Hindawi
Computational and Mathematical Methods in Medicine
Volume 2018, Article ID 6565241, 5 pages
https://doi.org/10.1155/2018/6565241

Research Article

Prediction of GPCR-Ligand Binding Using Machine Learning Algorithms

Sangmin Seo,[1] Jonghwan Choi,[1] Soon Kil Ahn,[2] Kil Won Kim,[2] Jaekwang Kim,[2] Jaehyuck Choi,[2] Jinho Kim,[3] and Jaegyoon Ahn[1]

[1] *Department of Computer Science and Engineering, Incheon National University, Incheon, Republic of Korea*
[2] *Department of Life Science, Incheon National University, Incheon, Republic of Korea*
[3] *Department of Chemistry, Incheon National University, Incheon, Republic of Korea*

Correspondence should be addressed to Jaegyoon Ahn; jgahn@inu.ac.kr

Received 4 September 2017; Revised 27 November 2017; Accepted 18 December 2017; Published 30 January 2018

Academic Editor: John Mitchell

We propose a novel method that predicts binding of G-protein coupled receptors (GPCRs) and ligands. The proposed method uses hub and cycle structures of ligands and amino acid motif sequences of GPCRs, rather than the 3D structure of a receptor or similarity of receptors or ligands. The experimental results show that these new features can be effective in predicting GPCR-ligand binding (average area under the curve [AUC] of 0.944), because they are thought to include hidden properties of good ligand-receptor binding. Using the proposed method, we were able to identify novel ligand-GPCR bindings, some of which are supported by several studies.

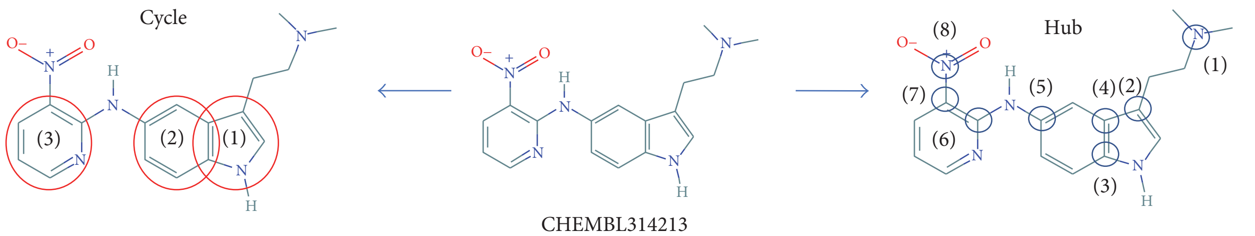
1. Introduction

G-protein coupled receptors (GPCRs) play an important role that involves detecting molecules or ligands from the outside of a cell and activating internal signal transduction pathways and cellular responses. Because the binding of an external ligand and a GPCR induces the coupling of a GPCR and G-proteins, which is followed by various forms of signal transduction, GPCRs have been extensively studied as important drug targets.

An important step for studying GPCRs as a drug target is the identification of drugs or ligands that bind to specific GPCR proteins. Naturally, many biochemistry or bioinformatics approaches to identification of drug-receptor binding have been proposed. Many of these approaches focus on calculating protein-ligand binding affinity [1–3]. However, because those methods rely on 3D structures of proteins or ligands, they cannot be used if the 3D structures of proteins or ligands are not known.

In the other class of protein-ligand identification methods, several features are extracted from proteins and ligands, and machine learning methods, such as SVM (Support Vector Machine) [4–6] or neural network [7], are applied to those features. The key principle of these approaches is that they make use of the similarities between proteins or targets. Jacob and Vert proposed a method that exploits the availability of known ligands for similar targets [6]. Geppert et al. clustered proteases into several groups and used SVM to predict ligands of proteins that have no known binding ligand [4]. Several network-based approaches have also been proposed. Iacucci et al. viewed the issue of protein-ligand binding identification as a protein-protein interaction (PPI) prediction problem [5], while Cheng et al. developed a network-based inference method to identify protein-ligand bindings [8].

