IL NUOVO CIMENTO Vol. 35 C, N. 5 Suppl. 1 Settembre-Ottobre 2012
DOI 10.1393/ncc/i2012-11337-y

Colloquia: PR PS BB 2011

Analysis of geometrical and topological attitude for protein-protein interaction

V. Cantoni(*), R. Gatti(**) and L. Lombardi(***)

Computer Vision Lab, Department of Computer Engineering and Systems Science
University of Pavia - Via Ferrata 1, 27100, Pavia, Italy

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

Summary. — Protein-protein interaction takes usually place on an extended area of the external molecules surfaces that are morphologically fitting. Geometric and topological congruence (*i.e.* concavity and convexity correspondences) is required to support the neighboring interaction of surface patches belonging to the two protein molecules. It is therefore important to adopt representations and data structures that can facilitate the analysis and the implementation of techniques for the evaluation of geometric and topological properties on extended surfaces. These areas of activity are usually roughly "planar" but with local concavity and complexity that must match each other for interacting. To this purpose we are suggesting a solution different from the one of ligand-protein interaction in which are involved a pocket and a small molecule. The solution here suggested is based on the concavity tree representation. Starting from the convex hull of the protein molecule a recursive process leads to a series of concavity and meta-concavity that allows reaching the detail level required. The consequence of the recursive process is obviously a hierarchical data structure (a tree) which at each level supports a complete description of a surface. Each node of the tree contains an array of features that support the geometrical, topological and biochemical properties of the correspondent surface patch.

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

(*) E-mail: virginio.cantoni@unipv.it
(**) E-mail: riccardo.gatti@unipv.it
(***) E-mail: luca.lombardi@unipv.it

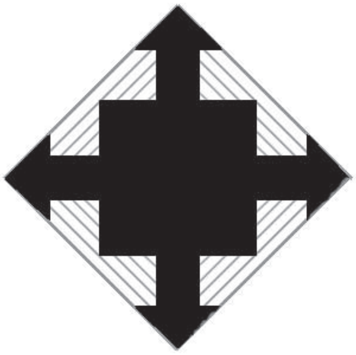

Fig. 1. – 2D travel depth example: grey lines represents pixel that share the same distance from the convex hull.

1. – Preliminary definitions

The objective is to develop methods for the forecast and the validation of the Protein-Protein Interactions (PPI), based on 3D geometrical and topological properties, by means of the mathematical morphology operators with the specific goal to inspect the regions of interface among two proteins. Adopting the Mathematical Morphology (MM) we plan also to tackle the problem of protein surface modeling.

1˙1. *MM for PPI*. – The basic idea is that the mathematical morphology (also called "Image Algebra" by S. R. Sternberg [1, 2]) developed at the Ecole Nationale Supérieure des Mines by G. F. Matheron and J. Serra [3] has not still been exploited as it should be for analyzing protein morphology. As an example considering the approach of M. J. Connolly [4, 5] for the protein surfaces modeling, this can be easily implemented with a few basic MM operators (just a "closing" operation, that is a sequence of two operations: "dilation" and "erosion"). But the address is more general: we believe that MM is the most appropriate tool for the morphological analysis of proteins and in general of biomolecules. In the literature the applications of MM in proteomics seem really sporadic [6-9]. In particular iterative or recursive applications of the basic MM operators allow to compute quantitative descriptions of the main surface features suitable for effective description and analysis. An emblematic example is the "propagation" process to cover completely a connected component and an important practical case of exploitation of this recursive process is the computation of the Distance Transform (DT) to evaluate the Travel Depth (TD).

