NICOGRAPH International 20xx, pp. 1234-1237

A Fast Pocket Extraction and Evaluation for Protein Surfaces

Ayaka Kaneko[1)] **Yukari Nakamura**[1)] **Takayuki Itoh**[1)]

1) Graduate School of Humanities and Sciences, Ochanomizu University
{ayaka, sincere, itot} (at) itolab.is.ocha.ac.jp

Abstract

Research of protein is pivotal to the drug discovery, since most of the drugs act upon proteins inside human body. Drugs act when they are close to the concave portions, so called “pockets”, of protein surfaces. Therefore, detection and analysis of the pockets are also important for the drug discovery. This paper presents a fast pocket extraction and evaluation technique for protein surfaces. Supposing protein surfaces are provided as triangular meshes, the method first applies mesh simplification to smooth small geometric features. It then detects concave portions from the simplified triangular meshes as pockets. The method then evaluates the pockets from the following viewpoints: geometric evaluation and chemical evaluation.

1. Introduction

Since the recent research of drug discovery has been incredibly increasing, there are many drugs produced around the world. As the discoveries expanded, more the problems including side-effects have caused. Here, I would like to explain how the side-effects cause. Most of the drugs work for the body by acting directly to proteins. Side-effects may cause by drugs acting to non-target proteins. Drugs act upon the concave portions of protein surfaces, so called “pockets”; therefore drugs often act upon the non-target pockets. On the other hand, there are proteins which have no pockets on their surfaces. It takes large computation time to analyze all the proteins because there are millions of amino acid composition pattern, which is protein. Therefore, we often suppose that proteins those do not have pocket-shaped portion on their surfaces have no need to worry about side-effects, and also will not be target proteins. In this study, we would like to detect pockets, exclude the proteins those do not evidently have pocket-shaped portion on their surfaces, and analyze the proteins those had pockets.

Researches on pocket detection are one of the cornerstones of modern drug discovery. Perot et al. [1] have summarized recent researches on “druggable” pockets and binding site. It reports several methods to search drug gable pockets, which can be divided in two major categories: geometric/probe algorithms and/or energy-based methods. Geometric pocket detection algorithms cover a variety of techniques, such as fitting of virtual spheres into the solvent-accessible space between protein atoms, and use of Delaunay triangulation or of the alpha-shapes approach to delineate cavities. Our study is one of the geometric approaches. Problem that hampers the analysis of pockets is the lack of standard definition of what constitutes a pocket. Therefore, there are variety of method-dependent geometric descriptions of binding pockets such as depth, size, volume, and amino acid composition. Another problem is that most of geometric approaches take large computation time.

We discussed the requirements of protein pocket discovery techniques with experts of a pharmaceutical company. We heard that geometric preciseness is not always important, since shapes of protein surfaces are not stable due to their molecular motions. Based on this discussion, we agreed with them that development of “rough but fast” pocket extraction techniques would be very fruitful for them.

This paper presents a geometry-based protein pocket extraction technique applying a mesh simplification technique. The technique is relatively fast, and easy to control the size of pockets to be detected. The paper also presents our work on protein pocket evaluation. The work consists of two trials: example-based evaluation based on geometric comparison of shapes with known druggable pockets, and criteria-based evaluation including depth, width, electric potential, and hydrophobicity.

2. Related Work

As discussed in Section 1, pocket discovery has been an important topic for protein druggability analysis. Many techniques have been presented, as surveyed in [1], and they are roughly categorized into geometry- and energy-based techniques. Energy-based techniques have been more major in the early stage of this field; however, many geometry-based techniques have been presented in these several years. Kawabata et al. [2] presented a technique which discovers concave portions of protein surfaces by rolling two sizes of spheres on them: this approach is good at intuitive parameter setting, while it may require large computation time. Halgren [3] presented another effective technique which generates grid points surrounding proteins and discovers pockets from the distribution of exterior grid-points. It is easy to implement, while pocket detection results may depend on the direction of the grid-points.

We previously presented a protein surface analysis method [4] which applies mesh simplification; however, it did not focus on discovery of pockets and binding sites.

3. Pocket Extraction and Evaluation

This section introduces our technique on pocket extraction and evaluation. Supposing protein surfaces are modeled as triangular meshes, the technique first simplifies the surfaces. It then roughly extracts concave portions of the simplified surfaces and projects them onto the original surfaces. Section 3.1 to 3.3 introduces the pocket extraction process.

We explored what kinds of evaluation schemes are useful for the extracted pockets. First trial is example-based evaluation which geometry compares the shapes of pockets with known druggable pockets. Second trial is criteria-based evaluation including depth, width, electric potential, and hydrophobicity. Section 3.4 to 3.5 introduces the pocket evaluation schemes.

