A Simple Topological Representation of Protein Structure: Implications for New, Fast, and Robust Structural Classification

David L. Bostick,[1] **Min Shen,**[2] **and Iosif I. Vaisman**[3*]

[1]*Department of Physics and Program in Molecular/Cell Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina*

[2]*Laboratory for Molecular Modeling, School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina*

[3]*Laboratory for Structural Bioinformatics, School of Computational Sciences, George Mason University, Manassas, Virginia*

ABSTRACT **A topological representation of proteins is developed that makes use of two metrics: the Euclidean metric for identifying natural nearest neighboring residues via the Delaunay tessellation in Cartesian space and the distance between residues in sequence space. Using this representation, we introduce a quantitative and computationally inexpensive method for the comparison of protein structural topology. The method ultimately results in a numerical score quantifying the distance between proteins in a heuristically defined topological space. The properties of this scoring scheme are investigated and correlated with the standard C_α distance root-mean-square deviation measure of protein similarity calculated by rigid body structural alignment. The topological comparison method is shown to have a characteristic dependence on protein conformational differences and secondary structure. This distinctive behavior is also observed in the comparison of proteins within families of structural relatives. The ability of the comparison method to successfully classify proteins into classes, superfamilies, folds, and families that are consistent with standard classification methods, both automated and human-driven, is demonstrated. Furthermore, it is shown that the scoring method allows for a fine-grained classification on the family, protein, and species level that agrees very well with currently established phylogenetic hierarchies. This fine classification is achieved without requiring visual inspection of proteins, sequence analysis, or the use of structural superimposition methods. Implications of the method for a fast, automated, topological hierarchical classification of proteins are discussed. Proteins 2004;56:487–501.**

Key words: **protein topology; hierarchical classification; protein structure comparison; computational geometry; protein evolution; computational geometry; Delaunay tessellation**

INTRODUCTION

The Motivation for Comparing Proteins

Perhaps the most popular tenet of proteomic biology is that the available structure space of proteins is much smaller than the available sequence space. That is, the mapping of a given protein structure to its sequence is not isomorphic.[1–4] Given that it is the goal of most of the world's "protein-ologists" to ultimately generate structural models for new protein sequences without invoking the laborious and time-consuming task of systematic experimental structure determination, it is difficult to know whether this lack of isomorphism comes as a blessing or a curse. On one hand, there are relatively few structures to which a given sequence might fold, thus boiling the task of predicting its structure down to what is known as "structure recognition"—guessing the sequence's structural category based on its similarity to the category's sequence. On the other hand, once an estimated model structure is generated, one cannot help but to question whether the "recognized" structure is the true native structure, since it is possible that the sequence similarity between a protein and its "category" or template might not imply a structural similarity in all cases.[3] Confronted with this dilemma in associating protein sequence with structure (and function) one is forced to focus on ways of characterizing relationships between analogous and remotely homologous proteins.[4–7]

Computational tools for the comparison of three-dimensional (3D) protein structures provide both experimental and theoretical biologists with means to more closely relate a given structure to its sequence, and to rationalize structural and mechanistic investigation of protein function. They also provide a means to organize the thousands of known protein structures, identify new types of protein architecture, and draw important evolutionary relationships between proteins.[5] These goals ulti-

*Correspondence to: Iosif I. Vaisman, Laboratory for Structural Bioinformatics, School of Computational Sciences, George Mason University, Manassas, VA 20110. E-mail: ivaisman@gmu.edu.

Received 14 October 2003; Accepted 7 January 2004

Published online 20 May 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20146

mately culminate in the formation of a hierarchical phylogenetic classification of proteins.

The classification of any set of objects into categories by similarity requires the establishment of a measure of similarity between pairs of these objects. Providing a fast, quantitative automated means of measuring similarity between protein pairs without any allusion to residue identity is the niche of such computational tools.[8]

Geometric Comparisons Versus Topological Comparisons

The traditional quantitative method for measuring similarity between protein pairs, introduced by Remington and Mathews,[9] involves the optimal superimposition of their structures' backbone coordinates by applying rigid body rotations and translations. Typically, the measure of similarity in this case is taken to be the distance root-mean-square deviation (RMSD) between alpha-carbons of the proteins' respective backbones after their structural alignment.[5,10,11] This sort of method provides a sound geometric basis for structural similarity and has been used in the construction of phenetic classifications of proteins.[5,8,12] However, topological information is sometimes lost due to the fact that it is possible for pairs of proteins to display common structural elements with disjoint backbone connectivity.[11,13] Therefore, in using such a method, one must cope with the fact that geometric equivalence is not identical to topological equivalence. Gap sizes in protein structural alignments are optimized in the assignment of equivalent residue pairs and a minimum RMSD between corresponding α-carbons of the aligned path is attained that sometimes does not reflect the similarities in the protein fold. Much work has been done to evade this drawback. Some studies extended the method of structural superimposition in order to provide comparisons between proteins that better reflect their topological differences (for example, the work of Falicov and Cohen[14]).

In order to successfully hierarchically classify proteins, it is generally understood that a method for pair-wise protein comparison that makes use of protein topological differences (as opposed to geometric ones such as those captured by structural alignment) is necessarry.[4,15] Hence the task of automated protein comparison should involve the decomposition of protein structure into global features representing topology whose elements can be compared.[15] Since protein structures are so complex, there are a variety of ways to represent any single protein's topology. Just as the answer to any question depends upon the manner in which the question is formulated, the comparison of proteins will be dependent upon the manner in which the proteins are represented. This simple truth has manifested itself in the emergence of several hierarchical classifications of proteins,[8] each one utilizing a different "measure" or combination of measures of protein similarity. Some rely on human expertise[16]—knowledge of function and visual inspection of structures as well as numerical measures of structural similarity and sequence homology[17] to determine classification. Such a classification is quite robust and often serves as a source for generating population statistics on the structures of proteins. Other classification methods rely on a combination of human-driven and automated means.[18,19] In the extreme case, a method of classification will make use of solely automated comparison of proteins.[5,20] Ideally, one would like to be able to rely on computer-automated comparison alone for a meaningful classification. However, since objective criteria for topological similarity have not yet been identified,[4] this is a work in progress. Such an objective topological similarity measure is necessary and sufficient for the assignment of two protein structures to the same fold.[4]

Representing the Sequence Measure as an Embedment of the Structural Euclidean Measure

In this work, we propose a method for comparing protein pairs that employs a representation based solely upon the topology of the protein core. The comparison is performed by classifying four-body clusters of nearest neighboring C_α atoms (in Euclidean space) according to their implicit topological significance. Information on the four bodies' separation in primary sequence (sequence space) is used with no allusion to residue identity. Other studies have investigated spatially neighboring residues and their relation to structural elements. Their focus has been on establishing how neighboring amino acid properties or identities affect the geometry of structural elements, or on relating sets of neighboring residues to particular secondary structures or protein conformations.[21,22] Brocchieri and Karlin[23] present a way of categorizing pairs of neighboring residues in proteins according to their separation along the primary sequence and investigating the pairs of residues (identity, hydrophobicity, charge, etc.) that occur in these categories. Another protein structure comparison method that combines geometric and topological aspects was developed by Carugo and Pongor.[24] In their work, the distribution of C_α–C_α distances between residues at given separations in primary sequence is used as a structural descriptor. While these methods can elucidate the link between tertiary folds and the physical constraints on the protein chain, pairs of nearest neighboring residues are not sufficient for the characterization of global topological structure.

The goal in concocting a topological description of a protein must inevitably reduce to linking the information pertaining to the closeness of residues in Euclidean space with the closeness of residues in sequence space. In this sense, the problem of topologically describing a protein is a problem of "topological embedding".[25] In light of this fact, we chose to represent the spatial closeness of residues in a protein by investigating the graph of the 3D Delaunay tessellation of its set of protein C_α atomic coordinates. In general, the tessellation of a set of points in any space is a division of the space into convex polytopes. The Voronoi tessellation, a construct that is dual to the Delaunay tessellation, of a set of points divides the space into space-filling convex polyhedra (in the case of three dimensions) that define the region of space closest to each point. A point in this tessellation whose Voronoi polyhedron

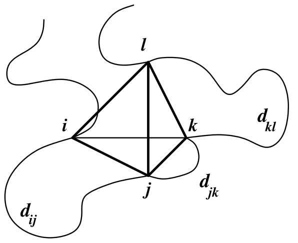

Fig. 1. **a**: Schematic diagram of the contribution of a single simplex to the 1000-tuple representation. The three far edges determine the element to which the simplex makes a statistical contribution.

shares a vertex with three other polyhedra is a natural nearest neighbor to the three points to which these polehedra belong. These nearest neighboring quadruplets define the vertices of space-filling tetrahedra. These tetrahedra are the convex polyhedra (simplices) of the Delaunay tessellation of the set of points. A Delaunay approach to identifying nearest neighboring clusters of residues has been described previously.[26,27] With this approach, we are left with the task of relating natural nearest neighboring residues in 3D Euclidean space with the closeness (or farness[28]) of residues in sequence space. The metric used to define the distance, itself, in sequence space between any two residues, i and j, in a protein is provided naturally by simply counting the number of residues falling between i and j.

