Vol.2018-MPS-118 No.44
Vol.2018-BIO-54 No.44
2018/6/15

Prediction of drug-target interactions with 3D structure information of target binding sites

Ruoming He[†1,†2,a)] Takashi Ishida[†1,†2,b)]

Abstract: Predicting drug-target interactions is an important step for drug design. Previous method to compare target pairwise similarities by comparing amino acid sequences is effective but containing limitation when dealing with remote homology sequences. Using 3D structure information is better since protein structures often decide the functions and the interaction modes of drug-target pairs. However, difficulties on getting 3D structures of target proteins make it tough to extract and analyze the binding site structures from the target protein structures. Moreover, rather than the whole structure, the binding site structure of a target decides more on the drugs it interacts with according to the hypothesis that targets with similar binding sites are easier to interact with the same drug. Thus, our approach applied target binding site similarities to represent target pairwise similarities by using homology search to get the 3D structures as well as extracting and comparing the binding sites structures. Finally, our method improved prediction accuracy compared with previous methods.

Keywords: Drug-target interaction prediction, Binding sites, Protein 3D structures

1. Introduction

Predicting drug-target interactions is an important part that helps new drug discovery. In addition, under the limited experimental environment, the rate of successful new drug development has decreased year by year, which makes it important to make use of existing or abandoned drugs (drug repositioning). Traditional human-depended method for selecting drug-target interactions is quite manpower-consuming and expensive, whereas using computer-aided methods will avoid experimental errors and give a series of scientific predictions. Moreover, the demand on dealing with huge amount of interaction data in a short time also leads to the trend of using computer-aided methods. Usually, computer-aided methods predict the interactions by analyzing the inner relationship between drug compounds and target proteins by catching characteristics based on the chemical functions or structure information with machine learning algorithms.

Nowadays, a number of computational methods are developed for making drug-target interaction predictions, which can be divided into network-based and machine learning-based methods. Network is an effective tool to predict possible drug-target associations. Cheng et al. [1] inferred drugtarget interactions based on the topology of the known interaction network with the drug similarity (DBSI), target similarity (TBSI) and network-based similarity (NBI) separately. Different from this, Chen et al. [2] applied network-based random walk with restart on a heterogeneous network (NRWRH). For machine learning methods, in the approach of Yamanishi et al. [3], they proposed a supervised bipartite graph learning approach by mapping the chemical space and the geometric space into a unified space and making use of the concept that interacting drugs and targets are close to each other while non-interacting drugs and targets are far away from each other. The probability of interaction is then calculated to express how close a pair of drug-target is in the mapping space. Bleakley and Yamanishi [4] further proposed a new supervised method, bipartite local model (BLM) by combining drug-based and target-based predictions so that transformed edge prediction problems into binary classification problems. In these supervised learning method, drug-target pairs are labeled as positive or negative samples according to whether the interactions between the drug and target have been confirmed or not. Therefore, some of the pairs that have not yet been confirmed will be regarded as negative samples, which will actually decrease the predictive accuracy. Instead, Xia et al. [5] applied a semi-supervised approach for local model learning and made improvements based on BLM method. Mei et al. [6] presented an improved version of BLM with neighbor-based interaction-profile inferring (BLMNII) that made it possible to provide a reasonable prediction for drug/target candidates that are currently new. Different from weighted profile methods, the interactions are used as label information to train the local model instead of being used directly in the final step to predict the interaction probability.

Besides developing the statistical and machine learning methods, enriching the data is another way to help with the prediction. For predicting the drug-target interactions, drug similarities, target similarities and known drug-target interactions are used as input data. The difficulty on selecting negative samples when using

[†1] Presently with School of Computing,Tokyo Institute of Technology, 2-12-1 W8-76 Ookayama, Meguro-ku, Tokyo 152–8550, Japan
[†2] Presently with Education Academy of Computational Life Sciences, Tokyo Institute of Technology, 2-12-1 W8-93 Ookayama, Meguro-ku, Tokyo 152–8550, Japan
[a)] he.r.aa@titech.ac.jp
[b)] ishida@c.titech.ac.jp

supervised machine learning algorithms comes from the blank of confirmed known drug-target interactions. To solve this problem, Shi et al. [7] thought of a concept of super-target to cluster similar targets following the concept, if a drug can interact with a target, it can interact with the super-target group containing that target. Using this approach, the method performanced better than previous methods. To calculate drug similarities, Yamanishi et al. [3] integrated the chemical structure information of compounds. Campillos et al. [8] calculated the similarities based on drug side-effect with 1018 drug side-effect relations and 746 market drugs.Shi et al. [7] incorporated a score based on the Anatomical Therapeutic Chemical (ATC) Classification System. To compare targets, Shi et al. [7] also introduced functional categories (FCs) with classified chemical reactions catalyzed by targets or the annotating functions of protein-coding genes while most of other researches calculated target protein similarities by calculating amino acid sequence similarities by Smith-Waterman algorithm [9].

