L), \ln(L_{\text{eff}}^{\alpha})$, and Φ^{β}. Each of these features derived from a protein may be correlated with its folding rate K_{f} through the following equations.

$$\ln\left(K_{\text{f}}^{(1)}\right) = a_1 + b_1 \ln(L) \tag{19.1}$$

$$\ln\left(K_{\text{f}}^{(2)}\right) = a_2 + b_2 \ln(L_{\text{eff}}^{\alpha}) \tag{19.2}$$

$$\ln\left(K_{\text{f}}^{(3)}\right) = a_3 + b_3 \Phi^{\beta} \tag{19.3}$$

where $K_{\text{f}}^{(i)}$ $(i = 1,\ 2,\ 3)$ are the protein folding rate constants predicted based on the length of protein, its α-helix related effective length, and its overall β-sheet propensity, respectively; while a_i and b_i are the corresponding parameters that can be determined through a training dataset by the following regression procedure [33].

First, let us just use the 80 proteins in the benchmark dataset $\mathbb{S}_{\text{bench}}$ (Appendix **A**) as the training data. Suppose the length, effective folding chain length, and overall β-sheet propensity for the k-th protein in the dataset are denoted by $L(k)$, $L_{\text{eff}}^{\alpha}(k)$, and $\Phi^{\beta}(k)$, respectively. In order to determine the coefficients of Eq. 19, let us define three objective functions given by:

$$\begin{cases} \Delta^{(1)} = \sum\limits_{k=1}^{80}\left\{\left[a_1 + b_1 \ln L(k)\right] - \ln\left[K_{\text{f}}(k)\right]\right\}^2 \\ \Delta^{(2)} = \sum\limits_{k=1}^{80}\left\{\left[a_2 + b_2 \ln L_{\text{eff}}^{\alpha}(k)\right] - \ln\left[K_{\text{f}}(k)\right]\right\}^2 \\ \Delta^{(3)} = \sum\limits_{k=1}^{80}\left\{\left[a_3 + b_3 \ln \Phi^{\beta}(k)\right] - \ln\left[K_{\text{f}}(k)\right]\right\}^2 \end{cases} \tag{20}$$

where $K_{\text{f}}(k)$ is the observed folding rate for the k-th protein in the dataset $\mathbb{S}_{\text{bench}}$ as given in Appendix **A**. The process of determining these coefficients is actually a process of

Table 2. The β-Strand Propensity Values for the 20 Native Amino Acids Converted According to the Max-Min Normalization Procedure of Eq. 18

Amino Acid Code		Propensity to form β-Strand	
Single Letter	Numerical Index u	Original $\Psi^{0}_{\beta,u}$	Max-Min Normalized $\Psi_{\beta,u}$
A	1	0.83	0.34
C	2	1.19	0.61
D	3	0.54	0.12
E	4	0.37	0.00
F	5	1.38	0.75
G	6	0.75	0.28
H	7	0.87	0.37
I	8	1.60	0.92
K	9	0.74	0.27
L	10	1.30	0.69
M	11	1.05	0.51
N	12	0.89	0.39
P	13	0.55	0.13
Q	14	1.10	0.54
R	15	0.93	0.42
S	16	0.75	0.28
T	17	1.19	0.61
V	18	1.70	1.00
W	19	1.37	0.75
Y	20	1.47	0.82

finding the minimum of $\Delta^{(i)}$ $(i=1,2,3)$, and hence can be easily obtained by the following equation:

$$\begin{cases} \dfrac{\partial\Delta^{(i)}}{\partial a_i}=0 \\ \dfrac{\partial\Delta^{(i)}}{\partial b_i}=0 \end{cases} \qquad (i=1,2,3) \tag{21}$$

Substituting Eq. 20 into Eq. 21, followed by using the data provided in Appendix **A** and the data derived therefrom as given in Appendix **B**, we can easily determine the coefficients in Eq. 19, as given below:

$$\begin{cases} a_1=32.4216, & b_1=-6.4077 \\ a_2=26.6906, & b_2=-5.5966 \\ a_3=30.7239, & b_3=-58.0109 \end{cases} \tag{22}$$

However, as explained below, the accuracy of a predictor is usually examined by the jackknife cross-validation in which the query sample should be in term excluded from the training dataset. Thus, instead of Eqs. 20-21, we should have:

$$\begin{cases} \Delta^{(1)}(k)=\sum\limits_{i\neq k}^{80}\left\{\left[a_1(k)+b_1(k)\ln L(i)\right]-\ln\left[K_{\mathrm{f}}(i)\right]\right\}^2 \\ \Delta^{(2)}(k)=\sum\limits_{i\neq k}^{80}\left\{\left[a_2(k)+b_2(k)\ln L_{\mathrm{eff}}^{\alpha}(i)\right]-\ln\left[K_{\mathrm{f}}(i)\right]\right\}^2 \qquad (k=1,\ 2,\ \cdots,\ 80) \\ \Delta^{(3)}(k)=\sum\limits_{i\neq k}^{80}\left\{\left[a_3(k)+b_3(k)\ln\Phi^{\beta}(i)\right]-\ln\left[K_{\mathrm{f}}(i)\right]\right\}^2 \end{cases}$$

(23)

$$\begin{cases} \dfrac{\partial\Delta^{(i)}(k)}{\partial a_i(k)}=0 \\ \dfrac{\partial\Delta^{(i)}(k)}{\partial b_i(k)}=0 \end{cases} \qquad (i=1,2,3;\quad k=1,\ 2,\ \cdots,\ 80) \tag{24}$$

The results thus obtained for $\left[a_1(k),\ b_1(k)\right]$, $\left[a_2(k),\ b_2(k)\right]$, and $\left[a_3(k),\ b_3(k)\right]$ are given in Appendix **C**.

All the above three formulae (Eqs. 19.1 – 19.3) can be used to predict the protein folding rates but they each reflect only one of the three features described above. To incorporate all these features into one predictor, let us consider the following equation:

$$\ln K_{\mathrm{f}}=\sum_{i=1}^{3}w_i\ln K_{\mathrm{f}}^{(i)} \tag{25}$$

where w_i is the weight that reflects the impact of the i-th formula on the protein folding rate. If the impacts of the three formulae were the same, we should have $w_i=1/3$ $(i=1,2,3)$. Since they are actually not the same, it would be rational to introduce some sort of statistical criterion to reflect their different impacts, as formulated below.

Given a system containing N statistical samples, we can define a cosine function as formulated by [34, 35]:

$$\Theta = \sum_{i=1}^{N} x_i y_i \bigg/ \left[\left(\sum_{i=1}^{N} x_i^2 \right) \left(\sum_{i=1}^{N} y_i^2 \right) \right]^{1/2} \tag{26}$$

where x_i and y_i are, respectively, the observed and predicted results for the i-th sample. Obviously, the cosine function is within the range of -1 and 1 [36]. When and only when all the predicted results are exactly the same as the observed ones, we have $\Theta = 1$. Suppose the value of the cosine function yielded with the i-th predictor in Eq. 19 on the benchmark dataset $\mathbb{S}_{\text{bench}}$ by the self-consistency test [37] is $\Theta\left(\ln K_{\text{f}}^{(i)}\right)$, which turned out to be

$$\Theta\left(\ln K_{\text{f}}^{(1)}\right) = 0.8938, \qquad \Theta\left(\ln K_{\text{f}}^{(2)}\right) = 0.9276, \qquad \Theta\left(\ln K_{\text{f}}^{(3)}\right) = 0.7145 \tag{27}$$

Then the weight w_i in Eq. 25 can be formulated as:

$$w_i = \frac{\Theta\left(\ln K_{\text{f}}^{(i)}\right)}{\sum_{j=1}^{3} \Theta\left(\ln K_{\text{f}}^{(i)}\right)} \qquad (i = 1,\ 2,\ 3) \tag{28}$$

which yields

$$w_1 = 0.3525, \qquad w_2 = 0.3658, \qquad w_3 = 0.2817 \tag{29}$$

Substituting Eq. 29 as well as Eqs. 19 and 22 into Eq. 25, we finally obtain

$$\ln K_{\text{f}} = 29.8470 - 2.2587\ln(L) - 2.0472\ln(L_{\text{eff}}^{\alpha}) - 16.3417\Phi^{\beta} \tag{30}$$

However, when the accuracy of Eq. 25 is examined by the jackknife cross-validation, by following the similar procedures in treating Eq. 19, we should instead have

$$\ln K_{\text{f}}(k) = A(k) + B(k)\ln(L) + C(k)\ln(L_{\text{eff}}^{\alpha}) + D(k)\Phi^{\beta} \tag{31}$$

where the values for $A(k)$, $B(k)$, $C(k)$, and $D(k)$ $(k = 1,\ 2,\ \cdots,\ 80)$ are given in Appendix **D**.

