International Journal of Pharmacology and Pharmaceutical Technology

Volume 1 | Issue 1 Article 4

January 2013

Surface To Surface Map Algorithm For Protein - Small Molecule Matching

Neha Gupta
Information Technology, Delhi College of Engineering, Delhi, India, neha.dce04@gmail.com

Megha Bajaj
Institute of Molecular Bioscience, The University of Queensland, Brisbane, Australia, micro.megha@gmail.com

Recommended Citation

Gupta, Neha and Bajaj, Megha (2013) "Surface To Surface Map Algorithm For Protein - Small Molecule Matching," *International Journal of Pharmacology and Pharmaceutical Technology*: Vol. 1 : Iss. 1 , Article 4.
DOI: 10.47893/IJPPT.2013.1001
Available at: https://www.interscience.in/ijppt/vol1/iss1/4

Surface To Surface Map Algorithm For Protein - Small Molecule Matching

Neha Gupta & Megha Bajaj
Information Technology, Delhi College of Engineering, Delhi, India
Institute of Molecular Bioscience, The University of Queensland, Brisbane, Australia
Email: neha.dce04@gmail.com, micro.megha@gmail.com

Abstract **-** Current methods for protein analysis are based on either sequence similarity or comparison of overall tertiary structure. These conserved primary sequences or 3-dimensional structures may imply similar functional characteristics. However, substrate or ligand binding sites usually reside on or near protein surface, so, similarly shaped surface regions could imply similar functions. Our current work includes development of an algorithm that would allow surface matching over specific regions on related proteins with an output equal to the match percentage between two proteins. Initial results indicate that we can successfully match a family of related active sites, and find their similarly shaped surface regions. This method of surface analysis could be extended to help us understand functional surface relationship between the proteins within which there is no relationship in sequence or overall structure.

Keywords- *Connolly Surface; K-mean clustering; Surface features;, Surface properties(shape index and radius of curvature)*

I. INTRODUCTION

Comparison of protein sequences and overall tertiary structures has added enormously to our understanding of structural, functional and evolutionary relationships between proteins. However, ligand binding sites usually occur on or near protein surface, so, similarly shaped surface regions could imply similar functions. Hence, comparing protein surfaces has power to reveal further functional relationships between proteins which might not be apparent from comparison of overall 3D structure. Various methods[1][2][3][4] have been developed describing shape properties of protein surfaces but they are not flexible enough to be able to recognize small variations caused to the protein surface due to conformational changes associated with mechanisms such as induced-fit. So, there is a need of an algorithm which estimates proteins surface similarity accurately, further contributing to our understanding of structural and functional relationships between proteins and, hence, become a powerful tool for prediction of function from structure.

The aim of the author was to develop an algorithm that would allow surface matching over specific regions on related proteins with an output equal to the match percentage between two proteins. For this, the author has implemented an algorithm which goes through various complex steps involving large number of mathematical calculations. In this paper, the steps involved and the results are listed.

II. STEPS INVOLVED

The algorithm used is four-phased executed sequentially i.e. output of one step is the input for the proceeding step. The four steps are:

1) To calculate the Connolly surface[5] of the 3D protein or the ligand binding site.

2) To compute the surface properties which include shape index and radius of curvature for each vertex on the Conolly Surface.

3) Select surface features on the basis of parent residues(the residues part of the surface to be compared)

4) To construct graphs for two proteins surfaces and compare them with each other. Aligning is done at the same time and the match percentage is calculated.

III. DEFINITIONS

A few definitions important for understanding the content of paper have been listed below:

A. Connolly Surface :

The solvent accessible surface(SAS) is the surface of protein which is in direct contact with the solvent. It is calculated by rolling a solvent ball over the protein and tracing the path which the center of the ball travels to form the SAS. In general, solvent accessible surface has many sharp crevices and sharp corners. In hope of obtaining a smoother surface, one can take the surface swept out by the front instead of the center of the solvent ball. This surface is the molecular surface (MS model), which is often called the Connolly's surface after Michael Connolly who developed the 1st algorithm for computing molecular surface [6].