However, it is estimated that around 60% of drug discovery projects fail because the target proteins are finally found to be not druggable. Therefore, understanding how a target protein works when interacting with a drug at the beginning will help to decrease the risk of predicting undruggable targets. When calculating the target protein similarities, it has limitation by calculating the amino acid sequence similarities rather than comparing 3D structures. Meanwhile, using the amino acid sequence similarities may involve the sequences that are remote homology sequences, which means that they are from different parts of the same protein with the same structure. Since not all residues on the protein surface participate in the interactions, instead there are specific locations that are known as binding sites for the interactions to happen. Those remote homology sequences may share the same binding sites when interacting with similar drugs. Thus, just calculating the amino acid sequence similarity is not sufficient to compare target protein similarities. Therefore, introducing binding sites similarity is crucial for improving the successful rate of drug-target interaction prediction. This point was also reviewed by Haupt and Schroeder [10].

Since similar drug compounds bind to similar target binding sites, the function and structure characteristics will be considered when extracting the binding sites. Methods for extracting binding sites are divided into geometric methods [11],[12],[13],[14],[15], and energy calculation methods [16],[17]. Geometric methods by analyzing the surface of the protein are more suitable for extracting binding sites from the target protein structures. Differences among those geometric methods are based on detection algorithms. For example, POCKET [11] and LIGSITE [13] search for protein-solvent-protein (PSP) events to represent the pockets along with both sides. SURFNET [12] marks the gap spheres between atoms and reduces the radii until no atom can intersect into and then retains the spheres with a bigger radius. FPOCKET [14] is based on voronoi tessellation and alpha spheres of the Qhull.

In this research, we calculated target protein binding site similarities considering chain information as well as other statistic values combining with the protein sequence similarity to avoid influence of tiny binding sites or incorrect binding sites. With our approach, the remote homology sequences will be calculated as the same binding site structure. In addition, those proteins that with similar sequences but actually share different binding sites can be picked out to decrease their influence. We tested our corrected dataset and new similarity definition on state of the current algorithm. The results showed using our method to calculate target similarity will help to increase the accuracy, especially on ion channel group. This increase on precision is important in drug-target prediction that related to peoples health.

2. Methods

To predict the unknown drug-target interaction pairs, the drug and target pairwise similarities as well as known interaction data are used as input of machine learning algorithm. To calculate the target similarity, we introduced binding site similarity. There are three steps for calculating the binding site similarities: searching 3D structures, extracting binding sites and comparing binding sites structures. The target proteins are originally noted with KEGG [18] ID and can be achieved with amino acid sequences from KEGG website. 3D structures are then calculated using these sequences by homology search and selected based on each protein with chain information. Binding sites structures are then extracted and interpreted into fingerprints and calculated with similarities. However, due to the size of binding sites, the calculation of similarity between small binding sites is not sufficient. Thus, previous sequence similarity is also considered to decrease the influence of this.

2.1 Search 3D structures

Using a series of amino acid sequence can definitely describe a unique protein, but it will cause problem when comparing the function or protein features between two sequences since both of the sequences may be a part of the same protein, which are known as homology sequences. To get avoid of this, we implemented homology search to the sequences by blast [19] at the beginning. Possible 3D structures of the sequence are ranked with statistic values (E-value, expect) and some values are used to describe the match degree (scores, identities, positives, gaps). Those homology sequences will return the same protein structure and the difference between homology sequences may exist on chain. Considering the chain information as well as the statistic values, the most possible 3D structure of the sequence is selected for furthering to extract the binding sites. Additionally, the protein structure similarities are also compared to confirm the structures we achieve match the original amino acid sequence.

