IL NUOVO CIMENTO Vol. 35 C, N. 5 Suppl. 1 Settembre-Ottobre 2012
DOI 10.1393/ncc/i2012-11338-x

Colloquia: PR PS BB 2011

A data structure for protein-ligand morphological matching

V. Cantoni, A. Gaggia and L. Lombardi

Department of Computer Engineering and Systems Science, University of Pavia - Pavia, Italy

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

Summary. — Pattern recognition techniques can be applied very profitably to proteomics because of the strong linkage of the protein's molecule morphology and proteins functionalities. In fact, geometric and topological congruence (concavity and convexity correspondences) can be often considered certainly not sufficient but in many cases necessary conditions. In this connection, considering that the "active sites" are always located in one of the biggest concavities (in one of the largest "pockets") and that the ligand must match this concavity, its effective part must be mainly convex. For this reason, the matching potential can be evaluated through an Extended Gaussian Image (EGI) shape representation. The original EGI, and a few extensions (namely Complex EGI and Enriched Complex EGI) representations and their correspondent concrete data-structures are here discussed. This data structure is then exploited for the implementation and evaluation of the matching stance between the small ligand molecule and a pocket of a protein macromolecule.

PACS 87.15.K- – Molecular interactions; membrane-protein interactions.
PACS 87.18.Xr – Proteomics.
PACS 87.85.mk – Proteomics.

1. – Preliminary statements

The methods developed for the comparison of protein surfaces aim to search sub-regions of a protein that are complementary, so as to match, to sub-regions of a second protein, as required by docking applications (for protein-ligand or protein-protein interactions). The objective is consequently to discover complementary regions (that is with concave and convex segment that match each others) between different proteins, and until now it has been pursued by *ad hoc* descriptors of patterns like spin image [1, 2], context shape [3] and harmonic shape [4], etc. A promising alternative approach, which we believe convenient to investigate, is the search for regions of interface that potentially corresponds to the active sites, through the EGI introduced for applications of photometry by B. K. P. Horn [5] in the '80 and which has been extended with a sequence of new definitions up to the Enriched Complex EGI in 2010. The EGI is the histogram of

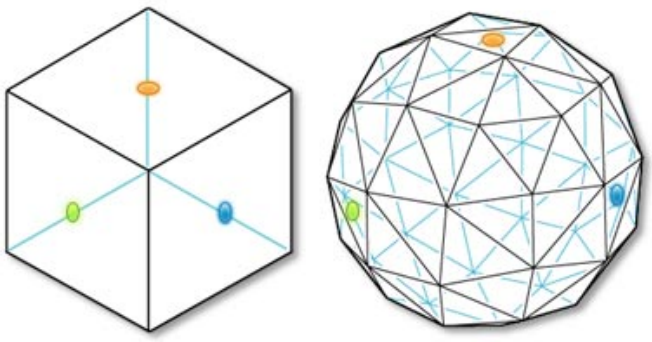

Fig. 1. – Extended Gaussian Image of a cube.

the orientations placed on the unitarian sphere and it constitutes a compact and effective representation of a 3D object as a protein or one of its part (the rotations of the object correspond to rotations of the EGI, the side surface of the object is the mass of the EGI, the barycentre of the EGI is in the origin, every hemisphere is balanced by the complementary hemisphere, etc.). Just for these important properties and for the simplicity and the operational handiness on the unitarian sphere herewith we want to show that the conditions for docking conducted on the EGI, are selective enough for screening purposes so as to avoid intensive and expensive experimentations. Moreover we think that the EGI should be a profitable approach because the hypothesis of model "basically" convex (obviously not always completely) can be applied in the search of matching between the convexities of the ligand/protein and the complement to the concavity of the second protein. However, if even the C-EGI, in presence of concavity [6], does not guarantee the biunivocity between representation and 3D object, the quoted sequence of improved solutions has been introduced just for removing the ambiguities that are introduced by concavities.

