Characterization of the Amino Acid Contribution to the Folding Degree of Proteins

Ernesto Estrada*

Safety and Environmental Assurance Centre (SEAC), Unilever, Colworth House, Sharnbrook, Beds, and RIAIDT, Edificio CACTUS, University of Santiago de Compostela, Spain

ABSTRACT **The folding degree index (Estrada, Bioinformatics 2002;18:697–704) is extended to account for the contribution of amino acids to folding. First, the mathematical formalism for extending the folding degree index is presented. Then, the amino acid contributions to folding degree of several proteins are used to analyze its relation to secondary structure. The possibilities of using these contributions in helping or checking the assignation of secondary structure to amino acids are also introduced. The influence of external factors to the amino acids contribution to folding degree is studied through the temperature effect on ribonuclease A. Finally, the analysis of 3D protein similarity through the use of amino acid contributions to folding degree is studied by selecting a series of lysozymes. These results are compared to that obtained by sequence alignment (2D similarity) and 3D superposition of the structures, showing the uniqueness of the current approach. Proteins 2004;54:727–737.**
© 2004 Wiley-Liss, Inc.

Key words: **graph theory; spectral moments; thermal expansion; secondary structure; protein similarity**

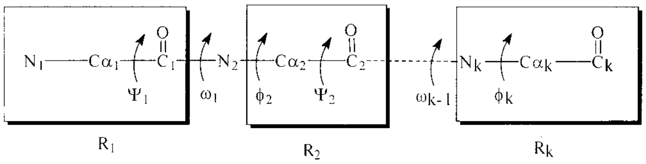
INTRODUCTION

The understanding of the molecular structure of proteins plays a key role for the comprehension of protein interactions and functions.[1–3] The necessary but not sufficient condition for attaining this objective is the determination of the protein structure by experimental or accurate theoretical methods.[4–6] This process entails the determination of the amino acid sequence as well as of the three-dimensional (3D) structure. However, in order to obtain valuable information about protein function, interaction or evolution, it is necessary to express this structural information in a way that permits their storage, manipulation and comparison with large pools of other structures in a fast and efficient way.[7] While these processes are very well developed for sequence analysis there is not the same level of development for the analysis of 3D structures. Protein sequences are represented as letter strings that are easily manipulated, stored and compared by different computer algorithms.[7–9] In contrast, to do similar quantitative comparisons between the 3D structure of proteins we have to "translate" this structural information into an appropriate mathematical form, such as a number, a vector or a matrix: the so-called molecular descriptors.[10–19] The superposition of two protein structures is also a way to obtain information about how similar they are in the 3D space, so the root mean square (RMS) of the superposition can be considered as a descriptor.

Recently, Estrada introduced a new molecular descriptor to characterize the 3D structure of protein chains.[20–22] This index represents in a condensed way the folding degree of a macromolecular chain, which is intimately related to the secondary and supersecondary structures of proteins,[21] and also to important protein properties and functions.[21,22] However, the reduction of the 3D structure of a protein to a single number, the folding degree index, represents a clear loss of information. One way to recover part of this information is to define the contribution of each amino acid to the folding degree index. This particular objective is covered by the current work in which we define the way to obtain these contributions and apply them to study its relation to secondary structure, influence of temperature on protein folding degree as well as to 3D protein similarity.

Theoretical Approach

Let P be a protein chain as represented in the following scheme:



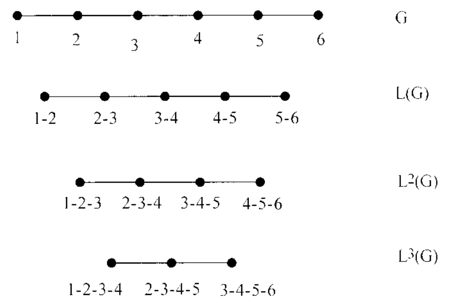
Here Ψ_i, ω_i and ϕ_i are the torsional or dihedral angles of the backbone chain of the protein. Ψ_i, describes rotation about the $C_{\alpha i} - C_i$ (= O) bond, ω_i defines rotation about the peptidic bond C_i (= O) $-$ N_{i+1}. The normal *trans* planar peptide bond has $\omega_i = 180°$. ϕ_i describes rotation about the $N_i - C_{\alpha i}$ bond.

This chain can be represented as a graph G = (V, E), where V is the set of vertices and E is the set of edges in the graph.[23] In this representation the atoms of the protein chain are represented as vertices in G and the covalent bonds connecting these atoms are represented as the edges

*E. Estrada's current address is University of Santiago de Compostela. E-mail: estrada66@yahoo.com

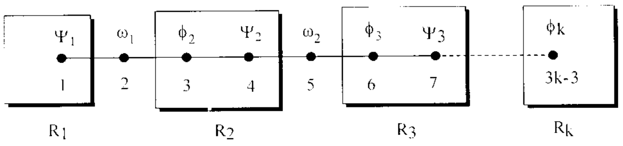
Received 24 April 2003; Accepted 22 July 2003

of G = (V, E). The graph G can be transformed into a set of other representations through the use of the iterated line graph sequence (ILGS) approach.[24] The first element of this sequence of graph representations is the proper graph G, the second one is the line graph of G, L(G), the next element corresponds to the line graph of the line graph that we call the second line graph of G, $L^2(G)$, the next element is the third line graph, $L^3(G)$, and so on. The vertices of each graph in the ILGS represent different structural elements of the protein chain P. For instance, the vertices of G represent the atoms, the vertices of L(G) represents pairs of atoms, the vertices of $L^2(G)$ represent triples of consecutive atoms, and the vertices in $L^3(G)$ represent quadruples of consecutive atoms in the protein chain. A graphical representation is given below:



Each of these elements in the ILGS graphs can be identified with geometrical elements of the chain. For instances, the pairs of adjacent atoms define covalent bonds in the chain, triples of consecutive atoms define bond angles and quadruples of consecutive atoms define dihedral or torsion angles.

According to the picture previously explained on the basis of the ILGS, the third line graph of a protein chain will correspond to the representation of its dihedral angles in the form given below schematically:



This graph can be represented mathematically through the use of the adjacency matrix, $A(L^3(G))$, using a function of the dihedral angles Ψ_i, ω_i, and ϕ_i as diagonal entries of the matrix. The adjacency matrix[23] is a square symmetric matrix whose non-diagonal entries are one's or zero's if the corresponding vertices of the graph are adjacent or not, respectively. In a chain having k residues (amino acids) the number of atoms, n, is equal to 3k and the number of dihedral angles is n − 3, which is equal to 3k − 3. Consequently, the order of this adjacency matrix is n − 3 or 3k − 3. For the sake of simplicity we will call simply **A** to the $\mathbf{A}(L^3(G))$ matrix.