2.2 Extract binding sites

Since not the whole structure of the protein will get involved in the interaction procedure, comparing the whole structure similarity is not accurate. Binding sites, with specific structure features to specific compounds, are more suitable to be compared to calculate the similarities between targets. Comparing the current methods for extracting binding sites structures, FPOCKET is the most practical method showing good performance by considering the geometric features. More importantly, it analyzes the druggability score which is one of the key elements that we referred

when selecting the possible binding sites.

2.3 Compare binding sites structures

Usually it is necessary to find a way to represent the 3D structures before comparing binding site structures. A typical method is to describe the 3D structures with C_α and C_β atoms. However, when the binding sites are focus on a single chain, the size of them will be too small to find enough atoms. Fuzcav [20] which represents a binding site structure with a 11-dimensional vector with 0 or 1 value in each bit of fingerprint, is chosen in our research. Following similarity calculation rule given in the original method we calculate the similarities of our protein datasets.

$$sim(A, B) = \frac{a}{min(nzA, nzB)} \quad (1)$$

As equation 1 shows: a is the number of common non-null counts in both fingerprints and nzA and nzB are the numbers of non-null counts in fingerprints A and B, respectively. With this method, it can calculate the similarities between two binding sites efficiently.

2.4 Datasets

The datasets used in our research are same with those in the research by Yamanishi et.al (2008). The datasets contain four groups of protein targets, which are enzyme, ion channel, G-protein-coupled receptor(GPCR) and nuclear receptor. Each group contains the drug compounds similarities considering chemical functions calculated by SIMCOMP and protein amino acid sequences similarities calculated by Smith-Waterman (SW) algorithm as well as the known drug-target interactions. In our research, we calculate the target protein similarities by calculating the binding site similarities following the three steps above. However, since the shortage of 3D structure information, there will be some target proteins that cannot achieve binding sites information. Moreover, there will be some binding sites that are too small to calculate the fingerprints. To simply compare the achievement of using binding site structures, in this research, we just remove those protein that cannot achieve binding site similarities from the original datasets. Table 1 shows the components of each group of datasets with the number of drugs (N_d), the number of targets (N_t) and the number of drug-target interactions (N_{d-t}).

Table 1 Data components of each target group

Dataset	Enzyme	Ion channel	GPCR	Nuclear receptor
N_d	445	210	223	54
N_t	419	122	71	24
N_{d-t}	186455	25620	15833	1296

The detailed datasets information with some statistic values describing the characteristics of each data group is shown in table 2. Regarding the results of the interactions, some of them are proved active (positive) drug-target interactions (N_{p_dt}) and others are unknown (negative) interactions (N_{n_dt}) , which will further be used as positive and negative label in machine learning algorithms. From the interaction data, we can find the number of negative interactions are far more than positive ones. Those negative interactions contain not only the true negative interactions but also the temporarily unknown interactions which will be clearly defined with the development of wet experiments. Number of drugs that interact with no target (N_{d-0t}) and only one target (N_{d-1t}) as well as number of targets that interact with no drug (N_{t-0d}) and only one drug (N_{t-1d}) are also counted. From these values, it can be found that active interactions take a tiny part, which is less than 0.05%. The small value of N_{d-0t} indicates there is a great space and significance for us to find out the interactions of those drugs with currently unknown disease target proteins. N_{d-1t} is smaller than N_{t-1d}, which proves that there exist some targets that interact with more than one drug, which also satisfies with the hypothesis that similar drugs can interact with the same protein. Therefore, finding out the target's properties which interacts with several drugs will help us to interprete the interaction between drugs and targets and help to find out unknown drugs with existing proteins.

Table 2 Statistics of each target group

Dataset	Enzyme	Ion channel	GPCR	Nuclear receptor
N_{p_dt}	2056	998	445	77
N_{n_dt}	184399	24622	15388	1219
N_{d-0t}	36	40	47	6
N_{d-1t}	231	64	87	38
N_{t-0d}	0	0	0	0
N_{t-1d}	173	10	26	8
Sparsity	0.989	0.961	0.9719	0.9406

The targets chosen for this research were processed from amino acid sequences to binding site fingerprints and then described with similarities. The details of data percentages in each step are shown in figure 1. It is shown that almost all of the sequences can be achieved with 3D structures, those with no 3D structures (the red part) takes less than 2% of the total. Meanwhile, among the 3D structures, a part of them do not have binding sites on specific chain (the yellow part). The lost data (the red and yellow part) in this research is just removed from the original dataset for comparing, but this can be improved with some statistic strategies in the future. Protein with fingerprints to describe their binding sites will be used by calculating the similarities that later used with other input data to make predictions.

