Estimation of important binding sites in compounds that interact with proteins

Tachibana, Kouhei
Fukazawa, Aoi
Nishide, Ryo
Ohkawa, Takenao

(Citation)
Computational Biology and Chemistry,93:107511

(Issue Date)
2021-08

(Resource Type)
journal article

(Version)
Accepted Manuscript

(URL)
https://hdl.handle.net/20.500.14094/90008514

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Computational Biology and Chemistry 93 (2021) 1–11

Estimation of important binding sites in compounds that interact with proteins

Kouhei Tachibana[a]

Aoi Fukazawa[a], Ryo Nishide[b], Takenao Ohkawa[a]

[a]Department of Information and Intelligence Engineering, Kobe University, Japan
[b]The Center for Data Science Education and Research, Shiga University, Japan

Abstract

Proteins are one of the important substances in understanding biological activity, and many of them express the function by binding to other proteins or small molecules (ligands) on the molecular surface. This interaction often occurs in the hollows (pockets) on the molecular surface of the protein. It is known that when pockets are similar in structure and physical properties, they are likely to express similar functions and to bind similar ligands. Therefore, exploring the similarity of the structure and physical properties in pockets is very useful because it leads to the discovery of new ligands that are likely to bind. In addition, exploring the important structure when binding to the protein significant spot in the ligand will provide useful knowledge for the development of new ligands.

In this study, we propose a method to search for proteins containing pockets that are structurally and physically similar to significant spot in the pocket of the analyzed protein, and to extract significant spots in the ligands that bind to them. We use feature points as data. Feature points are the 3-dimensional points that are extracted from 3D structure data of proteins with feature values quantifying hydrophobicity and electrostatic potential. The corresponding feature points are extracted by comparing structurally and physically the pockets of the search target proteins with the significant spot of the analyzed protein. By evaluating the similarity based on the comparison results of the feature values given to the extracted feature points, we search for proteins that are similar to the analyzed protein. From the ligands that bind to the searched proteins, atoms that are near the protein pocket and similar to the atoms in ligand binding to the analyzed protein are extracted. The site constituted by the extracted atoms is defined as a significant spot in the ligand.

As a result of classifying ligands binding to the protein by using the extracted significant spot in the ligand, the effectiveness of the proposed method was confirmed.

Keywords:
protein
ligand
important binding site

1. Introduction

Proteins are major substances that make up the human body and play an important role in life activities. It is known that this protein exerts its function by binding to low molecular weight compounds (hereinafter called ligands) and other proteins. Therefore, it is especially important to know what kind of substance a protein binds to and how it binds in order to know the function of protein expression.

Information on the atoms that make up the protein is stored in a database called the PDB[1] which stores a 3D structure of a protein, and it is possible to know the 3D coordinate information of an atom and the ligand to which it binds. In addition, information on the molecular surface of proteins is stored in a database called eF-site[2], and it is possible to obtain 3D polygon data having physical property information such as electrostatic potential and hydrophobicity of the molecular surface. By using this information, proteins can be treated as 3D point cloud data. In addition, data on ligands that bind to proteins are also stored in the PDB, and it is possible to know the types of atoms that make up the ligand and 3D coordinate information. By using the information obtained from each database, it is possible to investigate the binding state of protein and ligand.

The binding of a protein and a ligand is often carried out mainly at a recessed local site called a pocket on the molecular surface of the protein. The part of the pocket that is considered to be deeply involved in binding is called the protein significant binding site in this paper. The structure and properties of significant binding sites of proteins are especially important for investigating the binding state with ligands and other proteins.

In addition, it is known that proteins with pockets with similar structures and properties are likely to bind with similar ligands. In the ligand, the part that is deeply involved in binding is called the significant binding site of the ligand. By investigating the significant binding sites in a ligand, it is possible to narrow down the structures in which the ligand binding to the protein is likely to have. In this way, investigating important binding sites of ligands leads to useful knowledge for discovering newly bound ligands.

Therefore, in this study, we propose a method to search for proteins which have a highly similar pockets to the analyzed target protein (hereinafter referred to as query protein) and to extract important binding sites from the ligands that bind to those proteins. Since the important binding site of a protein is deeply involved in the binding, it is obvious that a ligand binding to a protein with a local site that has a high similarity to the important binding site of the protein is likely to be similar to a ligand that binds to a protein that has the important binding site. In other words, a ligand that binds to a protein with a site that has a high similarity to the important binding site of the query protein is likely to bind to the query protein. Therefore, the important binding site of the query protein is compared with the pockets of multiple proteins to search for proteins with highly similar pockets. When comparing, in the molecular surface data of protein obtained from eF-site, the points that represent the characteristic shape are used as the characteristic point data. This feature point is expressed by the feature value digitizing coordinate information obtained from the relationship with neighboring points and physical property information such as electrostatic potential and hydrophobicity. We aim to compare the important binding sites and pockets of the protein represented by this feature point data. Structure comparison is conducted exhaustively matching using 3D coordinate information, and property comparison is performed by comparing feature quantities. From the results of these two comparisons, we search for proteins with highly similar pockets.

It can be stated that the ligand that binds to the searched protein has a high possibility of binding to the query protein. By extracting important binding sites that are deeply involved in binding from these ligands, we can obtain structural candidates in which a ligand that binds to the query protein is likely to have structures. Since the binding important site of the ligand is a site which is deeply involved in the binding with the protein, it is considered that the site exists near the protein pocket in the bound state. Such sites can be excised by extracting the atoms of the ligand within a threshold distance from the feature points that make up the protein pocket. In addition, since it is known that ligands that bind to the same protein have a high degree of similarity, it is considered that the important binding site of the ligand has a high degree of similarity to the ligand that binds to the query protein. Such sites can be cut out by comparing the types and structures of atoms by matching. In other words, the site existing close to the protein pocket in the bound state and the one which is composed of atoms with high similarity to the ligand that binds to the query protein is extracted as the important binding site.

In addition, the important binding site of the ligand extracted by the proposed method is considered to be the structure which is possessed by most of the ligands that bind to the query protein. Based on this assumption, we also propose a method to determine whether a given ligand binds to the query protein by utilizing the important binding sites of the ligand. This attempt is used to evaluate the effectiveness of the proposed method.

The structure of this paper is as follows. In Section 2, we propose a method for estimating the important binding site of a ligand using the similarity of proteins, and in Section 3, we demonstrate the usefulness of the proposed method by experiments. And in Section 4 we will draw conclusions.