MATERIALS AND METHODS

Building a Raw Topological Representation

Using the information from the Delaunay tessellation of a protein's backbone, it is possible to build a statistical representation of that protein, which takes into account the way its sequence must "twist and turn" in order to bring each four-body residue cluster into contact. Each residue—*i, j, k,* and l of a four-body cluster comprising a simplex are nearest neighbors in Euclidean space as defined by the tessellation, but are separated by the three distances—d_{ij}, d_{jk}, and d_{kl}in sequence space (Fig. 1). Based on this idea, we build a 1000-tuple representation of a single protein by making use of two metrics: (1) the Euclidean metric used to define the Delaunay tessellation of the protein's C_α atomic coordinates and (2) the distance between residues in sequence space. A procedure similar to our previous work[27] was followed in the construction of our protein representation. We recapitulate the similarities in this work for brevity.

If we consider a tessellated protein with N residues integrally enumerated according to their position along the primary sequence, the length of a simplex edge in sequence space can be defined as

$$d_{ij} = j - i - 1 \tag{1}$$

where d_{ij} is the length of the simplex edge, $\overline{ij}$, corresponding to the ith and jth α-carbons along the sequence. If one considers the graph formed by the union of the simplex edge between the two points i and j and the set of edges between all d_{ij} points along the sequence between i and j, it is seen that the Euclidean simplex edge, $\overline{ij}$, can generally be classified as a far edge.[28] Every simplex in the protein's tessellation will have three such edges associated with its vertices: *i, j, k,* and l where *i, j, k,* and l are integers corresponding to C_α atoms enumerated according to their position along the primary sequence (see Fig. 1). Thus, we proceed to quantify the degree of "farness" in an intuitive way, by applying a transformation, T, which maps the length, d, of each edge to an integer value according to

$$T{:}d \;\rightarrow \begin{cases} 1 \; if \; d = 0 \\ 2 \; if \; d = 1 \\ 3 \; if \; d = 2 \\ 4 \; if \; d = 3 \\ 5 \; if \; 4 \leq d \leq 6 \\ 6 \; if \; 7 \leq d \leq 11 \\ 7 \; if \; 12 \leq d \leq 20 \\ 8 \; if \; 21 \leq d \leq 49 \\ 9 \; if \; 50 \leq d \leq 100 \\ 10 \; if \; d \geq 101 \end{cases} \tag{2}$$

The reasoning behind the design of the transformation is described elsewhere.[27] It is possible that a coarser categorization of the edge length, d, (for example, placing the edge length possibilities into 5–8 categories instead of 10) might be adequate for an effective protein representation. Indeed, the empirical optimization of the step function, T, will be required in order to test this hypothesis. However, for this work we have chosen to keep the "fine-grained" transformation of Equation (2), as in our previous work[27] in order to provide rigorous testing of our protein representation scheme without optimizing the transformation for coarseness. With this transformation, we construct an array that is representative of the distribution of combinations of segment lengths along the protein backbone giving rise to nearest-neighboring four-body residue clusters within the protein's structure as defined by the tessellation of its C_α atomic coordinates. Each simplex in the protein's tessellation contributes to a 3D array, M, where M_{npr} is the number of simplices whose edges satisfy the following conditions:

(a) The Euclidean length of any one simplex edge is not greater than 10 Å.
(b) $d_{y} = n$
(c) $d_{jk} = p$
(d) $d_{kl} = r$

Condition (a) is provided because simplices with a Euclidean edge length above 10 Å are generally a result of the positions of α-carbons on the exterior of the protein. We exclude contributions from these simplices to M, because they do not represent clusters of natural nearest neighbors due to the absence of solvent and other molecules around the protein in the tessellation. Figure 2 shows an example of a tessellated protein excluding simplices meeting condition (a) and a representative simplex along with its three segment lengths, d_{ij}, d_{jk}, and d_{kl} (as shown schematically in Figure 1). The data structure, M, contains 1000 elements. The number of elements is invariant with respect

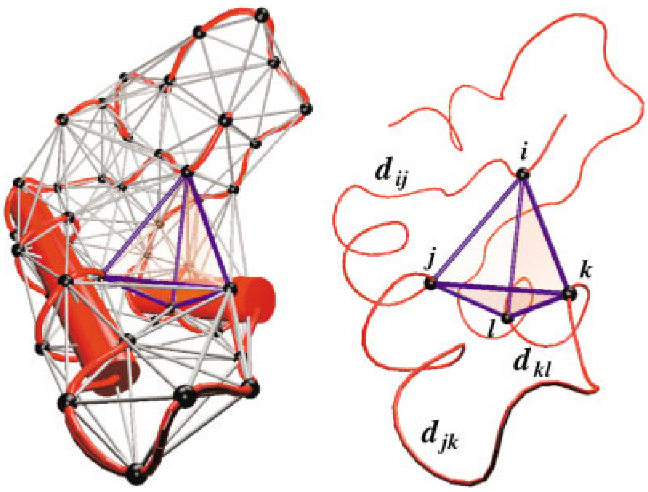

Fig. 2. **Left**: The resulting graph of the protein, Crambin (pdb identity 1ccn) after the removal of simplices from its Delaunay tessellation having an edge length greater than 10 Å. The remaining simplices' edges are represented as silver rods while the α-carbons are represented by black spheres. **Right**: Illustration of the contribution of a single representative simplex in Crambin to its 1000-tuple representation.

to the number of residues of the protein. In order to more easily conceptualize the mapping of the protein topology to the data structure, M, we rewrite it as a 1000-tuple vector:

$$\vec{M} = \{M_{000}, M_{001}, \ldots, M_{010}, M_{011}, \ldots, \ldots, M_{999}\} \qquad (3)$$

Raw and Normalized Scoring Schemes for Protein Comparison

Given that each element of this vector represents a statistical contribution to the global topology, a comparison of two proteins making use of this mapping must involve the evaluation of the differences in single corresponding elements of the proteins' 1000-tuples. We define, therefore, a raw topological score, Q, representative of the

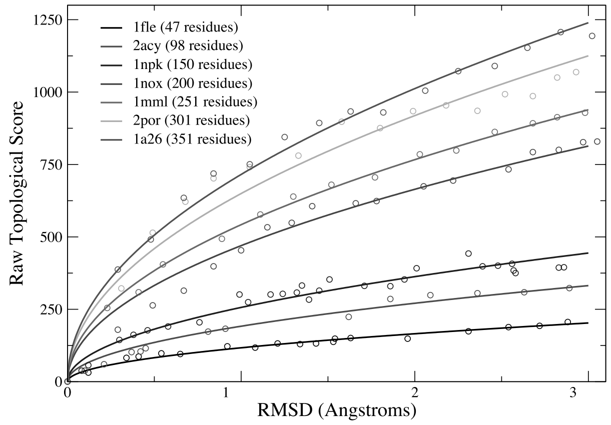

Fig. 3. Correlation of the raw topological score, Q, with standard C_α RMSD from the MD trajectory (in vacuum) of various proteins having different sequence length. For each trajectory, the initial structure (at t = 0) is compared to several other conformations in the trajectory up to an RMSD of ~3 Å. The trend-lines drawn are fitted to the data using a simple power-law expression ($f(x) = a\sqrt{x}$, where a is a fitting parameter) with an average correlation coefficient of ~0.99.

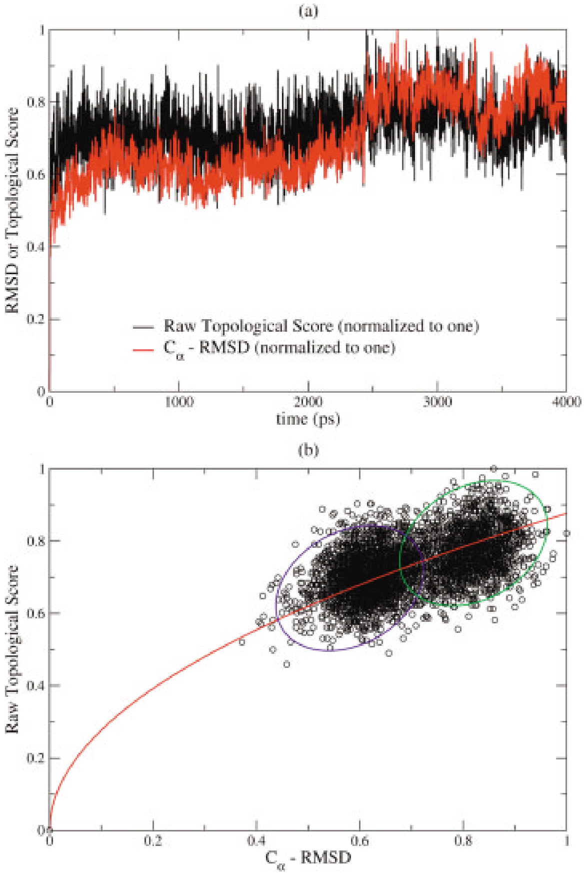

Fig. 5. **a**: Time-series of the raw topological score (black), Q, and RMSD (red) resulting from the MD simulation of halorhodopsin in a fully hydrated DPPC membrane. Each protein in the series is compared with the initial structure (at t = 0). The topological score and RMSD data are normalized according to the greatest value in the set in the interest of providing a clear comparison. Thus, the largest value in each respective set is unity. **b**: Correlation of topological score with RMSD. The green circle represents the region of data after the convergence of RMSD while the blue circle represents the region before convergence.

