Protein and Peptide Letters, Vol. 10, No. 5, pp. 435-439, 2003
Bentham Science Publishers Ltd.
0929-8665/03 $ 41.00 + .00

A SEQUENCE FUNCTION REVEALS NEW FEATURES IN β-PROTEIN FOLDING

Hui Shao and Zong-Hao Zeng*

Centre of Molecular Biology, Institute of Biophysics, Chinese Academy of Sciences, 15 Datun Road, Chaoyang District, Beijing 100101, China.

Abstract: When amino acid residues are represented by parameters describing their side chain lengths and polarities, a sequence function defined as the sum of the first two sequence autocorrelation functions is found to be negatively and linearly correlated with the logarithms of folding rates of β-proteins. The new function reveals new features in β-protein folding: larger residues slow down the folding while alternative distribution of polar-non-polar residues accelerates the folding.

INTRODUCTION

In a series of papers, it has been found that some simple quantities calculated from protein three-dimensional structures, such as the relative contact order [1], the long-range order [2], the total contact distance [3], the number of native contacts [4], and local secondary structure contents [5], can predict the folding rates of a set of single-domain fast folding proteins very well. Some of the quantities are evaluated only from the C_α atoms in the native structure (in fact, in our calculation, the first four quantities can all be evaluated from the C_α atoms only). In these analyses on correlations, the simple quantities reveal common features in protein folding. They proved that not all the atomic details are important in determine folding rates. This is in agreement with the common sense that for a system with many degrees of freedom some kind of regular phenomenological behaviors emerges at a certain place.

We have managed to construct functions of amino acid sequences that correlate with, and therefore can predict, folding rates. In a previous work, we proposed a simple quantity calculated from the amino acid sequence to predict the folding rates of α-helix proteins [6]. As there are only 5 data involved, this example is not very convincing. Here we report a new quantity, also calculated from amino acid sequences, which can predict the folding rates of 13 β-proteins.

DEFINITION OF SEQUENCE AUTOCORRELATION FUNCTION

First, a polar-length parameter is given to each residue as listed in Table 1. The sign of the polar-length parameter is the same with the polar parameter for each residue as in our previous paper [6], while its absolute value is the length of the residue's side chain. The length of a side chain is counted by the number of heavy-atom (non-hydrogen) covalent connections from C_α to the most distant heavy-atom. When there is more than one path connecting C_α to the most distant atom, the length of the shortest path is selected. For tryptophan, the most distant atom to C_α is $C_{\eta 2}$. For histidine, there are two atoms, $C_{\varepsilon 1}$ and $N_{\varepsilon 2}$, which are most distant to C_α. Zero is given to proline, as its side chain binds back to its main chain nitrogen.

Table 1. Polar-length parameters of amino acid residues.

A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	–6	–3	–3	2	–4	–4	0	–4	3	3	–5	4	5	0	–2	–2	6	6	2

Table 2. Values of *SP*, as well as that of G_0 and G_1 and the predicted logarithms of folding rates of β-proteins.

Sequence	Expr. lnk_f	G_0	G_1	*SP*	Pred. lnk_f
1pks (4-79) [7]	–1.05	1020	–45	975	–1.15
1srl (9-64) [8]	4.04	680	60	740	3.53
1nyf (84-141) [8]	4.54	718	–19	699	4.35
1mjc (2-70) [8]	5.24	697	–72	625	5.82
1csp (1-67) [9-11]	6.04	771	–67	704	4.25
2ait (1-74) [12]	4.20	744	–84	660	5.12
1shg (6-62) [8]	1.41	780	100	880	0.74
1wit (1-93) [13]	0.41	954	–98	856	1.22
1hng (2-98) [13]	2.89	1146	–237	909	0.16
1fnf_10 (1416-1509) [13]	5.48	888	–219	669	4.95
1tit (1-89) [13]	3.47	936	–234	702	4.29
1ten (803-891) [8]	1.06	947	–108	839	1.56
1fnf_9 (1326-1415) [14]	-0.91	961	–7	954	–0.74

