Pattern Recognition Letters 34 (2013) 559–563

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Protein motifs retrieval by SS terns occurrences

V. Cantoni [a,*], A. Ferone [b], O. Ozbudak [c], A. Petrosino [b]

[a] University of Pavia, Department of Electrical and Computer Engineering, Via A. Ferrata, 1, 27100 Pavia, Italy
[b] University of Naples Parthenope, Department of Applied Science, Centro Direzionale Isola C4, 80133 Napoli, Italy
[c] Istanbul Technical University, Department of Electronics and Communication Engineering, 34469 Istanbul, Turkey

ARTICLE INFO

Article history:
Received 20 July 2012
Available online 10 December 2012

Communicated by S. Sarkar

Keywords:
Protein motif retrieval
Secondary Structures
Similarity search
Generalized Hough Transform
Pattern recognition
Structural biology

ABSTRACT

This paper describes a new approach to the analysis of protein 3D structure based on the Secondary Structure (SS) representation. The focus is here on structural motif retrieval. The strategy is derived from the Generalized Hough Transform (GHT), but considering as structural primitive element, the triplet of SSs. The triplet identity is evaluated on the triangle having the vertices on the SS midpoints, and is represented by the three midpoints distances. The motif is characterized by the complete set of triplets, so the Reference Table (RT) has a tuple for each triplet. Tuples contain, beside the discriminant component (the three edge lengths), the mapping rule, i.e. the Reference Point (RP) location referred to the triplet. In the macromolecule to be analyzed, each possible triplet is searched in the RT and every match gives a contribution to a candidate location of the RP. Presence and location of the searched motif are certified by the collection of a number of contribution equal (obviously in absence of noise and ambiguities) to the RT cardinality (i.e. the number of motif triplets). The approach is tested on twenty proteins selected randomly from the PDB, but having a different number of SSs ranging from 14 to 46. The retrieval of all possible structural blocks composed by three, four and five SSs (very compact and completely distributed) have been conducted. The results show valuable performances for precision and computation time.

1. Introduction

Many evolutionarily and functionally meaningful links between proteins come to light through the analysis of their spatial 3D structures. Protein structure and morphology are significant to understand and predict their functionality (Shuoyong et al., 2007). Protein structure comparison is an important issue that helps biologists to understand various aspects of protein function and evolution. For this reason protein comparison and retrieval are basic issues that helps biologists to comprehend various aspects of the phylogenetic evaluation and of the tasks performed i.e. proteins role in the machinery of life.

The protein 3D structure is vitally important in many biological applications, such as rational drug design. The retrieval of a protein 3D structure can be achieved by different experimental and bioinformatics methods. To this aim, X-ray crystallography is a powerful tool although time-consuming, expensive, and not feasible for all proteins (e.g. so far very few membrane protein structures have been determined). Nuclear magnetic resonance (NMR) is another tool that can be employed to determine the 3D structures of membrane proteins, even though time-consuming and costly. In order to acquire the structural information in a timely manner, it is possible to adopt various bioinformatics tools (see, e.g. (Li et al., 2011; Ma et al., 2012; Wang and Chou, 2011; Chou et al., 1997; Wang and Chou, 2012) and a review Chou, 2005). The present study is devoted to develop a novel method to search a database of protein structures for 3D patterns of secondary structural elements.

Structural comparison and protein structure retrieval problems have been studied in the structural biology community. In most cases just representing the set of the protein by a set of SS elements. Can and Wang (2003) present a new method for conducting protein structure similarity searches and applies differential geometry knowledge on their 3D structure for extracting "signatures" such as curvature, torsion and SS type. Camoglu et al. (2003), to find similarities in protein database, build an indexing structure based on SS elements triplets by using R-tree. Chionh et al. (2003) propose the SCALE algorithm to compare protein 3D structures through matrices that utilizes angles and distances between SS elements. Krissinel and Henrick (2004) describe the Secondary Structure Matching (SSM) algorithm for comparison in 3D, including an original procedure for matching graphs built on the protein's SS elements, that is followed by an iterative 3D alignment of protein backbone C_α atoms. Chi et al. (2004) design a fast system for protein structural block retrieval by using image based distance matrices and multidimensional indices. The 1D string

* Corresponding author. Tel.: +39 0382 985358; fax: +39 0382 985373.
E-mail addresses: virginio.cantoni@unipv.it (V. Cantoni), alessio.ferone@uniparthenope.it (A. Ferone), ozbudak@itu.edu.tr (O. Ozbudak), alfredo.petrosino@uniparthenope.it (A. Petrosino).

0167-8655/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2012.12.003

representation of local protein structure retains a degree of structural information. This type of representation can be a powerful tool for comparison and classification. Friedberg et al. (2006) described the use of a particular structure fragment library, denoted as KL-strings, for the 1D representation of protein structure and developed an infrastructure for comparing structures with 1D representation. Shuoyong et al. (2007) developed a program, ProSMoS (Protein Structure Motif Search) to find fold-level structural similarities and to search for the presence of structural motifs. This package searches a library of protein structures for user defined 3D patterns of SS elements. Also a web server to make a pattern-based search, using interaction matrix representation of protein structures (Shuoyong et al. (2009)), has been developed. Albrecht et al. (2008) propose a different approach and apply data reduction techniques directly to the protein structure and convert 3D data into 2D so accelerating the structural comparisons. Zotenko et al. (2007) propose an approach to speed up protein comparison by mapping a protein structure to a high-dimensional vector and approximating structural similarity by suitable distances between the corresponding vectors. Zhang et al. (2009) by a transition probability matrix and some structural characteristic vectors of proteins developed FDOD (Function of Degree of Disagreement) a score scheme to measure the protein similarity. Nguyen and Madhusudhan (2011) propose a new algorithm, CLICK, to capture such similarities. This method optimally superimposes a pair of protein structures independently of their topology and can generally be applied to compare any pair of molecular structures represented in Cartesian coordinates as exemplified by the RNA structure superimposition benchmark. Cantoni and Mattia (2012) and Cantoni et al. (2012) made a study for retrieving structural motifs by using GHT and range tree. This approach is completely new, because the analysis is based on the 3D spatial distribution of the SS.