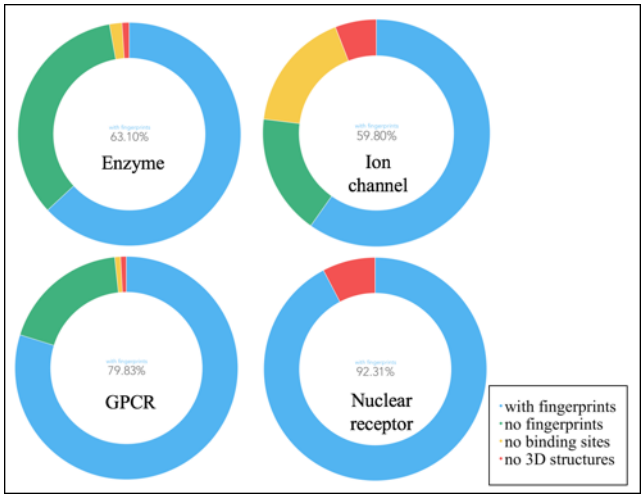

Fig. 1 Data distribution at each step for calculating the binding sites similarities

To decrease the influence of remote homology sequences and introduce the structure information, especially the binding site structures of target proteins, we proposed a method that used binding site similarities instead of protein sequence similarities. We further implemented our new pairwise similarities of targets on the current method to compare the results.

3. Results

To evaluate the prediction results based on different methods of calculating the similarities of targets, we adopted the same datasets of pairwise drug similarities and known drug-target interactions, the same procedure and assessment as those used in BLMNII [6]. We adopted 10-fold cross validation(10-CV) as the testing strategy in which targets in each dataset were split into 10 subsets of equal size, and 1 subset was used as testing set while the remaining 9 subsets were used as the training set. In each trail of cross validation, 1/9 targets were taken as testing targets, using interactions with known drugs that were labeled as positive while unknown drugs that were labeled as negative testing instances respectively. We then repeated the above procedure 10 times and evaluated the method on four groups of datasets. After that, Receiver Operator Characteristic (ROC) curve and Precision-Recall (PR) curve were used to describe the performance, and the area under each curve with values (AUC and AUPR,respectively) was also calculated. Usually, the method that achieves bigger AUC and AUPR values performance better. However, in drug-target interaction predictions, since the imbalance of the positive and negative data as well as the limitation of the datasets, AUPR will be more important which will decrease the influence of false positive data. Table 3 shows the AUC while table 4 presents the AUPR comparing our proposed method with the baseline method.

From the result values, our method of calculating the similarities of binding sites show good effectiveness on AUPR values than previous methods. Due to the limitation of the known interaction data, the AUPR value is more helpful on drug-target interaction predictions since it pays more attention on the true positive predicted drug-target pairs. In addition, future works on filling the blank data or introducing more information will help to increase the AUPR and even AUC.

Table 3 Comparison of AUC values among four groups of targets

Dataset	Enzyme	Ion channel	GPCR	Nuclear receptor
Proposed	0.93	0.96	0.89	0.86
Baseline	0.93	0.94	0.92	0.86

Table 4 Comparison of AUPR values among four groups of targets

Dataset	Enzyme	Ion channel	GPCR	Nuclear receptor
Proposed	0.50	0.61	0.28	0.48
Baseline	0.46	0.44	0.27	0.49

4. Discussion

4.1 Detailed analysis based on ROC curves

Our method showed good performance on ion channel group with improving both of AUC and AUPR value. Although on other groups of data, our methods didn't show an increase in AUC value, the ROC curves also gave us some inspirations on the improvement. With more interactions were predicted, noise data caused by incorrect similarities or non-zero values will cause the increase of TPR and decrease of FPR. From the ROC curve of Enzyme group in figure 4, when making enough predictions, our method's performance is better than the previous one. Thus, we can infer that when dealing with bigger size datasets, calculating the binding site similarities will help on avoiding false positive predictions. As figure 4 shows, there was some improvement in ion channel group while a decrease in both GPCR and nuclear receptor group.