The eigenvalues λ_i of the matrix **A** with the diagonal entries weighted by the cosines of the respective dihedral angles were used to define the folding degree index I_3 of a protein chain. Consequently, the main diagonal entries of this matrix are weighted by numbers in the range −1 to 1, which correspond to the values of the cosines of the dihedral angles. The cosine function is selected, among other things, because it takes the lowest value for the less folded structures, e.g., −1 for an extended structure with an angle of 180°, and the highest value for the most folded one, e.g., +1 for 0°. The I_3 index was defined as:[20–22]

$$I_3 = \frac{1}{n-3}\sum_{i=1}^{n-3} e^{\lambda_i} \tag{1}$$

The I_3 index can be expressed in terms of the spectral moments of the matrix **A** using the following transformation:

$$I_3 = \frac{1}{n-3}\sum_{i=1}^{n-3} e^{\lambda_i} = \frac{1}{n-3} \times \sum_{r=0}^{\infty}\left(\frac{\sum_{i=1}^{n-3}(\lambda_i)^r}{r!}\right) = \frac{1}{n-3}\sum_{r=0}^{\infty}\frac{\mu_r}{r!} \tag{2}$$

where $\mu_r = \sum_{i=1}^{n-3}(\lambda_i)^r = \mathbf{Tr}(\mathbf{A}^r)$ is the r-th spectral moment of **A**, and **Tr** means the trace, i.e., sum of diagonal entries of the corresponding matrix.

According to the proper definition of the spectral moments as the trace of the different powers of the corresponding matrix **A**,[25–27] we can express the r-th spectral moment as the following sum:

$$\sum_r \frac{\mu_r}{r!} = \sum_r \frac{\mu_r^{(1)}}{r!} + \sum_r \frac{\mu_r^{(2)}}{r!} + \sum_r \frac{\mu_r^{(3)}}{r!} + \cdots + \sum_r \frac{\mu_r^{(n-3)}}{r!} \tag{3}$$

where $\sum_r (\mu_r^{(i)}/r!)$ is the contribution of vertex i in $L^3(G)$ to the folding of the chain. The value of $\mu_r^{(i)}$ is simply the ii-diagonal entry of the matrix **A** raised to power r. We call to these elements the local spectral moments of the matrix **A**.[27] Spectral properties of graphs theoretic matrices have found interesting applications in protein studies in the works of Kunan and Vishveshwara.[28,29]

It is now straightforward to realize that the contribution of the amino acid (1) to the whole folding degree of the chain is given by the following expression (see Scheme 1):

$$R_1 = \sum_r \frac{\mu_r^{(1)}}{r!}, \tag{4}$$

and contributions of second and third amino acids are, respectively:

$$R_2 = \sum_r \frac{\mu_r^{(3)}}{r!} + \sum_r \frac{\mu_r^{(4)}}{r!}, \tag{5}$$

and

$$R_3 = \sum_k \frac{\mu_k^{(6)}}{k!} + \sum_k \frac{\mu_k^{(7)}}{k!} \tag{6}$$

In general, the contribution of non-terminal amino acids can be calculated by the following expression:

$$R_i = \sum_r \frac{\mu_r^{(3i-3)}}{r!} + \sum_r \frac{\mu_r^{(3i-2)}}{r!} \tag{7}$$

and that of terminal amino acids is given by:

$$R_k = \sum_r \frac{\mu_r^{(k)}}{r!} \tag{8}$$

There are other contributions coming from the vertices labelled as ω_i that correspond to peptide bonds (see Scheme 1). They account for couples of residues in the following way:

$$R_{1-2} = \sum_r \frac{\mu_r^{(2)}}{r!} \tag{9}$$

$$R_{2-3} = \sum_r \frac{\mu_r^{(5)}}{r!} \tag{10}$$

In general, the following expression will permit the calculation of any peptidic bond to the global folding degree of a protein chain:

$$R_{i-j} = \sum_r \frac{\mu_r^{(2i+j-2)}}{r!} \tag{11}$$

We recall that the following single relations exist between the number of residues, k, number of atoms, n, and the number of dihedral angles, h: n = 3k; h = n − 3 = 3k − 3.

METHODS

The crystallographic structures analyzed in the current work were taken from the Protein Data Bank (PDB) (http://www.rscb.org/pdb/).[30] The Cartesian coordinates of these structures were used to calculate the folding degree index and the amino acid contribution to folding using an in-house program. The percentages of secondary structures were taken from the PDBFINDER database (http://www.cmbi.kun.nl/gv/pdbfinder/).[31] They are reported here using the following DSSP codes (http://www.cmbi.kun.nl/gv/dssp/):[32] the hydrogen-bonded β-strands (extended strands) are represented by the letter E, α-helix by H, 3_{10}-helix by G, π-helix by I, isolated β-bridge by B, H-bonded turn (3-turn, 4-turn, or 5-turn) by T, and bend (five-residue bend centered at residue *i*) by S. The superposition of protein structures were carried out by using the Hyperchem computer software.[33] The superposition of structures was conducted by considering all atoms in the backbone of the proteins, i.e., the atoms contributing to the RMS are N, Cα, and C (carbonyl) of each amino acid. They were always carried out by considering only the backbone chain of the proteins and never the side chains. The alignments of primary sequences of proteins were done by using the program ALIGN from the database ExPasy (http://us.expasy.org). The fold topology of the proteins studied were obtained from the TOPS cartoons as given in the TOPS database (http://www.tops.leeds.ac.uk).

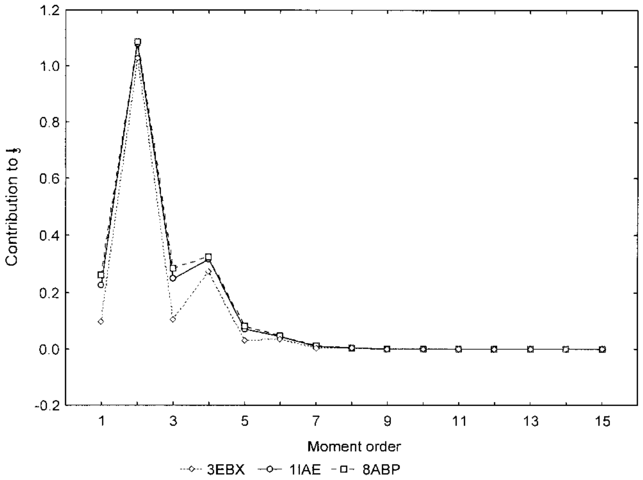


Fig. 1. Contribution of the spectral moments of different orders to the global folding degree index as a way to analyze their influence in the truncation of the moments at certain order (see text for explanation).

