Special Issue
"Protein-Ligand Interactions"

Vol. 13, pp. 139–147 (2016)
doi: 10.2142/biophysico.13.0_139

Biophysics and Physicobiology

https://www.jstage.jst.go.jp/browse/biophysico/

Regular Article

Effects of the difference in similarity measures on the comparison of ligand-binding pockets using a reduced vector representation of pockets

Tsukasa Nakamura[1,2] and Kentaro Tomii[1,2,3]

[1]*Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa-shi, Chiba 277-8562, Japan*
[2]*Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Koto-ku, Tokyo 135-0064, Japan*
[3]*Biotechnology Research Institute for Drug Discovery, National Institute of Advanced Industrial Science and Technology (AIST), Koto-ku, Tokyo 135-0064, Japan*

Received January 12, 2016; accepted June 6, 2016

Comprehensive analysis and comparison of protein ligand-binding pockets are important to predict the ligands which bind to parts of putative ligand binding pockets. Because of the recent increase of protein structure information, such analysis demands a fast and efficient method for comparing ligand binding pockets. Previously we proposed a fast alignment-free method based on a simple representation of a ligand binding pocket with one 11-dimensional vector, which is suitable for such analysis. Based on that method, we conducted this study to expand and revise similarity measures of binding pockets and to investigate the effects of those modifications with two datasets for improving the ability to detect similar binding pockets. The new method exhibits higher detection performance of similar binding pockets than the previous methods and another existing accurate alignment-dependent method: *APoc*. Results also show that the effects of the modifications depend on the difficulty of the dataset, implying some avenues for methods of improvement.

Key words: multidimensional scaling, alignment-free comparison, triangle descriptor, large-scale comparison

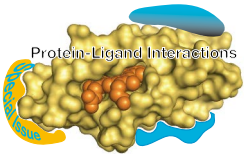
Abbreviations: MDS, multidimensional scaling; TPR, true positive rate; FPR, false positive rate; ROC, receiver operating characteristic; AUC, area under a ROC curve

Corresponding author: Kentaro Tomii, National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan.
e-mail: k-tomii@aist.go.jp

Elucidation of interactions between proteins and ligands such as small molecules is particularly important. They can be expected to facilitate the functional elucidation of proteins, with relations to metabolism, drug discovery, and drug repositioning. With the recent increase of a database of known protein structure, the Protein Data Bank (PDB) [1], we have huge amounts of structural information for approximately 300,000 known and 5.2 million unknown (estimated using a pocket identification program) ligand-binding pockets [2].

◀ ***Significance*** ▶

As the recent increase of protein structure information, comprehensive analysis of protein ligand-binding pockets demands an efficient method for comparing them. Based on our previous method using a vector representation of pockets, we conducted this study to expand and revise similarity measures of pockets by both changing the definition and increasing the number of pocket descriptors, and by using the PAM50 matrix. We investigated the effects of those modifications with two datasets, and found that the new method exhibits slightly higher detection performance of similar pockets than the previous method and an existing accurate alignment-dependent method.



Consequently, comprehensive comparison and classification of both known and predicted protein ligand-binding pockets provide important insights into predicting ligands and drug discovery. For such a comprehensive analysis, a fast pocket comparison method is extremely useful. Indeed, various approaches have already been proposed [3].

Pocket comparison methods are divisible into two classes, i.e., alignment-dependent and alignment-free methods [4]. Although alignment-dependent methods for pocket comparison perform structural alignment of binding residues, alignment-free methods are independent of the structural alignment. Such methods often use descriptors that represent binding residues in a pocket. Both methods have been developed to be applicable to large-scale comparison of binding pockets. For instance, Gao *et al.* developed a fast alignment-dependent method called *APoc* [5]. They argued that alignment-dependent methods are "generally more accurate, albeit slower than alignment-free methods." However, although alignment-free methods are unable to provide information of matched residues, they are efficient in terms of computational time. Consequently, these methods enable comparison of known binding pockets with numerous predicted ligand-binding pockets estimated using a pocket detection program. Furthermore, alignment-free methods are compatible with analysis of "flexible" binding pockets, and are readily applicable to binding pockets comprising multiple protein chains. It remains difficult to apply alignment-dependent methods directly to such pockets.

Considering these reasons, we assume that alignment-free methods can particularly contribute to protein–ligand interaction prediction and can enable the prediction of protein function. Therefore, we developed an alignment-free method that enables us to perform exhaustive comparison of both known and predicted ligand-binding pockets of 1,000,000 order [6], and to develop a database called PoSSuM that includes the comparison results [2,7]. Recently, we also proposed a fast method based on a simple representation, an 11-dimensional vector, of a ligand-binding pocket using a triangle descriptor defined by a set of three amino acids in a pocket and multidimensional scaling (MDS) [8]. In this method, the vector representation of a ligand-binding pocket is obtained using the linear combination of the occurrence frequency of triangles using their coordinates in a metric space. Using this method, one can calculate the similarity between two ligand-binding pockets merely by calculating the inner product of two reduced vectors.

For this study presented here, we sought to revise the definition of the triangle descriptor of ligand-binding pockets for improving the discriminative ability of our method. To define new similarity measures of binding pockets, we used an amino acid similarity matrix instead of the distance matrix used in our previous method to present physicochemical similarity between amino acids in pockets. We expanded the classes to consider the geometrical similarity between edges of the triangle descriptor, and modified the definition of those classes. In addition to these points, we also examined use of the increased number, instead of 11 used in our previous method, of dimensions in the results of MDS. The new method belongs to the alignment-free classes, exhibiting higher detection performance of similar binding pockets than our previous method and an existing fast sequence order-independent structural alignment method: *APoc*. This report describes that the effects of the modifications depend on the dataset difficulty. These results suggest some avenues for future development of similarity measures between binding pockets for improvement of pocket comparison methods.