A set of sequence autocorrelation functions, G_j, is calculated for each set of parameters as

$$G_j=\sum_i (g_i g_{i+j})$$

with g_i as the polar-length parameter of the i'th residue. The summation is taken over all the residues in a sequence. G_0 is the sum of squares of side chain lengths in a sequence. Clustering of polar, or non-polar, residues makes positive contributions to the values of G_1; whereas the alternative arrangement of polar and non-polar residues makes negative contributions to, and decreases its values.

The function, which is used to predict the folding rates of β-proteins, is defined as the sum of the zeroth and the first autocorrelation functions as

$$SP = G_0 + G_1.$$

RESULTS

The correlations of the autocorrelation functions G_j with the logarithms of folding rates, k_f, of 13 β-proteins (listed in Table 2) are systematically searched. Only *SP* is found to have significant correlations with β-proteins. The correlation coefficient (CC) of *SP* with $\ln k_f$ of the 13 β-proteins is –0.900. At the same time, it is found that *SP* is not correlated with the folding rates of α-proteins.

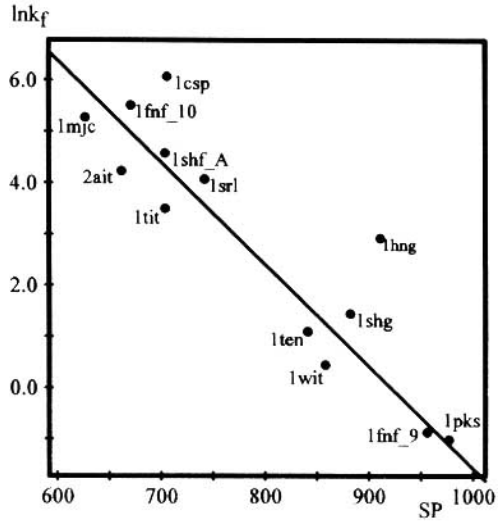

Figure 1. Relationship between *SP* and folding rates of β-proteins. The line shows the result of fitting for the 13 proteins in Table 2 with correlation coefficient –0.900.

When each G_j is used individually, the best three CC's are –0.598, –0.583 and 0.462 given out respectively by G_2, G_0 and G_8. The G_1 gives out a CC as less as –0.237. Therefore, when two terms work together, the combination of G_2 and G_0 might do the best. But, in fact, besides that given out by *SP*, the two best CC's are –0.74 and –0.70 given out by $G_2 – G_8$ and $G_0 + 1.5G_2$, respectively. It is noticeable that only G_0 with G_1 works cooperatively in such a way that they give out a CC (–0.900) better than the sum of CC's (–0.583 and –0.237) given out by individual G_0 and G_1, respectively.

DISCUSSIONS AND CONCLUSIONS

In the function *SP*, G_0 is the sum of squares of residue lengths, it is reasonable to regard that it is related to the sum of moments of inertial of each residue around the main chain. The more the number of larger residues is in a sequence, the slower the folding rate will be. The second term, G_1, will have more

negative value when more polar and non-polar residues in the sequence are alternatively distributed. This alternative distribution will accelerate the folding of β-proteins. A similar effect that the alternative distribution of polar and non-polar residues favors the formation of β-sheets has been noticed in the early days in investigating protein secondary structure prediction [15].

The residues cysteine and proline have special effects on protein folding by formation of non-native disulfide bonds or by cis-trans isomerization of Xaa-Pro peptide bonds, respectively. These effects were excluded from, or not observed in, the data listed in Table 2. Only the tendamistat (2ait) has two disulfide bonds among the 13 proteins, but they do not prevent efficient and rapid protein folding [12]. As only the typical two-state fast folding is the topic, the slow phase resulted from proline isomerization is not considered in experiment [13], or not used here, as well as in other works [1-6]. It is obvious that these special effects are the results of much more complex processes than those can be recoded in the simple polar-length parameters.

