A Visual Analytics of Geometric Distances Between Amino Acids and Surface Pockets of Proteins

Makiko Miyoshi, Ayaka Kaneko, Takayuki Itoh, Kei Yura
Ochanomizu University
Tokyo, Japan
{iqams, ayaka, itot}@itolab.is.ocha.ac.jp, yura.kei@ocha.ac.jp

***Abstract*—Protein is the major component of the organism. It has a unique three-dimensional structure formed 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. We started analyzing how "druggability" of proteins related to the location of amino acids in a pocket. This paper presents a visualization tool for distance analysis between pockets and the amino acid residue. Provided that a protein surface is described by a triangular mesh, this tool first identifies pockets on the protein surface, specifies the deepest point and outer loops of the pocket, and calculates distances between atoms of an amino acid residue and the deepest point or the outer loops of the pocket. The tool then visualizes the statistics of the distance calculation results by polyline charts and the distribution by scatterplots. This paper proposes a biological interpretation of the visualization results.**

Keywords-Visualization; Protein; Amino acid; Pocket identification

I. INTRODUCTION

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". Protein has a unique three-dimensional structure determined by its amino acid sequence. A concave (pocket) on the surface of a protein is one of the promising characteristics for druggability. Therefore, developing effective methods to discover and analyze protein pockets have been an active search target.

We proposed a technique for pocket extraction from protein surfaces [4], which requires less computation time than the existing techniques. This technique enables extracting appropriate concave portions of the protein surfaces; however, the extracted pockets are not necessarily druggable. We then proposed visualization techniques [2] to compare the shapes and chemical properties of pockets to discover criteria to divide pockets into druggable and undruggable ones. We tested the following four values; depth, width, electrostatic potential and hydrophobicity.

We proposed a technique for pocket extraction from protein surfaces [4], which requires less computation time than the existing techniques. This technique enables extracting appropriate concave portions of the protein surfaces; however, the extracted pockets are not necessarily druggable. We then proposed visualization techniques [2] to compare the shapes and chemical properties of pockets to discover criteria to divide pockets into druggable and undruggable ones. We tested the following four values; depth, width, electrostatic potential and hydrophobicity.

On the other hand, small molecules including drug compounds tend to interact with a certain type of amino acid [7]. We therefore conjectured that distances between a pocket and a amino acid residue can be fruitful information to examine druggability.

In this paper, we present a technique to discover a relationship between amino acid residue and druggability by visual analytics. First, our technique uses protein surface datasets downloaded from the protein surface database "eF-site" [8]. Then we extracted pockets from the dataset applying a quick extraction technique [4]. Next, we calculated the geometric features of the pockets. Finally, we visualized the distance data. We proposed three techniques to visualize the data. The first technique is polyline chart to find the preferred amino acid residue around druggable pocket. The second technique is scatterplots to find pairs of preferred amino acid residues around pockets. The third technique is matrices to summarize the druggability analysis applied to all the possible pairs of amino acid types in natural protein.

II. RELATED WORK

A. Pocket discovery

As discussed in Section 1, pocket discovery has been an important topic for protein druggability analysis. Many techniques have been proposed, as surveyed in [6], and they are roughly categorized into energy- and geometry-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. [3] presented a technique which discovers concave portions of protein surfaces by rolling two sizes of spheres on them. This approach is superior at the point that can operate a parameter intuitively. Halgren [1] presented another effective technique which generates grid points surrounding the protein 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