In the present study, we propose a novel protein-ligand binding prediction method, which makes use of the local and global structure of a ligand and amino acid motif sequence of a GPCR. The proposed approach is not dependent on the 3D structure of a receptor. Instead, it infers hidden properties of good ligand-receptor bindings, which are encoded as a random forest classifier. We performed extensive testing of various combinations of feature sets of receptor and ligand and found various machine learning algorithms that were effective. The proposed method shows



Cycle
(1) C=CNC=C
(2) C=CC=CC=C
(3) C=CC=CC=N

CHEMBL314213
Canonical Smile:
CN(C)CCC1=CNC2=C1C=C(C=C2)
NC3=C(C=CC=N3)[N+](=O)[O–]

Hub
(1) CNCC
(2) CC=CC
(3) NC=CC
(4) CC=CC
(5) C=CCN
(6) NC=CN
(7) C=CC[N+]
(8) C[N+]=O[O–]

Figure 1: Examples of hub and cycles and how they are encoded.

a high average area under the curve (AUC) of 0.944, which indicates that local and global structures of ligands and motif sequences of GPCR are good features for prediction of strong binding.

2. Methods

2.1. Data Sources and Preprocessing. We downloaded data for 3055 GPCRs, 276,324 ligands, and 811,601 GPCR-ligand bindings from the GLASS database [9]. Among these, we included in the analysis only 213,918 ligands with a molecular weight greater than 150 and less than 500, since ligands that are too light or too heavy are likely to be ions or proteins, respectively. Accordingly, the number of final GPCR-ligand bindings was reduced to 303,587.

To obtain classification features from GPCRs, we first measure 1-amino frequencies (total 20) and 2-amino frequencies (total 400) for each GPCR. We also count the motif frequency of each GPCR. The motifs are small amino acid sequences that are specific to GPCRs and not observed in non-GPCRs. We use the MEME Suite [10] to obtain these motifs. We then exclude motifs with frequencies in all GPCRs of 1, resulting in a number of final unique motifs of 1,929. After counting the frequencies of 1,929 motifs, we have a vector with a length of 2,529 (=20 + 400 + 1,929), for each GPCR.

To obtain features from ligands, we first use four measurements: molecular weight, XlogP (a measure of the molecule's lipophilicity and solubility), hydrogen bond donors, and hydrogen bond acceptors. We name these four measurements as *4chars*. We also use a canonical simplified molecular-input line-entry system (SMILES) of a ligand, which can be parsed into a graph of which nodes and edges represent atoms and bonding, respectively.

From a graph of a ligand, we obtain two important features: a hub and a cycle. We define a hub as a node that has connection to more than 3 nodes. A hub structure is encoded as "*hub_node–A=B≡C. . .*," where *hub_node* is a hub node, *A* is the satellite node first in alphabetical order, "–" is one of the bonding types between a hub node and *A*, *B* is located next to *A* in clockwise order, "=" is one of the bonding types between a hub node and *B*, *C* is located next to *B* in clockwise order, and so on. Figure 1 shows 8 hub structures in the ligand (CHEMBL314213) and how they are encoded.

Cycles are mostly various kinds of benzene rings. Depth-first search (DFS) is used to search cycles from a graph. A cycle is encoded as "*A–B=C . . .*," where *A* is the node first in alphabetical order, "–" is one of the bonding types between *A* and *B*, *B* is located next to *A* in clockwise order, "=" is one of the bonding types between *B* and *C*, and so on. Figure 1 shows 3 cycles in CHEMBL314213 and how they are encoded.

We were able to obtain 1,031 unique hub structures and 1,741 unique cycle structures from 213,919 ligands. We then count the frequencies of those hubs and cycles for each ligand, so each ligand has a vector with a length of 2,776 (=1,031 + 1,741 + 4).

Each GPCR-ligand binding has two vectors, one from a GPCR and the other from a ligand, and we concatenate those two vectors. Therefore, each GPCR-ligand binding now has one vector with a length of 5,306 (=2,529 + 2,776).