1˙2. *DT and TD*. – In biology, the minimum distance of a point from a reference surface is called travel depth. For its evaluation a distance transform is commonly used. In biology the travel depth description has been investigated by Coleman and Sharp [10, 11]. Travel depth is the value of distance associated on each point of a molecular surface from the surface's convex hull (a 2D example is shown in fig. 1). The shape and the properties of the molecular surface determine what interactions are possible with ligands and other macromolecules. In particular, the active sites are generally described as shallow and deep spots on the molecular volume. Experimentations have shown that active sites usually corresponds with areas on the surface having a high travel depth

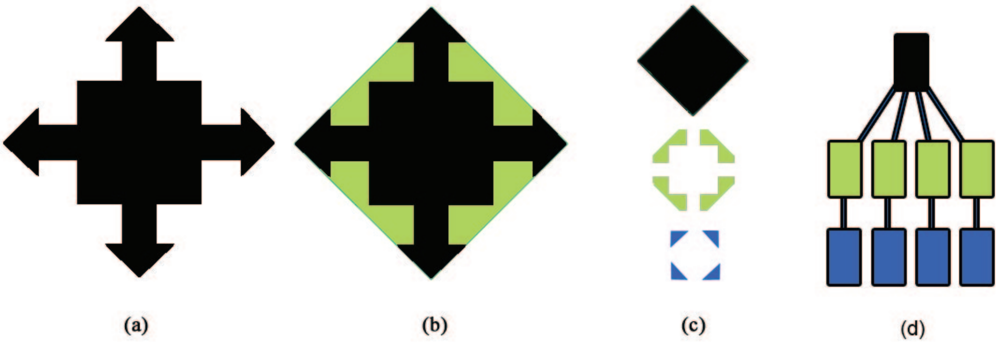

Fig. 2. – An object (a), its convex hull with green concavities (b), the actual shape of each node (c) and the corresponding three levels concavity tree (d).

value and thus they coincide with the bottom of protein's pockets. After computing some feature parameters which characterize the local surface geometry a suitable data structure to support these information and their topology must be defined: we are here proposing the Concavity Tree (CT) solution.

1˙3. *The CT data structure.* – CT has been introduced as data structure to represent image contours by [12] and further researched by Batchelor [13]. A CT is used for describing non-convex two-dimensional shapes. It is a rooted tree in which the root is a simple characterization of the whole object boundary, *i.e.* its convex hull. Each node of the next level describes the set of objects obtained by subtracting the object from the convex hull. Each next level is elaborated in the same way, iteratively. A node that represents a convex shape corresponds to a leaf in the tree, so it does not have any child.

Figure 2 shows an example of a shape (a), its convex hull, concavities, and meta-concavities (b), and its corresponding concavity tree (c). The shape generates five concavities as reflected in level one of the tree. The four leaf nodes in level one correspond to the highlighted triangular concavities shown in (d). Typically, each node in a concavity tree stores information pertinent to the part of the object the node is describing (a feature array for example), in addition to tree meta-data (like the level of the node; height, number of nodes, and number of leaves in the subtree rooted at the node, etc.). In the sequel we firstly introduce the analysis of protein surfaces for the detection of geometrical features and then we will present the CT representation as a mean for effective description, analysis and detection of protein active sites.

2. – Search for pockets and tunnels

In the discrete space the protein and the CH are defined in a cubic grid V of dimension $L \times M \times N$ voxels. Note that the grid is extended one voxel beyond the minimum and maximum coordinate of the SES in each orthogonal direction (in this way both SES and CH borders are inside the V border). The voxel resolution adopted is 0.25 Å, so as to be small enough to ensure that, with the used radii in biomolecules atoms, any concave depression or convex protrusion is represented by at least one voxel.

Let us call R the region between the CH and the SES (the concavity volume [14]), that is

$$R = CH \cap \overline{SES}. \tag{1}$$

Let us call B_{CH} the set of the border voxels of CH, that is

$$B_{CH} = CH - [CH \bullet K]\,, \tag{2}$$

where $\bullet$ represent the erosion operator of mathematical morphology and K the discrete unitarian sphere (in the discrete space a $3 \times 3 \times 3$ cube).

