Protein Fold Recognition Based Upon the Amino Acid Occurrence

Y.-h. Taguchi¹ and M. Michael Gromiha²

¹ Department of Physics, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo
112-8551, Japan

tag@granular.com

² 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

michael-gromiha@aist.go.jp

Abstract. We have investigated the relative performance of amino acid occurrence and other features, such as predicted secondary structure, hydrophobicity, normalized van der Waals volume, polarity, polarizability, and real/predicted contact information of residues, for recognizing protein folds. We observed that the improvement over other features is only marginal compared with amino acid occurrence. This is because amino acid occurrence, indirectly, can consider varieties of physical properties which are useful to discriminate protein folds. If we consider only proteins which are well aligned structurally with each other, the accuracy of discrimination is drastically improved. In order to discriminate protein folds more accurately, we need to consider anything other than structure alignment.

1 Introduction

Deciphering the native conformation of a protein from its amino acid sequence known as protein folding problem is a challenging task. The recognition of proteins of similar folds is a key intermediate step for protein structure prediction. Alignment profiles are widely used for recognizing protein folds [1,2]. Recently, Cheng and Baldi [3] proposed a machine learning algorithm using secondary structure, solvent accessibility, contact map and β -strand pairing for fold recognition, which showed the pair wise sensitivity of 27%. On the other hand, it has been reported that the amino acid properties are the key determinants of protein folding and are used for discriminating membrane proteins [4], identification of membrane spanning regions [5], prediction of protein structural classes [6], protein folding rates [7], protein stability [8] etc. Towards this direction, Ding and Dubchak [9] proposed a method based on neural networks and support vector machines for fold recognition using amino acid composition and five other properties, and reported a cross-validated sensitivity of 45 %.

Recently [10], we have used the amino acid occurrence (not composition) of proteins belonging to 30 major folds for recognizing protein folds. We have

developed a method based on linear discriminant analysis (LDA), which showed an accuracy of 37% for recognizing 1612 proteins from 30 different folds, which is comparable with other methods in the literature, in spite of the simplicity of our method and the large number of proteins considered.

In this paper, we have compared the performance of other features with that of amino acid occurrence. We have found that amino acid occurrence outperform other features to discriminate protein folds. Even if other features are considered together with amino acid occurrence, the ability to discriminate protein folds is hardly improved. On the hand, if we exclude pairs of proteins with poor 3D structural alignment, we have found that discrimination by amino acid occurrence is drastically improved. In conclusion, amino acid occurrence turns out to be the best feature to discriminate protein folds.

2 Materials and Methods

We have used three data sets to test the performance of our method. The first data set is that used by Ding and Dubchak [9]. It is available from their web site. The second data set is that used in the previous study [10]. It consists of 1612 amino acid sequence among which there are less than 25 % sequence identity. These amino acid sequence is taken from SCOP [11] and belong to one of major 27 folds. It is available from our prediction server [12]. We also used several feature (contact) vectors corresponding to these data set. The third one is taken from CATH [13]. It consists of 4146 amino acid sequences with less than 40 % mutual sequence identity. These belong to one of major 39 topologies. The selection of major 39 topologies is based upon Gubbi *et al* [14].

Since the method is described in our previous report [10], we have briefly outlined our methods. First, we have counted the number of amino acid residues in each amino acid sequence. Thus, we have 20 dimensional integer vector for each protein. Then, LDA is applied to this set of vectors. LDA we used is *lda* module in R [15]. Although there are many ways to weight the discrimination [10], in this paper we weight each fold(topology) equally. In other words, prior probability of each fold (topology) is assumed to be equal. As a measure of performance, we employ accuracy Q ,

$$Q = \frac{\sum_i TP_i}{N}, \quad (1)$$

where TP_i is the number of proteins correctly discriminated in i th fold (topology) and N is total number of proteins considered. In the following, Q by leave one out cross validation will be reported.

