Structure-Based Analysis of Protein Binding Pockets Using Von Neumann Entropy

Negin Forouzesh, Mohammad Reza Kazemi, and Ali Mohades

Laboratory of Algorithm and Computational Geometry, Department of Mathematics
and Computer Science, Amirkabir University of Technology, Tehran, Iran
{n.forouzesh,mr.kazemi,mohades}@aut.ac.ir

Abstract. Protein binding sites are regions where interactions between a protein and ligand take place. Identification of binding sites is a functional issue especially in structure-based drug design. This paper aims to present a novel feature of protein binding pockets based on the complexity of corresponding weighted Delaunay triangulation. The results demonstrate that candidate binding pockets obtain less relative Von Neumann entropy which means more random scattering of voids inside them.

Keywords: Protein binding site, Delaunay triangulation, Von Neumann entropy.

1 Introduction

Proteins mainly accomplish their biological activities in interaction with other molecules. Meanwhile these interactions, only some of surface atoms of the protein get involved. Thus identifying these regions, binding sites, help scientists to study the mechanism of interactions and protein performance very well. Moreover, identification of binding sites is known as the basis of structure-based drug design [1,2].

In recent years, various computational methods have been introduced and developed for the purpose of finding protein pockets which results in predicting protein binding sites. Generally these methods can be classified into two types: energy-based and geometry-based methods. Geometry-based algorithms themselves are classified to grid-based [3,4,5], sphere-based [6,7] and alpha-shape based [8,9] types. Besides, consensus methods [10,11] have been proposed recently in which some previous pocket detecting methods are combined together in order to improve the prediction success rate entirely.

Usually many pockets are found for each protein and definitely not all of them can be regarded as real binding sites. Therefore, it is necessary to evaluate those pockets according to a ranking method and report the top-rated cases. Pocket size, the number of atoms forming the pocket, is a widely used ranking criterion; however, past studies [12,4] elucidated that some real binding sites are disregarded when protein pockets are evaluated only by their size, especially when top ranked candidates have tiny diversity in size.

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, pp. 301–309, 2014.

Comprehensive studies have been performed recently to extract novel features of protein binding pockets beside their size to improve the past prediction results. In [4] the degree of the conservation of involved surface residues in pockets was studied to report TOP1 pocket among TOP3 largest cases. In addition to pocket size, distance from the protein centroid, sequence conservation and the number of hydrophobic residues are chosen as the ranking criteria both in combination with each other and solely in [13].

The most substantial issue is that although shape of pockets and related features have been examined in previous works, up to our knowledge, the arrangement of atoms consisting the protein pockets have not been considerably discussed before. The main contribution of this paper is examining the complexity of protein binding pockets in comparison with other pockets. This leads to achieve a novel feature of binding sites which finally help us to predict them. To accomplish that, in this study we make use of weighted Delaunay triangulation of protein atoms which results in a geometric graph sensitive to the location of each atom. Afterward, the complexity of those graphs are analyzed by a useful complexity measure, Von Neumann entropy. Results show that candidate binding pockets usually obtain less relative Von Neumann entropy and consequently more disorderliness in their structure.

2 Preliminaries

In this section, basic concepts that are necessary for next part are introduced. First, some computational geometry tools both for bare and weighted points are reviewed. Secondly, matrix representation of graphs and related definitions in linear algebra are discussed.

Given a set of finite points $P = \{p_1, p_2, \ldots, p_n\}$ in the space, called sites, the Voronoi diagram is the set of cells, V_i, $1 \leq i \leq n$, defined by:

$$V_i = \{p| \ |p - p_i| \ \leq \ |p - p_j| \ , \ 1 \leq j \ \leq n\}\,.$$

In other word, Voronoi diagram is a subdivision of the space into n cells. Each cell in this diagram corresponds to a site in P, under the condition that all points in cell V_i are closer to their corresponding site p_i rather than any other sites. If the sites lie in general position meaning that no three sites on a line, no four sites on a circle and no five sites on a sphere, then the dual graph of the Voronoi diagram results in a unique geometric graph called Delaunay triangulation in which sites are considered as vertices and edges are drawn between any two vertices whose corresponding cells are adjacent.

