Multiple Contact Network Is a Key Determinant to Protein Folding Rates

M. Michael Gromiha*

Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), AIST Tokyo Waterfront Bio-IT Research Building, 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan

Received December 3, 2008

Understanding the relationship between amino acid sequences and folding rates of proteins is an important task in computational and molecular biology. It has been shown that topological parameters, contact order, long-range order, and total contact distance relate well with protein folding rates. In this work, we have systematically analyzed the influence of amino acid residues that form multiple contacts in protein structures to folding rates of proteins. We observed an inverse relationship between the number of residues with multiple contacts and protein folding rates. Further analysis indicates that multiple contacts are influenced by hydrophobic residues, whereas the role is minimal between the residues that are capable of forming hydrogen bonds. The propensity of multiple contacts forming residues showed that aromatic and hydrophobic residues are dominant in two-state proteins, whereas the polar residues Ser and Thr are also preferred ones in three-state proteins. In addition, multiple contact forming residues are interconnected with each other through contact networks in protein structures. The comparison between slow and fast folding proteins demonstrated the presence of more multiple contact forming residues in slow folding proteins with a limit of 4−6 contacts/residue. These results have been reflected in amino acid sequences in the form of short-, medium-, and long-range contacts, which could discriminate slow and fast folding proteins with an accuracy of 96% using a 5-fold cross-validation method.

INTRODUCTION

The folding rate of a protein is the measure to understand the tendency of folding (slow/fast) from its unfolded state to its native three-dimensional structure. Studies of protein folding rates enhance our understanding of the variations in protein folding kinetics, which may lead to several pathologies such as prion and Alzheimer diseases. Fulton et al.[1] collected experimental data on protein folding rates and developed a protein folding database for understanding protein folding and stability.

As an advancement to understand/predict protein folding rates, Plaxco et al.[2] proposed the concept of contact order (CO) using information about average sequence separation of all contacting residues in the native state of two-state proteins and found a significant correlation between CO and folding rates of two-state proteins. Gromiha and Selvaraj[3] defined a novel parameter, long-range order (LRO) from the knowledge of long-range contacts (contact between two residues that are close in space and far in the sequence) in protein structure and established a simple statistical model for predicting protein folding rates. Recently it has been reported that LRO is the only parameter that shows excellent correlation with the folding rates of all structural classes of proteins.[4] These two parameters, CO and LRO, are incorporated into a new parameter, total contact distance (TCD), which showed a good relationship with protein folding rates.[5] All these parameters are derived from the knowledge of inter-residue interactions in protein three-dimensional structures.[6]

On the other hand, investigations have been carried out to understand/predict the folding rates of proteins from protein three-dimensional structures, secondary structure information, and amino acid sequences.[7] The prediction of protein folding rates from amino acid sequence includes the relationship with amino acid properties,[8−11] predicted secondary structures,[12] predicted inter-residue contacts,[13] amino acid composition,[14,15] and secondary structure length.[16]

In this work, we have systematically analyzed the contacts between amino acid residues in protein structures and revealed the importance of residues that form multiple contacts to determine protein folding rates. We found that the number of multiple contact forming residues explains better the folding rates of proteins than other structural/topological parameters, CO, LRO, and TCD. The multiple contacts are mainly influenced by hydrophobic residues, and the role of hydrogen bond forming residues is minimal. Further, we have dissected the limit of contacts between slow and fast folding proteins and analyzed the influence of short-, medium-, and long-range contacts to discriminate slow and fast folding proteins. We have utilized several machine learning techniques and showed that the slow and fast folding proteins could be discriminated with an accuracy of 96%.

MATERIALS AND METHODS

Experimental Folding Rates. We have used three sets of data in the present work: (i) 50 two-state proteins used in Gromiha et al.,[10] (ii) 27 two-state proteins with same experimental conditions,[17] and (iii) 25 three-state proteins.[10] Although some two-state proteins are common in data sets (i) and (ii), the data reported in Maxwell et al.[17] have been

* Corresponding author phone: +81-3-3599-8046; fax: +81-3-3599-8081; e-mail: michael-gromiha@aist.go.jp.