2. – From the EGI to ECEGI

A Gaussian Image is the mapping of all the normals of an object on a sphere of unitary radius (Gaussian Image): all the tails of the vectors are on the center of the sphere while the heads lie on the surface. The Extended Gaussian Image (EGI) [5] includes the associate area: each point on the sphere has a mass proportional to the area (fig. 1). The EGI is particularly suited to deal with the varying attitude of a 3D object in space: it is invariant to the object position and, in case of convex objects there is a bijective correspondence with their EGI. The basic idea is that each surface patch is mapped to a weighted point on the Gaussian sphere according to its surface normal and the value assigned to each orientation is the sum of area of all the surface patches having a common normal (fig. 1):

$$W_{\vec{d}} = \sum_{l=1}^{N_{\vec{d}}} A_{l,\vec{d}}, \tag{1}$$

where $\vec{d}$ is the direction associated with a point on the Gaussian sphere, $N_{\vec{d}}$ the total number of surface patches with normal $\vec{d}$ and $A_{l,\vec{d}}$ the area of the $l_t h$ surface patch with normal $\vec{d}$. Being a distribution with respect to surface orientation, as it has been said previously the EGI is in principle invariant to translation. Thus, in registering two 3D objects M and S, we can ignore the translation and determine the rotation between the shapes by just comparing their EGIs. Obviously $e(R)$ is the figure of merit of rotation R:

$$e(R) = \sum_{\vec{d}} \left(M_{\vec{d}} - S_{R,\vec{d}} \right)^2 . \tag{2}$$

The most important advantage of the Extended Gaussian Image is the position invariance, and its principal drawback is that the bijective correspondence is valid only in case of convex objects. To reduce the ambiguities that are generated by concave patches a new solution has been introduced adding a support function to the EGI: the signed distance of the oriented tangent plane from a predefined origin. In this new representation, called Complex EGI (CEGI) [6, 7], the weight at each patch, representing a discretized cell, is a complex number that encodes both area and distance. The magnitude of the complex number is the surface area of the object associated with that surface normal, while the phase, which supports the displacement information, is the signed distance of the surface patch from a predefined origin in the direction of its normal. The complex weight associated with surface patch A_i is $A_{i,nk}e^{jdk}$, where $A_{i,nk}$ is the area of patch A_i with the outward normal n_k, and d_k is the normal distance of the plane within which $A_{i,nk}$ lies to an assigned origin. For any given point in the CEGI corresponding to normal n_k, the magnitude of the point's weight is $\|A_{nk}e^{jdk}\|$. $A_n k$ is independent of the normal distance, and if the object is convex, the distribution of Ank corresponds to the conventional EGI representation. The following equation represents the CEGI complex weight:

$$W_{n\vec{k}} = \sum_{l=1}^{N_{\vec{k}}} A_{l,nk} e^{j d_{l,k}}. \tag{3}$$

The use of complex numbers allows the area and position information to be decoupled. Thus, in registering to determine the rotation between the shapes the comparison of their CEGIs can be obtained by

$$e(R) = \sum_{\vec{d}} \left(\left\| M_{\vec{d}} \right\| - \left\| S_{R,\vec{d}} \right\| \right). \tag{4}$$

Another improvement is given by the use of the Enriched Complex Extended Gaussian Image (ECEGI) [8]. Also in the ECEGI approach each patch of the 3D object surface contributes with a complex weight to the associated point orientation of the Gaussian sphere. However, while the CEGI uses only a scalar complex number, the ECEGI uses a vector of three complex numbers. The resultant weight is the sum of the contributions of all surface patches having the normal in common, where the exponent is given by the distance from each one from the coordinate planes. The magnitude of the ECEGI representation is translation-invariant. The ECEGI can be viewed as three independent

complex Gaussian spheres, each corresponding to the axis (x, y, z). The weight is, in this case, represented by three complex numbers given by:

$$
\begin{aligned}
W_{x,n\vec{k}} &= \sum_{i=1}^{N\vec{k}} A_{i,nk} e^{jX_{i,k}}, \\
W_{y,n\vec{k}} &= \sum_{i=1}^{N\vec{k}} A_{i,nk} e^{jY_{i,k}}, \\
W_{z,n\vec{k}} &= \sum_{i=1}^{N\vec{k}} A_{i,nk} e^{jZ_{i,k}}.
\end{aligned} \tag{5}
$$

