MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **65** (2011) 433-444

ISSN 0340 - 6253

Protein Structural Class Assignment Based on a Mixed Method

Liwei Liu [a], **Chunxin Yuan** [b], **Min Cai** [a]

[a] *College of Science, Dalian Jiaotong University, Dalian 116028, China*
[b] *School of Mathematical Science, Ocean University of China, Qingdao 266071,China*
e-mail: daliguowei@163.com

(Received December 15, 2009)

Abstract

In terms of the classification of the protein secondary structures, we have proposed a 2-D representation of protein secondary structure sequences. The representation is used to display, analyze, and compare the secondary structure sequences. Based on this representation, we constructed protein structural class tree by a mix method. The mix method extract the information of geometrical center of graphic and the information of leading eigenvalue. The structural class assignment deriving from the mix method closes to real structural class relation.

INTRODUCTION

It is well known that proteins are the executants of biological functions in life. Understanding protein structure is an important era of bioinformatics. A direct and valuable method to meet this challenge is protein structure comparison. Some methods of protein structure comparison have been proposed. For instance, Randic and Krilov used the D/D matrices to characterize the folding of five model proteins and applied the folding indices to measure the degree of similarity for molecules [1]-[3]. Liu and Wang constructed the partial order of the 3D model proteins based on the folding degree of 3-step path conformations [4]. Estrada offered the graph theory method. The application of this approach to the study of protein folding is reported [5]-[9]. Johannissen and Taylor applied a dynamic programming algorithm to compare the topological strings of protein and clustered them into a tree [10]. Gilbert et al. explored the alignment-free comparison of Topology

of protein structure (TOPS) diagrams [11]. Ferragina et al. used the USM to classify biological sequences and structures [12]. Liao et al. proposed a 6D representation of protein sequences consisting of 20 amino acids. Based on this 6D representation, they provided a proteome distance measure for constructing phylogenic tree [13]. In the paper [14], Jia et al. proposed a novel 2D representation for protein secondary structure sequences.

We outlined a graphic representation of protein secondary structure sequence [15]. In the paper, we introduce a 2D representation of protein secondary structure sequences, based on the 2D graphical representation, assigning the structural class to the protein, and discuss the advantage or disadvantage of the methods for predicting protein second structure.

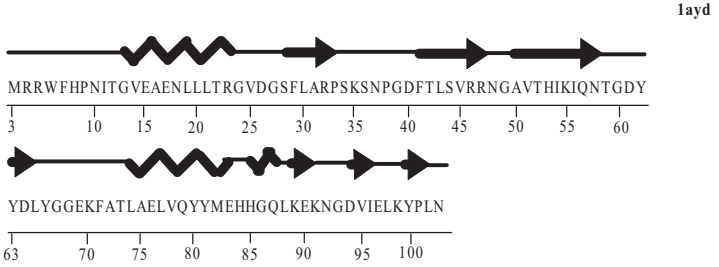

Figure 1. The secondary structures of the protein 1ayd

Figure 1 shows the secondary structures of protein, whose PDB code is 1ayd, and it belongs to $\alpha + \beta$ structural classes. In this graph, the secondary structures of a protein are defined by the local back-bone conformation at each position. Secondary structure elements of greatest interest include α-helices (wave) and β-strands (wide arrowhead). They are represented as H and E, respectively, in the 1D summary. Remaining positions are represented by C for coil. A secondary structure sequence [16] is a linear sequence defined over alphabet L = {C, H, E}, where H represents helix, E represents strand and the rest are represented by C (mainly coil and turn). For example, the subsequence and substructure corresponding to the position from the 84th to 93th in the protein 1ayd are illustrated in Figure 2.

