



ACADEMIC  
PRESS

Available online at [www.sciencedirect.com](http://www.sciencedirect.com)

SCIENCE @ DIRECT®

Biochemical and Biophysical Research Communications 304 (2003) 320–325

BBRC

[www.elsevier.com/locate/ybbrc](http://www.elsevier.com/locate/ybbrc)

## A new topological method to measure protein structure similarity

David Bostick<sup>a</sup> and Iosif I. Vaisman<sup>b,\*</sup>

<sup>a</sup> *Department of Physics and Program in Molecular/Cell Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA*

<sup>b</sup> *Bioinformatics and Computational Biology, School of Computational Sciences, George Mason University, Manassas, VA 20110, USA*

Received 14 March 2003

### Abstract

A method for the quantitative evaluation of structural similarity between protein pairs is developed that makes use of a Delaunay-based topological mapping. The result of the mapping is a three-dimensional array which is representative of the global structural topology and whose elements can be used to construe an integral scoring scheme. This scoring scheme was tested for its dependence on the protein length difference in a pairwise comparison, its ability to provide a reasonable means for structural similarity comparison within a family of structural neighbors of similar length, and its sensitivity to the differences in protein conformation. It is shown that such a topological evaluation of similarity is capable of providing insight into these points of interest. Protein structure comparison using the method is computationally efficient and the topological scores, although providing different information about protein similarity, correlate well with the distance root-mean-square deviation values calculated by rigid-body structural alignment.

© 2003 Elsevier Science (USA). All rights reserved.

**Keywords:** Protein structure; Protein similarity; Computational geometry; Delaunay tessellation

Since it has long been recognized that the mapping of protein sequence to structure is not, by nature, isomorphic, a major focus in the study of protein comparison has been the characterization of relationships between analogous and remotely homologous proteins [13,14,16,19]. It is thought that the inspection of such relationships will yield insights as to how residue properties affect the overall topology of proteins. It is also very natural that such inspection should reveal important information about what sorts of residues comprise the core structures of evolutionarily related proteins and how such conserved subsets of monomer spatial positions imply a particular protein function [21]. Ultimately, however, it is the topological arrangement of the atoms on which the energetic favorability of a particular conformation is based. Thus, the comparison of three-dimensional protein structures, whether they are obviously related via high sequence homology or similar due to convergent molecular evolution, is also an important step in gaining insight into general protein folding patterns [9].

Current automated means of measuring similarity in large databases of X-ray crystallographic structures have been quite successful in the classification of proteins into structural families by performing pairwise comparisons of protein structures without any allusion to residue identity [14,18,22]. These methods usually involve three-dimensional superimposition of protein pairs as rigid bodies using dynamic programming on a distance matrix generated from the atomic coordinates [12,24].

Although they provide a sound geometric basis for structural similarity using numerical scores such as the distance root-mean-square deviation (dRMSD) or a rough score of probability of alignment such as *z*-score, 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 [22]. This allows for the conclusion 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 dRMSD is attained that sometimes does not reflect the similarities in the protein fold [11,22]. Treatment of these cases involves various heuristic techniques

\* Corresponding author. Fax: 1-703-993-8401.

E-mail address: [ivaisman@gmu.edu](mailto:ivaisman@gmu.edu) (I.I. Vaisman).

in order to determine equivalent residue pairs as a basis for superimposition. Such techniques might employ a quantitative comparison of local geometric features on backbone fragments to combinatorially extend the initial alignment path [25], may individually identify segments of secondary structure for superimposition, or may make use of tools like “geometric hashing” in order to find matching partial structures for the superimposition without making an initial alignment [7]. This attainment of a best superimposition of rigid structures, as a result, consumes more computational time, especially when initial alignments are made.

These problems have been recognized and other means of structural comparison have evolved in order to more completely derive bases of protein relationships. In some cases these means are not fully automated such as those used in constructing the SCOP database [23], which uses biological knowledge of function and visual inspection of structures as well as numerical measures of structural similarity and sequence homology [5]. Such classifications are quite robust and often serve as sources for generating population statistics on the structures of proteins.

Other types of comparisons make use of the well-known fact that the subset of backbone atoms that are near the “core” of a protein structure will contain the greatest number of conserved coordinates among a family of related proteins [8,10,20,21]. Since many biologically significant pairwise similarities are drawn based upon particular elements of secondary structure, some methods of comparison are based solely on such elements, using vector representations of secondary structure fragments [1,2].

