ELSEVIER

Biophysical Chemistry 88 (2000) 81–101

Biophysical Chemistry

www.elsevier.nl/locate/bpc

Structural class prediction: an application of residue distribution along the sequence

T.S. Kumarevel[a,*], M. Michael Gromiha[b], M.N Ponnuswamy[c]

[a] *Agency of Industrial Science and Technology, National Institute of Bioscience and Human Technology, Laboratory of Molecular Genetics, 1-1 Higashi, Tsukuba Science City, Ibaraki 305-8566, Japan*

[b] *Tsukuba Institute, The Institute of Physical and Chemical Research (RIKEN), 3-1-1 Koyadai, Tsukuba, Ibaraki 305-0074, Japan*

[c] *Department of Crystallography & Biophysics, University of Madras, Guindy campus, Chennai − 600 025, India*

Received 24 May 2000; received in revised form 16 August 2000; accepted 16 August 2000

Abstract

Deciphering the native conformation of proteins from their amino acid sequences is one of the most challenging problems in molecular biology. Information on the secondary structure of a protein can be helpful in understanding its native folded state. In our earlier work on molecular chaperones, we have analyzed the hydrophobic and charged patches, short-, medium- and long-range contacts and residue distributions along the sequence. In this article, we have made an attempt to predict the structural class of globular and chaperone proteins based on the information obtained from residue distributions. This method predicts the structural class with an accuracy of 93 and 96%, respectively, for the four- and three-state models in a training set of 120 globular proteins, and 90 and 96%, respectively, for a test set of 80 proteins. We have used this information and methodology to predict the structural classes of chaperones. Interestingly most of the chaperone proteins are predicted under α/β or mixed folding type.

Keywords: Structural class; Globular proteins; Chaperones; Amino acid sequence; Residue distributions

* Corresponding author. Tel.: +81-298-61-6085; Fax: +81-298-61-6095.
E-mail address: kumaravel@nibh.go.jp (T.S. Kumarevel).

0301-4622/00/$ - see front matter
PII: S0301-4622(00)00201-5

1. Introduction

The three-dimensional conformation of many polypeptides is determined in vitro by their amino acid sequences. It means that the polypeptides can fold in vitro without the help of additional components [1]. Using the same assumption in vivo, folding and assembly of polypeptides occurs by an essentially spontaneous process. Recently, this trend has been changed, because of the discovery of chaperones, which helps the nascent proteins to attain their correct folding and prevent misfolding or aggregation at the molecular level [2–8]. The various functions and molecular mechanisms of these molecular chaperones have been studied and reviewed by many authors [9–15].

Although the chaperones were discovered some 15 years ago, the crystal structures are known for only four chaperone proteins, namely, Papd, Hsc70, GroEL, GroES and a complex of GroEL–GroES [16–20]. In our earlier work, we have analyzed the solvent accessibility of the mutants of Hsc70 ATPase fragment, hydrophobic and charged patches, short-, medium- and long-range interactions, and residue distributions in chaperones [21–24]. The residue distributions derived from the sequence and the residue–residue contacts in crystal structures of globular and chaperone proteins show a very good relationship, and the scores are higher between the sequence and structures of chaperones than those of globular proteins. These results paved the way to predict the secondary structure of a chaperone/chaperonin protein. Information about the secondary structure of a protein can be helpful in understanding its native folded state. To improve the accuracy level of secondary structures, successful structural class prediction may be very useful.

Amino acid sequence is the prime source of study in understanding the structural and functional aspects of proteins. The native structures (3-D) of proteins are believed to be coded in their amino acid sequences (1-D), and this process is referred to as 'protein folding'. Formation of secondary structures is believed to be an important intermediate step in this process. Depending on the proportion of various secondary structure components, proteins of known structures have been classified into four classes, namely all-α, all-β, $\alpha+\beta$, and α/β [25,26]. Information about the structural classes may be helpful in improving the accuracy levels of secondary structure prediction.

Several methods have been proposed for predicting the structural classes. Klein's group [27,28] used the multidimensional statistical techniques of discrimination for this purpose. In some other methods, the minimum distance rule was adopted [29,30]. Cid et al. [31] used Ponnuswamy's [32] normalized hydrophobicity values to determine the structural classes. Zhang and Chou [33] proposed a method based on optimization principles and obtained 83% accuracy in the prediction of structural classes; this level increased to 88% when $\alpha+\beta$ and α/β classes were treated as one (mixed) class. The same authors [34] have again proposed another method for predicting the folding types of globular proteins by characterizing their crystal states as vectors of 20-dimensional space, each type of amino acid falling on one vector. This method is reported to predict the protein class in a set of 35 proteins at an accuracy level of 91%. Gromiha and Ponnuswamy [35] proposed a method based on the surrounding hydrophobicity profile and obtained an accuracy level of 92%. At the same time, Chou [36] used the principle of least Mahalanobis distance and predicted with an accuracy of 99 and 95% for training set and test set proteins, respectively.

In this article, we have developed a simple procedure to predict the structural classes based on the amino acid residue distributions along the sequence. We constructed four-state (all-α, all-β, $\alpha+\beta$, and α/β) and three-state [all-α, all-β and mixed (combination of $\alpha+\beta$ and α/β in one class)] models for this purpose. This method identifies the structural class for the training set proteins to an accuracy of 93% for the four-state and 96% for the three-state models. For the test set proteins it gives an accuracy of 90 and 96% for the four- and three-state level predictions, respectively. Most of the chaperones are predicted under the α/β or mixed class.

2. Materials and methods

2.1. Computation of amino acid composition in proteins

The amino acid composition and percentage of residues in a given protein is defined as follows.

The number of amino acid of type i in a given protein is calculated by:

$$X_i = \sum_{j=1}^{N} X_{ij} \tag{1}$$

where $X_{ij} = 1$ if amino acid type at sequence position j is i, and $X_{ij} = 0$ otherwise.

Percentage of amino acid residue is:

$$P_i = X_i * 100/N \tag{2}$$

where N is the total number of amino acid residues in a protein.

The amino acid composition of a protein is expressed as (P_A, P_D, P_C,…, P_Y), where A, D, C,…,Y stand for the corresponding amino acid single letter code.

2.2. Computing the neighboring amino acid residue distributions

The composition of neighboring residues associated with every residue for each protein is calculated for a central residue, with one residue at each side, starting from N- to C-terminal end. A 20×20 matrix is formed by summing all the neighboring residues belonging to similar types of residues found in the whole sample separately, and further classifying each sum into 20 different groups. Thus, each element of this matrix, denoted by $M_{r,i}$ can computed as:

$$M_{r,i} = \sum_{i=1}^{n_i} m_{r,i} \tag{3}$$

where $m_{r,i}$ is the observed frequency of rth residue neighbored by the ith residue. Here i and r vary from 1 to 20 and n_i is the total number of ith-type residues found in the protein. The % of each type of neighboring residue associated with the ith-type, $D_{r,i}$, is calculated by dividing each element of the ith column by the sum of all of the elements of the respective row and multiplying by 100. That is:

$$D_{r,i} = \frac{M_{r,i}}{\sum_{i=1}^{20} M_{r,i}} \times 100 \tag{4}$$

Using this equation., we computed the $D_{r,i}$ values for each of the structural classes separately.

2.3. Structural class prediction

The neighboring amino acid composition for all four categories (all-α, all-β, α + β, α/β) of proteins is considered as the basic norm to predict the structural classes. These can be represented in the form of a 20×20 matrix, which is shown below:

$$D_{\alpha} = \begin{pmatrix} X_{1,1} & X_{1,2} & X_{1,3} & \cdots & X_{1,20} \\ X_{2,1} & X_{2,2} & X_{2,3} & \cdots & X_{2,20} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ X_{20,1} & X_{20,2} & X_{20,3} & \cdots & X_{20,20} \end{pmatrix} \tag{5}$$

$$D_{\beta} = \begin{pmatrix} X'_{1,1} & X'_{1,2} & X'_{1,3} & \cdots & X'_{1,20} \\ X'_{2,1} & X'_{2,2} & X'_{2,3} & \cdots & X'_{2,20} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ X'_{20,1} & X'_{20,2} & X'_{20,3} & \cdots & X'_{20,20} \end{pmatrix} \tag{6}$$

$$D_{\alpha+\beta} = \begin{pmatrix} X''_{1,1} & X''_{1,2} & X''_{1,3} & \cdots & X''_{1,20} \\ X''_{2,1} & X''_{2,2} & X''_{2,3} & \cdots & X''_{2,20} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ X''_{20,1} & X''_{20,2} & X''_{20,3} & \cdots & X''_{20,20} \end{pmatrix} \tag{7}$$

$$D_{\alpha/\beta} = \begin{pmatrix} X'''_{1,1} & X'''_{1,2} & X'''_{1,3} & \dots & X'''_{1,20} \\ X'''_{2,1} & X'''_{2,2} & X'''_{2,3} & \dots & X'''_{2,20} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ X'''_{20,1} & X'''_{20,2} & X'''_{20,3} & \dots & X'''_{20,20} \end{pmatrix} \quad (8)$$

The difference matrices A_{α}, A_{β}, $A_{\alpha+\beta}$, and $A_{\alpha/\beta}$ can be defined as the difference between the basic norms obtained (D_{α}, D_{β}, $D_{\alpha+\beta}$, and $D_{\alpha/\beta)}$ and the new (unknown) protein Y. As an example, A_{α} can be derived as:

$$A_{\alpha} = D_{\alpha} - \begin{pmatrix} Y_{1,1} & Y_{1,2} & Y_{1,3} & \dots & Y_{1,20} \\ Y_{2,1} & Y_{2,2} & Y_{2,3} & \dots & Y_{2,20} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ Y_{20,1} & Y_{20,2} & Y_{20,3} \dots & \dots & Y_{20,20} \end{pmatrix} \quad (9)$$

The standard deviation (S.D.) for all-α proteins can be computed using the formula:

$$\text{S.D.}_{\alpha} = \sqrt{\Sigma A_{\alpha}^2/N - (\Sigma A_{\alpha}/N)^2} \quad (10)$$

We have used Eq. (10) to compute the S.D. values for each of the structural classes.

