SEARCH AND RETRIEVAL OF 3D PROTEIN STRUCTURES

P. Daras [1], *D. Zarpalas* [2], *D. Tzovaras* [2] *and M. G. Strintzis* [1,2] Fellow, IEEE

[1]Information Processing Laboratory
Electrical and Computer Engineering Department
Aristotle University of Thessaloniki, Greece
540 06 Thessaloniki, Greece

[2]Informatics and Telematics Institute
1st Km Thermi-Panorama Road,
Thessaloniki 57001, Greece
Email:daras@iti.gr

ABSTRACT

In this paper a 3D shape-based approach is presented for the efficient search and retrieval of protein molecules. The method primarily relies on the geometric 3D appearance of the proteins, which is produced from the corresponding PDB files and secondarily on their primary and secondary structure. After proper positioning and alignment of the 3D proteins, a variation of the Generalized Radon Transform is applied to the 3D structures so as to produce descriptor vectors, which describe their shape. Further, descriptor vectors that characterize the primary and the secondary structure of the proteins are extracted. Then, different weight factors are assigned to each descriptor vector, which are introduced in a matching algorithm so as to retrieve the proteins with highest geometric similarity. A part of the FSSP/DALI database was used as ground truth in order to measure the efficiency of the proposed method.

1. INTRODUCTION

The structure of a molecule in 3D space is the main factor, which determines its chemical properties as well as its function. Therefore, the 3D representation of a residue sequence and the way this sequence is folding in the 3D space, is very important in order to be able to understand the "logic" in which a function or biological action of a protein is based on. Nowadays, due to the rapid development of X-Ray crystallography methods as well as the NMR spectrum analysis techniques, a high number of new 3D structures of molecules is determined. These 3D structures are stored in the worldwide repository Protein Data Bank (PDB) [1]. The number of the 3D molecular structure data increases rapidly since almost 200 new structures are stored per month in PDB. Today there are over than 24.000 3D protein and nucleic acid molecular models in this repository.

Structural classification schemata are already available for protein databases. The CATH [2] schema provides an hierarchical classification of the proteins into five categories: Class, Architecture, Topology, Homology Superfamily and Sequence. Although this system is mainly automatic, experts are often needed in the case of poor results providing from these algorithms. The SCOP schema (Structural Classification of Proteins) [3] also provides an hierarchical classification describing the structural and evolutionary relationship of the proteins. This classification includes the following categories: Family, Superfamily, Fold and Class. Another classification method is available from the FSSP database (Families of Structurally Similar Proteins) [4] where an algorithm (DALI) [5] for optimal pairwise alignment of protein structures is used.

The 3D coordinates of every protein are used for the creation of distance matrices that contain the distance between amino-acids (the distance between their $C^{alpha} - C^{alpha}$ atoms). These matrices are, firstly, decomposed into elementary formats, e.g. hexapeptidic-hexapeptidic submatrices. Similar formats make pairs and the emerging formats create new coherent pairs. Finally, a Monte Carlo procedure is used for the optimization of the similarity measure concerning the inner-molecular distances. This method is fully automatic and identifies structural resemblances and common structural cores accurately and sensitively, but it is very computationally expensive due to the many different alignments performed, the optimization procedures and the extremely high number of distances between amino-acids since a protein may consist of thousands of amino-acids.

What is desired, is an efficient algorithm that may act as a fast filter for further investigation and that may be restricted e.g. primarily to geometric aspects. Additionally, since most of the times proteins with similar 3D structure have similar functions, a geometric filtering may lead the biologists to discover and identify new possible functions. Up to now few methods have been reported that take into account primarily the geometric similarity of the proteins. In [6] a new classification method is proposed, which is based on the concept of 3D histograms for protein representation and a flexible similarity quadratic distance function. In [7] geometric features based on the moments [8] are extracted for each molecule along with characteristics of their primary and secondary structure. Both aforementioned meth-

ods were shown very good results using as ground truth the FSSP/DALI classification, with less complexity and much simplicity.

Inspired by these approaches a form of the Generalized Radon Transform [9] is used in this paper so as to extract descriptor vectors that describe the 3D shape of a protein molecule. Specifically, the Cylindrical Integration Transform (CIT) as well as its enhanced form (EnCIT), successfully used in 3D content-based search and retrieval [10], are proposed in this paper for identifying the geometric similarity of the proteins. That work is also extended here using some features, describing the primary and secondary structure of the proteins and also proposed in [7]. Further, weight factors are given to each descriptor vector assigning a higher value to the geometric ones and smaller to others. Finally, the descriptor vectors are inserted in a matching algorithm, which is used for the efficient retrieval of similar proteins.