1550-6037/14 $31.00 © 2014 IEEE
DOI 10.1109/IV.2014.38

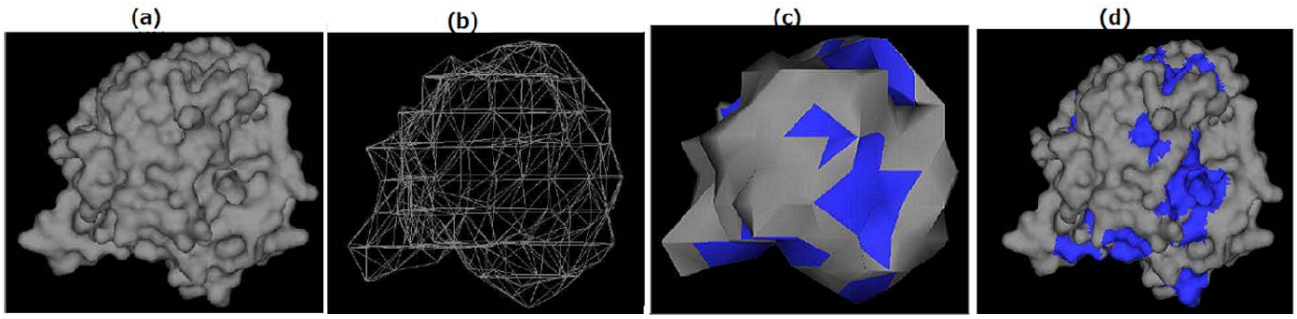

Figure 1. Example of a pocket extraction process. (a) Original protein surface. (b) Mesh simplification result. (c) Pocket candidate extraction result on (b). (d) Pocket projection result on the original mesh. Pockets are painted in blue.

analysis method [5] which applies mesh simplification; however, it did not focus on discovery of pockets and binding sites.

B. Interaction between amino acids and small molecules

The biological functions of proteins are now actively researched. Protein function is based on interactions between proteins and other molecules. The study in [7] surveyed interactions between proteins and small molecules. They implemented database named Het-PDB Navi [9]. They found preferred amino acid residues at the interaction sites of small molecules on the surface of protein. This study analyzes interaction between amino acid residues and small molecules; however, they did not analyze interactions between amino acid residues and pockets.

III. PROPOSED VISUAL ANALYTICS

We conjecture that distances between pockets and amino acids can be fruitful information to examine druggability from the above mentioned studies. Therefore, we propose a visual analytics scheme for the analysis of distances between pockets and amino acid residues in this paper. Our goal in this study is to discover a relationship between amino acid residue and druggability by the visual analytics.

This section describes our implementation of the visual analytics. Section A introduces the protein surfaces, and then Section B introduces the pocket extraction technique. Sections C to F propose the procedure of the visual analytics.

A. Protein surfaces

Our implementation applies protein surface datasets downloaded from the protein surface database “eF-site" [8]. This database collects the surfaces of proteins registered in PDB (Protein DataBank), by applying a Colony surface extraction technique [10]. 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.

B. 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. Fig. 1 shows the process flow of this technique applied to coagulation factor XA(PDB ID:1ezq).

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.

C. Distance calculation

This technique calculates the geometric features of the pockets by the following procedure. Firstly, the technique specifies the plane (P_o) that minimizes the sum of distances from vertices of the outer loop of a pocket. Secondly, it calculates the center position (Cp_o) and normal vector (Np_o) of the plane. Thirdly, this technique calculates the distance from vertices of the pocket to P_o, and identifies the deepest point of the pocket as the vertex which has the longest distance. We currently define the following two types of distances between a pocket and an amino acid in this study:

Distance 1 (d_1) is defined as the smallest distance between the vertices on the outer loop of the pocket and the atoms belonging to the amino acid residue.

Distance 2 (d_2) is defined as the smallest distance between the deepest point of the pocket and the atoms belonging to the amino acid residue.

Here, it is not mandatory to limit the distances as the two types above. We would further like to test different distance

calculation methods and find the appropriate definition in our future work.

D. *Visualization (1): Polyline chart for distance between pockets and amino acids*

We visualized the number of pockets close to the particular pockets by polyline charts, to find the preferred amino acid residue around druggable pockets. Here, we counted the number of druggable and undruggable pockets which are close to the particular pockets. In our visualization, the X-axis denotes the amino acid residues, and the Y-axis denotes the numbers of pockets. Colors of polylines represent the druggability of the pockets.