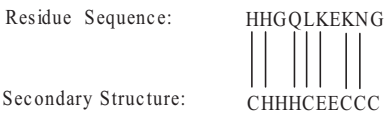

Figure 2. The substructures of the protein 1ayd

Secondary structure elements of greatest interest include α-helices (wave) and β-strands (wide arrowhead). They are represented as H and E, respectively, in the 1D summary. Remaining positions are represented by C for coil. In the paper [15], we let $G = g_1 g_2 \cdots$ be a protein secondary structure sequence. Then we have a map ϕ, which maps G into a plot set. Explicitly, $\phi(G) = \phi(g_1)\phi(g_2)\cdots$, where

$$\phi(g_i) = \begin{cases} (i,1) & \text{if } g_i = H \\ (i,0) & \text{if } g_i = C \\ (i,-1) & \text{if } g_i = E. \end{cases}$$

As usual, the corresponding plot set is named as characteristic plot set, and the curve connected by all plots in turn is characteristic curve.

According to three classifications, we can only obtain three representations, i.e. patterns HCE, EHC and CEH. They correspond to ϕ , $\phi^{'}$ and $\phi^{''}$, respectively, where

$$\phi^{'}(g_i) = \begin{cases} (i,1) & \text{if } g_i = E \\ (i,0) & \text{if } g_i = H \\ (i,-1) & \text{if } g_i = C. \end{cases}$$

$$\phi^{''}(g_i) = \begin{cases} (i,1) & \text{if } g_i = C \\ (i,0) & \text{if } g_i = E \\ (i,-1) & \text{if } g_i = H. \end{cases}$$

In this paper, based on this representation, a mixed method is introduced. This method can be used to extract the information of geometrical center of graphic and leading eigenvalues.

MATERIALS

Four structural classes of protein have been proposed by Levitt and Chothia [20], i.e. all-α, all-β, $\alpha+\beta$ and α/β, and any protein should be grouped into one of these classes. The all-α and all-β proteins are almost entirely composed of α -helices and β-strands, respectively. The $\alpha+\beta$ protein is a combination of α regions and β regions, which are largely separated and the β-strands are often antiparallel. while α/β protein consists of helices and strands that are alternatively mixed and the β-strands are often parallel. This classification basically describes almost all the structures of proteins, and has been accepted widely and used in the up-to-data literatures.

Table 1: The secondary structure sequences of the 12 different proteins

PDB code	Secondary structure sequences
1mba	CCCCHHHHHHHHHHHHHHHHCHHHHHHHHHHHHHHHHCHHHHHHCCCCCCCCHHHHH
	CCCCHHHHHHHHHHHHHHHHHHCCCHHHHHHHHHHHHHHHHHCCCCHHHHHHHHHH-
	HHHHHHCCCCCCCCHHHHHHHHHHHHHHHHHHCCC
1rcb	CCCCCHHHHHHHHHHHHHHCCCCCCCCCEEECHHHCCCCCCHHHHHHHHHHHHHHHHHH
	HCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCEEEHHHHHH-
	HHHHHHHHHHHHC
2hmqa	CCCCCCCCCCCHHHCCCCCCCCHHHHHHHHHHHHHHHHHCCCHHHHHHHHHHHHHHHH
	HHHHHHHCCCCCHHHHHHHHHHHHHHHHHCCCCCHHHHHHHHHHHHHHHCCHHHCCCC
1plc	CEEEECCCCCCCCEECCEEEECCCCEEEEEECCCCCCCCEECCCCCCCCCCHHHHCCCCC
	CCCCCCCCEEEEECCCCEEEEEECHHHCCCCCEEEEEEC
4fgf	CCEEEECCCCEEEEECCCCCEEEECCCCCHHHCEEEEEECCEEEEECCCCEEEECCCC
	CEEEECCCCHHHCEEEEECCCCCEEEEECCCCCCCCCCCCCCCCCCHHHCCCCCHH-
	HCEEEEC
1noa	CCCEEEEECCCCCCCCCEEEEEEECCCCCCEEEEEEEEEECCEEEECHHHCEEEECCCCC
	CEEEEEECCCEEEEECCCCCEEEEEECCCCCEEEEEECCCCCCCCCEECCCC
1sha	CCCCEEECCCHHHHHHHHCCCCCCCCCEEEEECCCCCCCEEEEEEEEECCCEEEEEEEEE
	ECCCCCEECCCCCEECCHHHHHHHHCCCCCCCCCCCCCCCCC
1ubq	CEEEEEECCCCEEEEECCCCCHHHHHHHHHHHHCCCHHHEEEEECCEECCCCCCCHHHCC
	CCCCEEEEEECCCCC
1ayd	CCCCCCCCCCHHHHHHHHHHCCCCEEEEECCCCCCCEEEEEEECCEEEEEEEECCCCE
	ECCCCCCCCCHHHHHHHHHHCHHHCEECCCEECCCCEECC
1wsya	CHHHHHHHHHHCCCCCCEEEEEEECCCCCHHHHHHHHHHHHHCCCCCEEEECCCC
2pgdI	CCCCEEEECCCHHHHHHHHHHHCCCCEEEECCCCHHHHHHHHCCCCCCCCEECCCHHHHH
	HHHCCCCCEEEECCCCCHHHHHHHHHHHHHCCCCCEEEECCCCCHHHHHHHHHHH-
	HCCCEEEEEEEECHHHHHHHCCEEEEEECCCCHHHHHHHHHHHCCCCCCCCCCCCCCCCC
	CCCCCCEECCHHH
2trxa	CCCEEECCCCCHHHHCCCCCCEEEEEEECCCCHHHHHHHHHHHHHHHHCCCCEEEEEEE
	CCCCCCHHHHCCCCCCCEEEEEECCEEEEEEECCCCHHHHHHHHHHHHC