In this paper, a method of comparison is suggested in which a scoring of protein pairs is based solely upon the topology of their cores and whose computational expenses are miniscule as compared to other methods that require superimposition of structures. The comparison is performed by classifying four-body nearest neighbor clusters of  $C_\alpha$  atoms according to their implicit topological significance. Information on the separation in primary sequence of the residues that are nearest neighbors in 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 [3,17]. Brocchieri and Karlin [6] 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.) which occur in these categories. 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 method we suggest uses the Delaunay tessellation of a set of points in 3D space that represent the amino acid residues. The tessellation allows for the delineation of a robust definition of nearest neighboring quadruplets of residues arranged at the vertices of irregular tetrahedra (Delaunay simplices). A quantitative topological mapping of these simplices is performed such that a representational three-dimensional (3D) array of integers is defined which is independent of the protein’s length. In the following section we describe these mapping and scoring schemes devised to quantify similarity between two proteins by comparing their resultant representative arrays.

## Materials and methods

*Delaunay tessellation of proteins.* In order to implement our quantitative study of protein topology, Delaunay tessellation was used in order to easily define nearest neighboring quadruplets of  $C_\alpha$  atoms. A geometric tessellation of a 3D set of points is a division of the space into space filling convex polytopes. In the case of the Voronoi tessellation of a set of points (a related geometric construct), the polytopes are polyhedra (termed, “Voronoi polyhedra”) that define the region of space closest to each point. A point in the Voronoi tessellation whose Voronoi polyhedron shares a vertex with three other polyhedra is considered to be a natural nearest neighbor to the three points to which the polyhedra belong. Such points define the vertices of irregular tetrahedra that are, similarly, space filling and are the convex polytopes of the Delaunay tessellation of the set of points. These space-filling tetrahedra are called Delaunay simplices. Delaunay tessellation has been successfully used for various types of protein structure analyses [26–28].

*Description of the mapped array.* Consider a tessellated protein structure with  $N$  residues enumerated with consecutive integers in an order corresponding to its primary sequence. The length of a simplex edge is defined as the number of  $\alpha$  carbons, which compose the segment of the tessellated protein chain between the simplex vertices that define that particular simplex edge. Hence, if the primary sequence is considered, then

$$d_{ij} = j - i - 1, \quad (1)$$

where  $d_{ij}$  is the length of the simplex edge,  $\overline{ij}$ , corresponding to the  $i$ th and  $j$ th  $\alpha$  carbons. Each simplex in the tessellation will then have three lengths associated with the four vertices:  $i, j, k$ , and  $l$ , where  $i, j, k$ , and  $l$  are integers and are all  $C_\alpha$  atoms whose enumeration is of the order of primary sequence.

A transformation,  $T$ , is then applied to each such length in the tessellation which maps the length to an integer value according to the following step function:

$$T : d \mapsto \begin{cases} 1 & \text{if } d = 0, \\ 2 & \text{if } d = 1, \\ 3 & \text{if } d = 2, \\ 4 & \text{if } d = 3, \\ 5 & \text{if } 4 \leq d \leq 6, \\ 6 & \text{if } 7 \leq d \leq 11, \\ 7 & \text{if } 12 \leq d \leq 20, \\ 8 & \text{if } 21 \leq d \leq 49, \\ 9 & \text{if } 50 \leq d \leq 100, \\ 10 & \text{if } d \geq 101. \end{cases} \quad (2)$$

Each simplex is then mapped into a 3D array,  $M$ , where  $M_{npr}$  is the number of simplices whose edges satisfy the following conditions:

- The euclidean length of any one simplex edge is less than 10 Å.
- $d_{ij} = n$ .
- $d_{jk} = p$ .
- $d_{kl} = r$ .

In this way, a heuristic topological mapping of the  $C_\alpha$  atoms of any protein results in a data structure that is independent of the protein size and sequence. Also, since simplices with a euclidean edge length greater than or equal to 10 Å generally correspond to  $C_\alpha$  atoms on the exterior of the protein, their exclusion in the mapping provides for a “core” structure representation of the protein.