Materials and Methods

New similarity measures for ligand-binding pocket comparison

We have proposed a simple representation, an 11-dimensional vector, of a ligand-binding pocket using triangle descriptor defined by a set of three amino acids in a pocket and MDS [8]. For this study, we intend to expand and revise similarity measures of our method and to elucidate their effects for improvement of the detection ability of similar binding pockets. We mainly emphasize and modify the following two points based on our previous method. i) Whereas our previous method used Miyata's amino acid distance matrix [9] to represent physicochemical distance or dissimilarity between two amino acids to be compared, we employ an amino acid similarity matrix to present physicochemical similarity between two amino acids in our new method presented here. ii) To represent the geometrical dissimilarity between edges of the triangle descriptor, in our previous method, we considered edges that correspond to the Cα–Cα distances of residue pairs of 1.0 Å to 13.6 Å, and classified them into five classes at 2.2 Å intervals. In our new methods, we extended Cα–Cα distances to be considered, ranging from 1.0 Å to 15.8 Å, and added a class of edges. Edges were classified into 6 classes at intervals of 2.2 Å. We assigned them Roman numerals (I, II, III, IV, V, and VI) in ascending order (hereinafter designated as 'interval set α'). We also investigated the effects of revising the interval distances which affect the classes of edges. We modified the intervals of 6 edge classes from 2.2 Å each to 1.0, 4.0, 6.36, 8.72, 11.08, 13.44, and 15.8 Å (hereinafter, 'interval set β'). According to these expansions and revisions, the definition of similarity between two triangle types is modified slightly (see below eq. (1)). In addition to these points, we also examined the use of a higher number of dimensions, instead of the 11 used in our previous method, in the results of MDS. Detailed explanations of modifications in our new methods are described below.

Enumerating possible triangle types

First, similarly to our previous study, we enumerated all possible triangle types. In this study, we eventually increased the classes of edge labels to 6 labels from the 5 labels used

in the previous study. However, vertex labels were the same as those used previously: 20 labels with one letter amino acid of 20 types. In all, we have the 295,240 triangle types for the 6 labels (and the 171,700 triangle types for the 5 labels [8]).

Definition of similarity between two triangle types

In our new method, for two triangle types of p and q, we defined the similarity s_{pq} consisting of two terms. They respectively denote physicochemical and geometrical similarity, as

$$s_{pq} \equiv \max\Big[\underset{\text{: 6 way superposition}}{\big(r(m_{AD}+m_{BE}+m_{CF}) + (1-r)(-1) \big)} \Big] \quad (1)$$
$$\big(f(AB, DE)+f(BC, EF)+f(CA, FD) \big)$$

Therein, m_{XY} represents a physicochemical similarity between two amino acids X and Y, defined with an amino acid substitution matrix. For this study, we used the PAM50 matrix [10], which is not rounded after a decimal point, because we assumed that residues consisting of a ligand-binding pocket are conservative for substituting amino acids. Also, the not-rounded PAM50 matrix yielded slightly better performance than the rounded (data not shown). In addition, A, B, and C respectively denote the vertices of the triangle type p; D, E, and F respectively denote those of triangle type q. r is a weighting factor, ranging from 0 to 1, for physicochemical and geometrical similarity terms in this equation. AB, BC, and CA denote the edges of the triangle type p; DE, EF, and FD denote those of triangle type q. Function f, which represents the geometrical dissimilarity between two edges X and Y, is defined as

$$f(edgeX, edgeY) \equiv |\text{value of the class for edgeX} - \text{value of the class for edgeY}|.$$

In this definition, the value of a class is given according to the assigned numerals for a class. For example, function f gives 4 when edge X belongs to class I and edge Y belongs to class V. Then, f is summed up for three edges and multiplied by –1 for converting dissimilarity to similarity (see eq. (1)). We regarded the maximum value of s_{pq} for all possible ways of superposition of triangle types considering rotation and reflection as similarity s_{pq} for two triangle types p and q.

Multidimensional Scaling (MDS)

To execute MDS, we calculated the similarities for all possible pairs of 295,240 triangle types based on the formulation described above. We were able to obtain a similarity matrix $\boldsymbol{S}$ between triangle types as a square matrix of order 295,240. We assumed a model by which the similarity between triangle types corresponds to the inner product, and used MDS to obtain the coordinates of each triangle type in a high-dimensional space (it is also called kernel PCA, if the similarity function is guaranteed to have positive (semi-) definite property) [11]. The procedures used for this study are summarized briefly as follows. First centering is performed over the previously described similarity matrix $\boldsymbol{S}$ to obtain the centered similarity matrix $\tilde{\boldsymbol{S}}$ because we want to obtain, eventually, those coordinates which have zero mean. The element of the centered similarity matrix $\tilde{\boldsymbol{S}}$ is obtainable as

$$\tilde{S}_{ij} = S_{ij} - \frac{1}{N}\sum_{a=1}^{N} S_{ia} - \frac{1}{N}\sum_{a=1}^{N} S_{aj} + \frac{1}{N^2}\sum_{a=1}^{N}\sum_{b=1}^{N} S_{ab}.$$

Then, the eigenvalue decomposition of $\tilde{\boldsymbol{S}}$ is performed, thereby yielding eigenvalue vector $\boldsymbol{\lambda}$ and a matrix of eigenvector $\boldsymbol{Z}$. Using $\boldsymbol{\lambda}$ and $\boldsymbol{Z}$, the coordinates of triangle types are then found using the following formula:

$$\boldsymbol{X} = (\sqrt{\lambda_1}\boldsymbol{z}_1, \sqrt{\lambda_2}\boldsymbol{z}_2, \sqrt{\lambda_3}\boldsymbol{z}_3, ..., \sqrt{\lambda_l}\boldsymbol{z}_l).$$

We used the randomized algorithm [12] to conduct large-scale singular-value decomposition for eigenvalue decomposition of $\tilde{\boldsymbol{S}}$ because the order of $\tilde{\boldsymbol{S}}$ is huge, and because the only necessary eigenvalues are those with a large absolute value.