The protein Y would be predicted to belong to the structural class for which the deviation is the lowest.

2.4. Database

2.4.1. Globular proteins

We have used Chou's criteria [36] to select the proteins from the database. It states:

- $\alpha \Rightarrow \alpha \geq 40\%$ $\beta \leq 5\%$;
- $\beta \Rightarrow \alpha \leq 5\%$ $\beta \geq 40\%$;
- $\alpha + \beta \Rightarrow \alpha \geq 15\%$ $\beta \geq 15\%$ with more than 60% anti-parallel β-sheets; and
- $\alpha/\beta \Rightarrow \alpha \geq 15\%$ $\beta \geq 15\%$ with more than 60% parallel β-sheets.

The data set that we used is the same as that of Chou [36] and it contains 30 representative proteins for each of the four structural classes ($4 \times 30 = 120$), and a set of 80 proteins for the test set. The data for the three-dimensional structures have been taken from the recent release of the Brookhaven Protein Data Bank [37,38].

2.4.2. Training set

2.4.2.1. 30 all-α proteins. 1AVHA, 1BABB, 1BRD, 1C5A, 1CPCA, 1CPCL, 1ECO, 1FCS, 1FHA, 1FIAB, 1HBG, 1HDDC, 1HIGA, 1LE4, 1LIG, 1LTSC, 1MBC, 1MBS, 1RPRA, 1TROA, 1UTG, 256BA, 2CCYA, 2LH1, 2LHB, 2MHBA, 2MHBB, 2ZTAA, 4MBA, 4MBN.

2.4.2.2. 30 all-β proteins. 1ACX, 1AYH, 1CD8, 1CDTA, 1CID, 1DFNA, 1HILA, 1HIVA, 1HLEB, 1MAMH, 1MONA, 1OMF, 1PHY, 1REIA, 1TEN, 1TLK, 1VAAB, 2ALP, 2AVIA, 2BPA2, 2HHRC, 2ILA, 2LALA, 2SNV, 3CD4A, 4GCR, 7APIB, 8I1B, 8FABA, 8FABB.

2.4.2.3. 30 α + β proteins. 1AAK, 1CTF, 1DNKA, 1EAF, 1HSBA, 1LTSA, 1LTSD, 1NRCA, 1OVB, 1POC, 1PPN, 1PRF, 1RND, 1SNC, 1TFG, 1TGSI, 2ACHA, 2ACT, 2BPA1, 2SNS, 2SSI, 3IL8, 3RUBS, 3SGBI, 3SICI, 4BLMA, 4TMS, 8CATA, 9RNT, 9RSAA.

2.4.2.4. 30 α/β proteins. 1ABA, 1CIS, 1CSEI, 1CTC, 1DHR, 1DRI, 1ETU, 1FX1, 1GPB, 1OFV, 1PAX, 1PFKA, 1PGD, 1Q21, 1SO1, 1SBP, 1SBT, 1TIMA, 1TMD, TREA, 1ULA, 1WSYB, 2HAD, 2LIV, 3GBP, 4FXN, 4CPA, 5P21, 8ABP, 8ATCA.

2.4.3. Test set

We have collected 80 unbiased, non-homologous proteins for the test set, which is not included in the training data set. The PDB codes are 1HBBA, 1IFA, 2PDE, 2HCOA, 1PPT, 1BBL, 1GCN, 1MLT, 1PP2, 451C, 1PPFE, 1RIA2, 1SHFA, 1TIE, 2ACHB, 2CTX, 2MEV1, 2SODO, 4CHA, 3EBX, 1PCY, 1HOE, 2FB4, 2APR, 3CAN, 3EST, 3COX, 3INS, 2LZM, 2CDV, 1DNKA, 1GLAG, 2MS2A, 1POC, 1PPBA, 1SHAA, 2AAA, 2PIA, 2SN3, 2TAAA, 4ENL, 1FDX, 2PAD, 1HIP,

3GAP, 2PAZ, 4FD1, 3RXN, 1HLA, 1HMG, 2CI2, 5PTI, 1SNS, 1PYP, 1SGT, 9PAP, 2CAB, 1PHH, 3GRS, 2CPP, 1GBP, 1MINA, 1NIPB, 1SBP, 1WSYA, 4ICD, 9RUBB, 1GD1O, 3ADK, 2GPD, 3GPK, 3LDH, 1SRX, 1ABP, 2LBP, 1CSE, 1GP1, 1GOX, 5ADH, 3PGK.

2.4.4. Chaperone sequences

The sequence information for a set of 91 chaperone proteins is obtained from the SWISS-PROT [39] database. The codes are CH60_ACYPS, CH60_AGRTU, CH60_AMOPS, CH60_ARATH, CH60_BACSU, CH60_BRUAB, CH60_CHLPN, CH60_CHLTR, CH60_CHRVI, CH60_CLOAB, CH60_CLOPE, CH60_COXBU, CH60_CYACA, CH60_ECOLI, CH60_HAEDU, CH60_LEGMI, CH60_LEGPN, CH60_MAIZE, CH60_MYCLE, CH60_MYCTU, CH60_PESAE, CH60_RICTS, CH60_SYNP7, CH60_SYNY3, CH60_THEP3, CH61_STRAL, CH62_STRAL, CH63_HELVI, CH60_BACST, CH60_BARBA, CH60_BRANA, CH60_LACLA, CH60_LEPIN, CH60_PLAFG, CH60_RHILV, CH60_STAAU, CH61_CUCMA, CH61_MYCLE, CH61_RHIME, CH61_STRCO, CH61_SYNY3, CH62_BRAJA, CH62_CUCMA, CH62_RHIME, CH63_BRAJA, CH63_RHIME, CH61_ECOLI, HS7A_YEAST, HSCB_ECOLI, HS82_BASCU, HSP7_YEAST, HSLU_BACSU, HSCA_ECOLI, HS9B_RAT, HS82_TOBAC, HS82_MAIZE, HS82_ASPFU, HS80_LYCES, DNAK_STRCO, DNAK_BRUOV, TCPZ_YEAST, TCPZ_MOUSE, TCPZ_HUMAN, TCPH_MOUSE, TCPG_YEAST, TCPG_MOUSE, TCPE_YEAST, TCPE_MOUSE, TCPE_AVESA, TCPD_YEAST, TF55_SULSH, TCPB_MOUSE, TCP1_YEAST, TCP1_RAT, TCP1_HUMAN, TCP1_DROME, TCP1_CRIGR, TCP1_ARATH, TCPD_MOUSE, TCPB_YEAST,

Table 1
Amino acid composition for the four structural classes of globular proteins and chaperones

Amino Acids	All α (%)	N[a]	All β (%)	N[a]	α + β (%)	N[a]	α/β (%)	N[a]	Globular (%)	N[a]	Chaperone (%)	N[a]
Ala	11.72	479	6.28	287	8.31	448	9.65	834	9.02	2048	10.79	5393
Asp	5.70	233	5.10	233	6.27	338	6.32	546	5.95	1350	6.41	3204
Cys	0.71	29	1.99	91	2.11	114	1.03	89	1.42	323	0.71	353
Glu	6.90	282	5.29	242	5.56	300	6.18	534	5.98	1358	8.34	4170
Phe	4.23	173	4.40	201	3.78	204	3.68	318	3.95	896	1.95	973
Gly	6.63	271	8.49	388	6.80	367	8.76	757	7.86	1783	8.45	4225
His	3.01	123	1.33	61	2.56	138	2.27	196	2.28	518	0.93	467
Ile	4.06	166	4.22	193	4.58	247	6.13	530	5.01	1136	7.08	3540
Lys	8.05	329	5.71	261	6.19	334	6.21	537	6.44	1461	8.01	4001
Leu	10.94	447	7.24	331	7.77	419	7.88	681	8.27	1878	8.71	4354
Met	2.52	103	1.57	72	1.93	104	2.29	198	2.10	477	2.76	1377
Asn	3.52	144	5.10	233	4.67	252	4.19	362	4.37	991	3.83	1914
Pro	2.86	117	4.72	216	5.06	273	4.08	353	4.23	959	2.86	1430
Gln	4.18	171	4.48	205	4.02	217	3.92	339	4.11	932	3.13	1562
Arg	4.16	170	3.94	180	4.75	256	4.27	369	4.30	975	4.56	2280
Ser	5.97	244	8.71	398	6.73	363	5.60	484	6.56	1489	5.08	2539
Thr	4.72	193	8.11	371	6.21	335	5.25	454	5.96	1353	5.73	2863
Val	6.19	253	7.02	321	6.58	355	7.49	647	6.94	1576	8.97	4481
Trp	1.10	45	1.75	80	1.46	79	1.32	114	1.40	318	0.20	98
Tyr	2.81	115	4.55	208	4.65	251	3.47	300	3.85	874	1.50	752
		4087		4572		5394		8642		22695		49976

[a] N is the number of residues.

TCPB_MOUSE, P60_RAT, P60_MOUSE, P60_HUMAN, P60_GRIGR, RUBB_BRANA, RUBB_ARATH, RUBA_RICCO, RUBA_WHEAT, RUBB_BRANA, RUBA_ARATH, RUB2_BRANA.

3. Results and discussion

3.1. Amino acid composition in globular proteins and chaperones

The amino acid composition of all-α, all-β, α + β and α/β, and 91 chaperone proteins are given Table 1.

In all-α proteins, Ala ranks as the most frequently occurring, followed by Leu, Lys, Glu, His and Met; Cys and Trp are found to be the least. In all-β proteins, Ser seems to be the highest occurring residue, followed by Thr, Val, Gly, Leu and Ala. The least occurring residues are His, Met and Trp. In α + β proteins, Ala appears to be the most abundant, followed by Leu, Gly, Ser and Val; Trp and Met residues are the least. The residues Ala, Gly, Leu, Val are the most and Cys, Trp are the least occurring residues in α/β proteins. When all the four different classes are compared, the residues Ala, Leu, Lys and Glu stand out as higher in all-α proteins, and Ser, Thr and Asn are the most abundant in all-β proteins. The residues Pro and Arg, and the residues Ile, Gly and Val are at higher levels in α + β- and α/β-type proteins, respectively. In general, the residues Leu, Gly, Ala, Ser, Thr, and Val are most abundant, whereas Cys, Trp, Met, and His, respectively are least abundant.