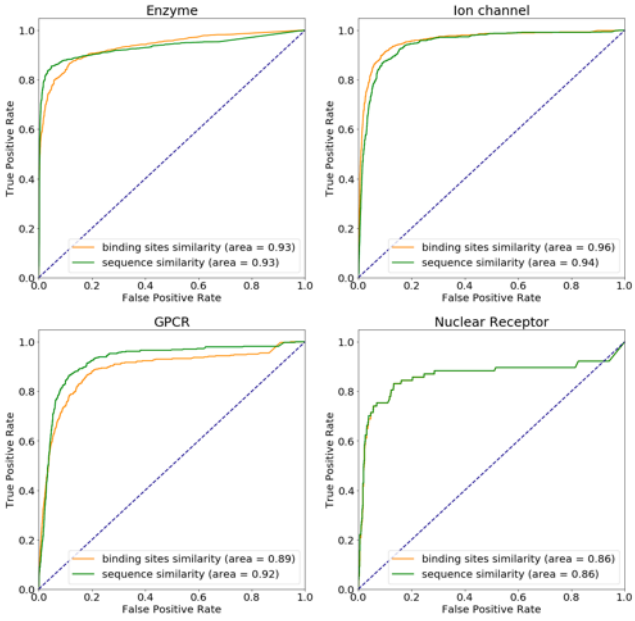

Fig. 2 Comparison of ROC curves on different target groups

One reason of this decrease may because of the size of the datasets. When calculating and comparing the binding sites structures, the loss of data will influence more on those originally small data groups. Thus, when comparing the similarities, it became not sufficient to compare among the small datasets. In the future, we can check this influence by replacing the blank values with real values based on statistic strategies.

4.2 Detailed analysis based on Precision-Recall curves

Comparing with AUC, AUPR is more useful to analyze the performance since it can decrease the influence of the highly-ranked false positive predictions when the number of pairs without known interactions are much more than the number of pairs with known interactions. The AUPR values were increased 5%-10% with our methods (figure 5), especially in ion channel group. For nuclear receptor group, the influence of using binding site similarities may not be obvious when combining with drug pairwise similarities. Additionally, the imbalance of known and unknown interactions may have a larger influence on this small group of data. In enzyme and GPCR group, at the beginning, our method showed a great advantage over the previous one, with more predictions, false positive predictions increased, where our

method's performance was worse than the previous one. It seems that with more predictions, our method contains more false positive results. This may be caused by the limitation of similarities calculated from binding site fingerprints. Among the dissimilar binding sites (with similarity equals 0), not all of them is really dissimilar binding site structures, since those without enough bits of fingerprints may also be regarded as dissimilar binding sites. In ion channel group, the percentage of binding sites with 0 bit fingerprints (0.287) is much smaller than GPCR group (0.310) and Enzyme group (0.537). Relatively, the performance of Ion channel group is shown better than other two groups. Between GPCR and Enzyme, the decreasing point also occurred later with GPCR group than Enzyme group. In future work, we can check this point by examining the bits of fingerprints. By dividing the data into binding sites with few bits and binding sites with enough bits but dissimilar, considering the latter with zero value while the former with a non-zero value, the calculation of binding sites similarities will be more sufficient and useful for making predictions.