Convert pockets into reduced vector representations

We defined $\boldsymbol{n}$ as a 295,240-dimensional vector based on the occurrence frequencies of triangle types at a ligand-binding pocket. All triangles that occur in a pocket with edge lengths of 1.0 Å to 15.8 Å are classified as one of 295,240 triangle types. Using $\boldsymbol{X}$ described in the previous section, we found the following.

$$\begin{aligned}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{n} &= (\sqrt{\lambda_1}\boldsymbol{z}_1, \sqrt{\lambda_2}\boldsymbol{z}_2, \sqrt{\lambda_3}\boldsymbol{z}_3, ..., \sqrt{\lambda_l}\boldsymbol{z}_l)^{\mathrm{T}}(n_1, n_2, n_3, ..., n_{295240})^{\mathrm{T}} \\ &= (\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, ..., \boldsymbol{x}_{295240})(n_1, n_2, n_3, ..., n_{295240})^{\mathrm{T}} \\ &= (w_1, w_2, w_3, ..., w_l)^{\mathrm{T}} = \boldsymbol{w}\end{aligned}$$

In the equations above, n_i represents the number of the i-th triangle in the list of triangle types. $\boldsymbol{w}$ stands for a vector representing a pocket (Fig. 1). To represent a pocket with a reduced vector based on the MDS result, we used the number of dimensions that satisfy a certain extent of cumulative contribution ratio calculated using only positive eigenvalues (Supplementary Fig. S1). For this study, we set the criteria of the cumulative contribution ratio as 0.98. We define similarity between two pockets i and j as a cosine distance between $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$. Therefore, the similarity can be found easily by calculating the inner product between normalized $\boldsymbol{w}_i$ and normalized $\boldsymbol{w}_j$. This procedure can be regarded as calculation of the weighted arithmetic mean over $\boldsymbol{X}$ weighted by $\boldsymbol{n}$.

Datasets

To optimize the weighting factor r in eq. (1) and also to analyze the performance of the methods described above, we used the two datasets used in our previous study [8]. One is *Ito138*. This difficult dataset comprises 138 known ligand-binding pockets. The dataset comprises pocket pairs that share the same types of small molecules in proteins with different global structures. The other is *APocS3* used in the study of *APoc* [5]. While this easy dataset originally comprised 38,066 pairs each in Subject and Control dataset, nevertheless, we noted that some of binding pocket are

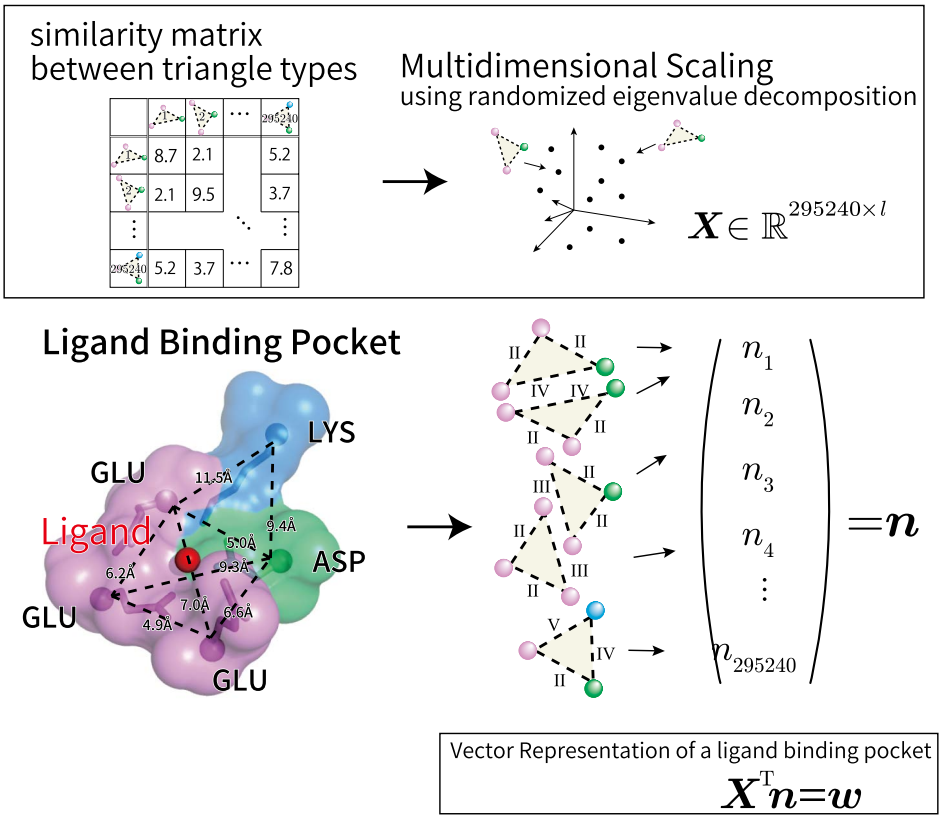


Figure 1 Schematic diagram of a vector representation of a ligand-binding pocket. Structural and amino acid information of a ligand-binding pocket are converted into a vector based on the MDS result.