The overall amino acid composition of chaperone and globular proteins are displayed in Fig. 1. The residues Ala, Glu, Ile, Lys, Val and Gly seem to have a higher % of occurrence in chaperones, while Phe, His, Trp, Tyr and Ser play that role for globular proteins. We have correlated the amino acid composition of the four structural classes with chaperone proteins, and the scores obtained are 0.968, 0.927, 0.956 and 0.982, respectively. Interestingly, most of the chaperone/chaperonin proteins fall into the α/β structural class.

3.2. Neighboring amino acid composition of globular proteins

The percentage of neighboring amino acid composition with one residue at each side of the

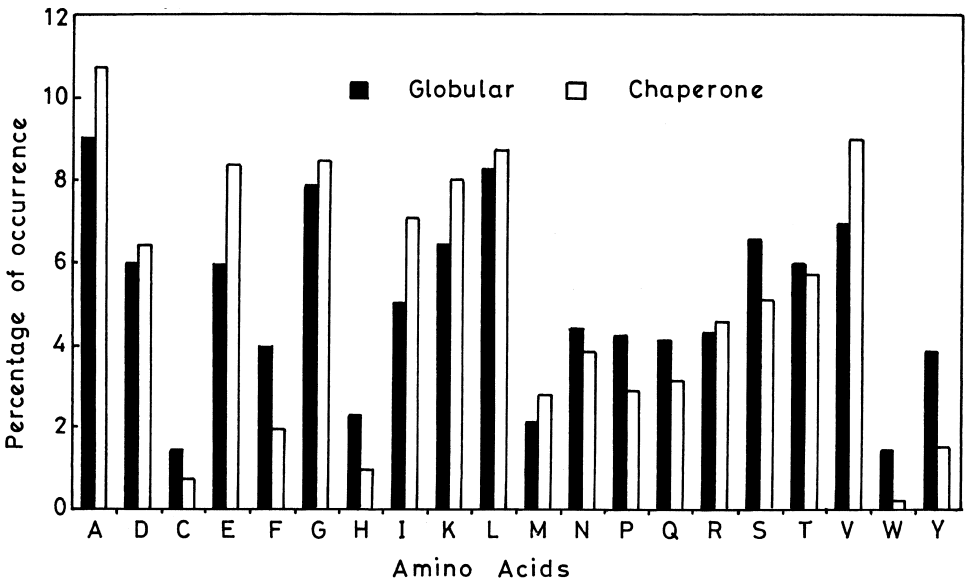

Fig. 1. Percentage amino acid composition in globular and chaperone proteins.

Table 2a
Neighboring amino acid compositions for all the residue types in all-α proteins

	Ala	Asp	Cys	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr	
Ala	14.9	7.61	0.42	7.51	2.64	7.4	3.07	4.02	6.77	10.25	2.75	3.91	2.33	5.39	2.64	5.81	2.96	5.5	1.59	2.54	946
Asp	15.87	8.48	1.52	6.09	5.65	5.87	1.3	5.0	6.3	9.78	2.61	3.48	4.57	3.26	3.48	3.91	4.13	5.65	1.09	1.96	460
Cys	6.9	12.07	6.9	5.17	0	1.72	8.62	6.9	8.62	15.52	1.72	1.72	1.72	1.72	1.72	1.72	3.45	8.62	0	5.17	58
Glu	12.59	4.96	0.53	9.57	4.26	5.14	1.77	2.48	7.45	13.65	1.95	4.26	3.72	3.55	2.84	6.74	6.38	4.43	1.6	2.13	564
Phe	7.23	7.51	0	6.94	5.78	10.69	1.45	4.34	10.98	10.12	1.16	4.62	2.89	2.31	7.23	4.34	4.62	3.47	0.58	3.76	346
Gly	13.71	5.02	0.19	5.6	7.14	4.25	4.63	3.67	7.92	10.04	2.32	4.44	2.7	2.9	3.86	4.83	5.98	7.34	1.16	2.32	518
His	11.98	2.48	2.07	4.13	2.07	10.33	4.96	2.48	11.57	14.46	0	0.41	4.55	2.07	3.72	9.09	3.31	7.44	0.83	2.07	242
Ile	11.52	6.97	1.21	4.24	4.55	6.06	1.82	5.45	9.09	10.3	2.42	3.94	4.24	5.15	5.45	5.15	4.85	4.55	0.91	2.12	330
Lys	9.73	4.41	0.76	6.38	5.78	6.23	4.26	4.56	9.12	11.09	4.1	3.34	2.58	2.28	3.8	5.17	3.95	7.6	0.76	4.1	658
Leu	10.9	5.06	1.01	8.65	3.93	6.18	3.93	3.82	8.2	8.54	2.36	3.37	2.58	3.03	5.28	6.07	5.84	7.3	0.9	3.03	890
Met	13.19	6.04	0	6.04	1.65	6.04	0	4.4	13.74	10.44	3.3	6.04	0	4.95	5.49	7.14	3.3	6.59	0	1.65	182
Asn	12.94	5.59	0.35	8.39	5.59	8.04	0.35	4.55	7.69	10.84	3.5	2.1	3.15	5.24	4.55	4.9	2.45	7.34	0.35	2.1	286
Pro	9.48	9.05	0.43	9.05	4.31	6.03	4.74	5.6	7.33	9.91	0	3.88	2.59	3.02	2.59	5.17	7.33	6.03	0.86	2.59	232
Gln	15.38	4.44	0.3	5.92	2.37	5.62	1.48	5.03	4.44	7.99	2.66	4.44	2.07	3.85	6.8	9.47	5.03	6.8	1.48	4.44	338
Arg	7.58	4.85	0.3	4.55	7.58	6.36	2.73	5.76	7.58	14.24	2.73	3.94	1.82	6.97	2.42	7.58	4.24	5.15	0.91	2.73	330
Ser	11.55	3.99	0.21	7.77	3.15	5.25	4.62	3.57	6.93	11.13	2.73	2.94	2.52	6.51	5.25	5.88	6.72	5.04	1.26	2.94	476
Thr	7.25	4.92	0.52	9.33	4.15	8.03	2.07	4.15	6.74	13.47	2.07	1.81	4.4	4.4	3.63	8.55	4.66	6.22	1.55	2.07	386
Val	10.73	5.26	1.01	5.06	2.43	7.69	3.44	3.04	10.12	12.35	2.43	4.25	2.83	4.45	3.44	4.86	4.86	7.29	1.62	2.83	494
Trp	16.67	5.56	0	10	2.22	6.67	2.22	3.33	5.56	8.89	0	1.11	2.22	5.56	3.33	6.67	6.67	8.89	2.22	2.22	90
Tyr	10.09	3.95	1.32	5.26	5.7	5.26	3.07	3.07	11.84	11.84	1.32	2.63	2.63	6.58	5.26	6.14	3.51	6.14	0.88	3.51	228
	947	458	57	562	345	527	242	329	655	881	192	286	233	331	333	480	383	499	90	224	8054

Table 2b
Neighboring amino acid compositions for all the residue types in all-β proteins

	Ala	Asp	Cys	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr	
Ala	9.32	6.45	1.79	5.56	3.41	7.53	1.08	2.69	6.09	6.27	0.54	5.91	5.73	3.76	3.76	8.06	9.14	7.17	1.61	4.12	558
Asp	7.83	5.65	1.52	6.52	5	9.57	0.65	6.09	5	3.91	3.2	3.26	4.13	3.48	4.78	7.61	7.61	6.52	2.17	5.43	460
Cys	5.62	4.49	1.69	3.93	2.81	8.43	2.81	5.62	4.49	5.06	0.56	5.06	3.37	6.18	5.06	10.11	11.8	5.62	1.12	6.18	178
Glu	6.78	6.36	1.48	6.78	4.03	10.81	1.06	4.03	5.3	5.3	2.33	4.03	5.08	3.6	3.6	6.14	5.51	8.26	2.54	6.99	472
Phe	4.77	5.78	1.26	4.77	5.03	7.79	1.26	3.77	4.27	5.78	1.26	6.53	3.52	3.77	5.03	9.8	13.07	6.03	0.75	5.78	398
Gly	5.47	5.73	1.95	6.64	3.91	9.64	1.56	3.65	5.86	7.16	1.04	5.99	3.78	4.56	5.47	9.9	6.64	6.12	0.52	4.43	768
His	4.92	2.46	4.1	4.1	4.1	9.84	4.92	3.28	5.74	8.2	1.64	4.1	5.74	2.46	5.74	4.92	9.02	9.02	3.28	2.46	122
Ile	3.91	7.81	2.6	4.95	3.91	7.29	1.04	4.17	6.77	7.03	1.3	4.95	6.25	5.73	3.13	8.59	9.64	5.47	1.56	3.91	384
Lys	6.78	4.46	1.55	5.04	3.29	8.91	1.36	5.04	4.07	7.36	2.52	6.78	6.2	4.07	2.13	6.78	8.14	9.3	3.1	3.1	516
Leu	5.3	2.73	1.36	3.94	3.48	8.33	1.52	4.09	5.76	6.36	0.61	6.21	4.55	5.61	5.61	8.64	12.27	7.58	1.36	4.7	660
Met	2.14	10.71	0.71	7.14	2.86	5.71	1.43	3.57	9.29	2.86	4.29	5	1.43	5	3.57	10.0	7.86	9.29	4.29	2.86	140
Asn	7.39	3.26	1.74	4.13	5.87	10	1.09	4.13	7.61	8.91	1.52	3.91	3.26	4.78	2.17	11.96	4.35	6.74	2.39	4.78	460
Pro	8.02	4.48	1.42	5.9	3.3	7.08	1.65	5.66	7.55	6.84	0.47	3.77	5.42	4.95	4.25	11.79	6.84	6.84	0.24	3.54	424
Gln	5.39	3.92	2.7	4.17	3.92	8.58	0.74	5.39	5.39	9.07	1.72	5.39	5.39	4.41	3.19	7.11	7.35	6.62	2.21	7.35	408
Arg	6.25	6.25	2.56	4.55	5.68	11.93	1.99	3.41	3.13	9.94	1.42	2.56	5.11	3.69	5.68	7.39	6.82	6.82	1.7	3.13	352
Ser	5.74	4.46	2.3	3.7	4.85	9.69	0.77	4.21	4.46	7.27	1.79	7.14	6.51	3.57	3.32	11.35	6.76	8.04	1.15	2.93	784
Thr	6.95	4.77	2.86	3.41	7.08	7.08	1.5	4.9	5.72	11.04	1.5	2.72	3.95	4.22	3.27	7.08	5.31	9.4	2.45	4.77	734
Val	6.45	4.72	1.57	6.29	3.77	7.39	1.73	3.3	7.55	7.86	2.04	4.87	4.4	4.09	3.62	9.91	10.85	5.03	1.57	2.99	636
Trp	5.77	6.41	1.28	7.05	1.92	2.56	2.56	3.85	10.26	5.77	3.85	7.05	0.64	5.77	3.21	5.77	11.54	6.41	0	8.33	156
Tyr	5.56	6.04	2.66	7.97	5.31	8.21	0.72	3.62	3.86	7.49	0.97	5.31	3.62	7.25	2.66	5.8	8.45	4.59	3.14	6.76	414
	566	463	176	471	396	772	122	381	514	656	142	460	421	403	353	784	735	637	158	414	9024

