PAPER

NJC

www.rsc.org/njc

Quantitative characterization of protein structure: application to a novel α/β fold

Goran Krilov[a] **and Milan Randić**[b]

[a] *Department of Chemistry, Columbia University, 3000 Broadway, New York, NY 10027 USA. E-mail: krilov@chem.columbia.edu*
[b] *3225 Kingman Rd., Ames, IA 50014 USA. E-mail: mrandic@msn.com*

Received (in St. Louis, USA) 5th April 2004, Accepted 28th June 2004
First published as an Advance Article on the web 5th November 2004

We apply a computational scheme, in which geometrical (through-space) and topological (through-bond) distances are combined to construct geometry-dependent structural D/D matrices, in order to develop a matrix invariant characterization of Top7, a *de novo* sequence design by Kuhlman *et al.* (B. Kuhlman, G. Dantas, G. C. Ireton. G. Varani, B. L. Stoddard and D. Baker, *Science*, 2003, **302**, 1364), which was shown to exhibit a novel α/β globular protein fold. A sequence of invariants ${}^{K}\boldsymbol{\Phi}$, known as the "folding profile" of a protein, are extracted from the D/D matrix and its related higher order representations, and a similarity measure is introduced allowing invariant-based quantification of similarity between protein forms. Quantitative comparison of the Top7 crystal structure with the computationally designed model is achieved for the full protein backbone and a specific substructural segment: Lys^{46}–Tyr^{76}. The approach appears particularly attractive as it allows both global comparisons, as well as characterization of local structural features of proteins.

1. Introduction

Molecular size and shape are the two most critical structural elements of molecules that determine many of their physico-chemical properties. Molecular size can often be well characterized by the number of atoms or the molecular weight, but molecular shape remains elusive to quantitative characterization. Although shape is a generally understood concept, quantification of the same is, however, far from straightforward. Therefore, a number of auxiliary shape features have been evoked, such as branching,[1] cyclicity,[2] complexity,[3] compactness,[4] flexibility,[5] and folding,[6–11] which are more susceptible to numerical characterizations. Hence we can speak of the degree of branching, the level of cyclicity, the degree of complexity, the degree of flexibility and therefore, the degree of folding of a molecule. One of the earliest non-trivial characterizations of molecular structure, recognized today as being based on graph theoretical concepts, was proposed by Harry Wiener.[4] He was able to construct good linear correlations for selective physico-chemical properties of alkanes and a number of alkane derivatives, using a graph theoretical notion of the so-called "topological" distance. "Topological" distance is in fact the graph theoretical distance measured over the molecular graph, in which the metric is defined in terms of the number of edges separating the two vertices.[12,13] Over time, particularly since the 1970s, numerous mathematical structural invariants have been suggested as molecular descriptors.[14–17] Moreover, it has been shown that some of these invariants (often somewhat incorrectly referred to as "topological indices" instead of "graph theoretical indices") can adequately account not only for molecular branching, cyclicity, complexity and similar property of molecular graphs, but also certain aspects of 3-D molecular structure,[18] as well as some elements of molecular shape.[19,20]

Our interest here is in numerical characterization of folded chain structures, of which proteins and DNA are the most prominent examples due to their biological importance. In particular we will consider the quantitative characterization of folded protein structures. We would like to stress that the well known "problem of protein folding" and our "characterization of folded proteins" are two distinct aspects of the study of protein shapes. In the former, one is interested in predicting protein geometry from a given list of primary amino acids, in the latter for a given protein of known geometry one is interested in characterizing its geometrical shape by a set of mathematical invariants. Here we will consider invariants that in addition to being sensitive to the assumed 3-D molecular structure, can also be interpreted as numerical parameters describing the degree of local folding in a structure. Thus, while "folded protein structure" is the *output* of the research into the folding of proteins, (*i.e.*, the "problem of protein folding"), "folded protein structure" is an *input* for our "folded protein structure" characterization.

There have been several approaches reported in the literature that utilize a combination of through-space and topological distances to characterize the properties of protein sequences. Plaxco *et al.*[21] have demonstrated a strong relationship between contact order (defined as the average sequence distance between all pairs of contacting residues normalized to the sequence length) and the folding rates of single domain proteins, with the proteins characterized by lower contact order exhibiting faster folding. A similar behavior was found by Gromiha and Selvaraj,[22] who showed that the folding rates correlate very well with the long range order parameter (LRO; defined as the average number of contacts between residues close in space but far apart in sequence), further demonstrating the important role of long range contacts in slowing down the folding rates. In a separate study, Dosztányi *et al.*[23] used neural networks to identify and predict sets of long range interacting residues, called stabilization centers, which have a particularly important role in stabilizing protein structures. The statistical analysis of a substantial database of proteins showed high structural and sequence conservation of stabilization centers

DOI: 10.1039/b405153j

over protein families, and pointed out their role in the formation of folding nuclei.