B. Principal Curvatures(K_{max} and K_{min}):

In differential geometry, the two principal curvatures at a given point of a surface are the eigenvalues of the shape operator at the point. They measure how the surface bends by different amounts in different directions at that point.

C. Radius of Curvature :

A positive number, c , to specify the amount, or 'intensity' of the surface curvature. It is defined as the curvedness as the distance from the origin in the **(**K1,K2)- plane. 'The scaling is such that the curvedness equals the absolute value of the reciprocal radius in the case of sphere. The curvedness-is inversely proportional with the size of the object. Whereas the shape index scale is quite independent of the choice of a unit of length, the curvedness scale is not. Curvedness has the dimension of reciprocal length.'[16] It is given by this formula :

$$C = \frac{\left(k_{max}^2 + k_{min}^2\right)^{1/2}}{2}$$

where k_{max} and k_{min} are principal curvatures.

D. Shape Index :

It is a number ranging from -1 to 1 and is scale invariant. The shape index captures the intuitive notion of 'local shape' particularly well. It is given by this formula :

$$S = -2 \arctan \frac{k_{max} + k_{min}}{k_{max} - k_{min}}$$

where k_{max} and k_{min} are principal curvatures

and faces formed by joining three edges.

IV. DETAILED EXPLANATION OF STEPS

STEP 1: Generating the Connolly Surface

To compute the Connolly surface, an input file including a list of the atom centres in 3D space and radius of each defining the complete 3D structure of a protein is needed. For example, a pdb file has a list of atom centres and their radii which define the corresponding protein. The algorithm began by generating a Connolly surface for ligand binding site or on entire protein using MSMS program[7] which triangulates the Connolly surface and gives an output listing all the vertices of the triangulated surface.

.

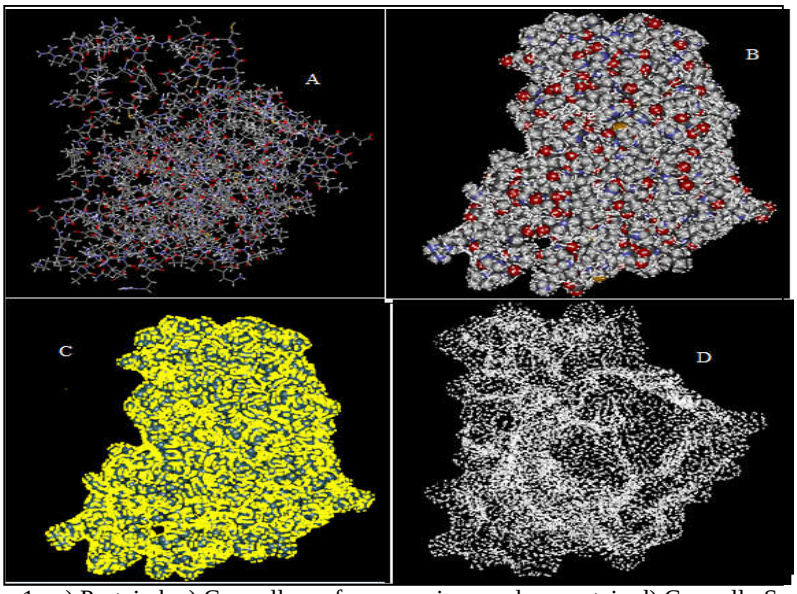

Figure 1 : a) Protein b,c) Connolly surface superimposed on protein d) Connolly Surface

. The radius of the probe can also be changed according to the precision required while running the executable of MSMS program. The calculation performed by the author was done by taking probe radius 1.4 Angstrom.