The transformation,  $T$ , was designed in such a way as to exploit the implicit topological significance of the array elements,  $M_{npr}$ . Generally, in cases where  $d_{ij}$ ,  $d_{jk}$ , and  $d_{kl}$  are small integers,  $M_{npr}$  will describe the significance of a quantitatively defined “tight” fold, whereas mappings to  $M_{npr}$  which involve large simplex edge lengths correspond to a quantitatively defined “loose” fold. For example, it is observed that  $M_{111}$  in most cases corresponds to a helical secondary structure.

Since each array element represents a contribution to the global topology, and neighboring elements are representative of very similar folds, a simple comparison of two proteins that utilizes such a mapping would involve the evaluation of the differences in single corresponding elements or local regions of the two protein arrays. Thus we use a “simple” scheme which was based upon differences in single corresponding elements of protein arrays  $M$  and  $M'$

$$Q = \sum_{r=1}^{10} \sum_{p=1}^{10} \sum_{n=1}^{10} |M_{npr} - M'_{npr}|, \quad (3)$$

where  $Q$  is the score value. We also tested a scheme that makes use of a “smoothed” comparison of arrays in which eight corresponding array neighbors were averaged and compared

$$Q_s = \frac{1}{8} \sum_{r=1}^9 \sum_{p=1}^9 \sum_{n=1}^9 \left| \sum_{ijk} M_{n+\delta_{ij}, p+\delta_{ij}, r+\delta_{ik}} - \sum_{ijk} M'_{n+\delta_{ij}, p+\delta_{ij}, r+\delta_{ik}} \right|, \quad (4)$$

where  $Q_s$  is a “smoothed” score and  $\delta_{ij}$  is the Kronecker  $\delta$  function. This scheme can also be expected to produce a score representing a more coarse depiction of the binning of the sequence distance between residues,  $d$ , than the transformation in Eq. (2).

Our method for protein comparison was tested on different sets of crystallographically determined protein structures selected from the Protein Data Bank. The testing was done in such a way as to demonstrate aspects of the scoring scheme that should be expected of a measure of topological similarity. The Delaunay tessellation of each set of coordinates was performed using the program, qhull (<http://www.geom.umn.edu/software/qhull>), which employs an algorithm called Quickhull [4]. The calculations and information processing for the mapping procedure of our comparison method were performed using a set of programs written in C and java languages.

## Results

The first of our tests involved protein structures taken from the sequence unique database used by WHATIF [15], a set of proteins with less than 30% sequence identity, less than 0.25  $R$ -factor, and resolution less than 2.5 Å. Fifty-one proteins having a sequence length in the interval of [80,120) were selected from the WHATIF database. An all-to-all comparison was performed within this selection using both simple and smooth scoring schemes in order to investigate the distribution of scores. Also, since the scoring schemes have an

obvious dependence on protein length implied by the relation,

$$N_s = \sum_{r=1}^{10} \sum_{p=1}^{10} \sum_{n=1}^{10} M_{npr}, \quad (5)$$

where  $N_s$  is the total number of simplices in a single tessellation, this set of proteins was divided into two smaller subsets. These subsets consisted of proteins with lengths within the intervals [80,100) and [100,120) and contained 19 and 32 proteins, respectively. The subsets were scored using an all-to-all comparison and the length difference between proteins in each comparison was noted in order to determine whether protein size plays a significant role in a comparison.

The distribution of scores obtained in our all-against-all simple scoring of the WHATIF data set of proteins with length [80,120) is fairly broad. In order to evaluate the contribution of the difference in protein sizes to the score we calculated the average score for each set of protein pairs with the same difference in length for the [80,100) and [100,120) subsets. The score dependence on protein length difference is shown for the two subsets in Fig. 1. This result shows that although there is a consistent shift in the average score values for the larger proteins, the score dependence on the length difference within a subset is negligible.

We also tested a set of proteins from the three families of structural neighbors of 4enl, 1cex, and 1bpi (PDB identification codes) taken from the FSSP database [12]. The FSSP or “Families of Structurally Similar Proteins” is a database containing the results of the alignments of the extended family of a set of representative protein chains from the PDB. Each family consists of all structural neighbors excluding “very close” homologs (with sequence identity greater than 70%). For this set,