The ensemble predictor formed by fusing the three individual predictors of Eq. 19 as formulated by Eq. 25 or Eq. 30 or Eq. 31 is called the **FoldRate**, which can yield much better prediction quality than the individual predictors as shown below.

IV. RESULTS AND DICSUSSIONS

In statistics the independent test, sub-sampling test, and jackknife test are the three cross-validation methods often used to examine the quality of a predictor [38]. To demonstrate the quality of **FoldRate**, we adopted the jackknife cross-validation on the benchmark dataset $\mathbb{S}_{\text{bench}}$ (see the Appendix **A**). During the jackknife cross-validation, each of protein samples in the benchmark dataset is in turn singled out as a tested protein and the predictor is trained by the remaining proteins. Compared with the other two cross-validation test methods, the jackknife test is deemed more objective that can always yield a unique result for a given benchmark dataset [37, 39], and hence has been increasingly used by investigators to examine the accuracy of various predictors (see, e.g., [40-54]).

In the current study, two kinds of scales are used to measure the prediction quality. One is the Pearson correlation coefficient (PCC) (see wikipedia.org/wiki/Correlation) and the other is the root mean square deviation (RMSD). They are respectively formulated as follows:

$$\text{PCC} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{N}(x_i - \bar{x})^2\right]\left[\sum_{i=1}^{N}(y_i - \bar{y})^2\right]}} \tag{32}$$

$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N}} \tag{33}$$

where x_i, y_i and N have the same meanings as Eq. 26, while $\bar{x}$ and $\bar{y}$ the corresponding mean values for the N samples. The meaning of RMSD is obvious; i.e., the smaller the value of RMSD, the more accurate the prediction. PCC is usually used to reflect the correlation of the predicted results with the observed ones: the closer the value of PCC is to 1, the better the correlation is. When all the predicted results are exactly the same as the observed ones, we have PCC=1 and RMSD=0.

Listed in Table **3** are the PCC and RMSD results obtained by the ensemble predictor **FoldRate** on the benchmark dataset $\mathbb{S}_{\text{bench}}$ *via* the jackknife cross-validation. For facilitating comparison, the corresponding results obtained by individual predictors are given in Table **3** as well.

As we can see from Table **3**, the overall PCC value yielded by the ensemble predictor of Eq. 25 is 0.88, which is the closest to 1 in comparison with those by the individual predictors in Eq. 19. Such an overall PCC value is even higher than 0.86 obtained for the same benchmark dataset by the method in which, however, the 3D structural information is needed [12]. Although the method developed recently by Ouyang and Liang could also be used to predict the protein folding rate without using the 3D structural information, the overall PCC value thus obtained would drop to 0.82 [12].

Moreover, it can be seen from Table **3** that the overall RMSD value for the ensemble predictor is the lowest one in comparison with those by the individual predictors. The highest correlation and lowest deviation results indicate that the **FoldRate** ensemble predictor formed by fusing individual predictors is indeed a quite promising approach.

V. CONCLUSIONS

FoldRate is developed for predicting protein folding rate. It is an ensemble predictor formed by fusing three individual predictors with each based on the size of a protein, its α-helix effect, and its β-sheet effect, respectively. Given a protein, all these effects can be derived from its sequence.

Table 3. Comparison of the Jackknife Cross-Validation Tested Results by Using Different Predictors on the Benchmark Dataset $\mathbb{S}_{\text{bench}}$

Predictor	Overall PCC (cf. Eq. 32)	Overall RMSD (cf. Eq. 33)
$\ln\left(K_\mathrm{f}^{(1)}\right)$ (cf. Eq. 19.1)	0.79	2.67
$\ln\left(K_\mathrm{f}^{(2)}\right)$ (cf. Eq. 19.2)	0.85	2.23
$\ln\left(K_\mathrm{f}^{(3)}\right)$ (cf. Eq. 19.3)	0.27	4.17
$\ln\left(K_\mathrm{f}\right)$ (cf. Eq. 25)	**0.88**	**2.03**

Therefore, **FoldRate** can be used to predict the folding rate of a protein according to its sequence information alone. **FoldRate** is freely accessible to the public *via* the web-site at www.csbio.sjtu.edu.cn/bioinf/FoldingRate/.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 60704047), Science and Technology Commission of Shanghai Municipality (Grant No. 08ZR1410600, 08JC1410600) and sponsored by Shanghai Pujiang Program.

APPENDIX A

The benchmark dataset $\mathbb{S}_{\text{bench}}$ consists of 80 proteins. The PDB codes listed below are just for the role of identity. In this study, only the protein sequences and their $\ln(K_\mathrm{f})$ values are used for developing the current predictor. See the text for further explanation.