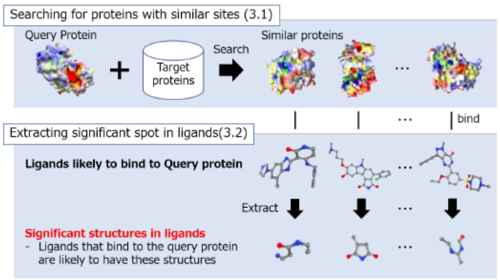

Searching for proteins with similar sites (3.1)
Query Protein
+
Target proteins
Search
Similar proteins
…
…
bind
Extracting significant spot in ligands(3.2)
Ligands likely to bind to Query protein
…
Extract
…
Significant structures in ligands
- Ligands that bind to the query protein are likely to have these structures
…

Figure 1. : Proposed method

2. Estimation of significant binding site using protein similarity in ligand and evaluation of the estimation method

When the protein pockets are structurally and physically similar, it is said that the binding ligands are also similar. In other words, a ligand that binds to a protein having a pocket which is similar to the pocket of the query protein is likely to bind to the query protein. Therefore, by investigating the important binding sites of these ligands, we can narrow down the candidate structures that most of the ligands that bind to the query protein have. In this way, narrowing down the candidates for the structure that a ligand binding to a protein will lead to another useful knowledge when developing a new binding ligand.

Based on the above, we propose a method for searching proteins with similar pockets and extracting the important binding sites of the ligands that bind to the protein. In addition, the important binding site of the extracted ligand is considered to be the structure possessed by most of the ligands that bind to the query protein. Therefore, we will search for ligands that are likely to bind to the query protein using the important binding sites of the ligand extracted by the proposed method.

Figure 1 shows the flow of the method proposed in Section 2. In Sections 2.1 and 2.2, we explain a method for estimating the important binding site in a ligand that utilizes the similarity of proteins. Section 2.1, describes a method for searching for proteins with pockets composed of feature points that are similar to the important binding sites of the query protein. Section 2.2, describes a method for extracting the important binding site of the ligand that binds to the searched protein. In Section 2.3, we discuss a method to search for a ligand that is likely to bind to the query protein, using the important binding site of the ligand estimated by the proposed method described in Sections 3.1 and 3.2.

2.1. Search for proteins with similar local sites

Important binding sites in proteins are composed of characteristic points that are deeply involved in binding with other substances. Therefore, it is considered that a protein having many feature points in the pocket that are structurally and physically similar to the feature points which constitute the important binding site is a highly similar protein. From this, in order to search for proteins with similar pockets, the binding important site of the query protein and the pocket of the protein to be compared are structurally and physically investigated to evaluate the similarity.

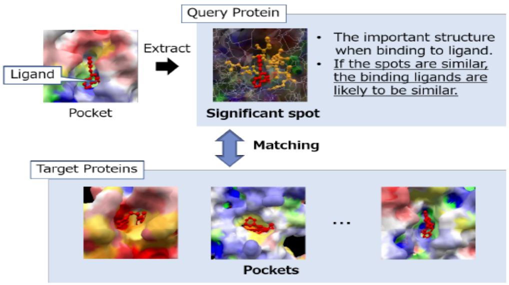

Query Protein
Ligand
Pocket
Extract
The important structure when binding to ligand.
If the spots are similar, the binding ligands are likely to be similar.
Significant spot
Matching
Target Proteins
…
Pockets

Figure 2. : Matching significant spot on query protein and pockets in target proteins

2.1.1. Extraction of local site by matching with important binding site

The binding important site of the query protein and the pocket of the protein (hereinafter referred to as the comparison target protein) are compared, in order to extract feature points that are structually and physically similar to the feature points that constitute the binding important site of the query protein. An image diagram of this extraction method is shown in Figure 2.

First, the structural comparison is performed by exhaustive matching using coordinate information, and then, physical properties are compared using the feature quantities assigned to the feature points that have been determined as structurally similar.

The current input data are the feature point data that make up the important binding sites of the query protein and the feature point data that make up the pocket of the comparison target protein. Also, the output is a match point list in which the IDs of feature point pairs associated by comparing the binding important site of the query protein and the pocket of the comparison target protein are described.

2.1.2. Structure and physical property comparison of matched feature point pairs

We examine in more detail the similarity with the important binding site of the query protein regarding the local site extracted by matching. The feature quantity has both positional information including information around the feature point and physical property information such as electrostatic potential and hydrophobicity. Therefore, the feature quantities that compose each site are compared, in order to investigate the similarity between the important binding site of the query protein and the local site extracted by matching. This makes it possible to know how similar the extracted parts are in terms of structure and properties. The input data are the match point list outputted in 2.1.1 and the feature point data that make up the important binding sites of the query protein and the pocket of the comparison target protein.

The match point list contains the IDs of the feature point pairs associated by matching. We read the IDs of the feature points (query feature points) that make up the important binding site of the query protein and the feature points (target feature points) that make up the local site extracted from the comparison target protein. We search the feature points holding these IDs from the feature point data in which the ID, the 3D coordinates, and the H-dimensional feature quantity are described. Then, we create an array of H-dimensional feature quantities held by the searched feature points. We call the arrays that store the IDs of the associated query feature points and target feature points *QueryID[i]* and *TargetID[i]*. Also, the arrays of the feature quantities held by the query feature points and the target feature points are called *QueryFeature[i][j]* and *TargetFeature[i][j]*, respectively. Algorithm 1 shows the specific processing for creating an array of feature quantities.