3 Results

3.1 Accuracy of Discrimination Using Amino Acid Occurrence and Other Features

Ding and Dubchak [9] has discriminated 27 folds for 311 proteins. They have used support vector machine (SVM) and/or neural networks (NN) with voting

Table 1. Accuracy Q as a function of used features

Features	Q [%]	References
Composition	35	
Composition + length	38	
Composition + five features	39	
Composition + five features	45	Ding and Dubchak [9]
Occurrence	42	Our previous report [10]
Occurrence + five features	44	

system. Features they used differ from amino acid occurrence. They have reported that their method achieved $Q = 45\%$ as 10-fold cross validation results. In the previous report [10], we have shown that our method can achieve $Q = 42\%$ in spite of simplicity of our method. Since ours are leave one out cross validation, our Q value does not have any statistical errors. However, Ding's value is 10-fold cross validation. If we consider this, our value $Q = 42\%$ is comparative with Ding's value $Q = 45\%$. Since our method uses solely amino acid occurrence while Ding and Dubchak used many other features than amino acid occurrence together, it is natural to expect that considering other features together with amino acid occurrence can improve accuracy Q . In Table 1, we have summarized Q as a function of used features. When we use more than two features to discriminate folds, we simply apply LDA to merged feature vectors. This means, if there are two features vectors $\mathbf{f}_n$ with n components and $\mathbf{f}_m$ with m features,

$$\mathbf{f}_n = (f_{n1}, f_{n2}, \dots, f_{nn}) \quad (2)$$

$$\mathbf{f}_m = (f_{m1}, f_{m2}, \dots, f_{mm}), \quad (3)$$

then we merge these two and apply LDA to

$$\mathbf{f}_{m+n} = (f_{n1}, f_{n2}, \dots, f_{nn}, f_{m1}, f_{m2}, \dots, f_{mm}). \quad (4)$$

Additional five features, i.e., predicted secondary structure, hydrophobicity, normalized van der Waals volume, polarity, polarizability, are those Ding and Dubchak [9] used. Since their method is sophisticated, and it utilized all features (i.e., composition + five features), their Q is better than us by 6 %. In spite of that, our simple method employing amino acid occurrence and five features has almost fulfilled this gap ($Q = 44\%$). If we consider the simplicity of our method, our method is even better than Ding and Dubchak's method.

If we see Table 1 more detail, we can find many interesting things. For example, if we consider only composition, Q is only 35 %, which is 7 % smaller than $Q = 42\%$ when we consider only occurrence. On the other hand, if we consider composition and length, i.e., the first 20 components of feature vectors consist of composition and the 21th component is amino acid length, Q raises from 35% to 38 %. In spite of that, if we consider composition and five features, Q becomes 39 %, which is as the same as $Q = 38\%$ when composition and length are considered. Thus, solely considering length is comparative with considering

all of five features. This definitely demonstrate the importance of considering amino acid length. This is the reason why considering occurrence instead of composition can improve accuracy by 7%.

3.2 SCOP

Since our method is simple, we can deal with larger data set. In our previous report [10] we have applied our method to 1612 proteins belonging to 30 major folds in SCOP. Q which we have achieved was 33 %. In this subsection, we have compared Q when considering other features than amino acid occurrence with Q when only amino acid occurrence is considered. In Table 2, we have listed Q obtained using other features than amino acid occurrence. Other features we used is average contacts in different sequence intervals, for example, 3-4, >4, 5-10, 11-20, 21-30, 31-40, 41-50 and >50. The contacts are predicted by several different contact prediction servers [16,17,18] and are taken from real structure. For some predicted contacts, number of amino acid sequence considered is less than 1612. Clearly, the real contact information outperformed in discrimination. On the other hand, the performance of amino acid occurrence is better than that with predicted contacts.

We also consider dipeptide occurrence. Since there are 20 amino acids, dipeptide occurrence n_{ij} are 400 kinds, where n_{ij} is the dipeptide occurrence for i th and j th amino acid ($1 \leq i, j \leq 20$). In Table 3, we have shown Q for considering dipeptide occurrence. Consideration of dipeptide occurrence does not improve Q at all. Even if we consider dipeptide occurrence together with amino acid occurrence, Q is not improved.