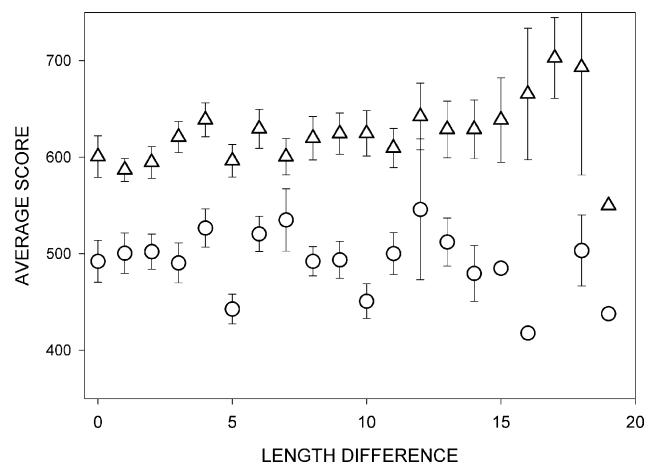


Fig. 1. Dependence of average topological score on the length difference between compared protein for the subsets from the WHATIF database having the ranges of protein length [80,100) (circles) and [100,120) (triangles).

the FSSP reported dRMSD of each of the three proteins in a one-against-all comparison with its neighbors was compared with our corresponding topological score up to a dRMSD of  $\sim 5$  Å. The results shown in Fig. 2 demonstrate an excellent correlation between the topological score and dRMSD values for close as well as for relatively distant (dRMSD 4–5 Å) structural neighbors. A similar correlation was seen for the coarser scoring scheme described by Eq. (4).

Trendlines are plotted in the figure along with the raw data. These trendlines intercept the origin since the structural representative in each family must necessarily have a score of zero when compared with itself. The lines then follow a power law through the representatives' structural neighbors, suggesting that there is a difference in scaling between the dRMSD values and the

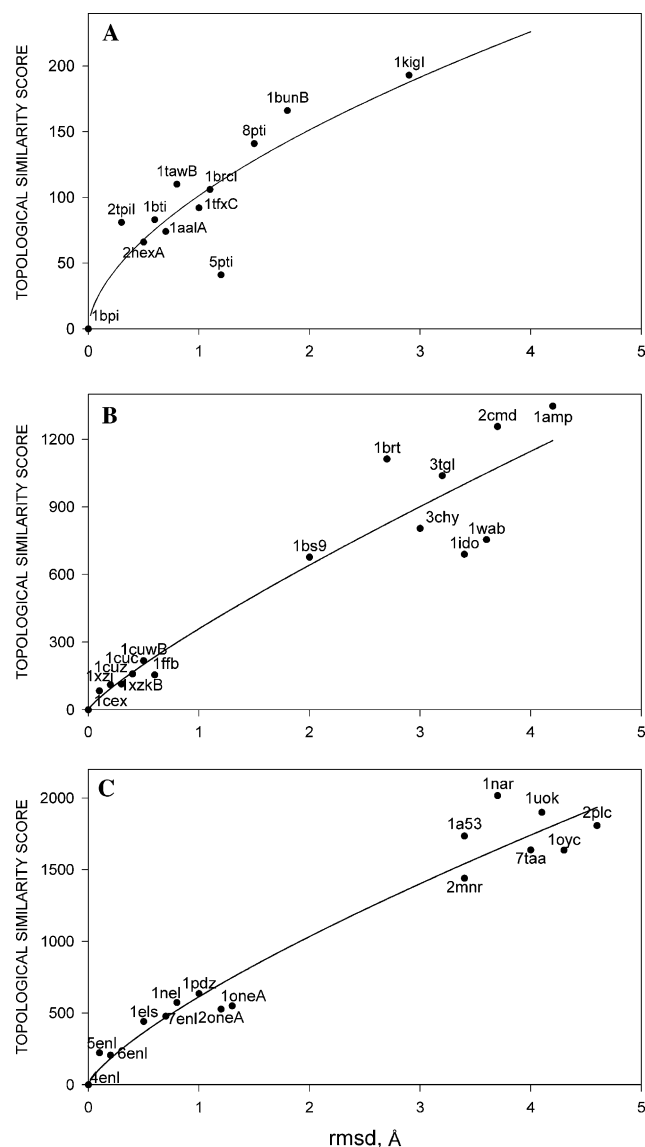


Fig. 2. Topological similarity score correlation with dRMSD for FSSP families: (A) 1bpi, (B) 1cex, and (C) 4enl.

topological score. Intuitively, one might explain this by noting that as neighbors become more dissimilar to their representative, the dRMSD changes more rapidly than our topological score. Since local geometry is matched in a gapped rigid alignment of protein backbones, the dRMSD will become large as local structural differences in the aligned residues become yet more different. On the other hand, the overall topology of similar proteins should remain more invariant as a representative's neighbors' structures become more locally dissimilar. A