1. PDB: 1APS, $\ln(K_\mathrm{f})$=-1.47

TARPLKSVDYEVFGRVQGVCFRMYAEDEARKIGVVGWVKNTSKGTVTGQVQGPEEKVNSM
KSWLSKVGSPSSRIDRTNFSNEKTISKLEYSNFSVRY

2. PDB: 1BA5, $\ln(K_\mathrm{f})$=5.91

KRQAWLWEEDKNLRSGVRKYGEGNWSKILLHYKFNNRTSVMLKDRWRTMKKL

3. PDB: 1BDD, $\ln(K_\mathrm{f})$=11.69

ADNKFNKEQQNAFYEILHLPNLNEEQRNGFIQSLKDDPSQSANLLAEAKKLNDAQAPKA

4. PDB: 1C8C, $\ln(K_\mathrm{f})$=6.95

ATVKFKYKGEEKQVDISKIKKVWRVGKMISFTYDEGGGKTGRGAVSEKDAPKELLQMLAKQ
KK

5. PDB: 1C9O, $\ln(K_\mathrm{f})$=7.20

QRGKVKWFNNEKGYGFIEVEGGSDVFVHFTAIQGEGFKTLEEGQEVSFEIVQGNRGPQAA
NVVKL

6. PDB: 1CSP, $\ln(K_\mathrm{f})$=6.54

LEGKVKWFNSEKGFGFIEVEGQDDVFVHFSAIQGEGFKTLEEGQAVSFEIVEGNRGPQAA
NVTKEA

7. PDB: 1DIV_c, $\ln(K_\mathrm{f})$=0.0

AAEELANAKKLKEQLEKLTVTIPAKAGEGGRLFGSITSKQIAESLQAQHGLKLDKRKIEL
ADAIRALGYTNVPVKLHPEVTATLKVHVTEQK

8. PDB: 1DIV_n, $\ln(K_\mathrm{f})$=6.61

KVIFLKDVKGKGKKGEIKNVADGYANNFLFKQGLAIEATPANLKALEAQKQKEQR

9. PDB: 1E0L, $\ln(K_\mathrm{f})$=10.37

ATAVSEWTEYKTADGKTYYYNNRTLESTWEKPQELK

10. PDB: 1E0M, $\ln(K_\mathrm{f})$=8.85

MGLPPGWDEYKTHNGKTYYYNHNTKTSTWTDPRMSS

11. PDB: 1ENH, $\ln(K_\mathrm{f})$=10.53

PRTAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEAQIKIWFQNKRAKI

12. PDB: 1FEX, $\ln(K_\mathrm{f})$=8.19

RIAFTDADDVAILTYVKENARSPSSVTGNALWKAMEKSSLTQHSWQSLKDRYLKHLRG

13. PDB: 1FKB, $\ln(K_\mathrm{f})$=1.45

VQVETISPGDGRTFPKRGQTCVVHYTGMLEDGKKFDSSRDRNKPFKFMLGKQEVIRGWEE
GVAQMSVGQRAKLTISPDYAYGATGHPGIIPPHATLVFDVELLKLE

14. PDB: 1FMK, $\ln(K_\mathrm{f})$=4.05

TFVALYDYESRTETDLSFKKGERLQIVNNTEGDWWLAHSLSTGQTGYIPSNYVAPS

15. PDB: 1FNF_9, $\ln(K_\mathrm{f})$=-0.92

DSPTGIDFSDITANSFTVHWIAPRATITGYRIRHHPEHFSGRPREDRVPHSRNSITLTNL
TPGTEYVVSIVALNGREESPLLIGQQSTV

16. PDB: 1G6P, $\ln(K_\mathrm{f})$=6.30

RGKVKWFDSKKGYGFITKDEGGDVFVHWSAIEMEGFKTLKEGQVVEFEIQEGKKGPQAAH
VKVVE

17. PDB: 1HDN, $\ln(K_\mathrm{f})$=2.69

FQQEVTITAPNGLHTRPAAQFVKEAKGFTSEITVTSNGKSASAKSLFKLQTLGLTQGTVV
TISAEGEDEQKAVEHLVKLMAELE

18. PDB: 1IDY, $\ln(K_\mathrm{f})$=8.73

EVKKTSWTEEEDRILYQAHKRLGNRWAEIAKLLPGRTDNAIKNHWNSTMRRKV

19. PDB: 1IMQ, $\ln(K_\mathrm{f})$=7.28

ELKHSISDYTEAEFLQLVTTICNADTSSEEELVKLVTHFEEMTEHPSGSDLIYYPKEGDD
DSPSGIVNTVKQWRAANGKSGFKQG

20. PDB: 1K8M, $\ln(K_\mathrm{f})$=-0.71

GQVVQFKLSDIGEGIREVTVKEWYVKEGDTVSQFDSICEVQSDKASVTITSRYDGVIKKL
YYNLDDIAYVGKPLVDIETEALKDLE

21. PDB: 1K9Q, $\ln(K_\mathrm{f})$=8.37

EIPDDVPLPAGWEMAKTSSGQRYFLNHIDQTTTWQDPRK

22. PDB: 1L2Y, $\ln(K_\mathrm{f})$=12.40

LYIQWLKDGGPSSGRPPPS

23. PDB: 1LMB, $\ln(K_\mathrm{f})$=8.50

LTQEQLEDARRLKAIYEKKKNELGLSQESVADKMGMGQSGVGALFNGINALNAYNAALLA
KILKVSVEEFSPSIAREIYEMYEAVS

24. PDB: 1MJC, $\ln(K_\mathrm{f})$=5.23

GKMTGIVKWFNADKGFGFITPDDGSKDVFVHFSAIQNDGYKSLDEGQKVSFTIESGAKGP
AAGNVTSL

25. PDB: 1N88, $\ln(K_\mathrm{f})$=3.0

KTAYDVILAPVLSEKAYAGFAEGKYTFWVHPKATKTEIKNAVETAFKVKVVKVNTLHVRG
KKKRLGRYLGKRPDRKKAIVQVAPGQKIEALEGLI

26. PDB: 1NYF, $\ln(K_\mathrm{f})$=4.54

TLFVALYDYEARTEDDLSFHKGEKFQILNSSEGDWWEARSLTTGETGYIPSNYVAPV

27. PDB: 1PGB_b, $\ln(K_\mathrm{f})$=12.0

TYKLILNGKTLKGET

28. PDB: 1PIN, $\ln(K_\mathrm{f})$=9.37

LPPGWEKRMSRSSGRVYYFNHITNASQWERP

29. PDB: 1PKS, $\ln(K_\mathrm{f})$=-1.06

GYQYRALYDYKKEREEDIDLHLGDILTVNKGSLVALGFSDGQEARPEEIGWLNGYNETTG
ERGDFPGTYVEYIGR

30. PDB: 1PRB, $\ln(K_\mathrm{f})$=12.90

IDQWLLKNAKEDAIAELKKAGITSDFYFNAINKAKTVEEVNALKNEILKAHA

31. PDB: 1PSE, $\ln(K_\mathrm{f})$=1.17

IERGSKVKILRKESYWYGDVGTVASIDKSGIIYPVIVRFNKVNYNGFSGSAGGLNTNNFA
EHELEVVG

32. PDB: 1QTU, $\ln(K_f)$=-0.36

SMAGEDVGAPPDHLWVHQEGIYRDEYQRTWVAVVEEETSFLRARVQQIQVPLGDAARPSH
LLTSQLPLMWQLYPEERYMDNNSRLWQIQHHLMVRGVQELLLKLLPDDRSPGIH

33. PDB: 1RFA, $\ln(K_f)$=7.0

NTIRVFLPNKQRTVVNVRNGMSLHDCLMKALKVRGLQPECCAVFRLLHEHKGKKARLDWN
TDAASLIGEELQVDFLD

34. PDB: 1SHG, $\ln(K_f)$=2.10

ELVLALYDYQEKSPREVTMKKGDILTLLNSTNKDWWKVEVNDRQGFVPAAYVKKLD

35. PDB: 1TEN, $\ln(K_f)$=1.06

DAPSQIEVKDVTDTTALITWFKPLAEIDGIELTYGIKDVPGDRTTIDLTEDENQYSIGNL
KPDTEYEVSLISRRGDMSSNPAKETFTT

36. PDB: 1URN, $\ln(K_f)$=5.76

VPETRPNHTIYINNLNEKIKKDELKKSLHAIFSRFGQILDILVSRSLKMRGQAFVIFKEV
SSATNALRSMQGFPFYDKPMRIQYAKTDSDIIAKM

37. PDB: 1VII, $\ln(K_f)$=11.51

LSDEDFKAVFGMTRSAFANLPLWKQQNLKKEKGLF

38. PDB: 1WIT, $\ln(K_f)$=0.41

KPKILTASRKIKIKAGFTHNLEVDFIGAPDPTATWTVGDSGAALAPELLVDAKSSTTSIF
FPSAKRADSGNYKLKVKNELGEDEAIFEVIVQ

39. PDB: 2A3D, $\ln(K_f)$=12.7

GSWAEFKQRLAAIKTRLQALGGSEAELAAFEKEIAAFESELQAYKGKGNPEVEALRKEAA
AIRDELQAYRHN

40. PDB: 2ACY, $\ln(K_f)$=0.84

EGDTLISVDYEIFGKVQGVFFRKYTQAEGKKLGLVGWVQNTDQGTVQGQLQGPASKVRHM
QEWLETKGSPKSHIDRASFHNEKVIVKLDYTDFQIVK

41. PDB: 2AIT, $\ln(K_f)$=4.21

TTVSEPAPSCVTLYQSWRYSQADNGCAETVTVKVVYEDDTEGLCYAVAPGQITTVGDGYI
GSHGHARYLARCL

42. PDB: 2CI2, $\ln(K_f)$=3.87

LKTEWPELVGKSVEEAKKVILQDKPEAQIIVLPVGTIVTMEYRIDRVRLFVDKLDNIAEV
PRVG

43. PDB: 2HQI, $\ln(K_f)$=0.18

TQTVTLAVPGMTCAACPITVKKALSKVEGVSKVDVGFEKREAVVTFDDTKASVQKLTKAT
ADAGYPSSVKQ

44. PDB: 2PDD, $\ln(K_f)$=9.69

IAMPSVRKYAREKGVDIRLVQGTGKNGRVLKEDIDAFLAGGA

45. PDB: 2PTL, $\ln(K_f)$=4.10

VTIKANLIFANGSTQTAEFKGTFEKATSEAYAYADTLKKDNGEYTVDVADKGYTLNIKFAG

46. PDB: 2ABD, $\ln(K_f)$=6.48

QAEFDKAAEEVKHLKTKPADEEMLFIYSHYKQATVGDINTERPGMLDFKGKAKWDAWNEL
KGTSKEDAMKAYIDKVEELKKKYGI

47. PDB: 2CRO, $\ln(K_f)$=5.35

QTLSERLKKRRIALKMTQTELATKAGVKQQSIQLIEAGVTKRPRFLFEIAMALNCDPVWL
QYGT

48. PDB: 1UZC, $\ln(K_f)$=8.68

PAKKTYTWNTKEEAKQAFKELLKEKRVPSNASWEQAMKMIINDPRYSALAKLSEKKQAFN
AYKVQTEK

49. PDB: 1CEI, $\ln(K_f)$=5.8

KNSISDYTEAEFVQLLKEIEKENVAATDDVLDVLLEHFVKITEHPDGTDLIYYPSDNRDD
SPEGIVKEIKEWRAANGKPGFKQG

50. PDB: 1BRS, $\ln(K_f)$=3.37

INTFDGVADYLQTYHKLPDNYITKSEAQALGWVASKGNLADVAPGKSIGGDIFSNREGKL
PGKSGRTWREADINYTSGFRNSDRILYSS

51. PDB: 2A5E, $\ln(K_{\mathrm{f}})$=3.50

EPAAGSSMEPSADWLATAAARGRVEEVRALLEAGALPNAPNSYGRRPIQVMMMGSARVAE
LLLLHGAEPNCADPATLTRPVHDAAREGFLDTLVVLHRAGARLDVRDAWGRLPVDLAEEL
GHRDVARYLRAAAGGTRGSNHARIDAAEGPSDIPD

