DOI 10.5012/bkcs.2010.31.8.2163

Bayesian Model for the Classification of GPCR Agonists and Antagonists

Inhee Choi,[†,‡,a] **Hanjo Kim,**[‡] **Jihoon Jung,**[§] **Ky-Youb Nam,**[‡] **Sung-Eun Yoo,**[#] **Nam Sook Kang,**[#,*] **and Kyoung Tai No**[†,§,*,b]

[†]*Institute of Life Science & Biotechnology, Yonsei University, Seoul 120-749, Korea*
[‡]*Bioinformatics & Molecular Design Research Center, Seoul 120-749, Korea*
[§]*Department of Life Science & Biotechnology, Yonsei University, Seoul 120-749, Korea. *E-mail: ktno@yonsei.ac.kr*
[#]*Korea Research Institute of Chemical Technology, Yuseong-Gu, Daejeon 305-600, Korea. *E-mail: nskang@krict.re.kr*

Received March 11, 2010, Accepted June 28, 2010

G-protein coupled receptors (GPCRs) are involved in a wide variety of physiological processes and are known to be targets for nearly 50% of drugs. The various functions of GPCRs are affected by their cognate ligands which are mainly classified as agonists and antagonists. The purpose of this study is to develop a Bayesian classification model, that can predict a compound as either human GPCR agonist or antagonist. Total 6627 compounds experimentally determined as either GPCR agonists or antagonists covering all the classes of GPCRs were gathered to comprise the dataset. This model distinguishes GPCR agonists from GPCR antagonists by using chemical fingerprint, FCFP_6. The model revealed distinctive structural characteristics between agonistic and antagonistic compounds: in general, 1) GPCR agonists were flexible and had aliphatic amines, and 2) GPCR antagonists had planar groups and aromatic amines. This model showed very good discriminative ability in general, with pretty good discriminant statistics for the training set (accuracy: 90.1%) and a good predictive ability for the test set (accuracy: 89.2%). Also, receiver operating characteristic (ROC) plot showed the area under the curve (AUC) to be 0.957, and Matthew's Correlation Coefficient (MCC) value was 0.803. The quality of our model suggests that it could aid to classify the compounds as either GPCR agonists or antagonists, especially in the early stages of the drug discovery process.

Key Words: Bayesian, Classification, GPCR, Agonists, Antagonists

Introduction

Since G protein-coupled receptors (GPCRs) constitute the largest family of eukaryotic signal transduction proteins that communicate across the membrane and, as such, are associated with a multitude of diseases that make members of these families important pharmacological targets, half of all modern drug targets are GPCRs.[1,2]

In addition to their widespread appearance and highly complex function in nature, GPCRs are modulated by a plethora of diverse endogenous and exogenous ligand.[3] In fact, several ligands for GPCRs are found among the worldwide top-100-selling pharmaceutical products.[1] It is believed that a receptor molecule exits in a conformational equilibrium between active and inactive biophysical state.[4] In general, GPCR ligands can be divided into three classes: (1) agonists, which binds selectively to active receptor conformations to cause a biological response; (2) inverse agonists, which decrease the proportion of active receptor states, and thereby reduce constitutive (basal) receptor activity; and (3) antagonists, which don't alter basal response by not disturbing the resting equilibrium as they bind with equal affinity to both active and inactive receptor conformations.[4,5]

A machine-learning method helps to overcome the difficulty, the cost and time-consuming problems in the discovery of novel chemical entities in the pharmaceutical industry. This kind of method takes as input a set of objects (the training set) that have previously been determined to be either active or inactive. These training-set molecules are then analyzed to develop a decision rule that can be used to classify new molecules (the test set) into one of the two classes.[6] Various machine-learning techniques have already been suggested to be used to increase the chances of identifying novel GPCR ligands; for example, back-propagation neural networks with BCUT descriptors,[7] a combination of structure-based and property-based parameters,[8] pattern recognition techniques,[9] similarity searching and dynamic compound mapping,[10] or self-organizing neural networks with RDF descriptors.[11] Besides a wide variety of these available methods, Bayesian concepts and methodology has existed for many years to analyze structure activity data or to predict chemical properties; however, its popularity as a tool for substructural analysis within drug discovery and structure-activity analysis is somewhat recent.[6,12]

Bayesian modeling is ideal for substructural analysis due to following reasons. First, it is fast and it scales linearly with large data sets with respect to the number of molecules. This is different from methods that try to fit the data where such methods nearly always scale greater than linearly. Second, it works for a few as well as many 'good' (for e.g., active) examples because the method is not a fitting method. Thus, it is also less affected by the "curse of dimensionality" when large numbers of descriptors are used. Third, the Bayesian model weights features by assigning greater significance to characteristics that appear to distinguish good samples from baseline samples. This is quite different from other clustering methods which uses static distance functions, such as the Tanimoto Distance between two fingerprints, where all bits are given equal weight. In such methods, a small number of bits representing features important for activity may be lost among a larger

[a] Current address: Tuberculosis Research Section, Laboratory of Clinical Infectious Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892, USA.
[b] Member of Translational Research Center for Protein Function Control, Korea.