The resultant weight at the point is then the sum of the contributions of all surface patches that are of the corresponding surface normal referred to each one of the coordinate planes. Thus, in registering to determine the rotation between the shapes the comparison of their ECEGIs can be obtained by

$$
\begin{aligned}
&e(R) = \\
&\sum_{\vec{d}} \left[\left(\left\| M_{x,\vec{d}} \right\| - \left\| S_{x,R,\vec{d}} \right\| \right)^2 + \left(\left\| M_{y,\vec{d}} \right\| - \left\| S_{y,R,\vec{d}} \right\| \right)^2 + \left(\left\| M_{z,\vec{d}} \right\| - \left\| S_{z,R,\vec{d}} \right\| \right)^2 \right].
\end{aligned} \tag{6}
$$

3. – EGI pocket and lingand representations

The aim, for docking applications, is the search for sub-regions that are complementary between different molecules. When we have a large molecule (receptor) and a small molecule (ligand), docking takes place in a protein cavity. In this connection the first sub-problem to be solved, in protein-ligand interfaces, is to develop the representations and the data structures suitable to support the computational methods that allow a quantitative evaluation of the protein-ligand matching on the basis mainly of their 3D structure and morphology. The EGI is a possible representation for the ligand molecule, with the correspondent data structure based on a first order statistic of the surface orientations. After the segmentation of the protein SES [9], the interface regions, which potentially can be active sites, are represented by an EGI. A given 3D molecule, modeled through its SES in a triangular mesh, is described by the set of triangles

$$
T = \{T_1, \ldots, T_m\}, T_l \subset R^3, \tag{7}
$$

where each T_l consists of a set of three vertices:

$$
T = \{P_{A,l}, P_{B,l}, P_{C,l}\}. \tag{8}
$$

Center, normal and area of each triangle T_l, namely g_l, d_l and A_l, respectively, can be computed by

$$
g_l = (P_{A,l} + P_{B,l} + P_{C,l})/3, \tag{9}
$$

$$
\vec{d_l} = (P_{C,l} - P_{A,l}) \times (P_{B,l} - P_{C,l}), \tag{10}
$$

$$
A_l = |(P_{C,l} - P_{A,l}) \times (P_{B,l} - P_{C,l})|/2. \tag{11}
$$

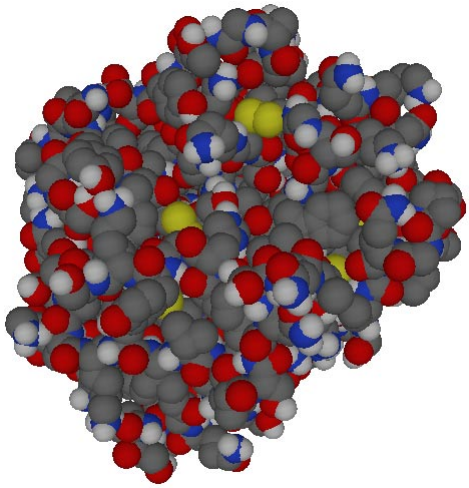

Fig. 2. – Trypsin molecole (pdbID 1TNL).