52. PDB: 1TIT, $\ln(K_{\mathrm{f}})$=3.6

IEVEKPLYGVEVFVGETAHFEIELSEPDVHGQWKLKGQPLTASPDCEIIEDGKKHILILH
NCQLGMTGEVSFQAANAKSAANLKVKEL

53. PDB: 1FNF_10, $\ln(K_{\mathrm{f}})$=5.48

DVPRDLEVVAATPTSLLISWDAPAVTVRYYRITYGETGGNSPVQEFTVPGSKSTATISGL
KPGVDYTITVYAVTGRGDSPASSKPISINYRT

54. PDB: 1HNG, $\ln(K_{\mathrm{f}})$=1.8

SGTVWGALGHGINLNIPNFQMTDDIDEVRWERGSTLVAEFKRKMKPFLKSGAFEILANGD
LKIKNLTRDDSGTYNVTVYSTNGTRILNKALDLRI

55. PDB: 1ADW, $\ln(K_{\mathrm{f}})$=0.64

THEVHMLNKGESGAMVFEPAFVRAEPGDVINFVPTDKSHNVEAIKEILPEGVESFKSKIN
ESYTLTVTEPGLYGVKCTPHFGMGMVGLVQVGDAPENLDAAKTAKMPKKARERMDAELAQ
VN

56. PDB: 1EAL, $\ln(K_{\mathrm{f}})$=1.3

FTGKYEIESEKNYDEFMKRLALPSDAIDKARNLKIISEVKQDGQNFTWSQQYPGGHSITN
TFTIGKECDIETIGGKKFKATVQMEGGKVVVNSPNYHHTAEIVDGKLVEVSTVGGVSYER
VSKKLA

57. PDB: 1IFC, $\ln(K_{\mathrm{f}})$=3.4

FDGTWKVDRNENYEKFMEKMGINVVKRKLGAHDNLKLTITQEGNKFTVKESSNFRNIDVV
FELGVDFAYSLADGTELTGTWTMEGNKLVGKFKRVDNGKELIAVREISGNELIQTYTYEG
VEAKRIFKKE

58. PDB: 1OPA, $\ln(K_{\mathrm{f}})$=1.4

KDQNGTWEMESNENFEGYMKALDIDFATRKIAVRLTQTKIIVQDGDNFKTKTNSTFRNYD
LDFTVGVEFDEHTKGLDGRNVKTLVTWEGNTLVCVQKGEKENRGWKQWVEGDKLYLELTC
GDQVCRQVFKKK

59. PDB: 1HCD, $\ln(K_{\mathrm{f}})$=1.1

GNRAFKSHHGHFLSAEGEAVKTHHGHHDHHTHFHVENHGGKVALKTHCGKYLSIGDHKQV
YLSHHLHGDHSLFHLEHHGGKVSIKGHHHHYISADHHGHVSTKEHHDHDTTFEEIII

60. PDB: 1BEB, $\ln(K_{\mathrm{f}})$=-2.20

TMKGLDIQKVAGTWYSLAMAASDISLLDAQSAPLRVYVEELKPTPEGDLEILLQKWENGE
CAQKKIIAEKTKIPAVFKIDALNENKVLVLDTDYKKYLLFCMENSAEPEQSLVCQCLVRT
PEVDDEALEKFDKALKALPMHIRLSFNPTQLEEQC

61. PDB: 1B9C, $\ln(K_{\mathrm{f}})$=-2.76

EELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTTF
VQCFSRYPDHMKQHDFFKSAMPEGYVQERTISFKDDGNYKTRAEVKFEGDTLVNRIELKG
IDFKEDGNILGHKLEYNYNSHNVYITADKQKNGIKANFKIRHNIEDGSVQLADHYQQNTP
IGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGIT

62. PDB: 1I1B, $\ln(K_{\mathrm{f}})$=-4.01

RSLNCTLRDSQQKSLVMSGPYELKALHLQGQDMEQQVVFSMSFVQGEESNDKIPVALGLK
EEKNLYLSCVLKDDKPTLQLESVDPKNYPKKKMEKRFVFNKIEINNKLEFESAQFPNWYI
STSQAENMPVFLGGTKGGQDITDFTMQFVSS

63. PDB: 1PGB_ab, $\ln(K_{\mathrm{f}})$=6.40

TYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE

64. PDB: 1UBQ, $\ln(K_{\mathrm{f}})$=5.90

QIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNI
QKESTLHLVLRLRGG

65. PDB: 1GXT, $\ln(K_{\mathrm{f}})$=4.39

TSCCGVQLRIRGKVQGVGFRPFVWQLAQQLNLHGDVCNDGDGVEVRLREDPETFLVQLYQ
HCPPLARIDSVEREPFIWSQLPTEFTIR

66. PDB: 1SCE, $\ln(K_f)$=4.17

PRLLTASERERLEPFIDQIHYSPRYADDEYEYRHVMLPKAMLKAIPTDYFNPETGTLRIL
QEEEWRGLGITQSLGWEMYEVHVPEPHILLFKREKD

67. PDB: 1HMK, $\ln(K_f)$=2.79

EQLTKCEVFQKLKDLKDYGGVSLPEWVCTAFHTSGYDTQAIVQNNDSTEYGLFQINNKIW
CKDDQNPHSRNICNISCDKFLDDDLTDDIVCAKKILDKVGINYWLAHKALCSEKLDQWLC

68. PDB: 3CHY, $\ln(K_f)$=1.0

DADKELKFLVVDDFSTMRRIVRNLLKELGFNNVEEAEDGVDALNKLQAGGYGFVISDWNM
PNMDGLELLKTIRADGAMSALPVLMVTAEAKKENIIAAAQAGASGYVVKPFTAATLEEKL
NKIFEKLGM

69. PDB: 1HEL, $\ln(K_f)$=1.25

VFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSR
WWCNDGRTPGSRNLCNIPCSALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQ
AWIRGCRL

70. PDB: 1DK7, $\ln(K_f)$=0.83

GMQFDRGYLSPYFINKPETGAVELESPFILLADKKISNIREMLPVLEAVAKAGKPLLIIA
EDVEGEALATLVVNTMRGIVKVAAVKAPGFGDRRKAMLQDIATLTGGTVISEEIGMELEK
ATLEDLGQAKRVVINKDTTTIIDGV

71. PDB: 1JOO, $\ln(K_f)$=0.30

TSTKKLHKEPATLIKAIDGDTVKLMYKGQPMTFRLLLVDTPETKHPKKGVEKYGPEASAF
TKKMVENAKKIEVEFDKGQRTDKYGRGLAYIYADGKMVNEALVRQGLAKVAYVYKPNNTH
EQLLRKSEAQAKKEKLNIWSEDNADSGQ

72. PDB: 2RN2, $\ln(K_f)$=1.41

LKQVEIFTDGSCLGNPGPGGYGAILRYRGREKTFSAGYTRTTNNRMELMAAIVALEALKE
HCEVILSTDSQYVRQGITQWIHNWKKRGWKTADKKPVKNVDLWQRLDAALGQHQIKWEWV
KGHAGHPENERCDELARAAAMNPTLEDTGYQVEV

73. PDB: 1RA9, $\ln(K_f)$=-2.46

ISLIAALAVDRVIGMENAMPWNLPADLAWFKRNTLDKPVIMGRHTWESIGRPLPGRKNII
LSSQPGTDDRVTWVKSVDEAIAACGDVPEIMVIGGGRVYEQFLPKAQKLYLTHIDAEVEG
DTHFPDYEPDDWESVFSEFHDADAQNSHSYCFEILERR

74. PDB: 1PHP_c, $\ln(K_f)$=-3.44

VLGKALSNPDRPFTAIIGGAKVKDKIGVIDNLLEKVDNLIIGGGLAYTFVKALGHDVGKS
LLEEDKIELAKSFMEKAKEKGVRFYMPVDVVVADRFANDANTKVVPIDAIPADWSALDIG
PKTRELYRDVIRESKLVVWNGPMGVFEMDAFAHGTKAIAEALAEALDTYSVIGGGDSAAA
VEKFGLADKMDHISTGGGASLEFMEGKQLPGVVALEDK