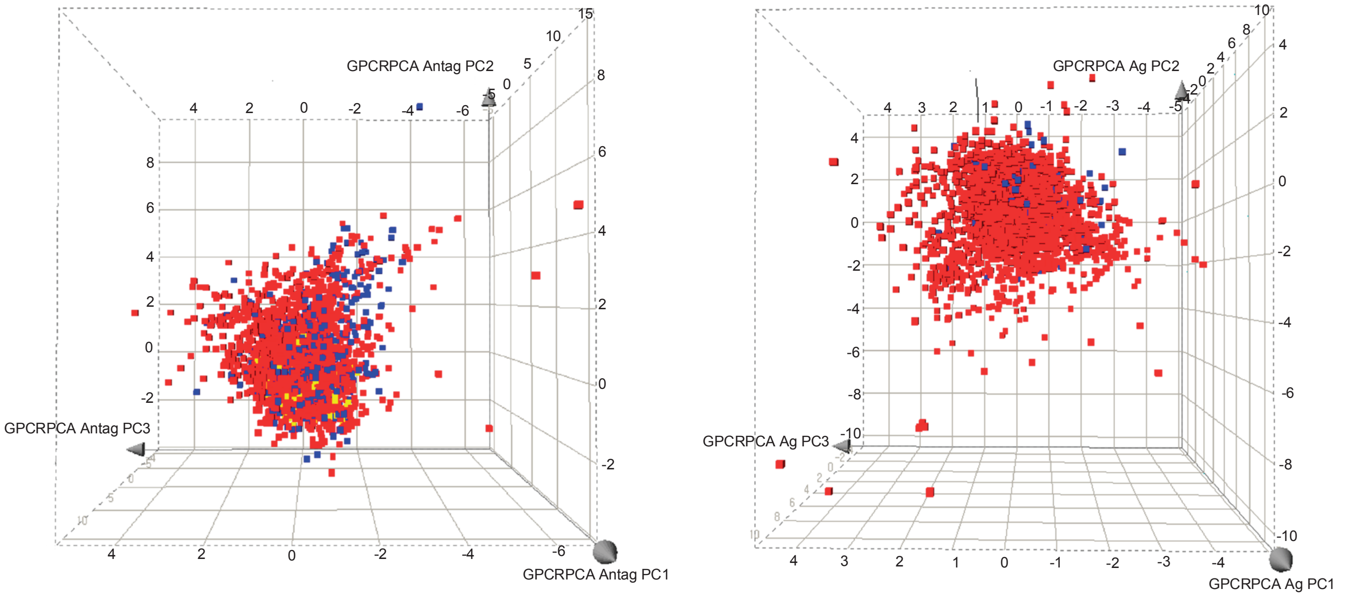


Figure 1. The PCA results of GPCR antagonists (A) and agonists (B) from GLIDA (red dots) database, Prous Science Integrity (blue dots) and Life Science Informatics Bonn website (yellow dots) in the form of a 3D scatter. The comparison is based on the first three principal components calculated from the molecular descriptors of SciTegic Pipeline Pilot and PreADMET.

number of bits representing less important features.[12] Fourth, it doesn't need tuning parameters beyond the selection of the input descriptors from which to learn.[12, 13] Last but not least, the model can display fragments that are active ('good') as well as inactive ('bad') within the dataset. Therefore, Bayesian modeling provides an ideal way to rapidly analyze data with a view to library development and compound prioritization.[12] Also, the fact that it can model broad classes of compounds and multiple modes of action can be represented in a single model,[13] is beneficial for GPCR-related research since there are so many GPCR ligands belonging to diverse classes and acting in various modes. Some successful uses of Bayesian models with large datasets have been for the classification of kinase inhibitors,[12] the prioritization of antitubercular agents,[14] and the prioritization of compound libraries toward natural product-likeness.[15]

Here we will report our investigation on a modified naïve Bayesian statistics implemented in SciTegic Pipeline Pilot,[16] and its application in the generation of generalized model that classifies GPCR agonists and antagonists. We will build our model with a whole slew of GPCR binding ligands that would cover wide range of subgroups in each class of GPCRs. For most of published works so far, only representative subgroups belonging to either class A and/or class B GPCR ligands were used. The main objective of these published works was either to classify GPCR and non-GPCR binding ligands[9] or to distinguish target- and family- selective GPCR antagonists.[10] Our model will be able to distinguish the mode of action of GPCR binding ligands, whether they are agonistic or antagonistic.