STEP 2: To compute the surface properties which include shape index and radius of curvature for each vertex on the Conolly Surface

The algorithm proceeds by calculating the surface shape properties(shape index and radius of curvature) at each of the vertices on the Connolly surface(calculated in the previous step) . This step is accomplished by using the GTS-GNU Triangulated Surface library[15] which calculates the principal curvatures at a surface patch , k_{min} and k_{max}[16]. Shape index (S) and Radius of curvature(R) in terms of principal curvatures is given by the following formula :

$$S = -2 \arctan \frac{k_{max} + k_{min}}{k_{max} - k_{min}} \text{ and}$$

$$R = \frac{\left(k_{max}^2 + k_{min}^2\right)^{1/2}}{2}$$

‘‘The shape index varies from −1 to 1 and describes the local shape at a surface point independent of the scale of the surface. A convex surface point with equal principal curvatures has a shape index of 1. A concave surface point with equal principal curvatures has a shape index of −1. A saddle surface point with principal curvatures of equal magnitude and opposite signs has a shape index of 0’’ (Duncan and Olson, 1993b).

Sample data obtained in this step is listed in Table 1.

STEP 3 : Select surface features on the basis of parent residues(the residues part of the surface to be compared)

In order to match the two surfaces, k-mean clustering [10] was performed where each chosen surface feature corresponds to a cluster. To decide whether a triangulated vertex should be added to a cluster or not, constraints were applied on shape index (S) and radius of curvature (R) after adding the new vertex. The constraints applied were as follows:

1. S-variance of cluster size(x) is less than 0.0453*pow(x, 0.0528)
2. R-variance of cluster size (x) is less than 0.1 if x <200, else less than 0.15
3. All the points in a cluster should be connected as per edge-criterion.

Also, the author introduced the residues restriction: User can give list of residues which are important. Now when the surface points are going to KMean clustering program, only the points which belong to the marked residues would be considered. The number of clusters can be decided at the run time to make sure atleast one cluster is assigned to each of the important residues.

If the constraints were not met, newly added vertex was removed from the cluster. In this step, clusters equal to the number of chosen features were obtained each having a 3D location, corresponding to the 3D location of the centre (mean) of the cluster. Each cluster also has an S and R value associated with it.

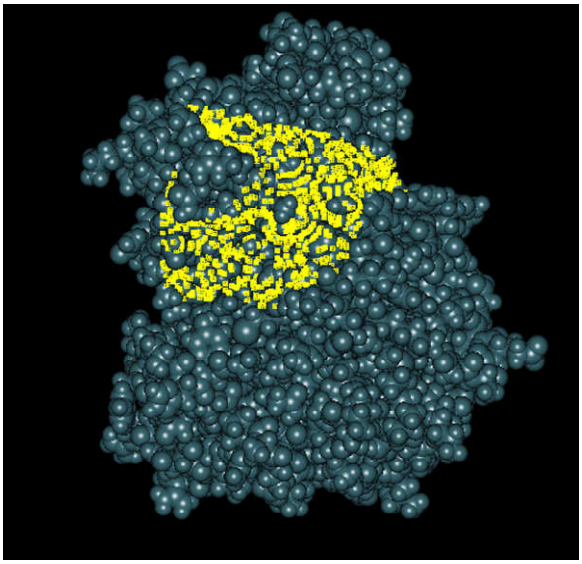

Figure 2 : Ligand binding Site for comparison

STEP 4 : To construct graphs for two proteins surfaces and compare them with each other. Aligning is done at the same time and the match percentage is calculated.

A mathematical graph with nodes as the cluster centres marked with properties S and R was constructed [14]. Edges comprised of an all-to-all joining of nodes with separation equal to the 3d distance in space. For both the proteins, 2 such graphs were formed and the maximal common sub-graph indicated a measure of the similarity between the two proteins. In this case, a match requires distances and shape properties to match within user-defined tolerances (typically 1.5 A_ for distances).