E. *Visualization (2): Scatterplots for distance between pockets and pairs of amino acids*

We then observed pairs of preferred amino acid residue around pockets. We applied the implementation of scatterplot-based visualization tool [2] to the observation of distances between the pockets and pairs of amino acid residue, by assigning the two distances to X- and Y- axes of the scatterplots.

Fig. 2 shows a snapshot of the visualization tool. The left part of the window shows a scatterplot in which each axis is assigned with one of the feature values of protein pockets. The features can be interactively selected by touching buttons placed on the right part of the window. Plots corresponding to “druggable" protein pockets are colored in red, those corresponding to “undruggable" in blue, those to “nearly undruggable (difficult)" in green, and otherwise in gray. A click on a plot in the scatterplot enables visualization of the corresponding pocket on the protein surface colored in blue at the center of the window.

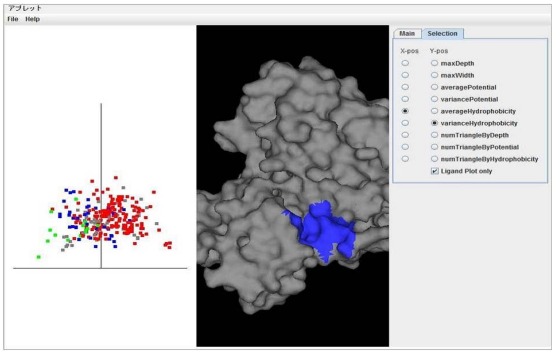

Figure 2. Snapshot of the visual analytics tool. The left side of the window shows a scatterplot for feature values of pockets of a protein. By specifying one of the plots, the corresponding pocket is colored in blue on the protein surface displayed at the center of the window.

F. *Visualization (3): Matrices for distance between pockets and pairs of amino acids*

We summarized the druggability analysis applied to all the possible pairs of amino acid types in natural proteins. In our implementation, 20 types of amino acids are arranged in both horizontally and vertically in the same order in a lower triangular matrix. Therefore, each element of the matrix denotes a value for a pair of amino acid type. We counted the numbers of pockets which are close to the particular pairs of amino acids. The total number of druggable pockets equals N_d and the number of pockets equals N_{all}. N_d divided by N_{all} equals ds. The element was colored in red if ds was larger than the predefined threshold value D, and it was colored in blue if ds was smaller than D. The brightness of each color was determined by the following S;

$$S = | ds - D | \times N_{all} \times \alpha \quad (1)$$

The result demonstrates that the druggable pockets tend to be close to specific pairs of amino acid types.

IV. RESULTS

In this section, we present examples of visualizing datasets using our implementations described above. Here, we prepared the following two datasets for the experiments:
Distance 1: A set of 31 proteins of which druggability was examined as druggable on SuperTarget [12].
Distance 2: A set of 60 proteins of which druggability was examined by Halgren [1].

Protein datasets in PDB format often contain records of “HETATM" which describe the coordinates of non-protein atoms/molecules in a 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 likely has specificity to a certain molecule and we name the pocket “reactive." A pocket without a molecule is hence named “non-reactive." We believe that the distinction can be a good indicator for druggable and non-druggable pockets.

In the experiments introduced in this section, we extracted pockets by a quick pocket extraction technique [4] applied on protein surface data, searched for non-protein atoms/molecules around the extracted pockets, and finally determined the reactivity of the pockets.

A. *Example of visualization (1)*

We visualized the distribution of amino acid residue around the pockets by polyline charts as described in Section 3.4. Fig. 3(left) shows the result applied to d_1 of Dataset 2, and Fig. 3(right) shows the result applied to d_2 of Dataset 2. We can observe little difference in the number of “undruggable" and “difficult" pockets. On the other hand, we can observe a certain difference in the number of “druggable" pockets. For example, Fig. 3(right) depicts that the number of pockets that taken into account the amino acid residue content around tryptophan is more than 3.5, while the number of pockets that taken into account the amino acid residue content around arginine is about 0.7.