Estrada has recently introduced a folding degree index computed from the spectral moments of the weighted adjacency matrix of the third line graph representing the cosines of dihedral angles of the protein chain.[24,25] In a series of papers he applied this approach to characterizing the binding affinity of antibodies,[26] secondary structure assignments and 3-D similarity,[27] and classification into domain classes and quantification of structural changes due to crystal packing and temperature.[28] Arteca, Tapia and coworkers have used descriptors derived from the probability distribution of projected bond-bond crossings[29–31] to characterize the fold diversity among proteins of equal chain length[32] and analyze molecular dynamics trajectories of proteins in vacuum.[33]

In this contribution we will examine a 92-residue globular α/β protein called Top7, which was designed through a novel computational sequence design method recently introduced by Baker *et al.*[34] Their method allows for optimal design of a sequence with an arbitrary fold topology, and the computational prediction of the 3-D structure of this protein. Top7 exhibits a novel fold topology that is not present in any of the proteins with structures currently available in the data bank, which makes it interesting to study from a structural point of view.[34] Moreover, the protein was successfully synthesized and the crystal structure was obtained, which was shown to be in excellent agreement with the computational model, with an overall RMSD of 1.2 Å.

2. Degree of folding of a molecule

We will outline the approach that leads to numerical characterization of molecular shape based on the local degree of folding, by reproducing the folding index Φ for the nine conformations of chains of six sp^2-hybridized CC bonds illustrated in Fig. 1. The data was taken from a review on graph theoretical characterization of 3-D molecular structures.[8] The nine structures considered are members of a set of possible conformations that can be superimposed on a hexagonal graphite lattice without overlapping of vertices. The numbers shown under each of the structures in Fig. 1 are the leading eigenvalues of the distance-distance D/D matrices divided by N, the number of vertices in the chain. The (i, j) matrix element of a D/D matrix is defined as the quotient of the geometrical and the graph theoretical (or topological) distance between the vertices i and j. The notion that the leading eigenvalue of a D/D matrix should parallel the degree of folding of a structure follows from the fact that if a structure is more folded, there will be numerous elements of the D/D matrix that are relatively small, since the through-space distance will be much smaller than the graph theoretical distance. As a consequence, the row sums of the matrix belonging to the more folded structure will be smaller than the row sums of the corresponding D/D matrix of a less folded structure. However, it is well known from matrix algebra that the leading eigenvalue of a matrix is bounded by the largest and the smallest row sum of the matrix. Thus, the smaller matrix elements will also result in a smaller leading eigenvalue of the matrix, which allows for an interpretation of the leading eigenvalue of the D/D matrix as an index of the degree of folding of a chain structure.

In Table 1 we illustrate the D/D matrices for the first and the last structures of Fig. 1. The first structure is the least folded chain, hence the corresponding D/D matrix shows relatively large numerical values, whereas the last structure is the most folded chain of Fig. 1, with its D/D matrix showing many elements that have relatively small magnitudes. Below each matrix in Table 1 we have included the average row sums, the leading eigenvalues, and the bounds of the leading eigenvalues for the two extreme cases of Fig. 1.

We should add that using a single index as molecular descriptor will most often be insufficient to discriminate among structures that may be visibly different but may have a very similar leading eigenvalue of the D/D matrix. For instance, this is the case with the two structures near the end of Fig. 1, with the values of the folding index of 0.697 and 0.696, respectively. As we see from Fig. 1, both structures have a similar value of the Φ index, but their appearance is distinct. Therefore, it appears desirable to have additional indices to supplement the information on molecular folding. One way to derive such additional descriptors is to consider closely related ${}^k(D/D)$ matrices, in which individual matrix elements of D/D matrix are raised to a power k. One can then use the leading eigenvalues of ${}^k(D/D)$ matrices, again normalized on the size of the chain. This leads to a sequence of indices ${}^K\boldsymbol{\Phi} = \{{}^1\Phi, {}^2\Phi, \ldots {}^k\Phi, \ldots {}^{K-1}\Phi, {}^K\Phi\}$, which we will refer to as the folding profile of the structure. In the following section, we have applied these concepts to derive a more satisfactory characterization of the Top7 fold. In addition, we compare our findings to those of Estrada,[26–28] whose recently introduced protein folding degree index parallels our approach.