TABLE I. Contribution of the Spectral Moments of Order 15 to the Global Folding Degree Index (I_3) in Several Proteins of Different Sizes

PDB code	Residues	Contribution of μ_{15} to I_3 (10^8)
8RXN	54	2.55
1CMB	104	3.19
4LYTA	129	3.87
1COB	151	1.96
4GCR	174	1.53
1IAE	200	3.46
1NBA	253	3.17
8ABP	305	3.49

RESULTS AND DISCUSSION

The first question that arises for the calculation of the amino acids contribution to the global folding degree of a protein chain is related to the number of spectral moments to be included in the calculation.[25,26] As can be seen in the right part of expression (2), the summation is carried out over an infinite number of spectral moments. This, in practice, is neither possible nor necessary. What we propose here is a truncation of this summation for a limited, and relatively small, number of spectral moments. The numerical values of the spectral moments increase rapidly with the order of the moment, r, but as can be seen in expression (2) they are divided by the factorial of the order, i.e., by r!, which also increase dramatically with the value of r. Consequently, if we plot the values of the contribution of $\mu_r/r!$ versus the order of the spectral moment we obtain a graph likes that shown in Figure 1 for three proteins with 62, 200, and 305 amino acids, respectively. In order to obtain the values of real contributions to the folding degree index we have divided each value of $\mu_r/r!$ by n − 3 as in expression (2). As can be seen in this graphic the contribution of spectral moments of orders higher than 10 is almost negligible and the truncation at a point after this value is justified without any loss of precision in the calculation of

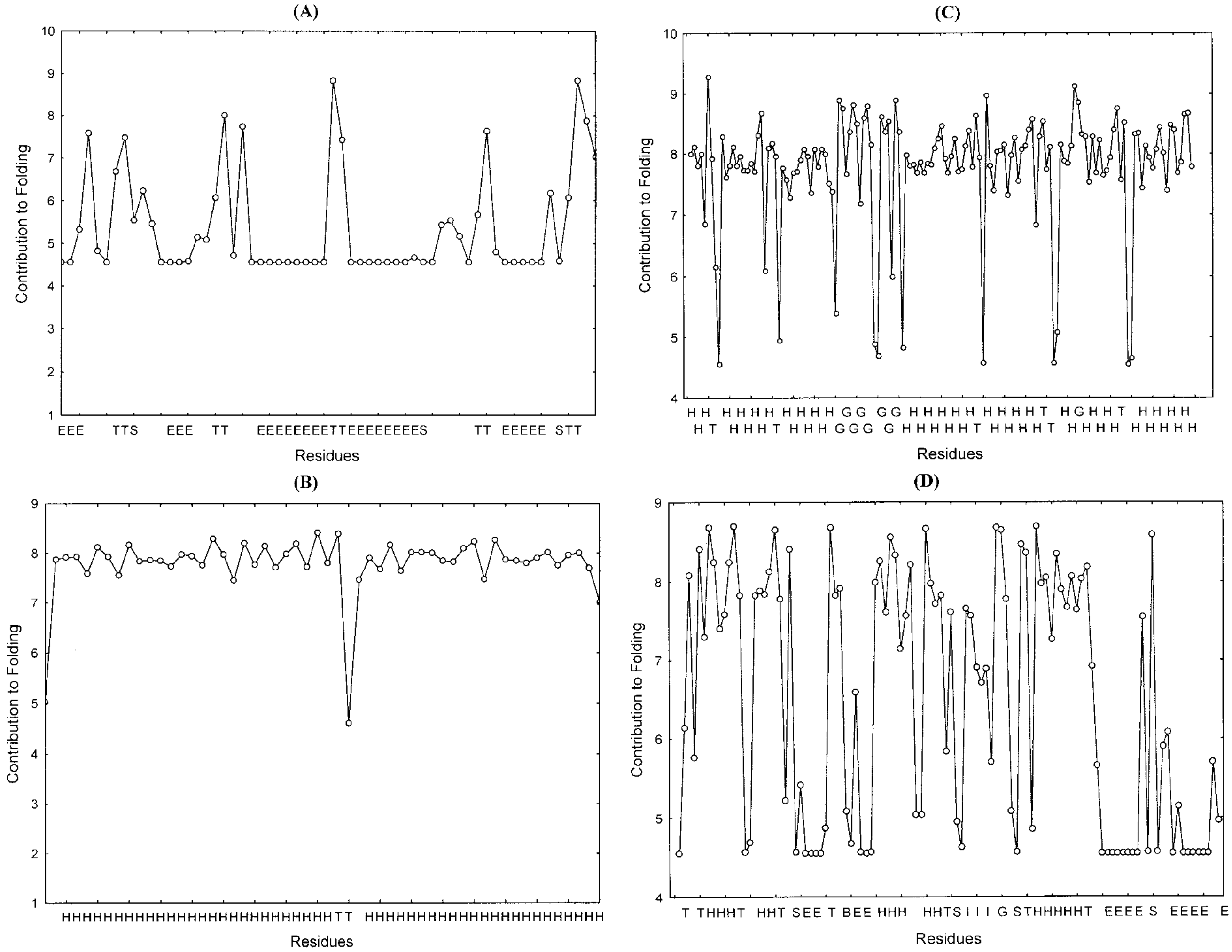


Fig. 2. Protein (3EBX) containing β sheets (designated by E) in the secondary structure (**A**), Protein (1ROP) containing α helices (designated by H) in the secondary structure (**B**), Protein (3SDHA) containing 3_{10} helices (designated by G) in the secondary structure (**C**) and Protein (1D7E) containing π helices (designated by I) in the secondary structure (**D**).

I_3. We recall that I_3 is normally reported with a precision of the fourth decimal place.

To avoid any loss of information we will use here and thereafter the first 15 spectral moments for the calculation of the amino acid contributions to the protein folding degree. In Table I we give the contributions of the spectral moment of order 15 to the value of I_3 for a series of eight proteins with different number of amino acids. We can see that the values of these contributions are in all cases under 10^{-7}, from which it is clear that there will be not any loss of information in the calculation of amino acids contributions to I_3.

Analysis of Secondary Structure

In a previous paper we shown that the folding degree index I_3 is linearly related to the secondary structure of proteins with different number of residues.[21] That is, a linear correlation exists between I_3 and the percentages of strand and helix, respectively. However, the use of the global folding degree index does not show the contribution of the different parts of a protein to their secondary structure.[21]

In Figure 2(A–D) we illustrate the contribution of each amino acid to the folding degree index in different proteins. At the bottom of these figures we give the DSSP code for the secondary structure assignation of each amino acid.[31] The first of these proteins is a mainly β protein (3EBX) that has 44% of strand in its secondary structure. It is observed that all β-strands (E) have very similar values of the amino acid contribution to the folding degree of about 4.5 with very small fluctuations. In fact, the calculation of the amino acid contributions to the I_3 index in a peptide having the standard values of the dihedral angles of a β-strand, i.e., $\rho = -139°$ and $\Psi = +135°$, gives 4.56. In contrast, most of the turns and bends have values over 6 showing a much greater variability than the β-strands. The second example [Fig. 2(B)] corresponds to a mainly α protein (1ROP), which has 89% of helix in its secondary structure. The amino acids in α-helixes (H) show contributions to the folding degree that are around 8 (average values for the two helices are 7.91 and 7.87). The value obtained for the hypothetical α-helix with $\rho = -57°$ and $\Psi = -47°$ is 7.97. These amino acid contributions

