PADS: Protein Structure Alignment Using Directional Shape Signatures*

S. Alireza Aghili, Divyakant Agrawal, and Amr El Abbadi

Department of Computer Science,
University of California-Santa Barbara,
Santa Barbara, CA 93106
{aghili, agrawal, amr}@cs.ucsb.edu

Abstract. A novel data mining approach for similarity search and knowledge discovery in protein structure databases is proposed. PADS (**P**rotein structure **A**lignment by **D**irectional shape **S**ignatures) incorporates the three dimensional coordinates of the main atoms of each amino acid and extracts a geometrical shape signature along with the direction of each amino acid. As a result, each protein structure is presented by a series of multidimensional feature vectors representing local geometry, shape, direction, and biological properties of its amino acid molecules. Furthermore, a distance matrix is calculated and is incorporated into a local alignment dynamic programming algorithm to find the similar portions of two given protein structures followed by a sequence alignment step for more efficient filtration. 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 analysis and knowledge discovery in large protein structures. The method has been compared with the results from CE, DALI, and CTSS using a representative sample of PDB structures. Several new structures not detected by other methods are detected.

Keywords: Shape Similarity, Protein Structure Comparison, Biological Data Mining, Bioinformatics.

1 Introduction

Protein structure similarity has been extensively used to highlight the similarities and differences among *homologous* three dimensional 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 of amino acid molecules, e.g. AALHSIAISAJSH. However, they also appear to conform into a three dimensional shape

* This research was supported by the NSF grants under CNF-04-23336, IIS02-23022, IIS02-09112, and EIA00-80134.

L. Zhou, B.C. Ooi, and X. Meng (Eds.): DASFAA 2005, LNCS 3453, pp. 17–29, 2005.

(*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. High sequence similarity implies descent from a common ancestral family, and the occurrence of many topologically superimposable substructures provides suggestive evidence of evolutionary relationship [8]. This is because the genetic mechanisms rarely produce topological permutations. For two given proteins, if the sequences are similar then the evolutionary relationship is apparent. However the three dimensional structure of proteins, due to their conformational and functional restraints, are much more resilient to mutations than the protein sequences. There exist functionally similar proteins which *sequence-level* similarity search fails to accurately depict the true similarity. Such cases introduce a big challenge and the necessity of incorporating *structure-level* similarity. Meanwhile, there are two main problems in protein structure similarity:

- *Complexity.* The problem of structure comparison is NP-hard and there is no exact solution to the protein structure alignment [9]. A handful of heuristics [4, 5, 6, 8, 12, 13, 14, 15, 18] have been proposed in which, to achieve 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. Currently the Protein Data Bank (PDB)[1] contains 27,112 protein structures (as of September 8^{th}, 2004.). The growth in the content of PDB demands faster and more accurate tools for structure similarity and the classification of the known structures.

We first provide the basic definitions of terms used throughout the paper in Table 1. In this paper, we consider both the sequence and structure of 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. These methods typically employ the three dimensional coordinates of the C_α atoms of the protein backbone and sometimes, in addition, the side chain comprising C_β atoms but exclude the other amino acid atoms when making global structural comparisons. When superimposing two protein structures, side chain conformations (coordinates of O, C, C_β, N, H atoms) may vary widely between the matched residues however the C_α atoms of the backbone trace and the corresponding SSEs are usually well conserved. However, there are situations where the local comparison of the side chain atoms can be of great significance, for instance, in the comparison of residues lining an active or binding sites especially when different *ligands*[1] are bound to the same or similar structures [10].

[1] *ligand*: An atom, molecule, or ion that forms a complex around a central atom.