Algorithm 1 Create array of feature value

```
Input: QueryID[i] ($0 \le i \le n$, $n$ is the number of match points)
Output: QueryFeature[i][j], TargetFeature[i][j] ($0 \le j \le H-1$)
for $i = 0$ to $n$ do
    if QueryID[i] = ID of the feature feature point that constitutes significant spot on query protein then
        for $j = 0$ to $H-1$ do
            QueryFeature[i][j] = $j$ th feature value of the feature point that constitutes important binding site on query protein
        end for
    end if
    if TargetID[i] = ID of the feature point that constitutes pocket of target protein then
        for $j = 0$ to $H-1$ do
            TargetFeature[i][j] = $j$ th feature value of the feature point that constitutes pocket of target protein
        end for
    end if
end for
```

Using the created array of feature quantities, we compare the feature quantities to examine the similarity of feature point pairs associated by matching. The features are H -dimensional histograms consisting of $n_{angle} \times n_{distA}$ -dimensional angle histogram, $n_{height} \times n_{distH}$ -dimensional height histogram, and $n_{esp} \times n_{dispP}$ -dimensional physical property histogram. Therefore, the Bhattacharyya coefficient, which represents the similarity between histograms, is used as an index that indicates the similarity of feature quantities. In other words, by calculating the Bhattacharyya coefficient for each of the angle histogram, height histogram, and physical property histogram that make up the feature quantity, we examine the similarity between the feature point pairs that make up the important binding site of the query protein and the pocket of the comparison target protein. Let $Bhattacharyya_{Angle}[i]$ be the Bhattacharyya coefficient for the angle histogram, $Bhattacharyya_{Height}[i]$ be the Bhattacharyya coefficient for the height histogram, and $Bhattacharyya_{Property}[i]$ be the Bhattacharyya coefficient for the physical property histogram ($0 \le i \le n-1, n$ is the number of matching points). The Bhattacharyya coefficient for each histogram is calculated by the following equation.

- Bhattacharyya coefficient for angle histogram: Use $0 \le j \le 49$ of $QueryFeature[j]$ and $TargetFeature[j]$.

$$Bhattacharyya_{Angle} = \sum_{j=0}^{49} \sqrt{QueryFeature[j] \times TargetFeature[j]} \tag{2.1}$$

$$\sum_{j=0}^{49} QueryFeature[j] = \sum_{j=0}^{49} TargetFeature[j] = 1 \tag{2.2}$$

- Bhattacharyya coefficient for height histogram: Use $50 \leq j \leq 109$ of $QueryFeature[j]$ and $TargetFeature[j]$.

$$Bhattacharyya_{Height} = \sum_{j=50}^{109} \sqrt{QueryFeature[j] \times TargetFeature[j]} \tag{2.3}$$

$$\sum_{j=50}^{109} QueryFeature[j] = \sum_{j=50}^{109} TargetFeature[j] = 1 \tag{2.4}$$

- Bhattacharyya coefficient for physical property histogram: Use $110 \leq j \leq 169$ of $QueryFeature[j]$ and $TargetFeature[j]$.

$$Bhattacharyya_{Property} = \sum_{j=110}^{169} \sqrt{QueryFeature[j] \times TargetFeature[j]} \tag{2.5}$$

$$\sum_{j=110}^{169} QueryFeature[j] = \sum_{j=110}^{169} TargetFeature[j] = 1 \tag{2.6}$$

The calculated Bhattacharyya coefficient takes a value from 0.0 to 1.0, and the greater the value is, the greater the similarity is. In addition, each histogram is normalized so that the total is 1. In this way, we calculate the similarity between the important binding sites of the query protein associated by matching and the feature point pairs composing the pocket of the comparison target protein.

2.1.3. Evaluation of similarity of extracted local parts

From the comparison results of feature point pairs, we evaluate the similarity between the important binding site of query protein and the pocket of the comparison target protein. Here, we define a new score by expressing the similarity as a numerical value in order to understand more easily. To calculate this score, we use the angle similarity, height similarity, and physical property similarity between the important binding site of the query protein and the local site in the pocket of the comparison target protein. These values show how similar the important binding site of the query protein and the local site in the pocket of the comparison target protein are in each item. By calculating these three similarities, it is possible to consider not the similarity for each feature point pair, but the structural/physical similarity between the important binding site of the query protein and the local site in the pocket of the comparison target protein. By calculating the score from these similarities, we evaluate the similarity between the important binding site of the query protein and the pocket of the comparison target protein.

We calculate the angle similarity, height similarity, and physical property similarity using the Bhattacharyya coefficient for each histogram calculated by comparing feature point pairs. The angle similarity, height similarity, and physical property similarity are obtained by dividing the cumulative value of each Bhattacharyya coefficient calculated for each feature point pair by the number of match points. In order to calculate these similarities, we use $Bhattacharyya_{Angle}[i]$, $Bhattacharyya_{Height}[i]$, $Bhattacharyya_{Property}[i]$ $(0 \leq i \leq n-1, n$ is the number of match points). Each similarity is calculated by the following equation.

- Angle similarity : we use $0 \leq i \leq n-1$ of $Bhattacharyya_{Angle}[i]$ (n is a match score).

$$Angle = \frac{\sum_{i=0}^{n-1} Bhattacharyya_{Angle}[i]}{n} \tag{2.7}$$

- Height similarity : we use $0 \leq i \leq n-1$ of $Bhattacharyya_{Height}[i]$ (n is a match score).

$$Height = \frac{\sum_{i=0}^{n-1} Bhattacharyya_{Height}[i]}{n} \tag{2.8}$$

- Physical property similarity : we use $0 \leq i \leq n-1$ of $Bhattacharyya_{Property}[i]$ (n is a match score).

$$Property = \frac{\sum_{i=0}^{n-1} Bhattacharyya_{Property}[i]}{n} \tag{2.9}$$

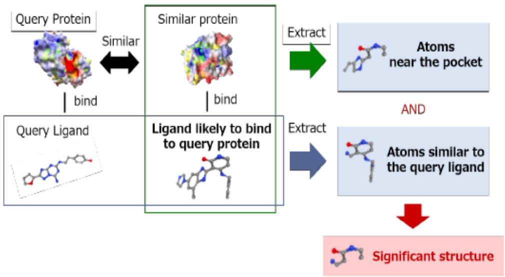


Figure 3. : The method of extracting significant structure in ligand likely to bind to the query ligand

We calculate the score by weighting each of these three similarities. If the number of feature point pairs associated by matching is too small, it is hard to say that the pocket of the comparison target protein is structurally and physically similar to the important binding site of the query protein. Therefore, we calculate a score for the comparison target proteins whose ratio of the number of match points based on the number of feature points that make up the important binding site of the query protein is S or more, and we set a score as 0 for the other comparison target proteins. Assuming the rate of the number of match points based on the number of feature points that make up the important binding site of the query protein is "*rate*", the equation for calculating the score is as follows.

$$Score = \begin{cases} Angle \times \omega_1 + Height \times \omega_2 + Property \times \omega_3 & (S \leq rate) \\ 0 \quad (S > rate) \end{cases} \tag{2.10}$$

Here, ω_1, ω_2, and ω_3 are weight parameters for Angle, Height, and Property, respectively. In this way, we calculate the score for the similarity between the important binding site of the query protein and the pocket of the comparison target protein.