10.1021/ci800440x CCC: $40.75 © 2009 American Chemical Society
Published on Web 04/02/2009

Table 1. Correlation between Multiple Contact Index and Protein Folding Rates[a]

data set	distance r_{ij}	residue separation |i-j|	minimum no. of contacts (n_c)	r
two-state 50	*7.5*	*12*	*4*	*−0.80*
hyd-hyd	7.5	13	2	−0.76
hyd-polar	7.5	7	3	−0.79
polar−polar	5.5	6	1	−0.72
HB-HB	6.0	16	1	−0.69
HB-nHB	7.0	13	1	−0.78
nHB-nHB	6.5	17	1	−0.67
CO				−0.64
LRO				−0.73
TCD				−0.73
two-state 27	8.0	11	2	−0.81
two-state 27	*7.5*	*12*	*4*	*−0.71*
three-state 25	*6.5*	*3*	*5*	*−0.83*

[a] Two-state 50:50 two-state proteins. Two-state 27:27 two-state proteins. Three-state 25:25 three-state proteins. Hyd: hydrophobic; HB: hydrogen bond; nHB: non-hydrogen-bond. CO: contact order; LRO: long-range order; TCD: total contact distance. *r*: linear correlation coefficient obtained between MCI and ln(k_f). The results obtained with the parameters given in eqs 1 and 2 are shown in italics.

unified with the same experimental conditions, and hence we have used the data set as such. The three-dimensional structures of proteins have been obtained from the Protein Data Bank.[18]

Multiple Contact Index (MCI). We have used three parameters to estimate the limit of multiple contacts. Two of them are the same as those we used earlier to define long-range order: the distance between amino acid residues in space and the sequence separation between them. The third one is the number of residues that have multiple contacts. Using the three parameters we define "multiple contact index" as given below

$$\begin{gathered} n_{ci} = \sum n_{ij};\ n_{ij}=1 \text{ if } r_{ij} < 7.5\ \text{Å}; |i\text{-}j|>12 \text{ residues; 0 otherwise} \\ \text{MCI} = \sum n_{mi}/N;\ n_{mi} = 1 \text{ if } n_{ci} \geq 4; \text{ 0 otherwise} \end{gathered} \tag{1}$$

where n_c is the number of contacts for each residue, and r_{ij} is the distance between the residues i and j. MCI gives the total number of residues that have more than 4 contacts with the conditions that the contacting residues are within the limit of 7.5 Å and are separated with at least 12 residues. As in other studies we have normalized the number of residues with multiple contacts in two-state proteins. For three-state proteins as the chain length is a major determinant,[19] we did not normalize with chain length. The MCI for three-state proteins is defined below:

$$\begin{gathered} n_{ci} = \sum n_{ij};\ n_{ij}=1 \text{ if } r_{ij} < 6.5\ \text{Å}; |i\text{-}j|>3 \text{ residues; 0 otherwise} \\ \text{MCI} = \sum n_{mi};\ n_{mi}=1 \text{ if } n_{ci} \geq 5; \text{ 0 otherwise} \end{gathered} \tag{2}$$

The MCI is further analyzed in terms of contacts between hydrophobic residues, polar residues, hydrophobic and polar residues and based on the classification of residues that are capable of forming hydrogen bonds.

MCI and Cliquishness. Micheletti[20] derived a parameter, cliquishness, using the concept of contact order[2] and an additional descriptor of the number of contacts. It is defined as[20]