Within the region R the following propagation is applied:

```
$D_i = \begin{Bmatrix} 1 \text{ if } i \in B_{CH} \\ 0 \text{ otherwise} \end{Bmatrix}$
$A = B_{CH};$
$N = (A \oplus K) \cap R;$
$E = N - A;$
while $E \neq \oslash$ $do$
    $\forall e \in E : d_e = \min_{n \in n_e}(d_n + w_n);$
    $A = N;$
    $N = (A \oplus K) \cap R;$
    $E = N - A;$
done
```

where:

- A represents the increasing set of voxels contained in R;
- E corresponds to the recruited set of near neighbors of A contained in R (*i.e.* the voxels reached by the last propagation step);
- $\min_{n \in n_e}(d_n + w_n)$; represents the minimum value among the distances d_e in the near neighbors belonging to D already defined, incremented by the displacement w_j between the locations (e, n).

The values in D represent the distance of each voxel of A from the border of B_{CH} and A corresponds to the connected component of R adjacent to the border. In order to separate the different pockets and tunnels the volume A must be partitioned into a set of disjoint segments $P_{SES} = \{P_1, \ldots, P_j, \ldots, P_N\}$, where N is the number of inlets. The partition must satisfy the following conditions:

$$P_i \cap P_j = \oslash, i \neq j, \tag{3}$$

$$P_1 \cup \ldots \cup P_j \cup \ldots \cup P_N = A. \tag{4}$$

The effective number of segments, that determines obviously the number and the morphology of pockets and tunnels, is found out on the basis of two heuristic parameters: i) the minimum travel depth value of the local tops TD_LT; ii) an evaluation of near neighbor pivoting effects PE_s. In fig. 3 the results of the segmentation process produced by our package on MTSP1 (MATRIPTASE) (PDB ID 1EAX). This process is executed in two steps: first an onward propagation, where the set of the pocket local tops is identified; and then a backward parallel propagation from each of the tops, which identifies all the pockets. A detailed description of the segmentation process is given in [15].

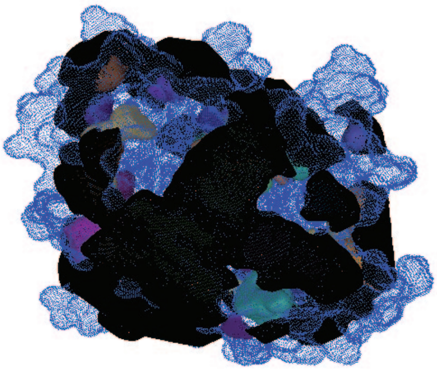

Fig. 3. – Pockets and inlets found after segmentation process on PDB ID 1EAX.

3. – Protein surface representation through CTs

The complex shape of the protein has been segmented into its components (*e.g.* a pockets set), and each pocket can be subsequently decomposed into simpler regions, and the complete description is given in terms of the region's features and their spatial relationship.

This process can be executed recursively. In this way a sequence of approximations is built, and, at each stage, this hierarchical structure can be effectively applied for analyzing and comparing complex shapes. A refinement of the analysis of these regions is performed to extract further details. The Convex Hull of each region at every level of detail is analyzed by the same process as that applied to the whole initial shape, going through concavities and meta-concavities. The process continues until all regions of the last level are convex, or a level of detail sufficiently fine is reached.

The final result is a hierarchical structure, the (meta) concavity tree. At each level the concavities can be analyzed and described on the basis of the feature vector computed at each node: obviously the features defined for concavities can also be computed for the meta-concavities.

The problem of defining an optimal set for feature selection is complicated because besides building robust models it is also important to simplify the amount of resources required to describe the data accurately, without ambiguity, in a very large set of redundant and relevant information. The expert can help, but can usually construct only a set of application-dependent features.

A rich set of general features that can be, in peculiar cases, partitioned in well-organized proficient subsets, is the following: i) Pocket Volume [16], ii) Surface-to-Volume Ratio, iii) Skewness and Kurtosis of Height Distribution, iv) Mouth Aperture (in details we consider area, perimeter and the perimeter-to-area ratio, Travel Depth [10,17], v) Top Peaks and Valleys, vi) Summit Density, Mean Summit Curvatures (both the average of the principal curvatures of peaks and valleys [18, 19]), vii) Interfacial Area Ratio, and viii) Residue Conservation [20] (the conservation score for each residue in a given protein can be obtained from the ConSurf-HSSP database [21]).