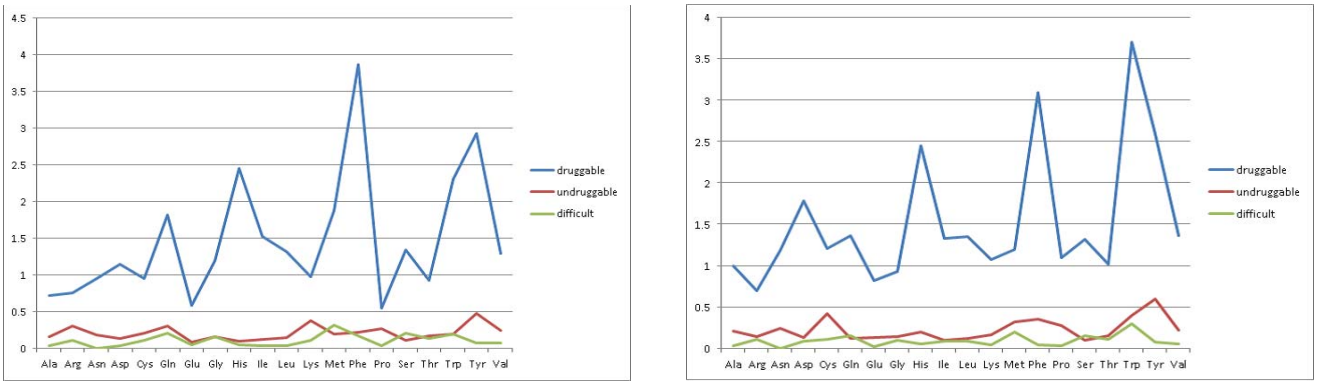

Figure 3. Polyline charts representing the number of pockets around particular amino acids. The X-axis denotes the amino acid residues, and the Y-axis denotes the number of pockets that taken into account the amino acid residue content. Colors of polylines represent the druggability of pockets. (Left) Distribution of d_1 with the Dataset 2. (Right) Distribution of d_2 with the Dataset2.

B. Example of visualization (2)

Next, we visualized the distribution of the distances between the pockets and the pairs of amino acid residue, by assigning two distances to X- and Y- axes of the scatterplot, as described in Section E. Here, we shaded the first quadrant, because the quadrant corresponds to the region where two amino acid residues are far from the pocket, and hence is not the target of our study.

In Fig. 4(left), the X-axis denotes the distance between the deepest point of the pocket and tyrosine residues, while the Y-axis denotes the distance between the deepest point of the pocket and valine residue. Unfortunately, we could not discover strong correlations between tyrosine and valine.

In Fig. 4(right), the X-axis denotes the distance between the deepest point of the pocket and asparagine residue, and the Y-axis denotes the distance between the deepest point of the pocket and tyrosine residue. The result clearly showed that the pockets with both asparagine and tyrosine residues nearby tend to be druggable. This result let us build a testable hypothesis that a deepest point of the pocket with both asparagine and tyrosine nearby are highly likely to be druggable.

C. Example of visualization (3)

We visualized the distribution of the pockets around pairs of amino acid residue as triangular matrices, as described in Section F. Fig. 5(left) shows the result applied to d_1 of dataset 2. Fig. 5(right) shows the result applied to d_2 of dataset 2. Fig. 6(left) shows the result applied to d_1 of dataset 1. Fig. 6 (right) shows the result applied to d_1 of dataset 1. These results demonstrate that the druggable pockets tend to be close to specific pairs of amino acid types. This result let us postulate that a pocket near a certain pair of amino acid types should be a good candidate for a reactive pocket, and this knowledge can be utilized for druggability prediction of the protein.