Let $P^w = \{p_1^w, p_2^w, \ldots, p_n^w\}$ be a set of weighted points where point p_i^w can be denoted as a spherical ball $b_i = b(z_i, r_i)$with center $z_i \in \mathbb{R}^3$and radius r_i . The distance of a point $x \in \mathbb{R}^3$ and a ball $b = b(z, r)$ is formulated as:

$$\pi_b\left(x\right) = \left|z - x\right|^2 - r^2.$$

Now the weighted Voronoi diagram (Power diagram) is defined by:

$$V_{b_i}\left(p\right) = \left\{p \in \mathbb{R}^3 \mid \pi_{b_i}\left(p\right) \leq \pi_{b_j}\left(p\right),\ 1 \leq j \ \leq n\right\}.$$

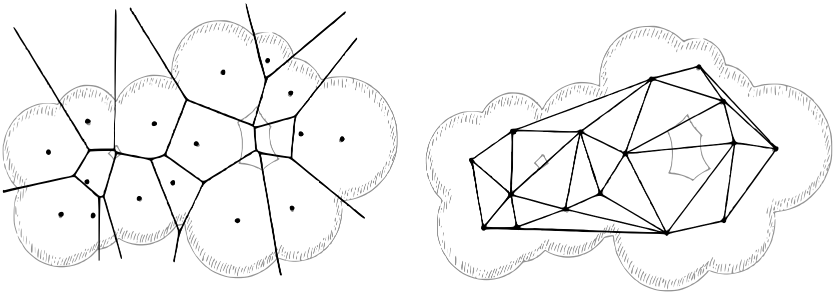

Fig. 1. Left: Weighted Voronoi diagram for a set of balls. Right: Weighted Delaunay triangulation for the same set of balls.

Such as the previous case, weighted Delaunay triangulation (Regular triangulation) can be obtained by the dual shape of weighted Voronoi diagram (see Fig 1).

Let $G(E, V)$ be a simple graph with n vertices and m edges. Adjacency matrix $A(G)$ is an $n \times n$ matrix in which $A_{uv} = A_{vu} = 1$ if two vertices u and v are adjacent and $A_{uv} = A_{uv} = 0$ otherwise. Degree matrix $D(G)$ is an $n \times n$ matrix in which diagonal element D_{uu} resembles the degree of vertex u and other elements are zero. Laplacian matrix $L(G)$ is defined as:

$$L\left(G\right) = D\left(G\right) - A\left(G\right).$$

Density matrix (G) is a normalized form of Laplacian matrix and is defined by the following relation:

$$\rho\left(G\right) = \frac{1}{tr\left[L\left(G\right)\right]} L\left(G\right)$$

where $tr[L(G)]$ is the sum of elements on the main diagonal of matrix $L(G)$. The trace of Laplacian matrix for each graph is equaled to the sum of all the vertices degree. Thus the previous relation is equaled to the following relation:

$$\rho\left(G\right) = \frac{1}{2m} L\left(G\right).$$

3 Methods and Materials

3.1 Protein Pocket Structure

We used CASTp [14] to detect protein pockets. In this method, protein atoms are modeled as spherical balls (weighted points). Weighted Voronoi diagram is computed for this set based on the concepts explained in preliminaries. Next, $ResB$ is defined as $ResB = \{V_b \cap b | b \in B\}$ (Fig 2.left). Like Delaunay triangulation, by connecting the centers of neighboring regions in $ResB$ a graph called $CpxB$ is acquired (Fig 2.middle). Obviously $CpxB \subseteq DelB$. Afterwards, spherical balls simultaneously get bigger based on the variation of a parameter α. $CpxB_\alpha$ grows as α increases until it gets to $DelB$. Pockets are informally defined as components in $DelB - CpxB$ which become voids before getting disappeared as their corresponding balls grow based on changes in α.

Briefly, Flow relation is utilized to find protein pockets in CASTp (Fig 2.right). Cell ρ has a flow to its neighboring cell σ if the Voronoi center of ρ locates in the opposite side of the plane passing through the common face between ρ and σ. Sinks are defined as cells containing their own Voronoi centers. Pockets are defined as a set of cells which directly or indirectly flow to a sink. Thus it is sufficient to find sinks and their corresponding flows to detect pockets. Fig 2 is adopted from [8].

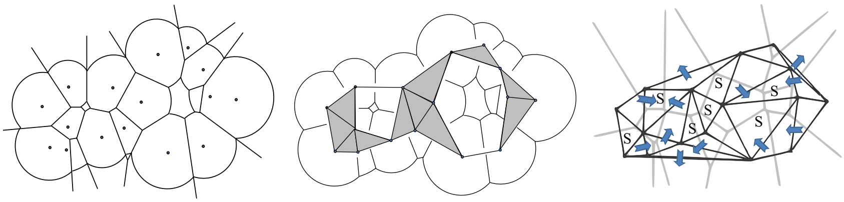

Fig. 2. Left:*Res B*, Middle: *Cpx B*, Right: Flow Relation (sinks are shown by S)

For each protein pocket we consider its graph consisting of vertices and edges corresponding to the centers of balls and Delaunay edges respectively. Then the adjacency matrix of these graphs is easily computed according to the definitions in preliminaries section.