2.2. Extraction of important binding sites in highly similar ligands

Proteins with similar pockets are said to bind similar ligands. Therefore, a ligand that binds to a protein that has a local site similar to the important binding site of the query protein is likely to bind to the query protein. Examining the site (important binding site) that is deeply involved in binding in such a ligand will lead to the investigation of candidates of the partial structure required for binding to the query protein. Since the important binding site of the ligand is deeply involved in the binding with the protein, it is considered that the site exists in the vicinity of the protein pocket when it is in the bound state. In addition, it is said that ligands binding to the same protein have similar structures and properties. We define a site composed of atoms existing close to the feature points that make up the protein pocket in the bound state and that are similar in structure and properties to the ligand (hereafter called the query ligand) that binds to the query protein as an important binding site in the ligand. Then, we extract such sites as important binding sites from the ligands (hereinafter referred to as binding candidate ligands) that bind to proteins to be judged as having high similarity to the query protein. Figure 3 shows an image of the method for extracting important binding sites from binding candidate ligands.

2.2.1. Extraction of binding site in ligand

A ligand binds by attracting a specific site on the protein. Therefore, in the bound state, the atoms in the vicinity of the feature points that make up the protein pocket are thought to be deeply involved in binding to the protein. The site composed of such atoms is called the binding site on the ligand. Then, we extract the atoms that make up the binding site of the ligand from the binding candidate ligand.

As input data for extraction, we use the 3D coordinate information of the feature points obtained from the feature point data that make up the pockets of the protein to have been judged as having high similarity to the query protein in Section 2.1. In addition, regarding the three-dimensional coordinates of the ligand atom, we use two kinds of information, the coordinate information written in the PDB file for the ligand and the coordinate information written in the protein PDB file. The two different sources of information are used because the coordinate information obtained from both are different. The former expresses the coordinate information about the ligand itself, while the latter expresses the coordinate information when it is bound to the protein. Therefore, the latter is used when investigating the positional relationship with the protein pocket, and the former, which is more general coordinate information, is used when extracting atoms as binding sites in the ligand.

We investigate the positional relationship using the 3D coordinate information on the protein feature points and ligand atoms obtained from these files. Let the 3D coordinates of the feature points that make up the protein pocket be (x_p, y_p, z_p) and the 3D coordinates of the atoms that make up the ligand be (x_l, y_l, z_l). We calculate the distance between the feature points that make up the protein pocket and the atoms that make up the ligand by the following equation.

$$d = \sqrt{(x_p - x_l)^2 + (y_p - y_l)^2 + (z_p - z_l)^2} \tag{2.11}$$

If the distance calculated by this equation is less than the threshold value *dis*, it is judged that the ligand atom exists near the pocket of the protein, and the information of the atom having the 3D coordinate information is registered in the list NeighborAtomList. The list NeighborAtomList is a list of atoms that make up the binding site of the ligand. We perform this calculation for all combinations of the feature points that make up the protein pocket and the atoms that make up the ligand. Also, an atom which binds to the atom registered in the list is also registered in the list NeighborAtomList as the atom that makes up the binding site in the ligand. This is because the atom registered in the list is likely to function as a functional group along with the atom that binds to it. In this way the site composed of the atoms registered in the list NeighborAtomList becomes the binding site of the ligand.

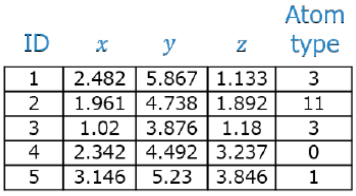
2.2.2. *Evaluation of ligand similarity by structure comparison*

Ligands that bind to highly similar proteins are considered to have high structural and property similarities. In other words, it can be said that the binding candidate ligand has a site that is structurally and physically similar to the query ligand. Therefore, we search and extract atoms that have high structural/property similarities to the atoms that make up the query ligand from the binding candidate ligands.

We conduct matching in order to do the structural and property ligands comparison. Here, as the data used for matching, we create and use new atomic data that holds the 3D coordinate information of the atoms that compose the ligand and the information about the types of atoms. We use the 3D coordinate information about the ligand atom in structural comparison. Regarding this 3D coordinate information, we use the coordinate information described in the PDB file for the ligand, that is, the coordinate information for the ligand alone. The types of atoms such as C (carbon) and N (nitrogen) represent the properties of the atoms. Therefore, if the atom types are the same, it is considered that they have similar functions in the ligand. Thus, we use information about the type of ligand atom for property comparison. The type of ligand atom is described in the mol2 file. A mol2 file is a file obtained by inputting the PDB file for the ligand used to obtain the 3D coordinate information into Open Babel[6]. Open Babel is software that converts between formats often used in the chemical field. To convert the types of ligand atoms described in this mol2 file into integer values as shown in Table 1 makes it easy to compare. Therefore, the input atomic data is shown as in Figure 4.

Table 1. Dataset for evaluation

Code	Atom type	Integer value
C.2	carbon sp2	0
O.2	oxygen sp2	1
O.3	oxygen sp3	2
C.3	carbon sp3	3
H	hydrogen	4
C.ar	carbon aromatic	5
N.pl3	nitrogen trigonal planar	6
C.cat	carbocation used only in a guadinium group	7
N.2	nitrogen sp2	8
O.co2	oxygen in carboxylate and phosphate groups	9
N.ar	nitrogen aromatic	10
N.am	nitrogen amide	11
C.1	carbon sp	12
S.O2	sulfone	13
N.3	nitrogen sp3	14
S.3	sulfur sp3	15
P.3	phosphorous sp3	16
S.2	sulfur sp2	17
F	fluorine	18
N.1	nitrogen	19
N.4	nitrogen sp3 positively charged	20



ID	x	y	z	Atom type
1	2.482	5.867	1.133	3
2	1.961	4.738	1.892	11
3	1.02	3.876	1.18	3
4	2.342	4.492	3.237	0
5	3.146	5.23	3.846	1

Figure 4. : Example of atom data

Each line shows information about one atom, and the contents are the atom ID, the 3D coordinates, and an integer value indicating the type of atom.