3. Folding of Top7 protein

We examined two models of the Top7 protein: the first is the model based on the X-ray crystal structure solution of the protein that is available from the PDB[41] (under the code 1qys) and the second is the structure based on the optimized computational model of the same protein provided by the Baker group.[34] We will refer to the two models briefly as the "crystal" model and the "*in silico*" (structure designed on a computer) model, respectively. In this study, we focus on the backbone structure with a residue level resolution, which determines the topology characteristic of a particular fold. However, the method that we use is easily extendable to an all-atom model of the protein as well, if desired. The positions of each of the residues are represented by the coordinates of the correspond-

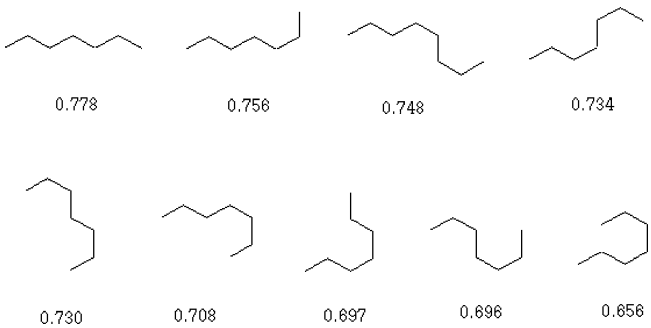

Fig. 1 Nine possible conformers of the seven carbon alkane chain superimposed on a hexagonal lattice, with no overlapping atoms. The values of the folding index Φ computed from the corresponding 7×7 D/D matrices are shown below each structure. The first structure is the least folded chain with the largest value of Φ, while the last structure is the most folded and hence exhibits the smallest value of Φ.

Table 1 The D/D matrix for the first, least folded chain (top) and the last, most folded chain (bottom) of Fig. 1. The rows and columns list the ratios between the topological and geometrical distances for the pairs of carbon atoms labelled 1–7. The unit of spatial distance is normalized to the length of the C–C bond, which defines the spacing of the hexagonal grid. The right-most column in each table shows the row sums of matrix elements for the particular row. Under each section of the table are included the average row sums, the leading eigenvalues, and the bounds of the leading eigenvalues for the two extreme cases of Fig. 1

	1	2	3	4	5	6	7	Row sum
Least folded chain								
1	0	1/1	$\sqrt{3}/2$	$\sqrt{7}/3$	$\sqrt{12}/4$	$\sqrt{19}/5$	$3\sqrt{27}/6$	5.351
2		0	1/1	$\sqrt{3}/2$	$\sqrt{7}/3$	$\sqrt{12}/4$	$\sqrt{19}/5$	5.485
3			0	1/1	$\sqrt{3}/2$	$\sqrt{7}/3$	$\sqrt{12}/4$	5.479
4				0	1/1	$\sqrt{3}/2$	$\sqrt{7}/3$	5.495
5					0	1/1	$\sqrt{3}/2$	5.479
6						0	1/1	5.485
7							0	5.351
Average row sum: 5.447; leading eigenvalue: $\lambda_1 = 5.446$; bounds: $5.352 < \lambda_1 < 5.496$								
Most folded chain								
1	0	1/1	$\sqrt{3}/2$	2/3	$\sqrt{3}/4$	1/5	$\sqrt{3}/6$	3.454
2		0	1/1	$\sqrt{3}/2$	2/3	$\sqrt{3}/4$	$\sqrt{7}/5$	4.494
3			0	1/1	$\sqrt{3}/2$	2/3	3/4	5.148
4				0	1/1	$\sqrt{3}/2$	$\sqrt{7}/3$	5.280
5					0	1/1	$\sqrt{3}/2$	4.831
6						0	1/1	3.299
7							0	4.315
Average row sum: 4.404; leading eigenvalue: $\lambda_1 = 4.592$; bounds: $3.300 < \lambda_1 < 5.281$								

ing C_α atoms. We construct the D/D matrix for the two protein models by computing the ratios of geometric and topological distances between pairs of C_α atoms. The topological distance is determined by the number of residues that separates the two pair members on the protein chain. The geometric distances were expressed by reduced distances obtained by dividing all distances by the maximum separation between neighboring C_α atoms in the protein chain. The reason behind this normalization is that in this way, all elements of the D/D matrix are equal to or less than 1, which ensures convergence for the elements of the higher order matrices. Subsequently, we generate the sequence of higher order ${}^k(D/D)$ matrices for $k = 2$ through 15. The leading eiegenvalues ${}^k\lambda$ of these matrices are computed and normalized to generate the folding profiles ${}^K\boldsymbol{\Phi}$, where $K = k_{\max}$ and ${}^k\Phi = {}^k\lambda/N$, N being the dimensionality of the ${}^k(D/D)$ matrix.