V. CONCLUSION

This paper presented visualization techniques to discover the relationship between amino acid residue and druggability. We extracted pockets from protein surfaces, and then we calculate the distances between pockets and amino acid residue. We, then, visualized the relationship between amino acid residue and druggability of pockets.

This paper introduced three types of visualization. The first one applied polyline charts to represent the number of pockets in consideration of the amino acid content. We found a certain amino acid residue are close to druggable pockets. The second one applied scatterplots to represent the distribution of the pairs of amino acid residues around pockets. We found certain pairs of amino acid residues tend to be close to druggable pockets. The third one applied triangular matrices to represent the statistics of the druggable pockets close to all the possible pairs of amino acid residue. This visualization result demonstrated that druggable pockets tend to be close to specific pairs of amino acid types.

Our potential future work is as follows. We would like to apply our technique to other datasets. Briefly, we would like to examine the amino acid types that constitute druggable pockets for drug compounds effective in each organ of human body. In addition, we would like to test various distance calculation methods and find appropriate definition of the distance. During this test we would like to observe whole amino acid types of pockets. We expect that the knowledge obtained in these the observations will contribute to drug development.

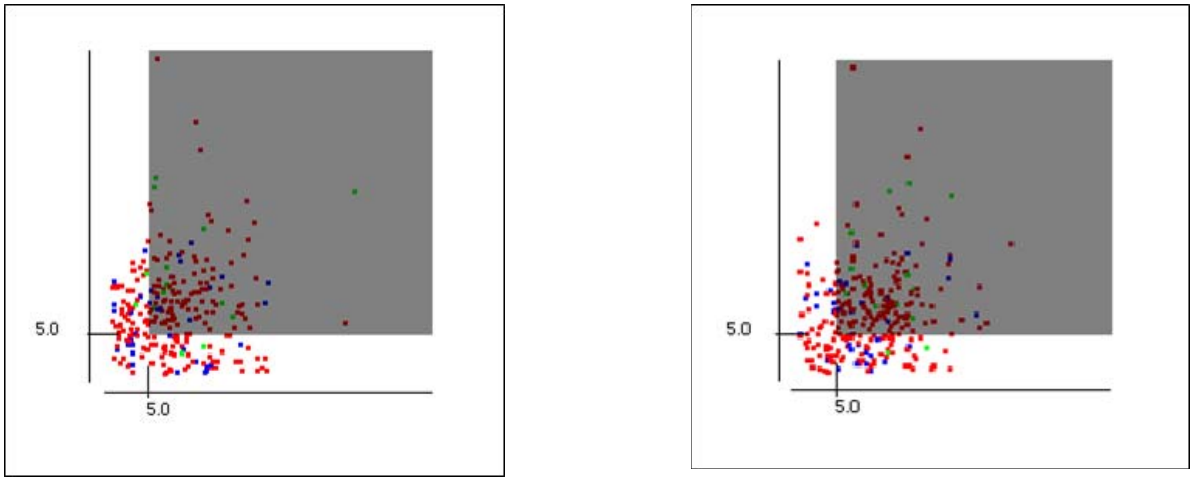

Figure 4. (Left) The X-axis is the distance between the bottom of a pocket and the closet Tyr residue, and the Y-axis is the distance between the pocket and Val residue. (Right) The X-axis is the distance between the bottom of a pocket and the closet Asp residue, and the Y-axis is the distance between the pocket and Tyr residue. Most of the pockets in proximity to both Asp and Tyr residues are druggable.

REFERENCES

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

[2] A. Kaneko, Y. Nakamura, T. Itoh, Visualization for Druggability Analysis of Protein Pockets, *NICOGRAPH 2012*, 1-8, 2012. (in Japanese)

[3] 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.

[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] 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.

[6] 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.

[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(1), 79-84, 2004.