It may be assumed that the consideration of dipeptide occurrence would improve the Q value. However, we observed that the Q value is less than that with

Table 2. Accuracy Q as a function of used feature

Feature	number of sequences	Q [%]	References
Occurrence	1612	33	Our previous report [10]
Composition	1612	26	Our previous report [10]
cornet	1530	22	Ref. [16]
nick	1555	13	Ref. [17]
gpcpred	1612	15	Ref. [18]
real structure	1612	50	

Table 3. Accuracy Q for discrimination by dipeptide

Feature	Q [%]	References
Occurrence	33	Our previous report [10]
Dipeptide	29	
Dipeptide+Occurrence	31	

amino acid occurrence. It might be due to the fact that dipeptide occurrence is not an independent information of amino acid occurrence, because

$$n_i = \sum_j n_{ij}. \quad (5)$$

Further, Table 3 shows that dipeptide occurrence cannot have more information than that solely amino acid occurrence can provide.

3.3 CATH

In order to see if our method can discriminate other fold classification than SCOP, we have considered topologies in CATH. In Table 4, we have shown the

Table 4. Accuracy Q [%] for SCOP and CATH with various sorts of definition of Q and weighting. Bold numbers are the same as those in other tables.

		with re-weighting over all fold	without re-weighting average over all fold	average
CATH	26	34	43	24
SCOP	33	32	37	28

comparison between CATH and SCOP. It is clear that the performance depends upon the definition of methods/accuracy which data base can be discriminated better by our method. Especially, CATH is very sensitive to the variety of definition of methods/accuracy used in these databases and our method could well discriminate the folds. The lowest Q is 26 % for CATH while the highest one is 43 %. Thus, the later is larger than the former by more than 50 %. This results show how difficult to decide what the *best* discrimination is. When we consider the definition of Q used in the present research, CATH ($Q = 26\%$) is harder to discriminate than SCOP ($Q = 33\%$). Although it is generally true that our method can discriminate folds no matter how they are defined, the performance is strictly dependent how we measure the goodness. Although we do not consider CATH in more detail here, one has to be careful how we can measure the goodness of discrimination.

4 Discussion

4.1 Why Does Occurrence Work So Well?

In the previous section, we have shown that our method (amino acid occurrence + LDA) can discriminate protein folds up to 30 to 40 % for up to thousands proteins and up to 40 folds (topology). We have also shown that considering other feature than amino acid occurrence generally can hardly improve accuracy Q . In this subsection, we would like to discuss why amino acid occurrence works well.

First of all, it is natural that occurrence is better than composition in contrast to the first impression. Suppose we have some protein belonging to one fold. Then, try to duplicate its amino acid sequence. Clearly, there will be very few possibility that duplicated protein belong to same fold. This discussion definitely show that composition cannot detect this effect at all, because duplication cannot change composition. The importance of protein length can be seen in Table 1. Consideration of protein length in addition to composition can improve Q by 3 %, which is as large as half of difference between composition and occurrence. In conclusion, we had better to consider occurrence than composition to discriminate protein folds.

Second, one may think it is strange that consideration other feature than occurrence cannot improve accuracy Q so much. However, *any* physical feature can be more or less expressed by amino acid occurrence. Thus, linear combination of amino acid occurrence can express more or less many of physical properties of proteins. In order to see this, we have computed the correlation coefficients between 49 physical, chemical energetic and conformational properties of each amino acid [19,20,21] and the first discriminate function for the second data set case (i.e., 1612 proteins belonging to 30 major folds in SCOP). Each property consists of 20 dimensional vector, like

$$\mathbf{P}^k = (P_1^k, P_2^k, \dots, P_i^k, \dots, P_{20}^k), \quad (6)$$

where P_i^k describe k th physical properties of i th amino acid. Since discriminant function is also 20 dimensional vector each component of which describe contribution from each amino acid, we can take correlation coefficient between them.

As can be seen in Table 5, 23 out of 49 properties have correlation coefficients with less than 5 % q -values (i.e., FDR corrected p -values). We can find 24 out of 49 properties have less than 5% q -value if we apply the same procedure to the third data set (CATH), although the number of commonly selected properties is as large as those by chance. (Not shown here). Thus, it is clear that linear discriminant function can express many of physical properties, at least, partly. Thus, even if we do not consider physical properties directly, amino acid occurrence can express them if some of physical properties are important for the discrimination of folds. This is the reason why the consideration of amino acid occurrence can discriminate folds (topologies for CATH) well. As another example of how well amino acid occurrence can express other physical properties, we consider contact information in Table 2. In order to check if amino acid occurrence can express contact information taken from real structure (the last row in Table 2), which achieved 50 % accuracy Q , we have applied multiple linear regression,

$$f^\ell = \sum_{i'} C_{i'} n_{i'}^\ell + C_0^\ell, \quad (7)$$

where f^ℓ is the contact information of l th protein, n_i^ℓ is i th amino acid occurrence for ℓ th protein. In Table 6, we have shown squared partial correlation coefficients based upon (7). Although these values are not so high, they are too large to

Table 5. Brief descriptions of 49 selected physico-chemical, energetic and conformational properties, their correlation coefficient with the first discriminate function for SCOP, and q -value. Asterisks in the last column shows q -value is less than 5 %.

