Journal of
Chemical Information
and Modeling

Article

pubs.acs.org/jcim

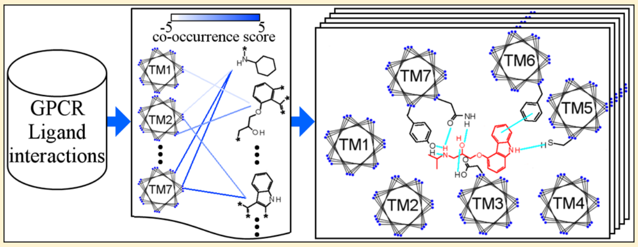
Chemical Genomics Approach for GPCR–Ligand Interaction Prediction and Extraction of Ligand Binding Determinants

Akira Shiraishi, Satoshi Niijima, J. B. Brown, Masahiko Nakatsui, and Yasushi Okuno*

Department of Systems Biosciences for Drug Discovery, Graduate School of Pharmaceutical Sciences, Kyoto University, Kyoto

S *Supporting Information*

ABSTRACT: Chemical genomics research has revealed that G-protein coupled receptors (GPCRs) interact with a variety of ligands and that a large number of ligands are known to bind GPCRs even with low transmembrane (TM) sequence similarity. It is crucial to extract informative binding region propensities from large quantities of bioactivity data. To address this issue, we propose a machine learning approach that enables identification of both chemical substructures and amino acid properties that are associated with ligand binding, which can be applied to virtual ligand screening on a GPCR-wide scale. We also address the question of how to select plausible negative noninteraction pairs based on a statistical approach in order to develop reliable prediction models for GPCR–ligand interactions. The key interaction sites estimated by our approach can be of great use not only for screening of active compounds but also for modification of active compounds with the aim of improving activity or selectivity.



INTRODUCTION

G-protein coupled receptors (GPCRs) belong to the largest group of seven transmembrane (7TM) spanning proteins involved in signal transduction,[1−3] and are among the most important target families in drug discovery.[4] GPCRs are known to interact with a variety of ligands as diverse as small molecular-weight ions, biogenic amines, nucleosides and nucleotides, peptide and protein hormones, and lipids and eicosanoids.[1−4] The GPCR superfamily has low sequence similarity in full length comparisons but contains seven highly conserved segments consisting of 25−35 consecutive residues within TM regions.[5−11] To date, hundreds of GPCR sequence motifs have been identified,[10,11] and individual motifs comprise structurally or functionally important sequences including the TM regions and ligand binding pockets. In particular, it is well-known that the rhodopsin-like GPCR family (class A), the largest in the GPCR superfamily, has ligand binding pockets within the TM regions,[2,4,12] and thus the class A ligands are often classified in terms of the similarity of the TM sequences in the known ligand target's TM region.[7,13] Information about the ligand binding modes in the TM regions is valuable for drug design and discovery, and therefore, various approaches have been developed for ligand binding site analysis.[14−18]

Virtual screening methods have played a pivotal role in exploration of the vast space of GPCR ligands. The structure-based approach[19] has provided structural insights into the ligand binding modes of some GPCRs.[14,20−23] Due to the difficulty in crystallization, however, only a handful of GPCR structures have been resolved.[24−41] The scarcity of available GPCR structures is the critical limitation of the structure-based approach. Until many more GPCRs can be crystallized, approaches relying on homology will remain a likely alternative for structure-based binding site analysis.[42]

On the other hand, the ligand-based approach[43] is being widely used due to its simplicity. However, it has difficulty in identifying new ligands when known ligands of an individual target of interest are scarce, and as a corollary, it is not applicable to orphan targets. This limitation hampers GPCR-wide analyses because there still remain more than one hundred orphan GPCRs.

Recently, the chemical genomics approach has received much attention as a state-of-the-art virtual screening strategy.[44,45] The authors previously developed a virtual screening method called CGBVS (chemical genomics-based virtual screening)[46] with the aim of finding new ligands for both well-studied and orphan targets. This method can be viewed as an integrative strategy, which leverages both ligand properties and protein sequence information. CGBVS and other related machine learning models based on ligand-target pattern analysis[47] have shown promising performance compared with the existing approaches. However, these statistical machine learning models have some limitations which we describe below.

First, there is ambiguity in the definition of noninteracting pairs. Both compound−protein interaction (CPI) pairs and noninteraction pairs are usually required to construct machine learning models. However, noninteraction pairs are often scarce in databases, and as such, compound−protein pairs without known interaction status have been selected and used as putative noninteraction pairs in most of the previous studies.[44−49] Yet, this strategy can deteriorate the prediction

Received: October 25, 2012

performance, because the randomly selected pairs might contain a fraction of positive CPI pairs. Second, there exist few ligand−target kernels that are constructed considering the physicochemical interaction between amino acids residues and ligand properties or chemical structures. Along with the question of which descriptors are suitable for representing ligands and proteins for interaction predictions, it is of fundamental importance to develop statistical models that enable identification of both chemical structures and amino acid properties that are associated with ligand binding.

To overcome these limitations, we have developed new GPCR-specific descriptions for interaction predictions, and statistical scores for evaluating the relevance of pairs of amino acid residues and chemical structures to GPCR−ligand interactions. Specifically, we propose a statistical approach which allows extraction of binding determinants together with predictions of ligand−target interactions on a GPCR-wide scale. We illustrate that the residue−fragment pairs extracted by our approach are more likely associated with ligand binding. Furthermore, in light of the second limitation mentioned above, we design a new strategy to select more plausible noninteraction pairs.

