3D PROTEIN CLASSIFICATION USING TOPOLOGICAL, GEOMETRICAL AND BIOLOGICAL INFORMATION

V. Tsatsaias[1], P. Daras[2] Member, IEEE and M.G. Strintzis[1,2] Fellow, IEEE

[1] Information Processing Lab
Department of Electrical & Computer Engineering
Aristotle University of Thessaloniki
Thessaloniki 54006 Greece

[2] Informatics and Telematics Institute
1^{st} km Thermi-Panorama Rd, Thessaloniki
570 01 P.O. Box 60361, Greece
e-mail: {tsatsaia, daras}@iti.gr
strintzi@eng.auth.gr

ABSTRACT

Computational approaches for protein classification have been proposed over the last years in order to speed up the analysis of the biological mechanics in living organisms. Most of the approaches tend to focus in geometrical comparison of the 3D molecules to reach their goals. In this paper a method suitable for partial (sub)graph matching of 3D proteinic models, in order to achieve fast and accurate classification, is proposed. The 3D objects are firstly segmented to their molecular structure. Then, descriptors are extracted for each segment using spherical harmonics algorithms, and graphs are constructed for the molecules. Next, a sub-graph matching procedure is utilized and the results are refined using biochemical properties to get biological meaningful classification. The experimental results proved that the proposed method achieves accurate classification of the proteinic data.

Index Terms— proteins, 3D representation, graph matching, classification

1. INTRODUCTION

Understanding the molecular engineering of life demands the decoding of the functions of proteins in a living organism. In order to achieve the latter, one has to classify proteins based on their function and their ability to interact with other molecules. Since the experimental techniques on this matter are particularly demanding in time and financial support, many computational technics have been proposed to solve this problem with more time-cost efficiency.

Recently, several researchers have investigated approaches for protein classification based on structural features of 3D protein models. The general idea behind the presented methods is that proteins with similar structures tend to take part in the same biological functions of an organism, thus by exploiting this, it is possible to classify proteins in common-function classes. More specifically, in [1], graphs which are built on the protein's secondary structure elements, are matched. This procedure is followed by an iterative 3D alignment of protein backbone C_α atoms leading to the tracing of the common structural features of two proteins. In [2], the proteinic molecule is represented as a graph, formed by a set of stereochemical groups, which contain both chemical and structural data. Then, a graph matching algorithm is used to find pairs of matching graphs. In [3], a theoretical framework based on bipartite graph matching is presented, to identify the best alignment between two proteins in 3D space in order to determine protein functionality. The graphs for this procedure consist of the backbone C_α atoms of each proteinic chain and a best correspondence of the two parts is found using a known algorithm [4].

Further, in [5], the G-Protein superfamily is being analyzed regarding the abilitiy of its members to interact with other molecules (ligands) by simulating the interacting process between a 3D proteinic model and a potential inhibitor using the internal coordinate method (ICM), that the same researchers have introduced. Another approach is proposed in [6], where the proteinic 3D model is compared to known active sites using a hash matching algorithm. By doing so, it is possible to test a protein for a great amount of probable active sites that correspond to an equal number of different functions. Finally, in [7], the Spherical Trace Transform is applied to proper positioned, in terms of translation and scaling, 3D structures in order to produce geometry-based descriptor vectors, which are rotation invariant and describe the 3D shape of the molecule. These vectors are enriched with attributes of the primary and secondary structure elements of the molecule and used to compare and classify the data.

The majority of the aforementioned methods aim to find structural similarities between two (or more) 3D protein molecules using various (different) descriptors for their features and algorithms to match these data. Although these approaches have shown promising results the demand for a computational method that classifies fast and accurate proteinic data is not completely satisfactory due to the fact that they do not use all

This work was supported by the VICTORY EC IST project.

the available information (topological, geometrical and biological) in order to describe and compare proteinic molecules.

In this paper, a partial graph matching method is proposed so as to achieve protein classification taking into account topological, geometrical and biochemical information contained in the pdb [8] files. Such a combination is extremely innovative since it is presented for the first time in the literature.

The rest of this paper is organized as follows: In section 2 the proposed method is described in detail, while in section 3 the experimental results are given. Finally, conclusions are drawn in section 4.

2. THE PROPOSED APPROACH

The proposed algorithm is as follows: Firstly, the pdb files are used so as to create the 3D protein representations. Secondly, each 3D protein is being segmented based on the type of secondary structures of the molecule. Further, for each segment the algorithm introduced in [9] is used, so as to extract geometrical descriptors invariant to geometric transformations. The segmentation process leads to a graph construction which is further enriched with the extracted geometrical descriptors and other topological information such as the angles between two neighboring edges. Then, a subgraph matching process [10] is utilized, in order to identify parts with topological and geometrical similarities. Finally, by taking into account biochemical data regarding the similarity between amino acids, such as the PAM250 Scoring Matrix [11], the results are being refined and the final classification is accomplished.