TABLE II. Average Values of the Amino Acid Contributions to Folding Degree for Different Secondary Structures in Proteins

PDB	Residues	Average	PDB	Residues	Average
α-helices					
1CZY(A)	2–14	8.0182	2PYP	11–15	8.0471
	28–36	7.9155		20–23	8.1490
1F9I(A)	11–14	8.0075		44–50	7.8902
	20–23	8.1170		54–57	8.0533
	44–50	7.8999		76–85	7.9198
	54–57	8.0092	1BYE(D)	14–23	7.7906
	76–85	7.9329		43–48	8.0704
1YAC(A)	32–48	7.8591		69–73	7.8509
	69–74	8.0571		91–94	7.8765
	97–101	7.9979		111–115	7.9499
	119–127	7.9195		129–151	7.7690
	144–157	7.8827		162–165	7.5985
	164–172	7.8469		168–173	7.5360
	179–188	7.8807		187–196	7.8609
	191–203	7.9276		202–205	8.1908
π-helices					
1CZY(A)	143–147	7.3363	2PYP	62–66	7.0174
1F9I(A)	62–66	7.0686	1BYE(D)	100–104	7.3666
1YAC(A)	114–118	7.3173			
3_{10}-helices					
1CZY(A)	87–89	8.4285	2PYP	68–70	8.3905
	121–123	8.3275	1BYE(D)	39–41	8.3276
1F9I(A)	15–17	8.2026		174–176	8.4620
	68–69	8.3804		178–181	8.4939
1YAC(A)	23–26	8.4111		183–185	8.3800
	89–91	8.5642			
	175–178	8.6449			
β-sheets					
1CZY(A)	21–25	4.5593	1YAC(A)	12–17	4.5619
	56–62	4.5599		53–58	4.5608
	73–81	4.5667		80–83	4.6995
	97–101	4.5597		106–111	4.5608
	110–114	4.5738		131–135	4.5601
	134–141	5.8821(?)		160–163	4.5736
	157–163	4.5599	2PYP	30–34	4.5648
1F9I(A)	29–34	4.7854		89–96	4.5613
	39–42	5.2119		103–110	4.6068
	89–96	4.5774		117–123	4.5605
	103–110	4.5970	1BYE	29–31	4.7551
	117–123	4.5602		56–59	4.5671
				62–65	4.6024

show more variability than those of amino acids in β-strands but there are not such fluctuations as for amino acids in turns and bends. The third example is that of a mainly α protein having three 3_{10} helices (G) in its structure (3SDHA) [Fig. 2(C)]. All α-helices in this protein have amino acid contributions to the folding degree that are under, or very close, to the value of 7.97 (8.01, 7.87, 7.74, 7.95, 7.99, 8.00, 8.02, 8.09). In contrast, the 3_{10} helices show values of amino acid contributions that are significantly over the value of 8 (8.37, 8.13, 8.78). It is known that 3_{10} helices are the most folded helices in proteins having standard values of the dihedral angles of $\phi = -49°$ and $\psi = -26°$.[38] The hypothetical 3_{10}-helix has an amino acid contribution to the folding degree of 9.30. The final example is a protein containing α-helices as well as a 3_{10} and a π helix (1D7E) [Fig. 2(D)]. As before, in this protein the contribution of α-helices are around the value of 8 (8.02, 8.07, 7.92, 8.04, 7.96) while the amino acids in the 3_{10} helix have average contributions of 8.37. However, the amino acids in the π helix (I) have contributions that average a value of 7.14, significantly under the values of all α-helices. This type of helices are known to be the less folded ones, having dihedral angles of $\phi = -57°$ and $\psi = -70°$ as compared to those of α-helices which are $\phi = -57°$ and $\Psi = -47°$.[38] The standard π -helix has an amino acid contribution to I_3 equal to 6.84. These simple examples show that the contribution of amino acids to the folding degree index reflects very well the secondary structure of the proteins as we also proved for the whole I_3 index.[21] We also obtained similar results for different types of proteins indicating the generality of our findings. In Table II we give the contribution of several α-, 3_{10}-, and π-helices as well as of β-sheets, which are present in four different proteins. As can be seen, the values of the amino acid contribution to the folding degree follow the patterns before mentioned giving values that are clearly distinguishable for each type of secondary structure. Another type of graphical information that could be produced using the I_3 index information is based on the map of ϕ and Ψ produced showing the value of the folding degree at each ϕ, Ψ point in the map. It, of course, will give a three-dimensional graphic that contains the information about the folding of the different regions of the protein in a detailed way. This kind of map will be studied in a forthcoming publication.

It is necessary at this point to say something about how the I_3 index and the amino acid contributions to it represent the folding degree of a protein. The I_3 index is based on the spectral moments of an adjacency matrix weighted in the main diagonal by the cosine of the dihedral angles of the protein backbone. Consider, for the sake of simplicity, the following example: put two boxes in any of four contiguous rooms. If we design by 1 a room filled by a box and by 0 those empty, we have the following three possibilities (eliminating those identical by symmetry): 1100; 1010; 1001. It is clear that the most compact organization of the boxes is the first one and the less compact is the last one. It is clear that the total number of boxes or the "density" of boxes are not enough to describe the "compactness" of the boxes in the rooms. If we consider, however, these organizations as graphs, i.e., weighted chains with weights 1 or 0, we can represent them by the adjacency matrices. Then, we can calculate the spectral moments of such alternatives. We can easily obtain that the first and second spectral moments are identical for the three conformations. The third spectral moment is identical for the first two ($\mu_3 = 11$) and takes a value of 8 for the third one. However, the fourth moment gives the following values: 32, 28, and 24, respectively. If we calculate a sort of I_3 index, which is simply the sum of these spectral moments divided by the factorial of the order of the moment, we obtain the higher value for the most compact organization of the boxes and the lowest value for the less compact one. It is straightforward to realize that the same is true if we, instead of the number of boxes in the rooms, consider cosine of dihedral angles in protein backbones. The highest values are obtained for those chains having more number of angles close to 0° (cosine equal to one) and less number of angles close to 180° (cosine equal to −1). At the same time, and more importantly, those chains having more number of consecutive (agglomeration) angles of the first type (as in the example of the boxes) will be considered as more folded. The explanation of why the spectral moments account for this kind of information resides in the fact that they count the number of weighted self-returning walks in the chain. By this mean, the "walker" visits the occupied places more times (ones in our example) if they are placed together instead of separated in the chain.