inadequate because of small number of binding residues, based on the default setting of *APoc* which is that "minimal number of pocket residue is 10". We omitted pocket pairs which could not be handled by *APoc* with default setting, and then confirmed that we reproduced the same ROC curves (Fig. 3B in [5]). For that reason, we used 37,956/26,527 pairs for Subject/Control dataset in this study. Coordinate files of binding pockets were obtained from the *APoc* website (http://cssb.biology.gatech.edu/APoc). We conducted all-against-all 9,453 (= $\binom{138}{2}$) comparisons for the *Ito138* dataset. For *APocS3*, we performed comparisons of the 64,483 ligand-binding pocket pairs for the Subject and Control datasets. We regarded ligand-binding pocket pairs that share the same (for *Ito138* & *APocS3*) and also similar (for *APocS3*) ligands as relevant pairs (Subject); otherwise we regarded pocket pairs as non-relevant pairs (Control). For a pocket pair with the same ligand, i.e., a relevant pair with a higher similarity score than a threshold value, it was regarded as a true positive (TP). Otherwise, it was regarded as a false negative (FN). If a pair with ligands that are not the same, i.e., non-relevant pair has a lower similarity score than the threshold value, then it was regarded as a true negative (TN). Otherwise, it was regarded as a false positive (FP). Here, a true positive rate (TPR), also designated as recall or sensitivity, is defined as TP/(TP+FN). A false positive rate (FPR), also designated as fall-out, is given as FP/(FP+TN). Then, the receiver operating characteristic (ROC) curve is used to present these results. Actually, ROC is a curve based on TPR against FPR at various thresholds. The area under the ROC curve (AUC) is used for performance evaluation of the methods.

Additionally, we used the third dataset, which we call *APocS3_LIGSITE*, to compare our new method with *APoc*. This dataset included pairs of predicted pockets generated by LIGSITE [13] based on pocket pairs in *APocS3*. In comparison with *APocS3*, pocket pairs were reduced to 34,511/17,408 pairs of the Subject/Control dataset because binding residues of some pockets could not be predicted correctly.

Results and Discussion

We investigated the effects of modifications in the following three points, i.e., the new similarity definition, the expansion of edge classes, and the revision of intervals of edge classes. Then we compared our new method with an existing fast sequence order-independent structural alignment method: *APoc*.

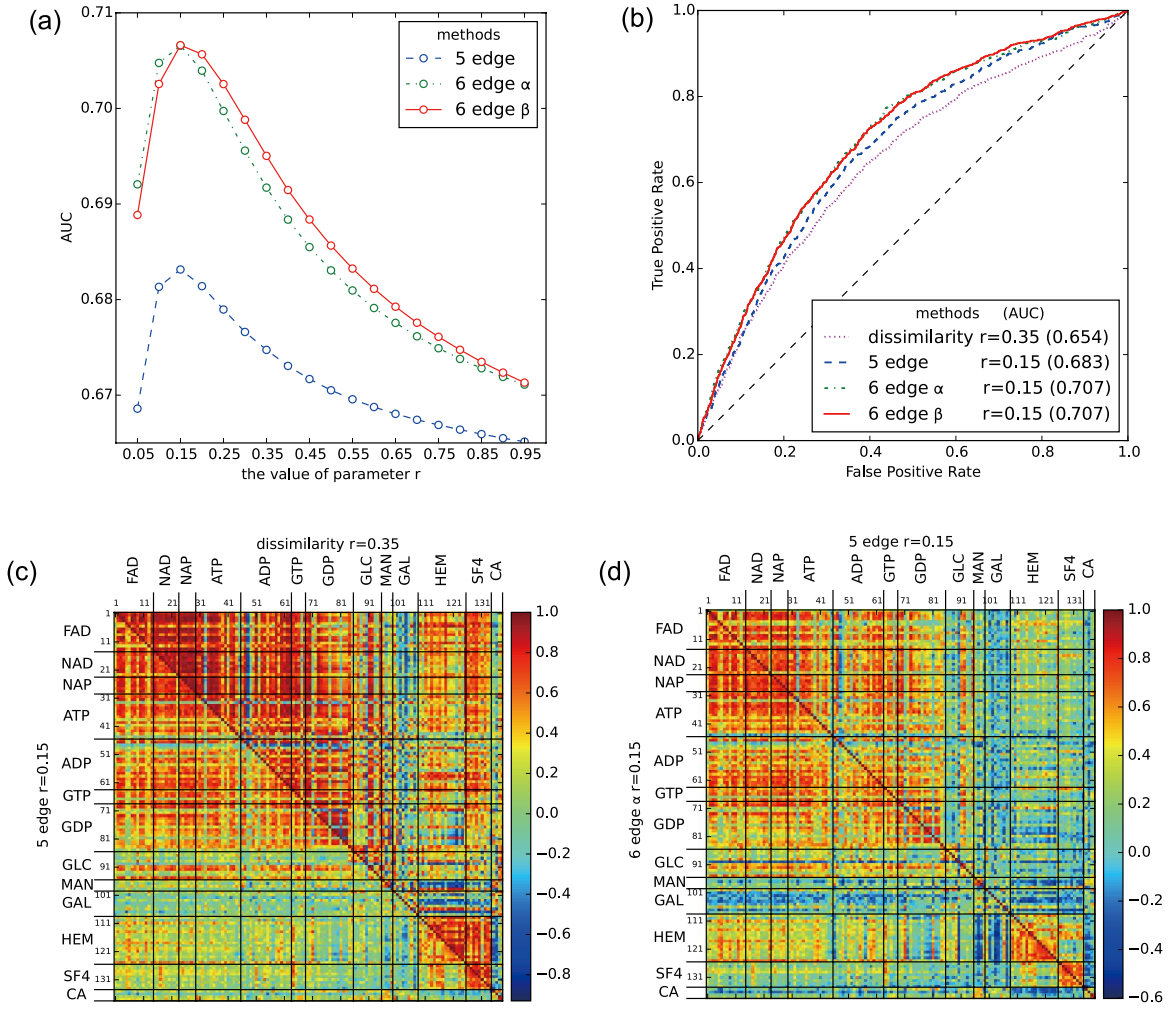


Figure 2 Benchmark results with *Ito138*. (a) Results of our new methods, including '5 edge', '6 edge (with interval set) α', and '6 edge (with interval set) β', with various values (0.05–0.95 with the sampling interval of 0.05) of the weighting factor *r* are shown. The X axis indicates *r*, and the Y axis indicates AUC values. (b) ROC curves of our new methods and our previous dissimilarity-based method (r=0.35) are shown. The X axis shows FPR. The Y axis shows TPR. (c) Heat maps to compare the new (5 edge class) similarity-based method (lower left) with the dissimilarity-based method (upper right) are shown. The color of each square in the map represents a similarity value for a pocket pair. Ligand abbreviations placed by axes correspond to the ligand to which a pocket binds. (d) Heat maps to compare the '6 edge α' method (lower left) with the '5 edge class' method (upper right) are shown.