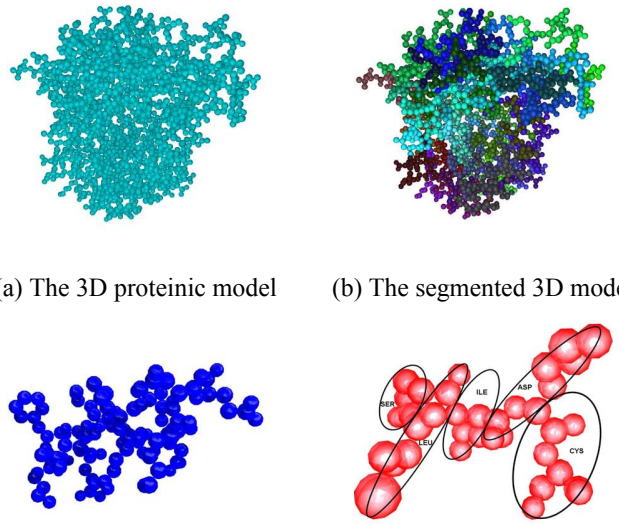
2.1. 3D Segmentation

Having as input a pdb file, only the amino acid sequence, the coordinates and the radii of the atoms C,O,N,S are taken into account so as to construct the 3D representation of the corresponding molecule (Fig.1(a)). The detailed procedure which leads to the 3D proteinic model is described in [7]. In order to segment the protein to a set of segments $\mathbf{S} = \{S_t,\ t = 1,\ldots,N\}$ the following assumptions are made:

- residues that belong to the same type of secondary structure form a segment S_t.
- w neighboring residues in the amino acid sequence, that do not belong to any type of secondary structure form a segment S_t.

In Fig.1(b) the segmented 3D proteinic molecule is depicted where the different segments are marked with different colors. In Fig.1(c), a segment formed from a helix consisting of 15 residues is shown, while in Fig.1(d), a segment of w = 5 residues, which do not belong to any type of secondary structure, is depicted.

Finally, each segment S_t is represented as a 3D function, expressed in spherical coordinates as $f_t(\theta,\phi)$, $t = 1,\ldots,N$ where N is the total number of segments.



(a) The 3D proteinic model (b) The segmented 3D model

(c) A helix segment (d) 5-residue segment

Fig. 1. The segmentation of protein 2dabA.

2.2. Segment Descriptor Extraction

When the segmentation process is accomplished, geometric descriptors are extracted for each segment S_t. To achieve the latter the method presented in [9] is followed, which is based on spherical harmonics approach. According to [9], the function $f_t(\theta,\phi)$ is decomposed in the sum of its harmonics:

$$f_t(\theta,\phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{m=l} a_{lm}^t Y_l^m(\theta,\phi) \qquad (1)$$

where $Y_l^m(\theta,\phi)$ are the spherical harmonics and a_{lm}^t are the harmonics coefficients grouped in vectors $\mathbf{a}_l^t = [a_{l,-l}^t,\ldots,a_{l,l}^t]$. Finally, the 3D segment is described by the Euclidean norms of vectors $\mathbf{a}_l^t$ forming the descriptor vector of segment t:

$$\mathbf{D}_t = \left[\ ||\mathbf{a}_0^t||\ ||\mathbf{a}_1^t||\ \ldots||\mathbf{a}_L^t||\ \right]^T \qquad (2)$$

where L is the total number of harmonics.

2.3. From Segments to Graphs

A graph can be mathematically represented as $G = \{V, E\ ,\{\mathbf{A}_i\}_{i=1}^r\ ,\ \{\mathbf{B}_i\}_{i=1}^s\}$ [10] where $V = \{(v)_p, p = 1,\ldots,n\}$ is the non-empty set of n vertices, E is the set of edges, $\mathbf{A}_i$ is the adjacency matrix that bears the $i-th$ edge attribute,

$\mathbf{B}_i$ is the value of the $i-th$ vertex attribute, r is the number of edge attributes and s is the total number of vertex attributes. In order to apply the graph theory, a proteinic graph consisted of $n = N$ vertices is formed. Every segment S_t is associated with each vertex $\mathbf{v}_t$. The edges are formed from the segments' connectivity so as to represent the proteinic sequence (Fig. 2). The edges are set to be undirected and not attributed. Thus, the graph that describes each protein p, is $G_p = \{V, E, \mathbf{A}, \{\mathbf{B_i}\}_{i=1}^{s}\}$, where $\mathbf{A}$ is a binary orthogonal adjacency matrix due to the hypothesis that if the edges are not attributed, the following assumption is made: all edges are single attributed with the same value (for simplicity reasons this value is set to 1).