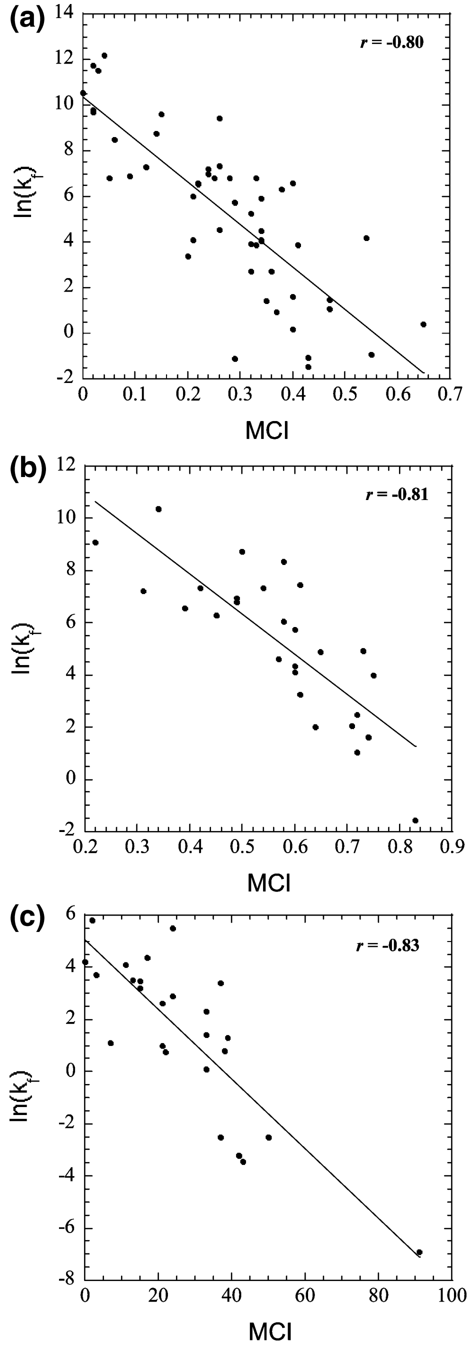

Figure 1. Relationship between MCI and ln(k_f) in two- and three-state proteins: (a) 50 two-state proteins, (b) 27 two-state proteins, and (c) 25 three-state proteins.

$$\text{Cliquishness(i)} = \sum \Delta_{ij}\Delta_{ij}\Delta_{ij}/[N_c(N_c\text{-}1)/2] \tag{3}$$

where N_c is the number of contacts for the residue, i. The cliquishness is properly defined only if the residue i is connected with at least two other residues.

On the other hand MCI is based on long-range order,[3] which mainly accounts for the number of long-range contacts with specific cutoffs for the spatial distance between two residues and the distance of separation in amino acid sequence. MCI is a measure of the number of residues, which have a minimum number of contacts with other residues.

Table 2. Propensity of Amino Acids To Form Multiple Contacts in Two- and Three-State Proteins

	propensity		
residue	2-state proteins (50)	2-state proteins (27)	3-state proteins (25)
Ala	1.04	1.03	1.10
Asp	0.55	0.73	0.57
Cys	1.61	1.57	2.19
Glu	0.54	0.74	0.55
Phe	1.24	1.22	0.93
Gly	0.88	0.88	1.01
His	1.02	1.02	0.87
Ile	1.60	1.32	1.59
Lys	0.71	0.88	0.74
Leu	1.10	1.15	1.15
Met	0.79	1.08	0.88
Asn	0.89	0.81	0.29
Pro	0.94	0.98	0.70
Gln	0.89	0.75	0.66
Arg	0.82	1.05	0.67
Ser	0.76	0.94	1.10
Thr	1.04	1.01	1.30
Val	1.87	1.35	1.61
Trp	1.50	1.25	1.67
Tyr	1.34	1.10	1.39

Table 3. Residues That Form Multiple Contacts and Those in Multiple Contact Networks in Chymotrypsin Inhibitor[a]

multiple contact forming residues	residues involved in multiple contact network
K3, T4, E5, W6, P7, L9, V10, G11, K12 S13, V14, A28, Q29, I30, I31, V32, L33, P34, V35, K44, I45, D46, R47, V48, R49, L50, F51, V52, D53, D56, N57, I58, A59, E60, V61, P62, R63, V64, G65	T4, E5, W6, P7, L9, V10, G11, K12 S13, V14, *A17, K18, I21, D24, K25, E27* A28, Q29, I30, I31, V32, L33, P34, V35, K44, I45, D46, R47, V48, R49, L50, F51, V52, D53, D56, N57, I58, A59, E60, V61, P62, R63, V64, G65

[a] The six residues that are not involved in multiple contacts are shown in italics.

We have used a minimum of 4 and 5 contacts, respectively, for two- and three-state proteins.

Further, the number of contacts for each residue has been used to define cliquishness, whereas the normalization is done with the chain length to define the MCI for two-state proteins.

Propensity of Residues To Form Multiple Contacts. The multiple contact forming propensity for the 20 amino acid residues in protein structures has been developed as follows: we have computed the frequency of occurrence of amino acid residues that form multiple contacts (f_{mc}) and in the protein as a whole (f_t). The propensity (P_{mc}) is calculated using the equation

$$P_{mc}(i) = f_{mc}(i)/f_t(i) \quad (4)$$

where i represents each of the 20 amino acid residues. The propensities have been normalized with the total number of multiple contact forming residues and total number of residues in all the considered proteins.