3.2 Von Neumann Entropy

Information theoretical methods are common to compare complexity of networks. Shannon entropy is one of the representative network complexities which is defined by:

$$H(X) = H(p_1,\ p_2,\ \ldots,\ p_n) = -\sum_i p_i \log p_i$$

where X is a random variable with probability distribution $p_1,\ p_2,\ \ldots,\ p_n$. Entropy is widely used in various spheres, such as biology and chemistry, to measure complexity of graphs [15,16]. One approach to accomplish that is partitioning graph vertices to some classes $\{X_i\}$ such that the ratio of partitions size to graph vertices, $\frac{|X_i|}{|X|}$, results in a probability distribution. Shannon entropy of graphs is calculated for that probability distribution by:

$$H(G) = -\sum_i \frac{|X_i|}{|X|} \log\left(\frac{|X_i|}{|X|}\right).$$

However, there is no unique form to partition graphs and consequently different entropic measures can be assigned to them. Von Neumann entropy which is calculated by eigenvalues of a graph Laplacian matrix is a useful complexity measure while studying graphs [17]. Formally, Von Neumann entropy for any density matrix ρ is defined by:

$$S(\rho) = -tr(\rho \log \rho) = -\sum_i \lambda_i log \lambda_i.$$

Density matrix can be computed for graphs by the definition provided in preliminaries. Based on [17], Von Neumann entropy for an arbitrary graph G increases as edges of the graph scatter more randomly. In a study [17] of four different kinds of graphs with maximum twenty vertices, the smallest entropy is obtained for complete graph then it increases for star, random and perfect matching respectively. By Von Neumann entropy the complexity of graphs can be computed directly from their structures.

While examining the complexity of pockets, it is important to ignore any other parameters, size of corresponding graphs for instance. On the other hand, Von Neumann entropy of graphs increases as they grow in size and this also happens when studying different pockets with different sizes. To avoid this problem, relative Von Neumann entropy is utilized instead of its primary form as:

$$S(\rho \parallel \sigma) = \sum_i p_i \left(\log p_i - \sum_j P_{ij} \log q_j \right)$$

in which $\{p_i\}$ and $\{q_j\}$ are the eigenvalues of matrices ρ and σ and $P_{ij} = \langle X_i,\ Y_j \rangle^2$ where $\{X_i\}$ and $\{Y_j\}$ are the eigenvectors of matrices ρ and σ respectively. The corresponding graph entropy relative to a matrix with same dimension n and maximum entropy, $\frac{1}{n} I_n$, is chosen to examine the pocket complexity independent from its size:

$$S\left(\rho \parallel \frac{1}{n} I_n\right) = \log n - S(\rho).$$

Obviously when the graph entropy gets closer to the entropy of $\frac{1}{n} I_n$, the relative entropy decreases. Therefore, less relative Von Neumann entropy means more complexity in the corresponding graph of pocket and consequently more disorderliness in its structure.

To demonstrate that relative Von Neumann entropy does not change by pocket size variation, we use the following result from [18] for almost every graph G,

$$S(G) = (1 + o(1)) \log n.$$

This consequently results in:

$$S\left(G \,||\, \frac{1}{n} I_n\right) = o(1) \log n.$$

By the definition as n grows, $o\,(1)$ approaches zero. Thus, the grow rate of relative Von Neumann entropy is almost less than the logarithm of its size, which grows slightly.

3.3 Test Dataset

A dataset of 48 bound/unbound structures first introduced in [4] is used to have a comprehensive test over both ligand-bound and unbound structures. A widely used method to check whether a pocket is the real binding pocket, is to measure if its geometry center is within $4A^{\circ}$ of the ligand atoms. Whenever more than one pocket meet the condition, the one which is closer to the ligand is reported. This method was firstly used in [4].

4 Results and Discussion

We used CASTp website [1] to detect protein pockets. Although it provides comprehensive information about pocket size, atoms and mouths, it does not give simplices of each pocket. Therefore, to have Delaunay edges of pockets more than their vertices reported by CASTp, we constructed those graphs by the use of Computational Geometry Algorithms Library (CGAL)[2] which provides access to efficient geometric algorithms in the form of C++ library.