The complete attribute vector of each node $\mathbf{v}_t$ is:

$$d_t = [\mathbf{D}_t, \ \omega] \tag{3}$$

where ω is the angle between two neighboring as depicted in Fig.2.

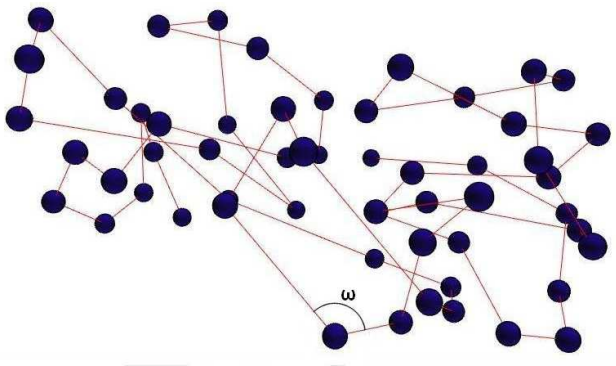


Fig. 2. The final graph that corresponds to the 3D model of protein 2dabA

2.4. Matching Method

Every protein is now represented as an undirected node attributed graph. In order to find the similarity between two proteins, for classification purposes, a graph matching technique is used. As proposed in [10], in order to match two undirected node attributed graphs $G_A = \{V_A, E_A, \{\mathbf{A}^A\}, \{\mathbf{B}_i^A\}_{i=1}^{s}\}$, $|V_A| = n_A$ and $G_B = \{V_B, E_B, \{\mathbf{A}^B\}, \{\mathbf{B}_i^B\}_{i=1}^{s}\}$, $|V_B| = n_B$, respectively, they must be aligned based on the Successive Projection Graph Matching Algorithm (SPGM) [10]. If G_A is partially matched to G_B, then a transformation matrix $\mathbf{P}$ that transforms G_A to G_B does exist. By utilizing the SPGM algorithm the best possible estimation $\mathbf{P}$ is computed. This estimation is referred as $\bar{\mathbf{P}}$. In SPGM the problem of attributed graph matching is reduced to an optimization problem. The matrix $\bar{\mathbf{P}}$ is calculated with an iterative procedure so as to minimize the following function:

Classes	Class Population
1a0cA	6
1abwA	94
1bbzA	4
1cnzA	18
1ycc	7
4icb	8
6mhtA	13
Total	150

Table 1. Dataset Population Table

$$J(\mathbf{p}) = -\frac{1}{2}\mathbf{p}^T\mathbf{X}\mathbf{p} - \frac{1}{2}\mathbf{p}^T\mathbf{I}\mathbf{p} - \mathbf{y}^T\mathbf{p} \tag{4}$$

In the above equation $\mathbf{p} = \mathbf{vec}(\bar{\mathbf{P}})$ (where $\mathbf{vec}(.)$ denotes the vectorization operation applied on matrix $\bar{\mathbf{P}}$), $\mathbf{X}$ is a compatibility matrix regarding the adjacency matrices, $\mathbf{y}$ is a compatibility vector regarding the nodes, and $\mathbf{I}$ is the appropriate identity matrix.

The proteins are compared in pairs and the $\bar{\mathbf{P}} = [\bar{p}_{ij}]$, where $\bar{p}_{ij}$ expresses the probability that the $\imath - th$ node of the first protein matches with the $\jmath - th$ node of the second protein, is calculated.

The final step is to combine the geometric and topological information derived previously with biochemical data. For this reason, the probability matrix $\bar{\mathbf{P}}$ is scanned so as to find contiguous pairs of nodes with high matching probability. A pair of nodes is consider matched if it has a matching score above a threshold (this threshold in the current application is selected to be 0.90). Furthermore, the number of contiguous pairs should be greater than 5% of the number of nodes of the smallest graph, so as to not take into account obsolete series of matched nodes. Then, the pairs of nodes that fulfill the aforementioned criteria are examined residue by residue and the corresponding score of the PAM250 matrix is retrieved. Finally, all the scores are normalized, into values between -1 and 1, and summed up to produce a compatibility score for the two chains. This procedure takes place for each of the valid chains of the two proteins. The scores that derive are added, providing us with a final compatibility score between the proteinic models at hand.

3. EXPERIMENTAL RESULTS

The proposed method was tested in terms of classification accuracy, on a dataset consisting of 150 proteins retrieved from the Protein Data Bank, as shown in Table 1. These proteins were classified using the DALI/FSSP method as the ground truth.