In this paper, a new approach for structural block retrieval based on protein SS comparison is proposed. Here, triangles joining the middle points of the SS triplets are considered as "structural elements" and all the block triangles are compared with all the macromolecule triangles. The focus of the paper is on the retrieval of an existing structural block completely and precisely known. The block can be defined without constraints such as adjacency, distance limits, homogeneity, etc. The only constraints is that the SS components exist in the protein macromolecule.

The rest of the paper is organized as following. Section II introduces the GHT and the triangle approaches. Section III represents the experiments and their results. In the final session IV a brief discussion and the future works are described.

2. Methodology

In this paper a novel approach, GHT-based, for motif retrieval is proposed. The GHT is used for comparison and search of structural similarity between a given structural block (a motif or a domain or the entire protein) and the proteins of a database like the PDB. Note that, if the searched structure is just a component of a protein (like a structural motif or a domain) the same algorithm supports the detection and the statistical distribution of these components.The primitive patterns to which is applied the cumulative voting procedure are triplets of SSs, that is the structural elements are the triangles having the vertices in the middle point of the SS triplets.

2.1. The triangular structural elements

In this algorithm we use SS triplets for motif retrieval in protein macromolecule. In three-dimension, middle points of three SSs are joined and an imaginary triangle is composed. So, through the SS triplets a local reference system is set up, e.g. having the origin in the triangle barycenter, the y-axis passing through the farthest vertex, the x-axis on the triangle plane, and the z-axis following the triangle plane normal (see Fig. 1).

The coordinates of the RP are determined with respect to this local reference system. A structural block, that in the sequel we name motif, is defined by a few SSs, and for each motif a RP is fixed in the center of gravity of the midpoints of these SSs. Being n the number of motif SSs, the number t of triplets/triangles is given by:

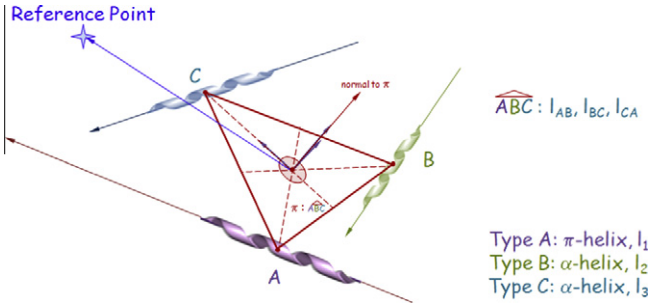

Fig. 1. Local reference system representation for the A, B, C triplet. The comparison parameters are the length of triangle edges (i.e. the mid-points distances). Other discriminant parameters can be considered such as: type of SS (π-helices, α-helices, β-strands, etc.), SS lengths (i.e. number of amino acids), types of amino, etc.

Table 1
Algorithm for the retrieval of all possible r motifs contained in a set of M proteins.

Input: Protein DSSP files; N_i: number of protein SSs; m: number of motif SSs
Output: Locations of candidate motifs in the accumulator A_{RP}, representing the parameter space.

```
for i = 1 to M do
    Calculate all m combinations of N_i : r = C(N_i, m)
    for j = 1 to r do
        Find the motif barycenter RP
        Calculate the number of motif triangles: c = C(m, 3)
        Calculate the number of protein triangles: p = C(N_i, 3)
        for k = 1 to c do
            Compute the edge lengths of motif triangle: d1_k, d2_k, d3_k //RT constituents
        for l = 1 to p do
            Compute the lengths of protein triangle: d1_l, d2_l, d3_l
            for k = 1 to c do
                if match(d1_k, d2_k, d3_k and d1_l, d2_l, d3_l) then A_RPl = A_RPl + 1
        Compute the peaks in HS
        Assign the position with the expected votes as candidate RP
```

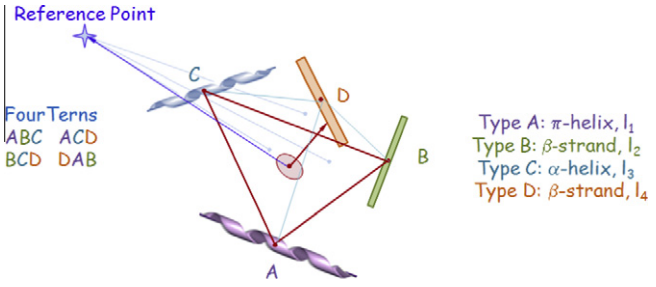

Fig. 2. A heterogeneous motif composed of two helices A and C, and two strands B and D. In this case $t = 4$, the corresponding triangles are shown on the left of the figure. In detail, it is represented the center of gravity of triangle ABC and it is shown the correspondence displacement, i.e. the RP position. If the motif is completely contained in the macromolecule the corresponding RP location receive one contribution for each of the four triangles, as shown (the other three contributions are just sketched).