One of the possible applications of this type of relationship between the secondary structure of proteins and the contribution of amino acids to the folding degree is related to the assignation of amino acids to specific secondary structure elements. In a recent paper, Fodje and Karadaghi have reanalyzed the occurrence, conformational features and amino acid propensities for the π-helix.[39] Among the proteins analyzed we selected (at random) the sarcoplasmic calcium-binding protein from sandworm (*Nereis diversicolor*) (2SCP) for calculating the contributions of amino acids to the folding degree. These authors reassign the region 56–62 as a π-helix while DSSP assign the region 45–59 as an α-helix and 61–63 as a 3_{10}-helix. The average value of the amino acid contribution for the regions assigned by DSSP are as follows: 7.8768 for the region 45–59 and 8.5630 for the region 61–63. If we analyze the region 56–62 proposed as a π-helix by Fodje and Karadaghi we obtain a value of amino acid contribution to folding degree of 7.5981. The π-helices that we have analyzed have values of amino acid contributions which are well-below this value, ranging from around 7.0 to 7.3. However, the average value obtained for amino acids in the region 45–59 corresponds with the values obtained for α-helices, which are in the range from around 7.6 to 8.2. The region between 61–63 is unequivocally identified by the amino acids contribution as a 3_{10}-helix, which corresponds with the assignation carried out by DSSP. If we analyze the dihedral angles for the region between amino acids 55 to 63 we can observe more agreement with the assignation carried out by DSSP than with that obtained by Fodje and Karadaghi (see Table III). This example illustrates the possibilities of application of the current approach in helping to assign the correct secondary structure of amino acids.

TABLE III. Dihedral Angles, Amino Acids Assignation to Secondary Structure According to DSSP or Fodje and Al-Karadaghi[39] and Contribution to Folding Degree for a Region of the Protein 2SCP

Residue	ϕ (°)	Ψ (°)	DSSP (F–K)[a]	Folding
55	−63.70	−22.14	H	8.6672
56	−60.84	−45.92	H(I)	7.8315
57	−71.64	−54.62	H(I)	6.7346
58	−50.34	−43.59	H(I)	8.4660
59	−99.50	−18.10	H(I)	7.5474
60	−120.23	−72.53	T (I)	5.2722
61	−60.78	−26.61	G(I)	8.6612
62	−59.26	−28.53	G(I)	8.6739
63	−75.87	−8.85	G	8.3538

[a] Assignation of amino acids to secondary structure, in parenthesis the assignation made by Fodje and Al-Karadaghi (Protein Eng 2002;15: 353–358).

The final aspect we want to remark concerning the calculation of amino acid contribution to the folding degree is related to the contribution coming from the peptide bonds. In the theoretical part of this work we make a separation of the contributions coming from the dihedral angles ϕ and ψ from those made by ω, i.e., the peptide bond. The main reason for this separation is the well known fact that the peptide bonds have almost constant values around 180°. The only one exception is the proline, which can adopt a *cis* conformation in which this angle is around 0°.[38] In Figure 3 we illustrate the contribution coming from peptide bonds in the protein 1RAT, which contains two prolines in *cis* conformation. The values of the contribution

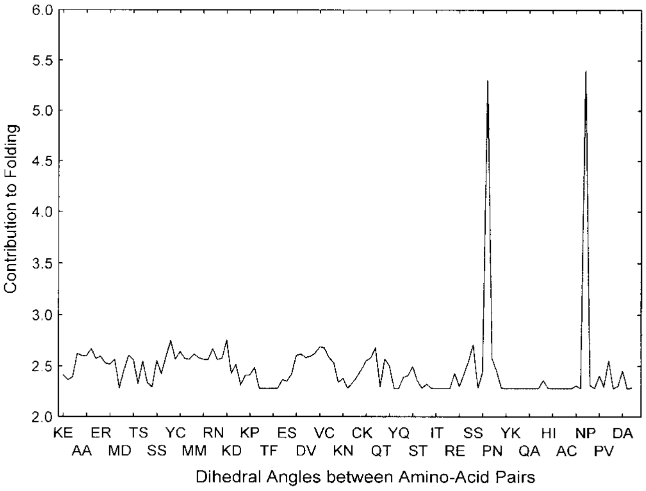


Fig. 3. Contribution of peptide bonds to the folding degree of the protein ribonuclease A (RAT1).

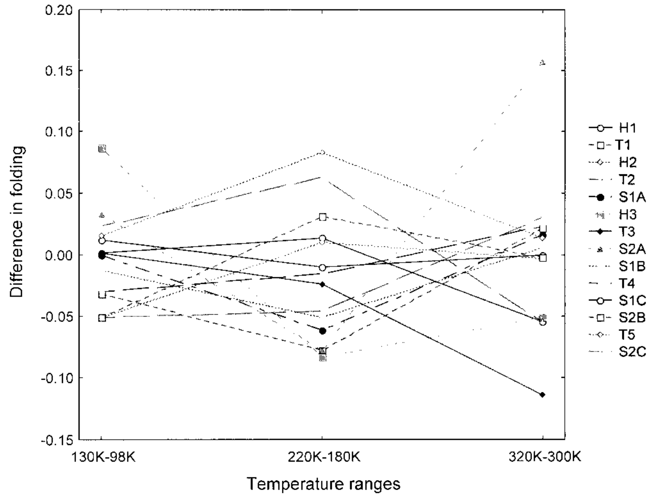


Fig. 4. Changes in folding degree of different regions of ribonuclease A at different temperatures.

of peptide bonds is almost constant around a value of 2.5 but the *cis* proline reach values of around 5.5.

Influence of Temperature on Amino Acid Folding Degree

Here we will first study the ribonuclease A determined at nine different temperatures from 98 to 320 K whose crystallographic structures were reported by Tilton et al.[40] In our previous report we showed that the folding degree index indicates a slight de-folding of the ribonuclease A that is linear with the increase of the temperature.[21] In the further analysis we will use the following definitions of secondary structure elements studied by Tilton et al.[40] (T, turn; H helix; S strand): H1 (residues 3–13); T1 (residues 14–23); H2 (residues 24–34); T2 (residues 35–41); S1A (residues 42–48); H3 (residues 50–60); T3 (residues 61–70); S2A (residues 71–75); S1B (residues 79–86); T4 (residues 87–95); S1C (residues 96–104); S2B (residues 105–110); T5 (residues 111–117); S2C (residues 118–124). These secondary structure elements were taken into account for calculating the differences in the average of amino acid contributions for structures close in temperature. This analysis shows that most of the secondary structure elements suffer a de-folding when the temperature is changed from 98K to 130K. On the contrary, the region H3-T3-S2A as well as S1C, T5, and S2C increase the folding as illustrated by the positive values of the differences in the amino acid contributions to the folding degree index. The greatest re-folding is observed for H3, which is bigger than the greatest de-folding observed for S2C. This helix start to de-fold when passing from 180K to 220K and keeps this trend when passing from 300K to 320K. Due to the anisotropic character of the thermal expansion it is not expected that each secondary structure element follow the same trend at different temperature ranges. This behavior is observed for instance for H2, S2A, S1C, and S2B. However, the trend observed when passing from 98K to 130K is maintained by 57% of the secondary structure elements when passing from 180K to 220K. For instance, H1 starts to re-folds from 98 to 130K and continues this trend from 180K to 220K. T1 starts to de-fold from 98 to 130K and continues de-folding from 180K to 220K. Similar behaviors to these are observed for T2, S1A, S1B, T4, T5, and S2C. However, around the temperature of 220K a change in these trends is observed. From the eight elements that have the same trend (increasing or decreasing folding) for the range 98K to 220K only T5 maintains it when passing from 300K to 320K. On the other hand, only three elements (H3, T3, and T5) maintain the same trend started at the range 180–220K when they are analyzed at the range 300–320K. The general trends for the 14 secondary structure elements is illustrated in Figure 4. It can be observed that around the temperature of 220K there is a change in the trend of the folding contribution of the amino acids indicating a change in the dynamic properties of the ribonuclease A crystal. The biphasic behavior of the thermal expansion of this protein was detected by Tilton et al. using the Debye-Waller factors.[40]