No.	Description	Corr. Coef.	q -value [%]	$q \leq 5\%$
1.	Compressibility	0.03	44.5	
2.	Thermodynamic transfer hydrophobicity	0.44	5.4	
3.	Surrounding hydrophobicity	0.72	0.3	*
4.	Polarity	0.47	4.5	*
5.	Isoelectric point	0.19	25.8	
6.	Equilibrium constant with reference to the ionization property	0.07	41.4	
7.	Molecular weight	0.05	42.8	
8.	Bulkiness	0.36	10.4	
9.	Chromatographic index	0.65	0.4	*
10.	Refractive index	0.19	25.8	
11.	Normalized consensus hydrophobicity	0.42	6.1	
12.	Short and medium range non-bonded energy	0.17	28.9	
13.	Long-range non-bonded energy	0.75	0.3	*
14.	Total non-bonded energy	0.68	0.3	*
15.	Alpha-helical tendency	0.08	39.4	
16.	Beta-helical tendency	0.68	0.3	*
17.	Turn tendency	0.54	2.3	*
18.	Coil tendency	0.45	5.0	*
19.	Helical contact area	0.24	21.8	
20.	Mean rms fluctuational displacement	0.66	0.3	*
21.	Buriedness	0.68	0.3	*
22.	Solvent accessible reduction ratio	0.68	0.3	*
23.	Average number of surrounding residues	0.67	0.3	*
24.	Power to be at the N-terminal of alpha helix	0.45	5.0	*
25.	Power to be at the C-terminal of alpha helix	0.58	1.2	*
26.	Power to be at the middle of alpha helix	0.04	44.5	
27.	Partial-specific volume	0.28	16.7	
28.	Average medium-range contacts	0.02	46.0	
29.	Average long-range contacts	0.71	0.3	*
30.	Combined surrounding hydrophobicity (globular and membrane)	0.72	0.3	*
31.	Solvent accessible surface area for denatured protein	0.13	33.7	
32.	Solvent accessible surface area for native protein	0.56	1.8	*
33.	Solvent accessible surface area for protein unfolding	0.51	3.0	*
34.	Gibbs free energy change of hydration for unfolding	0.31	14.4	
35.	Gibbs free energy change of hydration for denatured protein	0.42	6.1	
36.	Gibbs free energy change of hydration for native protein	0.49	3.7	*
37.	Unfolding enthalpy change of hydration	0.01	46.2	
38.	Unfolding entropy change of hydration	0.51	3.0	*
39.	Unfolding hydration heat capacity change	0.69	0.3	*
40.	Unfolding Gibbs free energy change of chain	0.16	29.9	
41.	Unfolding enthalpy change of chain	0.26	18.7	
42.	Unfolding entropy change of chain	0.50	3.1	*
43.	Unfolding Gibbs free energy change	0.37	9.7	
44.	Unfolding enthalpy change	0.35	10.7	
45.	Unfolding entropy change	0.33	11.9	
46.	Volume (number of non-hydrogen side chain atoms)	0.10	37.0	
47.	Shape (position of branch point in a side-chain)	0.20	25.6	
48.	Flexibility (number of side-chain dihedral angles)	0.22	23.4	
49.	Backbone dihedral probability	0.33	11.9	

neglect. Actually speaking, p -values for these is less than 1×10^{-14} . If we consider higher order,

$$f^\ell = \sum_{i'} \left[C_{i'} n_{i'}^\ell + C_{i'}^2 (n_{i'}^\ell)^2 \right] + C_0^\ell, \quad (8)$$

Table 6. Squared Partial Correlation Coefficients for (7) and (8)

Contact range	Squared Partial Correlation Coefficients	
	1st order (7)	2nd order (8)
3-4	0.29	0.37
>4	0.25	0.32
5-10	0.20	0.25
11-20	0.23	0.28
21-30	0.12	0.17
31-40	0.08	0.12
41-50	0.07	0.10
>50	0.19	0.25

squared partial correlation coefficients has increased (Here we have confirmed that Akaike Information Criterion (AIC) has decreased by considering higher order in order to avoid over fittings). This again demonstrates that amino acid occurrence can express physical properties which are useful for discrimination of protein folds.