The rest of the paper is organized as follows: In Section 2 the proposed descriptor vector extraction method and the matching algorithm are presented. Experimental results evaluating the efficiency of the proposed method are described in Section 3. Finally, conclusions are drawn in Section 4.

2. DESCRIPTOR EXTRACTION

A protein is mainly composed of Carbon (C), Nitrogen (N), Oxygen (O), Hydrogen (H) and Sulfur (S). In Figure 1 the 3D representation of a protein is shown. The colors used and the atomic radii are listed in Table 1.

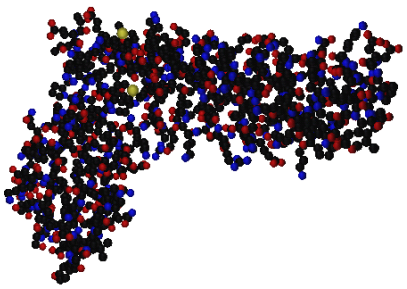

Fig. 1. The protein 1DD5.

The surface of each atom is triangulated. That way a protein P is comprised of a set of vertices $\mathbf{V}$ and a set of connections between the vertices.

Atom	Symbol	Radius (Å)	Color
$Carbon$	C	0.77	$Black$
$Nitrogen$	N	0.70	$Blue$
$Oxygen$	O	0.66	Red
$Hydrogen$	H	0.37	$White$
$Sulfur$	S	1.04	$Yellow$

TABLE 1: MAIN ATOMS OF A PROTEIN.

2.1. The 3D Generalized Radon Transform

Let $f(\boldsymbol{x})$, $\boldsymbol{x} = [x, y, z]$, the volumetric binary function of P, which is defined as:

$$f(\boldsymbol{x}) = \begin{cases} 1, & \text{when } \boldsymbol{x} \text{ lies within the 3D model's volume} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

Let, also, $\boldsymbol{\eta}$ be a unit length vector in $\Re^3$, $\boldsymbol{\eta} = [sin\phi cos\theta, sin\phi sin\theta, cos\phi]$ and l a real number.

The 3D Generalized Radon transform $R_f(\boldsymbol{\eta}, l)$ is a function which associates to each pair $(\boldsymbol{\eta}, l)$ the integral of $f(\boldsymbol{x})$ on the curve $C(\boldsymbol{\eta}, l) = \{\boldsymbol{x} | \psi(\boldsymbol{x}; \boldsymbol{\eta}, l) = 0\}$, where $\psi(\boldsymbol{x}; \boldsymbol{\eta}, l)$ denotes the transformation curve.

$$R_f(\boldsymbol{\eta}, l) = \int_{-\infty}^{+\infty} f(\boldsymbol{x}) \delta(\psi(\boldsymbol{x}; \boldsymbol{\eta}, l)) d\boldsymbol{x} \quad (2)$$

The Cylindrical Integration Transform (CIT) [10] is a function which associates only to each $\boldsymbol{\eta}$, since the radius of the cylinder l can be set to be Th, the integral of $f(\boldsymbol{x})$ on the cylinder $CYL(\boldsymbol{\eta})$:

$$CIT_f(\boldsymbol{\eta}) = \int_{\boldsymbol{x} \in CYL(\boldsymbol{\eta})} f(\boldsymbol{x}) d\boldsymbol{x} \quad (3)$$

where $x \in CYL(\boldsymbol{\eta})$ simply means: $\sqrt{|\boldsymbol{x}|^2 - |\boldsymbol{x} \cdot \boldsymbol{\eta}|^2} \leq Th$.

The discrete form of CIT, which will be used for the actual extraction of the shape descriptors is given by:

$$CIT(\boldsymbol{\eta}_i) = \sum_{\boldsymbol{x}_j \in CYL(\boldsymbol{\eta}_i)} f(\boldsymbol{x}_j) \quad (4)$$

$i = 1, \ldots, N_{CYL}$, $j = 1, \ldots, J$, where N_{CYL} is the total number of cylinders and J the total number of points $\boldsymbol{x}_j$.

