IL NUOVO CIMENTO Vol. 35 C, N. 5 Suppl. 1 Settembre-Ottobre 2012
DOI 10.1393/ncc/i2012-11334-2

Colloquia: PR PS BB 2011

Protein Gaussian Image (PGI)

A protein structural representation based on the spatial attitude of the secondary structure

V. Cantoni(1), A. Ferone(2), R. Oliva(2) and A. Petrosino(2)

(1) *University of Pavia - Pavia, Italy*
(2) *University of Naples "Parthenope" - Naples, Italy*

ricevuto il 30 Settembre 2011; approvato l' 1 Dicembre 2011

Summary. — A well-known shape representation usually applied for 3D object recognition is the Extended Gaussian Image (EGI) which maps the histogram of the orientations of the object surface on the unitary sphere. We propose to adopt an analogous "abstract" data-structure named Protein Gaussian Image (PNM) for representing the orientation of the protein secondary structures (*e.g.* helices or strands) which combines the characteristics of the EGI and the ones of needle maps. The "concrete" data structures is the same as for the EGI, with a hierarchy that starting with a discretization corresponding to the 20 orientations of the icosahedron facets, it is iteratively refined with a factor 4 at each new level (80, 320, 1280, ...) up to the maximum precision required. However, in this case to each orientation does not correspond the area of the patches having that orientation but the features of the protein secondary structures having that direction. Among the features we may include the versus (origin *versus* surface or vice versa), the length of the structure (*e.g.* the number of amino acids), biochemical properties, and even the sequence of the amino acids (stored as a list). We consider this representation very effective for a preliminary screening when looking in a protein data base for retrieval of a given structural block, or a domain, or even an entire protein. In fact, on this structure it is possible to identify the presence of a given motif, or also sheets (note that parallel or anti-parallel β-sheets are characterized by common or opposite directions of ladders). Herewith some known proteins are described with common typical motifs easily marked in the PGI.

PACS 87.18.Xr – Proteomics.
PACS 87.85.mk – Proteomics.
PACS 87.15.B- – Structure of biomolecules.
PACS 87.15.bd – Secondary structure.

1. – Preliminary statements and definitions

One of the most promising fields, on which the attention of different scientific communities is converging, is Structural Biology. Among these communities, certainly there is, for the consistent activity of the last half century on morphological analysis, the Pattern Recognition community.

In fact, a protein structure can be considered following a kind of hierarchical building process, that is well described by the F. Jacob aphorism: "Nature is a tinkerer and not an inventor" [1], *i.e.* 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. In proteins, a structural motif is a three-dimensional structural element which appears in a variety of molecules and usually consists of just a few elements. Several motifs packed together to form compact, local, semi-independent units are called domains. The size of individual structural domains varies from between about 25 up to 500 amino acids, but the majority, 90%, have less than 200 residues with an average of approximately 100 residues. The term family as it is used in taxonomy should not be confused with protein family which is a group of evolutionarily related proteins, that is: proteins in a protein family descend from a common ancestor and typically have similar three-dimensional structures, functions, and significant sequence. Note that it is also often used the term super–*, where * can stand for motif, or domain, or family, or fold, or class.

The main subject of this paper is the development of a new data structure for protein structural analysis that is derived from two basic object representation approaches, developed by the pattern recognition community in the early eighties, for recognition purposes: the Extended Gaussian Image (EGI) and the Needle Maps (NMs).

The EGI of a 3D object or shape is an orientation histogram that records the distribution of surface area with respect to surface orientation. First, each surface patch is mapped to a point on the unit Gaussian sphere according to its surface normal. The weight for each surface normal (a point on the Gaussian sphere) is the total sum of area of all the surface patches that are of that surface normal. Being a distribution with respect to surface orientation, EGI is in principle invariant to translation. Thus, in registering two 3D objects, we can ignore the translation initially and determine the rotation between the shapes by just comparing their EGIs. The EGI was introduced for applications of photometry by B. K. P. Horn [2] in the '80 and has been extended by K. Ikeuci [3-5] in the '90, and later by others [6, 7], up to the recent Enriched Complex EGI [8]. A needle map represent an object showing unit surface normals at points on the surface on a regular grid. Normals which point towards the view point are seen as dots, while tilted surface patches are represented by the unit vector pointing in the direction of steepest descent. To our knowledge the first which utilized both EGI and needle map has been the same Ikeuchi for determining the posture of objects represented by NMs [3]. More recently, an in depth analysis of the needle map representation merit of NMs can be found in [9, 10]. In detail, in the former it is said that "The needle-map is a valid representation for object recognition. In terms of dimensionality of the matching representation, it may be viewed as midway between model (3D) and appearance-based (2D) recognition. However, since a series of model needle-maps are needed for each object, it remains essentially an appearance-based technique. If we deal with unit normals, two values are sufficient to describe the direction of each normal, since the third component may be determined from the other two. Thus, matching can be performed using 2D vectors".