Here we will discuss the secondary structure sequences of the 12 proteins, shown in Table 1. Proteins 1mba, 1rcb and 2hmqa are almost entirely composed of α-helices, which means these belong to the α structural class. As for proteins 1plc, 4fgf and 1noa, β -strands are the main components, thus these should be the β structural class. Based on the classification stand, proteins 1sha, 1ubq and 1ayd should be thought as $\alpha + \beta$ structural class, while proteins 1wsya, 2pgdI and 2trxa should be grouped to α/β structural class.

FIRST METHOD

Using the 2D graphical representation [15], any protein secondary structure sequence could be described to a set of plots in 2D space, then for any sequence, we could have a set of points (x_i, y_i), $i = 1, 2, 3, ..., N$, where N is the length of the sequence, and the coordinates of the geometrical center of these points, denoted by x^0 and y^0, can be obtained by: $x^0 = \frac{1}{N}\sum_{i=1}^{N} x_i$, $y^0 = \frac{1}{N}\sum_{i=1}^{N} y_i$.

In this way,corresponding to three patterns HCE, EHC and CEH, we have three geometrical centers. Because the x coordinate of the geometric center is decided by the length of the secondary structure sequence, so these three geometric center's x coordinate is equal. The length of the secondary structure sequence is irrelevant to structure class, therefore, we only need to consider y coordinate. We take three y coordinates as three elements of a 3-component vector.

The similarities among such vectors can be got by this way: calculating the Euclidean distance between end points of the vectors. The Euclidean distance between end points of two vectors is smaller, which suggests the secondary structure sequences are much similar and hence the more possibly the corresponding two proteins belong to the same structural class. In Table 2, these Euclidean distances between 3-component vectors of the geometrical center are listed.

Then, this pair-wise distance matrix can be input to the neighbor program, PHYLIP package, to get a structural class tree. This structure class tree can be named as SC tree. In Figure 3, SC tree is drawn by Treeview Program. These proteins in the same cluster are considered to be the same structure class. From Figure 3, three α/β proteins(2trxa, 1wsya and 2pgdI) are separated into two branches and others are adjacent according to the structural class.

Table 2: The similarity/dissimilarity matrix for 12 protein secondary structure sequences based on the Euclidean distance between 3-component vectors of the geometrical center.