Effects of the new similarity definition between pockets

First, we evaluated the effectiveness of changing the similarity definition between two ligand-binding pockets. For evaluation, a new method was used with the number of edge classes set as five classes. The method of classifying them is the same as that of our previous method. We tested the weighting factor in every 0.05 sampling from 0.05 to 0.95 to define the optimized value of *r* using the *Ito138* dataset. Plots of the weighting factor vs. AUC are presented in Figure 2a as '5 edge'. According to this result, the best AUC is obtained with r=0.15. Therefore, we used this value to evaluate the effectiveness of the new similarity definition. Figure 2b presents ROC curves, i.e., plots of TPR vs. FPR for this evaluation and shows that the new similarity definition outperforms the previous dissimilarity definition. We identified the main reason behind the superiority of the new definition. Figure 2c presents actual similarity values for all-against-all 9,453 pairs as a heat map to compare the new similarity-based method (shown at the lower left) with our previous method (shown at the upper right). Each square in the graph corresponds to one similarity of a pair of pockets. Relative reddish/blueish in the color scale in a method is important. We found that the new method assigns lower similarity to pockets, especially to those which bind to HEM or SF4 with pockets which bind to the other ligands.

Next, we evaluated the effectiveness using *APocS3*. We tested weighting factor *r* in the manner described above. Plots of the weighting factor vs. AUC are presented in Figure 3a as '5 edge'. According to this result, the best AUC is given by 0.10. Therefore, we used it to evaluate the effective-

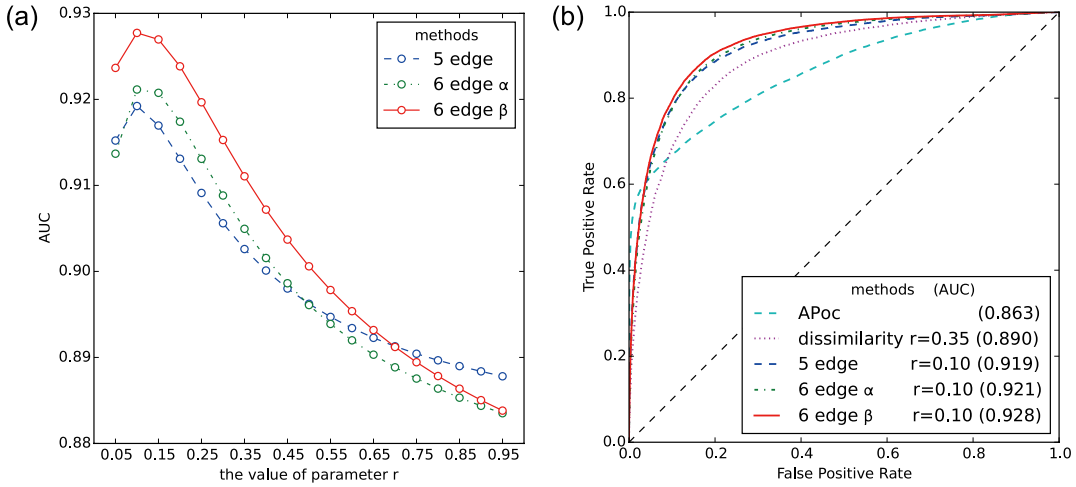


Figure 3 Benchmark results with *APocS3*. (a) AUC values of our new methods, including '5 edge', '6 edge α', and '6 edge β', with various values (0.05–0.95 with the sampling interval of 0.05) of the weighting factor *r* are shown. (b) ROC curves of our new methods, our previous dissimilarity-based method (*r*=0.35), and *APoc* are shown.

ness. Figure 3b presents ROC curves for this evaluation. It shows that the new similarity definition also outperforms previous dissimilarity definition. We considered that the parameter which gives the best AUC changed from 0.15 to 0.10 because the *APocS3* dataset is more conserved for pocket shape than *Ito138*. Therefore, the geometrical similarity term in eq. (1) is more weighted to distinguish a subject pair from a control pair.

Furthermore, to investigate the effects of similarity definitions, we compared the results obtained by combining previous/new definitions and different number of dimensions used in vector representation: i) the previous dissimilarity definition and new 139/137 dimensions, ii) the new similarity definition and previous 11 dimensions, and iii) the new similarity definition and the same criteria used in our previous study [8] for deciding the number of dimensions. Here, the criteria used in the previous study is "only use positive eigenvalues which are greater than the absolute value of negative minimum eigenvalue". These results are shown in Supplementary Table S1. When the same criteria used in the previous study was employed, the results with similarity are superior to ones with dissimilarity. Indeed, the results with similarity are superior to ones with dissimilarity except "11 dimensions (fixed)" in *APocS3*. We concluded that the improvements described above largely depends on the new similarity measure. We speculated that these improvements came from the difference of the matrix used for deriving the dissimilarity/similarity matrix. The Miyata's matrix which was used previously is just created from physicochemical characteristics in amino acids. On the contrary, the PAM50 matrix which was used for new measure reflects the evolutions of proteins and the concept based on our assumption that residues of pockets are less mutative.