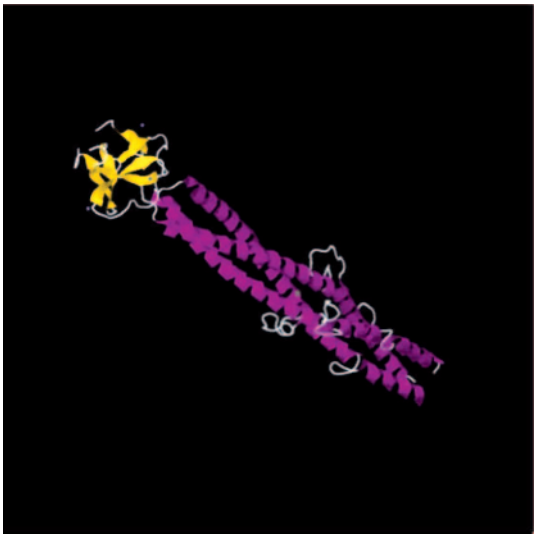

Fig. 1. – A picture generated by PyMOL on PDB file 1aa01 rotated by $\pi/2$ for format reasons.

Combining EGI and needle maps we are proposing a new data structure suitable for protein secondary structures representations and the analysis of their spatial 3D distribution. Something similar has been proposed by Yona G., and Kedem K [11], but at an higher-scale level, the protein residue backbone.

2. – The Protein Gaussian Image

There are several methods for defining protein secondary structure, but the Dictionary of Protein Secondary Structure (DSSP) [12] method is the most commonly used. The DSSP defines eight types of secondary structures, nevertheless, the majority of secondary prediction methods simplify further to the three dominant states: helix, sheet and coil. Namely, the helices include 3_{10}-helix, α-helix and π-helix; sheets or strands include extended strand (in parallel and/or antiparallel β-sheet conformation); finally, coils include hydrogen-bonded turn, bend, and amino acid residues which are not in any of the previous types. The structural analysis for protein recognition and comparison is conducted mainly on the basis of the two most frequent components [8]: the α-helices and the β-strands. There are many packages developed for segmenting the protein backbone in seconday structures (SSs). Every two years, the performance of current methods is assessed in the CASP experiment (Critical Assessment of Techniques for Protein Structure Prediction). The two on which we have based our analysis are the quoted DSSP and STRIDE [13]. These packages have been considered very reliable even if on average 4.8% of the target residues were differently assigned (this number reaching 12% for certain targets), and in this work we have considered these suitable performances.

DSSP and STRIDE both extract from the PDB segments 3D locations and attitudes, positions in the sequence of SSs and in particular also strands (constituting sheets), and many other information easily integrated the new data structure.

PGI is a representation in the Gaussian image in which each SS is mapped with a unit vector from the origin of the sphere having the orientation of the SS. Each point of the sphere surface contains the data orientation (length, location of starting and ending

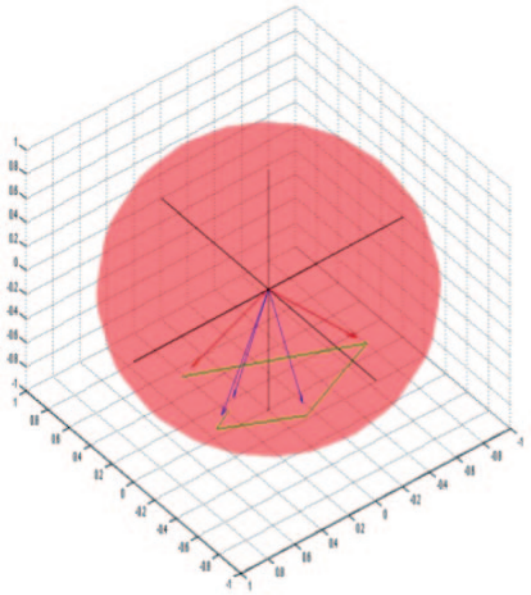

Fig. 2. – PGI of protein 1aa01.

residue, etc) of the existing protein SSs having the corresponding. The chain sequence of SS is recorded as a list which is mapped on the sphere surface (green line) as in figs. 1, 2 where protein 1aa01 is depicted with its PGI.

This 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, as for needle maps for general object representation, supports effectively the structural perception.