It is worth mentioning that weighted Delaunay triangulation is utilized as the basic graph according to its two convenient features for the purpose of predicting binding sites. First, since we want to examine the arrangement of points and its effect on forming binding sites, it is necessary to select a geometric graph. Delaunay triangulation is a geometric graph and is sensitive to the location of vertices. Second, there is an appropriate relation between Delaunay triangulation and void spheres in pockets. More precisely, every four vertices form a tetrahedron in Delaunay triangulation if the sphere passing through them is empty of any other points. Furthermore, these voids correspond to the regions in which the ligand atoms probably stand. Therefore, it is worthwhile to examine the distribution of them among protein surface when we want to predict the binding pockets. Indeed, Von Neumann entropy of Delaunay triangulation represents the distribution of voids on pocket surface. In a more comprehensive study, it is better to make use of dual complex of each pocket instead of its Delaunay triangulation to achieve a more accurate graph for each pocket.

Although we apply relative version of Von Neumann to ignore the effects of size, there are some small pockets with high entropy which are not eligible for being binding sites regarding their tiny available surface to interact with ligands. Figure 3 illustrates such examples.

To avoid taking those undesirable pockets to account as binding pockets, we narrowed down the list of pockets to 10 largest pockets of each sample in 48 bound/unbound dataset. The results are shown in Table 1.

[1] `http://cast.engr.uic.edu`

[2] `http://www.cgal.org`

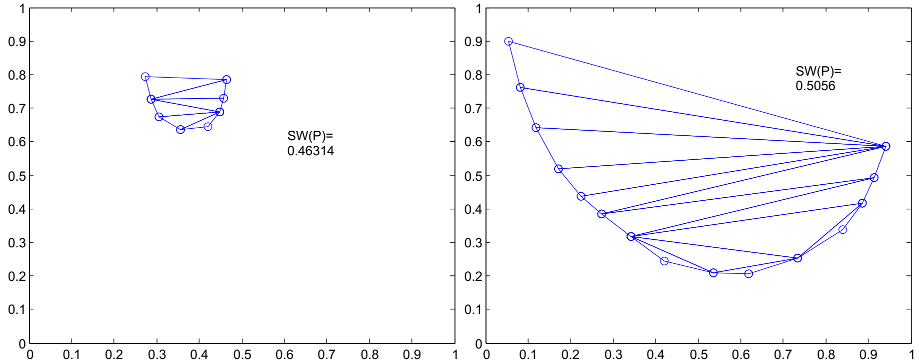

Fig. 3. Left: Extremely small pocket that has small relative Von Neumann entropy. Right: A more desired pocket.

Table 1. Prediction success rate of size and relative Von Neumann entropy

Ranking feature	Bound		Unbound	
	TOP1	TOP3	TOP1	TOP3
Pocket Size	67%	83%	58%	75%
Relative Von Neumann entropy	49%	77%	38%	83%

Table 2 shows the prediction success rate presented by different ranking methods adopted from [13] which is implemented on a same dataset. This is one of the recent studies about examining pockets properties and makes use of [19] for detecting protein pockets.

Table 2. Prediction success rate of different ranking features adopted from [13]

Methods	Unbound/Bound	
	TOP1	TOP3
Conservation score	57%	72%
Distance	56%	70%
Volume	44%	59%
Hidrophobic residues	30%	48%

We remind that pocket size is still the most successful feature to find binding pockets. Exploring novel features, preferably independent from size, can improve previous results. Pockets complexity measured by relative Von Neumann entropy can predict TOP3 pockets very well especially for unbound samples where the success rate even precedes previous results. In further studies, a hybrid criterion consisting of both size and entropy can be investigated.

Based on [17], Von Neumann entropy increases as graph edges scatter more randomly. Hence, for fixed number of edges, perfect matching and complete graph get maximum and minimum amount respectively. In particular, we have found that according to Table 1 in candidate binding pockets the edges of Delaunay triangulation and equivalently voids are scattered more uniformly than other pockets. In figure 4, two pockets with different edge distributions are shown. According to results, binding pockets are more likely to have a shape similar to left side figure rather than right side one.