2.2. Training Procedure. To obtain training data, we first randomly select 10,000 GPCR-ligand bindings and label them as positive. However, to train a classification model that distinguishes binding and nonbinding, we also need negative samples, which means nonbinding GPCR-ligand pairs. We used Algorithm 1 to generate 10,000 negative samples.

3. Results

We created 10 datasets, each of which consists of 10,000 positive and 10,000 negative samples, using all the attributes, all GPCR features and cycle, hub, and *4chars* for ligand features. To select a classification algorithm, we compared 10-fold cross-validation results using various algorithms including SVM [11], naïve Bayesian [12], classification and regression tree (CART) decision tree [13], random forest [14], and neural network [15], all of which are implemented in scikit-learn [16]. We can see that random forest shows the best classification accuracy.

Next we performed 10-fold cross-validation for 10 datasets using all combinations of feature sets, as shown

```
For each GPCR X from positive samples,
    Step 1. Randomly select GPCR Y from one of the positive samples;
    Step 2. If X and Y have common binding ligand(s), go to Step 1.
            Otherwise, randomly select one ligand Z among ligands that bind to Y,
            and make a negative sample with "X-Z"
    Step 3. If negative samples already have "X-Z", go to Step 1.
```

Algorithm 1: Algorithm for generating negative samples.

Table 1: Feature sets.

GPCR feature sets	Ligand feature sets
MF (motif frequency)	*4chars*
1AAF (1 amino acid frequency)	Hub
2AAF (2 amino acid frequency)	Cycle
MF & 1AAF	*4chars* & Hub
MF & 2AAF	*4chars* & Cycle
1AAF & 2AAF	Hub & Cycle
All	All

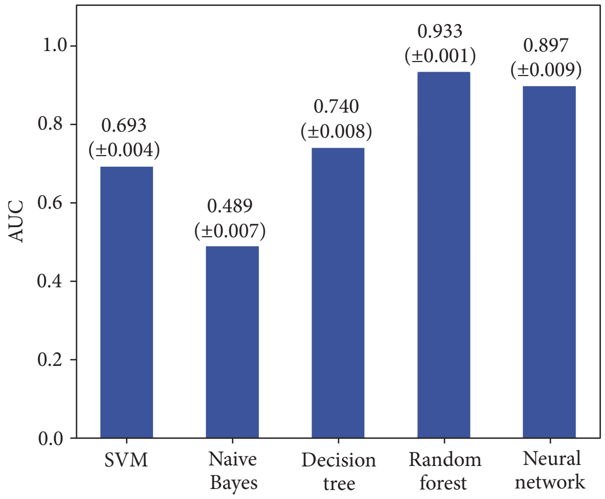


Figure 2: Classification results for various algorithms.

in Table 1. We used random forest, as it shows the best performance in Figure 2. Figure 3 shows the classification results.

Figure 3 shows that using all ligand feature sets results in the highest accuracy for all cases. Among ligand features, hub seems to be most important, since "Hub," "Hub & *4chars*," and "Cycle & Hub" are ranked 4, 3, and 2, respectively. Using "Cycle" was also relatively effective compared to "*4char*" which showed worst classification accuracy. These results show that structural data regarding ligands are key features.

Additionally, we can observe that the classification results using GPCR motifs only (first group) were the most accurate. This finding implies that motif sequences of GPCRs may play a role in ligand binding, and motif information can hold global structural information of GPCRs. We suspect that the amino acid frequency information may mask the motif information, since the classification results with GPCR motifs only show higher AUC than other cases. We can see that "MF & 1AAF" shows next best performance and "1AAF" shows better performance than "2AFF." Those results imply that 1AAF is more important than 2AAF. When only "*4chars*" are used, "MF & 1AAF" show much higher AUC than other cases, which means 1AAF can be supportive when ligand information is not enough. From those results, we can conclude that GPCR motifs are more efficient features than amino acid frequencies in prediction of binding.