2.2. Preprocessing

Before applying the proposed technique a canonical position is achieved for each P. First, the center of mass of P is calculated and each V is translated so that the new center of mass is at the origin. Next, the distance d_{max} between the new origin and the most distant vertex is computed, and P is

scaled so that $d_{max} = 1$. Finally, the Principal Component Analysis (PCA) algorithm is applied in order to achieve rotation normalization. The translated, rotated and scaled P is then placed into a bounding sphere, with radius d_{max}, which is partitioned in equal cube shaped voxels $\boldsymbol{v}$. Then, the discrete binary volume function $\hat{f}(\boldsymbol{v})$ of P is defined as:

$$\hat{f}(\boldsymbol{v}) = \begin{cases} 1, & \text{when } \boldsymbol{v} \text{ lies inside P} \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

2.3. Geometric feature extraction

After proper positioning of the 3D structure P, the procedure introduced in [10] is followed, producing the CIT vector with elements $CIT_f(\boldsymbol{\eta}_i)$, where $i \in S_C = \{1, ..., N_{CYL}\}$. The large and important amount of information contained in the $CIT_f(\boldsymbol{\eta})$ can be further exploited in order to enhance the CIT-based descriptor vectors. For this reason, the following set of functionals is used:

$$1.F_1(g) = \max\{g(t_i)\} \quad (6)$$

$$2.F_2(g) = \sum_{i=1}^{N} |g'(t_i)| \quad (7)$$

$$3.F_3(g) = \sum_{i=1}^{N} g(t_i) \quad (8)$$

$$4.F_4(g) = \max\{g(t_i)\} - \min\{g(t_i)\} \quad (9)$$

where g is a differentiable function, g' its derivative, $t_i, i = 1, \ldots, N$ are sample points for g and N is their total number.

The goal is the gradual reduction of the dimensions of $CIT_f(\theta, \phi)$, so as to produce a more compact representation of the descriptor vector. Therefore the descriptor vector $\mathbf{u}_{EnCIT}(i)$ is produced, where $i = 1, \ldots, N_{EnCIT}$.

2.4. Attribute feature extraction

Besides the geometric feature vectors, features that characterize the primary and secondary structure of a protein are also extracted [7]. More specifically, concerning the primary structure, the ratio of the amino-acids' occurrences relative to the total number of amino-acids (20 descriptors), the hydrophobic amino-acids ratio (1 descriptor) and the ratio of the helix types' occurrences (10 descriptors) containing in a protein, are calculated. Concerning the secondary structure, the number of Helices (1 descriptor), Sheets (1 descriptor) and Turns (1 descriptor), containing in a protein are calculated. The different features are listed in Table 2. All the aforementioned information is contained in each PDB file.

Geometric features	**Weight**
CIT	0.20
EnCIT	0.50
Secondary structure features	**Weight**
No of HELICES	0.04
No of SHEETS	0.03
No of TURNS	0.03
Primary structure features	**Weight**
Hydrophobic residue ratio	0.04
Helix Type	0.04
Residue ratio	0.12

TABLE 2: DESCRIPTORS AND THEIR WEIGHTS.

2.5. Matching algorithm

Let A, B, be two proteins. The descriptors are compared in pairs using the following formula (based on their L1-distance):

$$Similarity = (1 - \sum_{i=1}^{N_{total}} w_i \frac{|\mathbf{u}_A(i) - \mathbf{u}_B(i)|}{|\mathbf{u}_A(i) + \mathbf{u}_B(i)|/2}) \cdot 100\% \quad (10)$$

The dimension of the descriptor vectors is $N_{total} = N_{CYL} + N_{EnCIT} + 34$, where $N_{CYL} = 252$, and $N_{EnCIT} = 32$.

3. EXPERIMENTAL RESULTS

The proposed method was tested using a part of the database with proteins created by the DALI method. To evaluate the efficiency of the proposed method, three groups were formed with representatives the proteins 1a6m (consisting of 189 proteins), 1l92 (387 proteins) and 2cba (180 proteins) respectively. Each protein was used as a query object. The retrieval performance was evaluated in terms of "precision" and "recall", where precision is the proportion of the retrieved models that are relevant to the query and recall is the proportion of relevant models in the entire database that are retrieved in the query. In Figure 2 the overall averaged precision-recall is depicted using the geometrical descriptor vectors and the integrated descriptor vectors.

In figure 3 the precision-recall diagram of the representatives of each group is presented.

In order to measure the classification accuracy, the nearest neighbor was found after removing the query molecule from the group it belongs to. The overall classification accuracy was computed as the percentage of the correctly predicted class among all 756 proteins and it was found to be 99.6% since only 3 out of 756 proteins missed, using only the geometrical descriptor vectors. Further improvements were achieved introducing the descriptors from the

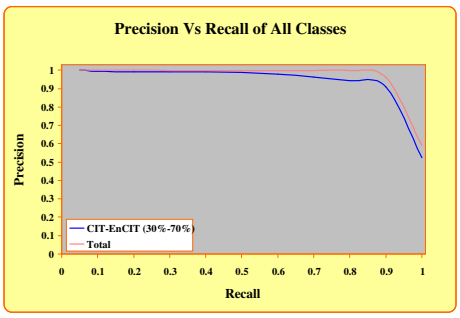

Fig. 2. Average precision-recall of all classes.