Experimental

Data sets. Experimentally determined GPCR agonists and antagonists were taken from GLIDA[17] and Prous Science Integrity databases.[18] GLIDA is a public GPCR-related Chemical Genomics database that is primarily focused on the integration of information between GPCRs and their ligands. It provides interaction data between GPCRs and their ligands, along with chemical information on the ligands, as well as biological information regarding GPCRs.[17] Prous Science Integrity enables researchers to manage and correlate chemistry and genomics data with experimental and clinical pharmacology results and with a knowledge base of disease entities.[18] Only the human full agonists and antagonists were retrieved but both partial and inverse agonists were retrieved from these databases. We have selected diverse ligands from many subgroups in order not to bias our dataset toward well-known GPCR class A ligands. In terms of biological diversity, our collections of compounds are known to belong the following GPCR classes: class A Rhodopsin like, class B Secretin like, class C Metabotropic glutamate/pheromone, and putative/unclassified Class A Orphan GPCRs. All peptides, ions, free radicals as well as dyes were removed in order to select small organic molecules only. Thus, 7345 compounds and 641 compounds from GLIDA (4742 agonists and 2603 antagonists) and Integrity Prous Science Integrity (192 agonists and 449 antagonists) databases, respectively, were collected. Additional 267 GPCR antagonists of biogenic amine receptors were obtained from the Life Science Informatics Bonn website (subgroups: dopamine, serotonin, and adrenergic GPCRs).[10]

Next, the chemical diversity was assessed and possible outliers were identified by principal component analysis (PCA) conducted on calculable molecular properties termed as "predefined set (ALogP, MW, No. of H donors, No. of H acceptors, No. of rotatable bonds, No. of atoms, No. of rings, No. of aromatic rings and No. of fragments)" in SciTegic Pipeline Pilot as well as molecular descriptors from PreADMET software, developed by BMDRC.[19] The PreADMET program provides rapid and reliable data of drug-likeness and ADME properties.

Table 1. PCA loadings obtained for selected variables with three principal components

Descriptors	GPCR Agonists			GPCR Antagonists		
	PC1	PC2	PC3	PC1	PC2	PC3
Constant	−6.282	−0.999	−0.752	−7.466	0.557	0.783
ALogP	0.104	0.315	−0.263	0.141	−0.213	0.302
Molecular_ Weight	0.004	0.0002	−0.0005	0.005	0.0004	0.0003
Num_H_Donors	0.179	−0.318	0.004	0.111	0.418	−0.288
Num_H_ Acceptors	0.149	−0.209	0.102	0.168	0.184	−0.101
Num_ Rotatable Bonds	0.0976	−0.045	−0.138	0.088	0.0819	0.115
Num_Atoms	0.0576	0.008	−0.003	0.067	−0.0002	−0.0007
Num_Rings	0.234	0.305	0.507	0.213	−0.354	−0.398
Num_ Aromatic Rings	0.331	0.229	0.215	0.270	−0.328	−0.213

Table 2. Number of compounds used in training and test sets

Data set	Training set	Test set
Agonist	1685	1632
Antagonist	1683	1627
Total	3368	3259

Table 3. Performance parameters, accuracy, sensitivity, specificity (in %), and MCC for two models corresponding to CYP3A4 training and test sets

Data Sets	Accuracy (Predictability) %	Sensitivity %	Specificity %	MCC[a]
Training	90.1% (3035/3368)	88.7% (1495/1685)	91.5% (1540/1683)	0.803
Test	89.2% (2906/3259)	85.3% (1392/1632)	93.1% (1514/1627)	0.786

$$^{a}MCC = \frac{(\mathrm{TP}\times\mathrm{TN})-(\mathrm{FP}\times\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FN})(\mathrm{TP}+\mathrm{FP})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

It can also calculate constitutional, electrostatic, physicochemical, geometrical and topological descriptors, which have been developed in response to need for rapid prediction of drug likeliness and ADME/Toxicity data. Here, PCA gives three significant PCs, which explains 87.2% of the variation in the data (54.1%, 21.7%, and 11.4%, respectively). The values of each molecular descriptor to respective PCs are shown in Table 1. The 3D scatter plot generated using Spotfire program[20] shows the distribution of compounds over the three first components in Figure 1. Several compounds that were outliers were removed from the respective data sources which could attribute to inaccurate classification by Bayesian model. Finally, we combined these sources and divided the data equally and randomly for training and test sets. The number of agonists and antagonists in both training and test sets are shown in Table 2.