Multiple Contact Forming Residues in Slow and Fast Folding Proteins. We have classified the proteins into slow and fast folding based on their folding rates. The proteins with $\ln(k_f) > 6$ are classified as fast folding, and others are assigned as slow folding proteins. The number of residues that have different numbers of contacts (say 1 to 15) have been computed, and the difference between slow and fast folding proteins has been delineated. The statistical significance of the results obtained in the present study has been verified with *t*-test and *p*-value by standard procedures.

Machine Learning Techniques. We have analyzed several machine learning techniques implemented in the WEKA program[21] for discriminating slow and fast proteins. This program includes several methods based on Bayes functions, neural networks, logistic functions, support vector machines, regression analysis, nearest neighbor methods, meta learning, decision trees, and rules. The details of these methods have been explained in our earlier article.[22] We have analyzed different classifiers and data sets to discriminate slow and fast folding proteins.

5-Fold Cross-Validation Method. We have performed a 5-fold cross-validation test for assessing the validity of the present work. In this method, the data set is divided into five groups: four of them are used for training, and the fifth is used for testing the method. The same procedure is repeated five times, and the average is computed to obtain the accuracy of the method.

RESULTS AND DISCUSSION

Relationship between Multiple Contact Index and Protein Folding Rates. We have computed the multiple contact index by varying different parameters, such as distance between two residues, distance of separation at sequence level, and number of contacts. The computed MCI for a set of proteins has been compared with their respective $\ln(k_f)$ values, and the parameters giving the highest correlation are shown in Table 1. We found that the MCI obtained with at least four contacts within a distance of 7.5 Å and separated by at least 12 residues (eq 1) showed the highest linear correlation of −0.80 with $\ln(k_f)$. The relationship between MCI and $\ln(k_f)$ is shown in Figure 1a. Interestingly, the distances between amino acid residues and sequence separation are similar to that obtained for defining long-range order.[3] The correlation increased from −0.73 to −0.80 due to the additional constraint on the minimum number of contacts for the amino acid residues. The correlations obtained with CO and TCD are −0.64 and −0.73, respectively, with the same data set. Using the data set of Maxwell et al.[17] and the same conditions we obtained a correlation of −0.71 (Table 1). The refinement of parameters increased the correlation up to −0.80, as shown in Figure 1b.

We have analyzed the tendency of hydrophobic residues to form multiple contacts in protein structures and to determine their folding rates. We found that multiple contacts between hydrophobic and polar residues attained the highest correlation of −0.79, whereas the correlation obtained with multiple contacts between hydrophobic residues or polar residues is −0.76 and −0.72, respectively. On the other hand, the classification of residues based on hydrogen bond (HB) forming capability showed correlations of −0.69, −0.67, and −0.78, respectively, for pairs of HB forming residues, non-HB forming residues and between HB and non-HB forming residues. In both hydrophobic and HB based classifications, the contacts between hydrophobic-polar or HB-nonHB forming residues showed the highest correlation. This observation emphasizes the importance of contacts between hydrophobic and polar residues for determining protein folding rates, which are reported to be important for the

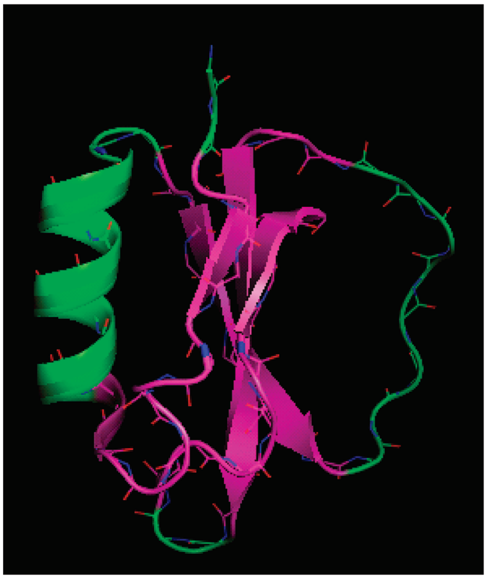

Figure 2. Multiple contact forming residues in chymotrypsin inhibitor (2CI2), which are shown in magenta.

stability of proteins.[23] Further, the minimum number of contacts is three in the case of contacts based on hydrophobicity of residues, whereas it is one for the classification of residues based on HB forming capability. This might be due to the chemical behavior of hydrophobic and hydrogen bonding interactions.