The amino acid contributions to the folding degree of the ribonuclease A were used to analyse the similarities of the structures of this protein determined at different temperatures. The values of the contribution were standardized and the similarity matrix using Euclidean distance and complete linkage was built for the nine structures described above. The dendogram is illustrated in Figure 5. It is observed that there are two main clusters of structures which have similarities between 20–30%. One of these clusters is formed by structures studied at temperatures 130, 160, and 180K (low temperatures) and the other cluster is formed by structures determined at temperatures 240, 260, and 300K (high temperatures). The structures obtained at temperatures 98, 220, and 320 K are quite dissimilar (unique) having similarities of less than 10%. The characteristic features of the structures determined at 98 and 320K are understood by the fact that these are the lowest and highest temperature studied and they represent the extreme cases of folding changes produced by the thermal expansion of this protein. However,

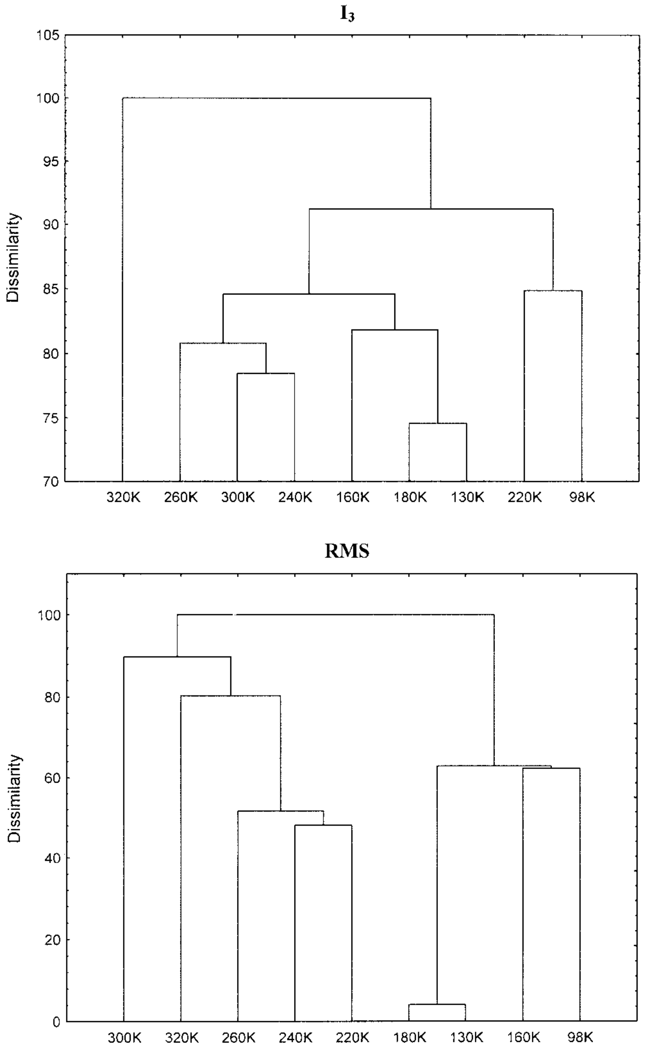


Fig. 5. Similarity/dissimilarity (Euclidean distance, complete linkage) of ribonuclease-A at nine different temperatures using folding degree (T_3) and chain superposition (RMS).

the uniqueness of the structure resolved at the intermediate temperature of 220K is only explained by the biphasic behavior observed by the thermal expansion of this protein, whose change of phase occurs around this temperature.[40] It has been stated that at this temperature a characteristic change occurs in the coordination shells of water molecules bound to the surface of the protein.[41,42] This pattern, however, is not reproduced by the similarity obtained by the 3D superposition of these structures also shown in Figure 5. The differences between these two methods for the analysis of similarity are analyzed in the following section.

3D Protein Similarity

In the last section we used the amino acid contribution to the folding degree to analyze the similarity/dissimilarity of a protein crystallized at different temperatures. The same approach can be used to analyze the similarity/dissimilarity among different proteins of the same family or even to compare different proteins pertaining at different protein families. For the sake of simplicity we will give an example here in which we compare the dissimilarity between six proteins of the same size, previously studied by Fleming and Richards.[16] The comparison of protein chains of different lengths could be done in a similar way to which the superposition of these structures is carried out, that is by making an alignment and then a superposition of the aligned sequences.

The PDB entries for the lysozymes studied here are: 4LYT(A), 2IHL, 135L, 1HHL, 1LMN, and 2EQL. They correspond to the egg white of hen (*Gallus gallus*), Japanese quail (*Coturnix coturnix japonica*), turkey (*Meleagris galloparo*), guinea fowl (*Numida meleagris*) as well as to rainbow trout (*Oncorhynchus mykiss*) kidney and horse (*Equus caballus*) milk, respectively.

Here we use three different method for analyzing the similarity/dissimilarity of these six proteins. The first is based on the alignment of the primary structures of the proteins. In doing so we used the ALIGN algorithm (see Methods) comparing the sequences from two by two and producing a similarity matrix whose elements are the percentage of identical residues in the proteins compared. The second method is based on the superposition of the structures and the calculation of their RMS error as the similarity index. The superposition is carried out for all pairs of structures and it is clear that a RMS of zero represents an absolute identity between the structures. As the values of RMS increase the similarity between the corresponding structures decreases, i.e., increase in dissimilarity. Finally, we calculated the Euclidean distances between all pairs of proteins using a vector representation of their folding degree. That is to say, a vector in which each entry corresponds to the normalized folding degree of the corresponding amino acid in the protein. As before, the lower the Euclidean distance the higher the similarity between the corresponding proteins. The results of the all three methods are given in Table IV.

According to the results of the alignment experiment the most similar structures correspond to 2IHL and 4LYT(A) (95.3%) as well as 4LYT(A) and 135L (94.6%). In fact, these three proteins together with 1HHL form a cluster of similarity with similarities over 85%. We recall that these four proteins are those corresponding to the birds, i.e., hen, Japanese quail, turkey, and guinea fowl. However, when we analyze the similarity produced by the superposition of the structures we can see that the most similar pair corresponds to 4LYT(A)-1LMN, that is the proteins corresponding to hen and to rainbow trout. These two proteins form a cluster together with 1HHL and 135L with RMS under 1.00.