75. PDB: 1PHP_n, $\ln(K_f)$=2.30

NKKTIRDVDVRGKRVFCRVDFNVPMEQGAITDDTRIRAALPTIRYLIEHGAKVILASHLG
RPKGKVVEELRLDAVAKRLGELLERPVAKTNEAVGDEVKAAVDRLNEGDVLLLENVRFYP
GEEKNDPELAKAFAELADLYVNDAFGAAHRAHASTEGIAHYLPAVAGFLMEKEL

76. PDB: 2BLM, $\ln(K_f)$=-1.24

DFAKLEEQFDAKLGIFALDTGTNRTVAYRPDERFAFASTIKALTVGVLLQQKSIEDLNQR
ITYTRDDLVNYNPITEKHVDTGMTLKELADASLRYSDNAAQNLILKQIGGPESLKKELRK
IGDEVTNPERFEPELNEVNPGETQDTSTARALVTSLRAFALEDKLPSEKRELLIDWMKRN
TTGDALIRAGVPDGWEVADKTGAASYGTRNDIAIIWPPKGDPVVLAVLSSRDKKDAKYDD
KLIAEATKVVMKALNMNGK

77. PDB: 1QOP_a, $\ln(K_f)$=-2.5

ERYENLFAQLNDRREGAFVPFVTLGDPGIEQSLKIIDTLIDAGADALELGVPFSDPLADG
PTIQNANLRAFAAGVTPAQCFEMLAIIREKHPTIPIGLLMYANLVFNNGIDAFYARCEQV
GVDSVLVADVPVEESAPFRQAALRHNIAPIFICPPPNNAADDDLLRQVASYGRGYTYLLS
RSGVTGAENRGPLHHLIEKLKEYHAAPALQGFGISSPEQVSAAVRAGAAGAISGSAIVKI
IEKNLASPKQMLAELRSFVSAMKAASR

78. PDB: 1QOP_b, $\ln(K_f)$=-6.9

TLLNPYFGEEFGGMYVPQILMPALNQLEEAFVSAQKDPEFQAQFADLLKNYAGRPTALTK
CQNITAGTRTTLYLKREDLLHGGAHKTNQVLGQALLAKRMGKSEIIAETGAGQHGVASAL
ASALLGLKCRIYMGAKDVERQSPNVFRMRLMGAEVIPVHSGSATLKDACNEALRDWSGSY
ETAHYMLGTAAGPHPYPTIVREFQRMIGEETKAQILDKEGRLPDAVIACVGGGSNAIGMF

ADFINDTSVGLIGVEPGGHGIETGEHGAPLKHGRVGIYFGMKAPMMQTADGQIEESYSIS
AGLDFPSVGPQHAYLNSIGRADYVSITDDEALEAFKTLCRHEGIIPALESSHALAHALKM
MREQPEEKEQLLVVNLSGRGDKDIFTVHDIL

79. PDB: 1BTA, $\ln(K_\text{f})$=1.11

KAVINGEQIRSISDLHQTLKKELALPEYYGENLDALWDCLTGWVEYPLVLEWRQFEQSKQ
LTENGAESVLQVFREAKAEGCDITIILS

80. PDB: 1L63, $\ln(K_\text{f})$=4.10

NIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAKSELDKAIGRNTNGVITKD
EAEKLFNQDVDAAVRGILRNAKLKPVYDSLDAVRRAALINMVFQMGETGVAGFTNSLRML
QQKRWDEAAVNLAKSRWYNQTPNRAKRVITTFRTGTWDAYK

APPENDIX B.

The values of the three special features derived from the 80 protein sequences in the benchmark dataset $\mathbb{S}_\text{bench}$ of Appendix **A.** See the text for further explanation.

PDB Code	$\ln(L)$ **(cf. Eq. 16)**	$\ln\left(L_\text{eff}^\alpha\right)$ **(cf. Eq. 15)**	Φ^β **(Eq. 17)**
1APS	4.6052	4.3438	0.4810
1BA5	3.9703	3.0445	0.4683
1BDD	4.1109	3.3673	0.3994
1C8C	4.1589	3.8067	0.4415
1C9O	4.1897	4.1744	0.4798
1CSP	4.2047	4.1897	0.4482
1DIV_c	4.5326	4.1744	0.4517
1DIV_n	4.0254	3.5553	0.4450
1E0L	3.6109	3.5835	0.4426
1E0M	3.6109	3.5835	0.4386
1ENH	3.9890	3.1781	0.4490
1FEX	4.0775	2.9444	0.4645
1FKB	4.6728	4.5747	0.4654
1FMK	4.0604	3.9890	0.4816
1FNF_9	4.4998	4.4886	0.4807
1G6P	4.1897	4.1744	0.4561
1HDN	4.4543	3.9703	0.4669
1IDY	3.9890	2.9957	0.4455
1IMQ	4.4543	3.6889	0.4321
1K8M	4.4659	4.4543	0.5021
1K9Q	3.6889	3.6636	0.4363
1L2Y	2.9957	2.9444	0.3965
1LMB	4.4659	2.8904	0.4519
1MJC	4.2341	4.2195	0.4580
1N88	4.5643	4.2767	0.4909
1NYF	4.0604	4.0073	0.4625
1PGB_b	2.7726	2.7081	0.4997
1PIN	3.5264	3.4340	0.4453
1PKS	4.3438	4.2905	0.4466
1PRB	3.9703	2.5649	0.4543
1PSE	4.2485	4.1744	0.5045
1QTU	4.7449	4.5747	0.4757
1RFA	4.3567	4.1744	0.4879
1SHG	4.0431	3.9890	0.4835
1TEN	4.4886	4.4773	0.4507
1URN	4.5643	4.2905	0.4874
1VII	3.5835	2.3026	0.4479
1WIT	4.5326	4.5109	0.4537
2A3D	4.2905	2.7081	0.4005
2ACY	4.5850	4.3438	0.4955

PDB Code	$\ln(L)$ (cf. Eq. 16)	$\ln\left(L_{\text{eff}}^{\alpha}\right)$ (cf. Eq. 15)	Φ^{β} (Eq. 17)
2AIT	4.3041	4.1744	0.5049
2CI2	4.1744	3.9703	0.5081
2HQI	4.2767	3.9318	0.4906
2PDD	3.7612	3.1781	0.4660
2PTL	4.1431	3.8067	0.4727
2ABD	4.4659	3.5264	0.4093
2CRO	4.1744	3.2581	0.5014
1UZC	4.2341	3.2581	0.4213
1CEI	4.4427	3.6889	0.4309
1BRS	4.4998	4.2195	0.4549
2A5E	5.0499	4.5109	0.4155
1TIT	4.4886	4.4543	0.4532
1FNF_1	4.5326	4.5218	0.4975
1HNG	4.5643	4.4427	0.4846
1ADW	4.8122	4.5951	0.4414
1EAL	4.8442	4.7185	0.4708
1IFC	4.8828	4.7536	0.4741
1OPA	4.8903	4.7707	0.4794
1HCD	4.7791	4.7362	0.4397
1BEB	5.0562	4.9053	0.4573
1B9C	5.4116	5.3083	0.4706
1I1B	5.0239	5.0039	0.4607
1PGB_a	4.0254	3.6636	0.4692
1UBQ	4.3307	4.0943	0.4776
1GXT	4.4998	4.2485	0.5071
1SCE	4.5747	4.2627	0.4451
1HMK	4.8122	4.4067	0.4880
3CHY	4.8675	4.3307	0.4590
1HEL	4.8598	4.4188	0.4828
1DK7	4.9836	4.6347	0.4800
1JOO	5.0039	4.7185	0.4387
2RN2	5.0434	4.6250	0.4637
1RA9	5.0689	4.8598	0.4617
1PHP_c	5.3891	4.9053	0.4588
1PHP_n	5.1648	4.6821	0.4529
2BLM	5.5607	5.0876	0.4491
1QOP_a	5.5910	4.9488	0.4657
1QOP_b	5.9713	5.4848	0.4552
1BTA	4.4998	3.4657	0.4785
1L63	5.0876	4.3694	0.4831