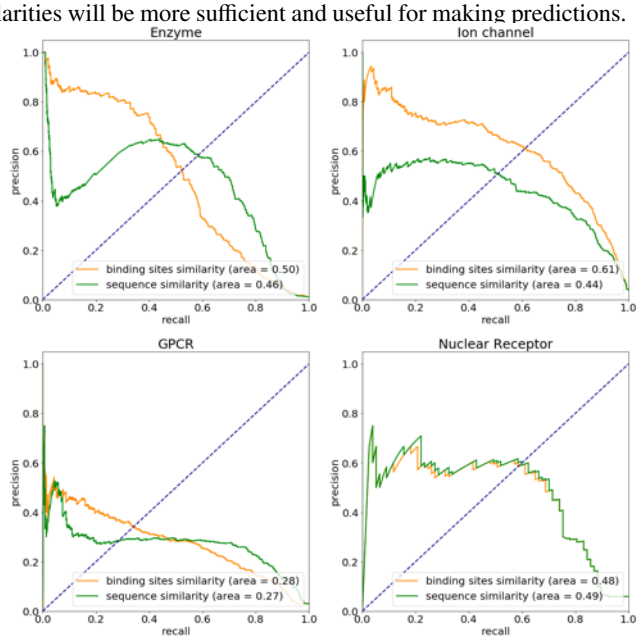

Fig. 3 Comparison of Precision-Recall curves on different target group

4.3 Importance of protein binding site structure information

By using homology search to calculate the 3D structures of proteins, the structures should be checked to make sure it can represent the original sequence. We applied TM-align measurement, which calculates the similarities between protein 3D structures and then compared the TM-align with SW-values. As figure 2 shows, the distributions of two measurements are mainly the same, just using TM-align enlarge the values. Thus, using homology search, we successfully interpreted those sequences into 3D structures. Among them, the different data containing homology structures, which has a SW-value but TM-align equals to 1. Those sequences are dangerous if we didnt deal with before making predictions. For example, in Nuclear Receptor group, target proteins *hsa2103* and *hsa2104* which we can also infer from their id that they are homology structures are proved with high SW-value (0.721) and TM-align (1). However, for those unconspicuous data, for example, *hsa8856* and *has9970* are actually homology structures that even on the same chain of the structure, that TM-align equals 1 while SW-value equals to 0.325. Those data can be dealt with by our method.

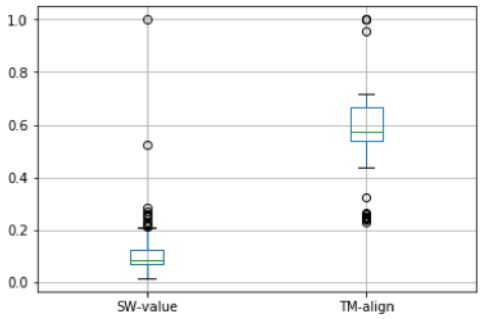

Fig. 4 Comparison of distributions of SW-value and TM-align

Other than this, sequences with low SW-values are considered as dissimilar proteins, but indeed they are similar in 3D structures. For example, the target protein hsa10161 and hsa3362 in GPCR group (SW-value=0.051, TM-align=0.668). We checked their 3D structures and found that part of structure of hsa10161 is very similar to structure of hsa3362 (Figure 3), which is ignored due to using the amino acid sequence to describe a protein. By confirming of those structures, from which the binding site structures were extracted can be proved to be more credible for representing the protein.

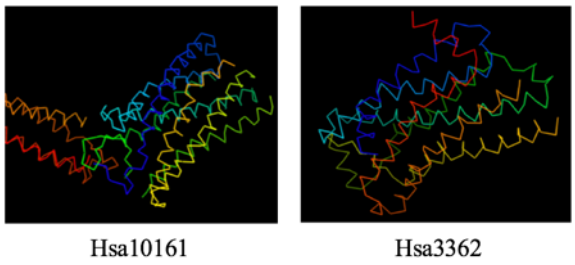

Fig. 5 Example of low sequence similarity but high structure similarity

5. Conclusion

In this paper, we developed a new method to predict drug-target interactions and presented the advantages of using similarities of binding site structures over using protein amino acid sequence similarities. Most of the previous researches used protein sequence similarities to describe the relationship of pairwise targets because sequence information can be easily achieved from the websites and similarities between them can be effectively calculated by dynamic algorithms. However, there are two problems using sequence similarities. One is sequence similarities cannot show the structure and function features, which actually influence most when a target interacting with drugs. The other problem is using sequence similarities includes remote homology sequences, which show different sequences but actually are with the same structure. Binding sites are specific places that the interactions of drugs and targets happen. By using binding site structures similarities, it will be suitable for describing the feature of the target

that interact with the drug. Thus, it will be more accurate to calculate the similarities between targets by calculating their binding site similarities. Moreover, for extracting the binding sites, we have to calculate the 3D structures of the protein using homology search, which can decrease the influence of the remote homology structures and calculate them as the same structure.

Some new drug-target interactions are found and some previous predicted interactions are checked based on our method. The improvement on AUPR value proves the good performance of our proposed method. By analyzing the reason for influence on AUC and AUPR, we found the data size and also the information imbalance on fingerprint similarities of binding sites can both influence the results. Thus, with further correction and modification in the future, the AUC and AUPR may be increased more significantly.

Future works can be focused on optimizing the calculation of the binding site similarities especially when dealing with small binding sites with few bits of fingerprints. Moreover, specializing a way to correct the protein similarity with both protein sequence similarities and binding site similarities as well as other datasets that can describe the target pairwise similarities can enrich the datasets and help with the prediction results.

Acknowledgments We thank Mr. Tomohiro Ban for giving technical supports.