We perform matching about the query ligand and the binding candidate ligand using this atomic data. We compare the structures by repeating matching exhaustively using 3D coordinate information, as described in Section 2.1.1. Regarding the comparison of properties, if the integer values representing the kinds of atoms are similar, we judge to be similar,while if they are not similar, we judge not to be similar. In this way, we conduct matching by comparing the structure and properties. The output of this matching is a match point list in which the IDs of the atom pairs associated by comparing the query ligand and the binding candidate ligand are written. Those atoms of the binding candidate ligands with the atom IDs listed in this match point list are the atoms that have structual/property similarity to the atoms of the query ligand. Therefore, we extract the atoms of the binding candidate ligands with the atom IDs listed in the match point list from the atom data that is an input data, and register in the list MatchAtomList. This list MatchAtomList is a list of atoms that have

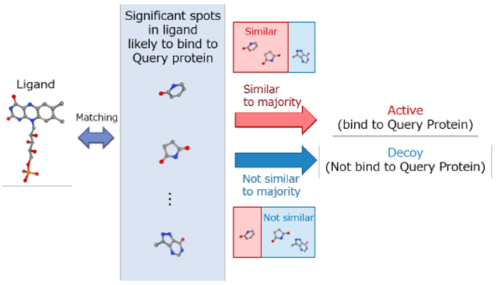


Figure 5. : Classification of ligands with estimated significant spot of ligand

structural/property similarity to the atoms that make up the query ligand.

2.2.3. *Extraction of important binding site in ligand*

The important binding site in the ligand is the site that constitutes the binding site in the ligand and is composed of atoms that are similar in structure and properties to the query ligand. Therefore, we select the atoms extracted in Section 2.2.1 and Section 2.2.2 as important binding sites in the binding candidate ligand. We use the list NeighborAtomList outputted in Section 2.2.1 and the list MatchAtomList outputted in Section 2.2.2 as input data. Both lists contain atoms that have 3D coordinate information about a single ligand. Therefore, if there is an atom whose coordinate information matches in both lists, it constitutes a binding site in the ligand, and it is an atom that is similar in structure and property to the query ligand. Thus, based on the 3D coordinate information, we search for the atoms commonly registered in these two sequences and output as the atoms that constitute the important binding site in the binding candidate ligand.

2.3. *Distinction of binding compounds using the important binding sites of the estimated ligand*

Let us consider a method of judging whether or not a given ligand binds to a query protein by using the important binding sites of the binding candidate ligand outputted by the proposed method. In the proposed method, we searched a protein with a pocket similar to the query protein, and we defined the ligand that binds to the searched protein as a binding candidate ligand that is likely to bind to the query protein. From these binding candidate ligands, we extracted important binding sites composed of atoms that are considered to be deeply involved in binding structurally and physically. The important binding site in the extracted binding candidate ligand is considered to be the site possessed by most of the ligands that bind to the query protein. In addition, this important binding site was extracted only the number of proteins judged to have pockets similar to the query protein. In other words, there are multiple structural candidates that the ligand binding to the query protein is likely to have. Based on these facts, if the given ligand has a structure similar to the majority of the important binding sites in the binding candidate ligand extracted by the proposed method, we judge that it binds to the query protein. The outline of this method is shown in Figure 5.

For comparison of the important binding sites in the given ligand and the binding candidate ligand, we do the same matching as described in Section 2.2.2. Using the atom data of both ligands as input, we attempt the structures and properties comparison, and a match point list for the associated atoms is obtained as output. When the ratio of the number of match points based on the number of atoms constituting the binding important site in the binding candidate ligand is greater than a certain value, the given ligand is considered to be similar to the binding important site in the binding candidate ligand. We will make this judgement for important binding sites in each binding candidate ligand. As a result, if it is similar to the majority of the important binding sites, it is judged to be an active compound (Active) that binds to the query protein, otherwise, it is judged as an inactive compound (Decoy). By such method, it enables to search for a ligand that binds to the query protein.

3. Evaluation Experiment, Result and Discussion

3.1. *Experimental settings*

3.1.1. *Experimental method*

Evaluation experiments are conducted to show the efficiency of the proposed method. In addition, docking simulation is conducted using the docking simulation tool AutoDockVina 1.1.2[8] (hereinafter referred to as ADV), and comparison with the experiment of the proposed method is performed. First, we explain the experiments of the proposed method. The structure of the ligand, which is important for binding to a certain pocket of the query protein, is extracted by the method proposed in 2.1 and 2.2. For this pocket, the one in which the information on the ligand to be bound registered in the PDB is selected. In addition, when multiple ligands that bind to the query protein are registered, important binding sites in the ligand are extracted for each pocket that binds each ligand. By using the important binding site in the extracted ligand, it can be determined whether the given ligand binds to the pocket of the query protein, in other words, whether it is active or inactive to the query protein. The results of this discriminatory process are evaluated.

Next, we explain the experiment using ADV. In ADV, when calculating whether or not a protein and a ligand are docked, an empirical score function based on the physicochemical relationship between molecules is used to evaluate the docking result, and a score representing the binding affinity is outputted. The ligands given by the scores are ranked, and those ligands at the top are judged to be active for the query protein.

3.1.2. *Data Set*

In both the proposed method and the ADV experiment, DUD-E[7] is used. DUD-E is a dataset of active and inactive compounds for each of the 102 proteins. 51 out of 102 proteins whose ligands did not span over multiple protein chains were used as query proteins in this experiment. The ligands known to bind to each of the 51 query proteins are shown in Table 2. Thus, the structures of the ligands that were important for binding to the 90 pockets to which these ligands bind were extracted. In addition, active and inactive compounds for each query protein were used as test compounds to distinguish whether they were active or inactive. Since the number of data of active and inactive compounds was very large, a total of 60 compounds including 30 active and 30 inactive compounds were used in this experiment.

3.1.3. *Parameter Settings*

The parameters of the proposed method in this experiment are set as follows.