In three-state proteins we observed a different trend in the relationship between multiple contact forming residues and protein folding rates. Within the spatial distance of 6.5 Å, sequence separation of more than 3 residues and a minimum of 5 contacts (eq 2) showed the highest correlation of −0.83 (Table 1). The relationship between MCI and ln(k_f) is shown in Figure 1c. We inferred from this result that the medium- and long-range contacts influence protein folding rates, and the number of contacts formed within the limit of 6.5 Å revealed the presence of clusters, which are key determinants of protein folding rates in three-state proteins.

In earlier studies, the physical origin behind the correlations between various topological measures and protein folding rates has been discussed with a topomer search model[24] and loop closure principles.[25] The definition of MCI seems to follow the principles of loop closure as in the case of contact order and long-range order.

Influence of Chain Length on Protein Folding Rates. We have analyzed the influence of chain length on protein folding rates in two directions: (i) based on the normalization with chain length in two- and three-state proteins and (ii) the combination of two- and three-state proteins together. The MCI obtained with eq 1 for a set of 50 two-state proteins showed a correlation of −0.80 with ln(k_f), whereas the correlation was −0.58 without normalization. A similar trend is also observed for the set of 27 two-state proteins (r is −0.71 and −0.47, with and without normalization, respectively). On the other hand, the MCI obtained with eq 2 for a set of 25 three-state proteins showed a correlation of −0.83. The normalization of MCI with chain length decreased the correlation to −0.42. Further, the combination of 50 two-state and 25 three-state proteins showed correlations of −0.69 and −0.70 between ln(k_f) and MCI obtained with and without normalization of chain length, respectively. These results agree with the previous observation that the chain length is a major determinant in three-state proteins.[19]

Multiple Contact Forming Propensity of Amino Acid Residues in Two- and Three-State Proteins. We have computed the multiple contact forming propensity of the 20 amino acid residues by comparing the residues that can form multiple contacts and total number of residues in a protein (eq 4). The results obtained with a data set of 50 two-state proteins are presented in Table 2. We observed that the aromatic and other hydrophobic residues prefer to form multiple contacts, which reveals the formation of hydrophobic clusters and/or aromatic−aromatic interactions. Among the hydrophobic residues Val has the highest preference followed by Ile. Trp has a higher preference compared with Tyr and Phe. We observed a similar trend with the data set of Maxwell et al.[17]

In Table 2, we also included the results for 25 three-state proteins. In three-state proteins Cys has the highest preference, which indicates the formation of disulfide bridges and contacts with other residues. Interestingly, the polar residues Thr and Ser prefer to form multiple contacts in three-state proteins along with hydrophobic residues. Further, we observed that the propensity of Cys to form multiple contacts is universally high for both two- and three-state proteins.

Contact Networks among the Residues That Form Multiple Contacts. We have collected the information about all multiple contact forming residues and examined the possibility of forming a network. This has been done with a procedure as follows: for a residue A with multiple contacts we have detected all the contacting residues (say B, C, D, etc.). Then we have checked whether the residues B, C, D, etc. have multiple contacts. This procedure is repeated for all the multiple contact forming residues in a protein. As an example, the residues that form more than 2 contacts with a sequence separation of >11 residues within 8 Å in chymotrypsin inhibitor (2CI2) are presented in Table 3. In addition, the residues that are involved in multiple contacts are also included. The comparison of these two sets of data showed that only six residues in the contact network are not involved in multiple contacts. The residues involved in multiple contact networks in 2CI2 are shown in Figure 2. We noticed that all the residues are located in the interior of the protein, which mainly has β-strand conformation. We have analyzed the behavior in all the considered 27 proteins, and we observed that the multiple contact forming residues are involved in the contact network, which may play a key role in protein folding rates.

Influence of Multiple Contact Forming Residues in Slow and Fast Folding Proteins. We have analyzed the influence of residues with different numbers of contacts on the folding rates of two- and three-state proteins. In this analysis, we have chosen a common cutoff distance (8 Å) and residue separation (>2 residues) as well as specific cutoffs with highest correlation (Table 1). In Figure 3a we show the percentage of residues with different numbers of contacts in a data set of 27 two-state proteins. We observed that the percentage of residues with up to seven contacts is higher in fast folding proteins than slow folding proteins, whereas the number of residues with more than seven contacts is high in slow folding proteins. A similar trend is also observed with the data set of 50 two-state proteins (Figure 3b) and 25 three-state proteins (Figure 3c). We have