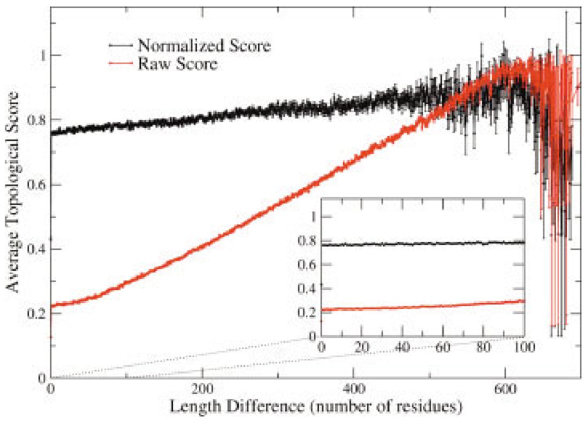

Fig. 6. Dependence of $\underline{Q}$ (black) and Q (red) on the length difference between compared proteins. Note that the standard error (represented by the error bars) in the score becomes greater in either case as the length difference between compared proteins increases.

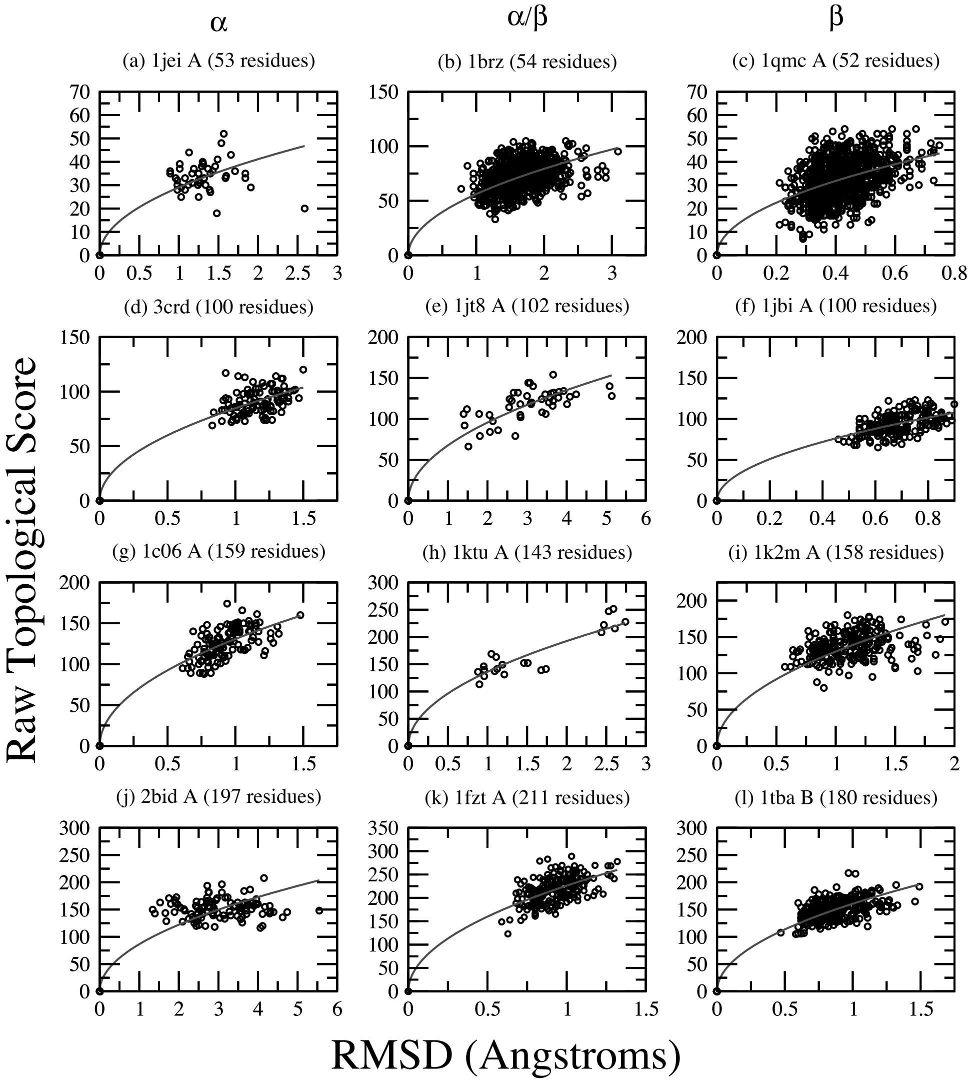

Fig. 4. Correlation of the raw topological score, *Q*, with standard C_α RMSD from twelve sets of protein configurations determined by NMR spectroscopy. The trend-lines follow the same functional form as in Figure 2 with an average correlation coefficient of ~0.90.

topological distance between any two proteins represented by data structures, $\vec{M}$ and $\vec{M}'$ as the supremum norm,

$$Q \equiv |\vec{M} - \vec{M}'|_{\sup} \equiv \sum_{i=0}^{999} |M_i - M_i'| \tag{4}$$

This norm is topologically equivalent to the Euclidean norm[25] and has the added advantage that it is less computationally expensive to calculate.

This topological score has an obvious dependence on the sequence length difference between the two proteins being compared due to the following implicit relation for a single protein representation,

$$N_s = \sum_{i=0}^{999} M_i \tag{5}$$

where N_s is the number of simplices with no edge having a Euclidian length greater than 10 Å, and the M_i are the elements of the protein representation. In other words, since N_s is proportional to the number of residues in the protein, the difference in the length between two compared proteins might provide a systematic extraneous contribution to their score, *Q*, in Equation (4). This is not to say that the sequence length of a protein does not play a role in its topology. In fact, the length should be quite crucial.[27] However, the length dependence of our score implied by Equation (5) is endemic to our protein representation (derived from its tessellation), and not due to protein topology itself. This length dependence may be removed by first normalizing the vector representation as follows:

$$\underset{\rightarrow}{M} = \frac{\vec{M}}{|\vec{M}|} \tag{6}$$

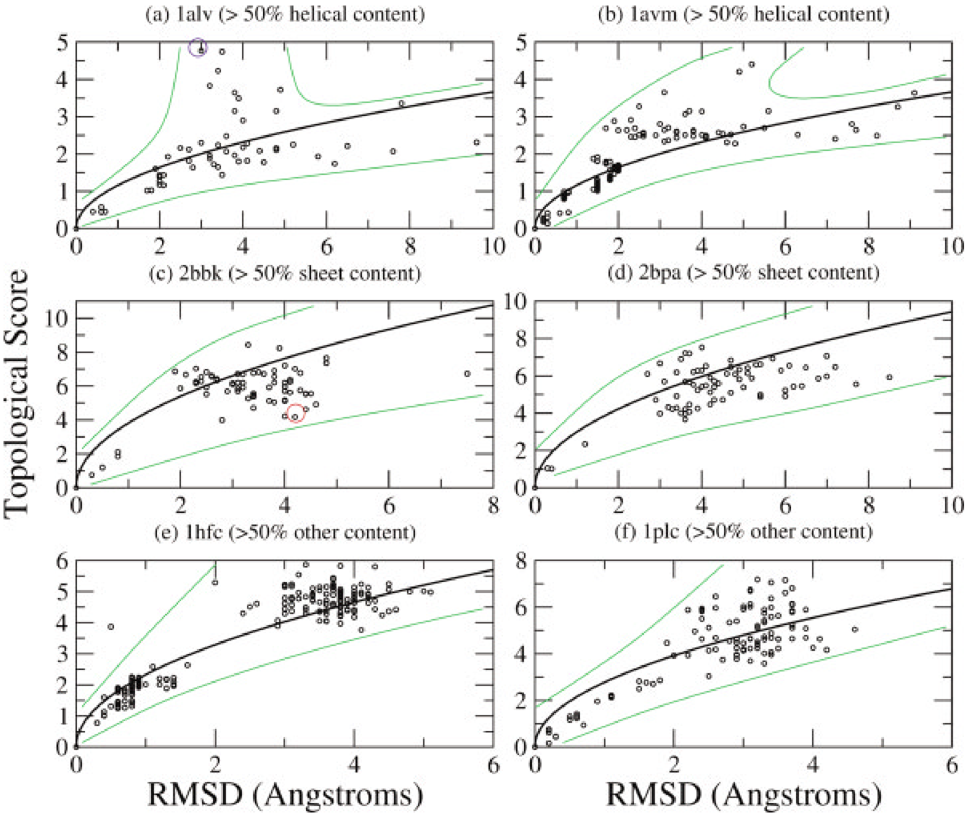

Fig. 7. The correlation of normalized topological scoring, $\underline{Q}$, with standard C_α RMSD of FSSP protein families having various proportions of secondary structure. While the power-law trend-lines can be fitted to the data very well (with an average correlation coefficient of ~0.89), the "global trend," for each plot is very distinctive for each type of secondary structure content.

resulting in the unit-vector representation, $\underset{\rightarrow}{M}$. The corresponding normalized topological score,

$$\underline{Q} \equiv |\underset{\rightarrow}{M} - \underset{\rightarrow}{M}'|_{\text{sup}} \tag{7}$$

can be expected to be less sensitive to the chain length difference between the two proteins being compared. Despite normalization, however, this score should still have an inherent dependence on the length difference between the compared proteins. A protein's structure must be dependent on the length of its sequence, because the number of configurational degrees of freedom in a polymer's structure is proportional to the number of residues it possesses. Such a dependence on the size of compared proteins is even apparent in geometric methods of comparison such as structural alignment, and in some cases, has been accounted for.[29]

We designed a set of tests for the topological scoring schemes in Equations (4) and (7) to evaluate their capability to describe changes in the topology of single proteins upon conformational change and differences in topology within families of structurally related proteins. The sensitivity due to the expected systematic length dependence [Equation (5)] of our topological comparison was also tested in both schemes. We then moved further to probe the sensitivity of our topological depiction to the secondary structure of proteins under comparison. Finally, we evaluated the capability of our protein comparison method to provide a means for the hierarchical classification of proteins. We compared this capability to that of other well known automated and human-derived classifications, and ultimately, investigated the way such a hierarchical clustering might look.