The total area of the mesh A is given by cumulating the area of each single triangle:

$$A = \sum_{l=1}^{m} A_l, \tag{12}$$

where the Gaussian sphere is partitioned into a number of cells m. A tessellated sphere with uniform and isotropic subdivision is needed. These properties are obviously satisfied by the projection of a regular polyhedron onto the sphere. Adopting the highest order regular polyhedron, the icosahedron with twenty triangular cells as a basis (but it provides a too coarse sampling of the orientations), and proceeding further by dividing iteratively each triangular cells into four smaller triangles according to the well-known geodesic dome [5] constructions, the required level of resolution can be achieved: being n the number of iterative subdivision steps, the cells number is $m = 10 \times 22n + 1$, and the area (solid angle) of the single cells is $\pi/10 \times 22n - 1$, respectively. The corresponding data structure is consequently a hierarchical one (in which each cell of one level contains, other than the specific orientation, the four pointers to cells of the subsequent level) and the searching strategy of the orientation histogram values becomes a hierarchical process. Given two candidates molecules dual parts (*i.e.* a cavity and a ligand) the aim is to find if they are morphologically (or geometrically) compatible, in other words we look for the rigid motion that could bring the protrusion into the cavity.

4. – Two practical examples

Two couples of protein-ligand molecules are shown: the first couple is the complex Trypsin (pdbID 1TNL) with a Trans-2-Phenylcyclopropylamine (pdbID TPA), the second one is the complex Camphor 5-monooxygenase (pdbID 3L62) with a Protoporphyrin IX containing Fe (pdbID HEM). For the former couple fig. 2 represents the 1TNL molecule; fig. 3 represents respectively the wireframe of the couple pocket chosen as a binding candidate and the ligand TPA properly aligned and fig. 4 the correspondent EGIs

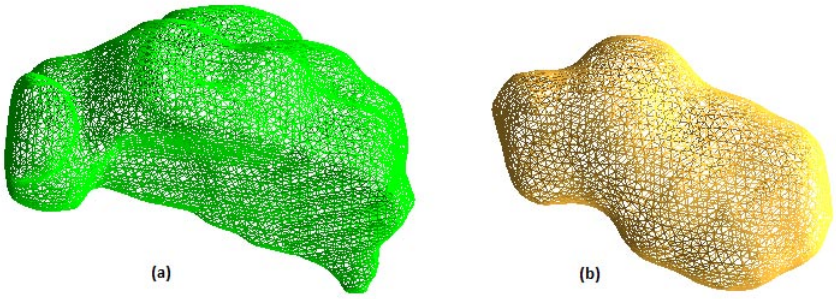

Fig. 3. – (a) Pocket candidate of 1TNL protein and (b) TPA ligand properly aligned.

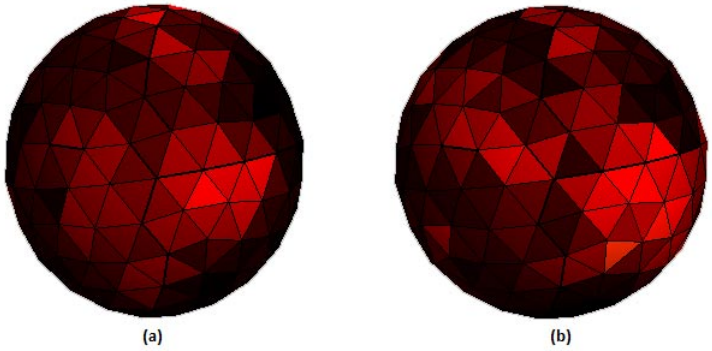

Fig. 4. – EGIs of (a) Pocket candidate of 1TNL protein and (b) TPA ligand properly aligned.

(here used instead of ECEGI for visualization purpose) of the registered couple; finally in fig. 5 the resultant molecule(s). In fig. 6 to 9 are shown the analogous representation for the latter couple.

