Druggability Analysis and Prediction based on Geometric Distances Between Amino Acid Residues and Protein Surface Pockets

Makiko Miyoshi, Ayaka Kaneko, Takayuki Itoh
Ochanomizu University
Tokyo, Japan
Email: {iqams, ayaka, itot} @itolab.is.ocha.ac.jp

Kei Yura
Ochanomizu University
Tokyo, Japan
Email: yura.kei@ocha.ac.jp

Masahiro Takatsuka
The University of Sydney
Sydney, Australia
Email: masa.takatsuka@icloud.com

***Abstract*—Protein is the major component of the organism. A concave (pocket) on a protein surface is known to be the best target for a drug to react. We previously presented a study on distance analysis between pockets and amino acid residue. We firstly identified pockets on the protein surface and then calculated distances between atoms of an amino acid residue and the deepest points or the outer loops of the pockets. We extracted proteins which at least one of the pockets are close to arbitrary pairs of amino acid residues, calculated the ratios of druggable proteins, and visualized the distribution of the ratios as a colored matrix. We suggested from the visualization results that particular pairs of amino acid residues may affect the druggability of the proteins in our previous study. This paper presents an extension of our study to explore the relevance between druggability of proteins and distances between a set of amino acid residues and protein surface pockets. Our technique treats the pockets as 20-dimensional vectors consisting of distances to each of amino acid residues, and applies GeodesicSOM with the set of the vectors. Spherical maps generated by GeodesicSOM are used to visualization of distribution of the pockets in the 20-dimensional vector space, and estimation of druggability of proteins with the 20-dimensional vectors of the pockets.**

***Keywords*-Protein, Druggability, Visualization, Self-Organizing Map.**

I. Introduction

Protein is the major component of the organism. It has a unique typical three-dimensional structure determined by its amino acid sequence. A concave (pocket) on the surface of a protein is known to be the best target for a drug to react. Reactivity between proteins and drug compounds is often called "druggability", and proteins that have relatively higher reactivity with the drug compounds are called "druggable proteins".

pocket discovery has been an important topic for protein druggability analysis. Many techniques have been presented, as surveyed in [5], 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 [1] 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 proposed a technique for pocket extraction from protein surfaces [4], which requires less computation time than existing techniques. This technique just extracts well-sized concave portions of the protein surfaces; however, the extracted pockets are not necessarily druggable.

On the other hand, small molecules including drug compounds tend to combine certain amino acid residues [7]. They found preferred amino acid residues than other amino acid residues at interaction site. We supposed this knowledge may be a powerful hint to analyze the druggability of proteins, and have been studying distances between pockets and amino acid residues can be fruitful information to diagnose druggability.

We presented a visualization technique to discover the relationship between druggability of proteins and distances between amino acid residues and protein surface pockets [3]. The technique firstly extracts pockets from a set of input protein surfaces, and then calculates the geometric features of the pockets, including distances between amino acid residues and the extracted pockets. We developed three types of visualization components to represent the results. One of the visualization components, matrices to summarize the druggability analysis applied to all the possible pairs of amino acid types, were especially interesting in our experiment. The visualization results suggested that particular pairs of amino acid residues might affect the druggability of proteins, since the red columns depicting high druggability concentrated at a particular portion of the matrix, as shown in Figure 1.

There are 20 different amino acids most commonly exist in nature [1], and most of proteins contain these 20 amino acids. However, our above study just demonstrated the

[1] http://www.proteinstructures.com/Structure/Structure/amino-acids.html

978-1-5090-2305-9/16 $31.00 © 2016 IEEE
DOI 10.1109/NicoInt.2016.4

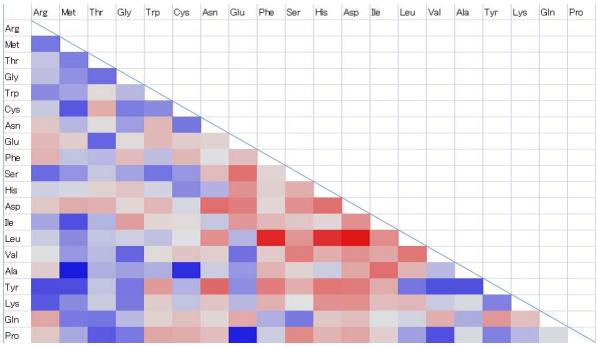

Figure 1. Representation of the druggability analysis in a lower triangular matrix [3]. Both horizontal and vertical axes denote types of amino acid residue lined in the same order. Colors of the columns denote the druggability of pockets close to a pair of amino acid residues. Redness denotes the high druggability, blueness denotes the low druggability, and saturation denotes the number of corresponding pockets.