Bayesian model development and validation. Laplacian modified Bayesian statistics available in SciTegic Pipeline Pilot (version 7.0)[16] was used to develop predictive model. The model is generated by computing the specified descriptors and a two class Bayesian categorization model is built based on the molecular descriptors calculated on the fly.[13] Pipeline Pilot provides proprietary descriptors *via* following fragmentation scheme: each atom is represented by a string of extended connectivity values, calculated using a modified Morgan algorithm. There are two different circular substructure descriptors available: Extended Connectivity Fingerprints (ECFPs) and Functional Connectivity Fingerprints (FCFPs).[21] The descriptor used in this study is FCFP_6.[22] It is a 2D fingerprint where the atom types are abstracted to the role that the atom plays in the molecule. The generation of an FCFP fingerprint for a molecule involves with the assignment of an atom code for the role of each heavy (non-hydrogen) atom in the molecule (HBA, HBD, positively ionized or positively ionizable, negatively ionized or negatively ionizable, aromatic, and halogen) and its neighbors.[13,22] The learning process generates a large set of Boolean features from the input descriptors, then collects the frequency of occurrence of each feature in the "good" subset and in all data samples. Weight is calculated for each feature using a Laplacian-adjusted probability estimate. The weights are summed to provide a probability estimate, which is a relative predictor of the likelihood of that sample being from the "good" subset.[12]

Once a model is built, every molecule is given a prediction score based on the contributions from each constitutive feature. This enables the compounds to be ranked in order of their probability of having either GPCR agonistic or antagonistic activities. In addition, each compound can be classified as GPCR agonist or antagonist depending on whether its score is greater than (positive) or less than (negative) a predetermined classification cut-off value (0), respectively. The test set compounds are subsequently used for prediction and an external validation purposes using the derived model.[13]

Once all the samples had predictions, an enrichment plot was generated, and the percentage of true category members captured at a particular percentage cutoff. The enrichment result table (Supplementary Table 1) shows the percentage of samples that are in that particular category, the number of category members, and the percentage of true members found. For example, in a column labeled "1%" would be the percentage of true category members (e.g., actives) that were found in the top 1% of the list, when sorted by the model score.[14] The percentile results table shows the cutoff needed to capture a particular percentage of the "good" samples (Supplementary Table 2). For each cutoff, it shows the estimated percentages of false positives and true negatives for the "non-good" samples. This table is designed to assist in picking the cutoff value that best balances the desire to capture as many "good" samples as possible, while keeping the number of false positives at a minimum. The rates shown in this table are estimates derived from the cross-validated data.[16] The category statistics table shows, for each category statistics derived from the cross-validated predictions of the model built for that category as applied to members of that category and non-members of that category (Supplementary Table 3). For each group, the number of members/nonmembers (N) is given; the mean prediction for each subset (Mean); and the estimate standard deviation of the predictions for each subset (StdDev).[16]

Results and Discussion

The use of Bayesian statistics to model for the classification of general (multiclasses) GPCR agonists or antagonists has been investigated. One of the goals of this study was to analyze GPCR agonist and antagonist relevant structures to classify GPCR agonist and antagonists. The second goal was to provide a GPCR agonist/antagonist likeness score to establish a prediction model for the probability of a molecule to be either GPCR agonist or antagonist.

The predictive abilities of this model were assessed by various statistics such as accuracy, sensitivity, specificity, and Matthew's correlation coefficient (MCC) (Table 3). The model was able to correctly classify nearly 90% of the 3259 test set compounds (1392 out of 1632 GPCR agonists and 1514 out of 1627 GPCR antagonists), which is a fairly good prediction. In addition, Matthews Correlation Coefficient (MCC) values were nearly +1 in for both training and test sets where a coefficient of +1 represents a perfect prediction.[23,24] In addition, the strength of our model is that it reflects diverse classes and subtypes of GPCRs whereas other published models were built usually with class A type ligands.