This approach results particularly fruitful in proteomics in which the morphology plays a fundamental role for studying protein-protein interaction. The solvent excluded

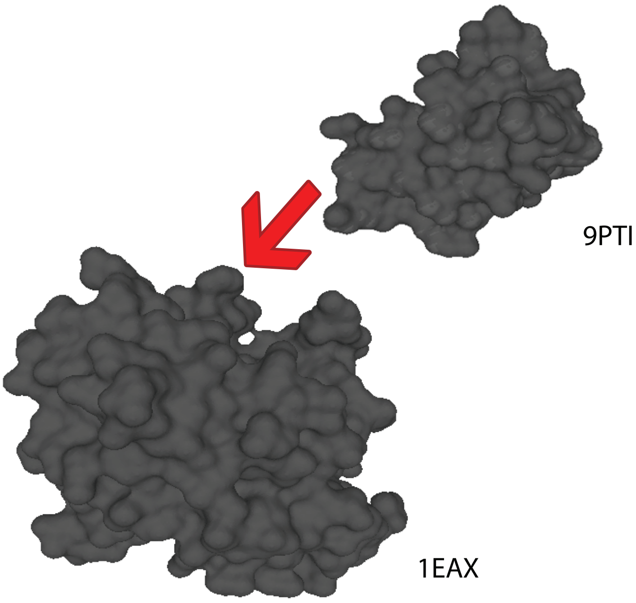

Fig. 4. – Example of two interacting proteins: MTSP1 (MATRIPTASE) (PDB ID 1EAX) and basic pancreatic trypsin inhibitor (PDB ID 9PTI).

surfaces of a couple of interacting protein molecules (MTSP1 (MATRIPTASE) (PDB ID 1EAX) and basic pancreatic trypsin inhibitor (PDB ID 9PTI)) are shown in fig. 4. The main binding site that is involved in the docking working as a ”receptor”, detected with our ”protein inspector” software package is shown in fig. 5. In figs. 6 and 7 the useful subsets of the concavity trees representing the intreacting parts of PDB ID 1EAX and PDB ID 9PTI are shown.

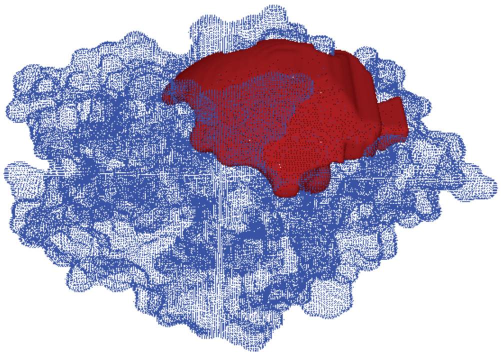

Fig. 5. – The binding site of PDB ID 1EAX (note that it is its main pocket) involved in the interaction with PDB ID 9PTI.

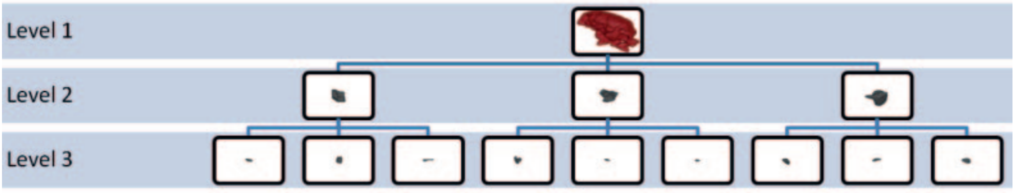

Fig. 6. – The CT hierarchical representation of the binding site of PDB ID 1EAX.

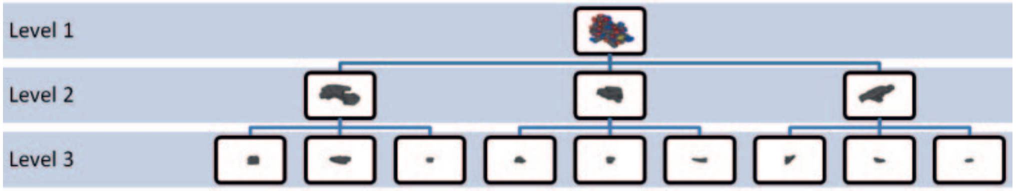

Fig. 7. – The CT hierarchical representation of the parts of PDB ID 9PTI interacting with the binding site of PDB ID 1EAX represented in fig. 6.