References

[1] Cheng, F., Liu, C., Jiang, J., Lu, W., Li, W., Liu, G., Zhou, W., Huang, J. and Tang, Y.: Prediction of drug-target interactions and drug repositioning via network-based inference, *PLoS computational biology*, Vol. 8, No. 5, p. e1002503 (2012).

[2] Chen, X., Liu, M.-X. and Yan, G.-Y.: Drug–target interaction prediction by random walk on the heterogeneous network, *Molecular BioSystems*, Vol. 8, No. 7, pp. 1970–1978 (2012).

[3] Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W. and Kanehisa, M.: Prediction of drug–target interaction networks from the integration of chemical and genomic spaces, *Bioinformatics*, Vol. 24, No. 13, pp. i232–i240 (2008).

[4] Bleakley, K. and Yamanishi, Y.: Supervised prediction of drug–target interactions using bipartite local models, *Bioinformatics*, Vol. 25, No. 18, pp. 2397–2403 (2009).

[5] Xia, Z., Wu, L.-Y., Zhou, X. and Wong, S. T.: Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces, *BMC systems biology*, Vol. 4, No. 2, BioMed Central, p. S6 (2010).

[6] Mei, J.-P., Kwoh, C.-K., Yang, P., Li, X.-L. and Zheng, J.: Drug–target interaction prediction by learning from local information and neighbors, *Bioinformatics*, Vol. 29, No. 2, pp. 238–245 (2012).

[7] Shi, J.-Y., Yiu, S.-M., Li, Y., Leung, H. C. and Chin, F. Y.: Predicting drug–target interaction for new drugs using enhanced similarity measures and super-target clustering, *Methods*, Vol. 83, pp. 98–104 (2015).

[8] Campillos, M., Kuhn, M., Gavin, A.-C., Jensen, L. J. and Bork, P.: Drug target identification using side-effect similarity, *Science*, Vol. 321, No. 5886, pp. 263–266 (2008).

[9] Smith, T. and Waterman, M.: [a]Identification of Common Molecular Subsequences. [o] J, *Molecular Biology*, Vol. 147, pp. 195–197 (1981).

[10] Haupt, V. J. and Schroeder, M.: Old friends in new guise: repositioning of known drugs with structural bioinformatics, *Briefings in bioinformatics*, Vol. 12, No. 4, pp. 312–326 (2011).

[11] Levitt, D. G. and Banaszak, L. J.: POCKET: a computer graphies method for identifying and displaying protein cavities and their surrounding amino acids, *Journal of molecular graphics*, Vol. 10, No. 4, pp. 229–234 (1992).

[12] Laskowski, R. A.: SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions, *Journal of molecular graphics*, Vol. 13, No. 5, pp. 323–330 (1995).

[13] Hendlich, M., Rippmann, F. and Barnickel, G.: LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins, *Journal of Molecular Graphics and Modelling*, Vol. 15, No. 6, pp. 359–363 (1997).

[14] Le Guilloux, V., Schmidtke, P. and Tuffery, P.: Fpocket: an open source platform for ligand pocket detection, *BMC bioinformatics*, Vol. 10, No. 1, p. 168 (2009).

[15] Venkatachalam, C. M., Jiang, X., Oldfield, T. and Waldman, M.: LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites, *Journal of Molecular Graphics and Modelling*, Vol. 21, No. 4, pp. 289–307 (2003).

[16] Laurie, A. T. and Jackson, R. M.: Q-SiteFinder: an energy-based method for the prediction of protein–ligand binding sites, *Bioinformatics*, Vol. 21, No. 9, pp. 1908–1916 (2005).

[17] Goodford, P. J.: A computational procedure for determining energetically favorable binding sites on biologically important macromolecules, *Journal of medicinal chemistry*, Vol. 28, No. 7, pp. 849–857 (1985).

[18] Kanehisa, M. and Goto, S.: KEGG: kyoto encyclopedia of genes and genomes, *Nucleic acids research*, Vol. 28, No. 1, pp. 27–30 (2000).

[19] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic acids research*, Vol. 25, No. 17, pp. 3389–3402 (1997).

[20] Nakamura, T. and Tomii, K.: Protein ligand-binding site comparison by a reduced vector representation derived from multidimensional scaling of generalized description of binding sites, *Methods*, Vol. 93, pp. 35–40 (2016).