The following section presents the results from these tests of the capabilities of our topological representation of proteins. These assessments of the scoring method involve the pair-wise comparison of various sets of protein structures. All protein structures compared were taken from the PDB[30] at http://www.pdb.org. Each structure file was subjected to rigorous parseability criteria[31] before its use in our study. Protein comparisons were performed using a Pentium III, 930 MHz processor, utilizing a Linux kernel version 2.4.2-2. The programs for our comparisons and analysis were written in the C programming language. All programs performing Delaunay tessellation implemented the algorithm described by Watson.[32]

In order to observe how our scoring schemes describe changes in the topology of single proteins upon conformational change, molecular dynamics (MD) trajectories of individual proteins were generated. In this work, all protein simulations performed in vacuum made use of the MD module of the SYBYL 6.4 software package (TRIPOS Associates, St. Louis, MO). An MD trajectory of halorhodopsin in an explicit, hydrated DPPC bilayer was generated using the GROMACS package[33,34] with a time step of 2 fs. The forcefield parameters for lipids were based on the work of Berger[35] and those of the protein were provided by GROMACS. The forcefield describing the retinal moiety in the interior of the protein were adopted from previous studies.[36,37] Periodic boundary conditions were applied in all three dimensions and long range electrostatics were handled using PME.[38] The system was simulated in an NPT ensemble with a temperature of 325 K maintained using the Nose-Hoover scheme[39] with a thermostat relaxation time of 0.5 ps. The Parrinello-Rahman[40] pressure coupling scheme with a barostat relaxation time of 2.0 ps was used to maintain a pressure of 1 atm.

The halorhodopsin system consisted of 1 protein molecule, 1 chloride ion, 1 sodium ion, 172 DPPC molecules, and 10660 SPC water molecules. An initial DPPC bilayer was constructed using a protocol outlined by Tu et al,[41] and the experimentally determined structure of halorhodopsin[42] was placed in the center of this bilayer. All DPPC and water molecules that overlapped the atomic positions of the protein model were then removed from the initial hydrated bilayer. The atoms of the protein and ions were loosely restrained to their initial positions while simulating the system for 3 ns. Over this period the volume of the simulation cell converged along with the area of the bilayer–protein system (~66.5 nm^2) indicating the relaxation of the solvent and bilayer molecules around the protein. Restraints on the protein were then removed and an additional 4-ns run was performed on the system. Snapshots were saved at intervals of 1 ps throughout the calculation for subsequent analysis.

RESULTS AND DISCUSSION

Sensitivity to Conformational Differences

In order to ascertain the way in which our topological comparison scheme quantifies differences in topology upon conformational change within a single protein, MD simulations were performed on seven different proteins in vacuum. The purpose of these simulations was merely to provide a set of structures that were "distorted" from their initial crystallographic conformations. We do not claim to obtain any biologically relevant information from these simulations. The proteins were selected such that their sequence lengths covered a range of values (~50 to ~350 residues) in order to evaluate how the topological score, Q, depends on protein size.

Figure 3 shows the correlation of Q with standard C_α RMSD for various protein structures along each of the seven trajectories. Structures along the trajectories were compared to their corresponding initial configurations (experimental structures). The data points are spaced fairly evenly along the RMSD axis so as to provide a clear picture of the correlation. A very obvious correlation between the two measures of protein similarity is observed. The trend for each protein intersects the origin in the trivial case of comparing a protein to itself. It is then seen to follow a power-law as each structure becomes more deformed. This trend is explained by a difference in scaling between RMSD and Q. Generally, we should expect this difference in scaling when comparing a global topological similarity score to a score based on a local structural alignment.[27] The global topology of a protein should remain more invariant as its related structures become yet more locally dissimilar.[27] Thus, we see that with a large change in RMSD, an increasingly smaller change in Q occurs, and the resulting trend-line is monotonic and concave-down.

The topological score's dependence on the sequence length implied by Equation (5) is demonstrated clearly in Figure 3. We see that as the length increases, the topological score increases. Hence, the trend in the Q:RMSD correlation, itself, becomes more pronounced as the number of residues in a protein becomes larger. In this test of the raw score, Q, we compared a protein in each of the seven sets to a deformed conformation of itself for each data point. Thus, the trend of the score is preserved since the length difference between the two compared proteins is zero.

Twelve sets of protein conformations determined by NMR were obtained from the protein data bank (PDB) for topological comparison with the raw score. These sets were selected to observe whether or not experimentally determined conformations of a protein would show an adherence to the trend seen in Figure 3. Four sets of three proteins were selected with sequence lengths of approximately 50, 100, 150, and 200. These different levels of sequence length were selected in order to demonstrate the consistency of the length dependence of the raw score. Within these levels of sequence length, one of the proteins had mostly α-helical content, one possessed mostly β-sheet content, and the third contained a combination of α and β secondary structure.

An all-against-all comparison was carried out between all possible pairs of proteins within each of the 12 sets of conformations. The raw topological score is plotted against RMSD for each set in Figure 4. Again, it is seen that the raw topological score increases with the size of the protein. The power-law trend demonstrated in Figure 3 is also observed in Figure 4. The average correlation coefficient for the trend-lines was ~0.90. The persistence of the trend is also seen to be invariant over the secondary structure of the sets of conformations. The data are not as evenly distributed over RMSD as in Figure 3, but this is due to the fact that sets of NMR structures are more representative of a statistical ensemble. Thus, the RMSD and topological scores should be expected to vary evenly about an average value causing us to observe a more clustered set of points in the correlation plots. The trend in the correlation is still observed, because of the difference in scaling between the two different structural comparison methods.

A realistic extended time-series of structures for the protein, halorhodopsin (pdb identifier 1e12) was also analyzed in order to observe how the topological score evolved during an extensive MD equilibration. Halorhodopsin is a membrane embedded chloride ion pump with seven membrane-spanning helices. To mimic a realistic environment for this protein, it was equilibrated in a fully hydrated dipalmitoylphosphatidylcholine (DPPC) bilayer at a temperature of 325 K and a pressure of 1 atm (Bostick and Berkowitz, unpublished results, 2003). The initial protein structure in the equilibration trajectory was compared to all other structures in the time-series. A plot of the raw topological score and C_α RMSD for every structure in this analysis is shown in Figure 5(a). It can be seen that the topological score is relatively invariant when compared to the RMSD throughout the entire trajectory. The RMSD converges to its final value after ~2500 ps. Again, this supports the notion that global topology is more invariant than local geometry when a protein is limited to small conformational differences. In this case the small differences are a result of the structure's relaxation in a hydrated membrane. Figure 5(b) shows the correlation of Q with the RMSD. The data forms two clusters: points representing structures occurring before convergence of the RMSD (circled in blue) and points representing structures after convergence (circled in green). Again, the power-law trend-line fits this set of points well. Upon observing the features of the raw topological score and comparison with standard C_α RMSD, we see that the protein comparison method we describe has many properties that can be expected of a global, topological comparison. In a gapped, optimal, rigid alignment of protein backbones, the RMSD increases as local structural differences in the aligned C_α become larger. Alternatively, our measure of global topological differences will remain relatively invariant upon local structural changes. Meaningful topological differences in compared proteins would be accompanied by large jumps in the topological score.

Removal of Length Dependence From the Topological Score

There is still the issue, however, of the raw score's dependence on the sequence length difference between the proteins under comparison. The previous examples of the topological scoring capabilities were not affected by this issue, since they have involved the comparison of single proteins with different conformations of themselves. Thus, the next logical aim would be to evaluate how the removal of the length dependence of our score implied by Equation (5) affects the comparison of proteins.

A representative set of protein structures was taken from the WHATIF database.[43] This set of proteins contains 1424 chains with less than 30% sequence identity, less than 0.25 R-factor, and a resolution of less than 2.5 Å. We performed an all-against-all comparison on this set using both topological comparison scores, Q and $\underline{Q}$. The result of these comparisons is shown in Figure 6. The normalized score, $\underline{Q}$, is seen to be free of much of the length difference dependence while the raw score, Q, shows a marked dependence. Nonetheless, on average, both scores increase with the length difference of the compared proteins.

Generally, a protein's structure must be dependent on the length of its sequence, because the number of configurational degrees of freedom in a polymer's structure is proportional to the number of residues it possesses. Thus the topological differences in compared proteins should increase gradually with their length difference. The expected gradual increase is seen for $\underline{Q}$, but is more dramatic for Q. The inset in Figure 5 shows the two scores' behavior over small length differences. It can be seen that a small length difference between two compared proteins might be safely neglected in the case of the raw comparison score, Q. Despite the difference in the two scores' global dependence on the length difference, the increase in the conformational freedom that comes with a greater sequence length manifests itself similarly in both scores via the standard error. The fluctuation in the topological score becomes greater as the sequence length difference between the compared proteins becomes larger. This is because there is a large average difference in the number of conformations available to either one of the proteins under comparison.

Topological Characterization Within Families

In the previous sections, our focus was on the comparison of different conformations of the same protein. We now investigate the application of our topological characterization method to different structures within families of structurally related proteins. This requires that we make use of the normalized topological score, $\underline{Q}$, because it is free of the artificial length dependence intrinsic to the raw score.