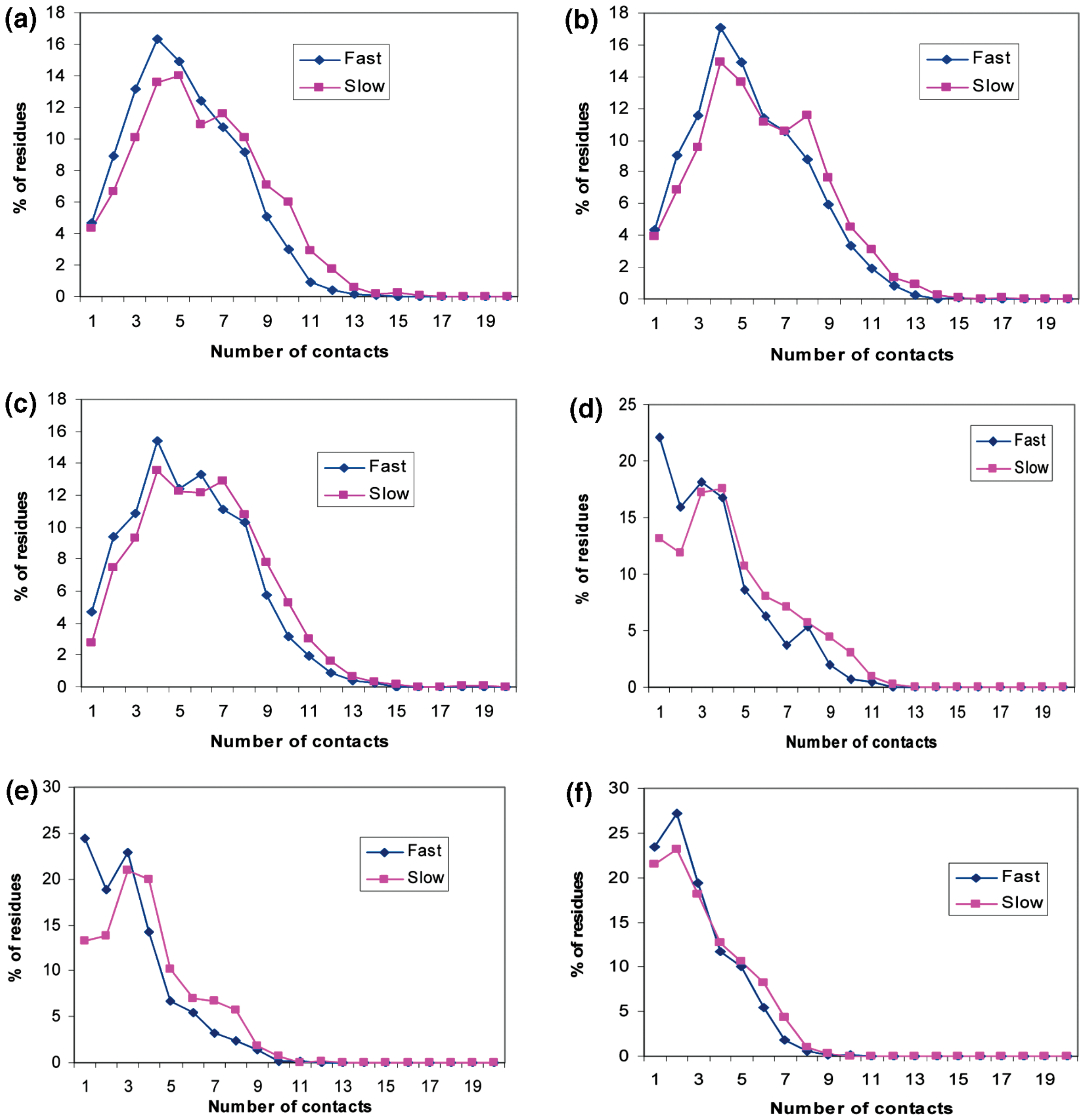

Figure 3. Percentage of residues with different numbers of contacts in two- and three-state proteins. The *p*-value and *t*-test data are shown in parentheses. (a) 27 two-state proteins with the same cutoffs (p = 1.3e-6; t = 13.5), (b) 50 two-state proteins with the same cutoffs (p = 2.4e-6; t = 14.5), (c) 25 three-state proteins with the same cutoffs (p = 2.9e-6; t = 13.5), (d) 27 two-state proteins with specific cutoffs (8 Å; >11 residues; p = 2e-4; t = 7.3), (e) 50 two-state proteins with specific cutoffs (7.5 Å; >12 residues p = 6e-4; t = 5.7), and (f) 25 three-state proteins with specific cutoffs (6.5 Å; >3 residues p = 1.5e-4; t = 19). The same cutoff is the one used commonly for all the three sets of proteins (8 Å; >2 residues). The specific cutoff is the one obtained for each set of proteins to attain the highest correlation coefficient.