A major strength of Bayesian modeling is its ability to rank compounds according to their probability of being active. This ranking is important when prioritizing compounds for screening or for further development.[13] Thus, the ranking of active compounds in the ordered list of the test set is an indicator of the quality of the model conveniently visualized by the Receiver Operating Characteristic (ROC) plot (Figure 2) and the Enrichment curve (Figure 3). The enrichment curve plots the number of active compounds recovered *versus* the proportion of the database screened. The straight diagonal line shows how many active compounds would be recovered in a random, unbiased screening. If the samples are rank-ordered according to their likelihood of exhibiting activity and then screened, the active compounds should be found more rapidly than if they are screened at random and the plot appears as the curve shown.[12]

It is important to validate that the process would build a useful model if it were given data "sufficiently similar" to the samples in the training set.[22] This was done using leave-one-out cross-validation method. In this procedure, one sample was left out, and a model was built using the remaining samples; that model was used to predict the scores for the left-out samples. This was repeated until all samples had a prediction. The samples were then sorted by decreasing score, and the ROC plot was used to estimate the predictability of the modeling process. The ROC plot shows that a large number of true positives can be discovered with only a few false positives. A perfect model has the area under the curve (AUC) of 1.0; the ROC AUC score of our model was 0.957, which means, if given an agonist and an antagonist, and if one used the model score to guess which one the agonist was, one would be right 95.7% of the time.[13] Both of these graphs demonstrate the high discriminating power of our Bayesian model generated for the prediction of GPCR agonists and antagonists.

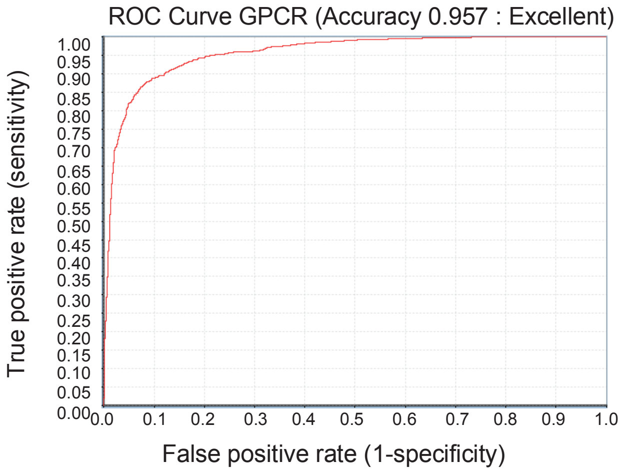


Figure 2. The performance of the model is depicted graphically by Receiver operating characteristic (ROC) plot.

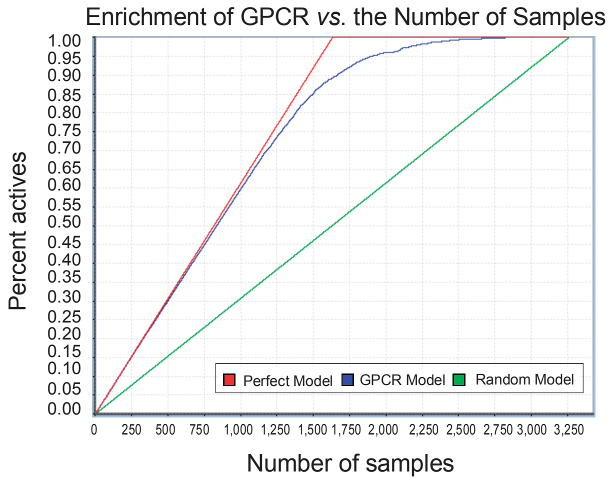


Figure 3. The enrichment plot of the percentage of GPCR agonists ("good" compounds) found (Y-axis) against the percentage of compounds screened (X-axis).

The model statistics show that our model has good enrichment rates (almost 90% of the "good" compounds occurred in the top 50% of the list) as well as percentile results (less than 50% cutoff value will lead to less than 10% false positives), while the category statistics indicate quite big separation between GPCR agonists and antagonists (Supplementary Table 3).

The representative GPCR agonistic and GPCR antagonistic features derived from FCFP_6 descriptors from this model and its frequency associated with a good compound are shown in Table 4. All 47 and 43 features of GPCR agonists and antagonists, respectively, are shown in Supplementary Figure 1. A cumulative score of feature contributions to "GPCR agonist" likeness is computed. Scores must therefore be interpreted by likelihood.[12] That is, if compounds A's Bayesian score is 90 and compounds B's score is 70, a correct interpretation is that compound A is more likely to be an agonist than compound B.[12] The normalized probability scores of the fingerprint features most closely associated with agonistic activity (G1-G20) range

Table 4. Scaffolds representing GPCR agonistic and antagonistic features from FCFP-6