Not all factors affecting protein folding can be fully included in the simple function *SP*. Observation of Figure 1 reveals that the proteins are clustered into two groups at the left-upper and the right-lower corners, respectively. Inside each group, the distribution of data is somewhat random. As such, there must exist other functions independent of *SP* to taking account the larger deviations of 1hng and 1csp to the straight line. *SP* is rather good at explain why proteins with similar structures may have very different folding rates. There are 5 lg-like domains including 1fnf_9, 1fnf_10, 1tit, 1ten and 1wit and they all distributed very close to the straight line in Figure 1. The extreme example is the two domains, 9 and 10, of human fibronectin (1fnf). Their C_α atoms can be overlapped to each other with a root mean square deviation of 0.98 Å or 1.01 Å for backbone atoms (All data calculated from the 82 well aligned residues), while their folding rates have about a 600-fold difference. This is properly explained by their difference in amino acid sequence as evaluated by the values of G_0, G_1 and *SP* listed in Table 2.

The correlation between *SP* and folding rates cannot be extended to α and α/β proteins. The correlation of *SP* with folding rates is restricted to only one structural class. It is necessary to find a sequence function that can differentiate structural classes for constructing a unified function for predicting folding rates of proteins in all structural classes. We hope our efforts will speed up the searching for such functions.

ACKNOWLEDGEMENTS

Grant sponsor: National Natural Science Foundation of China; Grant number 30080004.

REFRENCES

[1] Plaxco, K. W. Simons, K. T. and Baker, D. (**1998**) *J. Mol. Biol.*, *277*, 985-994.
[2] Gromiha, M. M. and Selvaraj, S. (**2001**) *J. Mol. Biol.*, *310*, 27-32.
[3] Zhou, H. and Zhou, Y. (**2002**) *Biophys. J.*, *82*, 458-463.
[4] Makarov, D. E. Keller, C. A. Plaxco, K. W. and Metiu, H. (**2002**) *Proc. Natl. Acad. Sci.* USA., *99*, 3535-3539.
[5] Gong, H. Isom, D. G. Srinivasan, R. and Rose, G. D. (**2003**) *J. Mol. Biol.*, *327*, 1149-1154.
[6] Shao, H. Peng, Y. Zeng, Z.-H., (**2003**) *Protein and Peptide Lett.*, *10*, 277-280.

[7] Guijarro, J. I. Morton, C. J. Plaxco, K.W. Campbell, I. D. and Dobson, C. M. (**1998**) *J. Mol. Biol.*, *276*, 657-667.
[8] Jackson, S. E. (**1998**) *Fold Des.*, 3, R81-91.
[9] Debe, D. A. and Goddard, W. A. 3rd. (**1999**) *J. Mol. Biol.*, *294*, 619-625.
[10] Perl, D. Welker, C. Schindler, T. Schroder, K. Marahiel, M. A. Jaenicke, R. and Schmid, F. X. (**1998**) *Nat. Struct. Biol.*, *5*, 229-235.
[11] Reid, K. L. Rodriguez, H. M. Hillier, B. J. and Gregoret, L. M. (**1998**) *Protein Sci.*, *7*, 470-479.
[12] Schonbrunner, N. Koller, K. P. and Kiefhaber, T. (**1997**) *J. Mol. Biol.*, *268*, 526-538.
[13] Clarke, J. Cota, E. Fowler, S. B. and Hamill, S. J. (**1999**) *Structure Fold Des. 7*, 1145-53.
[14] Plaxco, K. W. Spitzfaden, C. Campbell, I. D. and Dobson, C. M. (**1997**) *J. Mol. Biol.*, *270*, 763-770.
[15] Lim, V. I. (**1974**) *J. Mol. Biol.*, *88*, 857-872.

Received on June 16, 2003, accepted on July 28, 2003.