APPENDIX C

The values of $\left[a_1(k),\, b_1(k)\right]$, $\left[a_2(k),\, b_2(k)\right]$, and $\left[a_3(k),\, b_3(k)\right]$ determined according to Eqs. 23-24 by excluding (jackknifing) the k-th protein sample in term from $\mathbb{S}_{\text{bench}}$ of Appendix **A**. See the text for further explanation.

k	PDB Code	$a_1(k)$	$b_1(k)$	$a_2(k)$	$b_2(k)$	$a_3(k)$	$b_3(k)$
1	1APS	32.346	-6.378	26.619	-5.567	30.032	-56.397
2	1BA5	32.536	-6.430	27.196	-5.709	30.824	-58.293
3	1BDD	31.978	-6.324	26.324	-5.519	28.039	-52.327
4	1C8C	32.340	-6.393	26.623	-5.585	30.341	-57.233

k	PDB Code	$a_1(k)$	$b_1(k)$	$a_2(k)$	$b_2(k)$	$a_3(k)$	$b_3(k)$
5	1C9O	32.318	-6.389	26.687	-5.608	31.371	-59.530
6	1CSP	32.357	-6.396	26.694	-5.608	30.460	-57.491
7	1DIV_c	32.411	-6.396	26.693	-5.587	31.231	-58.983
8	1DIV_n	32.423	-6.408	26.704	-5.599	30.426	-57.413
9	1E0L	32.225	-6.367	26.431	-5.545	29.669	-55.877
10	1E0M	32.500	-6.424	26.537	-5.566	29.881	-56.286
11	1ENH	32.042	-6.333	26.498	-5.554	29.920	-56.433
12	1FEX	32.259	-6.377	26.990	-5.664	30.754	-58.197
13	1FKB	32.390	-6.398	26.707	-5.602	30.689	-57.873
14	1FMK	32.631	-6.448	26.698	-5.597	30.935	-58.503
15	1FNF_9	32.436	-6.398	26.602	-5.567	30.128	-56.619
16	1G6P	32.375	-6.399	26.688	-5.605	30.579	-57.754
17	1HDN	32.436	-6.408	26.734	-5.602	30.696	-57.925
18	1IDY	32.228	-6.370	26.860	-5.634	30.067	-56.696
19	1IMQ	32.381	-6.408	26.620	-5.583	30.238	-57.006
20	1K8M	32.466	-6.405	26.613	-5.570	29.875	-56.111
21	1K9Q	32.490	-6.422	26.559	-5.571	29.960	-56.440
22	1L2Y	32.690	-6.465	26.367	-5.524	27.525	-51.232
23	1LMB	32.376	-6.411	27.003	-5.667	30.285	-57.170
24	1MJC	32.425	-6.408	26.700	-5.606	30.666	-57.916
25	1N88	32.420	-6.407	26.693	-5.598	30.916	-58.448
26	1NYF	32.588	-6.440	26.685	-5.596	30.716	-58.012
27	1PGB_b	33.447	-6.629	26.607	-5.578	34.256	-65.939
28	1PIN	32.513	-6.427	26.525	-5.562	29.953	-56.466
29	1PKS	32.611	-6.434	26.644	-5.573	31.663	-59.880
30	1PRB	31.788	-6.282	26.578	-5.571	29.976	-56.627
31	1PSE	32.631	-6.443	26.692	-5.590	30.612	-57.760
32	1QTU	32.316	-6.377	26.625	-5.576	30.334	-57.073
33	1RFA	32.344	-6.397	26.687	-5.607	31.767	-60.394
34	1SHG	32.830	-6.487	26.741	-5.602	30.617	-57.764
35	1TEN	32.435	-6.403	26.671	-5.590	31.151	-58.838
36	1URN	32.446	-6.421	26.729	-5.616	31.459	-59.693
37	1VII	32.038	-6.327	27.236	-5.723	29.724	-56.035
38	1WIT	32.412	-6.397	26.651	-5.584	31.095	-58.704
39	2A3D	32.081	-6.353	26.482	-5.549	27.409	-50.996
40	2ACY	32.392	-6.395	26.662	-5.585	30.381	-57.236
41	2AIT	32.448	-6.412	26.690	-5.599	31.824	-60.470
42	2CI2	32.542	-6.430	26.705	-5.598	31.851	-60.524
43	2HQI	32.647	-6.445	26.821	-5.614	30.196	-56.810
44	2PDD	32.217	-6.366	26.598	-5.576	30.849	-58.446
45	2PTL	32.551	-6.432	26.746	-5.606	30.790	-58.177
46	2ABD	32.395	-6.409	26.727	-5.604	30.992	-58.577
47	2CRO	32.443	-6.412	27.028	-5.670	32.064	-61.014
48	1UZC	32.236	-6.376	26.666	-5.591	29.746	-55.965
49	1CEI	32.395	-6.407	26.705	-5.599	30.701	-57.964
50	1BRS	32.422	-6.407	26.692	-5.598	30.803	-58.155
51	2A5E	32.783	-6.499	26.769	-5.622	32.188	-61.088
52	1TIT	32.422	-6.408	26.749	-5.617	30.806	-58.165
53	1FNF_1	32.428	-6.415	26.852	-5.649	31.889	-60.633
54	1HNG	32.409	-6.401	26.690	-5.596	30.565	-57.645
55	1ADW	32.367	-6.393	26.675	-5.592	31.662	-59.916
56	1EAL	32.416	-6.406	26.754	-5.615	30.586	-57.654
57	1IFC	32.584	-6.451	26.912	-5.661	30.741	-58.053

k	**PDB Code**	$a_1(k)$	$b_1(k)$	$a_2(k)$	$b_2(k)$	$a_3(k)$	$b_3(k)$
58	1OPA	32.445	-6.414	26.788	-5.625	30.503	-57.491
59	1HCD	32.386	-6.398	26.750	-5.614	31.653	-59.907
60	1BEB	32.185	-6.348	26.568	-5.562	31.108	-58.665
61	1B9C	32.330	-6.386	26.726	-5.606	30.331	-56.990
62	1I1B	31.997	-6.300	26.429	-5.524	30.951	-58.283
63	1PGB_a	32.443	-6.412	26.678	-5.594	30.870	-58.406
64	1UBQ	32.377	-6.401	26.669	-5.598	31.096	-58.896
65	1GXT	32.419	-6.409	26.702	-5.604	32.017	-60.895
66	1SCE	32.434	-6.413	26.703	-5.604	30.852	-58.267
67	1HMK	32.491	-6.427	26.710	-5.604	30.810	-58.208
68	3CHY	32.406	-6.403	26.666	-5.586	30.860	-58.221
69	1HEL	32.419	-6.407	26.671	-5.590	30.462	-57.404
70	1DK7	32.453	-6.416	26.695	-5.598	30.413	-57.281
71	1JOO	32.416	-6.406	26.692	-5.597	31.892	-60.399
72	2RN2	32.557	-6.442	26.722	-5.606	30.724	-57.946
73	1RA9	32.159	-6.342	26.535	-5.552	30.846	-58.099
74	1PHP_c	32.187	-6.351	26.463	-5.532	31.069	-58.551
75	1PHP_n	32.805	-6.502	26.796	-5.628	30.942	-58.423
76	2BLM	32.844	-6.509	26.749	-5.613	31.529	-59.590
77	1QOP_a	32.622	-6.455	26.555	-5.559	30.612	-57.599
78	1QOP_b	32.089	-6.330	26.230	-5.474	31.610	-59.620
79	1BTA	32.429	-6.402	27.207	-5.704	30.469	-57.408
80	1L63	32.905	-6.529	26.731	-5.612	30.978	-58.600

APPENDIX D

The values of $A(k)$, $B(k)$, $C(k)$, and $D(k)$ $(k=1,\ 2,\ \cdots,\ 80)$ determined according to Eqs. 31 by excluding (jackknifing) the k-th protein sample in term from $\mathbb{S}_{\text{bench}}$ of Appendix **A**. See the text for further explanation.