The numerical values for the folding profiles corresponding to the crystal structure and the computational *in silico* model of Top7 are shown in the first two columns of Table 2. In Fig. 2(b) we have plotted the numerical folding indices against the exponent k. One can observe from Table 2 that the differences in the "degree of folding" between the actual X-ray structure and computational *in silico* model are rather small. Hence, the computed folding characterization of Top7 protein provides an additional numerical confirmation of the close similarity between the two models, which is apparent from the low RMSD of the superimposed structures shown in Fig. 2(a).[34,35] A close look at Fig. 2(b) shows that for increasing values of k, the computed folding profiles show but a minor difference. Moreover, inspection of the plot of Fig. 2(b) for higher values of the exponent k indicates that the *in silico* model shows a slightly slower convergence of ${}^k\Phi$ as $k \rightarrow \infty$. This means that the computational model corresponds to a somewhat less compactly folded structure. The differences are very small but nevertheless they allow us to speculate that the native structure reflected by the folding profile of the "crystal" model is able to achieve better packing and to form a more compact structure, than the minimum energy *in silico* model predicted computationally. In other words, perhaps not surprisingly, even though modelling techniques have advanced considerably over the last few years, nature still does a better job (in folding structures) than we are able to do computationally! If this trend is found to hold for other protein structures, we would have an important novel hypothesis for the 3-D structures of proteins: "Protein 3-D geometry is such that they achieve the maximum possible folding in the minimum energy conformation." In other words: Proteins fold as much as possible but not more!

The relationship between the energetics of protein structures and the degree of folding was also observed by Estrada in his study of steroid-DB3 antibody affinity.[26] In particular, he showed that the principal determinant of the binding affinity is the change in compactness and folding increment of the two antibody chains. This effect was also observed in more general studies by Getzoff *et al.*,[36] Pellequer *et al.*,[37] and Coulon *et al.*[38] and is likely related to the hydrophobic interactions between the steroids and DB3.[26] While the above studies support our hypothesis, a more extensive analysis of proteins from multiple families is necessary before we can ascertain the generality of this trend.

Table 2 The values for the folding profile of the Top7 protein. Each row corresponds to ${}^k\Phi$, the folding index derived from the k-th order matrices ${}^k(D/D)$. The first two columns show the profiles of the complete 92-residue backbone for the crystal structure and computational *in silico* model, respectively. The last two columns show the profiles of Lys^{46}–Tyr^{76}, the 30-residue core α/β hairpin segment, for the crystal structure and the *in silico* model

	Full chain		Lys^{46}–Tyr^{76}	
k	Crystal	*in silico*	Crystal	*in silico*
1	0.25994	0.26287	0.4440	0.4509
2	0.13154	0.13423	0.2863	0.2937
3	0.09888	0.09968	0.2339	0.2421
4	0.08287	0.08397	0.2043	0.2134
5	0.07150	0.07313	0.1821	0.1920
6	0.06273	0.06484	0.1642	0.1747
7	0.05570	0.05822	0.1492	0.1602
8	0.04994	0.05283	0.1365	0.1480
9	0.04515	0.04836	0.1256	0.1374
10	0.04113	0.04462	0.1162	0.1283
11	0.03771	0.04146	0.1080	0.1204
12	0.03479	0.03898	0.1009	0.1134
13	0.03229	0.03685	0.0947	0.1074
14	0.03013	0.03500	0.0892	0.1021
15	0.02839	0.03338	0.0844	0.0974

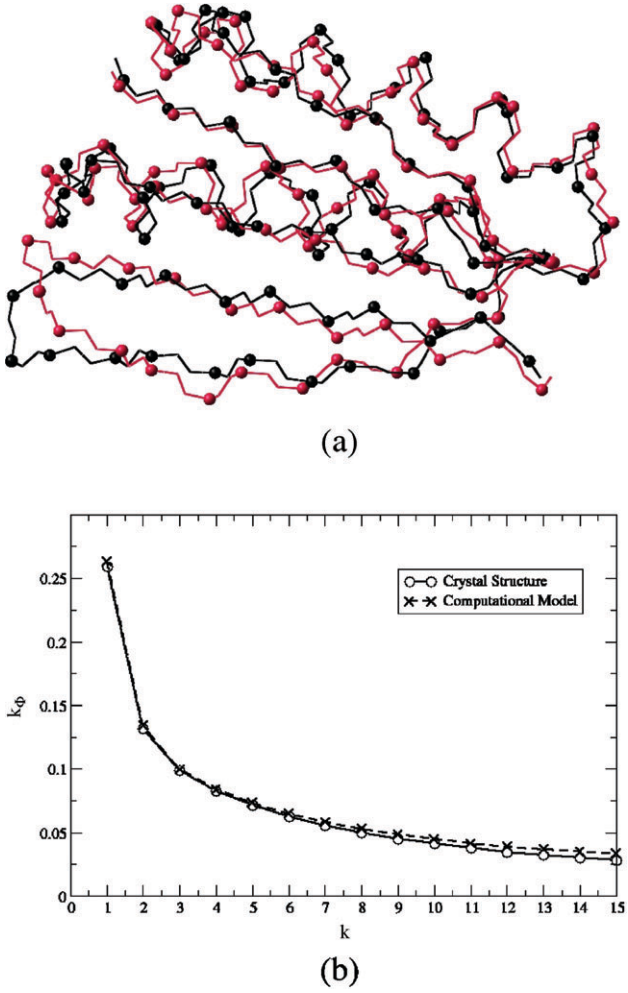

Fig. 2 The folding profiles for the full 92-residue Top7 chain. In (a) we show the superimposed backbone models of the Top7 crystal structure (black) and the computationally predicted *in silico* structure (red). The positions of C_α used to compute the $^k(D/D)$ matrices are shown as spheres. The folding profiles $^K\boldsymbol{\Phi}$ for the crystal structure (solid line and empty circles) and the *in silico* predicted structure (broken line and crosses) are shown in (b). The excellent overlap of the superimposed structures and the close match of the respective folding profiles both point to highly similar structures.