In conclusion, in contrast to the intuition, amino acid occurrence can express, at least partially, variety of physical properties with which protein folds can be discriminated.

4.2 Folds vs Structural Alignments

Although we have considered many other features than amino acid occurrence, accuracy Q cannot be improved so much. This is because the amino acid occurrence can have ability to express other physical features as discussed in the previous subsection. In this subsection, we try to estimate the relationship between the goodness of structural alignment and the goodness of fold recognition. If structural alignment between proteins belonging to the same fold is poor, it is natural that fold recognition is not successful.

In order to check this point, we have employed the third data set (4146 proteins belonging to 39 major folds). We have randomly picked up 100 pairs of proteins from each of intra/inter topology pairs. For example, when we consider inter topology pairs from topology I and J , pairs of proteins are taken such that one of pair belongs to topology I while another of pair belongs to topology J . On the other hand, when considering intra topology pairs, both of proteins are taken from the same topologies.

Then for selected pairs, structural alignment has been done using Matalign [22] which can get structural alignment even for chopped sequence which frequently appears in CATH. We employ $Nscore$ as a measure of goodness of structural alignment,

$$Nscore \equiv \frac{3N_a}{1 + \Delta \min(\text{length1}, \text{length2})}, \quad (9)$$

where N_a is number of aligned residues, Δ is root mean squared deviation (RMSD) and length1 and length2 are number of residues of two aligned

proteins. Larger *Nscore* means better structural alignment. Then, we have found many intra topology pairs have less than or equal to *Nscores* of inter topology pairs (Fig. 1(a)). This means, solely the goodness of structural alignment cannot decide if a pair of proteins belong to the same topology or not.

Here, we have considered the pair of topologies (3 10 129) and (3 30 360) for which our method get the least accuracy *Q*. Then by applying structural alignment to all pairs among these two topologies, we have excluded intra protein pairs which have poorer structural alignment. Although *Nscores* within (3 10 129) are always larger than those between two topologies (Fig. 1 (b)), those

Fig. 1. Comparisons of histogram of *Nscore*. (a) Inter topology (left half) vs intra topology (right half) for all of 39 topologies considered. Each histrgam is normalized such that total area is unity no matter how many pairs are considered. Alignment has been done randomly sampled 100 pairs of proteins for each of pairs of topologies. (b) *Nscore* between topologies (3 10 129) and (3 30 360) (left half) vs that within (3 10 129) (right half) (c) *Nscore* between topologies (3 10 129) and (3 30 360) (left half) vs that within (3 30 360) (right half) (d) *Nscore* within (3 10 129) (right half) vs that within (3 30 360) (right half).

Table 7. Sensitivity and accuracy for pair wise discrimination between topologies (3 10 129) and (3 30 360). Before : before exclusion of poorly structural aligned pairs of proteins. After : after exclusion of poorly structural aligned pairs of proteins.

	Sensitivity		
	(3 10 129)	(3 30 360)	Q
Before			
Number of Proteins	19	20	39
	0.58	0.40	0.48
After			
Number of Proteins	19	6	25
	0.76	0.83	0.76

within (3 30 360) are not (Fig. 1 (c)). Clearly, those within (3 30 360) are less than those within (3 10 129) (Fig. 1 (d)). Then we found that our method can have better Q by exclusion of pairs of proteins that have poor suructurally alignment. In Table 7, we have shown the comparison of sensitivity and accuracy Q before and after exclusion of badly aligned pairs of proteins. Although no proteins are removed from topology (3 10 129), 14 out of 20 proteins are excluded from (3 30 360). Then sensitivity and accuracy drastically increases. Especially, it is remarkable that sensitivity for (3 10 129) also increases although there are no proteins removed. Thus, it is clear that our method can discriminate topologies if they are well structurally aligned. In other words, we have to consider something other than structural alignment to discriminate topologies in CATH. It is very important to find what we should consider.