Six protein families were selected from the FSSP (Families of Structurally Similar Proteins) database[43] for topological evaluation. We selected families that span various levels of secondary structural content. The representatives of these families are as follows: 1alv and 1avm (having greater than 50% α-helical content), 2bbk and 2bpa (having greater than 50% β-sheet content), and 1hfc and 1plc (having at least 50% content that is classified as neither α-helical nor β-sheet). The FSSP database contains the results of the alignments of the extended family of each of these representative chains. Each family in the database consists of all structural neighbors excluding very close homologs (proteins having a sequence identity greater than 70%). The topological score was calculated for each representative in a one-against-all comparison with its neighbors. All of the scores are plotted against RMSD for each of the families in Figure 7.

The power-law trend can be seen for all families, although it is less pronounced than that seen in Figure 3. Perhaps the most striking feature of the plots is the particular shape of the correlations in each secondary structural content group. These "global trends" are outlined in green on the plots. The families with mainly helical content seem to follow a common trend until ~2 Å RMSD [see Fig. 7(a, b)], after which the trend "splits" in two. This indicates that many proteins with high helical content may be well superimposed while, topologically, they may be very different. The families with mainly sheet content also display a characteristic trend. The data tend to follow the fitted line very well, with a cluster of protein pairs possessing a low RMSD. A second cluster of protein pairs is centered around 3 Å RMSD in the case of 2bbk and around 4 Å in the case of 2bpa [see Fig. 7(c, d)]. Furthermore, the data tend to more sparsely populate the region along the trend-line than in the case of proteins with more helical content. This implies that both global topology (as indicated by our score) and local geometry (as indicated by standard RMSD) change in the same abrupt manner within families having high β-sheet content. In the tested families having a content majority classified as "neither" α-helical nor β-sheet, the correlation follows a pattern possessing traits of both "highly helical" and "highly sheet" trends [see Fig. 7(e, f)]. The data populate the region along the trend-line well as in the case of highly helical families. However, there is no "split" in the trend, although there is slight deviation from the fitted line at higher RMSD (hinting at a helical behavior). A demonstration of the difference in the type of information one can glean from our topological score as opposed to a standard calculation of C_α RMSD after superimposition can be seen if we view the difference in two related proteins that are close according to RMSD but far according to topological score (with respect to the other proteins in the family). The data point corresponding to the comparison of chain A of 1alv to 1thg is encircled in Figure 7(a). These proteins and their optimal alignment as determined by the program, CE,[44] are rendered in Figure 8(a). It is easy to see that the similarity implied by the RMSD is due to the alignment of substructures. If the proteins are to be aligned, their vast difference in sequence length requires that 1alv be aligned to some fragment of 1thg. It can also be seen that while 1alv contains almost no β-sheet, 1thg contains a very large portion of β-sheet (shown in yellow). Also, by virtue of the difference in their sequence length, one should expect the two proteins to be topologically different. This difference is reflected in the topological score. An additional example is encircled in Figure 7(c) in which we compare two proteins that are relatively remote according to RMSD, but close according to topological score. Figure 8(b) shows these two proteins (2bbk, chain H and 1qlg, chain A) and their optimal alignment. Qualitatively, their topological similarity is evident. The chain of 2bbk contains joined segments of sheet and three very small fragments of helix while the chain of 1qlg contains the same sheet motif, and two small helical fragments. Their alignment, while optimal, leads to a large RMSD relative to other family members, but their topological comparison classifies them as more closely related.

Now that we have established that the topological score, in most cases, can give information about changes in local geometry just as RMSD, although by different means, we wish to say a few words about the differences in the behavior of the trend in the correlation between the topological score and RMSD in the test cases shown in Figures 3, 4, 5, and 7. Figure 3 shows cases where the topological score versus RMSD correlation follows the power-law trend almost perfectly. In these cases, the scores do not represent the sampling of an equilibrium system (a protein in solvent as in Figure 5). Instead, the initial protein configurations for the simulations (in vacuum) are protein structures that would be representative of an equilibrium if they were in solvent. Thus, as time evolves, the structures move toward their "would be" equilibrium states in a vacuum environment, but never reaching it. The plots in Figure 3, therefore, show a sampling of the differences in topology (via topological score) and geometry (via RMSD) along one possible pathway from the solvent-equilibrium state to the vacuum-equilibrium state.

The same idea applies to the correlation plots for the FSSP families in Figure 7, but in a different way. In this case, we do not view a trajectory of a single protein. Instead, we view a set of structures that sample topological and geometric possibilities in a broader manner than a set of structures from an equilibrium ensemble (i.e., we are sampling many different proteins of similar secondary structure content), but in a narrower manner than a set of structures from a non-redundant set of structures (i.e., the structures used to generate the data in Figure 6).

In the cases of the MD simulation of halorhodopsin (Fig. 5) and of the NMR structures (Fig. 4), there is a strong power-law trend in the correlation plots, but generally, the variance of both the RMSD and the topological score is much smaller than in the cases of the simulations in vacuum (Fig. 3) and of the FSSP families (Fig. 7). In the halorhodopsin simulation, the initial structure was a crystal structure in a realistic environment, which was allowed to relax and conform to the simulation conditions (i.e., temperature and pressure). If the crystal structure is very close to its equilibrium structure in the simulation conditions (unlike in the case of the MD simulations in vacuum for Figure 3), then the central limit theorem would imply that the RMSD and the topological score should each display a set of values with a roughly Gaussian distribution. Thus, since the simulation of halorhodopsin and the sets of NMR structures represent systems sampling a nearly equilibrium ensemble, the more tightly clustered sets of points in Figures 4 and 5 are reasonable. The difference in scaling between the RMSD and the topological score still produce a correlation that follows a power-law trend. In summary it's all a matter of how the structures in question sample structural space. Thus the differences among the figures showing correlation among the topological score and the RMSD are reasonable given the details of the particular test for each figure.

Hierarchical Classification of Proteins

Now that we have shown what sort of information the topological score can give us about the similarity of proteins under comparison, we focus on its classificatory ability. In order to see how well our scoring scheme reflects current hierarchical protein classifications, we devised a test to determine if a given protein's structural neighbors according to our score is consistent with that one might find in the classifications of, for example, SCOP[16] (Structural Classification of Proteins) or CATH[19] (Class, Architecture, Topology, and Homologous Superfamily). To this end, a consistency on a "general" topological level was first investigated.

A set of 995 protein chains whose corresponding PDB files met parseability criteria[31] were selected out of 3285

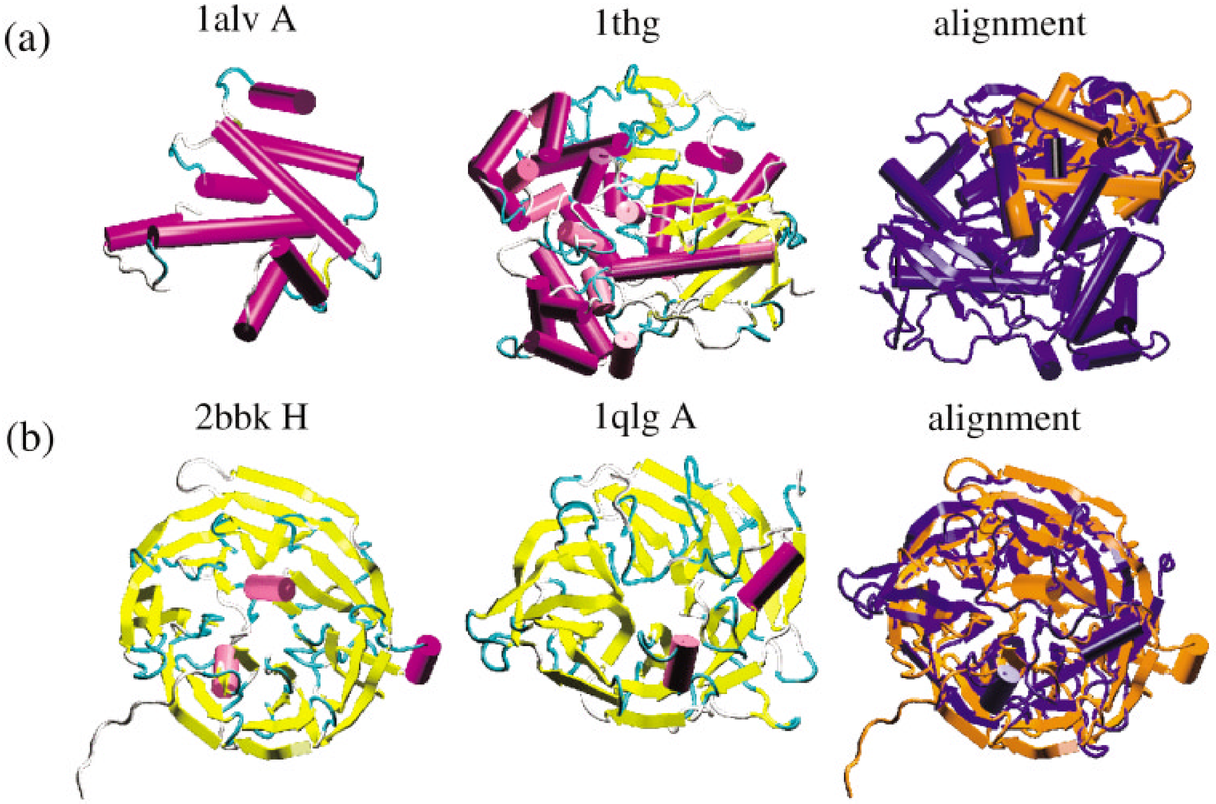

Fig. 8. Rendered structures and superimposition of the FSSP neighbors, a: 1alv, chain A and 1thg [corresponding to the data point in Figure 6(a) circled in blue] and (b) 2bbk, chain H and 1qlg, chain A [corresponding to the data point circled in red in Figure 6(c)]. The helical portions of the proteins are shown as cylinders and the sheet portions are shown as thick ribbons with arrows. Elements of secondary structure are colored: purple corresponds to helical structure, yellow corresponds to sheet structure, and light blue corresponds to loop or turn. In the alignments (lower portion of the figures), the structures 1alv A and 2bbk H are colored orange, while 1thg and 1qlg A are colored blue.