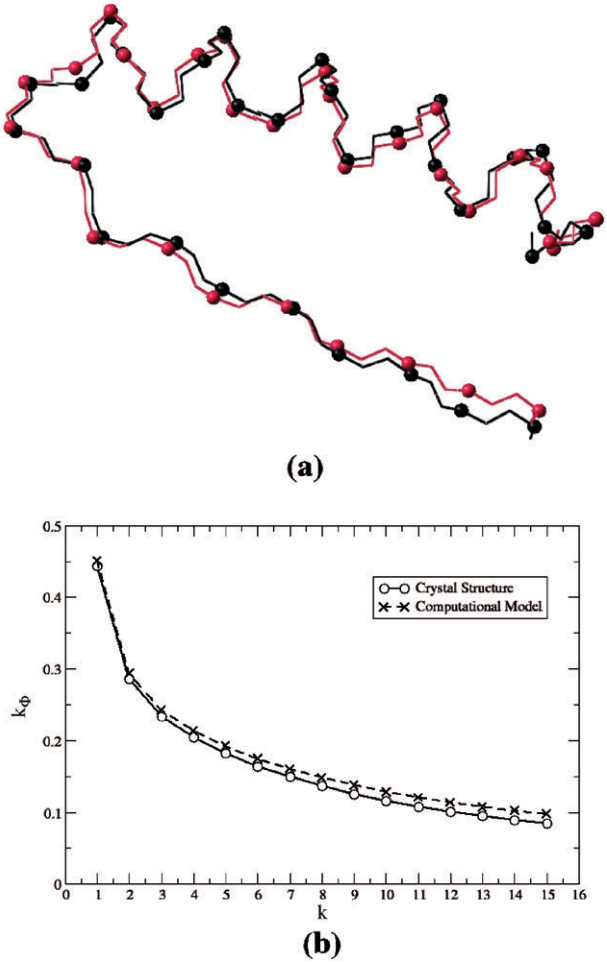

Fig. 3 The folding profiles for the 30-residue core α/β hairpin segment of the Top7 chain, consisting of residues Lys^{46}–Tyr^{76}. In (a) we show the superimposed backbone models of the crystal structure (black) and the computationally predicted *in silico* structure (red) of the segment. The positions of C_α used to compute the $^k(D/D)$ matrices are shown as spheres. The folding profiles $^K\boldsymbol{\Phi}$ for the crystal structure (solid line and empty circles) and the *in silico* predicted structure (broken line and crosses) are shown in (b). Although the superimposed structures show a remarkable degree of overlap, the folding profiles are better able to capture the small differences between the two models in this smaller structure.

4. Local folding features

The outlined computational procedure for characterization of protein folds is quite general in the sense that it applies to large and small proteins. The products of our analysis are "global" invariants of a chain structure that have been extracted from the D/D matrices, which to a degree reflect the "average" matrix properties. Hence, it is clear that as the size of the matrices increases, one is not likely to get more and more specific indices but rather profiles that are less and less sensitive to differences in underlying structures—all due to the fundamental "averaging" process of extracting suitable matrix invariants. The latter was also noted by Estrada, who was able to recover some of the lost information by analyzing the local contributions to his folding degree index by individual amino acids.[27] We anticipate, therefore, that D/D matrices may be more informative for characterizing *local* protein features, such as individual topological motifs of proteins or structures of active sites, which consist of fewer residues and are associated with smaller D/D matrices.

We tested the above premise by considering the core segment of Top7, which consists of the residues Lys^{46}–Tyr^{76}, comprising a well defined α/β hairpin motif.[34] The computational model closely matches the crystal structure, which is evident from the superimposed backbone structures shown in Fig. 3(a). In the last two columns of Table 2 we show the $^k\Phi$ values for both models of this core segment, while in Fig. 3(b) the magnitude of the folding index is plotted against the exponent k. Again we see that there is close overlap of the computed indices, which confirms that the two hairpin motifs, the "natural" and the "predicted", show a very similar degree of folding. Likewise, we observe that the natural structure is somewhat more compactly folded, indicating a slightly better packing of the hydrophobic core than obtained computationally. A comparison of data in Table 2 with the global folding profile indicates that the α/β hairpin motif is a less folded structure than the full Top7 chain, with the leading index of the folding profile being 0.444 *vs.* 0.260. Since the segment has a considerably higher helix content than the full chain (60% *vs.* 39%), this is in apparent contrast to Estrada's observations that helix residues constitute the largest per-residue contribution to local folding.[27] However, we note that, although related, our definition of the folding profile is different from that of Estrada. While the latter emphasizes local folding, such as that of the coils of a helix, the folding profile derived from D/D matrices is more sensitive to long range interactions between topologically distant residues, and hence lends more weight to tertiary fold features. The folding of individual