In order to evaluate the proposed method, a classification experiment was performed as follows:

- the proteins were compared in pairs and the final similarity score was stored in a matrix.

- the scores were sorted and the proteins were classified using the nearest neighbor approach.

The above experiment proved that the proposed method can efficient classify the total dataset with a percentage accuracy of $97,33\%$. This result can be considered particularly accurate since it performs near to ground truth. It can also be used as a filtering procedure before any complex biological procedure takes place.

Apart from the classification performance, the efficiency of the proposed shape comparison method was evaluated in terms of retrieval performance. In this case, its model of the database was used as query and the retrieved proteins were ranked in terms of same similarity to the query. For the presentation of the results the Precision-Recall curve [7] was used. the results are depicted in Fig. 3. It can be inferred that the proposed method retains high performance in all values of recall.

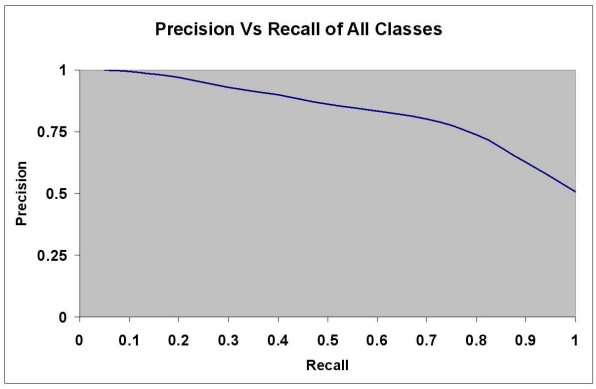


Fig. 3. Precision-Recall curve

4. CONCLUSIONS

In this paper a novel method that combines geometrical, topological and biochemical data in order to compare and classify proteinic data was proposed. The 3D representation of each protein was derived from the PDB file and segmented in order to create a proteinic graph. Each segment was described using the spherical harmonics coefficients and these descriptors were used as the graph nodes' attributes. During the matching process, firstly an attributed graph matching algorithm was applied and then, a similarity metric, which efficiently combines geometrical topological and biochemical information, was computed.

The experimental results were found particularly encouraging as the classification accuracy outperforms 97%.

5. REFERENCES

[1] E.Krissinel and K.Henrick, "Secondary-structure matching (ssm), a new tool for fast protein structure allignment in three dimensions," *Acta Cryst D60,2256-2268*, October 2004.

[2] M.Jambon, A.Imberty, G.Deléage, and C.Geourjon, "A new bioinformatic approach to detect common 3d sites in protein structures," *PROTEINS: Structure, Function, and Genetics 52:127-145*, 2003.

[3] Y.Wuang, F.Makedon, and J.Ford, "A bipartite graph matching framework for finding correspondences between structural elements in two proteins," in *Proceedings of the 26th annual International Conference of the IEEE EMBS*, San Francisco, CA, USA, September 2004.

[4] H.W.Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, pp. 83–97, 1955.

[5] C.N.Cavassotto, A.J.W.Orry, and R.A.Abagyan, "Structure-based identification of binding sites, native ligands and potential inhibitors for g-protein coupled receptors," 2003.

[6] C.C.Chen, J.T.Tu, P.K.Chang, B.Y.Chen, R.H.Liang, and M.Ouhyoung, "Protein function prediction by matching 3d structural data," in *Proceedings of NICOGRAPH International*, Hsinchu, Taiwan, 2004, pp. p.113 – p.120.

[7] P.Daras, D.Zarpalas, A.Axenopoulos, D.Tzovaras, and M.G.Srintzis, "Three-dimensional shape-structure comparison method for protein classification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 3 Issue 3, pp. 193–207, July 2006.

[8] H.Berman, J.Westbrook, Z.Feng, G.Gilliland, T.Bhat, H.Weissig, I.Shindyalov, and P.Bourne, "The protein data bank," *Nucl. Acid. Res. 28*, pp. 235–242, 2000.

[9] M.Kazhdan, T.Funkhouser, and S.Rusinkiewicz, "Rotation invariant spherical harmonic representation of 3d shape descriptors," in *Symposium on Geometry Processing*, 2003, pp. 167–175.

[10] B.J. van Wyk and M.A. van Wyk, "A pocs-based graph matching algorithm," *IEEE Transactions On Pattern Analysis and Machine Intelligence*, vol. 26 No. 11, pp. 1526–1530, November 2004.

[11] R.M.Schwartz and M.O.Dayhoff, "Matrices for detecting distant relationships.," *In M. O. Dayhoff, editor, Atlas of Protein Sequences National Biomedical Research Foundation*, pp. 353–358, 1979.