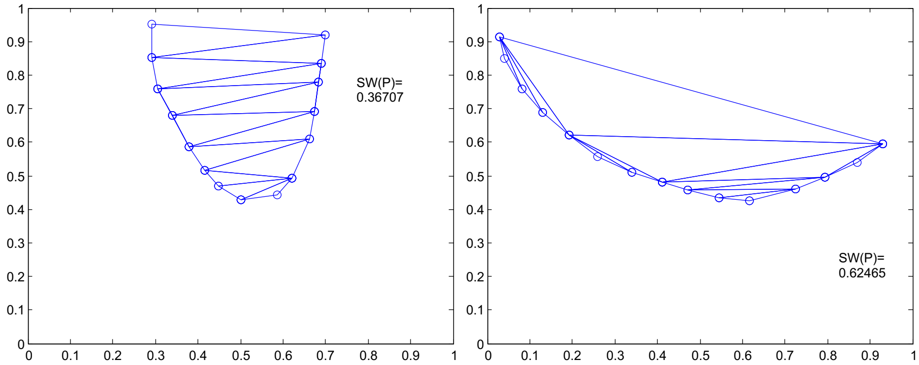

Fig. 4. Left:A pocket with uniform distribution of edges Right: A pocket with irregular distribution of edges

5 Conclusion

In this study we examined a novel feature in protein binding pockets based on the complexity of corresponding geometric graphs. CASTp was utilized for detecting pockets and weighted Delaunay triangulation was considered as pockets graphs. It was illustrated that binding pockets usually acquire less relative Von Neumann entropy which means more regular distribution of Delaunay edges and consequently uniform scattering of voids. Regarding small dependency of relative Von Neumann to the size of graphs, one can merge them in future studies to propose a comprehensive scoring criterion.

Acknowledgments. We would like to thank Dr. Fatemeh Zare-Mirakabad for her advice and assistance in analyzing biological aspects of this study. Besides, we appreciate Ms. Kamelia Jamaati for her valuable technical support in implementing C++ code.

References

1. Seco, J., Luque, J., Barril, X.: Binding Site Detection and Druggability Index from First Principles. Journal of Medicinal Chemistry 52, 2363–2371 (2009)
2. Pérot, S., Sperandio, O., Miteva, M.A., Camproux, A., Villoutreix, B.O.: Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. Drug Discovery Today 15, 656–667 (2010)
3. Hendlich, M., Rippmann, F., Barnickel, G.: LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins. J. Mol. Graph. Model. 15, 359–363 (1997)
4. Huang, B., Schroeder, M.: LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. BMC Structural Biology 6, 19–29 (2006)
5. Weisel, M., Proschak, E., Schneider, G.: PocketPicker: analysis of ligand binding-sites with shape descriptors. Chemistry Central Journal 1 (2007)
6. Laskowsk, R.A.: SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J. Mol. Graph. 13, 323–330, 307–308 (1995)
7. Brady, G.P., Stouten, P.F.: Fast prediction and visualization of protein binding pockets with PASS. J. Comput. Aided Mol. Des. 14, 383–401 (2000)
8. Edelsbrunner, H., Facello, M., Liang, J.: On the definition and the construction of pockets in macromolecules. Descrete Applied Mathematics 88, 83–102 (1998)
9. Le Guilloux, V., Schmidtke, P., Tuffery, P.: Fpocket: an open source platform for ligand pocket detection. BMC Bioinformatics 10 (2009)
10. Haung, B.: MetaPocket: a meta approach to improve protein ligand binding site prediction. OMICS 13, 325–330 (2009)
11. Zhang, Z., Li, Y., Lin, B., Schroeder, M., Huang, B.: Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. Bioinformatics 27, 2083–2088 (2011)
12. Laskowski, R.A., Luscombe, N.M., Swindless, M.B., Thornton, J.M.: Protein clefts in molecular recognition and function. Protein Science 5, 2438–2452 (1996)
13. Gao, J., Liu, Q., Kang, H., Cao, Z., Zhu, R.: Comparison of Different Ranking Methods in Protein-Ligand Binding Site Prediction. International Journal of Molecular Science 13, 8752–8761 (2012)
14. Dundas, J., Ouyang, Z., Tseng, J., Binkowski, A., Turpaz, Y., Liang, J.: CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucleic Acids Res. 34, W116–W118 (2006)
15. Dehmer, M., Barbarini, N., Varmuza, K., Graber, A.: A large scale analysis of information-theoretic network complexity measures using chemical structures. PLoS One 4, e8057 (2009)
16. Dehmer, M., Mowshowitz, A.: A history of graph entropy measures. Information Science 181, 57–78 (2011)
17. Passerini, F., Severini, S.: Quantifying complexity in networks: The Von Neumann entropy. IJATS 4, 58–67 (2009)
18. Du, W., Li, X., Li, Y., Severini, S.: A note on the von Neumann entropy of random graphs. Linear Algebra and its Application (2010)
19. Dai, T., Liu, Q., Gao, J., Cao, Z., Zhu, R.: A new protein-ligand binding sites prediction method based on the integration of protein sequence conservation information. BMC Bioinformatics 12(suppl. 14), S9 (2011)