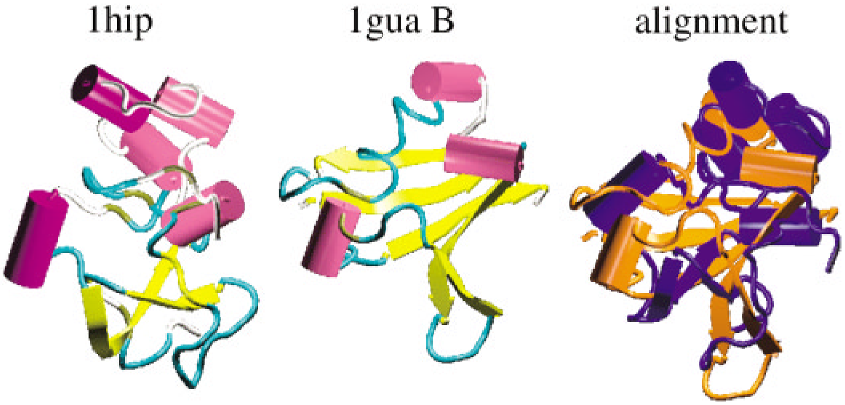

Fig. 9. Rendered structure of 1hip classified by CATH as "neither α nor β," but as "α-β" according to its neighbor, chain B of 1gua, as determined by our topological method. The details of rendering and color are the same as those for Figure 7. In the alignment (lower portion of the figure), 1hip is colored blue and 1gua B is colored orange.

chains from the CATH domain list, version 2.4, containing a set of representative domains from sequence families clustered at 35% sequence identity (available at ftp://ftp.biochem.ucl.ac.uk/pub/cathdata/v2.4/). These proteins were classified into the general classes, α, β, α-β, or neither α nor β. The following protocol was followed in order to determine the topological comparison's capacity to discriminate between the CATH designated classes: For any given protein, A, in the set,

1. Find the lowest score, Q, resulting from the topological comparison of A with all other proteins in the set. This

TABLE I. Classification of Proteins into CATH-type and SCOP-type Classes

a.	CATH class[a]	f	f/f_{random}
	α	0.957	3.75
	β	0.914	3.54
	α − β	0.826	1.86
	neither	0.333	7.92
b.	**SCOP class[b]**	f	f/f_{random}
	α	0.840	3.36
	β	0.760	3.04
	α + β	0.610	2.44
	α/β	0.730	2.92

[a] The fraction, f, of proteins classified by topology in a manner conforming to the CATH classification. f_{random} is the fraction of proteins in a given class that one would expect to classify in a manner conforming to CATH by random guessing, calculated as N_i/N_{tot}, where N_i is the number of proteins in class i (α, β, α−β, or neither) and N_{tot} is the total number of proteins classified.

[b] The fraction of proteins classified by topology in a manner conforming to the SCOP classification along with the ratio, f/f_{random}, calculated as in footnote a.

"lowest score" corresponds to the topological neighbor, B, of A.

2. The CATH classification of B is taken to be the topological classification of A.

The results of this classification are summarized in Table I(a). Classification of the set of proteins was, overall, very consistent with CATH.

The fraction (hit-rate or hit-probability), f, of proteins classified by topology in a manner conforming to the CATH classification has the lowest value for the "neither α nor β" class. Nonetheless, this hit-rate is quite significant in light of the probability of classifying proteins from this class randomly (f/f_{random} = 7.92). This means that the sample space of the "neither" class is small, so that it is easier to find neighbors of "neither" proteins in the α, β, and α-β classes. An example of such a case is shown for one of the proteins in our set, 1hip, in Figure 9. The CATH database places this protein in the class, neither α nor β, while its nearest neighbor, chain B of 1gua, is in the class, α-β, as determined by our topological classification test. The similarities in topology between these proteins are apparent. Both structures contain eight secondary structural elements (1hip has five helix segments and three sheet segments, while 1gua B has three helix segments and five sheet segments). In addition, both structures have a region of helices that come together with a sheet region. The main difference between the proteins is that 1hip has much more loop or turn content where 1gua B has more sheet content. The fact that our test protocol for classification places some proteins into different classes than CATH is not due entirely to the topological nature of the score. Note that only 995 proteins were selected out of 3285 in the CATH list. If the remainder of the list had met our parseability criteria, the probability of finding a topological neighbor in the corresponding CATH class would have increased. One would expect, due to the nature of the classification protocol outlined above, that if the entire set of CATH domains were used in such a test, rather than the set of 3285 sequence representatives, the probability of classifying the domains in a conforming manner would increase. However, a topological classification can still be expected to be slightly different despite this increase in likelihood. It is interesting that our topological method can find a reasonable neighbor for a given protein despite the sparseness of the set of proteins used in the classification.

TABLE II. The Fraction of Proteins Classified into Folds by Topology in a Manner Conforming to the SCOP Fold Classification[†]

SCOP fold	f	f/f_{random}
Toxins' membrane translocation domains (1)	0.714	47.6
Membrane all-alpha (2)	1.00	1.29
Light-harvesting complex subunits (3)	0.650	15.1
Transmembrane beta-barrels (4)	0.970	6.74
Leukocidin (pore-forming toxin) (6)	0.875	50.9

[†] The classification is for proteins under SCOP class f (membrane and cell surface proteins and peptides). Again, the ratio, f/f_{random} is calculated as in Table I. Folds not included in the analysis under class f, were not included because parseability criteria allowed too few files for analysis.

A similar experiment was performed for a set of proteins from the SCOP classes of α, β, α + β, and α/β. We randomly selected 100 proteins from each of these classes in the set of proteins classified by SCOP version 1.61 (http://www.berkeley.edu/parse/html). The results of the test are shown in Table I(b). The conformity of the topological classification to the SCOP classification is not as favorable as for the test of the set from CATH, but all observed hit-rates are still quite significant. The described experiments provide a preliminary test of the ability of our topological representation to classify proteins into general structural categories. A true classification effort would involve a much larger set of proteins. Nonetheless, our results demonstrate the feasibility of such a classification. In order to further demonstrate the capability of our topological comparison to categorize proteins, we probed more deeply into the SCOP hierarchy.

We used our scheme to categorize all the proteins having parseable PDB files from the SCOP class f, "membrane and cell surface proteins and peptides." The result of this categorization shown in Table II demonstrates a remarkable conformity to the SCOP hierarchy. This can be expected because we exhaustively categorized all possible proteins of the class in contrast to the partial set we used for the general categorization result shown in Table I. A 100% hit-rate was obtained for the fold, "membrane all-alpha." This fold is particularly easy to categorize by human inspection of the structure, but our method makes this categorization without visualization or alignment of structures. The remainder of the proteins has a very high hit-rate as well. The lowest fraction of proteins classified by topology in a manner conforming to SCOP belongs to the fold category of light-harvesting complex subunits. However, this fraction is still extremely significant, with f/f_{random} = 15.1.

TABLE III. The Fraction of Proteins Classified Into Families by Topology in a Manner Conforming to the SCOP Family Classification[†]

SCOP family	f	f/f_{random}
Seven-helix membrane receptors (1)	0.938	10.6
Photosynthetic reaction centre, L-, M- and H-chains (2)	1.00	4.24
Cytochrome c oxidase-like (3)	0.991	3.21
F1F0 ATP synthase subunits (4)	0.929	23.9
Aquaporin-like (5)	1.00	45.0
Cytochrome bc1 transmembrane subunits (8)	0.958	7.19
Fumarate reductase respiratory complex transmembrane subunits (9)	1.00	18.0
Calcium ATPase (10)	1.00	90.1
Oligomeric gated channels (11)	0.737	14.0
* Photosystem I (12)	0.00	0.00 ($f_{\text{random}} = 0.0194$)
C1C chloride channel (13)	1.00	36.0
* Formate dehydrogenase N, cytochrome (gamma) subunit (14)	0.00	0.00 ($f_{\text{random}} = 0.00556$)

[†] The classification is for proteins under SCOP class f (membrane and cell surface proteins and peptides), fold 2 (membrane all-alpha), and superfamily 1 (membrane all-alpha). Again, the ratio, f/f_{random} is calculated as in Table I. Families not included in the analysis under class f, were not included because parseability criteria allowed too few files for analysis. Families with $f = 0$ (indicated by an asterisk *) could not be classified in a manner conforming to SCOP because too few proteins exist within the family such that a nearest topological neighbor could be found easily.