relationship between druggability and distances to one or two amino acid residues from the pockets. Our next interest is how distances to three or more amino acid residues are related to druggability of proteins. To address this interest, we treated the distances from a pocket to each of amino acid residues as a 20-dimensional vector, and applied GeodesicSOM [6] to a set of vectors corresponding to a set of pockets. We can visualize the distribution of pockets in the 20-dimensional vector space by displaying a spherical map generated by GeodesicSOM, and observe how druggable or undruggable pockets form clusters in the vector space. Also, we developed a technique to estimate the druggability of proteins by calculating the distances from their pockets to their amino acid residues and compare with the spherical map of GeodesicSOM. This paper introduces the processing flow of our technique and experimental results with the dataset of 60 proteins published by Halgren [1].

II. Processing Flow

This section describes the processing flow of the presented technique. The technique firstly extracts pockets from protein surfaces, and calculates distances between the bottom of the pockets and amino acid residues of the protein.

A. Pocket extraction

Our implementation extracts pockets from the protein surface datasets by applying a quick extraction technique [4]. This technique goes through the following procedures, and extracts pockets from protein surfaces.

1) Apply a mesh simplification technique using an implicit surface to get rough geometry by smoothing small bumps, and consequently only larger geometric features remain.
2) Extract peptide sizes of the concave portions on the simplified triangular mesh.
3) Project the concave portions extracted from the simplified triangular mesh onto the original triangular mesh as pocket candidates.
4) Remove the unnecessary parts of the projected pocket candidates.

Our implementation applies protein surface datasets downloaded from the protein surface database "eF-site" [8].

B. Distance calculation

We calculate the distance between a pocket and an amino acid residue as follows. The technique firstly specifies the plane that minimizes the sum of distances from vertices of the outer loop of a pocket. It then calculates the distance from vertices of the pocket to the plane, and identifies the deepest point of the pocket as the vertex which has the largest distance value. In this study we define the distance as the smallest distance between the deepest point of the pocket and the atoms belonging to the amino acid residue.

C. Druggability estimation and visualization

Our previous study [3] visualized the statistics of distances between a pocket and one or two amino acid residues, and explored the relationships between druggability of proteins and the statistics of distances. Meanwhile, we may need to analyze the relationships between the druggability and distances of three or more amino acid residues to the pockets, to discover further new knowledge. We extended the previous study by treating a pocket as a 20-dimensional real value vector with a categorical value, $p_i = \{d_{i1}, d_{i2}, ..., d_{i20}, b_i\}$. Here, p_i is the i-th pocket, and d_{ij} is the distance from the i-th pocket to the j-th amino acid residue. b_i is a categorical value indicating the druggability of the belonging protein, which takes "druggable", "difficult", "undruggable", or "unknown".

We applied GeodesicSOM [6] to the set of pockets of protein surfaces. GeodesicSOM is a kind of spherical SOM (Self-Organizing Map), where SOM is a unsupervised neural network which learns the characteristics of a set of multi-dimensional vectors and non-linearly maps the input data onto low-dimensional spaces. Our challenge in this study includes the following two issues:

- Visualization of pockets divided according to the distance values and druggability.
- Estimation of druggability of proteins from the set of pockets those druggability are already known.

We firstly divided the pockets into two groups $P_1 = \{p_1, ..., p_{n1}\}$ and $P_2 = \{p_{n1+1}, ..., p_{n1+n2}\}$, Here, P_1 is a set of pockets those druggability is known ("druggable", "difficult" or "undruggable"), treated as a training dataset. Meanwhile, P_2 is another set of pockets those druggability is "unknown", treated as a test dataset. Numbers of pockets in P_1 and P_2 are n_1 and n_2 respectively.

Our study supposes to input P_1 to GeodesicSOM and calculates the positions of the pockets on a spherical map.

We can use this map to visualize how druggable (or undruggable) pockets are clustered according to the distances to amino acid residues. Also, we apply this map to estimate druggability of the test datasets by the following procedure.