Table 2c
Neighboring amino acid compositions for all the residue types in α + β proteins

	Ala	Asp	Cys	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr	
Ala	11.39	6.61	2.39	5.24	5.01	6.61	1.82	4.1	6.04	8.09	1.14	4.21	5.24	3.19	2.96	7.74	4.9	8.43	1.37	3.53	878
Asp	8.48	4.46	2.53	4.32	5.21	9.67	2.08	4.61	5.65	7.14	1.49	3.13	5.95	4.02	5.36	5.21	6.99	7.59	1.49	4.61	672
Cys	9.82	7.59	3.57	5.36	2.23	10.71	2.23	1.79	4.91	8.04	2.23	6.25	4.91	2.68	4.91	6.7	5.8	4.02	1.79	4.46	224
Glu	7.67	4.83	2.0	5.67	3	6.5	2.83	4.33	6.5	6.67	1.83	5.67	6.17	5.67	5.5	5.83	6.5	7.67	1.5	3.67	600
Phe	10.45	8.71	1.24	4.48	1.99	6.72	2.24	5.97	6.72	4.73	1.99	4.73	3.98	4.98	6.22	7.46	6.47	5.47	0.75	4.73	402
Gly	7.95	8.9	3.15	5.34	3.84	5.21	2.33	3.97	6.16	6.3	0.55	3.7	4.25	4.52	4.93	7.81	7.26	6.16	1.23	6.44	730
His	5.84	5.11	1.82	6.2	3.28	6.2	2.92	5.11	6.57	10.22	1.46	1.82	6.2	5.47	5.47	5.11	10.95	5.47	1.09	3.65	274
Ile	7.58	6.35	0.82	5.33	4.92	5.94	2.87	3.89	7.58	8.61	1.02	5.33	4.1	4.1	4.71	5.74	7.17	6.35	2.25	5.33	488
Lys	8.05	5.78	1.82	5.62	4.1	6.69	2.74	5.62	9.73	8.05	3.04	5.02	4.56	2.89	3.95	5.02	5.62	5.47	1.52	4.71	658
Leu	8.6	5.81	2.18	4.84	2.3	5.57	3.39	5.08	6.3	9.2	1.57	5.08	3.63	4.48	5.69	7.63	6.9	6.3	0.97	4.48	826
Met	4.5	5.0	2.5	5.5	4	2	2	2.5	10	6	6	4	7	6	6	6	8	8.5	1.5	3	200
Asn	7.46	4.23	2.82	6.85	3.83	5.44	1.01	5.24	6.05	8.47	1.61	5.44	7.26	3.23	5.24	6.25	5.85	6.45	1.61	5.65	496
Pro	8.61	7.33	2.2	6.78	2.93	5.68	3.3	4.03	5.49	5.86	2.56	6.59	5.13	3.3	3.66	5.68	4.03	8.61	1.65	6.59	546
Gln	6.51	6.28	1.4	7.91	4.65	7.21	3.49	4.65	4.42	8.6	3.02	3.72	4.19	1.86	6.51	6.98	7.21	5.58	3.02	2.79	430
Arg	5.1	7.06	2.16	6.47	4.9	7.06	2.94	4.51	5.1	9.02	2.35	5.1	3.92	5.49	2.35	6.47	5.88	6.86	1.37	5.88	510
Ser	9.41	4.92	1.97	4.92	4.21	8.15	1.97	3.93	4.63	8.85	1.83	4.07	4.35	4.21	4.63	8.43	6.46	5.76	1.83	5.48	q712
Thr	6.78	7.08	1.81	5.87	3.92	7.98	4.52	5.27	5.72	8.43	2.41	4.37	3.31	4.67	4.52	6.93	5.12	6.02	1.2	4.07	664
Val	10.37	7.24	1.28	6.53	3.13	6.39	2.13	4.4	5.26	7.39	2.41	4.55	6.68	3.41	4.97	5.54	5.68	5.97	1.7	4.97	704
Trp	7.59	6.33	2.53	5.7	1.9	5.7	1.9	6.96	6.33	5.06	1.9	5.06	5.7	8.23	4.43	8.86	5.06	7.59	0	3.16	158
Tyr	6.05	6.25	2.02	4.44	3.83	9.27	2.02	5.24	6.05	7.46	1.21	5.65	7.26	2.42	6.05	8.06	5.44	7.06	1.01	3.23	496
	876	673	222	598	405	727	275	489	657	826	204	497	539	431	511	714	663	706	157	498	10668

Table 2d
Neighboring amino acid compositions for all the residue types in α/β proteins

	Ala	Asp	Cys	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr	
Ala	9.98	5.9	1.03	5.66	2.98	9.67	1.95	7.06	7.79	8.7	2.62	3.95	3.1	4.99	3.77	5.35	4.56	7.24	1.28	2.43	1644
Asp	8.97	5.31	1.19	5.22	3.66	7.88	1.74	6.68	6.32	9.07	2.2	4.03	4.58	4.03	4.4	6.32	4.58	8.06	2.11	3.66	1092
Cys	9.55	7.3	1.12	2.25	3.37	11.24	1.12	8.99	5.62	10.67	0.56	5.06	5.06	2.25	5.06	5.06	5.06	8.43	0	2.25	178
Glu	8.77	5.28	0.38	6.79	4.72	7.83	1.32	6.42	5.94	8.3	2.17	3.49	5.57	3.77	5	4.72	6.23	7.17	1.79	4.34	1060
Phe	7.73	6.31	0.95	7.89	3.79	9.94	2.37	3.47	7.1	6.78	1.89	4.26	3.79	4.26	3.31	5.99	5.99	8.52	0.79	4.89	634
Gly	10.54	5.7	1.33	5.5	4.18	8.62	2.32	6.5	5.57	6.7	2.06	3.25	4.18	3.18	4.44	6.3	5.5	7.49	1.26	5.37	1508
His	8.25	4.9	0.52	3.61	3.87	9.02	1.55	5.93	6.19	7.22	2.32	3.09	7.73	3.09	5.15	8.25	5.67	8.51	1.29	3.87	388
Ile	10.91	6.93	1.52	6.45	2.09	9.3	2.18	6.83	5.5	7.21	2.09	4.27	3.51	4.84	3.7	5.31	5.6	7.21	1.52	3.04	1054
Lys	11.98	6.32	0.94	5.94	4.15	7.83	2.36	5.47	5.94	9.34	2.17	3.96	3.3	3.77	4.25	5.09	4.81	7.83	1.13	3.4	1060
Leu	10.55	7.3	1.40	6.49	3.17	7.45	2.06	5.68	7.23	8.55	1.92	4.5	3.24	3.54	4.42	4.87	5.53	8.19	1.18	2.73	1356
Met	11.29	6.45	0.27	5.91	2.96	8.33	2.42	5.38	5.65	6.99	2.15	5.65	6.18	2.15	6.45	4.57	2.96	9.68	1.34	3.23	372
Asn	9.31	6.11	1.25	5.42	3.75	6.81	1.67	6.39	5.83	8.47	2.92	3.89	5.69	5.97	4.31	3.89	5.28	7.08	2.5	3.47	720
Pro	7.39	7.1	1.28	8.38	3.41	8.95	4.26	5.11	4.97	6.25	3.41	5.82	0.85	3.84	3.98	6.11	6.25	7.39	0.99	4.26	704
Gln	12.5	6.55	0.6	5.95	4.02	7.14	1.93	7.59	5.8	7.14	1.19	6.4	4.02	2.68	5.21	5.06	5.21	6.85	0.3	3.87	672
Arg	8.83	6.52	1.22	7.2	2.85	9.1	2.72	5.3	5.98	8.15	3.4	4.21	3.8	4.76	3.53	4.89	5.43	7.07	1.63	3.4	736
Ser	9.04	7.17	0.94	5.2	3.95	9.88	3.33	5.82	5.51	6.86	1.77	2.91	4.47	3.53	3.64	8.52	5.51	6.34	1.35	4.26	962
Thr	8.28	5.52	0.99	7.28	4.19	9.16	2.43	6.51	5.63	8.28	1.55	4.19	4.86	3.86	4.42	5.85	4.42	7.51	1.88	3.2	906
Val	9.21	6.81	1.16	5.88	4.18	8.9	2.55	5.88	6.35	8.67	2.79	3.95	4.02	3.56	4.02	4.72	5.26	9.13	0.93	2.01	1292
Trp	9.21	10.09	0	8.33	2.19	8.33	2.19	7.02	5.26	7.02	2.19	7.89	3.07	0.88	5.26	5.7	7.46	5.26	0.88	1.75	228
Tyr	6.69	6.69	0.67	7.69	5.35	13.55	2.51	5.35	6.02	6.02	2.01	4.18	5.02	4.35	4.18	6.86	4.85	4.35	0.67	3.01	598
	1649	1088	178	1062	633	1509	390	1054	1057	1356	384	715	703	670	732	965	903	1290	228	598	17164

central residue is calculated as described in Section 2 (Eqs. (5)–(8)). The calculated values for the 30 representative proteins in each of the structural classes, all-α, all-β, α + β and α/β, are given in Table 2a–d. The column in each table represents different kinds of neighboring residues associated with the same type, and the row shows the same kind of neighboring residues associated