- Angle histogram parameters: $n_{angle} = 5$, $n_{distA} = 10$.
- Height histogram parameters: $n_{height} = 6$, $n_{distH} = 10$.
- Property histogram parameters: $n_{esp} = 4$, $n_{hp} = 3$, $n_{distP} = 5$.
- Rate of match score when calculating score showing pocket similarity: $S = 1.0$

Table 2. Data set for evaluation

Protein	Ligand	Protein	Ligand
1b9v-A	NAG, RA2	2oj9-A	BMI
1bcd-A	FMS	2ojg-A	19A
1c8k-A	CPB, PLP	2owb-A	626, ACT
1d3g-A	ACT, BRE, DDQ, FMN, ORO	2oyu-P	BOG, HEM, IMS, NAG
1e66-A	HUX	2qd9-A	LGF
1j4h-A	SUB	2rgp-A	HYZ
1q4x-A	G24	2zdt-A	46C, GOL
1qw6-A	3AR, H4B, HEM	3bgs-A	DIH
1r9o-A	FLP, GOL, HEM	3biz-A	61E
1sj0-A	E4D	3bkl-A	ACT, FUC, GOL, KAW, NAG
1sqt-A	UI3	3bqd-A	DAY
1udt-A	VIA	3bwm-A	DNC, SAM
1uyg-A	PU2	3bz3-A	YAM
1vso-A	AT1, GOL	3cqw-A	CQW
1zw5-A	IPE, ZOL	3e37-B	ED5, SUC
2aa2-A	AS4, BOG, GOL	3eml-A	STE, ZMA
2am9-A	DTT, GOL, TES	3eqh-A	5BM, ADP
2ayw-A	BEN, GOL, MES, ONO	3g0e-A	B49
2azr-A	982	3hmm-A	855
2e1w-A	FR6	3krj-A	ACT, KRJ
2hv5-A	NAP, ZST	3l3m-A	A92
2i78-B	KIQ	3lq8-A	88Z
2ica-A	2IC	3m2w-A	L8I
2nnq-A	T4B	3nxo-A	D2B, NDP
2of2-A	547	3ny8-A	CLR, JRZ, OLA, OLC, PGE
2oi0-A	283	-	-

- Weighting parameter when calculating score showing pocket similarity: $\omega_1 = 0.3$, $\omega_2 = 0.2$, $\omega_3 = 0.5$
- A threshold for the distance between the feature points that make up the protein pocket and the atoms that make up the ligand: $dis = 10\text{Å}$

3.1.4. Evaluation value and comparison method

The AUC is calculated from the result of discrimination whether the given ligand is active or inactive, and the method is evaluated. AUC is the area under the ROC curve created from the true positive rate (ratio of active ligands correctly identified as active) and false positive rate (ratio of false ligands incorrectly classified as active). An AUC value of 1.0 indicates that the discrimination performance is high, 0.5 implies random discrimination, and 0.0 implies that the classification performance is low. This enables to evaluate the discrimination performance of the method. In the proposed method, as a result of matching of the given ligand and the binding important site of the ligand, if the ratio of the number of match points based on the number of atoms constituting the binding important site of the ligand turns into a certain value p or higher, it is judged that the given ligand is similar to the important binding site in the ligand. The AUC value is calculated by changing the threshold parameter p related to this ratio. In ADV, it is calculated whether or not a protein and a ligand are docked, and a score indicating the binding affinity is outputted. The ligands given by the scores are ranked, and the top ranked ligands are judged to be active for the query protein. Here, the AUC value in these methods can be calculated by changing the parameter that determines whether the top x% of the ranked ligands are active compounds.

3.2. Experimental Result

Table 3 shows the average AUC values when the proposed method and Autodock Vina 1.1.2., are used for the DUD-E.

Table 3. Comparisons of proposed method, and Autodock Vina1.1.2. on the DUD-E

Method	Average AUC
Proposed method	0.6078
AutodockVina1.1.2.	0.5892

From Table 3, the average AUC value of the proposed method is greater than that of Autodock Vina 1.1.2., and the results show significant discrimination performance.

Docking is a method for estimating the positional relationship of a protein and a ligand in a bound state on a computer in order to predict its binding affinity. On the other hand, the proposed method is a method of estimating active compounds by paying attention to the frequency of data. Thus, the experiment using the proposed method with a completely different approach, paved way to obtain results comparable to the generally used basic docking method. From this, it can be stated that investigating the important binding site in the ligand leads to obtaining important knowledge for searching the binding ligand. Table 4 shows the AUC values calculated from the results of distinguishing the ligands that bind to the pockets of the proteins in the dataset by the experiments using the proposed method and ADV. In the proposed method, the AUC value is calculated for each combination of protein and ligand in order to distinguish between active and inactive compounds based on the binding important site of each ligand that binds to a certain query protein, while in the case of ADV, the AUC value is obtained for each type of protein in order to determine the binding affinity by directly performing docking simulation with a certain protein and ligand which is an active compound or an inac-

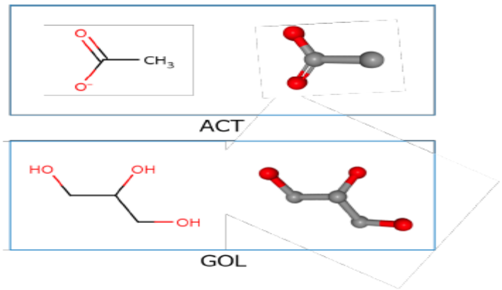


Figure 6. : Structures of ACT and GOL

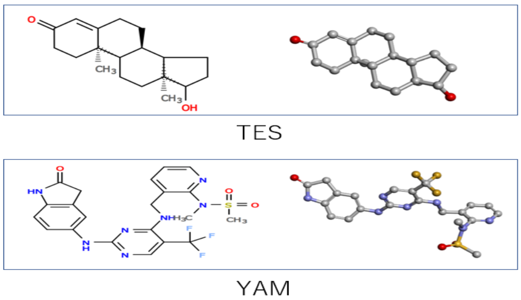


Figure 7. : Structures of TES and YAM

tive compound. Contemplating further improvements of the proposed method, in order to obtain better results than docking simulation tools with higher discrimination performance than ADV in the future, in the next section, we will discuss the results of the proposed method with much details.

3.3. Discussion

3.3.1. On ligand structure

In the pocket where the ligands ACT and GOL bind, all the ligands were judged to be inactive, and there were many cases where they could not be distinguished. Even in the pockets where discrimination was possible, the AUC was below 0.5. The reason for this is that ACT and GOL are very small molecules, as shown in Figure 6. If the query ligand is a small molecule, it will be difficult to capture the features when comparing the ligands with the method described in Section 2.2.2. Therefore, it is considered that the discrimination may not be performed correctly. In order to solve this problem, it is necessary to increase the features used for property comparison in the ligand matching as described in Section 2.2.2. In ligands matching, the property comparison is currently to check whether the atom types match. This is based on the idea that the property of the ligand can be expressed by the type of constituent atoms. In addition to that, however, better results may be obtained by using characteristics such as the type of functional group to which an atom belongs and the charge of the atom to compare properties. Thus, one improvement plan is to increase the features used to compare the properties of the ligands. Moreover, in the extraction of ligand-constituting atoms in the vicinity of the feature points that constitute the important binding site of the protein described in Section 2.2.1, it may be possible that adjusting the threshold of the distance regarded as the neighborhood can improve the discrimination performance.

