STRUCTURAL BIOLOGY AND PATTERN RECOGNITION

V. Cantoni,[1] A. Ferone,[2] O. Ozbudak,[3] and A. Petrosino[2]

[1] *University of Pavia, Department of Electrical and Computer Engineering, Via A. Ferrata, 1, 27100, Pavia, Italy*

virginio.cantoni@unipv.it

[2] *University of Naples Parthenope, Department of Applied Science, Centro Direzionale Isola C4, 80133, Napoli, Italy*

{alfredo.petrosino,alessio.ferone}@uniparthenope.it

[3] *Istanbul Technical University, Department of Electronics and Communication Engineering, Ayazaga Campus, 34469, Maslak, Istanbul, Turkey*

ozbudak@itu.edu.tr

Abstract In this paper we present two tracks based on the exploitation of protein structure comparison in 3D: the first for retrieving a structural block (small subset or even an entire macromolecule) from a protein data base, and the second for protein supervised classification. The first track is based on the Generalized Hough Transform (GHT) which is applied adopting two different strategies. The former algorithm uses exhaustive matching of all possible subsets of the macromolecule having the same cardinality of the structural block model with the model itself. The second strategy exploits the retrieval of co-occurrences of two (but may be extended to three, or more) basic components for contributing to the motif recognition. The later approach is based on a protein peculiar representation through an Extended Gaussian Image (EGI). In the EGI the histogram of the orientations of an object surface is mapped on the unitary sphere. We propose to adopt a similar data-structure named Protein Gaussian Image (PGI) for representing the orientation of the protein Secondary Structures (SSs) (helices and sheets), storing the features of the SS. For the taxonomy of proteins into given classes, on the basis of the PGI representation, a neural network implementing the Kohonen self-organizing feature maps is adopted. We consider both these approaches very effective for a preliminary screening in looking on the PDB and for confirming the SCOP classification.

Keywords: Kohonen maps, Generalized Hough transform, Protein classification, Protein motif retrieval, Protein structure comparison, Self organizing maps for structured data.

Introduction

Structural biology is a branch of life science concerned with the study of the structure of biological macromolecules like proteins and nucleic acids. Note that, the structure of a protein gives much more insights in the function of the protein than its amino acid sequence. To look for patterns among the diverse sequences that give rise to particular shapes many strategies have been developed by the bioinformatics community. Proteins functions are conditioned by their spatial structures, so protein structure comparison is important for understanding the evolutionary relationships among proteins, predicting protein functions, and predicting protein structures from amino acid sequences.

Proteins are formed by two basic regular 3D structural patterns called SSs: helices and sheets [1]. A structural motif is a compact three-dimensional protein structure referring to a small specific combination of secondary structural elements which appears in a variety of molecules. These elements are often called super SSs. Note that, while the spatial sequence of elements is the same in all instances of a motif, they may be encoded in any order: in this sense sequence is sometimes misleading and the structure analysis may give much more insights [2]. The segmentation of the protein backbone in SSs is achieved through common packages such as DSSP and STRIDE [3,4,5]. On the average 4.8% of the residues are differently assigned.

Several motifs are packed together to form compact, local, semi-independent units called domains, i.e. with more interactions within them than with the rest of the protein. Therefore, a structural domain forms a compact 3D structure, independently stable, and can be determined by two characteristics: its compactness and its extent of isolation.

Many proteins consist of several structural domains to form multi-domain and multifunctional molecules in which each domain may fulfill its function independently, or concurrently with its neighbors. Many domains could have once existed as independent proteins. Multi-domain proteins are likely to have emerged from a selective pressure during evolution to create new functions. Various proteins have diverged from common ancestors by different combinations and associations of domains.

This hierarchical makeup of macromolecules is quite explicit in the F. Jacob's aphorism: *Nature is a tinkerer and not an inventor* [6], that is new sequences are adapted from pre-existing ones rather than invented, in fact motifs and domains are the common material used by nature to generate new sequences.

A structural motif is a 3D structural element and usually consists of just a few SSs, each one with an average of approximately 5 and 10 residues for sheets and helices, respectively. Several motifs are packed together to form domains, the size of individual structural domains varies from between about

25 up to 500 amino acids, but the majority, 90%, has less than 200 residues with an average of approximately 100 residues. Note that, it is often used the term super-*, where * can stand for motif, or domain, or family, or fold, or class.