topological motifs is perhaps of more interest than the global fold of a protein. Nevertheless, we consider the global folding profile as an important characterization of proteins, since different proteins may have individual local structural motifs of similar form. For example, it has been pointed out that although Top7 is a protein of novel fold, it "differs from natural folds only by the topological connections of the secondary-structural elements—the order in which the helices and strands are linked together to each other".[35] Hence, a numerical characterization provides an additional tool to identify similar structural elements. This idea was also utilized by Estrada, who showed in a recent study how the folding degree can be used to discriminate between protein families, hence allowing for the possibility of automatic classification of proteins into structural domain classes.[28]

5. On the similarity of protein forms

An important aspect of the characterization method illustrated in this study is that it provides the means for quantifying the similarity between protein structures. To accomplish the latter, a measure of similarity between matrix invariants needs to be defined. One can consider a folding profile sequence ${}^{K}\boldsymbol{\Phi} = \{{}^{1}\Phi, {}^{2}\Phi \ldots {}^{k}\Phi, \ldots {}^{K-1}\Phi, {}^{K}\Phi\}$, as components of a k-dimensional vector. The degree of similarity between the profiles is then defined as the Euclidean distance between the folding profile vectors:

$$d_K = \sqrt{\sum_{k=1}^{K} \left({}^{k}\Phi - {}^{k}\Phi'\right)^2} \qquad (1)$$

The values of d_K between the crystal structure and the computational *in silico* model of Top7 backbone for the full protein and the core region described in the previous section were computed for several values of k. The results are shown in Fig. 4. The distance measure shows an almost linear increase with k for the core region whereas it depends only weakly on k when considering the full protein. The latter is likely due to additional averaging performed in extracting invariants from the larger structure.

Comparing only a single pair of structures (the "crystal" and the "*in silico*" models) of the Top7 protein without reference to data (currently not available) on other proteins of similar size allows at best only limited conclusions to be drawn on the degree of similarity or dissimilarity between the two models. Nevertheless, we are confident that the present calculations can be interpreted as pointing to two highly similar protein forms. This confidence is based on the result obtained earlier[7,9] for a set of "toy" proteins of Li *et al.*,[39] model proteins consisting of 27 points embedded in a $3 \times 3 \times 3$ cube illustrated in Fig. 5. In Table 3 we show the similarity/dissimilarity table for the five "toy" proteins, labelled A–E, based on 10-component vectors constructed from ${}^{k}(D/D)$ matrices using the average row sums as matrix invariants, which were computed in a previous study.[7,9] A different coding of the folding of model proteins A–E was described using the geometry of three successive edges in graphs, which may represent covalent bonds or folding directions.[40] From Table 3, we find very small entries for pairs (B, E) and (C, D). While the similarity between B and E is not apparent and may be fortuitous (which is always possible when invariants are used for characterization of structures) the similarity between C and D appears plausible, giving credence to the computed level of similarity based on the folding invariants. Moreover, all the entries relating to B, apparently the structure with the most different folding, are relatively large. It is apparent from the data in Table 3 that character-

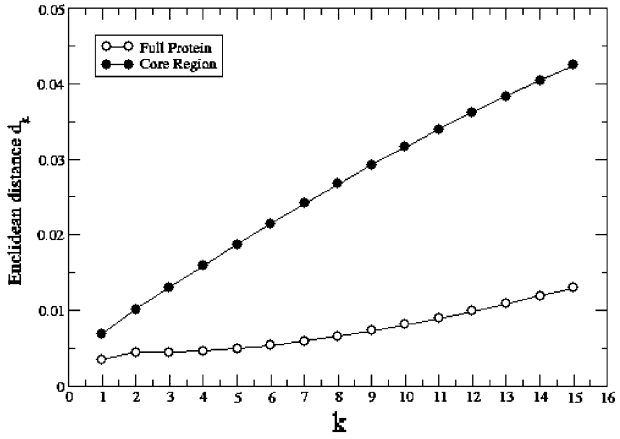

Fig. 4 The Euclidean distance between the folding profiles of the crystal structure and the *in silico* model for the full 92-residue Top7 chain (empty circles) and the 30-residue core α/β hairpin segment (solid circles). The distance measure shows an almost linear increase with k for the core region whereas it depends only weakly on k when considering the full protein.