On the other hand, the AUC value was 0.7 or more in the ligand-binding pockets such as TES and YAM, indicating that the discrimination performance was high. The reason for this is that TES and YAM are large molecules and have many structures that can be important binding sites. By having many important binding sites on the ligand side, it seems that it became easier to grasp the characteristics when comparing ligands. In addition, TES and YAM have a complicated structure, and when experiments are performed using AutoDock Vina, problems such as time consuming and high calculation cost may occur, so the proposed method can be said to be more useful. In the future, in order to further improve the accuracy of discrimination performance, it will be necessary to increase the characteristics used for property comparison even for large ligand molecules as well as for small molecules.

3.3.2. On protein type

CATH[9] is a database that classifies proteins based on their 3D structure. Table 5 shows the classification of proteins in the dataset by CATH.

Proteins with ∗ are called membrane proteins, which exist in cells and play important roles as sensors for nutrient transport and sensory organs. It has been difficult, however, to analyze their 3D structure, compared to water-soluble proteins. The discrimination performance in the pockets of such membrane proteins was low in the experiment.

However, with respect to proteins belonging to the type called Retinoid X Receptor, the AUC value of the discrimination result in most pockets were greater than 0.6. The average AUC value calculated for the retinoid X receptor alone is 0.726, which is higher than the average AUC value for ADV. These results indecate that the approach of the proposed method is useful for proteins belonging to the retinoid X receptor. In addition, regarding proteins belonging to the type called (phosphorylase kinase domain 1), 13 out of 20 have an AUC value of 0.5 or more. Thus, the dataset contained multiple proteins that belonged to the types such as retinoid X receptor and phosphorylase kinase, which could be said to have obtained significant results. Therefore, it may be necessary to expand the data set in order to conduct useful discrimination for proteins belonging to more types.

3.3.3. On important binding site of protein

The important binding site of a protein is composed of points that are matched by multiple proteins. Figure 8 shows a graph of the number of referenced proteins and the number of feature points that make up the important binding sites of proteins. The horizontal axis shows the number of referenced proteins, and the vertical axis shows the number of feature points that constitute the important binding site of the protein. The blue dots refer to proteins with AUC values greater than 0.5, and the orange dots refer to proteins with AUC values less than 0.5.

Figure 8 shows that significant results were not obtained for proteins with more than 15 reference proteins. This is probably because if the number of referenced proteins is too large, structures that are common to all proteins will be extracted. In addition, the AUC value tends to be less than 0.5 even when the number of feature points that constitute the important binding site is less than 10. This is probably because, when comparing the binding important site of a protein with the pockets of other proteins, if the number of feature points constituting the binding important site is small, it can be considered that there is high similarity to any pocket. In other words, when selecting the feature points that make up the important binding site of a protein, it is important that the number of reference proteins is not too large and the number of feature points is not too small. Therefore, the proposed method can be improved by changing the method of selecting the fea-

Table 4. Results by the proposed method and AutodockVina1.1.2.

Protein	Ligand	AUC(P.M)	AUC(ADV)	Protein	Ligand	AUC(P.M)	AUC(ADV)
1b9v-A	NAG	0.797	0.428	2oi0-A	283	0.768	0.647
	RA2	0.398		2oj9-A	BMI	0.718	0.737
1bcd-A	FMS	-	0.492	2ojg-A	19A	0.491	0.59
1c8k-A	CPB	0.74	0.636	2owb-A	626	0.529	0.54
	PLP	0.689			ACT	-	
1d3g-A	ACT	-	0.826	2oyu-P	BOG	0.349	0.524
	BRE	0.644			HEM	0.669	
	DDQ	0.38			IMS	0.552	
	FMN	0.458			NAG	0.34	
	ORO	0.76		2qd9-A	LGF	0.57	0.542
1e66-A	HUX	0.633	0.616	2rgp-A	HYZ	0.578	0.458
1j4h-A	SUB	0.661	0.523	2zdt-A	46C	0.823	0.644
1q4x-A	G24	0.763	0.441		GOL	-	
1qw6-A	3AR	0.443	0.573	3bgs-A	DIH	0.397	0.629
	H4B	0.564		3biz-A	61E	0.755	0.667
	HEM	0.583		3bkl-A	ACT	-	0.47
1r9o-A	FLP	0.714	0.614		FUC	0.457	
	GOL	0.366			GOL	-	
	HEM	0.509			KAW	0.535	
1sj0-A	E4D	0.776	0.618		NAG	0.696	
1sqt-A	UI3	0.697	0.733	3bqd-A	DAY	0.606	0.621
1udt-A	VIA	0.744	0.604	3bwm-A	DNC	0.917	0.467
1uyg-A	PU2	0.538	0.291		SAM	0.38	
1vso-A	AT1	0.449	0.536	3bz3-A	YAM	0.826	0.691
	GOL	-		3cqw-A	CQW	0.789	0.779
1zw5-A	IPE	-	0.198	3e37-B	ED5	0.678	0.762
	ZOL	-			SUC	0.413	
2aa2-A	AS4	0.652	0.482	3eml-A	STE	0.384	0.563
	BOG	0.818			ZMA	0.527	
	GOL	-		3eqh-A	5BM	0.743	0.529
2am9-A	DTT	0.779	0.79		ADP	0.472	
	GOL	0.521		3g0e-A	B49	0.473	0.592
	TES	0.89		3hmm-A	855	0.833	0.652
2ayw-A	BEN	0.582	0.64	3krj-A	ACT	-	0.53
	GOL	0.371			KRJ	0.524	
	MES	0.411		3l3m-A	A92	0.756	0.771
	ONO	0.696		3lq8-A	88Z	0.593	0.638
2azr-A	982	0.592	0.693	3m2w-A	L8I	0.642	0.769
2e1w-A	FR6	0.501	0.493	3nxo-A	D2B	0.787	0.693
2hv5-A	NAP	0.487	0.612		NDP	0.601	
	ZST	0.523		3ny8-A	CLR	0.592	0.547
2i78-B	KIQ	0.632	0.529		JRZ	0.878	
2ica-A	2IC	0.547	0.569		OLA	-	
2nnq-A	T4B	0.758	0.542		OLC	0.582	
2of2-A	547	0.482	0.55		PGE	0.636	