5. – Conclusions

An effective way for computing the EGI is of great interest for the analysis of proteins convex and concave segments. This paper presents a research on the way, with a proposal of a suitable concrete data structure. We got profitable solutions but, up to now, some phases of the analysis have tested only partially; nevertheless, the new methods look not only quite efficient but also very fast. There is a drawback: variations of patterns which we often find in the joining area of proteins, change the position of local blocks in the histogram of the orientations, then obviously in the EGI; it is certainly necessary to experiment the efficiency of the proposed approach and eventually to study a particular strategy for the management of data block in the EGI to take into account of this structural flexibility.

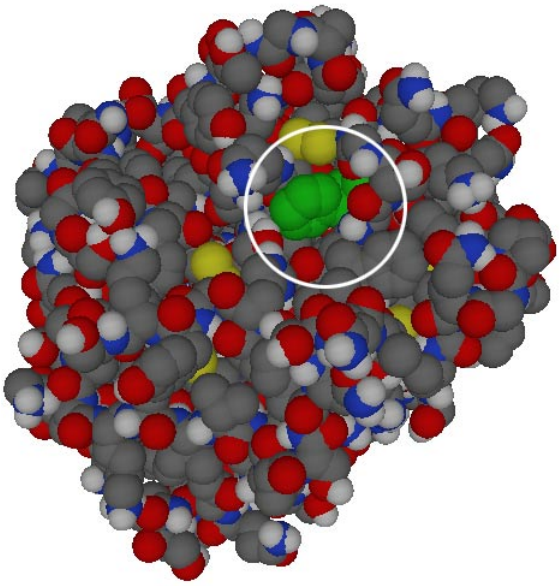

Fig. 5. – The resultant complex with the ligand surrounded by a white circle.

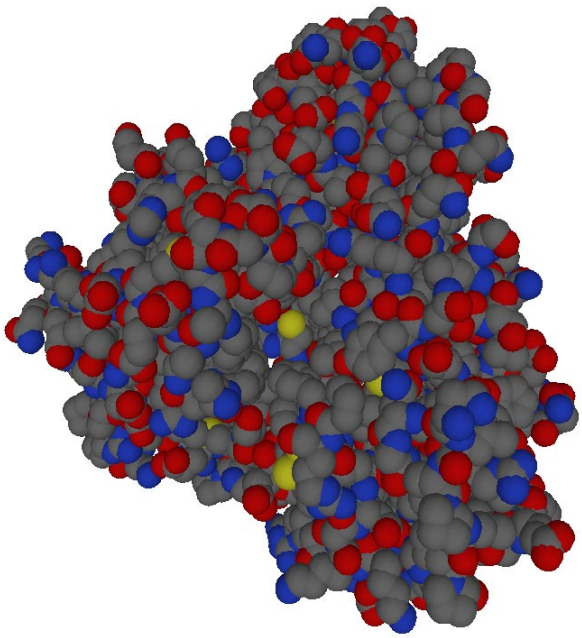

Fig. 6. – Camphor 5-monooxygenase (pdbID 3L62).

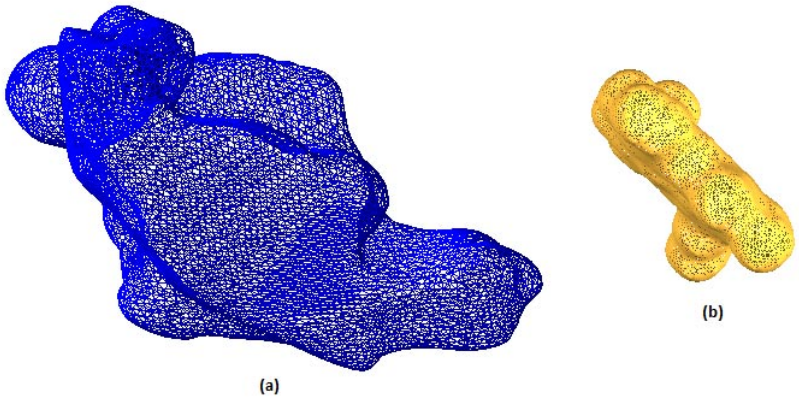

Fig. 7. – (a) Pocket candidate of 3L62 protein and (b) HEM ligand properly aligned.