Effects of increasing the number of edge classes

Next, we increased the number of edge classes from 5 to 6, according to the expansion of Cα–Cα distances of residue pairs, ranging from 1.0 Å to 15.8 Å, used as the triangle edge. Edges were classified into 6 classes at intervals of 2.2 Å (interval set α). First the effects of this modification were evaluated using *Ito138*. Figures 2a, b show that 6 edge classes outperform 5 edge classes, which suggests that addition of edge classes engenders better ability to recognize similar binding pockets. In Figure 2d, as heat maps, we compared individual result obtained using the '6 edge α' method (shown in the lower left) with it using the '5 edge class' method (shown in the upper right). In this case we also found that discrimination of HEM binding pockets from SF4 binding pockets is improved in the '6 edge α' method compared with the '5 edge class' method. Squares which correspond to HEM binding pockets vs SF4 binding pockets are less reddish/orangish in '6 edge alpha class' than '5 edge class', and boxes which correspond to HEM binding pockets vs HEM binding pockets are more reddish. Similarly, the discriminate power of GDP binding pockets from other pockets by the '6 edge α' method is slightly better than that of the '5 edge class' method. We suppose that discrimination of HEM binding pockets and SF4 binding pockets became better because the maximum value of edge length changed to 15.8 Å. The HEM binding pocket is commonly large. The distance between some binding residues which face each other through HEM is about 15 Å.

We also evaluated the effectiveness of our new method using an easy dataset: *APocS3*. Figures 3a, b show that the '6 edge α' method also outperforms 5 edge method. Similar but slightly different results were found with the results described above obtained with *Ito138*. Regarding AUC values along with the weighting factor *r* (Fig. 3a), the '6 edge α'

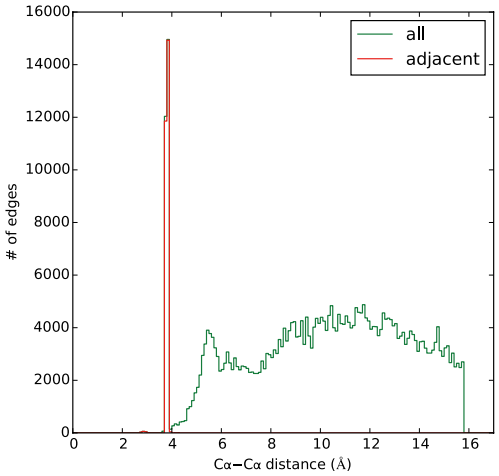


Figure 4 Distribution of the edge length (Cα–Cαdistance) of triangles taken from all pockets in *Ito138*.

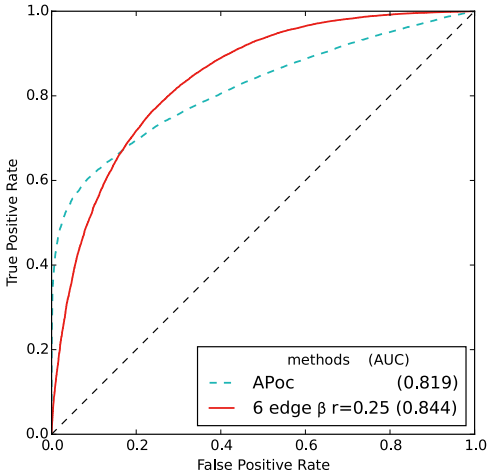


Figure 5 Benchmark results with *APocS3_LIGSITE*. ROC curves of our new method and *APoc* are shown.

method is better than the 5 edge method, except for the range of r=0.05, and 0.50 and more. With the new similarity definition, both methods showed maximum AUC values at r=0.10.

Effects of the intervals of class of edge labels

Next, we examined the procedure used to decide the intervals and modified the intervals of 6 edge classes from 2.2 Å each to 1.0, 4.0, 6.36, 8.72, 11.08, 13.44, and 15.8 Å (interval set β). The setting of every 2.2 Å interval is the same as that of our original method [6]. The first interval was set to 4.8 Å, the distance originated from the FuzCav method [14]. However, we investigated the frequency of the edge length of triangles taken from all pockets in the *Ito138* dataset (Fig. 4). In the figure, the green line shows the frequencies of all edge lengths. The red line shows the frequency of edge lengths which comprise two adjacent residues in a chain. According to this figure, almost all edges shorter than about 4.0 Å comprise adjacent residues. Thus, we considered it natural to set the first interval as 4.0 Å based on the difference of chemical characteristics between adjacent residues, or lack thereof. First the effectiveness of this modification was evaluated using *Ito138*. Figures 2a, b show that this modification was not so influential, in terms of AUC values, to our new method on *Ito138*, probably because *Ito138* comprises pockets that are too diverse to be affected by this improvement. On the other hand, Figures 3a, b show that the '6 edge β' method is better than that of the '6 edge α' method on *APocS3*. Moreover, comparing with *APoc*, in the low-FPR region, *APoc* showed higher performance. However, in the region higher than 3.7% FPR, the '6 edge β' method showed higher performance than that of *APoc*. The result that *APoc* is superior in low FPR region is also shown in our original method PoSSuM, which is also alignment-free method (Supplementary Fig. S2). Here, we used the encoding set 3, which showed the highest AUC among the first four sets (1~4), used in PoSSuM for labeling residues in a pocket [6]. We found that the '6 edge β' method achieved the highest AUC among the methods. Note that, in this experiment, we used 26,407 pairs out of 26,527 ones of the Control dataset, due to lack of triangles, based on the definition of the encoding set 3, in some pockets (some amino acids are assigned no vertex labels). Thus, it is noteworthy that our new method can handle all pairs because all 20 types of amino acids are assigned vertex labels.