Table 3
The most (M) and least (L) neighboring residues around all the residue types observed in the different classes of globular proteins

	Most preferred (M)					Least preferred (L)				
(a) All α-proteins										
Ala	A	L	D	E	G	C	W	P	Y	F
Asp	A	L	D	K	E	W	H	C	Y	M
Cys	L	D	K	H	V	F	W	N	M	G
Glu	L	A	E	K	S	C	W	H	M	Y
Phe	K	G	L	D	A	C	W	M	H	Q
Gly	A	L	K	V	F	C	W	Y	M	P
His	L	A	K	G	S	M	N	W	Y	C
Ile	A	L	K	D	G	W	C	H	Y	M
Lys	L	A	K	V	E	C	W	Q	P	N
Leu	A	E	L	K	V	W	C	M	P	Y
Met	K	A	L	S	V	P	W	C	H	Y
Asn	A	L	E	G	K	H	W	C	Y	N
Pro	L	A	E	D	K	M	C	W	Y	P
Gln	A	S	L	R	V	C	H	W	P	F
Arg	L	F	K	A	S	C	W	P	R	Y
Ser	A	L	E	K	T	C	W	P	M	Y
Thr	L	E	S	G	A	C	W	N	Y	M
Val	L	A	K	G	V	C	W	M	F	Y
Trp	A	E	L	V	S	C	M	N	Y	W
Tyr	K	L	A	Q	S	W	M	C	P	N
(b). All β-proteins										
Ala	A	T	S	G	V	M	H	W	C	I
Asp	G	A	S	T	E	H	C	W	M	N
Cys	T	S	G	Q	Y	M	W	C	F	H
Glu	G	V	Y	E	A	H	C	M	W	R
Phe	T	S	G	N	V	W	M	C	H	P
Gly	S	G	L	E	T	W	M	H	C	I
His	G	T	V	L	R	M	Y	D	Q	W
Ile	T	S	D	G	L	H	M	W	C	R
Lys	V	G	T	L	S	H	C	R	M	Y
Leu	T	S	G	V	L	M	C	W	H	D
Met	D	S	K	V	T	C	H	P	A	Y
Asn	S	G	L	K	A	H	M	C	R	W
Pro	S	A	K	G	L	W	M	C	H	F
Gln	L	G	T	Y	S	H	M	W	C	R
Arg	G	L	S	T	V	M	W	H	C	N
Ser	S	G	V	L	N	H	W	M	C	Y
Thr	L	V	F	S	G	H	M	W	N	C
Val	T	S	L	K	G	C	W	H	M	Y
Trp	T	K	Y	E	N	W	P	C	F	G
Tyr	T	G	E	L	Q	W	R	C	M	H

Table 3 (*Continued*)

	Most preferred (M)					Least preferred (L)				
(c) α + β -proteins										
Ala	A	V	L	S	G	M	W	H	C	R
Asp	G	A	V	L	T	M	W	H	C	N
Cys	G	A	L	D	S	I	W	F	H	M
Glu	A	V	L	G	T	W	M	C	H	F
Phe	A	D	S	K	G	W	C	M	F	H
Gly	D	A	S	T	Y	M	W	H	C	N
His	T	L	K	G	P	W	M	C	N	H
Ile	L	K	A	T	D	C	M	W	H	I
Lys	K	A	L	G	D	W	C	H	Q	M
Leu	L	A	S	T	K	W	M	C	F	H
Met	K	V	T	P	Q	W	H	G	C	I
Asn	L	A	P	E	V	H	M	W	C	Q
Pro	A	V	D	E	N	W	C	M	F	H
Gln	L	E	G	T	S	C	Q	Y	W	M
Arg	L	G	D	V	S	W	C	R	M	H
Ser	A	L	S	G	T	M	W	C	H	I
Thr	L	G	D	S	A	W	C	M	P	F
Val	A	L	D	P	E	C	W	H	M	F
Trp	S	Q	A	V	I	W	M	F	H	C
Tyr	G	S	L	P	V	W	M	C	H	Q
(d) α / β-proteins										
Ala	A	G	L	K	V	C	W	H	Y	M
Asp	L	A	V	G	I	C	H	W	M	Y
Cys	G	L	A	I	V	W	M	C	H	Y
Glu	A	L	G	V	E	C	H	W	M	N
Phe	G	V	E	A	K	W	C	M	H	R
Gly	A	G	V	L	I	W	C	M	H	Q
His	G	V	S	A	P	C	W	H	M	N
Ile	A	G	L	V	D	C	W	M	F	H
Lys	A	L	G	V	D	C	W	M	H	P
Leu	A	L	V	G	D	W	C	M	H	Y
Met	A	V	G	L	R	C	W	M	Q	H
Asn	A	L	V	G	I	C	H	W	M	Y
Pro	G	E	A	V	D	P	W	C	F	M
Gln	A	I	G	L	V	W	C	M	H	Q
Arg	G	A	L	E	V	C	W	H	F	Y
Ser	G	A	S	D	L	C	W	M	N	H
Thr	G	A	L	V	E	C	M	W	H	Y
Val	A	V	G	L	D	W	C	Y	H	M
Trp	D	A	E	G	N	C	Q	W	Y	M
Tyr	G	E	S	A	D	C	W	M	H	Y

with different types of residues. The sum of each row and column are listed at the end of every row and column, respectively.

3.2.1. All-α proteins

In all-α proteins, Ala is predominantly neighbored by the residues Ala, Leu, Asp, Glu and Gly, and least neighbored by Cys, Trp, Pro, Tyr and Phe. For Asp, the residues Ala, Leu, Asp, Lys, Glu are the most prominent, and Trp, His, Cys, Tyr and Met are the least prominent neighbors. Leu, Asp, Lys, His, Val and Ala are the most

prominent and Phe, Trp, Asn, Met and Gly are the least prominent around Cys. It is interesting to note that Gly and Cys are, respectively, the most and least abundant residue around Ala, and Gly is the least preferred one around Cys. In all-α proteins, Cys residues have twice the average value around Asp; around Cys, His and Cys are observed two and nine times higher than the average, respectively; and around the His residue, twice the average value of Cys is observed. In all-α proteins, the residue Cys is the least neighboring residue, followed by Trp and Met. In a similar fashion, the most and the least favored neighboring residues around all the residue types are given in Table 3a. For convenience, and to analyze the results, the neighboring residues associated with each type have been divided into two groups according to their % distributions. The most (M) and least (L) neighbored residues for all the residue types are also listed in order of their preference.

3.2.2. *All-β proteins*

The most and least neighbored residues around all the residue types are given in Table 3b. Around Ala, the residues Ala, Thr, Ser, Gly, and Val are the most, and Met, His, Trp, Cys and Ile are the least favored. In all-β proteins, the residues Thr,

Table 4
Structural class prediction for all α proteins[a]

Protein PDB	Standard deviation				Predicted structural class	
	All α	All β	α + β	α/β	4-state model	3-state model
1AVHA	6.1*	6.65	6.43	6.21	α	α
1BABB	7.44*	7.93	7.87	7.65	α	α
1BRD	6.55*	7.24	7.03	6.77	α	α
1C5A	9.72*	10.72	10.39	10.23	α	α
1CPCA	10.03*	10.49	10.39	10.28	α	α
1CPCL	9.54*	9.81	9.76	9.74	α	α
1ECO	9.05*	9.27	9.28	9.1	α	α
1FCS	8.94*	9.39	9.31	9.18	α	α
1FHA	9.04*	9.3	9.25	9.23	α	α
1FIAB	9.47*	9.69	9.6	9.62	α	α
1HBG	9.68*	10.21	10.03	9.94	α	α
1HDDC	12.53*	12.93	12.83	12.81	α	α
1HIGA	11.92*	12.14	12.06	12.01	α	α
1LE4	11.84*	12.19	12.09	12.03	α	α
1LIG	11.91*	12.24	12.12	12.06	α	α
1LTSC	12.45*	12.57	12.55	12.57	α	α
1MBC	12.03*	12.18	12.16	12.13	α	α
1MBS	12.22*	12.5	12.46	12.38	α	α
1RPRA	12.8*	13.11	13.06	13.1	α	α
1TROA	12.87*	13.13	13.03	13.06	α	α
1UTG	12.51*	12.75	12.67	12.7	α	α
256BA	12.07*	12.3	12.24	12.23	α	α
2CCYA	12.24*	12.42	12.28	12.27	α	α
2LH1	11.88*	12.07	11.9	11.95	α	α
2LHB	12.57*	12.85	12.73	12.67	α	α
2MHBA	12.77*	13.08	12.97	12.84	α	α
2MHBB	12.24	12.26	12.26	12.15*	α/β	Mixed
2ZTAA	13.96*	14.15	14.1	14.03	α	α
4MBA	13.16*	13.41	13.35	13.19	α	α
4MBN	12.63	12.87	12.79	12.56*	α/β	Mixed

[a] Rate of correct prediction for: 4-state model 28/30 = 93.33%; and 3-state model 28/30 = 93.33%.

Gly, Ser are highly favored around all the residue types, whereas His, Met and Cys are the least favored residues.

3.2.3. $\alpha + \beta$ proteins

Ala, Val, Leu, Ser and Gly are the most, and Arg, Cys, His, Trp and Met are the least neighbored residues around Ala. A study of Table 3c reveals that the residues Leu, Ala, Val and Gly are higher and Trp, Met, His and Cys are lower in the order of preferences around all the residue types in the case of $\alpha + \beta$ proteins.

3.2.4. α / β proteins

In α/β proteins, Gly, Ala and Val are the most, and Cys, Trp and Met are the least favored residues around all the residue types. The least favored residues are very similar to those observed for all-α proteins.