	1mba	1rcb	2hmqa	1plc	4fgf	1noa	1sha	1ubq	1aya	1wsya	2pgdI	2trxa
1mba	0	0.0904	0.4525	1.3262	1.2890	1.3652	0.9686	0.8292	1.0602	0.6723	0.6279	0.3847
1rcb		0	0.4889	1.2780	1.2379	1.3087	0.9470	0.7885	1.0136	0.6147	0.5718	0.3488
2hmqa			0	1.1099	1.0915	1.1946	0.6516	0.6260	0.8572	0.6032	0.5628	0.3483
1plc				0	0.0629	0.1525	0.4964	0.5031	0.2663	0.6696	0.7092	0.9429
4fgf					0	0.1210	0.4986	0.4744	0.2363	0.6258	0.6668	0.9076
1noa						0	0.6175	0.5706	0.3382	0.6941	0.7377	0.9892
1sha							0	0.2487	0.3097	0.4712	0.4819	0.6009
1ubq								0	0.2384	0.2310	0.2519	0.4445
1ayd									0	0.4121	0.4488	0.6766
1wsya										0	0.0459	0.3094
2pgdI											0	0.2636
2trxa												0

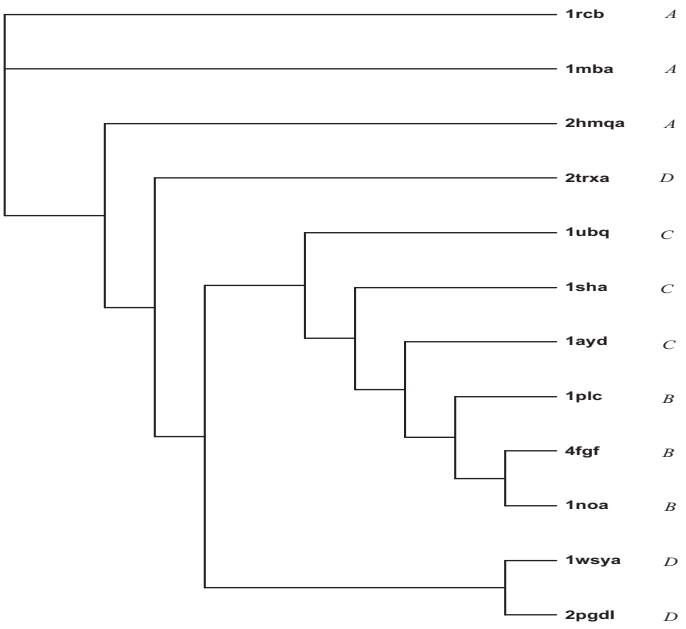

Figure 3. SC tree (obtained the first method), A, B, C, D indicate all-α class, all-β class, $\alpha + \beta$ class and α/β class, respectively.

SECONDARY METHOD

In this section, we will transform the 2D representation of the protein secondary structure sequences into another mathematical object, a matrix. Once we have a matrix to represent the protein secondary structure sequences, we can use some of matrix invariants as descriptors of the secondary structure sequences. One of the matrices is the M/M matrix, which is the same as Randic's paper [17]. So that, $m_{i,j} = \frac{d_{i,j}}{|i-j|}$, where $d_{i,j}$ is the Euclidean distance between a pair of vertices. The M/M matrix's eigenvalues, and in particular its leading eigenvalue can be used as descriptor of a protein secondary structure sequence. Among all eigenvalues of a matrix, the leading eigenvalue often plays a special role. In the case of the adjacency matrix of trees, based on a result by Lovász and Pelikán [18], one suggested [19] that the leading eigenvalue λ be viewed as an index of molecular branching. Then, corresponding to these three patterns HCE, EHC and CEH, three leading eigenvalues could be obtained. We take three leading eigenvalues as three elements of a 3-component vector.

Next, we will consider the 12 protein secondary structure sequences and their distances between 3-component vectors. By uniting all these distances into a matrix, a pair-wise distance matrix is got, this matrix, shown in Table 3, contains the distance information about the 12 protein secondary structure sequences. At last, similar to the first method, this matrix can be also input to the PHYLIP package for getting a SC tree. In Figure 4, SC tree is drawn by Treeview Program. From this figure, we find that three $\alpha + \beta$ proteins(1ubq, 1sha and 1ayd) are adjacent, but other proteins separated into two branches.