The most interesting comparison, however, is that between the two 3D similarities as it is well known that 2D similarity is not necessarily related to the 3D one. When we analyze the similarities produced by the amino acid contributions to folding degree index we see that the most similar pair is that corresponding to 135L and 1HHL, that is, the proteins of Japanese quail and turkey. They form a cluster together with 2IHL and 1LMN, but the last one shows greater Euclidean distances with 135L and 2IHL. It is interesting that this clustering does not include the

TABLE IV. Similarity and Dissimilarities Between Lysozymes of Different Species as Obtained by Three Different Methods

	2IHL	135L	1HHL	1LMN	2EQL
Sequence similarity by ALIGN					
4LYT(A)	**95.3**	94.6	91.5	60.5	49.2
2IHL		92.2	88.4	60.5	48.5
135L			86.8	59.7	46.9
1HHL				58.9	49.2
1LMN					50.0
RMS superposition					
4LYT(A)	1.033	0.629	0.882	**0.622**	1.787
2IHL		0.997	1.224	1.095	1.957
135L			0.716	0.813	1.762
1HHL				0.985	1.875
1LMN					1.797
Folding degree similarity					
4LYT(A)	15.0	14.5	15.1	16.0	20.6
2IHL		11.2	11.5	13.4	21.3
135L			**10.2**	12.8	19.8
1HHL				11.9	20.3
1LMN					19.8

protein 4LYT(A), which has distances with the rest greater than those in the cluster. All the three methods coincide in recognizing the protein 2EQL corresponding to the horse as the most dissimilar of all the six proteins studied.

It is clear that both 3D method of protein similarity gives different results for the same series of proteins. This means that they account for different structural similarities in the proteins and hopefully that they could be used to help solve different problems related to protein similarity. The values of RMS of the superposition is known to be particularly affected by compactness, i.e., if two random structures are more compact than others they will have a lower RMS. We have shown in our previous paper that compactness is something the folding degree is not measuring, i.e., compactness and folding degree are independent as shown by the complete lack of correlation between I_3 and the radius of gyration for 60 different proteins.[21] Our main interest here is to provide an explanation of the similarities rank produced by the amino acid contributions to the folding degree index. In doing so we have selected the TOPS diagram for each protein,[34–37] which is a comparatively simple level of analyzing the protein topology. This is a highly simplified description of the protein fold that includes only the sequence and secondary structure elements as well as their relative spatial positions and approximate orientation. In Figure 6 we give the TOPS cartoons for the six proteins studied here. As we can easily observe, the cartoons for 135L and 1HHL are almost identical corresponding with the high similarity found by the folding degree method presented here. We recall that these two proteins formed a cluster with 2IHL and 1LMN, which have similar TOPS diagrams to 135L and 1HHL. However, 4LYT(A) shows a different picture in the fold according to its TOPS cartoon, additionally 2EQL is easily recognized as a unique pattern compared to those of the rest of proteins. These two features are also well recognized in a numerical way by the folding degree similarity. In Table IV we can see the highest dissimilarities of 4LYT(A) and 2EQL based on the folding degree similarity index. Accordingly, for the case of the proteins analyzed, the amino acid contributions to the folding degree and the

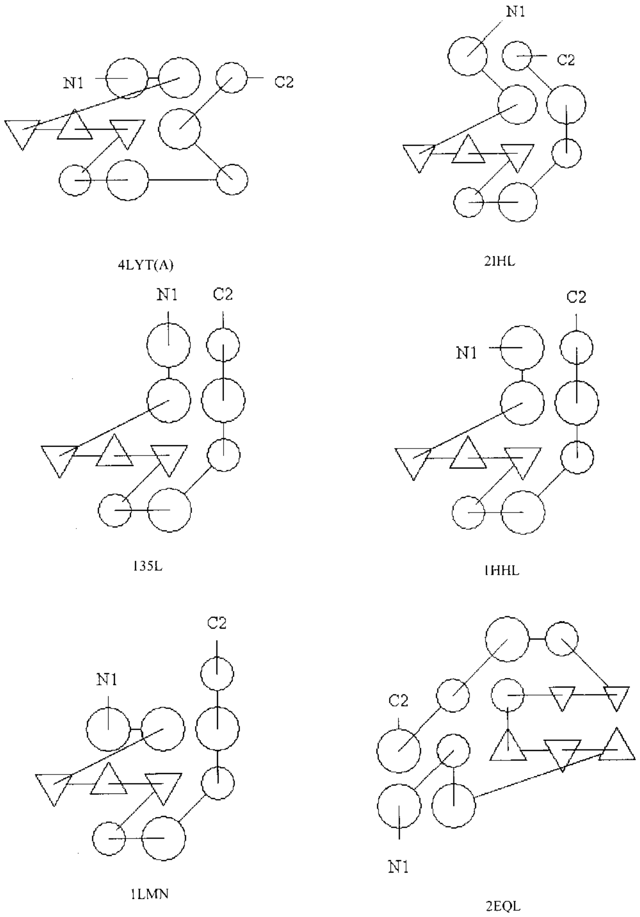


Fig. 6. TOPS cartoons for the six lysozymes studied in the current work.

TOPS cartoons are like the two sides of a coin. The first gives a numerical representation of the global folding of the proteins and the second one gives their graphical representations. More work is necessary, however, in order to extend this observation to a general rule valid for all proteins. However, we hope this will be the case taking into consideration the nature of both methods, which try to represent the topology of a protein using information coming from their primary and secondary (and supersecondary) structures.

CONCLUSIONS

We have introduced a method for quantifying the contribution of each amino acid to the global folding degree of a protein. This method is based on the mathematical properties of the I_3 index previously developed as a measure of folding degree of a protein chain. We have shown here that the amino acid contribution to the folding degree is able to account for the differences in folding of the different elements of secondary structures in proteins. One of the potential applications of this approach is in helping the assignation of amino acids to specific secondary structure or as a way for checking this assignation carried out by other methods. The last application was exemplified by the analysis of one of the π-helices assigned by Fodje and Al-Karadaghi.[39]

Another important application of the approach developed here is the analysis of the amino acids response to external factors in terms of their contribution to the protein folding degree. External factors are understood here as any external stress that can affect the protein folding, such as temperature, pH, ligand binding, interaction with other proteins, DNA, etc. In the current work we studied the influence of the temperature on the amino acids folding degree, showing that the current approach is able to account for the main changes that are taking place in the protein.

Finally, we have shown that the current approach is useful in extracting 3D structural information of proteins that is valuable in analyzing protein similarity. We have analyzed the similarity between some 3D structures of proteins using their amino acid contributions to the folding degree and we have compared them with 2D similarity as well as 3D similarity obtained by structure superposition. The extension of this approach to the similarity analysis of proteins of different chain lengths is straightforward and will provide a tool for the analysis of protein function and evolution on the basis of their folding degree.