1. Structural Blocks Retrieval

In recent years we have developed starting from traditional pattern recognition techniques new approaches for retrieving a model (motif, or domain, or ..., or an entire protein) within a protein or in the entire Protein Data Base (PDB) [7], by using structure comparison in 3D.

The General Hough Transform Approach

Our approach for structural block retrieval exploits the GHT [8,9]. A first method implements a peculiar exhaustive matching and the second one directly uses co-occurrences of two SSs for block retrieving. Both these approaches use three parameters for comparison [10]. These are midpoint distance, axis distance and angle related to two SSs in 3D. Midpoint distance (Md) is the Euclidean distance between middle points of two SSs, axis distance (Ad) is the shortest distance between the axis of two SSs and the angle is the angle (φ) between the two SSs translated to present a common extreme (see Fig. 1).

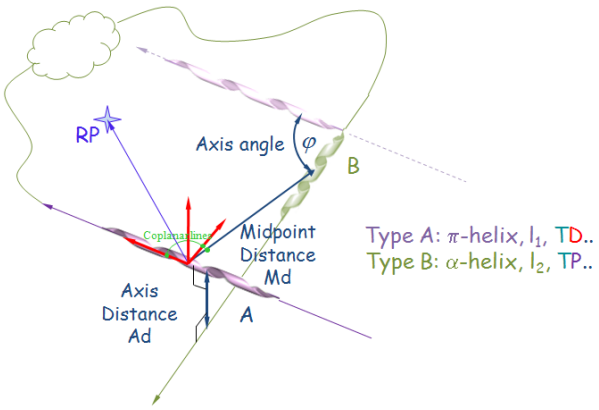

Figure 1. RP parameter terns for the GHT approaches: Md, Ad and φ.

In both methods the barycenter of the model is assigned as Reference Point (RP) and in order to find the RP in the macromolecule a GHT voting process is performed. These methods are similar for what refers to the basic process and adopt the same parameter space but differ in the voting process. The first method compares the model with all the possible model instances in the macro-

molecule or protein and for every correspondence a weighted contribution is given to the candidate barycenter. The second one is based directly on the GHT and compares every SS couple of the model with all the possible couples in the macromolecule. For every compatible correspondence a vote is given to the point which is figured out with a special mapping rule [11,12]. In both methods, after the voting process, the point which has the maximum number of votes is candidate as RP (it is obviously known the expected number of model co-occurrences).

As an example in the sequel are shown some results with the two approaches for a known protein and a common typical four SS motif (the Greek Key one) (see Figs. 2 and 3). The experimental results show that the RP was determined with efficiency and precision in both the GHT approaches. We consider this representation very effective for a preliminary screening in looking on the PDB.

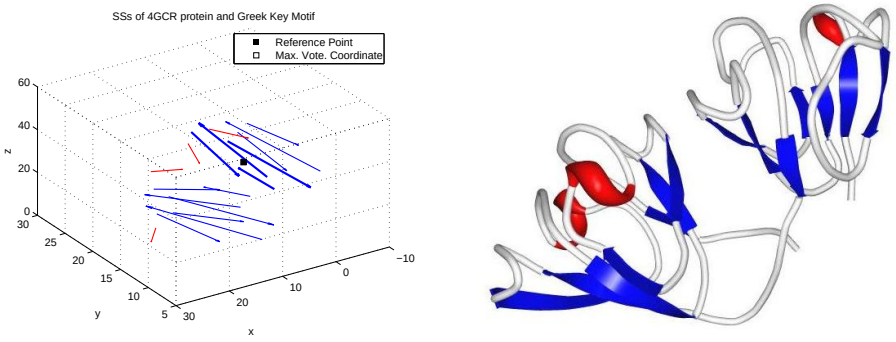

Figure 2. SSs of the 4GCR protein. On the right a picture generated by PyMOL on PDB file 4GCR. Blue lines are β-sheets and red lines are α-helices. Bold lines correspond to the searched Greek Key Motif formed by four SSs (residues from 34 to 62). The result refers to the exhaustive matching, the RP and the maximum vote coordinates are almost coincident.