3.1 Protein Surfaces

Our method uses protein surface datasets downloaded from the protein surface database "eF-site" [5]. This database collects the surfaces of proteins registered in PDB (Protein DataBank), by applying a Colony surface extraction technique [6]. We can freely obtain the protein surfaces as triangular meshes in XML format, containing vertices, edges connecting pairs of the vertices, and triangles enclosed by sets of three edges.

3.2 Mesh Simplification

Our technique aims to detect adequately-sized concave regions ignoring smaller bumps. It applies a mesh simplification technique using an implicit surface to get rough geometry by smoothing small bumps, and so only larger geometric frames will remain. Our implementation of the mesh simplification step generates a grid which surrounds the protein surface, and assigns scalar values to the grid-points. It assigns scalar values to the grid-points based on the distances to the closest vertices. Our implementation then generates isosurfaces as the simplified protein surfaces, by applying the Marching Cubes [7] method. This step forms polygons inside the grid-elements by connecting the points on the grid-edges where the scalar values are zero Consequently, numbers of triangles are significantly reduced to accelerate the pocket extraction step. Also, it is flexible to control the sizes of detected pockets just by changing the intervals of grid-lines.

3.3 Pocket Extraction and Projection

Then, we give attributes to each vertex. Let vertices of the simplified triangular mesh v_i, its position (x_i, y_i, z_i), and its normal vector $(n_{x_i}, n_{y_i}, n_{z_i})$. Our method applies to following Equation (1) to v_i, and its adjacent vertices.

$$t = n_{x_i}(x - x_i) + n_{y_i}(y - y_i) + n_{z_i}(z - z_i)$$

Here, $(x, y, z)$is the position of an adjacent vertex, which is connected with v_i by an edge. Our method assigns the attribute "convex" if all of the *t* values are positive, the attribute "concave" if all of the values are negative, to v_i. In other words, our method assigns "convex" if all adjacent vertices of v_i are interior the tangent plane ofv_i, and assigns "concave" if all adjacent vertices of v_i are exterior the tangent plane of v_i. Otherwise, it assigns "others".

Next, our method simply assigns "concave" to the triangles which are connected to one or more "concave" vertices. It treats regions consisting of sets of adjacent "concave" triangles as pocket candidates.

Our method then projects the pocket candidates extracted from a simplified triangular mesh onto the original triangular mesh. Let triangles of the original meshb_i, and triangles of simplified mesha_i. Our method simply specifies a_i, which is the closest to b_i, and copies the attributes of a_i to the matched of b_i. The technique treats the portions on the original triangular mesh where pocket candidates are projected as pockets.

3.4 Example-based Pocket Evaluation

We implemented a geometric feature calculation for the pocket including the following procedure. Firstly, it specifies the plane that minimizes the sum of distances from vertices of the outer loop of a pocket, and calculates the center position and normal vector of the plane. This section describes this plane *Po*, the center position C_{Po}, and the normal vector N_{Po}. At the same time, it evenly generates sample points on the triangles of a concave portion. The technique then calculates the distance from sample points to *Po*, as well as the distance from the sample points to the line which is parallel to N_{Po} and passes C_{Po}. The technique treats the histogram of the two distances as a feature vector of a pocket.

At the same time, the technique supposes that users collect sample pockets which are really observed well-shaped and druggable. It calculates the geometric feature values of the sample pockets and stores to a database as a preprocessing.

Our technique calculates the cosine similarity between the geometric feature values of a pocket and stored sample pocket, and treats the maximum cosine value as the score of the pocket. We treat high score pockets as druggable pockets.

3.5 Criteria-based Pocket Evaluation

We observed the following four values to discover good criteria to divide proteins into druggable and undruggable ones. Here, the criteria discovered by this observation can be implemented to the score calculation of example-based pocket

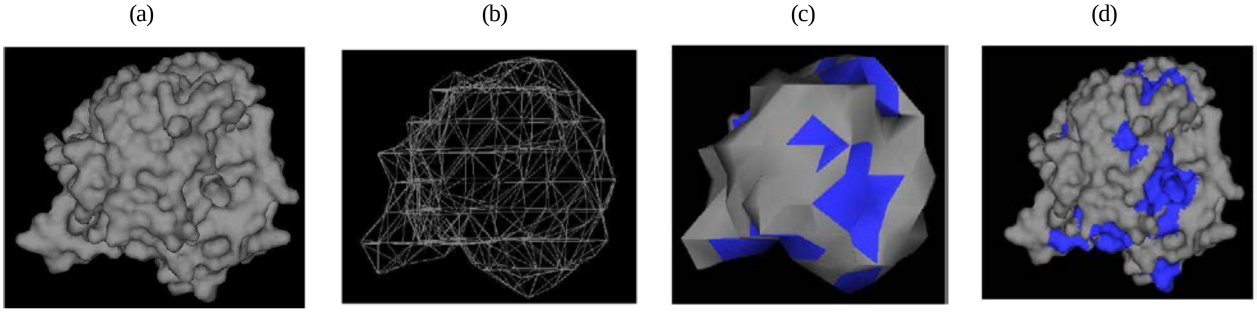

Figure 1. Example of pocket extraction process. (a) Original protein surface. (b) Mesh simplification result. (c) Pocket candidate extraction result on the simplified mesh. (d) Pocket projection result on the original mesh. Pockets are painted as blue.