5 Conclusion

In this paper, we have investigated the relative performance of amino acid occurrence and other features to recognize protein folds. We found that consideration of other features than amino acid occurrence cannot improve accuracy Q so much. The reason is because amino acid occurrence can have ability to consider variety of physical properties which are useful to discriminate protein folds. It is conformed that our method can better discriminate topologies if proteins within each topology have good structural alignment. In order to improve accuracy Q , we have to consider something other than structural alignment.

Acknowledgement

This work has been supported by the Grant-in-Aid for Creative Scientific Research No.19500254 of the Ministry of Education, Culture, Sports, Science and Technology (MEXT) from 2007 to 2008. We are grateful for their support.

References

1. Shi, J., Blundell, T.L., Mizuguchi, K.: FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.* 310, 243–257 (2001)
2. Zhou, H., Zhou, Y.: Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins* 58, 321–328 (2005)
3. Cheng, J., Baldi, P.: A machine learning information retrieval approach to protein fold recognition. *Bioinformatics* 22, 1456–1463 (2006)
4. Gromiha, M.M., Suwa, M.: A Simple statistical method for discriminating outer membrane proteins with better accuracy. *Bioinformatics* 21, 961–968 (2005)
5. Hirokawa, T., Boon-Chieng, S., Mitaku, S.: SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics* 14, 378–379 (1998)
6. Chou, K.C.: Prediction of protein structural classes and subcellular locations *Curr. Protein Pept. Sci.* 1, 171–208 (2000)
7. Gromiha, M.M., Selvaraj, S., Thangakani, A.M.: A Statistical method for predicting protein unfolding rates from amino acid sequence. *J. Chem. Inf. Model* 46, 1503–1508 (2006)
8. Gromiha, M.M., Oobatake, M., Kono, H., Uedaira, H., Sarai, A.: Relationship between amino acid properties and protein stability: Buried Mutations. *J. Protein Chem.* 18, 565–578 (1999)
9. Ding, H.Q.D., Dubchak, I.: Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* 17, 349–358 (2001)
10. Taguchi, Y.-h., Gromiha, M.M.: Comparison of amino acid occurrence and composition for predicting protein folds. *IPSJ SIG Technical Report 2007-BIO-008*, pp. 9–16 (2007)
11. Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C.: SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536–540 (1995)
12. PROLDA: <http://www.granular.com/PROLDA/>
13. Pearl, F.M., Bennett, C.F., Bray, J.E., Harrison, A.P., Martin, N., Shepherd, A., Sillitoe, I., Thornton, J., Orengo, C.A.: The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Research* 31, 452–455 (2003)
14. Gubbi, J., Shilton, A., Parker, M., Palaniswami, M.: Protein Topology Classification Using Two-Stage Support Vector Machines. *Genome Informatics* 17, 259–269 (2006)
15. R: <http://www.R-project.org/>
16. Olmea, O., Valencia, A.: Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Folding Design* 2, S25–S32 (1997). Fariselli, P., Casadio, R.: A neural network based predictor of residue contacts in proteins. *Protein Eng.* 12, 15–21(1999)
17. Nick and Thomas' Protein Contact Prediction Server, <http://foo.acmc.uq.edu.au/nick/Protein/contact.html>
18. MacCallum, R.M.: Striped sheets and protein contact prediction. *Bioinformatics* 20(suppl. 1), I224–I231 (2004)
19. Gromiha, M.M., Oobatake, M., Sarai, A.: Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. *Biophysical Chemistry* 82, 51–67 (1999)

20. Gromiha, M.M., Oobatake, M., Kono, H., Uedaira, H., Sarai, A.: Importance of Mutant Position in Ramachandran Plot for Predicting Protein Stability of Surface Mutations. *Biopolymers* 64, 210–220 (2002)
21. Gromiha, M.M.: Importance of Native-state Topology for Determining the Folding Rate of Two-state Proteins. *J. Chem. Inf. Comp. Sci.* 43, 1481–1485 (2003)
22. Zeyar, A., Kian-Lee, T.: MatAlign: Precise Protein Structure Comparison by Matrix Alignment. *J. Bioinform. Comp. Biol.* 6, 1197–1216 (2006)