Table 3: The similarity/dissimilarity matrix for 12 protein secondary structure sequences based on the Euclidean distance between 3-component vectors of the leading eigenvalue.

	1mba	1rcb	2hmqa	1plc	4fgf	1noa	1sha	1ubq	1aya	1wsya	2pgdI	2trxa
1mba	0	0.7427	1.8254	1.9630	2.1335	0.8618	1.4384	1.1199	1.4445	1.5383	1.6237	5.0907
1rcb		0	1.0830	1.2212	1.3914	0.1451	0.6972	0.3794	0.7025	0.7959	0.8812	4.3481
2hmqa			0	0.1740	0.3222	0.9755	0.3990	0.7111	0.3884	0.2932	0.2087	3.2653
1plc				0	0.1715	1.1044	0.5246	0.8432	0.5190	0.4262	0.3422	3.1318
4fgf					0	1.2756	0.6955	1.0139	0.6892	0.5956	0.5108	2.9604
1noa						0	0.5807	0.2646	0.5888	0.6837	0.7693	4.2358
1sha							0	0.3186	0.0220	0.1063	0.1905	3.6552
1ubq								0	0.3255	0.4199	0.5056	3.9731
1ayd									0	0.0957	0.1811	3.6479
1wsya										0	0.0857	3.5533
2pgdI											0	3.4677
2trxa												0

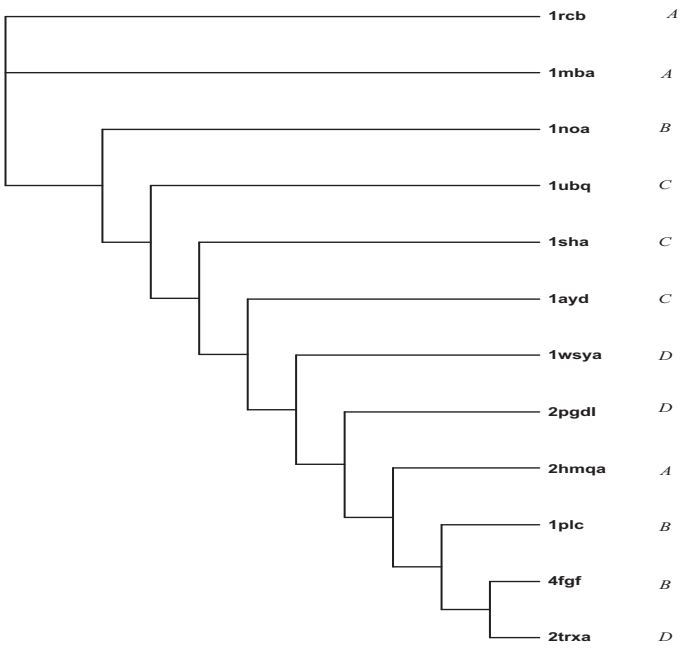

Figure 4. SC tree (obtained the second method)

Consequently, two sets of parameters will be obtained, and the first one reflects the difference of center positions represented by the Euclid distance between the geometric centers, while the second indicates the different trends of 2D graphs displayed by the related eigenvalues [21].

But we should realize that the constructed structure class trees got by both of these methods are not same as the real structure class relation. A possible way to improve the results is to unit these two methods, then we can get a mixed method. At first, we put the distance in Tables 2 and 3 be normalized. The so-called normalization is that the maximum distance of Table is set to 1, others divided by the maximum distance. For example, the maximum distance in Table 2 is 1.3652, so, the distance between 1mba and 1rcb is $0.0904/1.3652 = 0.0602$. Then we get the sum of these two normalized matrices, called as mixed matrix.

Table 4: The similarity/dissimilarity matrix for 12 protein secondary structure sequences based on the Euclidean distance of the mix method.