For the theoretical prediction of structural class and secondary structure of proteins, the information presented above is helpful in understanding the most and least favored residue types (nearest neighbor) around all the residue types, depending on the structural class.

Table 5
Structural class prediction for all β proteins[a]

Protein PDB	Standard deviation				Predicted structural class	
	All α	All β	$\alpha + \beta$	α/β	4-state model	3-state model
1ACX	10.71	10.27*	10.4	10.43	β	β
1AYH	8.91	8.45*	8.57	8.56	β	β
1CD8	9.75	9.23*	9.5	9.49	β	β
1CDTA	10.75	10.51*	10.64	10.69	β	β
1CID	9.44	9.22*	9.4	9.43	β	β
1DFNA	11.78	11.62*	11.79	11.8	β	β
1HILA	9.85	9.32*	9.67	9.66	β	β
1HIVA	10.45	10.02*	10.32	10.23	β	β
1HLEB	11.81	11.52*	11.78	11.72	β	β
1MAMH	9.97	9.46*	9.78	9.76	β	β
1MONA	11.39	10.85*	11.13	11.24	β	β
1OMF	9.74	9.48*	9.54	9.72	β	β
1PHY	10.53	10.21*	10.31	10.49	β	β
1REIA	11.07	10.5*	10.8	10.97	β	β
1TEN	11.6	10.92*	11.28	11.38	β	β
1TLK	11.19	10.65*	10.98	10.96	β	β
1VAAB	9.15	8.83*	8.89	8.95	β	β
1ALP	9.32	8.89*	9.05	9.05	β	β
1AVIA	10.73	10.15*	10.47	10.64	β	β
2BPA2	8.78	8.26*	8.54	8.68	β	β
2HHRC	7.68	7.21*	7.37	7.6	β	β
2ILA	7.52	7.43	7.39*	7.47	$\alpha + \beta$	Mixed
2LALA	8.33	7.93*	8.01	8.06	β	β
2SNV	8.57	7.91*	8.17	8.13	β	β
2CD4A	9.23	8.56*	8.87	8.93	β	β
4GCR	8.67	8.1*	8.32	8.28	β	β
7APIB	10.17	9.71*	9.95	9.81	β	β
8II1B	9.35	9*	9.18	9.19	β	β
8FABA	9.63	9.11*	9.32	9.32	β	β
8FABB	8.79	8.36*	8.54	8.62	β	β

[a] Rate of correct prediction for 4- and 3-state models $29/30 = 96.66\%$.

3.3. Structural class prediction of globular proteins

The computed neighboring amino acid composition for each residue type (Table 2a–d) is considered as the norm to predict the structural class. The prediction of the structural classes has been performed for two sets of proteins, namely the training set and the test set. The prediction for the former is a re-substitution test for checking the self-consistency of this method, while that for the latter is a cross-validation test for checking its extrapolating effectiveness. This method gives better results, which are comparable to those already available.

3.4. Structural class prediction for the training set

We computed the neighboring amino acid composition for all the training set proteins, i.e. 30 representative proteins for each class (all-α, all-β, α + β, and α/β). Then we subtracted these 20 × 20 matrices from the norms of all-α, all-β, α + β and α/β proteins. Afterwards, the normal standard deviations (S.D.) for the four classes were calculated using Eq. (10). The lowest value indicates the structural class to which the particular protein belongs. The standard deviations computed for all the classes of proteins are presented in Tables 4–7. We found that 28 all-α proteins

Table 6
Structural class prediction for all α + β proteins[a]

Protein PDB	Standard deviation				Predicted structural class	
	All α	All β	α + β	α/β	4-state model	3-state model
1AAK	7.86	7.81	7.51*	7.56	α + β	Mixed
1CTF	10.72	11.08	10.82	10.67*	α/β	Mixed
1DNKA	7.3	7.28	7.02	6.96*	α/β	Mixed
1EAF	7.64	7.87	7.4*	7.41	α + β	Mixed
1HSBA	6.99	6.92	6.53*	6.65	α + β	Mixed
1LTSA	9.08	8.88	8.48*	8.61	α + β	Mixed
1LTSD	10.09	9.79	9.51*	9.56	α + β	Mixed
1NRCA	10.81	10.43	10.23*	10.33	α + β	Mixed
1OVB	10.83	10.47	10.25*	10.32	α + β	Mixed
1POC	9.55	9.17*	9.19	9.3	β	β
1PPN	9.16	8.77	8.71*	8.8	α + β	Mixed
1PRF	9.47	9.26	9.18	9.17*	α/β	Mixed
1RND	8.69	8.45	8.28*	8.3	α + β	Mixed
1SNC	8.86	8.85	8.59*	8.65	α + β	Mixed
1TFG	8.77	8.73	8.56*	8.59	α + β	Mixed
1TGSI	9.82	9.64	9.57*	9.66	α + β	Mixed
2ACHA	9.07	8.9	8.83*	8.86	α + β	Mixed
2ACT	8.99	8.7	8.5*	8.56	α + β	Mixed
2BPA1	7.18	6.74	6.53*	6.75	α + β	Mixed
2SNS	7.84	7.8	7.39*	7.72	α + β	Mixed
2SSI	9.58	9.37	9.13*	9.42	α + β	Mixed
3IL8	10.57	10.27	10.06*	10.37	α + β	Mixed
3RUBS	9.75	9.43	9.27*	9.61	α + β	Mixed
3SGBI	10.02	9.72	9.59*	9.92	α + β	Mixed
3SICI	9.96	9.7	9.46*	9.76	α + β	Mixed
4BLMA	9.89	9.78	9.43*	9.71	α + β	Mixed
4TMS	7.91	7.75	7.42*	7.62	α + β	Mixed
8CATA	5.66	5.31	4.99*	5.19	α + β	Mixed
9RNT	8.23	7.99	7.81*	7.92	α + β	Mixed
9RSAA	7.87	7.49	7.29*	7.61	α + β	Mixed

[a]Rate of correct prediction for: 4-state model 26/30 = 86.33%; and 3-state model 29/30 = 96.66%.

are correctly predicted out of 30, and the accuracy of prediction is 93%. In the case of all-β and α/β, 29 proteins are correctly predicted out of 30, with an accuracy of 97%. The all-β protein, 2ICA was wrongly predicted as a mixed type, and the mixed class protein 1TMD was assigned as all-β class. In the case of α + β proteins, only 26 were correctly predicted, with an accuracy of 87%. Altogether, 112 proteins are correctly predicted out of 120, with the rate of correct prediction being 93%. In most cases, α + β and α/β proteins are mixed judiciously and the structural class prediction is computed (three-state model). This model predicts with an accuracy level of 96% (115/120).

3.5. Structural class prediction for the test set

We have collected 80 unbiased, non-homologous proteins for the test set, which are not included in the training data set (Table 8). We applied the same methodology to these proteins and the rate of correct prediction is 90% (72/80) for the four-state and 96% (77/80) for the three-state level. It is worth noting that the three-state model predicts both training and test set proteins

Table 7
Structural class prediction for all α/β proteins[a]

Protein PDB	Standard deviation				Predicted structural class	
	All α	All β	α + β	α/β	4-state model	3-state model
1ABA	8.75	8.53	8.33	8.24*	α/β	Mixed
1CIS	10.93	11.14	10.82	10.6*	α/β	Mixed
1CSEI	10.93	10.92	10.68	10.43*	α/β	Mixed
1CTC	8.01	7.78	7.53	7.45*	α/β	Mixed
1DHR	6.83	6.73	6.54	6.28*	α/β	Mixed
1DRI	6.48	6.57	6.28	5.95*	α/β	Mixed
1ETU	7.47	7.31	7.26	7.04*	α/β	Mixed
1FX1	8.73	8.85	8.69	8.31*	α/β	Mixed
1GPB	3.71	3.78	3.32	2.99*	α/β	Mixed
1OFV	7.1	7.04	6.89	6.41*	α/β	Mixed
1PAX	7.98	7.84	7.77	7.44*	α/β	Mixed
1PFKA	6.67	6.55	6.45	6.19*	α/β	Mixed
1PGD	5.53	5.45	5.4	5.16*	α/β	Mixed
1Q21	5.71	5.87	5.57	5.3*	α/β	Mixed
1SO1	7.11	6.99	6.93	6.59*	α/β	Mixed
1SBP	6.6	6.54	6.5	6.18*	α/β	Mixed
1SBT	7.18	6.93	6.95	6.64*	α/β	Mixed
1TIMA	6.23	6.15	6.05	5.79*	α/β	Mixed
1TMD	4.63	4.04*	4.13	4.07	β	β
1TREA	5.77	6.03	5.81	5.51*	α/β	Mixed
1ULA	5.46	5.15	5.09	4.99*	α/β	Mixed
1WSYB	5.77	5.72	5.58	5.28*	α/β	Mixed
2HAD	4.85	5.15	4.86	4.79*	α/β	Mixed
2LIV	5.5	5.68	5.55	5.24*	α/β	Mixed
3GBP	5.32	5.58	5.39	4.99*	α/β	Mixed
4FXN	7.05	7.18	6.98	6.56*	α/β	Mixed
4CPA	6.39	6.27	6.13	5.93*	α/β	Mixed
5P21	6.31	6.09	5.92	5.7*	α/β	Mixed
8ABP	6.56	6.39	6.32	6.04*	α/β	Mixed
8ATCA	7.22	7.11	7.03	6.89*	α/β	Mixed

[a] Rate of correct prediction for 4- and 3-state models 29/30 = 96.66%. **Overall prediction**: 4-state model 112/120 = 93.33%; and 3-state model 115/120 = 95.83%.