Table III shows the results of an additional categorization done at a low level of the SCOP hierarchy for class f (membrane and cell surface proteins and peptides), fold 2 (membrane all-alpha), and superfamily 1 (membrane all-alpha). The analysis of this protein set was also exhaustive. In addition, on average, the hit-rate is yet higher for finding a categorization conforming to SCOP. In many cases it is 100%. A few families with $f = 0$ could not be classified in a manner conforming to SCOP because too few proteins exist within the family such that a nearest topological neighbor could not be found easily. Part of the reason for this is that many of the proteins are, again, "weeded out" because of our PDB file parseability criteria, thereby creating a partially incomplete set of proteins. Nonetheless, the ability of our topological scheme to produce a categorization having such an agreement with the SCOP hierarchy is very well demonstrated by this test.

As a final demonstration of the classificatory power of the topological representation we created phylogenetic trees for the following SCOP superfamilies: (1) **class:** alpha, **fold:** DNA/RNA-binding 3-helical bundle, **superfamily:** "Winged Helix" DNA-binding domain (containing 65 parseable structures), (2) **class:** beta, **fold:** Prealbumin-like, **superfamily:** aromatic compound dioxygenase (containing 103 parseable structures), (3) **class:** alpha and beta, **fold:** Flavodoxin-like, **superfamily:** Flavoproteins (containing 106 parseable structures), and (4) **class:** alpha, **fold:** Globin-like, **superfamily:** Globin-like (containing 762 parseable structures). An all-against-all distance matrix of topological scores (Q) was generated for each superfamily of proteins and used as input for phylogenetic classification as determined by the neighbor-joining algorithm[45] and implemented by the program, neighbor, from the PHYLIP suite of packages for the inference of phylogenies (made freely available at http://evolution.genetics.washington.edu/phylip.html). The resulting phylogenetic tree for superfamily 1 is shown in Figure 10 and the trees for superfamilies 2, 3, and 4 are presented in the supplementary material. Trees 1–3 are annotated on the right according to the SCOP family, protein name, and in some cases, the species. It is easily seen that in almost all cases, different families occupy different branches of the tree. In the cases where the species is included in the annotation, it is easy to see that the species are invariantly clustered together within a family. The topological classification has the uncanny ability to group together proteins with similar PDB identities (e.g., 2irf G, 2irf H, 2irf I, 2irf J, 2irf K, and 2irf L – the interferon regulatory factors in Figure 10) with no a priori knowledge of their kinship, allusion to sequence similarity, or preliminary structural alignment.

A similar test was performed for the SCOP multi-domain class of proteins. Three folds were exhaustively clustered using the same method as for constructing the tree of Figure 10. The multi-domain folds classified were the following: (1) beta-Lactamase/D-ala Carboxypeptidase (177 parseable structures), (2) Heme-linked Catalases (47 parseable structures), and (3) Sugar Phosphatases (120 parseable structures). The resulting phylogenetic trees were also annotated according to the SCOP classification (supplemental material). The consistencies with the SCOP classification are again apparent.

Computational Effort

Given that the topological scoring calculation consists of the simple expression given in Equation (7) involving only simple addition of real numbers, the rate limiting element is due to the Delaunay tessellation of the proteins under comparison. This procedure's complexity is $N\log N$, where N is the number of tessellated points. We calculated the CPU time for an all-against-all comparison of 100 proteins using the topological scheme to find Q for each pair within the set. The average CPU time per topological comparison was observed to be 0.39 s. The same all-against-all comparison was done using the program CE[44] to find the RMSD between structure pairs after their optimal superimposition giving an average of 7.86 s per pair-wise comparison. Thus, our topological comparison was ~20 times faster than the calculation of RMSD for the comparison of

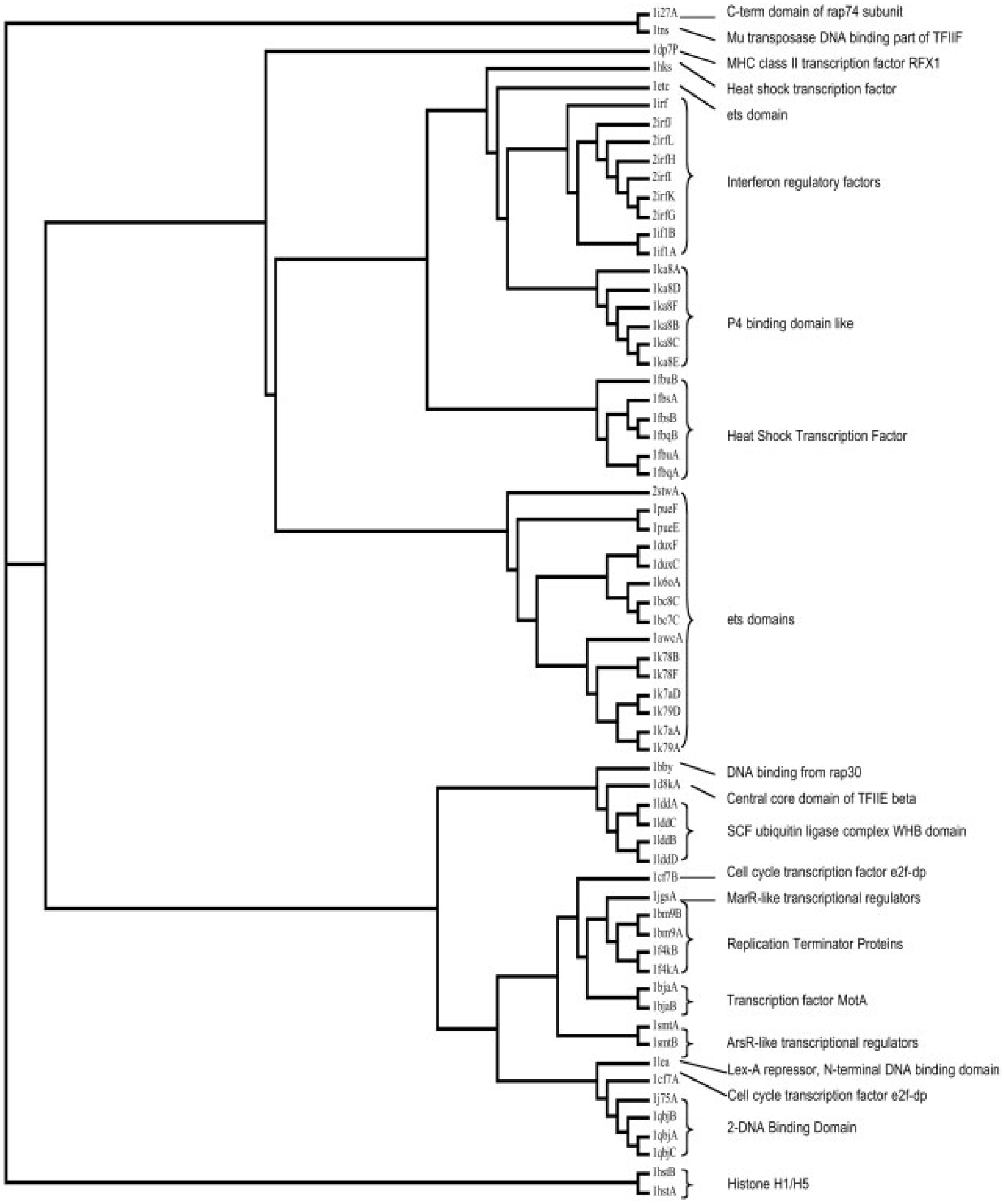
1i27A
1tns
1dp7P
1hks
1etc
1irf
2irfJ
2irfL
2irfH
2irfI
2irfK
2irfG
1if1B
1if1A
1ka8A
1ka8D
1ka8F
1ka8B
1ka8C
1ka8E
1fbuB
1fbsA
1fbsB
1fbqB
1fbuA
1fbqA
2stwA
1pueF
1pueE
1duxF
1duxC
1k6oA
1bc8C
1bc7C
1awcA
1k78B
1k78F
1k7aD
1k79D
1k7aA
1k79A
1bby
1d8kA
1lddA
1ldC
1lddB
1lddD
1cf7B
1jgsA
1bm9B
1bm9A
1f4kB
1f4kA
1bjaA
1bjaB
1smtA
1smtB
1lea
1cf7A
1j75A
1qbjB
1qbjA
1qbjC
1hstB
1hstA
C-term domain of rap74 subunit
Mu transposase DNA binding part of TFIIF
MHC class II transcription factor RFX1
Heat shock transcription factor
ets domain
Interferon regulatory factors
P4 binding domain like
Heat Shock Transcription Factor
ets domains
DNA binding from rap30
Central core domain of TFIIE beta
SCF ubiquitin ligase complex WHB domain
Cell cycle transcription factor e2f-dp
MarR-like transcriptional regulators
Replication Terminator Proteins
Transcription factor MotA
ArsR-like transcriptional regulators
Lex-A repressor, N-terminal DNA binding domain
Cell cycle transcription factor e2f-dp
2-DNA Binding Domain
Histone H1/H5

Fig. 10. Phylogenetic tree as determined via the neighbor-joining algorithm and an all-against-all topological comparison of the proteins in the SCOP hierarchy falling below the level of *class*: a (alpha), *fold*: 4 (DNA/RNA-binding 3-helical bundle), and *superfamily*: 5 ("Winged Helix" DNA-binding domain). The tree is annotated on the right according to SCOP family, protein name, and in some cases, the species.

proteins. Aside from the conceptual benefits one might gain from the use of a topological protein comparison, such a difference in computational efficiency provides for a very attractive method for fast, automated comparison and classification of proteins.