4. – Conclusions

A new data structure has been introduced that supports the surface analysis for matching and protein-protein interaction. We are now planning an intensive quantitati ve analysis of the effectiveness of this new representation approach for practical problems and to evaluate the performances on established benchmarks such as CAPRI [22].

REFERENCES

[1] Sternberg S. R., "Language and architecture for parallel image processing", in *Proceedings of the Conference on Pattern Recognition in Practice*, 1980.
[2] Sternberg S. R., *Overview of image algebra and related issues* (Academic Press) 1985.
[3] Serra J., *Image Analysis and Mathematical Morphology* (Academic Press) 1988.
[4] Connoly M. L., *Science*, **221** (1983) 709.
[5] Connoly M. L., *J. Appl. Cryst.*, **16** (1983) 548.
[6] Masuya M., *Shape Analysis of Protein Molecule and Legand-Recepter Docking Studies Using Mathematical Morphology* (The University of Tokyo) 1996.
[7] Takeshi K., Multi-scale pocket detection on protein surface using 3d image processing technique, pp. 49–56 (2006) IPSJ SIG Technical Reports.
[8] Chang H. T., Liu C. H., Fan T. C., Da M., Chang T. and Pai T. W., Estimation and extraction of predictive linear epitopes by mathematical morphology approaches, in *Proceedings of The Sixth Asia Pacific Bioinformatics Conference*, 2008.
[9] Prisant M., "Ray representation formalism for geometric computations on protein solid models", in *Applied Computational Geometry Towards Geometric Engineering*, edited by Ming Lin and Dinesh Manocha, *Lect. Notes Comp. Sci.*, Vol. **1148** (Springer Berlin, Heidelberg) 1996, pp. 79–90.
[10] Coleman R. G. and Sharp K. A., *J. Mol. Biol.*, **362** (2006) 441.
[11] Coleman Ryan G. and Sharp Kim A., *J. Chem. Inf. Model.*, **50** (2010) 589.
[12] Sklansky J., *IEEE Trans. Comput.*, **21** (1972) 1355.
[13] Batchelor B. G., *IEE J. Comput. Digital Tech.*, **2** (1978) 157.

[14] Borgefors G. and Sanniti Di Baja G., *Comput. Vision Image Understand.*, **63** (1996) 145.

[15] Cantoni V., Gatti R. and Lombardi L., "Segmentation of ses for protein structure analysis", in *Proceedings of the 1st International Conference on Bioinformatics* (2010) pp. 83–89.

[16] Laskowski R., Luscombe N. M., Swindells M. B. and Thornton J. M., *Protein Sci.t*, **2438** (1996) .

[17] Giard J., Rondao-Alface P. and Macq B., "Fast and accurate travel depth estimation for protein active site prediction", in *Image Processing: Algorithms and Systems*, San Diego, CA, 2008.

[18] Coleman R. G., Burr M. A., Souvaine D. L. and Cheng A. C., *Proteins*, 2005.

[19] Cantoni V., Gatti R. and Lombardi L., "Towards protein interaction analysis through surface labeling" in *Proceedings of the 15th International Conference of Image Analysis and Processing* (2009) pp. 604–612.

[20] Glaser F., Morris R. J., Najmanovich R. J., Laskowski R. A. and Thornton J. M., *Proteins*, **62** (2006) 479.

[21] Glaser Fabian, Rosenberg Yossi, Kessel Amit, Pupko Tal and Ben-tal Nir, *Proteins*, **58** (2005) 610.

[22] Janin J., Henrick K., Moult J., Eyck L. T., Sternberg M. J., Vajda S., Vakser I. and Wodak S. J., *Proteins*, **1** (2003) 2.