Greater tolerances allow matches between more distantly related surfaces and account for surface flexibility. Lesser tolerances will have more accuracy, but at the same time over-precision lead to anomalous results.

TABLE I : SAMPLE VERTEX COORDINATES AND SURFACE PROPERTIES

X-coordinate	Y-coordinate	Z-coordinate	Shape index	Radius of curvature
43.1600	57.2280	26.2730	-0.3100	0.2953
43.1630	57.2150	26.2790	-0.4363	0.2875
44.2170	57.1240	25.5320	-0.7292	0.3530
22.5130	48.6460	29.6830	-1.0000	0.2071
21.4070	48.2750	30.3250	-0.3745	0.0848

The product graph[8] was first calculated which was then used to find the maximal common subgraph using the Bron and Kerbosch algorithm [9] to detect cliques. When a match was discovered by the method above, the surfaces were superimposed in 3D to align the two surfaces. This was done using matrix algebra [11][12][13] to obtain the best rotation to relate the two sets of points.

RESULT: The proposed algorithm is a comprehensive approach to match selected parts or regions of two protein surfaces. This algorithm has been implemented and results found were in coherence with the expected results. Specific ligand binding sites or patches on protein surfaces can be successfully compared using the proposed algorithm.

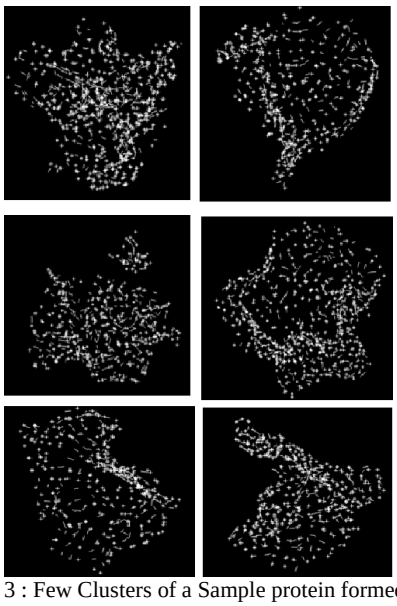

Figure 3 : Few Clusters of a Sample protein formed after applying K-mean clustering algorithm and applying constraints mentioned

The match percentage calculated can be used to identify functionally related surfaces in much more distantly related proteins.

FUTURE SCOPE: Further scope includes testing of proteins in which there is a functional surface relationship but no relationship in sequence or overall structure, and different types of functional sites. Also, chemical descriptors to the surface, including charge, hydrophobicity, and residue/atom identity can be added to the algorithm which will increase the power of our method in describing functional features of proteins.

REFERENCES

[1] Orengo, C.A., Taylor, W.R., 1996. Methods Enzymol. 266,617–635.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Holm, L., Sander, C., 1993. J. Mol. Biol. 233, 123–138.

[3] Alexandrov, N.N., 1996. Protein Eng. 9, 727–732.

[4] Michel F. Sanner,Arthur J. Olson,Jean-Claude Spehner, Reduced Surface: An Efficient Way to Compute Molecular Surfaces

[5] Koch, I., Lenauer, T., Wanke, E., 1996. J. Comp. Biol. 3,289–306.Electronic Publication: Digital Object Identifiers (DOIs): Article in a journal:

[6] Bron, C., Kerbosch, J., 1971. Commun. ACM 16 (9), 1973.

[7] David M. Mount, KMlocal: A Testbed for *k*-means Clustering Algorithms

[8] Russell, R.B., Barton, G.J., 1992. Proteins 14, 309–323.

[9] McLachlan, A.D., 1972. Acta Crystallogr., Sect. A 28, 656.

[10] Kabsch, W., 1978. Acta Crystallogr., Sect. A 34, 827–828.

[11] Harary, F., 1967. Graph Theory. Addison–Wesley, London

[12] http://gts.sourceforge.net/

[13] Jan J Koenderink and Andrea J van Doorn, Surface shape and curvature scales

◈◈◈