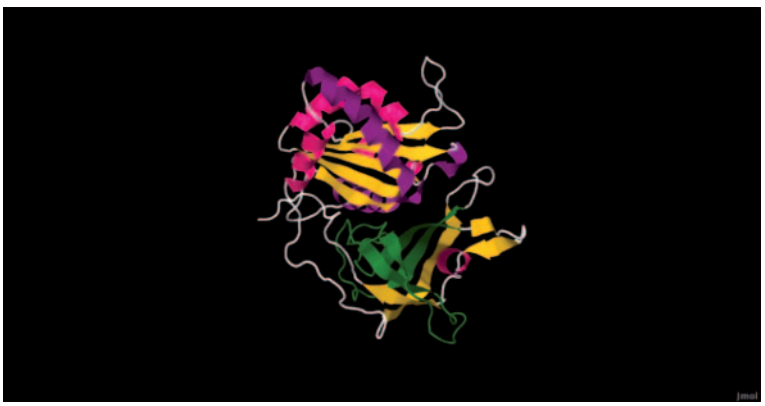

Fig. 3. – A picture generated by PyMOL on PDB file 1FNB rotated by $\pi/2$ for format reasons. In green the Greek-key motif (residues 56-116).

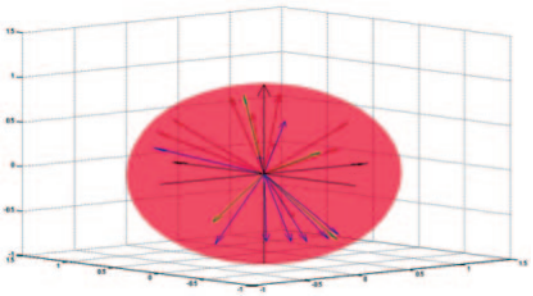

Fig. 4. – Protein Gaussian Image of protein 1FNB. Green arrows represent the Greek-key motif.

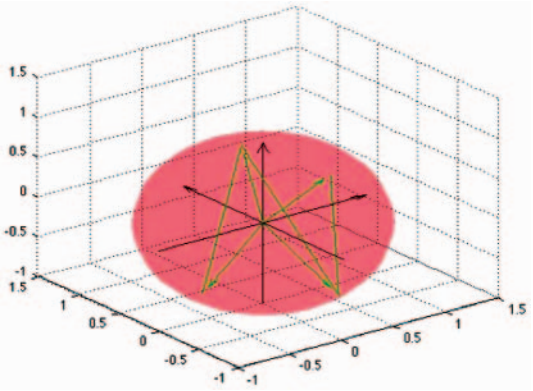

Fig. 5. – Protein Gaussian Image of Greek-key motif contained in protein 1FNB. Green arrows represent the Greek-key motif, while the green line show the sequence of SS.

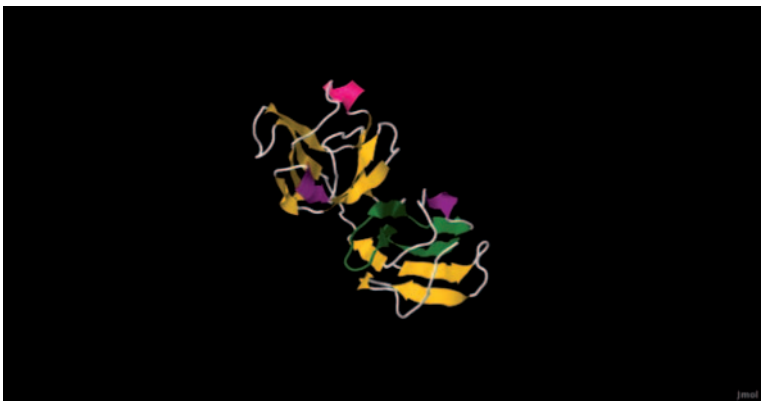

Fig. 6. – A picture generated by PyMOL on PDB file 4GCR rotated by $\pi/2$ for format reasons. In green the Greek-key motif (residues 34-62).

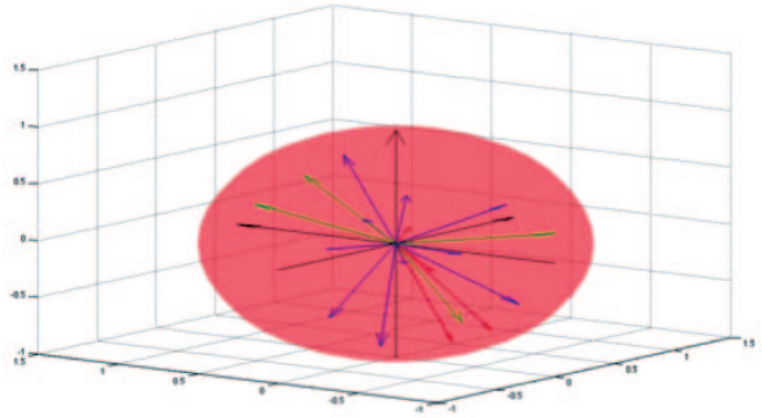

Fig. 7. – Protein Gaussian Image of protein 4GCR. Green arrows represent the Greek-key motif.