Table 8
Structural class prediction for the test set protein[a]

Protein PDB	Standard deviation				Observed	Predicted	
	All-α	All-β	α + β	α/β		4-state	3-state
1HBBA	9.64*	10.41	10.13	9.84	α	α	α
1IFA	9.21*	9.69	9.59	9.51	α	α	α
2PDE	12.73*	13.2	13.15	12.92	α	α	α
2HCOA	11.46*	12.1	11.97	11.68	α	α	α
1PPT	12.69*	12.97	12.98	12.8	α	α	α
1GCN	14.47*	14.7	14.75	14.61	α	α	α
1MLT	15.27*	15.64	15.61	15.43	α	α	α
1PP2	14.4*	14.63	14.6	14.43	α	α	α
1BBL	9.29	9.5	9.28	9.27*	α	α/β	Mixed
451C	12.9*	13.53	13.43	13.25	α	α	α
1PPFE	11.99*	12.29	12.21	12.2	β	α	α
1RIA2	9.21	9*	9.13	9.16	β	β	β
1SHFA	10.84	10.65*	10.8	10.86	β	β	β
1TIE	9.85	9.56*	9.68	9.74	β	β	β
2ACHB	11.6	11.38*	11.5	11.54	β	β	β
2CTX	11.56	11.2*	11.45	11.46	β	β	β
2MEV1	9.58	9.18*	9.28	9.43	β	β	β
2SODO	9.94	9.34*	9.5	9.58	β	β	β
4CHA	9.01	8.67*	8.72	8.78	β	β	β
3EBX	11.27	10.92*	10.93	11.15	β	β	β
2APR	4.62	3.58*	3.84	3.75	β	β	β
3CAN	4.30	3.42*	3.56	3.83	β	β	β
1PCY	10.91	10.45*	10.61	10.74	β	β	β
1HOE	11.54	11.24*	11.42	11.43	β	β	β
2FB4	10.25	9.94*	10.07	10.03	β	β	β
3COX	7.74	7.3	7.18*	7.24	α + β	α + β	Mixed
1DNKA	7.47	7.23	6.92*	7.15	α + β	α + β	Mixed
1GLAG	6.94	6.87	6.47*	6.5	α + β	α + β	Mixed
2MS2A	7.61	7.6	7.39*	7.5	α + β	α + β	Mixed
3EST	4.79	4.03	3.84*	4.06	α + β	α + β	Mixed
1POC	8.02	7.8	7.72*	7.97	α + β	α + β	Mixed
3INS	5.17	4.82	4.75*	5.04	α + β	α + β	Mixed
1PPBA	6.56	6.52	6.22	6.2*	α + β	α/β	Mixed
1SHAA	7.71	7.7	7.63*	7.66	α + β	α + β	Mixed
2LZM	4.72	4.49	4.36*	4.43	α + β	α + β	Mixed
2AAA	5.49	5.17	5.05*	5.19	α + β	α + β	Mixed
2PIA	5.41	5.4	5.17*	5.29	α + β	α + β	Mixed
2CDV	5.14	4.99	4.82*	4.83	α + β	α + β	Mixed
2SN3	10.28	10.26	10.02*	10.16	α + β	α + β	Mixed
2TAAAA	5.07	4.71	4.47*	4.64	α + β	α + β	Mixed
4ENL	5.43	5.69	5.25*	5.29	α + β	α + β	Mixed
1FDX	8.1	7.94	7.73*	7.77	α + β	α + β	Mixed
2PAD	10.14	10.23	10.05*	9.95	α + β	α + β	Mixed
2HIP	10.52	10.76	10.48*	10.57	α + β	α + β	Mixed
3GAP	8.13	8.17	8.06*	8.13	α + β	α + β	Mixed
2PAZ	7.88*	8	7.91	7.9	α + β	α	α
4FD1	8.89	9	8.87*	8.92	α + β	α + β	Mixed
3RXN	9.91	9.96	9.86	9.85*	α + β	α/β	Mixed
1HLA	7.29	7.18	6.92*	7.19	α + β	α + β	Mixed
1HMG	6.16	5.83	5.74*	5.88	α + β	α + β	Mixed

Table 8 (*Continued*)

Protein PDB	Standard deviation				Observed	Predicted	
	All-α	All-β	α + β	α/β		4-state	3-state
2CI2	10.1	10.06	9.96*	9.86	α + β	α + β	Mixed
5PTI	11.72	11.72	11.57*	11.74	α + β	α + β	Mixed
1SN3	13.44	13.4	13.21*	13.26	α + β	α + β	Mixed
1PYP	10.4	10.22	10.04*	10.18	α + β	α + β	Mixed
1SGT	10.35	10.21	10.03*	10.1	α + β	α + β	Mixed
9PAP	8.73	8.42	8.27*	8.37	α + β	α + β	Mixed
2CAB	8.41	8.18	8.08*	8.13	α + β	α + β	Mixed
1PHH	6.78	6.85	6.56*	6.73	α + β	α + β	Mixed
3GRS	6.36	6.05	5.98	5.94*	α + β	α/β	Mixed
2CPP	6.04	5.81	5.68*	5.75	α + β	α + β	Mixed
1GBP	4.73	4.68	4.43	4.28*	α/β	α/β	Mixed
1MINA	4.33	4.01	3.78	3.69*	α/β	α/β	Mixed
1NIPB	5.08	5.07	4.87	4.46*	α/β	α/β	Mixed
4ADH	3.58	3.26	3.13	3.03*	α/β	α/β	Mixed
1SBP	4.81	4.66	4.39	4.21*	α/β	α/β	Mixed
1WSYA	4.87	5.12	4.76	4.61*	α/β	α/β	Mixed
4ICD	4.44	4.44	4.25	4.02*	α/β	α/β	Mixed
9RUBB	4.72	4.68	4.49	4.35*	α/β	α/β	Mixed
1GD1O	5.48	5.55	5.43	5.19*	α/β	α/β	Mixed
3ADK	6.61	6.7	6.42*	6.46	α/β	α + β	Mixed
3DFR	3.58	3.80	3.39	3.09*	α/β	α/β	Mixed
2GPD	5.49	5.19	5	4.91*	α/β	α/β	Mixed
3PGK	6.02	6.28	5.97	5.88*	α/β	α/β	Mixed
3LDH	5.17	5.08	4.81*	4.85	α/β	α + β	Mixed
1SRX	7.56	7.55	7.44	7.41*	α/β	α/β	Mixed
1ABP	6.23	6.14	6.12	6.04*	α/β	α/β	Mixed
2LBP	6.33	6.17	6.07	5.96*	α/β	α/β	Mixed
1CSE	7.29	7.22	7.22	7.02*	α/β	α/β	Mixed
1GP1	7.62	7.6	7.57	7.44*	α/β	α/β	Mixed
1GOX	7.36	7.12	7.07	6.9*	α/β	α/β	Mixed

[a] Rate of correct prediction for: 4-state model 72/80 = 90%; 3-state model 77/80 = 96.25%.

with an accuracy level of 96%. Furthermore, the present method correctly predicted the structural class of the proteins 2PIA, 4ENL and 1WSYA, which were wrongly predicted by Chou's [36] method.

3.6. Structural class prediction of chaperones

We made an attempt to predict the secondary structural class of chaperones with the same norms used for globular proteins. For this analysis, we have taken 91 chaperone proteins, of which 11% are predicted to be classed under all-α folding type, 7% belonging to the α + β and the remaining 82% predicted under α/β. When computed based on the three-state model, 11% of proteins were predicted as under the all-α type, whereas the remaining 89% were in the mixed type (Table 9). It is interesting to note that most of the chaperone proteins come under the category of α/β, or mixed folding type.

4. Conclusion

The amino acid composition and neighboring residue distributions of amino acid residues in all the four structural classes of globular proteins have been analyzed and the results are compared with chaperones. We have developed a method

Table 9
Structural class prediction of chaperone proteins

Protein	Standard deviation				Predicted structural class	
	All α	All β	α + β	α/β	4-state	3-state
P60_CRIGR	5.83	6.2	5.86	5.47*	α/β	Mixed
P60_HUMAN	5.87	6.14	5.81	5.44*	α/β	Mixed
P60_MOUSE	5.85	6.21	5.86	5.49*	α/β	Mixed
P60_RAT	5.86	6.22	5.88	5.5*	α/β	Mixed
RUBA_ARATH	7.75	8.12	7.9	7.5*	α/β	Mixed
RUBA_BRANA	8.39	8.86	8.63	8.25*	α/β	Mixed
RUBA_RICCO	8.55	9.07	8.85	8.48*	α/β	Mixed
RUBA_WHEAT	8.07	8.51	8.2	7.89*	α/β	Mixed
RUBB_ARATH	7.17	7.55	7.3	7.04*	α/β	Mixed
RUBB_BRANA	7.22	7.69	7.39	7.13*	α/β	Mixed
RUB2_BRANA	7.56	7.97	7.78	7.44*	α/β	Mixed
DNAK_BRUOV	7.68*	8.25	8.01	7.72	α	α
DNAK_STRCO	8.8	9.21	9.03	8.7*	α/β	Mixed
DNAK_ECOLI1	9	9.31	9.1	8.77*	α/β	Mixed
DNAK_MOUSE	8.78	8.95	8.81	8.57*	α/β	Mixed
HSP60_YEAST	8.63	8.87	8.75	8.47*	α/β	Mixed
CH60_ACYPS	8.62	8.97	8.87	8.6*	α/β	Mixed
CH60_AGRTU	8.96	9.43	9.19	8.93*	α/β	Mixed
CH60_AMOPS	8.49	8.79	8.69	8.33*	α/β	Mixed
CH60_ARATH	8.96	9.31	9.14	8.93*	α/β	Mixed
CH60_BACSU	8.68	9.03	8.84	8.57*	α/β	Mixed
CH60_BRUAB	8.86	9.26	9.06	8.84*	α/β	Mixed
CH60_CHLPN	7.96*	8.35	8.23	8	α	α
CH60_CHLTR	8.11*	8.5	8.34	8.14	α	α
CH60_CHRVI	7.64	7.98	7.83	7.55*	α/β	Mixed
CH60_CLOAB	7.98	8.3	8.05	7.9*	α/β	Mixed
CH60_CLOPE	8.94	9.18	9.11	8.9*	α/β	Mixed
CH60_COXBU	8.54	8.72	8.61	8.35*	α/β	Mixed
CH60_CYACA	8.92	9.1	9	8.74*	α/β	Mixed
CH60_ECOLI	8.89	9.14	9.08	8.73*	α/β	Mixed
CH60_HAEDU	9.23	9.52	9.4	9.09*	α/β	Mixed
CH60_LEGMI	8.46	8.5	8.45	8.14*	α/β	Mixed
CH60_LEGPN	8.47	8.52	8.46	8.14*	α/β	Mixed
CH60_MAIZE	7.52	7.77	7.61	7.37*	α/β	Mixed
CH60_MYCLE	8.38	8.63	8.58	8.24*	α/β	Mixed
CH60_MYCTU	8.42	8.68	8.63	8.28*	α/β	Mixed
CH60_PESAE	9.18	9.46	9.31	9.01*	α/β	Mixed
CH60_RICTS	8.98	9.02	8.85	8.66*	α/β	Mixed
CH60_SYNP7	9.45	9.78	9.59	9.29*	α/β	Mixed
CH60_SYNY3	9.57	9.82	9.67	9.33*	α/β	Mixed
CH60_THEP3	10.32	10.61	10.4	10.06*	α/β	Mixed
CH61_STRAL	9.37	9.68	9.5	9.1*	α/β	Mixed
CH62_STRAL	9.43	9.72	9.56	9.24*	α/β	Mixed
CH63_HELVI	9.36	9.51	9.38	9.16*	α/β	Mixed
CH60_BACST	9.49	9.8	9.57	9.25*	α/β	Mixed
CH60_BARBA	9.5	9.74	9.53	9.28*	α/β	Mixed
CH60_BRANA	8.98	9.07	9.01	8.73*	α/β	Mixed
CH60_LACLA	9.12	9.23	9.05	8.83*	α/β	Mixed
CH60_LEPIN	9.76	9.84	9.72	9.48*	α/β	Mixed
CH60_PLAFG	9.12	9.39	9.22	9.06*	α/β	Mixed