ACKNOWLEDGMENTS

Thanks are given to the program Ramon y Cajal of the Ministerio de Ciencia y Tecnologia, Spain for partial financial support. Constructive comments by an anonymous referee contributed to a significant improvement of the results presented here.

REFERENCES

1. Veselovski AV, Ivanov YD, Ivanov AS, Archakov AI, Lewi P, Janssen P. Protein-protein interactions: mechanisms and modification by drugs. J Mol Recognit 2002;15:404–422.
2. Sinha N, Smith-Gill SJ. Protein structure to function via dynamics. Protein Peptide Lett 2002;9:367–377.
3. Luque I, Leavitt SA, Freire E. The linkage between protein folding and functional cooperativity: Two sides of the same coin? Annu Rev Biophys Biomol Struct 2002;31:235–256.
4. Kanelis V, Forman-Kay JD, Kay LE. Multidimensional NMR methods for protein structure determination. IUBMB Life 2001;52: 291–302.
5. Werten PJL, Remigy HW, de Groot BL, Fotiadis D, Philippsen A, Stahlberg H, Grubmuller H, Engel A. Progress in the analysis of membrane protein structure and function. FEBS Lett 2002;529:65–72.
6. Rost B. Review: Protein secondary structure prediction continues to rise. J Struct Biol 2001;134:204–218.
7. Eddy SR. Multiple-alignment and sequence searches. Bioinformatics: A Trend Guide 1998;5:15–18.
8. Setubal J, Meidanis J. Introduction to Computational Molecular Biology. Boston: PWS; 1997. 296 p.
9. Wang JTL, Shapiro BA, Shasha D, editors. Pattern discovery in biomolecular data. Tools, techniques, and applications. Oxford: Oxford University Press; 1999. 251 p.
10. Willet P. Searching for pharmacophoric patterns in databases of three-dimensional chemical structures. J Mol Recognit 1995;8:290–303.
11. Wade RC, Gabdoulline RR, De Rienzo F. Protein interaction property similarity analysis. Int J Quantum Chem 2001;83:122–127.
12. Arteca GA. Overcrossing spectra of protein backbones: characterization of three-dimensional molecular shape and global structural homologies. Biopolymers 1993;33:1929–1941.
13. Arteca GA, Mezey PG. The shapes of backbones of chain molecules: three-dimensional characterization by spherical shape maps. Biopolymers 1992;32:1609–1612.
14. Richards FM. Areas, volumes, packing and protein structure. Ann Rev Biophys Bioeng 1977;6:151–176.
15. Pattabiraman N, Ward KB, Fleming PJ. Occluded molecular surface: analysis of protein packing. J Mol Recognit 1995;8:334–344.
16. Fleming PJ, Richards FM. Protein packing: dependence on protein size, secondary structure and amino acid composition. J Mol Biol 2000;299:487–498.
17. Randic M, Krilov G. Characterization of 3D sequence of proteins. Chem Phys Lett 1997;272:115–119.
18. Randic M, Krilov G. On the characterization of the folding of proteins. Int J Quantum Chem 1999;75:1017–1026.
19. Byatutas L, Klein DJ, Randic M, Pisanski T. Foldedness in linear polymers: a difference between graphical and Euclidean distances. DIMACS Series Disc Math Theor Comp Sci 2000;51:39–61.
20. Estrada E. Characterization of 3D molecular structure. Chem Phys Lett 2000;319:713–718.
21. Estrada E. Characterization of the folding degree of proteins. Bioinformatics 2002;18:697–704.
22. Estrada E. Application of a novel graph-theoretic folding degree index to the study of steroid-DB3 binding affinity. Computat Biol Chem 2003;27:305–313.
23. Aldous JM, Wilson RJ. Graphs and Applications. London: Springer; 1999. 444 p.
24. Estrada E. Generalized spectral moments of the iterated line graph sequence. A novel approach to QSPR studies. J Chem Inf Comput Sci 1999;39:92–95.
25. Estrada E. Spectral moments of the edge adjacency matrix in molecular graphs. 1. Definition and applications to the prediction of physical properties of alkanes. J Chem Inf Comput Sci 1996;36: 844–849.
26. Estrada E. Spectral moments of the edge adjacency matrix in molecular graphs. 2. Molecules containing heteroatoms and QSAR applications. J Chem Inf Comput Sci 1997;37:320–328.
27. Estrada E, Molina E. Novel local (fragment-based) topological molecular descriptors for QSPR/QSARR and molecular design. J Mol Graph Modell 2001;20:54–64.
28. Kannan N, Vishveshwara S. Identification of side-chain clusters in protein structures by a graph spectral method. J Mol Biol 1999;292:441–464.
29. Brida KV, Kannan N, Vishveshwara S. Analysis of homodimeric protein interfaces by graph-spectral methods. Protein Eng 2002;15: 265–277.
30. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig

H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.

31. Hooft RWW, Sander C, Vriend G. The PDBFINDER database: A summary of PDB, DSSP and HSSP information with added value. CABIOS 1996;12:525–529.
32. Kabsh W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers1983;22:2577–2637.
33. Hyperchem, Version 6.0. HyperCube Inc. Gainesville, FL; 2000.
34. Gilbert D, Westhead D, Viksna J, Thornton JM. A computer system to perform structure comparison using TOPS representations of protein structure. Comp Chem 2001;26:23–30.
35. Gilbert D, Westhead DR, Nagano N, Thornton JM. Motif-based searching in TOPS protein topology databases. Bioinformatics 1999;15:317–326.
36. Westhead DR, Slidel TWF, Flores TPJ, Thornton JM. Protein structural topology: automated analysis, diagrammatic representation and database searching. Protein Sci 1999;8:897–904.
37. Westhead DR, Hutton DC, Thornton JM. An atlas of protein topology cartoons available on the World Wide Web. Trends Biochem Sci 1998;23:35–36.
38. Fersht A. Structure and mechanism in protein science. New York: W.H. Freeman, 1999. 631 p.
39. Fodje MN, Al-Karadaghi S. Occurrence, conformational features and amino acid propensities for the pi-helix. Protein Eng 2002;15: 353–358.
40. Tilton RF, Dewan JC, Petsko GA. Effects of temperature on protein structure and dynamics: X-ray crystallographic studies of the protein ribonuclease A at nine different temperatures from 98 to 320K. Biochemistry 1992;31:2469–2481.
41. Fraunfelder H, Hartmann H, Karplus M, Kuntz ID, Kuriyan J, Parak F, Petsko GA, Ringe D, Tilton RF, Connolly MML, Max N. Thermal expansion of a protein. Biochemistry 1987;26:254–261.
42. Peterson-Kennedy SE, McGourty JL, Kalweit JA, Hoffman BM. Temperature dependence and ligation effects on long-range electron transfer in complementary [Zn,Fe^{III}] haemoglobin hybrids. J Am Chem Soc 1986;108:1739–1746.