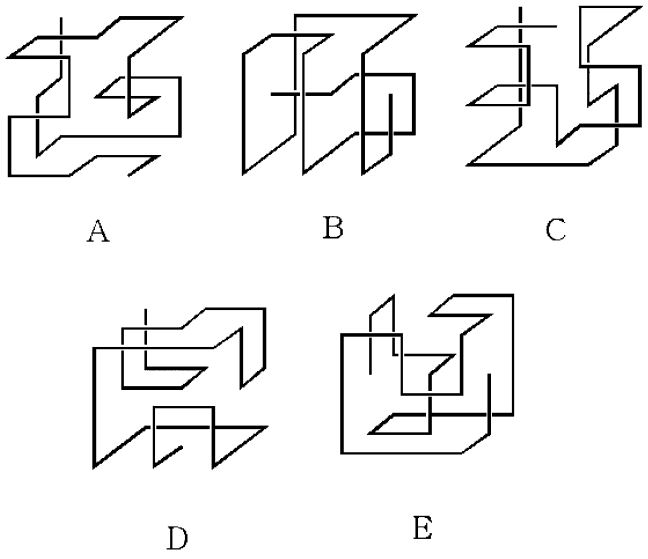

Fig. 5 The five degenerate minimum energy conformations of the 27-residue lattice model protein of Li *et al.*[39] The residues have a binary (hydrophilic-hydrophobic) model and the conformations are restricted to a cubic lattice.

Table 3 The similarity/dissimilarity of the 27-residue model protein structures superimposed on the cubic grid shown in Fig. 5. The five proteins are labelled A–E and the similarity is computed from the Euclidean distances between the molecular profiles derived from distance matrices.[9] Note the small entries for the distances between the (B, E) and (C, D) pairs, signifying a high degree of similarity using this measure

	A	B	C	D	E
A	0	0.1373	0.0644	0.0666	0.1376
B		0	0.0730	0.0708	0.0006
C			0	0.0023	0.0732
D				0	0.0711
E					0

ization of structural models of long chains of different folding patterns is sufficiently sensitive to variations in skeletal forms. Hence, the small difference in the "distance" between the ${}^{K}\Phi$ vectors for the crystal model and the *in silico* model conclusively points to small variations in their geometry.

One should note that the similarity measure based on the folding profile may lead to a different conclusion than a more common measure of structure similarity based on the RMSD. This is due to the fact that the folding profile contains both distance and connectivity information, whereas the RMSD measure is based on the geometry alone. This was also pointed out by Estrada in his study of similarity between lysozymes of different species, where he contrasted the results based on the RMSD similarity measure with the ones obtained from a Euclidean distance measure derived from a sequence of higher order spectral-moment-based folding degrees.[27]

6. Concluding remarks

In the present article we illustrated a numerical characterization of folded protein structures that can be subsequently used for storing information on proteins, similarity searching and clustering, and for constructing structure-function relationships. It may be too early to speculate that the geometrical forms assumed by proteins are those of maximal folding. However, at least in the case of globular proteins, that goal may parallel a more obvious structural feature, that of minimizing the overall outside surface with respect to volume, in an effort to bury the hydrophobic residues. This is supported by recent studies by Estrada and others who have noted a strong correlation between the energy of the folded structure and degree of compactness.[26,36–38] Nonetheless, a comprehensive analysis of a large database of protein structures is necessary before one can establish this relationship. We plan to address this question as well as the applicability of the folding profile to characterize different families of folds in a future study. However, of more interest for the comparative study of proteins and their functionality are local structural features of proteins. In particular, a large number of diverse protein structures can be decomposed into a limited number of structural motifs. A rational quantification of folding in these substructures would be very helpful for classification, manipulation, storage and comparison as well as structure-function analysis of proteins. The present approach indicates that our folding profile approach equally applies to local features of proteins, here illustrated by the characterization of the core α/β hairpin segment. However, further work is necessary to fully validate these claims, which we plan to report in a future study.

The main limitation of the folding profile method in its current application is the computational cost involved in extracting eigenvalues of large matrices, which can be considerable for very large proteins. In addition, as with any invariant approach, there is some loss of information associated with representation of a structure by a small set of descriptors that constitute the folding profile. Hence, although we demonstrated the sensitivity of our model to small structural fluctuations, it remains to be seen how successful it will be, for example, in characterizing the structural details of active sites that strongly influence the bioactivity of proteins.