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

[9] Het-PDBNavi.,
http://hetpdbnavi.nagahama-i-bio.ac.jp/index.php

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

[11] Protein DataBank,
http://www.rcsb.org/pdb/home/home.do/

[12] SuperTarget,
http://www.insilico.charite.de/supertarget/

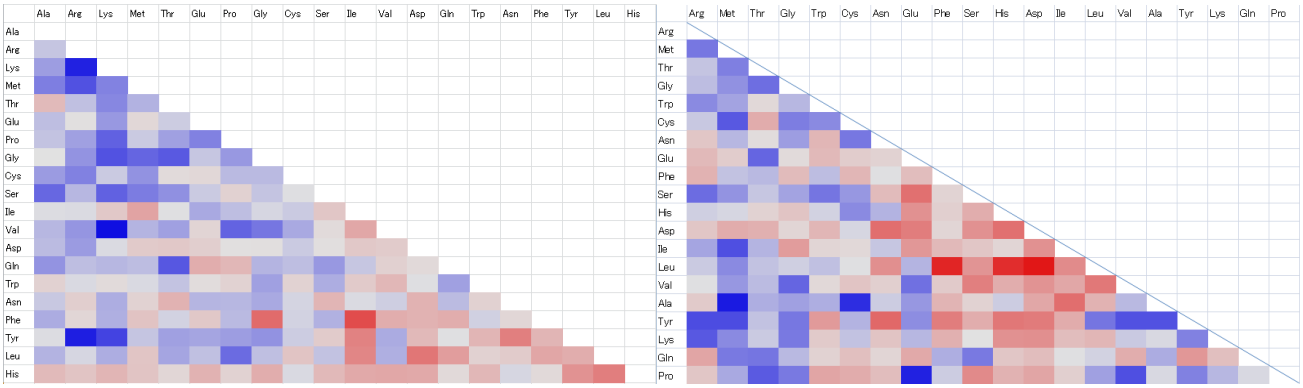

Figure 5. Representation of the druggability analysis in a lower triangular matrix. Both horizontal and vertical axes denote types of amino acid residue lined in the same order. Colors of the columns denote the druggability of pockets close to a pair of amino acids. Redness denotes the high druggability, blueness denotes the low druggability, and saturation denotes the number of corresponding pockets.

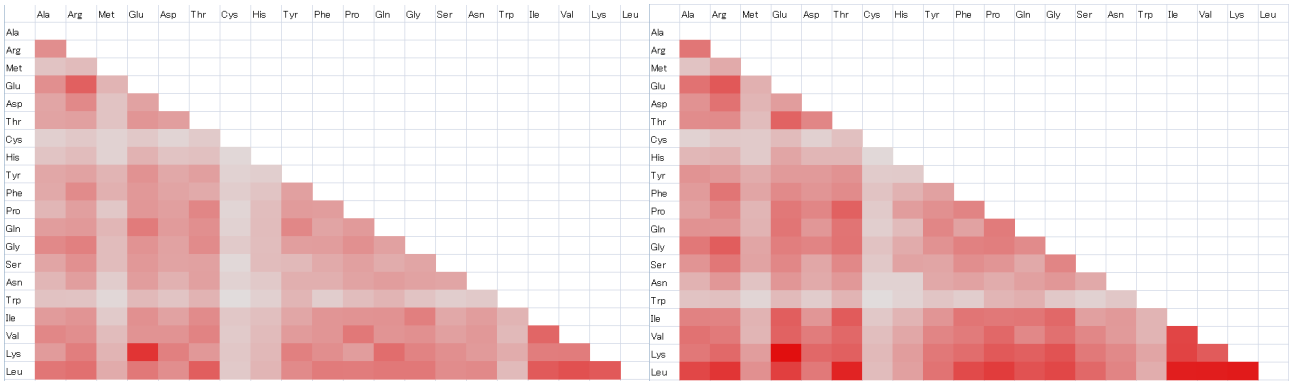

Figure 6. Representation of the druggability analysis. Matrix is similarly structured as Fig. 5; however, this result does not represent the lowness of the druggability. Redness denotes the high druggability, and saturation denotes the number of corresponding pockets.