2. Protein Gaussian Images and Neural Network Approach

A second approach for model retrieval applies the Extended Gaussian Image (EGI) [13,14] which maps on the unitary sphere the histogram of the orientations of the object surface. We propose to adopt a similar "abstract" data-structure named Protein Gaussian Image (PGI) [15] for representing the orientation of the protein SSs (helices and sheets). The "concrete" data structure is the same as for the EGI, however, in this case the points of the Gaussian sphere do not contain the area of the patches having that orientation, but fea-

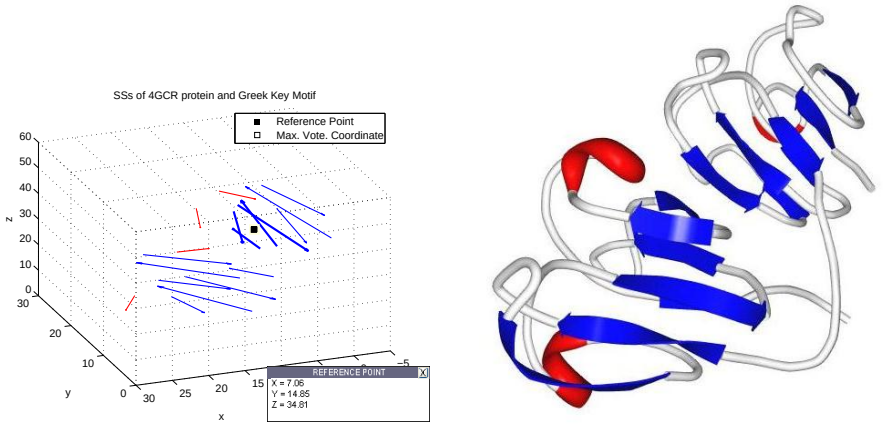

Figure 3. SSs of the 4GCR protein. Blue lines are β-sheets and red lines are α-helices. Bold lines correspond to the searched Greek Key Motif formed by four SSs. The result refers to the co-occurrence of SS couple, the RP and the max. vote coordinates are coincident.

tures of the SSs having that direction. Among the features we may include the versus (origin versus surface or vice versa), the length of the structure (number of amino acids), biochemical properties, and even the sequence of the amino acids (such as in a list). Moreover the chain sequence of SSs is recorded as a list which is mapped on the sphere surface. In Figure 4, an example of a protein (4GCR) is represented as a PGI.

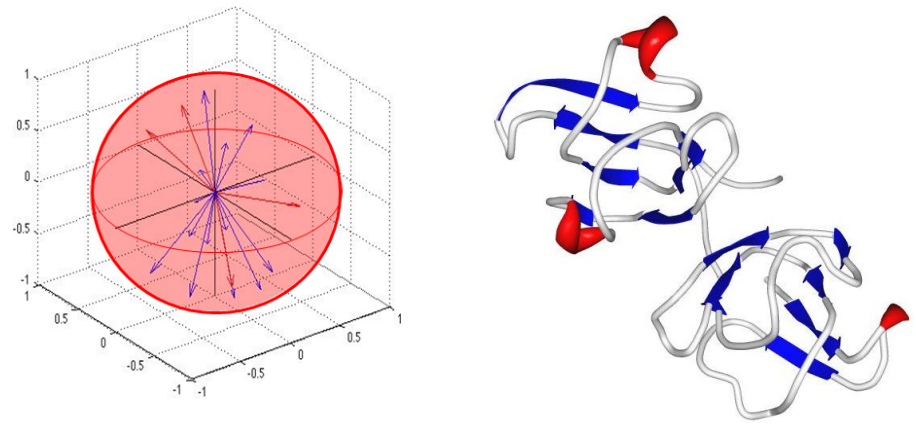

Figure 4. Left Protein Gaussian Image of protein 4GCR. Red arrows represent the Greek Key motif. Right SSs of the 4GCR protein. Blue lines are β-sheets and red lines are α-helices.