We compared the prediction performance of the proposed approach with Cyscore [1]. Since Cyscore requires 3D structure of the receptors, we generated GPCR with known 3D structure from PDB (http://www.rcsb.org) [17] and their interacting ligand pairs of which number is 110,186. The same number of negative samples was also generated according to Algorithm 1. Then we randomly selected three training datasets which is composed of 10,000 positive and 10,000 negative samples. The results of Cyscore are affinity scores of GPCR-ligand pairs, so we can figure out TPR (True Positive Rate) and FPR (False Positive Rate) values varying the affinity thresholds, draw ROC curve, and calculate AUC. Figure 4 shows that our AUC values of the proposed approach are much better than those of Cyscore.

Lastly, we randomly selected 100 GPCRs that do not have binding ligands provided in the GLASS database and predicted ligands that may bind to them (provided in the Supplementary Materials (available here)). To measure the prediction accuracy, we downloaded protein-ligand bindings from BindingDB [18]. We selected 9 GPCRs which are common in BindingDB, Cyscore, and the proposed bindings and counted the number of predictions found in the BindingDB. To count true negatives (which is not predicted by methods and also not in BindingDB), we define them as X − (true positives + false positives + false negatives), where X is a set of all possible bindings of 9 GPCRs and ligands which are bound by more than one of 9 GPCRs in BindingDB. Figure 5 shows ROC curves for the predicted bindings of the proposed method and Cyscore, and we can see that AUC of the proposed method is much higher than that of Cyscore.

Among our predictions which were not found in BindingDB, we were able to identify some interesting bindings, one of which is "psychosine receptor (Q8IYL9), quercetin." Studies have suggested that quercetin may have therapeutic benefit in cardiovascular diseases [19]. The psychosine receptor is a receptor for the glycosphingolipid psychosine (PSY) and several related glycosphingolipids. Interestingly,

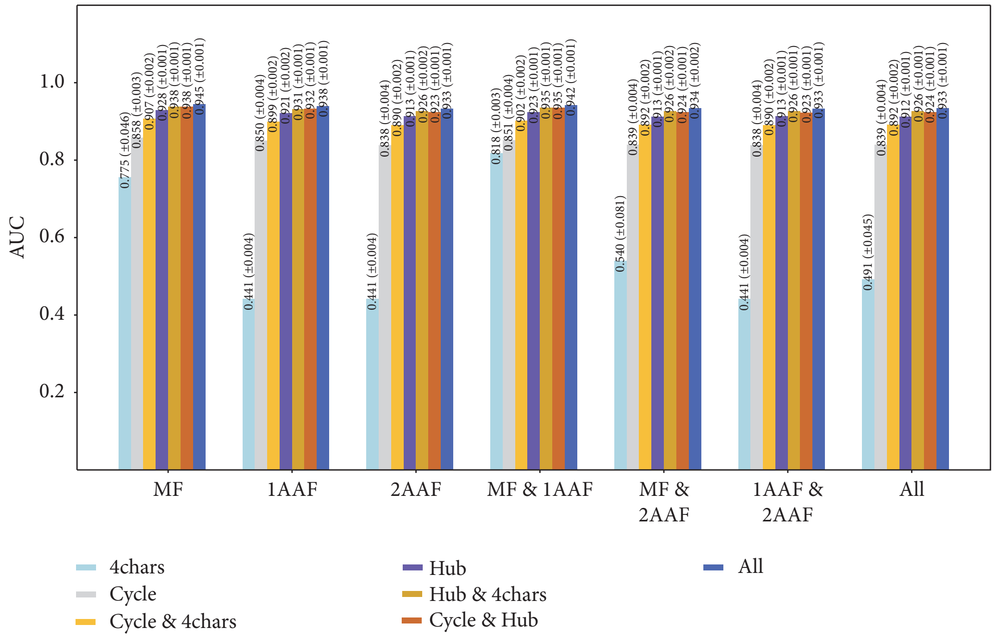


Figure 3: Classification results for all combinations of feature sets.