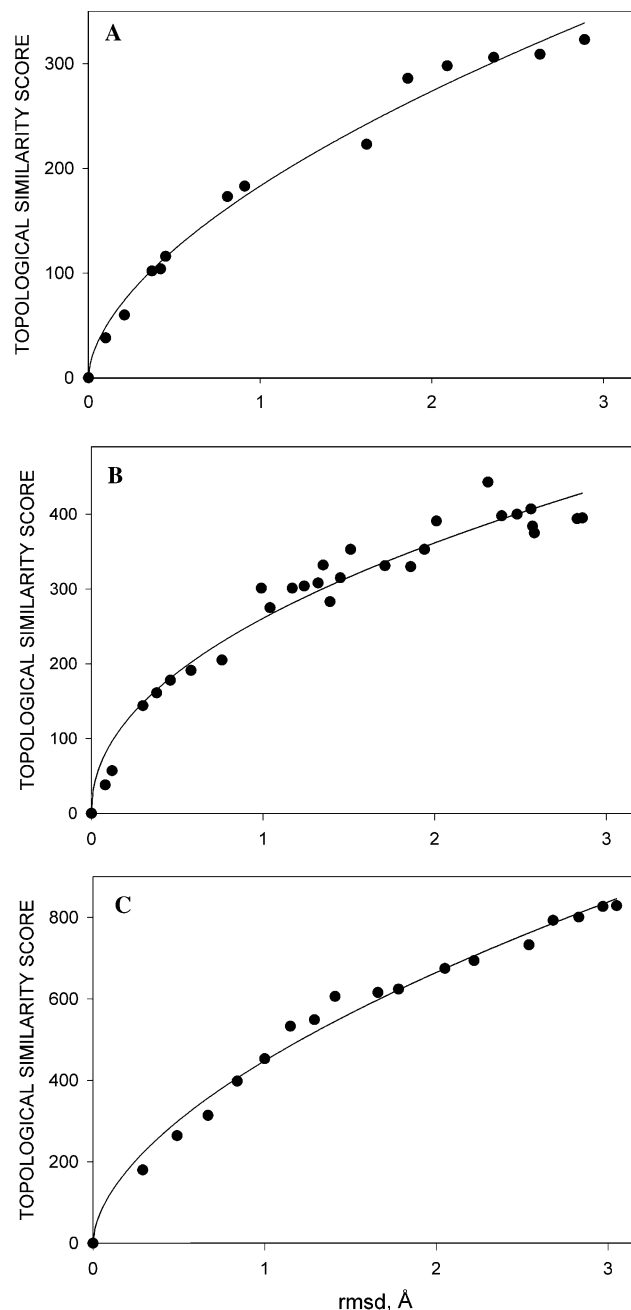


Fig. 3. Topological similarity score correlation with dRMSD for structures from the molecular dynamics trajectory of (A) 2acy, (B) 1nox, and (C) 2por for dRMSD up to  $\sim 3$  Å.

measure of topological similarity should, therefore, correlate with dRMSD in a manner such as that we show in Fig. 2. Significant changes in global topology would be marked by large jumps in the trend of the correlation.

In order to observe how our topological score correlates with the standard dRMSD when comparing different protein conformations, three proteins of different lengths were taken from the WHATIF database: 2acy (98 residues), 1nox (200 residues), 2por (301 residues), and 1a26 (351 residues). The proteins were selected to evenly cover a range of typical protein sizes (from ~100 to ~300 residues long). For each of these proteins, a series of molecular dynamics simulations was performed in vacuum using the dynamics simulation module of the SYBYL 6.4 software package (TRIPOS Associates, St. Louis, MO). The goal of the simulations was to produce several distorted structures for each protein with dRMSD from the original PDB coordinates ranging from 0.0 Å to approximately 3.0 Å as determined by the alignment program, CE [25].

Topological scoring shows a very distinctive sensitivity to topological differences within the three sets of different protein conformations generated from their original coordinates in the PDB. The results for these sets are shown in Fig. 3. The sensitivity is very much the same in character as that seen when observing the differences in members of a protein family with its representative (see Fig. 2). Here again, a clear relationship between the dRMSD determined using the CE algorithm and our topological scoring is apparent. Our molecular dynamics calculations allowed us to probe different levels of dRMSD at a finer level than that we see in our data on protein families. Thus, the trend in the correlation due to the different scaling of the topological score as compared to the dRMSD is more pronounced.