	1mba	1rcb	2hmqa	1plc	4fgf	1noa	1sha	1ubq	1aya	1wsya	2pgdI	2trxa
1mba	0	0.2121	0.6901	1.3570	1.3633	1.1693	0.9921	0.8274	1.0604	0.7947	0.7789	1.2818
1rcb		0	0.5708	1.1760	1.1801	0.9871	0.8307	0.6521	0.8805	0.6066	0.5919	1.1096
2hmqa			0	0.8472	0.8628	1.0666	0.5557	0.5982	0.7042	0.4994	0.4532	0.8965
1plc				0	0.0798	0.3286	0.4667	0.5341	0.2971	0.5742	0.5867	1.3059
4fgf					0	0.3392	0.5018	0.5467	0.3085	0.5754	0.5887	1.2463
1noa						0	0.5664	0.4700	0.3634	0.6427	0.6915	1.5567
1sha							0	0.2448	0.2312	0.3661	0.3904	1.1582
1ubq								0	0.2385	0.2517	0.2838	1.1061
1ayd									0	0.3207	0.3643	1.2122
1wsya										0	0.0504	0.9246
2pgdI											0	0.8743
2trxa												0

The mixed matrix is shown in Table 4, and based on this matrix, the SC tree could be constructed in Figure 5. From this figure, we can find that proteins of a class are adjacent, and the distances between all α proteins and all β proteins are furthest, which suggests the SC tree derived from the mixed method is close to real structure class relation. Compared these methods, the fist method only considers the geometrical center of the graph, and the second method just takes into account the degree of folding of the

curve of graph, obviously these two methods lose the information of protein secondary structure sequences. The information derived from the mixed method manifests more full-scale. So the result got from the mixed matrix is more close to the real structural class relation.

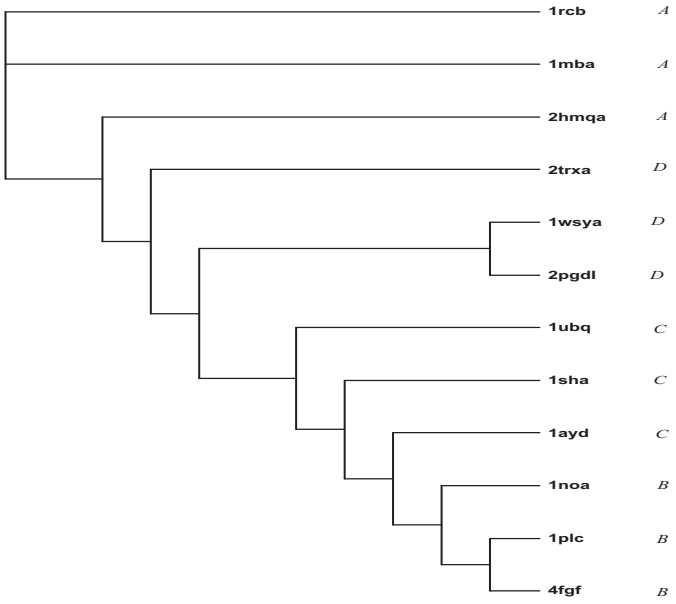

Figure 5. SC tree (obtained the mix method)

CONCLUSION

In this paper, we integrate protein secondary structure sequences into the mixed method to cluster analysis. It is appropriate for the research on different protein class. We provide the mixed method for the comparison of protein structures based on both geometrical center and leading eigenvalue. The main advantage is that this mixed method can extract more full-scale information. Our result demonstrated 12 protein secondary structure sequences. All proteins clustered together according to structural class, and the distance between all α protein and all β protein is farthest. This result show that the SC tree deriving from the mixed method closes to real structural class relation. How about the result? The first method only consider the geometrical center of the graphic; the

second method only consider the degree of folding of the curve of graph. The information derive from the mixed manifest more full-scale. So, the result derived the mixed method more close to real structural class relation. Unlike most existing structure comparison methods, the proposed method does not require multiple alignment, and the computation is simple.

The method used to construct the protein classification in SCOP or CATH is essentially the visual inspection and comparison of structures though various automatic tools are used to make the task manageable and help provide generality. Our method used to construct the protein classification is research of fully automated classification of protein structural classes.