The proposed data structure is complete (no information is lost for an analytic analysis) and effective from the computational viewpoints (only two reference coordinates are needed), but also supports effectively the structural perception. In order to validate the effectiveness of the PGI representation of the protein structure, we propose to employ the Self Organizing Maps–Structured Data (SOM-SD) [16] framework for structured data in a practical structural learning problem, where each protein is represented by a PGI. The aim of the SOM learning algorithm [17] is to learn a feature map which, given a vector in the input space returns a point in the output space. This is obtained in the SOM by associating each point in the output space to a different neuron. Given an input vector $\mathbf{v}$, the SOM returns the coordinates, within the output space, of the neuron with the closest weight vector. Thus, the set of neurons induces a partition of the input space, where input vectors that are close to each other will activate neighbor neurons. SOM-SD represents an extension of the SOM framework, where the input space is a structured domain and the computational framework is similar to that defined for recursive neural networks [18].

The dataset employed in the classification task is composed of 45 proteins classified by SCOP (Structural Classification of Proteins) [19] as belonging to the class *Alpha and beta proteins* (a/b). Three folds have been considered, namely *Flavodoxin-like*, *RibonucleaseH-likemotif* and *TIMbeta/alpha-barrel*, and for each fold, 15 proteins have been chosen. The task consists in grouping proteins or side chains belonging to the same fold.

The first test has been conducted considering the whole protein as a pattern, i.e., each protein is represented by a PGI. It can be observed in Table 1 that the results are quite good in terms of clustering performance. Even though this measure does not take into account the desired clustering outcome, the result is supported by the good retrieval performance which reflects a reduced confusion in the mapping of each pattern. The classification performance, reflecting the performance with respect to the desired clustering outcome, shows less accurate results, but with an interesting peak at 74.82%.

Table 1. Performance of a 200×200 SOM–SD considering the whole proteins.

	Test Set								
Learning Rate	1			1.25			1.5		
Iterations	40	60	80	40	60	80	40	60	80
Retrieval	84.99	84.08	84.86	86.46	87.69	79.79	79.29	82.31	79.82
Classification	72.65	56.95	56.91	59.39	70.65	60.73	65.91	64.30	74.82
Clustering	0.79	0.85	0.83	0.82	0.81	0.85	0.83	0.79	0.80

The second test has been performed considering as patterns the single side chains of each protein, i.e., each side chain is represented by a PGI. From Table 2 it can be noted how this "reduced" representation yields better results

Table 2. Performance of a 200 × 200 SOM–SD considering the single side chains of each protein.

	Test Set								
Learning Rate	1			1.25			1.5		
Iterations	40	60	80	40	60	80	40	60	80
Retrieval	74.39	81.67	79.63	92.72	92.35	93.40	92.36	92.34	93.85
Classification	75.58	76.37	77.64	85.11	84.14	84.17	85.03	85.78	86.42
Clustering	0.80	0.80	0.80	0.79	0.80	0.79	0.79	0.80	0.79

in terms of classification and retrieval performance while performing slightly worst with respect to clustering performance. In particular, the clustering performance is almost the same but with a higher confidence reflected by the higher retrieval performance. The interesting result concerns the classification performance that is much higher considering only the side chain.

We consider the PGI representation very effective for a fast protein classification.

3. Conclusions

Protein functions are determined by their spatial structure, for this reason it is important to learn structure-function relationship in the protein universe by comparing their structures and retrieving similar motifs, domains, proteins. This paper pursues two relevant goals: i) retrieving a structural block (composed by a number of SSs ranging from four to tenth -the motif case-, up to tenths to hundred -domain and entire protein case) from a macromolecule or the PDB using two different approaches; ii) classifying proteins into different known protein classes. The new proposed solutions appeared quite effective.

For blocks retrieval it can be concluded that GHT is a very efficient method for protein motif matching. It is also worth to point out that the GHT is indeed suited for parallel implementation (for example at the motif level - e.g. more motifs can contribute concurrently - but also at the lower SS level - e.g. helices and sheets can contribute in parallel), then the technique can be easily implemented on parallel machines, so reducing consistently the computation time. The data base under examination can contain all the about 80 thousand proteins at present included in the PDB, and the objective is to return on-line the hypotheses of similarity for a specific structural element. Obviously, this method can be speeded-up, because as described, the selected primitive is only the type of SS, but other information can be exploited as well, such as for example, the amino acid composition or/and the length in term of number of residues, etc., that up-to-now have not been considered. The results confirmed that, even in this simple implementation, the candidate RP is located with a good precision.

For the classification goal, these preliminary results demonstrate the quality and the effectiveness of the PGI representation and that this solution supports the efficient exploitation of neural network standard strategies such as the Kohonen maps. In this case, we have selected a supervised approach for testing: a well known SCOP class and three of its folds (the three with the highest cardinalities) also in this case the results look very promising.