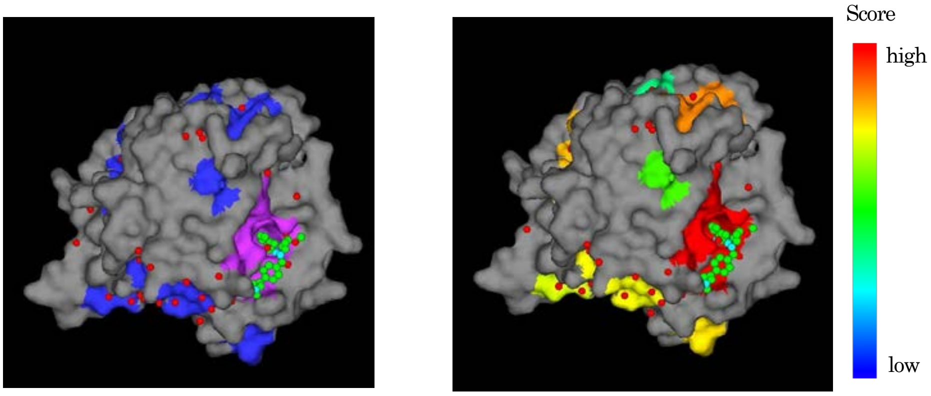

Figure 2. Example of example-based pocket geometry evaluation. (Left) Non-protein atoms drawn as colored small spheres. Pockets around concentrated non-protein atoms are painted as pink. Their pockets are defined as sample pockets. (Right) Pocket geometry evaluation results. Here, pockets are colored based on their scores; red or orange pockets are high score pockets. That means pockets painted in red or orange colors are geometrically similar to at least one of sample pockets.

evaluation described in Section 3.4.

(1) Depth: We calculated the maximum distance from sample points on the pocket to *Po* as the depth of the pocket.

(2) Width: We projected the sample points and vertices on the outer loop onto *Po*. We calculated the minimum distance from a sample point to the vertices on *Po*, and treated the maximum value of them as the width of the pocket.

(3) Electric potential: Protein surface datasets downloaded from eF-site contain electric potential values at their vertices. We calculated the average and variance of electric potential values of the vertices of pockets.

(4) Hydrophobicity: Protein surface datasets downloaded from eF-site contain hydrophobicity values at their vertices. We calculated the average and variance of hydrophobicity values of the vertices of pockets.

4. Results

We implemented the technique with JDK (Java Development Kit) 1.6.0, and executed on Lenovo ThinkPad T420s (CPU 2.7GHz Dual Core, RAM 8.0GB) with Windows 7.

4.1 Pocket Extraction

Figure 1 shows an example of protein surface, mesh simplification, pocket extraction, and pocket projection, using the surface of protein 1EZQ containing 18860 vertices. Here, pocket extraction results strongly depend on intervals of grids described in Section 3.2. Our current implementation applies five interval values, 4.0, 5.0, 6.0, 7.0, and 8.0 angstrom. Computation time also depends on intervals of grids: we measured that computation time was 1.855 to 7.462 seconds. Mesh simplification process occupied a large part of computation time in our experiment. We would like to apply an accelerated isosurface extraction technique [8] to reduce the computation time.

4.2 Example-based Pocket Evaluation

Many of protein datasets downloaded from PDB contain non-protein atoms which remained after protein crystallization process. Experts focus on binding sites around concentrated non-protein atoms because they are often druggable. Based on this knowledge, we extracted pockets around the concentrated non-protein atoms as sample pockets as shown in Figure 2(Left). We then calculated similarity between each pocket and the extracted sample pockets, and visualized the results as shown in Figure 2(Right). We observed the results and subjectively evaluated the results are good. We would like to validate how this approach can contribute to pocket druggability test as a future work.

4.3 Criteria-based Pocket Evaluation

We calculated depth, width, electric potential, and hydrophobicity of pockets around concentrated non-protein atoms, and visualized by scatter plots as shown in Figures 3 and 4. Unfortunately we have not discovered strong correlations between depth, width, or electric potential and druggability. On the other hand, Figure 4 denotes that most of proteins which have pockets with positive average hydrophobicity values are druggable. We would like to test with more variety of values to explore correlations between such values and druggability.