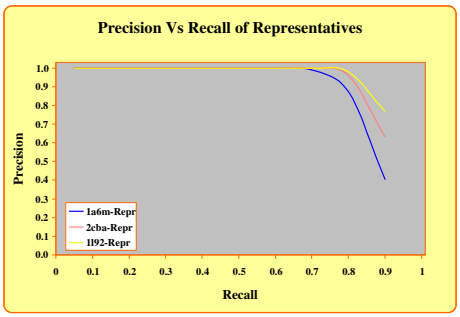

Fig. 3. Precision-recall of the representatives.

secondary and primary structure of the proteins where the percentage of the correct classified proteins was 100%.The time needed for an one-to-all comparison of a query protein descriptor vector to all (755) descriptor vectors is 50 seconds. The time needed for the complete preprocessing procedure, from the creation of the 3D structure up to the final normalization step, is approximately 3 min. On the other hand, for the comparison of a query protein to 385 proteins, using the DALI method, an entire day is needed [4]. Thus, the approach followed performs near to ground truth with less complexity and much simplicity and it might be efficiently used as a filter before taking into account the chemical properties of the proteins.

4. CONCLUSIONS

In this paper a novel approach for the retrieval of similar 3D proteins is proposed. The approach consists of two steps: an offline and an online. In the offline step the proteins taken from a PDB file are visualized creating spheres. The spheres are triangulated. Next, each protein is translated, scaled, rotated and voxelized. Finally, one transform, originated from the Generalized Radon Transform is applied and feature vectors describing the geometric characteristics of each protein are extracted. Additionally, feature vectors, which correspond to the protein's primary and secondary structure extracted as well. In the online step, a matching algorithm is applied to the feature vectors assigning a different weight factor to each vector.

Experiments performed evaluating the efficiency of the proposed method using as ground truth the DALI database, in terms of precision-recall and classification accuracy. The retrieval results using the proposed method were very close to those of the ground truth with less complexity and much simplicity. Thus, the proposed method can be efficiently used as a filter before the complex chemical properties of the proteins taken under consideration.

5. REFERENCES

[1] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, "The Protein Data Bank", *Nucleic Acids Research*, 28 pp. 235-242, 2000.

[2] C.A. Orengo, A.D. Michie, D.T. Jones, M.B. Swindells and J.M. Thornton, "CATH - A hierarchic classification of protein domain structures", *Structure*, 5(8), pp. 1093-1108, 1997.

[3] T.J.P. Hubbard, A.G. Murzin, S.E. Brenner and C. Chothia. "SCOP: a structural classification of proteins database", *Nucleic Acids Research*, 25, pp. 236-239, 1997.

[4] L. Holm, and C. Sander, "The FSSP database: fold classification based on structure-structure alignment of proteins", *Nucleic Acids Research*, 24, pp. 206-210, 1996.

[5] L. Holm, and C. Sander, "Touring Protein Fold Space with Dali/FSSP", *Nucleic Acids Research*, 26, pp. 316-319, 1998.

[6] M. Ankerst, G. Kastenmuller, H.P. Kriegel and T. Seidl, "Nearest Neigbor Classification in 3D Protein Databases", *In Proc. of the 7th Int. Conf. on Intelligent Systems for Molecular Biology (ISMB 1999)*, Heidelberg, Germany, 1999.

[7] C. Zhang and T. Chen, "Retrieval of 3D Protein Structures", *In Proc. of IEEE Int. Conf. on Image Processing (ICIP 2002)*, Rochester, NY, U.S.A., Sep. 2002.

[8] C. Zhang and T. Chen, "Efficient feature extraction for 2D/3D objects in mesh representation", *In Proc. of IEEE Int. Conf. on Image Processing (ICIP 2001)*, Vol. 3, pp. 935-938, Thessaloniki, Greece, 7-10 Oct. 2001.

[9] P. Daras, D. Zarpalas, D. Tzobaras and M.G. Strintzis, "Watermarking of 3D Modes for Data Hiding", *In Proc. of IEEE Int. Conf. on Image Processing (ICIP 2004)*, Singapore, 24-27 Oct. 2004.

[10] P. Daras, D. Zarpalas, D. Tzobaras and M.G. Strintzis, "3D Model Watermarking for Indexing Using the Generalized Radon Transform", *3D Data Processing, Visualization and Transmission (3DPVT 2004)*, Thessaloniki, Greece, Sept 6-9, 2004.