k	**PDB Code**	$A(k)$	$B(k)$	$C(k)$	$D(k)$
1	1APS	29.5992	-2.2482	-2.0364	-15.8870
2	1BA5	30.1004	-2.2666	-2.0884	-16.4211
3	1BDD	28.8002	-2.2292	-2.0189	-14.7405
4	1C8C	29.6856	-2.2535	-2.0430	-16.1225
5	1C9O	29.9914	-2.2521	-2.0514	-16.7696
6	1CSP	29.7511	-2.2546	-2.0514	-16.1952
7	1DIV_c	29.9869	-2.2546	-2.0437	-16.6155
8	1DIV_n	29.7684	-2.2588	-2.0481	-16.1732
9	1E0L	29.3855	-2.2444	-2.0284	-15.7406
10	1E0M	29.5810	-2.2645	-2.0360	-15.8558
11	1ENH	29.4162	-2.2324	-2.0317	-15.8972
12	1FEX	29.9076	-2.2479	-2.0719	-16.3941
13	1FKB	29.8320	-2.2553	-2.0492	-16.3028
14	1FMK	29.9829	-2.2729	-2.0474	-16.4803
15	1FNF_9	29.6518	-2.2553	-2.0364	-15.9496
16	1G6P	29.7888	-2.2556	-2.0503	-16.2693
17	1HDN	29.8601	-2.2588	-2.0492	-16.3175
18	1IDY	29.6556	-2.2454	-2.0609	-15.9713
19	1IMQ	29.6699	-2.2588	-2.0423	-16.0586
20	1K8M	29.5951	-2.2578	-2.0375	-15.8065
21	1K9Q	29.6077	-2.2638	-2.0379	-15.8991
22	1L2Y	28.9221	-2.2789	-2.0207	-14.4321
23	1LMB	29.8215	-2.2599	-2.0730	-16.1048
24	1MJC	29.8353	-2.2588	-2.0507	-16.3149

k	**PDB Code**	$A(k)$	$B(k)$	$C(k)$	$D(k)$
25	1N88	29.9014	-2.2585	-2.0477	-16.4648
26	1NYF	29.9013	-2.2701	-2.0470	-16.3420
27	1PGB_b	31.1728	-2.3367	-2.0404	-18.5750
28	1PIN	29.6014	-2.2655	-2.0346	-15.9065
29	1PKS	30.1612	-2.2680	-2.0386	-16.8682
30	1PRB	29.3717	-2.2144	-2.0379	-15.9518
31	1PSE	29.8898	-2.2712	-2.0448	-16.2710
32	1QTU	29.6759	-2.2479	-2.0397	-16.0775
33	1RFA	30.1121	-2.2549	-2.0510	-17.0130
34	1SHG	29.9792	-2.2867	-2.0492	-16.2721
35	1TEN	29.9648	-2.2571	-2.0448	-16.5747
36	1URN	30.0767	-2.2634	-2.0543	-16.8155
37	1VII	29.6296	-2.2303	-2.0935	-15.7851
38	1WIT	29.9336	-2.2549	-2.0426	-16.5369
39	2A3D	28.7168	-2.2394	-2.0298	-14.3656
40	2ACY	29.7295	-2.2542	-2.0430	-16.1234
41	2AIT	30.1659	-2.2602	-2.0481	-17.0344
42	2CI2	30.2122	-2.2666	-2.0477	-17.0496
43	2HQI	29.8254	-2.2719	-2.0536	-16.0034
44	2PDD	29.7762	-2.2440	-2.0397	-16.4642
45	2PTL	29.9315	-2.2673	-2.0507	-16.3885
46	2ABD	29.9264	-2.2592	-2.0499	-16.5011
47	2CRO	30.3554	-2.2602	-2.0741	-17.1876
48	1UZC	29.4971	-2.2475	-2.0452	-15.7653
49	1CEI	29.8364	-2.2585	-2.0481	-16.3285
50	1BRS	29.8699	-2.2585	-2.0477	-16.3823
51	2A5E	30.4155	-2.2909	-2.0565	-17.2085
52	1TIT	29.8916	-2.2588	-2.0547	-16.3851
53	1FNF_1	30.2365	-2.2613	-2.0664	-17.0803
54	1HNG	29.7975	-2.2564	-2.0470	-16.2386
55	1ADW	30.0863	-2.2535	-2.0456	-16.8783
56	1EAL	29.8293	-2.2581	-2.0540	-16.2411
57	1IFC	29.9900	-2.2740	-2.0708	-16.3535
58	1OPA	29.8286	-2.2609	-2.0576	-16.1952
59	1HCD	30.1179	-2.2553	-2.0536	-16.8758
60	1BEB	29.8269	-2.2377	-2.0346	-16.5259
61	1B9C	29.7169	-2.2511	-2.0507	-16.0541
62	1I1B	29.6656	-2.2207	-2.0207	-16.4183
63	1PGB_a	29.8910	-2.2602	-2.0463	-16.4530
64	1UBQ	29.9282	-2.2564	-2.0477	-16.5910
65	1GXT	30.2145	-2.2592	-2.0499	-17.1541
66	1SCE	29.8920	-2.2606	-2.0499	-16.4138
67	1HMK	29.9028	-2.2655	-2.0499	-16.3972
68	3CHY	29.8708	-2.2571	-2.0434	-16.4009
69	1HEL	29.7651	-2.2585	-2.0448	-16.1707
70	1DK7	29.7721	-2.2616	-2.0477	-16.1361
71	1JOO	30.1746	-2.2581	-2.0474	-17.0144
72	2RN2	29.9062	-2.2708	-2.0507	-16.3234
73	1RA9	29.7319	-2.2356	-2.0309	-16.3665
74	1PHP_c	29.7782	-2.2387	-2.0236	-16.4938
75	1PHP_n	30.0821	-2.2920	-2.0587	-16.4578
76	2BLM	30.2440	-2.2944	-2.0532	-16.7865
77	1QOP_a	29.8365	-2.2754	-2.0335	-16.2256
78	1QOP_b	29.8108	-2.2313	-2.0024	-16.7950
79	1BTA	29.9667	-2.2567	-2.0865	-16.1718
80	1L63	30.1037	-2.3015	-2.0529	-16.5076

REFERENCES

[1] C. B. Anfinsen, H. A. Scheraga, "Experimental and theoretical aspects of protein folding", *Adv. Protein. Chem.*, vol. 29, pp. 205-300, 1975.

[2] A. Aguzzi, "Unraveling prion strains with cell biology and organic chemistry", *Proc. Natl. Acad. Sci. USA,* vol. 105, pp. 11-12, 2008.

[3] C. M. Dobson, "The structural basis of protein folding and its links with human disease", *Philos. Trans. R Soc. Lond B. Biol. Sci.*, vol. 356, pp. 133-145, 2001.

[4] S. B. Prusiner, "Prions", *Proc. Natl. Acad. Sci. USA,* vol. 95, pp. 13363-13383, 1998.

[5] L. L. Qiu, S. A. Pabit, A. E. Roitberg, S. J. Hagen, "Smaller and faster: The 20-residue Trp-cage protein folds in 4 microseconds", *J. Chem. Soc.,* vol. 124, pp. 12952-12953, 2002.

[6] M. E. Goldberg, G. V. Semisotnov, B. Friguet, K. Kuwajima, O. B. Ptitsyn, S. Sugai, "An early immunoreactive folding intermediate of the tryptophan synthease beta 2 subunit is a 'molten globule", *FEBS. Lett.,* vol. 263, pp. 51-56, 1990.

[7] K. W. Plaxco, K. T. Simons, D. Baker, "Contact order, transition state placement and the refolding rates of single domain proteins", *J. Mol. Biol.,* vol. 277, pp. 985-994, 1998.

[8] D. N. Ivankov, S. O. Garbuzynskiy, E. Alm, K. W. Plaxco, D. Baker, A. V. Finkelstein, "Contact order revisited: influence of protein size on the folding rate", *Protein. Sci.*, vol. 12, pp. 2057-2062, 2003.

[9] H. Zhou, Y. Zhou, "Folding rate prediction using total contact distance", *Biophys. J.,* vol. 82, pp. 458-463, 2002.

[10] M. M. Gromiha, S. Selvaraj, "Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction", *J. Mol. Biol.,* vol. 310, pp. 27-32, 2001.

[11] B. Nolting, W. Schalike, P. Hampel, F. Grundig, S. Gantert, N. Sips, W. Bandlow, P. X. Qi, "Structural determinants of the rate of protein folding", *J. Theor. Biol.,* vol. 223, pp. 299-307, 2003.

[12] Z. Ouyang, J. Liang, "Predicting protein folding rates from geometric contact and amino acid sequence", *Protein Sci.,* vol. 17, pp. 1256-1263, 2008.

[13] D. N. Ivankov, A. V. Finkelstein, "Prediction of protein folding rates from the amino acid sequence-predicted secondary structure", *Proc. Natl. Acad. Sci. USA,* vol. 101, pp. 8942-8944, 2004.