## Discussion

We have suggested in this work a method for the comparison of experimentally determined protein structures based upon topology. The array,  $M$ , as we depict, can be thought of as a representation 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 folded structure. The comparison of one protein's array to another, therefore, amounts to comparing the proteins' distributions of (primary structure) segment length triplets occurring with each nearest neighboring four-body residue cluster.

The shape of our score's correlation with standard measures of dRMSD within a family suggests its topological nature—it remains more invariant than dRMSD measures of similarity due to the invariant nature of the

topology within a family. Thus it may capture elements of structural similarity that structural alignments might miss. Nonetheless, the fact that it correlates so well with standard dRMSD measurements is evidence of its ability to describe similarity between proteins at least as well as these measurements, except in a topological way.

Using pair-wise topological scores as similarity measures we can develop a new, topologically based, hierarchical classification of protein structures. An agglomerative clustering method has been used to build dendrograms for various test sets of protein structures. The results can be compared to those based on conventional (various types of dRMSD) similarity measures. An example of such hierarchical clustering is shown in Fig. 4. Ten proteins were clustered using topological similarity score as a distance measure between all pairs of structures. The results correlate well with conventional protein classification schemes. Proteins that are grouped together according to their topological similarity are also neighbors according to the FSSP hierarchical classification [12]. Thus, such a representation of protein structure has very strong implications for the classification of the fold space of experimentally determined protein structures. Since the data structure of our topological representation of a protein is the same for any protein, the comparison of these data structures is facilitated. Such facility makes the evaluation of similarity computationally very fast when compared to standard structural alignment. This makes the topological method a very interesting prospect for the classification

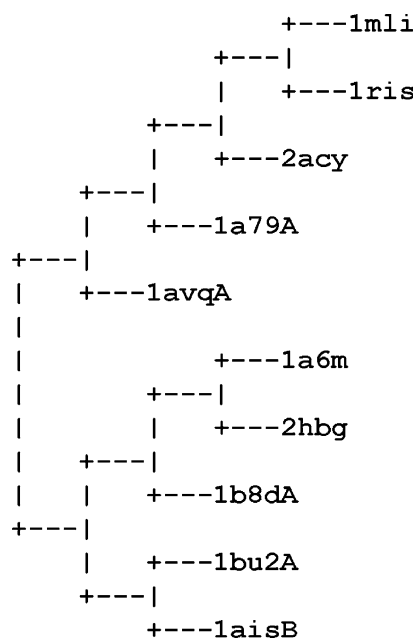


Fig. 4. Hierarchical classification of a set of 10 proteins based on agglomerative clustering utilizing a distance matrix comprised of pairwise similarity scores.

cation of large sets of structures. Indeed, it is the subject of our future study.

It will also be interesting to determine if the intrinsic length dependence of our characterization of proteins depicted in Eq. (5) can be removed. It has already been seen that standard dRMSD measurements have a dependence on the sequence length of the compared proteins. Naturally, a protein's structure is very dependent upon its sequence length because the addition of residues to a protein increases its structural degrees of freedom. Nonetheless, it is of interest to all methods of protein comparison to remove the dependence on sequence length to such an optimal extent as to most effectively quantify structural similarity. Hence, further work will focus on the parameterization of scoring functions and development of more elaborate mapping schemes such as those described in Eq. (2) to improve the precision and efficiency of this methodology.

### Acknowledgments

This work was supported in part by the Faculty Development Award in Bioinformatics from the Pharmaceutical Research and Manufacturers of America Foundation (I.I.V., 1998–1999). We thank Zhibin Lu for his help with some of the programming aspects of this work.