Table 9 (*Continued*)

Protein	Standard deviation				Predicted structural class	
	All α	All β	α + β	α/β	4-state	3-state
CH60_RHILV	9.34	9.76	9.52	9.28*	α/β	Mixed
CH60_STAAU	8.99	9.35	9.09	8.93*	α/β	Mixed
CH61_CUCMA	9.2	9.53	9.29	9.06*	α/β	Mixed
CH61_MYCLE	9.01	9.14	8.88	8.74*	α/β	Mixed
CH61_RHIME	9.47	9.69	9.41	9.28*	α/β	Mixed
CH61_STRCO	10.53	10.86	10.64	10.4*	α/β	Mixed
CH61_SYNY3	10.53	10.79	10.6	10.39*	α/β	Mixed
CH62_BRAJA	10.36	10.58	10.39	10.17*	α/β	Mixed
CH62_CUCMA	9.8	10.05	9.82	9.67*	α/β	Mixed
CH62_RHIME	9.65*	10.01	9.74	9.67	α	α
CH63_BRAJA	9.18*	9.61	9.34	9.2	α	α
CH63_RHIME	9.06	9.39	9.15	8.98*	α/β	Mixed
TCPD_MOUSE	8.61	8.61	8.5*	8.5*	α/β	Mixed
TCPB_MOUSE	8.12	8.12	7.96	7.92*	α/β	Mixed
TCPB_YEAST	7.82	7.82	7.64*	7.68	α + β	Mixed
TCP1_ARATH	7.17	7.39	7.09*	7.1	α + β	Mixed
TCP1_CRIGR	7.44	7.74	7.36	7.34*	α/β	Mixed
TCP1_DROME	7.64	7.81	7.56	7.52*	α/β	Mixed
TCP1_HUMAN	7.59	7.84	7.54	7.46*	α/β	Mixed
TCP1_RAT	7.61	7.88	7.58	7.48*	α/β	Mixed
TCP1_YEAST	7.67	7.8	7.64	7.5*	α/β	Mixed
TCPB_MOUSE	7.53	7.75	7.51	7.4*	α/β	Mixed
TF55_SULSH	8.54	8.66	8.56	8.3*	α/β	Mixed
TCPD_YEAST	8.58	8.47	8.37*	8.39	α + β	Mixed
TCPE_AVESA	7.67	7.58	7.43	7.4*	α/β	Mixed
TCPE_MOUSE	7.68	7.61	7.43	7.38*	α/β	Mixed
TCPE_YEAST	7.76	7.69	7.54	7.49*	α/β	Mixed
TCPG_MOUSE	6.22	6.07	5.93	5.85*	α/β	Mixed
TCPG_YEAST	5.28	5.44	5.18	5.17*	α/β	Mixed
TCPH_MOUSE	5.07	5.14	4.98	4.8*	α/β	Mixed
TCPZ_HUMAN	5.84	6	5.94	5.84*	α/β	Mixed
TCPZ_MOUSE	5.79	6.02	5.93	5.81*	α/β	Mixed
TCPZ_YEAST	6.22*	6.44	6.24	6.23	α	α
HS80_LYCES	5.21*	5.27	5.23	5.26	α	α
HS82_ASPFU	7.67	7.72	7.46*	7.58	α + β	Mixed
HS82_MAIZE	5.54*	5.68	5.58	5.62	α	α
HS82_TOBAC	5.95	6.02	5.9*	5.96	α + β	Mixed
HS9B_RAT	10.21*	10.38	10.25	10.28	α	α
HSCA_ECOLI	6.76*	7.06	6.81	6.86	α	α
HSLU_BACSU	6.32	6.39	6.24*	6.25	α + β	Mixed
HSP7_YEAST	6.5	6.67	6.49	6.45*	α/β	Mixed
Predicted results						
4-state model	10		6	75		
3-state model	10			81		

based on the neighboring residue distribution along the sequence, which predicts the structural class at an accuracy level of 96%. The present method is simple and does not require high computational power. Our structural class prediction for chaperones reveals that most of them come under the category of α/β, or mixed folding type.

References

[1] C.B. Anfinsen, Science 181 (1973) 223–230.
[2] G. Fisher, F.X. Schmid, Biochemistry 29 (1990) 2205–2212.
[3] R.B. Freedman, Cell 57 (1989) 1069–1072.
[4] R.B. Freedman, Trends Biochem. Sci. 9 (1984) 438–441.
[5] R.J. Ellis, Nature 328 (1987) 378–379.
[6] R.J. Ellis, Science 250 (1990) 954–959.
[7] R.J. Ellis, S.M. Vander Vies, Annu. Rev. Biochem. 60 (1991) 321–347.
[8] J.P. Hendrick, F.U. Hartl, Annu. Rev. Biochem. 62 (1993) 349–384.
[9] M.J. Gething, J. Sambrook, Nature 355 (1992) 33–45.
[10] M. Gierasch, Annu. Rev. Biophys. Biomol. Struct. 23 (1994) 645–669.
[11] R.J. Ellis, Curr. Opin. Struct. Biol. 4 (1994) 117–122.
[12] F.U. Hartl, Nature 381 (1996) 571–580.
[13] W.A. Fenton, A.L. Horwich, Protein Eng. 6 (1997) 743–760.
[14] J.E. Rothman, Cell 59 (1989) 591–601.
[15] T.J.P. Hubbard, C. Sander, Protein Eng. 4 (1991) 711–717.
[16] K.M. Flaherty, C. Deluca-Flaherty, D.B. McKay, Nature 346 (1990) 623–628.
[17] A. Holmgren, C.I. Branden, Nature 342 (1989) 248–251.
[18] K. Braig, Z. Otwinski, R. Hegda et al., Nature 371 (1994) 578–586.
[19] J.F. Hunt, A.J. Weaver, S.J. Landry, L. Gierasch, J. Deisenhofer, Nature 379 (1996) 37–45.
[20] Z. Xu, A.L. Horwich, P.B. Sigler, Nature 388 (1997) 741–750.
[21] T.S. Kumarevel, M.M. Gromiha, M.N. Ponnuswamy, Biophys. Chem. 71 (1998) 99–111.
[22] T.S. Kumarevel, M.M. Gromiha, M.N. Ponnuswamy, Biophys. Chem. 75 (1998) 105–113.
[23] T.S. Kumarevel, M.M. Gromiha, M.N. Ponnuswamy, Curr. Sci. 75 (1998) 1000–1002.
[24] T.S. Kumarevel, M.M. Gromiha, M.N. Ponnuswamy, Prep. Biochem. Biotech. (in press).
[25] M. Levitt, C. Chothia, Nature 261 (1976) 552–557.
[26] J.S. Richardson, D.C. Richardson, Principles and patterns of protein conformation, in: G.D. Fasman (Ed.), Prediction of Protein Structure and Principles of Protein Conformation, Plenum Press, New York, 1989, pp. 1–98.
[27] P. Klein, Biochem. Biophys. Acta 874 (1986) 205–215.
[28] P. Klein, C. DeLisi, Biopolymers 25 (1986) 1659–1672.
[29] P.Y. Chou, in: G.D. Fasman (Ed.), Prediction of Protein Structure and Principles of Protein Conformation, Plenum Press, New York, 1989, pp. 549–586.
[30] H. Nakashima, K. Nishikawa, T. Ooi, J. Biochem. 99 (1986) 152–162.
[31] H. Cid, M. Bunster, M. Canales, F. Garitura, Protein Eng. 5 (1992) 373–375.
[32] P.K. Ponnuswamy, M. Prabhakaran, P. Manavalan, Biochem. Biophys. Acta. 623 (1980) 301–316.
[33] C.T. Zhang, K.C. Chou, Protein Sci. 1 (1992) 401–408.
[34] K.C. Chou, C.T. Zhang, J. Protein Chem. 12 (1993) 169–178.
[35] M.M. Gromiha, P.K. Ponnuswamy, Int. J. Peptide Protein Res. 45 (1995) 225–241.
[36] K.C. Chou, Proteins 21 (1995) 319–344.
[37] F.C. Bernstein, T.F. Koetzle, J.B. Williams et al., J. Mol. Biol. 112 (1977) 535–542.
[38] E. Abola, F.C. Bernstein, S.H. Bryant, T.F. Koetzle, J. Weng, Protein data bank, in: F.H. Allen, G. Bergerhoff, R. Sievers (Eds.), Crystallographic Databases — Information content, Software Systems, Scientific Applications, , 1987, pp. 107–132.
[39] A. Bairoch, B. Boeckmann, Nucleic Acids. Res. 20 (1992) 2019–2022.