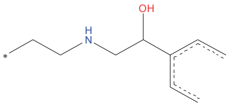
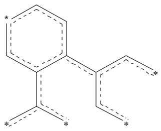
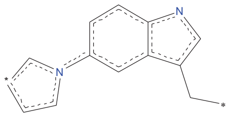
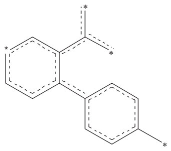
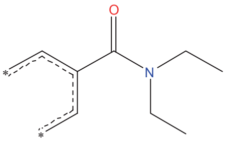
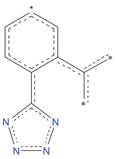
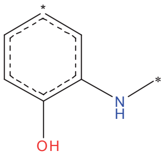
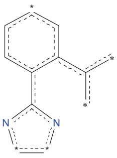
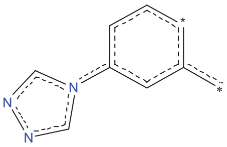
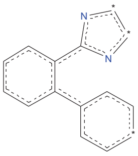
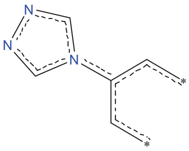
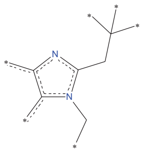
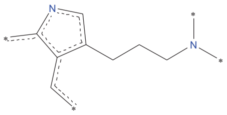
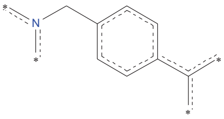
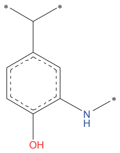
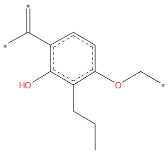
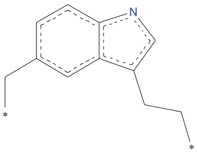
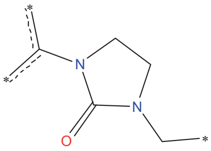
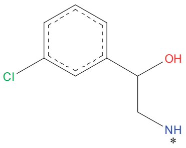
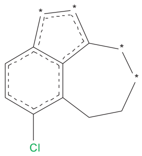
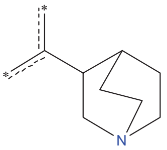
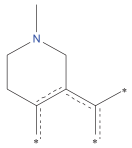
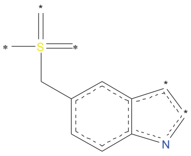
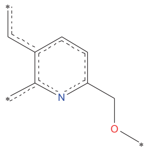
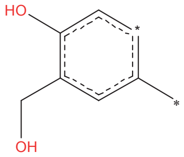
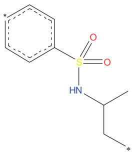
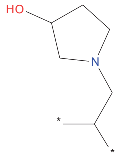
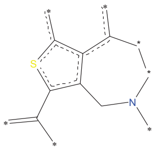
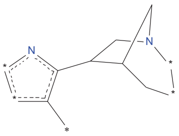
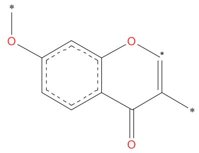
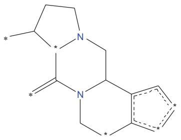
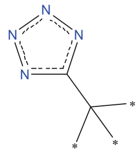
GPCR agonistic features from FCFP-6			GPCR antagonistic features from FCFP-6		
ID	Feature	Bayesian score	ID	Feature	Bayesian score
G1		0.703	B1		−3.833
G4		0.703	B3		−3.823
G7		0.702	B4		−3.801
G8		0.702	B5		−3.801
G10		0.701	B8		−3.779
G13		0.701	B9		−3.779
G14		0.700	B10		−3.597
G17		0.698	B20		−3.041
G18		0.698	B24		−2.969
G19		0.698	B25		−2.891

Table 4. Continued

GPCR agonistic features from FCFP-6			GPCR antagonistic features from FCFP-6		
ID	Feature	Bayesian score	ID	Feature	Bayesian score
G21		0.693	B29		−2.864
G24		0.692	B31		−2.864
G28		0.689	B33		−2.864
G39		0.685	B35		−2.836
G46		0.680	B43		−2.807
G48		0.678	B46		−2.746

The stars in the sketches represent ‘any atom’.

between +0.703 and +0.678. The model shows that FCFP_6 features B1 and B20 (normalized probability score ranges between −3.833 and −2.682) were associated with GPCR antagonistic activity.