CONCLUSION

This work introduces a representation of protein structure that captures the way in which a protein's sequence writhes into its 3D structure. The representation is, in effect, a vector whose elements describe the distribution of combinations of sequence segment lengths along the protein backbone that give rise to four-body clusters of nearest-neighboring residues within the protein's folded structure. Comparison of two representative vectors is a fast $O(n)$ algorithm that results in a score reflecting the distance between a pair of proteins in a heuristically defined topological space. This renders our topological scoring procedure faster than any automated method utilizing a structural alignment protocol. Currently, there is a need to stringently parse PDB structure files for quality[31] before implementing our comparison because all C_α atoms of a protein are necessary in order to tessellate its structure and, consequently, build its topological representation. This presents a difficulty when attempting to perform exhaustive comparisons within a large database of structures for subsequent classification. However, this problem might easily be overcome with known algorithms for the inference of atomic coordinates when presented with partial protein structures.[46–49]

Our work shows that the topological score between protein pairs correlates with the standard measure of C_α RMSD after optimal superimposition of the structures in a way that can be expected of a topological measure of similarity. The topological score remains more invariant than the RMSD due to the invariant nature of topology with respect to differences in local geometry among the structures. Nonetheless, we have demonstrated that the topological score can lend insight to structural differences on a local geometric level in a manner similar to the RMSD. Furthermore, we have demonstrated that our topological comparison can isolate cases where topology implies structural relatedness while RMSD measurement does not [as in Figure 8(b)]. The method can also discern structural dissimilarity, where RMSD measurement implies similarity [as in Figure 8(a)].

Also, the topological comparison we describe here allows for the identification of structural neighbors within substantially large sets of proteins on the levels of class, fold, superfamily, family, protein, and species that are in agreement with established hierarchical classifications.[16,19] Of course, our method of classification uses different criterion than SCOP or CATH, so there might be differences in the resulting hierarchy. However, in cases where the topologically determined category of a protein does not agree with standard classification, the topological classification has a demonstrated ability to provide an extremely reasonable alternative (as in Figure 9). Finally, our topological comparison method demonstrates a capacity to independently provide a very reasonable hierarchical classification even on the advanced level of superfamilies. The independent classification is also extremely compatible with modern standard hierarchical classifications.[16] With its demonstrated classificatory power, the simple topological representation for proteins we outline in this work is a very interesting prospect for the classification and characterization of protein structures.

ACKNOWLEDGMENTS

This work was supported in part by the Faculty Development Award in Bioinformatics from the Pharmaceutical Research and Manufacturers of America Foundation and by the NSF grant ITR/MCB 0112896 (a collaborative project with Alex Tropsha and Tamar Schlick). We would also like to acknowledge Bala Krishnamoorthy for participation in the development of parsing codes for PDB files, Sagar Pandit for useful discussions pertaining to the formalities of graph theory and topology, and Max Berkowitz' research group for providing the molecular dynamics trajectory of halorhodopsin for topological scoring.

REFERENCES

1. Govindarajan S, Recabarren R, Goldstein R. Estimating the total number of protein folds. J Mol Biol 1999;35:408–414.
2. Wang ZX. A re-estimation of the total number of protein folds and superfamilies. Protein Eng 1998;11:621–626.
3. Koehl P, Levitt M. Sequence variations within protein families are linearly related to structural variations. J Mol Biol 2002;323:551–562.
4. Koonin EV, Wolf YI, Karev GP. The structure of the protein universe and genome evolution. Nature 2002;420:218–223.
5. Holm L, Sander C. Mapping the protein universe. Science 1996;273: 595–603.
6. Holm L, Sander C. Searching protein structure databases has come of age. Proteins 1994;19:165–173.
7. Karlin S, Brendel V, Bucher P. Significant similarity and dissimilarity in homologous proteins. Mol Biol Evol 1992;9:152–167.
8. May AC. Toward more meaningful hierarchical classification of protein three-dimensional structures. Proteins 1999;37:20–29.
9. Remington SJ, Mathews BW. A systematic approach to the comparison of protein structures. J Mol Biol 1980;140:77–199.
10. Kikuchi T. Similarity between average distance maps of structurally homologous proteins. J Protein Chem 1992;11:305–320.
11. Mizuguchi K, Go N. Seeking significance in three-dimensional protein structure comparisons. Curr Opin Struct Biol 1995;5:377–382.
12. Holm L, Sander C. Dali: a network tool for protein structure comparison. Trends Biochem Sci 1995;20:478–480.
13. Young MM, Skillman AG, Kuntz ID. A rapid method for exploring the protein structure universe. Proteins 1999;34:317–332.
14. Falicov A, Cohen FE. A surface of minimum area metric for the structural comparison of proteins. J Mol Biol 1996;258:871–892.
15. Taylor WR. A "periodic table" for protein structures. Nature 2002;416:657–660.
16. Murzin AG, Brenner SE, Hubbard TJP, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
17. Brenner SE, Chothia C, Hubbard TJ, Murzin AG. Understanding protein structure: using scop for fold interpretation. Methods Enzymol 1996;266:635–643.
18. Mizuguchi K, Deane CM, L BT, Overington JP. HOMSTRAD: a database of protein structure alignments for homologous families. Protein Sci 1998;7:2469–2471.
19. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH-a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
20. Holm L, Sander C. The FSSP database of structurally aligned protein fold families. Nucleic Acids Res 1994;22:3600–3609.
21. Karlin S, Zuker M, Brocchieri L. Measuring residue associations in protein structures. possible implications for protein folding. J Mol Biol 1994;239:227–248.
22. Azarya-Sprinzak E, Naor D, Wolfson HJ, Nussinov R. Interchanges of spatially neighbouring residues in structurally conserved environments. Protein Eng 1997;10:1109–1122.
23. Brocchieri L, Karlin S. How are close residues of protein structures distributed in primary sequence? Proc Natl Acad Sci USA 1995;92:12136–12140.
24. Carugo O, Pongor S. Protein fold similarity estimated by a probabilistic approach based on C^α -C^α distance comparison. J Mol Biol 2002;315:887–898.
25. Munkres JR. Topology a first course. New Jersey: Prentice-Hall, Inc.; 1975. 413 p.
26. Singh RK, Tropsha A, Vaisman, II. Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues. J Comput Biol 1996;3:213–221.
27. Bostick D, Vaisman II. A new topological method to measure protein structure similarity. Biochem Biophys Res Commun 2003; 304:320–325.
28. Pandit SA, Amritkar RE. Characterization and control of small-world networks. Phys Rev E 1999;60:1119–1122.
29. Carugo O, Pongor S. A normalized root-mean-square distance for comparing protein three-dimensional structures. Protein Sci 2001; 10:1470–1473.
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. Krishnamoorthy B, Tropsha A. Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations. Bioinformatics 2003;19:1540–1548.
32. Watson DF. Computing the N-dimensional Delaunay tessellation

with application to Voronoi polytopes. The Computer Journal 1981;24:167–172.

33. Berendsen HJC, van der Spoel D, van Drunen R. GROMACS: A message-passing parallel molecular dynamics implementation. Comput Phys Commun 1995;91:43–56.
34. Lindahl E, Hess B, van der Spoel D. Gromacs 3.0: a package for molecular simulation and trajectory analysis. J Mol Model 2001;7: 306–317.
35. Berger O, Edholm O, Jahnig F. Molecular dynamics simulations of a fluid bilayer of dipalmitoylphosphatidylcholine at full hydration, constant pressure, and constant temperature. Biophys J 1997;72: 2002–2013.
36. Nina M, Roux B, Smith JC. Functional interactions in bacteriorhodopsin: a theoretical analysis of retinal hydrogen bonding with water. Biophys J 1995;68:25–39.
37. Roux B et al. Thermodynamic stability of water molecules in the bacteriorhodopsin proton channel: a molecular dynamics free energy perturbation study. Biophys J 1996;71:670–681.
38. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen L. A smooth particle mesh Ewald method. J Chem Phys 1995;103: 8577–8593.
39. Nose S, Klein ML. Constant pressure molecular dynamics for molecular systems. Mol Phys 1983;50:1055–1076.
40. Parrinello M, Rahman A. Polymorphic transitions in single crystals: a new molecular dynamics method. J Appl Phys 1981;52:7182–7190.
41. Tu K, Tobias DJ, Klein ML. Constant pressure and temperature molecular dynamics simulation of a fully hydrated liquid crystal phase dipalmitoylphosphatidylcholine bilayer. Biophys J 1995;69: 2558–2562.
42. Kolbe M, Besir H, Essen L, Oesterhelt D. Structure of the light-driven chloride pump halorhodopsin at 1.8 angstrom resolution. Science 2000;288:1390–1396.
43. Hooft RWW, Sander C, Vriend G. Verification of protein structures: side-chain planarity. J Appl Crystallogr 1996;29:714–716.
44. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. Protein Eng 1998;11:739–747.
45. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987;4:406–425.
46. Hornischer K, Blocker H. Grafting of discontinuous sites: a protein modeling strategy. Protein Eng 1996;9:931–939.
47. Kazmierkiewics R, Liwo A, Sheraga HA. Energy-based reconstruction of a protein backbone from its alpha-carbon trace by a Monte-Carlo method. J Comput Chem 2002;23:715–723.
48. Kazmierkiewicz R, Liwo A, Sheraga HA. Addition of side chains to a known backbone with defined side-chain centroids. Biophys Chem 2003;100:261–280.
49. Liwo A, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA. Calculation of protein backbone geometry from alpha-carbon coordinates based on peptide-group dipole alignment. Protein Sci 1993;2:1697–1714.