Acknowledgements

We would like to thank Prof. Bruce J. Berne of Columbia University for providing computational resources and sponsoring this work. This work was supported by NIH grant GM-4330 to Bruce J. Berne

References

1 M. Randić, *J. Am. Chem. Soc.*, 1975, **97**, 6609.
2 T. Pisanski, D. Plavšić and M. Randić, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 520.
3 S. H. Bertz, *J. Am. Chem. Soc.*, 1981, **103**, 3599.
4 H. Wiener, *J. Am. Chem. Soc.*, 1947, **69**, 17.
5 L. B. Kier, *Quant. Struct.-Act. Relat.*, 1989, **8**, 218.
6 M. Randić, L. M. DeAlba and A. F. Kleiner, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 277.
7 M. Randić and G. Krilov, *Chem. Phys. Lett.*, 1997, **272**, 115.
8 M. Randić and M. Razinger, in *From Chemical Topology to Three-Dimensional Geometry*, ed. A. T. Balaban, Plenum Press, New York, 1997, pp. 159–236.
9 M. Randić and G. Krilov, *Int. J. Quantum Chem.*, 1999, **75**, 1017.
10 L. Bytautas, D. J. Klein, M. Randić and T. Pisanski, *DIMACS Series: Discrete Math. Theor. Comput. Sci.*, 2000, **51**, 39.
11 M. Randić, M. Vračko, M. Novič and S. C. Basak, *MATCH Commun. Math. Chem.*, 2000, **42**, 181.
12 F. Harary, *Graph Theory*, Addison–Wesley, Reading, MA, 1969.
13 F. Buckley and F. Harary, *Distance in Graphs*, Addison–Wesley, Redwood City, CA, 1990.
14 M. Randić, in *The Encyclopedia of Computational Chemistry*, eds. P. V. R. Scheleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer III and P. R. Schreiner, John Wiley & Sons, Chichester, UK, 1998, pp. 3018–3032.
15 A. T. Balaban, in *Topological Indices and Related Descriptors in QSAR and QSPR*, eds. J. Devillers and A. T. Balaban, Gordon and Breach, Amsterdam, The Netherlands, 1999, pp. 403–453.
16 *Topological Indices and Related Descriptors in QSAR and QSPR*, eds. J. Devillers and A. T. Balaban, Gordon and Breach, Amsterdam, The Netherlands, 1999.
17 R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors: Methods and Principles in Medicinal Chemistry*, eds. R. Mannhold, H. Kubinyi and H. Timmerman, Wiley–VCH, Weinheim, Germany2000, vol. **11**, p. 1.
18 A. T. Balaban, in *From Chemical Topology to Three-Dimensional Geometry*, ed. A. T. Balaban, Plenum Press, New York, 1997, pp. 1–24.
19 L. B. Kier, *Quant. Struct.-Act. Relat.*, 1986, **5**, 1.
20 M. Randić, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 607.
21 K. W. Plaxco, K. T. Simons and D. Baker, *J. Mol. Biol.*, 1998, **277**, 985.
22 M. M. Gromiha and S. Selvaraj, *J. Mol. Biol.*, 2001, **310**, 27.
23 Z. Dostányi, A. Fiser and I. Simon, *J. Mol. Biol.*, 1997, **272**, 597.
24 E. Estrada, *Chem. Phys. Lett.*, 2000, **319**, 713.
25 E. Estrada, *Bioinformatics*, 2002, **18**, 697.
26 E. Estrada, *Comput. Biol. Chem.*, 2003, **27**, 305.
27 E. Estrada, *Proteins*, 2004, **54**, 727.
28 E. Estrada, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1238.
29 G. A. Arteca and P. G. Mezey, *Biopolymers*, 1992, **32**, 1609.
30 G. A. Arteca, *Biopolymers*, 1993, **33**, 1929.
31 G. A. Arteca, *Phys. Rev. E*, 1995, **51**, 2600.
32 G. A. Arteca and O. Tapia, *J. Chem. Inf. Comp. Sci.*, 1999, **39**, 642.
33 G. A. Arteca and O. Tapia, *Int. J. Quantum Chem.*, 2000, **80**, 848.
34 B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard and D. Baker, *Science*, 2003, **302**, 1364.
35 S. Borman, *Chem. Eng. News*,, Nov. 24, 2003, **81**, 11.
36 E. D. Getzoff, H. M. Geysen, S. J. Rodda, H. Alexander, J. A. Tainer and R. A. Lerner, *Science*, 1987, **235**, 1191.

37 J. L. Pellequer, S. W. W. Chen, V. A. Roberts, J. A. Tainer and E. D. Getzoff, *J. Mol. Recognit*., 1999, **12**, 267.
38 S. Coulon, J. L. Pellequer, T. Blachere, M. Chartier, E. Mappus, S. W. W. Chen, C. Y. Cuilleron and D. Baty, *J. Mol. Recognit*., 2002, **15**, 6.
39 H. Li, R. Helling, C. Tang and N. Wingreen, *Science*, 1996, **273**, 666.
40 A. T. Balaban and C. Rücker, *J. Chem. Inf. Comput. Sci*., 2001, **41**, 1145.
41 H. M. Berman, J. Westbrook, Z. Feng, G Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res*., 2000, **28**, 235–242.