

Protein Structure Alignment using Geometrical Features*

S. Alireza Aghili[†]
Department of Computer
Science
University of California Santa
Barbara
Santa Barbara, CA 93106
aghili@cs.ucsb.edu

Divyakant Agrawal
Department of Computer
Science
University of California Santa
Barbara
Santa Barbara, CA 93106
agrawal@cs.ucsb.edu

Amr El Abbadi
Department of Computer
Science
University of California Santa
Barbara
Santa Barbara, CA 93106
amr@cs.ucsb.edu

ABSTRACT

A novel approach for similarity search on protein structure databases is proposed which incorporates the three dimensional coordinates of the main atoms of each amino acid and extracts a geometrical signature along with the direction of the given amino acid. As a result, each protein is presented by a series of feature vectors representing local geometry, shape, direction, and secondary structure assignment of its amino acid constituents. Furthermore, a residue-to-residue distance matrix is calculated and is incorporated into a local alignment dynamic programming algorithm to find the similar portions of two given proteins and finally a sequence alignment step is used as the last filtration step. The optimal superimposition of the detected similar regions is used to assess the quality of the results. The proposed algorithm is fast and accurate and hence could be used for the analysis of large protein structure similarity.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous;
I.5 [Pattern Recognition]: Models, Design Methodology;
J.3 [Life and Medical Sciences]: Biology and Genetics

General Terms

Design, Algorithms, Performance

Keywords

Shape Similarity, Protein Structure Alignment, Biological Databases, Biological Data Mining

1. INTRODUCTION

Protein structure similarity has been extensively used to highlight the similarities and differences among *homologous* protein structures. The corresponding applications include *drug discovery*, *phylogenetic analysis*, and *protein classification* which have attracted tremendous attention and have been broadly studied within the past decade. The proteins have a primary *sequence*, which is an ordered sequence

*This research was supported by the NSF grants under EIA00-80134, IIS02-23022, and IIS02-09112.

[†]To whom all the correspondences should be forwarded.

of amino acid molecules, e.g., AKHFDDYAAGDTHKLF. However, they also appear to conform into a three dimensional shape (*fold*) which is highly conserved in the protein evolution. The fold of a protein strongly indicates its functionality and the potential interactions with other protein structures. Meanwhile, the protein sequences as well as their structures may change over time due to mutations during evolution or natural selection. Extensive sequence similarity implies descent from a common ancestral family and the occurrence of many topologically superimposable substructures provides suggestive evidence of evolutionary relationship [5]. This is because the genetic mechanisms are less likely to produce topological permutations. For two given proteins, if the sequences are similar then the evolutionary relationship is apparent. However the problem becomes much more challenging when the sequences are very different from each other while the given proteins appear to demonstrate substantial functional similarities. The three dimensional structures of proteins due to structural, conformational and functional restraints placed on them, are much more resilient to mutations than the protein sequences. As a result, the inspection of structural similarity among various protein fragments may help to better understand the differences in the observed functionalities and potentially their evolutionary relationships. There are two main problems in the study of protein structure similarity:

- *Complexity*. The problem of structure comparison is NP-hard and there is no exact solution to the protein structure alignment [6]. A handful of heuristics [2, 3, 4, 5, 8, 9, 10] have been proposed in which, for the best result, the similarity might need to be evaluated using a series of techniques in conjunction. However, none of the proposed methods can guarantee optimality within any given precision. There are always cases where one heuristic fails to detect, while some of the others succeed!
- *Curse of Dimensionality*. The total number of discovered protein structures has been growing exponentially. Currently¹ the Protein Data Bank (PDB)[1] contains 27,112 protein structures. The growth in the content of PDB demands faster and more accurate tools for structure similarity and the classification of the known structures.

Copyright is held by the author/owner.

CIKM'04, November 8–13, 2004, Washington, DC, USA.
ACM 1-58113-874-1/04/0011.

¹as of September 7th, 2004.

In this paper, we consider both the geometrical features and the corresponding amino acid sequences of the protein chains for more efficient similarity comparison. The main goal of protein structure similarity is to superimpose two proteins over the maximum number of *residues* (amino acids) with a minimal distance among their corresponding matched atoms. Distances between the atom coordinates, residual feature vectors or molecular properties are often used to compare protein structures. The molecular properties (e.g., *physical properties, local conformations, distance from gravity center, position in space, global/local direction in space, side chain orientation, and secondary structure type*) or their corresponding relationships are considered either separately or in combination with each other as a basis for structural comparison.