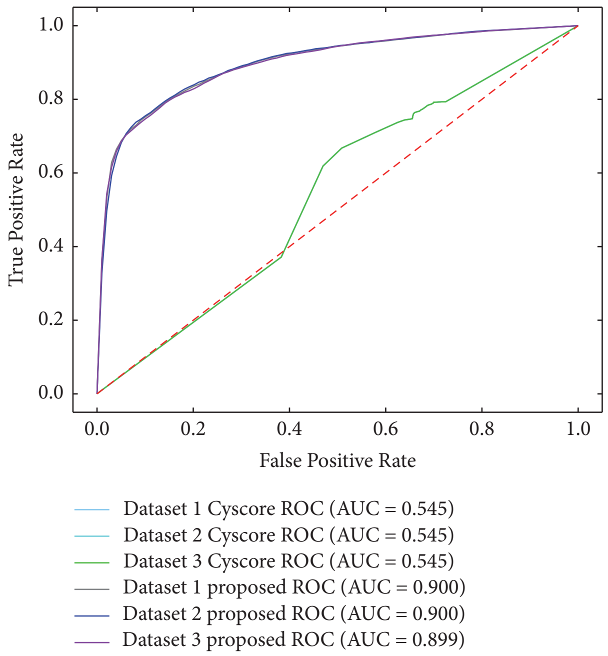


Figure 4: Comparison of prediction performance.

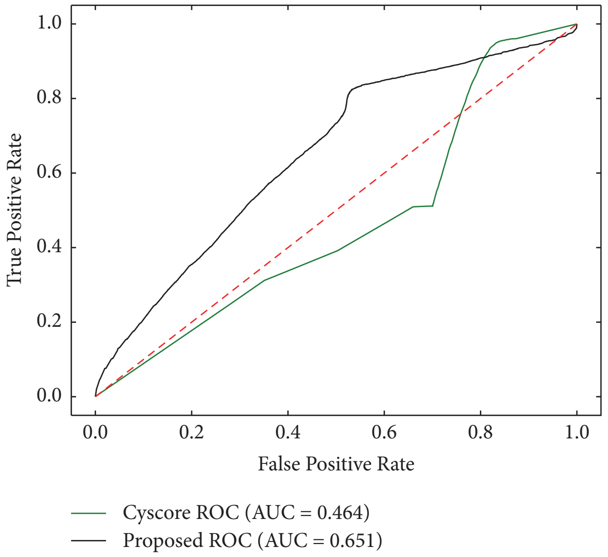


Figure 5: ROC curves for the predicted bindings.

we were able to determine that GPR65 is associated with cardiovascular diseases [20]. Our prediction that quercetin binds to the psychosine receptor may assist in elucidating the role of GPCR65 in the cardiovascular diseases. Another interesting binding is “Frizzled-2 (Q14332), garcinone E.” Frizzled-2 is a receptor for Wnt proteins and has been reported to be elevated in metastatic liver, lung, colon, and breast cancer cell lines [21]. We found that garcinone E has already been proposed as a potential drug to treat certain types of cancer [22]. Our prediction that garcinone E binds to Frizzled-2 supports the rationale behind this suggestion and further suggests the hypothesis that Frizzled-2 may be inhibited by binding of garcinone E and may thus have a therapeutic effect in these types of cancer cells.

4. Discussion

In this paper, we propose a novel protein-ligand binding prediction method. The first step in this method is the extraction of local and global features from ligands and GPCR protein sequences. These features include hub and cycle structures of ligands and amino acid motifs of GPCRs. Our experimental results show that these new features were effective in predicting GPCR-ligand binding, and these features have the potential to enable us to discover hidden properties of good ligand-receptor bindings. Our method showed much higher AUC than Cyscore, and the suspected reason of low AUC of Cyscore is that it is not optimized for prediction of GPCR-ligand binding. Using the proposed method, we were able to identify novel ligand-GPCR bindings, some of which are supported by several previous studies.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

The authors acknowledge the financial support of Incheon National University of Korea through Converging Research Program (2015-1829).

Supplementary Materials

Each supplementary material file includes the ligands which are predicted to be bound by each GPCR. They are tab delimited files where first and second columns are chemical name and InChI Key, respectively (first row has the total number of predicted ligands). *(Supplementary Materials)*

References

[1] Y. Cao and L. Li, "Improved protein-ligand binding affinity prediction by using a curvature-dependent surface-area model," *Bioinformatics*, vol. 30, no. 12, pp. 1674–1680, 2014.