Performance comparison with *APocS3_LIGSITE*

Finally, we compared the '6 edge β' method with *APoc* using the *APocS3_LIGSITE* dataset. Figure 5 shows that, in the low-FPR region, *APoc* showed higher performance. However, in the region higher than 17% FPR, the '6 edge β' method showed higher performance than that of *APoc*. Additionally, in the perspective of AUC, the '6 edge β' method showed higher performance than that of *APoc*.

Expedient examples of our method

We present examples that demonstrate the usefulness of our new method by comparison to *APoc*. First, we show an example from *Ito138* dataset. The interferon-inducible p47 resistance GTPases from mouse (PDBID: 1TQ4 [15]) and the alpha1,3-fucosyltransferase with GDP from *H. pylori* (2NZX [16]) have the same ligand: GDP, though the two proteins possess different global structures; P-loop containing nucleoside triphosphate hydrolases fold (1TQ4) and UDP-Glycosyltransferase/glycogen phosphorylase fold (2NZX). Our new method gave 0.864 as the similarity score for this pocket pair (the higher the similarity score, the more likely the pair is composed of pockets to which the same/similar ligand bind). It is noteworthy that *APoc* gave 0.772 as the *p*-value for this pair (the lower the *p*-value, the more

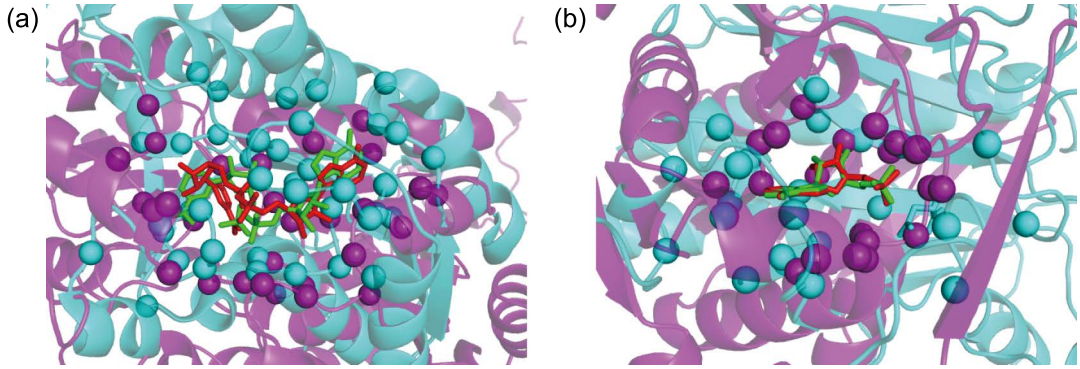


Figure 6 Superpositions of two expedient examples detected by our method. The superpositions are obtained by minimizing RMSD between the each ligands. (a) 1AD3 (cyan and green) and 1ZBQ (purple and red), which bind the same ligand: NAD. The Ca atoms of binding residues are depicted as spheres. Ligand RMSD is 3.05 Å (b) 12AS (cyan and green) and 1EFV (purple and red), which bind the same ligand: AMP. Ligand RMDS is 0.97 Å.

likely the pair is composed of pockets to which the same/similar ligand bind), and gave 0.307 as the raw "Pocket Similarity score (PS-score)" (the higher the PS-score, the more likely the pair is composed of pockets to which the same/similar ligand bind).

We present two more examples from the *APocS3* dataset. The aldehyde dehydrogenase from rat (1AD3 [17]; ALDH-like fold) and the 17-beta-hydroxysteroid dehydrogenase type 4 from human (1ZBQ; NAD(P)-binding Rossmann fold) have the same ligand: NAD (Fig. 6(a)). As discussed in the *Discussion* and *Conclusion* sections in the paper about *APoc*, this is an example of dissimilar pockets with different ligand conformations. *APoc* assigned this pair of pockets a *p*-value of 0.418 (and gave 0.325 as the PS-score), even though our new method produced a similarity score of 0.872. We regard this fact as demonstrating the effect of usefulness of our alignment-free method, which can accommodate the pocket conformation change associated with the ligand conformation change. Furthermore, the asparagine synthetase from *E. coli* (12AS [18]) and the electron transfer flavoprotein from human (1EFV [19]) have the same ligand: AMP (Fig. 6(b)). Similarly, as discussed in the paper related to *APoc*, this is an example of dissimilar pockets and similar ligand conformations. Whereas *APoc* gave 0.212 as a *p*-value (and gave 0.337 as the PS-score) for this pocket pair, our new method showed a similarity score of 0.731. We regard this feature as demonstrating the usefulness of this alignment-free method, which can vaguely represent the circumstances related to a ligand.

Computational Time

It is explained in the report about *Apoc* that *APoc* requires 73 seconds as the total computation time (not including input file preparation) for pocket comparison using 1000/1000 randomly selected pairs of pockets from the Subject/Control sets. Thus, APoc consumes 0.037 s for one pair comparison. In contrast, using the same size of pairs randomly selected from the datasets, our new method consumes only 0.00000106 s ($=1.06\times10^{-6}$) for one pair comparison with 137-dimensional vectors, although 0.007 s are needed to produce 137-dimensional vectors by counting up the number of triangles from pocket data (not including input file preparation). Figure 7 shows that the estimation of computational times. If 300,000 known plus 5.2 million unknown pockets are used as pocket data, our new method consumes 200 days just using single thread whereas *APoc* needs almost 30,000 years which is also just using single thread. We speculate that we can achieve 5.5 million pockets comparison using this new method in a few days when we employ a hundred multi-threads.