2. THE PROPOSED METHOD

We have proposed a novel method for fast and accurate protein structure similarity using geometrical signatures. The algorithm not only exploits the topological properties of the amino acid and protein structures, but also incorporates the biological properties of the amino acids into account. The algorithm starts by identifying the geometrical properties of each amino acid of the given proteins along with their directions and the corresponding biological features. As a result, each protein structure is represented by a series of directional shape signature feature vectors, one for each amino acid. In the next step, a score matrix is constructed on the corresponding feature vectors. A local *structural alignment* [11] based on shape, direction and biological features, detects the optimal local matching regions among the two proteins. For each of the locally matched regions (pertaining to length and score constraints), a *sequence alignment* is performed to facilitate a visualization of the sequence similarities. Thereafter, the best locally matched regions are topologically superimposed. The corresponding *RMSD* (*Root Mean Square Distance*) value, length of the aligned fragments, and sequence alignment score are reported for the assessment of the quality of the match. A *linear time* least-square solution to superimpose the ordered sets of protein feature vectors is applied. We sort the results based on their extent (L) and RMSD value and report a list of top alignments with the best scores φ , where $\varphi = L/RMSD$.

3. EXPERIMENTAL EVALUATIONS

We implemented our proposed technique using *Java 1.4.1* and ran experiments on an *Intel Xeon 2.4 GHz* with *1GB* of main memory. Our experiments incorporated a representative of the PDB database using the PDBSELECT² method [7] which does not contain any homologue protein pairs. The PDBSELECT database is an archive of 2216 *non-homologue* protein chains with a total number of 352855 residues (as of December 2003). Each of the protein pairs from the PDBSELECT protein database has less than 25% sequence identity (non-homologue). Protein pairs with low sequence similarity may not be efficiently compared solely based on a sequence-level similarity procedure and therefore introduce a challenging problem where the combination of structure and sequence alignment may be very helpful. We incorporated a combination of structural and sequence alignment

²For more information please refer to <http://homepages.fh-giessen.de/hg12640/pdbselect/>

for efficient protein similarity comparison. In the experiments, we discovered motifs not reported by other alignment tools such as CE [10], DALI [8], and CTSS [3]. One such motif discovered by our technique was between 1AKT: (made of 147 residues and 1108 atoms) and 1CRP: (made of 166 residues and 2619 atoms) protein chains (having 8.9% sequence identity) with RMSD 0.58 Å. The results were reported after finding the best similar fragments, followed by the optimal superimposition of the structures of the corresponding matched fragments. The fragment score (φ) was incorporated to denote the *quality* of the matched fragments and the best aligned fragment is the one with the highest fragment score. One interesting observation was that the matched fragment pairs using our technique with high φ were those having a higher level of similarity to their corresponding aligned fragments in DALI. We incorporated DALI to validate the quality of our results, while DALI is designed with very insightful domain expertise and is expected to return biologically meaningful results. Our results were either similar to that of DALI, or consisted of the combination of some consecutive fragment pair outputs of DALI. In general, the experimental evaluations depicted highly *accurate* and *consistent* results compared with DALI, CE, and CTSS protein structure similarity methods, while running only in fractions of a second.

4. REFERENCES

- [1] Protein data bank(pdb). <http://www.rcsb.org/pdb/holdings.html>, 2004.
- [2] P. Bradley, P. Kim, and B. Berger. Trilogy: Discovery of sequence-structure patterns across diverse proteins. *Proc. Natl. Academy of Science*, 99(13):8500–5, 2002.
- [3] T. Can and Y. Wang. Ctss: A robust and efficient method for protein structure alignment based on local geometrical and biological features. In *IEEE Computer Society Bioinformatics Conf.*, pages 169–179, 2003.
- [4] O. Çamoğlu, T. Kahveci, and A. Singh. Towards index-based similarity search for protein structure databases. In *IEEE Computer Society Bioinformatics Conf.*, pages 148–158, 2003.
- [5] J. Gibrat, T. Madej, and S. Bryant. Surprising similarities in structure comparison. *Current Opinion Structure Biology*, 6(3):377–85, 1996.
- [6] A. Godzik. The structural alignment between two proteins: is there a unique answer? *Protein Sci.*, 5:1325–1338, 1996.
- [7] U. Hobohm, M. Scharf, and R. Schneider. Selection of representative protein data sets. *Protein Science*, 1:409–417, 1993.
- [8] L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. *J. Molecular Biology*, 233(1):123–138, 1993.
- [9] G. Lua. Top: a new method for protein structure comparisons and similarity searches. *J. Applied Crystallography*, 33(1):176–183, 2000.
- [10] I. Shindyalov and P. Bourne. Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. *Protein Engineering*, 11(9):739–747, 1998.
- [11] R. Smith and M. Waterman. Identification of common molecular subsequences. *J. Mol. Bio.*, 147(1):195–197, 1981.