[2] M. A. Khamis, W. Gomaa, and W. F. Ahmed, "Machine learning in computational docking," *Artificial Intelligence in Medicine*, vol. 63, no. 3, pp. 135–152, 2015.

[3] Q. Liu, C. K. Kwoh, and J. Li, "Binding affinity prediction for protein-ligand complexes based on β contacts and B factor," *Journal of Chemical Information and Modeling*, vol. 53, no. 11, pp. 3076–3085, 2013.

[4] H. Geppert, J. Humrich, D. Stumpfe, T. Gärtner, and J. Bajorath, "Ligand prediction from protein sequence and small molecule information using support vector machines and fingerprint descriptors," *Journal of Chemical Information and Modeling*, vol. 49, no. 4, pp. 767–779, 2009.

[5] E. Iacucci, F. Ojeda, B. de Moor, and Y. Moreau, "Predicting receptor-ligand pairs through kernel learning," *BMC Bioinformatics*, vol. 12, article 336, 2011.

[6] L. Jacob and J.-P. Vert, "Protein-ligand interaction prediction: an improved chemogenomics approach," *Bioinformatics*, vol. 24, no. 19, pp. 2149–2156, 2008.

[7] A. Givehchi and G. Schneider, "Multi-space classification for predicting GPCR-ligands," *Molecular Diversity*, vol. 9, no. 4, pp. 371–383, 2005.

[8] F. Cheng, Y. Zhou, W. Li, G. Liu, and Y. Tang, "Prediction of chemical-protein interactions network with weighted network-based inference method," *PLoS ONE*, vol. 7, no. 7, Article ID e41064, 2012.

[9] W. K. B. Chan, H. Zhang, J. Yang et al., "GLASS: a comprehensive database for experimentally validated GPCR-ligand associations," *Bioinformatics*, vol. 31, no. 18, pp. 3035–3042, 2015.

[10] T. L. Bailey, M. Boden, F. A. Buske et al., "MEME SUITE: tools for motif discovery and searching," *Nucleic Acids Research*, vol. 37, no. 2, pp. W202–W208, 2009.

[11] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *Journal of Machine Learning Research (JMLR)*, vol. 5, pp. 975–1005, 2004.

[12] H. Zhang, "The optimality of naive bayes," *AA*, vol. 1, p. 3, 2004.

[13] L. Breiman, *Classification and Regression Trees*, CRC press, 1984.

[14] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[15] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient BackProp," in *Neural networks: Tricks of the trade*, G. Montavon, G. B. Orr, and K.-R. Müller, Eds., vol. 7700 of *Lecture Notes in Computer Science*, pp. 9–48, Springer, Berlin, Germany, 1998.

[16] F. Pedregosa, G. Varoquaux, and A. Gramfort, "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[17] H. M. Berman, J. Westbrook, Z. Feng et al., "The protein data bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[18] M. K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, and J. Chong, "BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology," *Nucleic Acids Research*, vol. 44, no. 1, pp. D1045–D1053, 2016.

[19] F. Perez-Vizcaino, J. Duarte, and R. Andriantsitohaina, "Endothelial function and cardiovascular disease: effects of quercetin and wine polyphenols," *Free Radical Research*, vol. 40, no. 10, pp. 1054–1065, 2006.

[20] S. Golz, U. Brüggemeier, A. Geerts, and U. Brüggemeier, "Diagnostics and therapeutics for diseases associated with the gpr65 receptor," *Google Patents*, 2003.

[21] T. S. Gujral, M. Chan, L. Peshkin, P. K. Sorger, M. W. Kirschner, and G. Macbeath, "A noncanonical frizzled2 pathway regulates epithelial-mesenchymal transition and metastasis," *Cell*, vol. 159, no. 4, pp. 844–856, 2014.

[22] C.-K. Ho, Y.-L. Huang, and C.-C. Chen, "Garcinone E, a xanthone derivative, has potent cytotoxic effect against hepatocellular carcinoma cell lines," *Planta Medica*, vol. 68, no. 11, pp. 975–979, 2002.