5. Conclusion

This paper presented a technique to extract pockets from protein surface. The technique first generates the smoothed protein surface by applying a mesh simplification method. It then detects the adequately-sized concaves as pocket candidates, and finally projects to the original protein surfaces. The paper also discussed what kind of evaluation is useful to determine druggability of protein pockets. This discussion is still in an early stage and we would like to explore more variety of schemes.

5.References

[1] S. Perot, O. Sperandio, M. A. Miteva, A.-C. Camproux, B. O. Villoutreix, Druggable Pockets and Binding Site Centric Chemical Space: A Paradigm Shift in Drug Discovery, Drug Discovery Today, 15(15-16), 656-667, 2010.

[2] T. Kawabata, N. Go, Detection of Pockets on Protein Surfaces Using Small and Large Probe Spheres to Find Putative Ligand Binding Sites, Proteins: Structure, Function, and Bioinformatics, 68(2), 516-529, 2007.

[3] T. A. Halgren, Identifying and Characterizing Binding Sites and Assessing Druggability, Journal of Chemical Information and Modeling, 49(2), 377-389, 2009.

[4] K. Nishiyama, T. Itoh, PROTEIN: A Visual Interface for Classification of Partial Reliefs of Protein Molecular Surfaces, The Institute of Image Electronics Engineering of Japan, 37(3), 181-188, 2008.

[5] eF-site, http://ef-site.hgc.jp/eF-site/index.jsp.

[6] Molecular Surface Package, http://connolly.best.vwh.net/

[7] W. E. Lorensen, H. E. Cline, Marching Cubes: A High Resolution 3D Surface Construction Algorithm, ACM SIGGRAPH, 21(4), 163-169, 1987.

[8] T. Itoh, Y. Yamaguchi, K. Koyamada, Fast Isosurface Generation Using the Volume Thinning Algorithm, IEEE Transactions on Visualization and Computer Graphics, 7(1), 32-46, 2001.

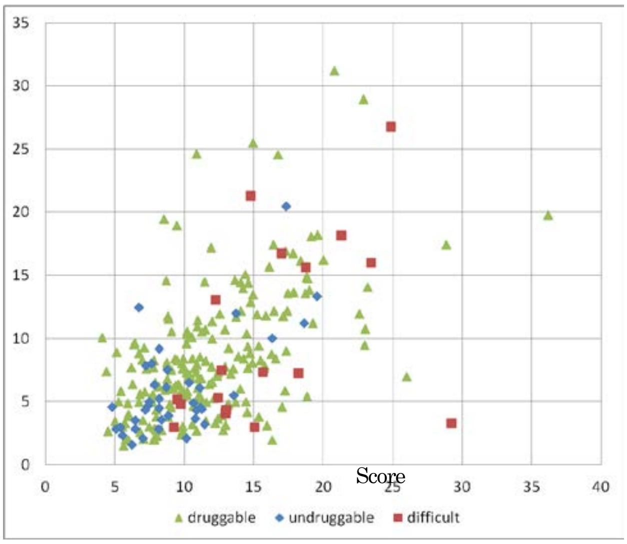

Figure 3. Depth (X-axis) and width (Y-axis) of pockets around concentrated non-protein atoms . Strong correlations between these values and druggability have not been discovered yet.

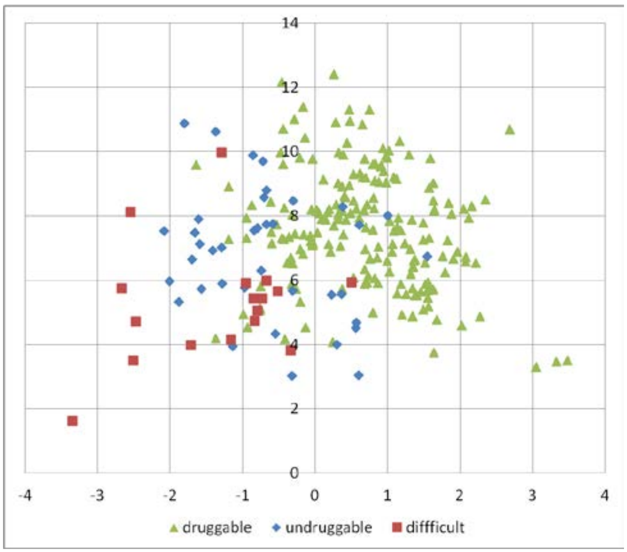

Figure 4. Average (X-axis) and variance (Y-axis) of hydrophobicity of pockets around concentrated non-protein atoms . Most of proteins which have pockets with positive average hydrophobicity values are druggable.