To estimate the druggability of a pocket p_k in P_2, our druggability estimation technique firstly calculates distances between p_k and any of the pockets belonging to P_1, and specifies the pocket p_j in P_1 which is the closest to p_k. Here, the distance is calculated as an Euclidian distance between the 20-dimensional vectors defined as follows:

$$dist_{jk} = \sqrt{\sum_{i=1}^{20}(d_{ji} - d_{ki})^2} \tag{1}$$

The technique determines that the druggability of p_k is unpredictable when the distance between p_k and p_j exceeds the user-defined threshold $distmax$. Otherwise, we apply the following procedure.

This section calls the neuron of GeodesicSOM which p_j belongs as "winner neuron". Figure 2 illustrates an example structure around the winner neuron corresponding to a dot painted in orange. We suppose some of adjacent neurons painted in blue in this figure are associated with other pockets, and the others are not associated with any pockets. Characters "d", "f", and "u" in this figure denote druggability of the associated pockets, "druggable", "difficult", and "undruggable" respectively.

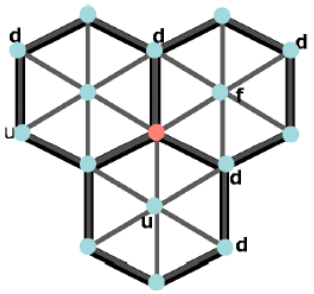

Figure 2. Illustration of the winner neuron (painted in orange) and adjacent neurons (painted in blue). Characters denote druggability of pockets associated to some of the adjacent neurons.

The technique then calculates the distances between p_k and pockets associated to the adjacent neurons by applying the equation (1). It extracts the pockets those distances are smaller than $distmax$, and divides them to the following three groups:

P_{ad}: group of "druggable" pockets.
P_{af}: group of "difficult" pockets.
P_{au}: group of "undruggable" pockets.

We define the possibility of druggability by the following equations:

$$\begin{aligned} psbd_k &= \sum_{i \in P_{ad}} \frac{1}{dist_{ki}} \\ psbf_k &= \sum_{i \in P_{af}} \frac{1}{dist_{ki}} \\ psbu_k &= \sum_{i \in P_{au}} \frac{1}{dist_{ki}} \end{aligned} \tag{2}$$

where $psbd_k$, $psbf_k$, and $psbu_k$ are possibilities of "druggable", "difficult" and "undruggable" respectively. We simply estimate the druggability of the pocket p_k as "druggable" if $psbd_k$ is the maximum. Similarly, we determine as "difficult" if $psbf_k$ is the maximum, or "undruggable" if $psbu_k$ is the maximum.

III. Experiment

We tested our technique with a set of 60 proteins of which druggability was examined by Halgren [1]. Chemical structures of all the 60 proteins are published by PDB (Protein DataBank), and geometry of their surfaces are published by eF-site.

Protein datasets in PDB format often contain records of "HETATM" which describe the coordinates of non-protein atoms/molecules in protein crystal. These atoms/molecules except for water molecules tend to bind with specificity to the protein. Therefore when a molecule is found in a pocket, the pocket has specificity to a certain molecule and we name the pocket "reactive". A pocket without a molecule is hence named "non-reactive". In the experiment, we extracted pockets [4] on the surfaces of the 60 proteins, searched for non-protein atoms/molecules around the extracted pockets, and finally extracted the reactive pockets as the input dataset.

We applied all the pockets in the input dataset to GeodesicSOM. Figure 3 shows a visualization example. Characters denote druggability of pockets and their positions on the GeodesicSOM, where black or gray are randomly assigned to the characters just to improve the readability. There were just a small number of pockets of "difficult" or "undruggable" proteins in this dataset, indicated as "f" or "u" in the visualization result, since reactive pockets often make proteins "druggable". Colors depict geodesic distances among adjacent neurons. Regions painted in cold colors depict that input vectors of adjacent pockets inside these regions are actually similar.

This visualization result shows that the pockets form several clusters corresponding to regions painted in cold colors according to the vectors of distances to amino acid residues. It suggests that clustering according to these distance vectors may bring knowledge regarding the protein druggability. Also, the result shows that undruggable pockets form several small clusters as indicated by pink circles in Figure 3. These clusters may suggest the characteristics of reactive pockets which cannot make the proteins "druggable".