### References

- [1] V. Alesker, R. Nussinov, H.J. Wolfson, Detection of non-topological motifs in protein structures, *Protein Eng.* 9 (1996) 1103–1119.
- [2] N.N. Alexandrov, D. Fischer, Analysis of topological and nontopological structural similarities in the PDB: new examples with old structures, *Proteins* 25 (1996) 354–365.
- [3] E. Azarya-Sprinzak, D. Naor, H.J. Wolfson, R. Nussinov, Interchanges of spatially neighbouring residues in structurally conserved environments, *Protein Eng.* 10 (1997) 1109–1122.
- [4] C.B. Barber, D.P. Dobkin, H. Huhdanpaa, The quickhull algorithm for convex hulls, *ACM Trans. Math. Software* 22 (1996) 469–483.
- [5] S.E. Brenner, C. Chothia, T.J. Hubbard, A.G. Murzin, Understanding protein structure: using scop for fold interpretation, *Methods Enzymol.* 266 (1996) 635–643.
- [6] L. Brocchieri, S. Karlin, How are close residues of protein structures distributed in primary sequence? *Proc. Natl. Acad. Sci. USA* 92 (1995) 12136–12140.
- [7] D. Fischer, O. Bachar, R. Nussinov, H. Wolfson, An efficient automated computer vision based technique for detection of three-dimensional structural motifs in proteins, *J. Biomol. Struct. Dyn.* 9 (1992) 769–789.
- [8] D. Fischer, C.J. Tsai, R. Nussinov, H. Wolfson, A 3D sequence-independent representation of the protein data bank, *Protein Eng.* 8 (1995) 981–997.
- [9] D. Fischer, H. Wolfson, S.L. Lin, R. Nussinov, Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similarities: potential implications to evolution and to protein folding, *Protein Sci.* 3 (1994) 769–778.
- [10] I. Gelfand, A. Kister, C. Kulikowski, O. Stoyanov, Geometric invariant core for the V(L) and V(H) domains of immunoglobulin molecules, *Protein Eng.* 11 (1998) 1015–1025.
- [11] L. Holm, C. Sander, Protein structure comparison by alignment of distance matrices, *J. Mol. Biol.* 233 (1993) 123–138.
- [12] L. Holm, C. Sander, The FSSP database of structurally aligned protein fold families, *Nucleic Acids Res.* 22 (1994) 3600–3609.
- [13] L. Holm, C. Sander, Searching protein structure databases has come of age, *Proteins* 19 (1994) 165–173.
- [14] L. Holm, C. Sander, Mapping the protein universe, *Science* 273 (1996) 595–603.
- [15] R.W.W. Hoof, C. Sander, G. Vriend, Verification of protein structures: side-chain planarity, *J. Appl. Crystallogr.* 29 (1996) 714–716.
- [16] S. Karlin, V. Brendel, P. Bucher, Significant similarity and dissimilarity in homologous proteins, *Mol. Biol. Evol.* 9 (1992) 152–167.
- [17] S. Karlin, M. Zuker, L. Brocchieri, Measuring residue associations in protein structures. Possible implications for protein folding, *J. Mol. Biol.* 239 (1994) 227–248.
- [18] T. Kikuchi, Similarity between average distance maps of structurally homologous proteins, *J. Protein Chem.* 11 (1992) 305–320.
- [19] E.V. Koonin, Y.I. Wolf, G.P. Karev, The structure of the protein universe and genome evolution, *Nature* 420 (2002) 218–223.
- [20] J.V. Lehtonen, K. Denessiouk, A.C. May, M.S. Johnson, Finding local structural similarities among families of unrelated protein structures: a generic non-linear alignment algorithm, *Proteins* 34 (1999) 341–355.
- [21] Y. Matsuo, S.H. Bryant, Identification of homologous core structures, *Proteins* 35 (1999) 70–79.
- [22] K. Mizuguchi, N. Go, Seeking significance in three-dimensional protein structure comparisons, *Curr. Opin. Struct. Biol.* 5 (1995) 377–382.
- [23] A.G. Murzin, S.E. Brenner, T.J.P. Hubbard, C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.* 247 (1995) 536–540.
- [24] C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells, J.M. Thornton, CATH-A hierarchic classification of protein domain structures, *Structure* 5 (1997) 1093–1108.
- [25] I.N. Shindyalov, P.E. Bourne, Protein structure alignment by incremental combinatorial extension (CE) of the optimal path, *Protein Eng.* 11 (1998) 739–747.
- [26] R.K. Singh, A. Tropsha, I.I. Vaisman, Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues, *J. Comput. Biol.* 3 (1996) 213–221.
- [27] H. Wako, T. Yamato, Novel method to detect a motif of local structures in different protein conformations, *Protein Eng.* 11 (1998) 981–990.
- [28] R. Zimmer, M. Wohler, R. Thiele, New scoring schemes for protein fold recognition based on Voronoi contacts, *Bioinformatics* 14 (1998) 295–308.