Table 5. Classification by CATH

Superfamily	Protein
3′5′-cyclic nucleotide phosphodiesterase, catalytic domain	1udt-A
Aldolase class I hline Alpha carbonic anhydrase hline Alpha/Beta hydrolase fold, catalytic domain	1d3g-A 1bcd-A 1e66-A, 2i78-B
Calycin beta-barrel core domain	2nnq-A
Chitinase A; domain 3	1j4h-A
Collagenase (Catalytic Domain)	2oi0-A
Cytochrome P450	1r9o-A
Dihydrofolate Reductase, subunit A	3nxo-A
Farnesyl Diphosphate Synthase	1zw5-A
Glycogen Phosphorylase B	1c8k-A
Glycosyltransferase	3e37-A
Histidine kinase-like ATPase, C-terminal domain	1uyg-A
Laminin	2oyu-P
Metal-dependent hydrolases	2e1w-A
NADP-dependent oxidoreductase domain	2hv5-A
Neuraminidase	1b9v-A
Nitric Oxide Synthase; Chain A, domain 1	1qw6-A
Periplasmic binding protein-like II	1vso-A^*
Phosphorylase Kinase; domain 1	2of2-A, 2oj9-A, 2ojg-A, 2owb-A, 2qd9-A, 2rgp-A, 2zdt-A, 3biz-A, 3bz3-A, 3cqw-A, 3eqh-A, 3g0e-A, 3hmm-A, 3krj-A, 3lq8-A, 3m2w-A
Poly(ADP-ribose) polymerase, regulatory domain	3l3m-A
Protein tyrosine phosphatase superfamily	2azr-A
Retinoid X Receptor	1q4x-A, 1sj0-A, 2aa2-A, 2am9-A, 3bqd-A
Rhodopsin 7-helix transmembrane proteins	3eml-A^*, 3ny8-A^*
Trypsin-like serine proteases	1sqt-A, 2ayw-A
Vaccinia Virus protein VP39	3bwm-A
von Willebrand factor, type A domain	2ica-A

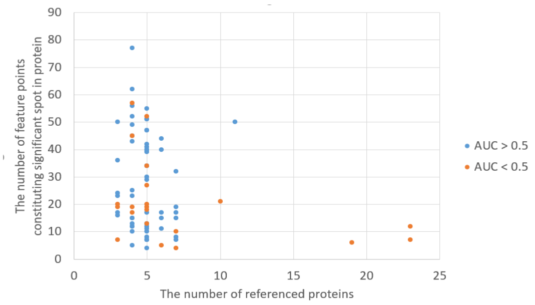


Figure 8. : The number of referenced proteins and the number of feature points constituting significant spot in protein

ture points that make up the important binding site of the protein so as to satisfy this condition.

4. Conclusion

In this paper, we proposed a method to search for proteins that are highly similar to query proteins and to extract important binding sites of ligands that bind to them. In the proposed method, the binding important site of the query protein and the pocket of the comparison target protein are compared by matching, and we evaluate the similarity using the feature quantity for the associated sites. From the ligands that bind to highly similar proteins searched in this way, by extracting the atoms that are close to the feature points that make up the protein pocket and that are highly similar to the query ligand, the important binding site in the ligand is identified. The important binding site of the extracted ligand is considered to be the structure possessed by most of the ligands that bind to the query protein. Based on this idea, we also investigated a method to distinguish ligands that are likely to bind to query proteins using the important binding sites of the ligand extracted by the proposed method.

In order to confirm the usefulness of the proposed method, we conducted an experiment to distinguish whether the ligand given by the important binding site in the extracted ligand is active or inactive. The docking method called AutodockVina 1.1.2. was used as a comparison method. As a result, the average AUC value of the proposed method was significant than that of AutodockVina 1.1.2., which indicates that the proposed method has discrimination performance comparable to the basic docking method.

Future challenges include increasing the features used to compare the properties of ligands and expanding the dataset to increase the number of protein types. Another modification is to change the method of selecting the feature points that make up the important binding site of the protein. By extending the research in such respects, it can be expected that the discrimination performance of active compounds by the important binding sites of the ligand extracted by the proposed method will be improved. This will lead to gaining useful knowledge when developing new ligands.

Acknowledgments

The authors thank the editors and the anonymous reviewers for their helpful comments and suggestions on the quality improvement of our paper.

References

[1] Berman, H.M., Westbrook, J., Feng, Z. Gilliland, G. Bhat, T.N. Weissig, H. Shindyalov, I.N. Bourne, P.E.: ”The Protein Data Bank”. Nucleic Acids Research Vol. 28, 235-242 (2000)

[2] Kinoshita, K., Nakamura, H.: eF-site and PDBjViewer: database and viewer for protein functional sites.. Bioinformatics Vol. 20, 1329-1330 (2004)

[3] Fukazawa, A., Tamori, R., Nishide, R., Ohkawa, T.: Extraction of Feature Values Using Protein Molecular Surface Data and Prediction of Significant Spot for Preferential Binding of Ligands. Proceedings of the 7th IIAE International Conference on Intelligent Systems and Image Processing. po.271-277 (2019).

[4] Nishimura, H., Ohkawa, T.: A New Biclustering Algorithm with Exclusive Random Selection of Columns for Predicting Recognition Spots on Protein Molecular Surfaces. International Journal of Bioscience, Biochemistry and Bioinformatics, Vol. 8, No. 1, pp. 11-19 (2018)

[5] Ho, H. T., Gibbins, D.: Curvature-based Approach for Multiscale Feature Extraction from 3D Meshes and Unstructured Point Clouds. IET Computer Vision Vol. 3, No. 4, 201-212 (2009)

[6] O′Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T.: Open Babel: An Open Chemical Toolbox. Journal of Cheminformatics, Vol. 3, No. 33, (2011)

[7] Mysinger, M. M., Carchia, M., Irwin, J. J., Shoichet, B. K.: Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. Journal of Medicinal Chemistry. Vol. 55, No. 14, pp. 6582-6594 (2012)

[8] Trott, O., Olson, A. J.: AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization and Multithreading. Journal of Computational Chemistry. Vol. 31, No. 2, pp. 455-461 (2010)

[9] Dawson, N.L., Lewis, T.E., Das, S., Lees, J.G., Lee, D., Ashford, P., Orengo, C.A., Sillitoe, I.: CATH: An Expanded Resource to Predict Protein Function through Structure and Sequence. Nucleic Acids Research. Vol. 45, Issue D1, pp. 289-295. (2017).