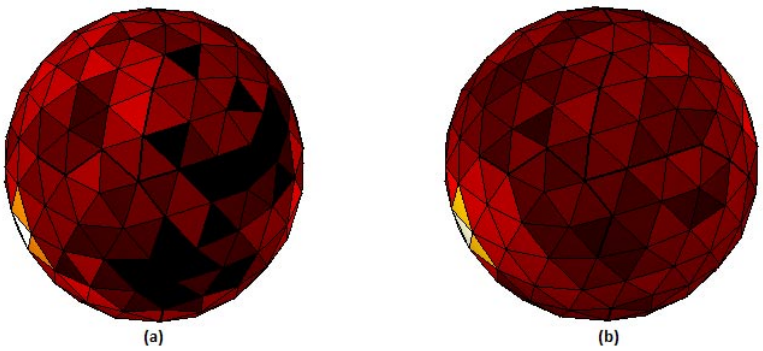

Fig. 8. – EGIs of (a) Pocket candidate of 3L62 protein and (b) HEM ligand properly aligned.

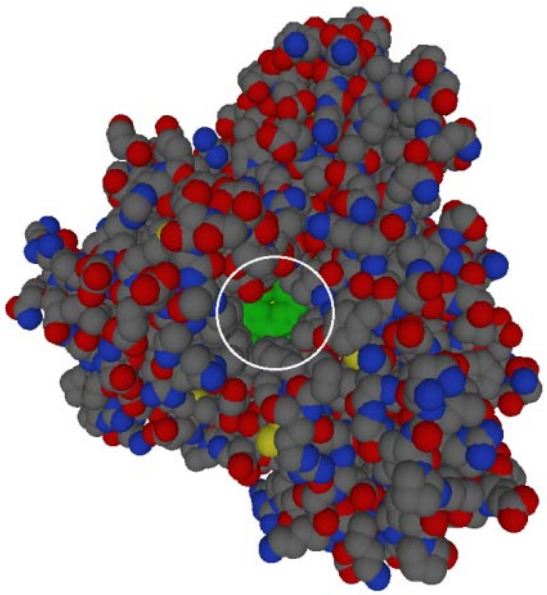

Fig. 9. – The resultant complex with the ligand surrounded by a white circle.

REFERENCES

[1] Shulman-Peleg A., Nussinov R. and Wolfson H. J., *J. Mol. Biol.*, **339** (2004) 607.
[2] Bock M. E., Garutti C. and Guerra C., *J. Comput. Biol.*, **14** (2007) 285.
[3] Frome A., Daniel H., Kolluri R., Bulow T. and Malik J., *Recognizing objects in range data using regional point descriptors*, in *Proceedings of the European Conference on Computer Vision (ECCV), May 2004*, pp. 224–237.
[4] Glaser F., Morris R. J., Najmanovich R. J., Laskowski R. A. and Thornton J. M., *Proteins*, **62** (2006) 479.
[5] Horn Berthold K. P., *Proc. IEEE*, **72** (1984) 1671.
[6] Kang S. B. and Ikeuchi K., *Determining 3-d object pose using the complex extended Gaussian image*, in *Proceedings of the 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Lahaina, Maui, Hawaii, June 1991*, pp. 580–585.
[7] Kang S. B. and Ikeuchi K., *IEEE Trans. Pattern Anal. Machine Intell.*, **15** (1993) 707.
[8] Zhaozheng H., Ronald C. and Kenneth S. M. F., *Mach. Vis. Appl.*, **21** (2010) 177.
[9] Cantoni V., Gatti R. and Lombardi L., *Segmentation of ses for protein structure analysis*, in *Proceedings of the 1st International Conference on Bioinformatics, BIOSTEC 2010*, pp. 83–89.