Generally speaking from these analyses, GPCR agonists are composed of flexible fragments such as chains and hydrophilic groups like either aliphatic amines or hydroxyl groups. These top ranking scaffolds reflected the common pharmacophoric features of various GPCR binding ligands. For example, nonpeptide human urotensin-II agonists (Figure 4) share the following common pharmacophoric features: (i) a central nitrogen-rich scaffold such as triazole core (G10, G13), and benzimidazole like core (G4), (ii) a protonable nitrogen atom at the extremity of an aliphatic side chain (G14), (iii) one aromatic ring connected to the scaffold by linker groups of variable length (G1, G14, G19).[25]

On the other hand, GPCR antagonists have planar structures usually composed of two or three rings and aromatic amines. A well known motif found in both GPCR and non-GPCR drugs is the biphenyl substructure. For example, this motif is found in potent and selective EP3 antagonists developed by Merck and dual NK1/NK2 antagonist developed by Novartis.[3] This biphenyl fragment has been found within the top 10 antagonistic scaffolds such as B3 or B8. The common pharmacophoric groups of non-peptide human urotensin-II antagonists (Figure 5) are hydrogen-bond acceptor and donor, ionizable group (basic amine), and aromatic hydrophobic features.[26] These groups were found within the top 20 antagonistic fingerprint features.

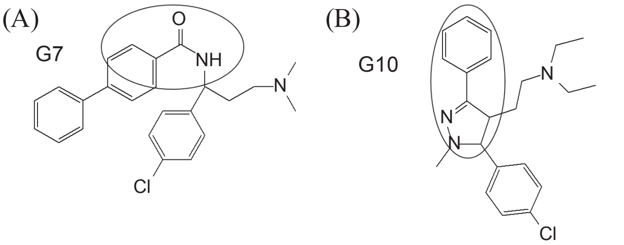


Figure 4. Several nonpeptide U-II agonists, with the GPCR agonistic features predicted by the model highlighted by a circle.

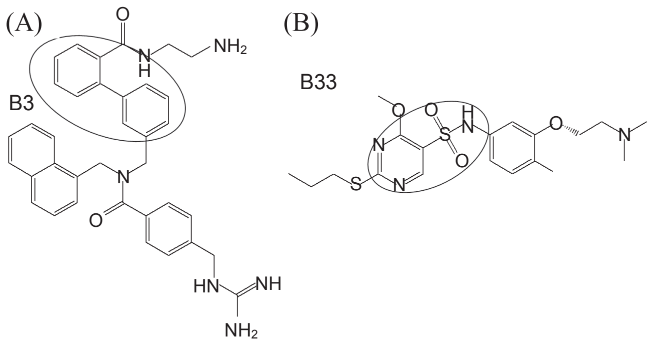


Figure 5. Several nonpeptide U-II antagonists, with the GPCR antagonistic features predicted by the model highlighted by a circle.

Among GPCR antagonistic features found from our analysis, fragments like sulfonamide group and benzodiazepine-like scaffold are usually found in clinically used antagonists.[3]

GPCRs are often the first ranking for medicinal chemists concerning the druggability of a target. The present study is a ligand-based retrospective analysis of the classification of GPCR agonistic as well as antagonistic activities using a Bayesian statistical approach. Model was built with thousands of structurally diverse compounds which were experimentally determined as GPCR agonists and antagonists. This work demonstrated how GPCR agonist or antagonist-like libraries can be generated for smart screening using this model. This model also allows identification of structural features that are associated with GPCR activities. Scaffolds conferring GPCR activities were identified; GPCR agonists are composed of flexible fragments and aliphatic amines, whereas GPCR antagonists have planar structures and aromatic amines. The Bayesian model reached about 90% of accuracy for both the training and test sets indicating strong predictive quality of the model. Thus, the general performance of this method seems to be satisfactory with significant activity enrichment.

Though many ligand-based methods allow for the retrieval of novel or alternative molecular scaffolds, the optimization of the initial hits is often considered challenging since ligand-based methods generally lack any information on how the potential ligands might bind to the receptor binding site.[27] To compensate this issue, we are planning to classify these determined GPCR agonistic and antagonistic fragments to their corresponding GPCR targets more in depth. This way structure-selectivity relationship of each GPCR class could be explored. Additional further works will be related to confirming the applicability of this technique to newly synthesized compounds once those biological assay results become available.

So far, drugs have still only been developed to affect a very small number of the GPCRs, and the potential for drug discovery within this field is enormous.[1] This Bayesian model could be applied to predict compounds in order to assist earlier identification of either GPCR agonist or antagonist during the preliminary step of drug discovery. This model could be applied to prioritize compounds for screening or to optimally select compounds from third-party data collections. Predicted scaffolds could be applied and provide information in chemical synthesis stage.

Acknowledgments. This work was supported by Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2006-005-J04501), a grant from Translational Research Center for Protein Function Control (2010-0001931), and a grant (CBM33-B2000-01-00-00) from the Center for Biological Modulators of the 21st Century Frontier R&D Program, the Ministry of Science and Technology, Korea.