3. – Two practical examples

Two couples of motif-protein molecules are shown: 1FNB and 4GCR. For the former couple fig. 3 represents the molecule, fig. 4 shows the PGI (in green the Greek-key motif is highligthed) while fig. 5 shows the PGI of the motif with the linked sequence of SSs. In figs. 6-8 the analogous representation for the latter couple are shown.

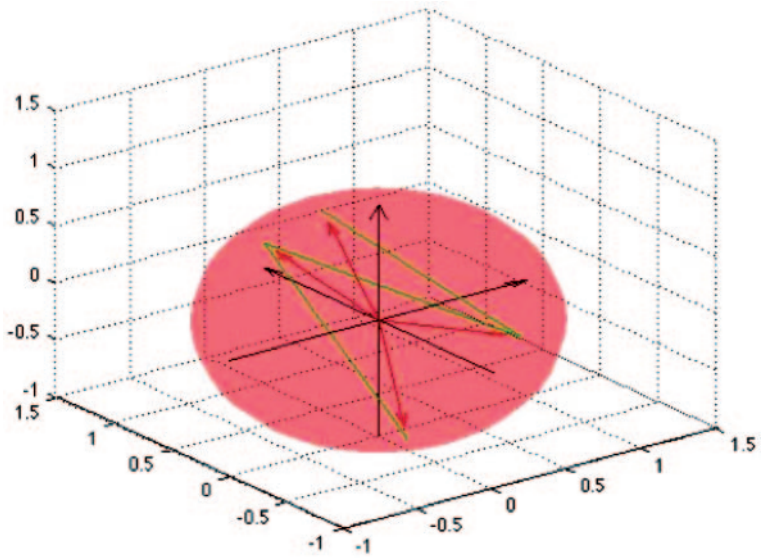

Fig. 8. – Protein Gaussian Image of the Greek-key motif contained in protein 1FNB. Green arrows represent the Greek-key motif, while the green line show the sequence of SS.

4. – Conclusion

A new data structure has been introduced that supports both artificial and human analysis of protein structure. We are now planning an intensive quantitative analysis of the effectiveness of this new representation approach for practical problems such as alignment or even of structural block retrieval at different level of complexity: from basic motifs composed of a few SSs, to domains, up to units.

REFERENCES

[1] Jacob F., *Science*, **196** (1977) 1161.
[2] Horn B. K. P., *Proc. IEEE*, **72,** 12 (1984) 1671.
[3] Ikeuchi K., *Determining the Attitude of an Object from a Needle Map using the Extended Gaussian Image*, M.I.T. A.I. Laboratory, Memo Number 714.
[4] Kang S. B. and Ikeuchi K., *Proc. IEEE TPAMI*, (1993) 707.
[5] Shum H., Hebert M. and Ikeuchi K., *Proc. IEEE-CVPR-1996*, (1996) 526.
[6] Matsuo H. and Iwata A., *3-D Object Recognition Using MEGI Model from Range Data*, in *Proceedings of the 12th International Conference on Pattern Recognition* (1994) pp. 843–846.
[7] Wang D., Zhang J., Wong H. S. and Li Y., *3D Model Retrieval Based on Multi-Shell Extended Gaussian Image*, in *VISUAL 2007, LNCS 4781* (2007) pp. 426–437.
[8] Zhaozheng H., Chung R. and Fung S. M. K., *Machine Vision and Applications*, **21** (2010) 177.
[9] Worthington P. L. and Hancock E. R., *View Synthesis from Needle-Maps*, in *15th International Conference on Pattern Recognition (ICPR'00)*, **4** (2000) pp. 110–113.
[10] Worthington P. L. and Hancock E. R., *Proc. IEEE TPAMI*, **21,** 12 (1999) 1250.
[11] Yona G. and Kedem K., *J. Comput. Biol.*, **12,** 1 (2005) 1232.
[12] Kabsch W. and Sander C., *Biopolymers*, **22** (1983) 2577.
[13] Heinig M. and Frishman D., *Nucl. Acids Res.*, **32** (2004) W500-2.