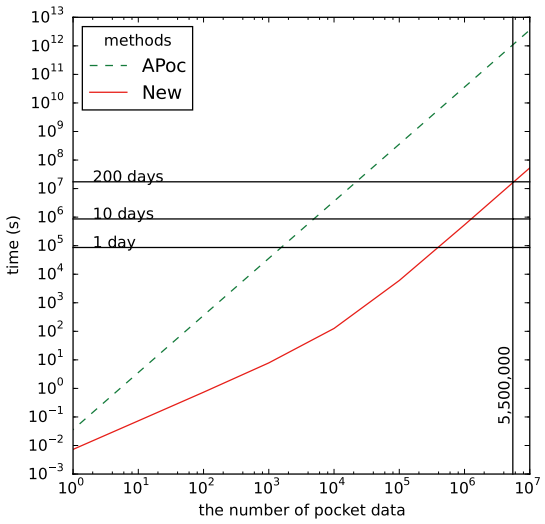


Figure 7 Estimated computation time of *APoc* and our new method. Both the X axis and the Y axis are shown in logarithmic scale.

Conclusion

Based on our previous method, for improving the ability to detect similar ligand-binding pockets, we expanded and revised similarity measures of pockets. We observed the effectiveness of those modifications with two different datasets, *Ito138* and *APocS3*, in comparison with our previous method. We also found that the effectiveness of the modifications depend on the difficulty of the dataset. These results should be considered for future development of pocket comparison methods. The method proposed herein showed higher detection performance of similar binding pockets than an existing fast sequence order-independent structural alignment method: *APoc*. Because of its succinct representation, our new method is expected to be useful for large-scale comparison of binding pockets to infer ligands and functions of proteins.

Acknowledgments

We thank Dr. Jun-Ichi Ito for helpful discussion. We also thank Prof. Tsuda for helpful comments. This study was partially supported by the Platform Project for Supporting in Drug Discovery and Life Science Research (Platform for Drug Discovery, Informatics, and Structural Life Science) from Japan Agency for Medical Research and Development (AMED).

Conflict of Interest

The authors declare that they have no conflict of interest, financial or otherwise, in relation to this study.

Author Contributions

T. N. conducted computational experiments and analysis of data, and drafted the manuscript. K. T. supervised the project, contributed to interpretation of data, and wrote the manuscript.

References

[1] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

[2] Ito, J., Ikeda, K., Yamada, K., Mizuguchi, K. & Tomii, K. PoSSuM v. 2.0: data update and a new function for investigating ligand analogs and target proteins of small-molecule drugs. *Nucleic Acids Res.* **43**, D392–D398 (2015).

[3] Konc, J. & Janežič, D. Binding site comparison for function prediction and pharmaceutical discovery. *Curr. Opin. Struct. Biol.* **25**, 34–39 (2014).

[4] Gao, M. & Skolnick, J. A comprehensive survey of small-molecule binding pockets in proteins. *PLoS Comput.* Biol. **9**, e1003302 (2013).

[5] Gao, M. & Skolnick, J. APoc: large-scale identification of similar protein pockets. *Bioinformatics* **29**, 597–604 (2013).

[6] Ito, J., Tabei, Y., Shimizu, K., Tomii, K. & Tsuda, K. PDB-scale analysis of known and putative ligand-binding sites with structural sketches. *Proteins* **80**, 747–763 (2012).

[7] Ito, J., Tabei, Y., Shimizu, K., Tsuda, K. & Tomii, K. PoSSuM: a database of similar protein-ligand binding and putative pockets. *Nucleic Acids Res.* **40**, D541–D548 (2012).

[8] Nakamura, T. & Tomii, K. Protein ligand-binding site comparison by a reduced vector representation derived from multi-dimensional scaling of generalized description of binding sites. *Methods* **93**, 35–40 (2016).

[9] Miyata, T., Miyazawa, S. & Yasunaga, T. Two types of amino acid substitutions in protein evolution. *J. Mol. Evol.* **12**, 219–236 (1979).

[10] Dayhoff, M. O. & Schwartz, R. M. Chapter 22: A model of evolutionary change in proteins. *Atlas Protein Seq. Struct.* (1978).

[11] Shawe-Taylor, J. & Nello, C. *Kernel methods for pattern analysis* (Cambridge university press, Cambridge, 2004).

[12] Halko, N., Martinsson, P. G. & Tropp, J. A. Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.* **53**, 217–288 (2011).

[13] Huang, B. & Schroeder, M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol.* **6**, 19 (2006).

[14] Weill, N. & Rognan, D. Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites. *J. Chem. Inf. Model.* **50**, 123–135 (2010).

[15] Ghosh, A., Uthaiah, R., Howard, J., Herrmann, C. & Wolf, E. Crystal structure of IIGP1: a paradigm for interferon-inducible p47 resistance GTPases. *Mol. Cell* **15**, 727–739 (2004).

[16] Sun, H. Y., Lin, S. W., Ko, T. P., Pan, J. F., Liu, C. L., Lin, C. N., *et al.* Structure and Mechanism of Helicobacter pylori Fucosyltransferase. A BASIS FOR LIPOPOLYSACCHARIDE VARIATION AND INHIBITOR DESIGN. *J. Biol. Chem.* **282**, 9973–9982. (2007).

[17] Liu, Z. J., Sun, Y. J., Rose, J., Chung, Y. J., Hsiao, C. D., Chang, W. R., *et al.* The first structure of an aldehyde dehydrogenase reveals novel interactions between NAD and the Rossmann fold. *Nat. Struct. Biol.* **4**, 317–326 (1997).

[18] Nakatsu, T., Kato, H. & Oda, J. Crystal structure of asparagine synthetase reveals a close evolutionary relationship to class II aminoacyl-tRNA synthetase. *Nat. Struct. Biol.* **5**, 15–19. (1998).

[19] Roberts, D. L., Frerman, F. E. & Kim, J. J. P. Three-dimensional structure of human electron transfer flavoprotein to 2.1-Å resolution. *Proceedings of the National Academy of Sciences* **93**, 14355–14360. (1996).