also analyzed the results with different cutoff distances for proteins in different data sets. We noticed the difference in the cutoff contacts (four) and the performance is similar to that observed with common criteria indicating the influence of multiple contacts in slow folding proteins (Figure 3d-f). Although the percentage of residues with multiple contacts is low in protein structures, they play a vital role to form clusters, which may be a key determinant in protein folding rates.

We have verified the significance of the difference between fast and slow folding proteins using statistical tests (*p*-value and *t*-test). We found that in all sets of two- and three-state proteins (Figure 3a-f) the *p*-value is less than 1e-4 and the *t*-test value lies in the range of 5 to 20. This analysis validates the significance of the results obtained in the present study.

Discrimination of Slow and Fast Folding Proteins. We have computed the parameters, amino acid occurrence, residue pair preference, and short-, medium-, and long-range contacts as well as the combination of contacts using amino acid sequence information. The residues within the sequential interval of 3 residues from the central one are termed short-range contacts, 3−4 residues are medium-range contacts, and more than 4 residues are termed long-range contacts.[6] We have used a window size of up to 21 residues (10 residues on both sides) in this work. On the other hand, the proteins have been classified into slow ($\ln(k_f) < 3$/s), medium (3/s < $\ln(k_f) < 6$/s), and fast ($\ln(k_f) > 6$/s) folding based on their folding rates. The parameters amino acid occurrence, residue pair preference, and short-, medium-, and long-range contacts have been used as input features and are related to protein folding rates. We found that the performance is better with contacts than amino acid occurrence/residue pair preference. We have used a data set of 75 proteins for three-state classification (slow, medium, and fast) and obtained an accuracy of 62% with 5-fold cross-validation. We noticed that the confusion is mainly between fast and medium and medium and slow. The fast and slow folding proteins can be distinguished correctly with the accuracy of 96%. When we tested with a set of 27 proteins used in Maxwell et al.,[17] we obtained an accuracy of 96.3% in distinguishing slow and fast folding proteins; only one fast folding protein (CheW) was misclassified as a slow folding protein. The classification of proteins based on slow and fast folding will have several applications in genome-wide annotations. The refinement of parameters and attempts to improve prediction accuracy are in progress.

CONCLUSIONS

We have systematically analyzed the relationship between residues forming multiple contacts and protein folding rates. We observed that the fast folding proteins have fewer multiple contact forming residues compared with slow folding proteins. The propensity of amino acids to form multiple contacts indicates the involvement of hydrophobic residues, and they tend to form contact networks. The information about short-, medium-, and long-range contacts carries more information than amino acid composition or residue-pair preference for discriminating fast and slow folding proteins. We have utilized the contact information from amino acid sequence, which could discriminate the slow and fast folding proteins at an accuracy of 96% with a 5-fold cross-validation test. The refinement of parameters, improvement of prediction performance, and applications to genomic sequences are under progress.

ACKNOWLEDGMENT

The author wishes to thank Dr. Martin Frith for reading the manuscript and the anonymous reviewers for constructive comments. This research was partially supported by Strategic International Cooperative Program, Japan Science and Technology Agency (JST).

REFERENCES AND NOTES

(1) Fulton, K. F.; Bate, M. A.; Faux, N. G.; Mahmood, K.; Betts, C.; Buckle, A. M. Protein Folding Database (PFD 2.0): an online environment for the International Foldeomics Consortium. *Nucleic Acids Res.* **2007**, *35*, D304–307.

(2) Plaxco, K. W.; Simons, K. T.; Baker, D. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **1998**, *277*, 985–94.

(3) Gromiha, M. M.; Selvaraj, S. 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.* **2001**, *310*, 27–32.

(4) Istomin, A. Y.; Jacobs, D. J.; Livesay, D. R. On the role of structural class of a protein with two-state folding kinetics in determining correlations between its size, topology, and folding rate. *Protein Sci.* **2007**, *16*, 2564–2569.