Acknowledgements: This work was partially supported by the National Natural Science Foundation of China(Grant No.10871219) and the Science Research Project of Educational Department of Liaoning Province of China (2009A125)

The author thanks the anonymous referees for their valuable suggestions and support.

References

[1] M. Randić, G. Krilov, Characterization of 3–D sequences of proteins, *Chem. Phys. Lett.* **272** (1997) 115–119.

[2] M. Randić, G. Krilov, On a characterization of the folding of proteins, *Int. J. Quantum. Chem.* **75** (1999) 1017–1026.

[3] G. Krilov, M. Randić, Quantitative characterization of protein structure: application to a novel α/β fold, *New. J. Chem.* **28** (2004) 1608–1614.

[4] L. Liu, T. Wang, Novel characterization of the folding of proteins, *Int. J. Quantum. Chem.* **107** (2007) 1970–1974.

[5] E. Estrada, Characterization of 3D molecular structure, *Chem. Phys. Lett.* **319** (2000) 713–718.

[6] E. Estrada, Characterization of the folding degree of proteins, *Bioinformatics* **18** (2002) 697–704.

[7] E. Estrada, Characterization of the amino acid contribution to the folding degree of proteins, *Proteins* **54** (2004) 727–737.

[8] E. Estrada, A protein folding degree measure and its dependence on crystal packing, protein size, secondary structure, and domain structural class, *J. Chem. Inf. Comput. Sci.* **44** (2004) 1238–1250.

[9] E. Estrada, E. Uriarte, S. Vilar, Effect of protein folding on the stability of protein–ligand complexes, *J. Proteo. Res.* **5** (2006) 105–111.

[10] L. O. Johannissen, W. R. Taylor, Protein fold comparison by the alignment of topological strings, *PEDS* **16** (2003) 949–955.

[11] D. R. Gilbert, F. Rossello, G. Valiente, M. Veeramalai, Alignment–free comparison of TOPS strings, London algorithmics and stringology, in: J. Daykin, M. Mohamed, K. Steinhofel (Eds.), *London Algorithmics and Stringology*, volume 8 of Texts in Algorithmics, College Publications, **ch. 11** (2007) pp. 177–197.

[12] P. Ferragina, R. Giancarlo, V. Greco, M. Veeramalai, G. Manzini, G. Valiente, Compression–based classification of biological sequences and structures via the universal similarity metric: experimental assessment, *BMC Bioinformatics* **8** (2007) 252.

[13] B. Liao, J. Luo, R. Li, W. Zhu, Novel method for analyzing proteome, *Int. J. Quantum. Chem.* **107** (2007) 1295–1300.

[14] C. Jia, T. Liu, X. Zhang, S. Yan, Protein secondary structure class assignment on the basis of a new graphic representation, *Int. J. Quantum. Chem.* **109** (2009) 819–825.

[15] L. Liu, T. Wang, 2D representation of protein secondary structure sequences and its applications, *J. Comput. Chem.* **27** (2006) 1119–1124.

[16] C. T. Zhang, R. Zhang, S curve, a graphic representation of protein secondary structure sequence and its application, *Biopolymers* **53** (2000) 539–549.

[17] M. Randić, M. Vračko, N. Lerš, D. Plavšić, Novel 2–D graphical representation of DNA sequences and their numerical characterization, *Chem. Phys. Lett.* **368** (2003) 1–6.

[18] L. Lovasz, J. I. Pelikan, On the eigenvalues of trees, *Period. Math. Hung.* **3** (1973) 175–182.

[19] D. M. Cvetković, I. Gutman, Note on branching, *Croat. Chem. Acta* **49** (1977) 115–121.

[20] M. Levitt, C. Chothia, Structural patterns in globular proteins, *Nature* **261** (1976) 552–558.

[21] C. Yuan, L. Liu, T. Wang, C. Li, On property of the invariant of graphical representations of DNA sequences, *J. Math. Chem.* **43** (2008) 1177–1183.