[14] M. M. Gromiha, A. M. Thangakani, S. Selvaraj, "FOLD-RATE: prediction of protein folding rates from amino acid sequence", *Nucleic Acids Res.,* vol. 34, pp. W70-74, 2006.

[15] K. C. Chou, "Review: Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems", *Biophys. Chem.,* vol. 35, pp. 1-24, 1990.

[16] K. C. Chou, "Graphical rules in steady and non-steady enzyme kinetics", *J. Biol. Chem.,* vol. 264, pp. 12074-12079, 1989.

[17] S. X. Lin, K. E. Neet, "Demonstration of a slow conformational change in liver glucokinase by fluorescence spectroscopy", *J. Biol. Chem.*, vol. 265, pp. 9670-9675, 1990.

[18] W. H. Beyer CRC Handbook of Mathematical Science, 6th ed, Chapter 10, CRC Press, Inc.: Boca Raton, Florida, 1988; p. 544.

[19] K. C. Chou, S. P. Jiang, W. M. Liu, C. H. Fee, "Graph theory of enzyme kinetics: 1. Steady-state reaction system", *Sci. Sin.,* vol. 22, pp. 341-358, 1979.

[20] K. C. Chou, S. Forsen, "Graphical rules for enzyme-catalyzed rate laws", *Biochem. J.,* vol. 187, pp. 829-835, 1980.

[21] K. C. Chou, W. M. Liu, "Graphical rules for non-steady state enzyme kinetics", *J. Theor. Biol.*, vol. 91, pp. 637-654, 1981.

[22] G. P. Zhou, M. H. Deng, "An extension of Chou's graphical rules for deriving enzyme kinetic equations to system involving parallel reaction pathways", *Biochem. J.*, vol. 222, pp. 169-176, 1984.

[23] D. Myers, G. Palmer, "Microcomputer tools for steady-state enzyme kinetics", *Bioinformatics, (original: Comput. Appl. Biosci.),* vol. 1, pp. 105-110, 1985.

[24] K. C. Chou, F. J. Kezdy, F. Reusser, "Review: Steady-state inhibition kinetics of processive nucleic acid polymerases and nucleases", *Anal. Biochem.*, vol. 221, pp. 217-230, 1994.

[25] J. Andraos, "Kinetic plasticity and the determination of product ratios for kinetic schemes leading to multiple products without rate laws: new methods based on directed graphs", *Can. J. Chem.*, vol. 86, pp. 342-357, 2008.

[26] O. V. Galzitskaya, D. N. Ivankov, A. V. Finkelstein, "Folding nuclei in proteins", *FEBS Lett*., vol. 489, pp. 113-118, 2001.

[27] J. S. Richardson, "The anatomy and taxonomy of protein structure", *Adv. Protein Chem.,* vol. 34, pp. 167-339, 1981.

[28] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices", *J. Mol. Biol.*, vol. 292, pp. 195-202, 1999.

[29] K. C. Chou, G. Nemethy, M. S. Pottle, H. A. Scheraga, "The folding of the twisted ß-sheet in bovine pancreatic trypsin inhibitor", *Biochemistry,* vol. 24, pp. 7948-7953, 1985.

[30] K. C. Chou, G. Nemethy, M. Pottle, H. A. Scheraga, "Energy of stabilization of the right-handed beta-alpha-beta crossover in proteins", *J. Mol. Biol.,* vol. 205, pp. 241-249, 1989.

[31] P. Y. Chou, G. D. Fasman, "Prediction of secondary structure of proteins from amino acid sequences", *Adv. Enzymol. Relat. Subjects Biochem*., vol. 47, pp. 45-148, 1978.

[32] K. C. Chou, H. A. Scheraga, "Origin of the right-handed twist of beta-sheets of poly-L-valine chains", *Proc. Natl. Acad. Sci. USA,* vol. 79, pp. 7047-7051, 1982.

[33] K. C. Chou, "Using pair-coupled amino acid composition to predict protein secondary structure content", *J. Protein Chem.,* vol. 18, pp. 473-480, 1999.

[34] K. C. Chou, C. T. Zhang, "A correlation coefficient method to predicting protein structural classes from amino acid compositions", *Eur. J. Biochem.,* vol. 207, pp. 429-433, 1992.

[35] J. J. Chou, "Predicting cleavability of peptide sequences by HIV protease via correlation-angle approach", *J. Protein Chem.,* vol. 12, pp. 291-302, 1993.

[36] J. J. Chou, "A formulation for correlating properties of peptides and its application to predicting human immunodeficiency virus protease-cleavable sites in proteins", *Biopolymers*, vol. 33, pp. 1405-1414, 1993.

[37] K. C. Chou, H. B. Shen, "Review: recent progresses in protein subcellular location prediction", *Anal. Biochem.,* vol. 370, pp. 1-16, 2007.

[38] K. C. Chou, C. T. Zhang, "Review: prediction of protein structural classes", *Crit. Rev. Biochem. Mol. Biol.,* vol. 30, pp. 275-349, 1995.

[39] K. C. Chou, H. B. Shen, "Cell-PLoc: A package of web-servers for predicting subcellular localization of proteins in various organisms", *Nat. Protoc.,* vol. 3, pp. 153-162, 2008.

[40] X. B. Zhou, C. Chen, Z. C. Li, X. Y. Zou, "Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes", *J. Theor. Biol.,* vol. 248, pp. 546-551, 2007.

[41] Y. S. Ding, T. L. Zhang, "Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier", *Pattern Recognit. Lett.,* vol. 29, pp. 1887-1892, 2008.

[42] G. Y. Zhang, H. C. Li, B. S. Fang, "Predicting lipase types by improved Chou's pseudo-amino acid composition", *Protein Pept. Lett.,* vol. 15, pp. 1132-1137, 2008.

[43] H. Lin, "The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition", *J. Theor. Biol.,* vol. 252, pp. 350-356, 2008.

[44] F. M. Li, Q. Z. Li, "Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach", *Protein Pept. Lett.,* vol. 15, pp. 612-616, 2008.

[45] G. Y. Zhang, B. S. Fang, "Predicting the cofactors of oxidoreductases based on amino acid composition distribution and Chou's amphiphilic pseudo amino acid composition", *J. Theor. Biol.,* vol. 253, pp. 310-315, 2008.

[46] H. Lin, H. Ding, F. B. Feng-Biao Guo, A. Y. Zhang, J. Huang, "Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition", *Protein Pept. Lett.,* vol. 15, pp. 739-744, 2008.

[47] Y. S. Ding, T. L. Zhang, Q. Gu, P. Y. Zhao, K. C. Chou, "Using maximum entropy model to predict protein secondary structure with single sequence", *Protein Pept. Lett.,* vol. 16, pp. 552-560, 2009.

[48] H. Ding, L. Luo, H. Lin, "Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition", *Protein Pept. Lett.,* vol. 16, pp. 351-355, 2009.

[49] Z. H. Lin, H. L. Wang, B. Zhu, Y. Q. Wang, Y. Lin, Y. Z. Wu, "Estimation of affinity of HLA-A*0201 restricted CTL epitope based on the SCORE function", *Protein Pept. Lett.,* vol. 16, pp. 561-569, 2009.

[50] L. Nanni, A. Lumini, "A further step toward an optimal ensemble of classifiers for peptide classification, a case study: HIV protease", *Protein Pept. Lett.*, vol. 16, pp. 163-167, 2009.

[51] X. Shao, Y. Tian, L. Wu, Y. Wang, J. L. N. Deng, "Predicting-DNA-andRNA-binding proteins from sequences with kernel methods", *J. Theor. Biol.*, vol. 258, pp. 289-293, 2009.

[52] X. Xiao, P. Wang, K. C. Chou, "GPCR-CA: A cellular automaton image approach for predicting G-protein-coupled receptor functional classes", *J. Comput. Chem.*, vol. 30, pp. 1414-1423, 2009.

[53] J. Y. Yang, Z. L. Peng, Z. G. Yu, R. J. Zhang, V. Anh, D. Wang, "Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation", *J. Theor. Biol.*, vol. 257, pp. 618-626, 2009.

[54] X. Xiao, P. Wang, K. C. Chou, "Predicting protein quaternary structural attribute by hybridizing functional domain composition and pseudo amino acid composition", *J. Appl. Crystallogt.*, vol. 30, pp. 1414-1423, 2009.

Received: March 31, 2009 Revised: May 11, 2009 Accepted: May 12, 2009