(5) Zhou, H.; Zhou, Y. Folding rate prediction using total contact distance. *Biophys. J.* **2002**, *82*, 458–63.

(6) Gromiha, M. M.; Selvaraj, S. Inter-residue interactions in protein folding and stability. *Prog. Biophys. Mol. Biol.* **2004**, *86*, 235–77.

(7) Gromiha, M. M.; Selvaraj, S. Bioinformatics approaches for understanding and predicting protein folding rates. *Curr. Bioinf.* **2008**, *3*, 1–8.

(8) Gromiha, M. M. Importance of native state topology for determining the folding rate of two-state proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1481–1485.

(9) Gromiha, M. M. A Statistical Model for Predicting Protein Folding Rates from Amino Acid Sequence with Structural Class Information. *J. Chem. Inf. Model.* **2005**, *45*, 494–501.

(10) Gromiha, M. M.; Thangakani, A. M.; Selvaraj, S. FOLD-RATE: prediction of protein folding rates from amino acid sequence. *Nucleic Acids Res.* **2006**, *34*, W70–W74.

(11) Huang, J. T.; Tian, J. Amino acid sequence predicts folding rate for middle-size two-state proteins. *Proteins* **2006**, *63*, 551–554.

(12) Ivankov, D. N.; Finkelstein, A. V. Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 8942–8944.

(13) Punta, M.; Rost, B. Protein folding rates estimated from contact predictions. *J. Mol. Biol.* **2005**, *348*, 507–512.

(14) Ma, B. G.; Guo, J. X.; Zhang, H. Y. Direct correlation between proteins' folding rates and their amino acid compositions: an ab initio folding rate prediction. *Proteins* **2006**, *65*, 362–372.

(15) Huang, L.-T.; Gromiha, M. M. Prediction of protein folding rates using quadratic response surface models. *J. Comput. Chem.* **2008**, *29*, 1675–1683.

(16) Huang, J. T.; Cheng, J. P.; Chen, H. Secondary structure length as a determinant of folding rate of proteins with two- and three-state kinetics. *Proteins* **2007**, *67*, 12–17.

(17) Maxwell, K. L.; Wildes, D.; Zarrine-Afsar, A.; De Los Rios, M. A.; Brown, A. G.; Friel, C. T.; Hedberg, L.; Horng, J. C.; Bona, D.; Miller, E. J.; Vallée-Bélisle, A.; Main, E. R.; Bemporad, F.; Qiu, L.; Teilum, K.; Vu, N. D.; Edwards, A. M.; Ruczinski, I.; Poulsen, F. M.; Kragelund, B. B.; Michnick, S. W.; Chiti, F.; Bai, Y.; Hagen, S. J.; Serrano, L.; Oliveberg, M.; Raleigh, D. P.; Wittung-Stafshede, P.; Radford, S. E.; Jackson, S. E.; Sosnick, T. R.; Marqusee, S.; Davidson, A. R.; Plaxco, K. W. Protein folding: defining a "standard" set of experimental conditions and a preliminary kinetic data set of two-state proteins. *Protein Sci.* **2005**, *14*, 602–616.

(18) Berman, H.; Henrick, K.; Nakamura, H.; Markley, J. L. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.* **2007**, *35*, D301–303.

(19) Galzitskaya, O. V.; Garbuzynskiy, S. O.; Ivankov, D. N.; Finkelstein, A. V. Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics. *Proteins* **2003**, *51*, 162–166.

(20) Micheletti, C. Prediction of folding rates and transition-state placement from native-state geometry. *Proteins* **2003**, *51*, 74–84.

(21) Witten, I. H.; Frank, E. *Data Mining: Practical machine learning tools and techniques*; 2nd ed., Morgan Kaufmann: San Francisco, 2005.

(22) Gromiha, M. M.; Suwa, M. Discrimination of outer membrane proteins using machine learning algorithms. *Proteins* **2006**, *63*, 1031–1037.

(23) Gromiha, M. M. Important inter-residue contacts for enhancing the thermal stability of thermophilic proteins. *Biophys. Chem.* **2001**, *91*, 71–77.

(24) Makarov, D. E.; Plaxco, K. W. The topomer search model: A simple, quantitative theory of two-state protein folding kinetics. *Protein Sci.* **2003**, *12*, 17–26.

(25) Weikl, T. R. Loop-closure principles in protein folding. *Arch. Biochem. Biophys.* **2008**, *469*, 67–75.

CI800440X