We can conclude that both the proposed methods are simple to implement and then computationally efficient, but for what refers robustness with respect to the other approaches we need to experiment on more complex structures, and with an extended statistical performance evaluation. In the future work we will use these methods for more complex structures, adopting an extended statistical performance evaluation, and we will compare these methods quantitatively to other methods in terms of complexity, efficiency and speed.

References

[1] Eisenberg, D., (2003), *The Discovery of Alpha-Helix and Beta-Sheet, the Principal Structural Features of Proteins*, Proc. Of the National Academy of Sciences of the United States of America, vol.100, no.20, pp. 11207-11210.

[2] Frishman D., Argos P., (1995), *Knowledge-Based Protein Secondary Structure Assignment Proteins: Structure, Function, and Genetics*, 23:566-579.

[3] Kabsch W., Sander C., (1983), *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features*. Biopolymers 22 (12), pp. 2577-2637.

[4] http://swift.cmbi.kun.nl/gv/dssp/

[5] Heinig M., Frishman D., (2004), *STRIDE: a Web server for secondary structure assignment from known atomic coordinates of proteins*, Nucl. Acids Res., 32, W500-2, http://webclu.bio.wzw.tum.de/stride/

[6] Jacob F., (1977), *Evolution and tinkering*, *Science*, vol.96, no.1, pp. 1161-1166.

[7] http://www.pdb.org/

[8] Ballard, D.H., (1981), *Generalizing the Hough transform to detect arbitrary shapes*, Pattern Recognition, Vol. 13, N. 2.

[9] Sloan, K.R. Jr. and D.H. Ballard., (1980), *Experience with the generalized Hough transform*, Proc., 5th Int'l. Conf. Pattern Recognition & Image Processing, Miami Beach, FL.

[10] Cantoni, V., A. Ferone, O. Ozbudak and A. Petrosino., (2012), *Motif Retrieval by Exhaustive Matching and Couple Co-occurrences*, CIBB'12, inpress.

[11] Cantoni, V., A. Ferone, O. Ozbudak and A. Petrosino., (2012), *Search of Protein Structural Blocks through Secondary Structure Triplets*, IPTA'12, submitted.

[12] Cantoni V., Mattia E., Essay: *Hough transform for structural motif retrieval, Definitions: Hough Transform; Range Tree*; in Encyclopedia of Systems Biology, Werner Dubitzky, Olaf Wolkenhauer, Kwang-Hyun Cho, Hiroki Yokota (Eds.), Springer Science+Business Media LLC.

[13] Horn B.K.P., (1984), *Extended Gaussian images*, Proc. IEEE, Vol. 72, N. 12, pp. 1671-1686.

[14] Cantoni V., Lombardi L., Gaggia A., Essay: *Extended Gaussian Image for pocket-ligand matching, Definitions: EGI; CEGI e ECEGI*; in Encyclopedia of Systems Biology, Werner Dubitzky, Olaf Wolkenhauer, Kwang-Hyun Cho, Hiroki Yokota (Eds.), Springer Science+Business Media LLC.

[15] Cantoni V., Ferone A. and Petrosino A., Protein Gaussian Image (PGI) - *A Protein Structural Representation Based on the Spatial Attitude of Secondary Structure*. New Tools and Methods for Pattern Recognition in Complex Biological Systems, in press.

[16] Hagenbuchner M., Sperduti A., Ah Chung Tsoi, (2003), *A self-organizing map for adaptive processing of structured data*, Neural Networks, IEEE Transactions on , vol.14, no.3, pp. 491- 505.

[17] Kohonen T., (1994), *Self Organizing Maps*, Springer Series in Information Sciences, Springer: Espoo, Finland.

[18] Sperduti A., (2000), *A tutorial on neurocomputing of structures*, in Knowledge-Base Neurocomputing, I. Cloete and J. M. Zurada, Eds. Cambridge, MA: MIT Press, pp. 117-152

[19] Murzin A. G., Brenner S. E., Hubbard T., Chothia C., (1995), *SCOP: a structural classification of proteins database for the investigation of sequences and structures*, J. Mol. Biol. 247, pp. 536-540.