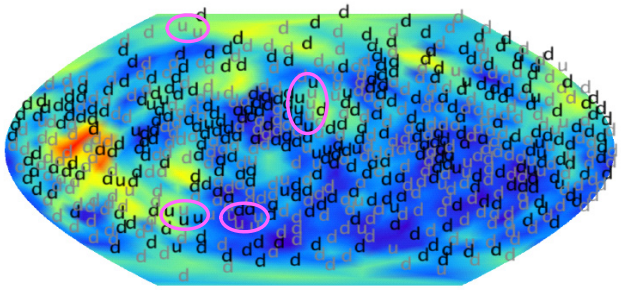

Figure 3. Reactive pockets mapped to GeodesicSOM. Pink circles indicate that pockets of "undruggable" proteins form several small clusters on GeodesicSOM.

We also tested our druggability estimation technique. We randomly divided 60% of the pockets to P_1 (training dataset), and the rest to P_2 (test dataset). Here, P_2 consisted of 474 pockets, including 391 pockets of "druggable" proteins, 35 pockets of "difficult" proteins, and 48 pockets of "undruggable" pockets.

In many use cases, we would like to divide "druggable" proteins and others, or "undruggable" proteins and others. Therefore, we aggregated the estimation results for pockets estimated as "druggable" or "undruggable" respectively, as shown in Tables I and II. Precision and recall for extraction of pockets of "druggable" proteins were 0.869 and 0.969, respectively. Meanwhile, for extraction of pockets of "undruggable" proteins, precision and recall were 0.250 and 0.571. We archived good precision and recall for extraction of "druggable" proteins. On the other hand, result for "undruggable" proteins was poor. We suppose that one of the reasons for the poor result was that number of pockets of "undruggable" pockets was too small. We would like to find larger datasets and test this technique again.

Table I
STATISTICS FOR POCKETS ESTIMATED AS "DRUGGABLE".

	Known as "druggable"	Known as others	total
Estimated as "druggable"	379	57	436
Estimated as others	12	26	38
Total	391	83	474

Table II
STATISTICS FOR POCKETS ESTIMATED AS "UNDRUGGABLE".

	Known as "undruggable"	Known as others	total
Estimated as "undruggable"	12	9	21
Estimated as others	36	417	453
Total	48	426	474

We had the same experiment five times: random division of pockets into P_1 and P_2, generation of spherical maps with P_1, and estimation of druggability of pockets in P_2. Precision and recall were almost similar through the five results.

IV. CONCLUSION

This paper presented a technique for visualization and estimation of protein druggability applying GeodesicSOM. The technique treats pockets of protein surfaces as 20-dimensional vectors consisting of distances to each of amino acid residues. Spherical maps generated by GeodesicSOM can be used for visualization of distribution of pockets in the 20-dimensional vector space, and for estimation of druggability of new proteins.

We had an experiment with the technique applying 60 proteins introduced by Halgren [1]. We could visualize clusters of "druggable" or "undruggable" pockets in the vector space; however, we have not yet discussed the reasoning of the clusters. We would like to analyze and discuss how the clusters formed as future work. We also tested the estimation of druggability of proteins We could archive a good result for extraction of "druggable" proteins; however, the result for extraction of "undruggable" proteins was poor. We need to look for larger datasets to archive more reliable results.

REFERENCES

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

[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] M. Miyoshi, T. Itoh, K. Yura, A Visual Analytics of Geometric Distances Bewteen Amino Acids and Surface Pockets of Proteins, *18th International Conference on Information Visualisation (IV2014)*, 164-169, 2014.

[4] Y. Nakamura, A. Kaneko, T. Itoh, An Accelerated Pocket Extraction and Evaluation Technique for Druggability Analysis with Protein Surfaces, *ACM SIGGRAPH ASIA*, Poster Session, 2011.

[5] 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.

[6] Y. Wu, M. Takatsuka, Spherical Self-Organizing Map Using Efficient Indexed Geodesic Data Structure, Neural Networks, 19(6), 900-910, 2006.

[7] A. Yamaguchi, K. Iida, N. Matsui, S. Tomoda, K. Yura, M. Go, Het-PDB Navi. :A Database for Protein-Small Molecule Interactions, *Journal of Biochemistry*, 135, 79-84, 2004.

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