Supporting Information. All supplementary Tables 1-3 and Figure 1 are available *via* the Internet, http://newjournal.kcsnet.or.kr.

References

1. Fredriksson, R.; Lagerström, M. C.; Lundin, L.-G.; Schiöth, H. B. *Mol. Pharmacol.* **2003**, *63*, 1256.
2. Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A.; Rasmussen, S. G.; Thian, F. S.; Kobilka, T. S.; Choi, H. J.; Kuhn, P.; Weis, W. I.; Kobilka, B. K.; Stevens, R. C. *Science* **2007**, *318*, 1258.
3. Bleicher, K. H.; Green, L. G.; Martin, R. E.; Rogers-Evans, M. *Curr. Opin. Chem. Biol.* **2004**, *8*, 287.
4. Brink, C. B.; Harvey, B. H.; Bodenstein, J.; Venter, D. P.; Oliver, D. W. *Br. J. Clin. Pharmacol.* **2004**, *57*, 373.
5. Ellis, C. *Nat. Rev. Drug Discov.* **2004**, *3*, 577.
6. Chen, B.; Harrison, R. F.; Papadatos, G.; Willett, P.; Wood, D. J.; Lewell, X. Q.; Greenidge, P.; Stiefl, N. *J. Comput. Aided Mol. Des.* **2007**, *21*, 53.
7. Manallack, D. T.; Pitt, W. R.; Gancia, E.; Montana, J. G.; Livinstone, D. J.; Ford, M. G.; Whitley, D. C. *J. Chem. Inf. Comp. Sci.* **2002**, *42*, 1256.
8. von Korff, M.; Steger, M. *J. Chem. Inf. Comp. Sci.* **2004**, *44*, 1137.
9. Balakin, K. V.; Lang, S. A.; Skorenko, A. V.; Tkachenko, S. E.; Ivashchenko, A. A.; Savchuk, N. P. *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 1553.
10. Vogt, I.; Ahmed, H.; Auer, J.; Bajorath, J. *Mol. Divers.* **2008**,*12*, 25.
11. Selzer, P.; Ertl, P. *QSAR Comb. Sci.* **2005**, *24*, 270.
12. Xia, X.; Maliski, E. G.; Gallant, P.; Rogers, D. *J. Med. Chem.* **2004**, *47*, 4463.
13. Hassan, M.; Brown, R. D.; Varma-O'Brien, S.; Rogers, D. *Mol. Divers.* **2006**, *10*, 283.
14. Prathipati, P.; Ma, N. L.; Keller, T. H. *J. Chem. Inf. Model.* **2008**, *48*, 2362.
15. Ertl, P.; Roggo, S.; Schuffenhauer, A. *J. Chem. Inf. Model.* **2008**, *48*, 68.
16. Scitegic Inc., 9665 Chesapeake Dr., Suite 401, San Diego, CA 92123, USA, PipeLine Pilot 7.0.1, 2008, version 7.0.1.
17. Okuno, Y.; Tamon, A.; Yabuuchi, H.; Niijima, S.; Minowa, Y.; Tonomura, K.; Kunimoto, R.; Feng, C. *Nucl. Acids Res.* **2008**, *36*, D907.
18. Science Integrity®. http://integrity.prous.com © Prous Science, a Thomson Scientific business, Stamford, CT, USA, 2007.
19. PreADMET 2.0, BMDRC, Seoul, Korea, 2008.
20. Spotfire® DecisionSite® 9.1.1. http://spotfire.tibco.com, Somerville, MA, USA, 2008.
21. Sciabola, S.; Carosati, E.; Cucurull-Sanchez, L.; Baroni, M.; Mannhold, R. *Bioorg. Med. Chem.* **2007**, *15*, 6450.
22. Rogers, D.; Brown, R. D.; Hahn, M. *J. Biomol. Screen.* **2005**, *10*, 682.
23. Matthews, B. W. *Biochim. Biophys. Acta* **1975**, *405*, 442.
24. Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A.; Nielsen, H. *Bioinformatics* **2000**, *16*, 412.
25. Lescot, E.; Bureau, R.; Rault, S. *Peptides* **2008**, *29*, 680.
26. Lescot, E.; Sopkova-de Oliveira Santos, J.; Dubessy, C.; Oulyadi, H.; Lesnard, A.; Vaudry, H.; Bureau, R.; Rault, S. *J. Chem. Inf. Model.* **2007**, *47*, 602.
27. Radestock, S.; Weil, T.; Renner, S. *J. Chem. Inf. Model.* **2008**, *48*, 1104.