■ MATERIALS AND METHODS

CPI Data. According to the GRAFS system,[50] GPCRs are classified into five groups: rhodopsin (class A), secretin (class B), glutamate (class C), adhesion, and frizzled/taste2.[2−4] In particular, class A is the largest and most studied group, and while it is known that ligands for class A GPCRs are chemically diverse, little is known about the structural basis of ligand binding in connection with that diversity.[2−4] Therefore, we focused on class A GPCRs in the present study.

Compound−protein interaction (CPI) pairs with class A GPCRs were collected from the GVK Biosciences database.[51] This commercial database collects CPI pairs from the literature and patent information and provides the information as tags embedded in the SD chemical representation. From this database, we utilized 150 357 agonist-labeled CPI pairs and 477 763 antagonist-labeled CPI pairs with respect to 238 class A GPCRs, and the information about the GPCR sequences was obtained from GPCRDB.[1]

Kernel Methods. We utilized kernel methods to evaluate CPI pair similarities. The kernel method is often used in the machine learning field because the method can be used in nonlinear pattern analysis can refer to nonvectorial objects, and can easily integrate heterogeneous data.[52] In this paper, kernels for CPI pairs, so-called ligand−target kernels, are composed by integrating chemical similarities and protein similarities.

Chemical Kernels. We used the Tanimoto kernel with ECFP6[53] as one similarity method, and a radial basis function (RBF) kernel with 655 Dragon descriptors[54] as comparative similarity functions for internal validation. For external validation, the Tanimoto kernel with ECFP6 was used. The ECFPs were calculated using Pipeline Pilot,[55] and the Dragon descriptors were calculated using Dragon version 5.5.[54]

Protein Kernels. To measure protein similarity, a total of seven protein kernels (labeled below a−h) were used (Figure 1). The similarity was defined with full length amino acid sequences or multiply aligned TM region sequences.

It can be observed that small molecules bind within the extracellular cavity of the TM helices. For example, 11-cis retinal in bovine rhodopsin,[24] carazolol in the beta adrenergic receptor 2,[25] ZM-241385 in the adenosine A2 receptor,[26] [(3*S*)-3-amino-4-(3-hexylanilino)-4-oxobutyl]phosphonic acid in Sphingosine 1-phosphate receptor 1,[27] and vorapaxar in Proteinase-activated receptor 1,[28] etc.[29−33] are all small molecules binding to TM helices. This fact and the fact that similar molecules bind to similar receptors[7,15,16] endorse the application of statistical analysis to multiple sequence alignments of those helices or parts thereof to identify ligand binding residues. It has also been shown that for some receptors which bind large native ligands, such as the luteinizing hormone receptor, low molecular weight compounds can be designed so as to competitively bind in among the TM helices.[28]

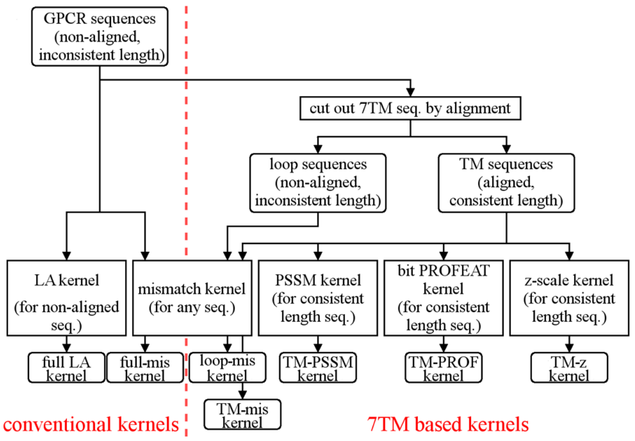

GPCR sequences (non-aligned, inconsistent length)
cut out 7TM seq. by alignment
loop sequences (non-aligned, inconsistent length)
TM sequences (aligned, consistent length)
LA kernel (for non-aligned seq.)
mismatch kernel (for any seq.)
PSSM kernel (for consistent length seq.)
bit PROFEAT kernel (for consistent length seq.)
z-scale kernel (for consistent length seq.)
full LA kernel
full-mis kernel
loop-mis kernel
TM-PSSM kernel
TM-PROF kernel
TM-z kernel
TM-mis kernel
conventional kernels
7TM based kernels

Figure 1. Workflow of protein kernel computations: (full) full length; (mis) mismatch; (PROF) bit PROFEAT; (z) z-scales.

The concept of designing small molecules competitive with large natural ligands also exists in the opioid family of receptors,[35−37] where, for example, JDTic is a large ligand of the kappa opioid receptor[34] whose binding pocket could be competitively targeted.

Recently, X-ray crystallography and NMR structure analysis have shown that peptide receptor native ligands neurotensin and substance P partially bind their respective target receptors at the TM pocket.[38,39] These results suggest pattern extraction techniques could be used for those receptors as well.

For these reasons, we compared kernels with full length sequences (below a, d), TM-only sequences (b, e, f, g), and loop-only sequences (c) (Figure 1). Each sequence kernel is as follows.

(a) Mismatch Kernel with Full Length Sequences (Mismatch-Full Kernel). The mismatch kernel is a class of string kernels which compares sequence strings representing *k*mer subsequences. The mismatch kernel allows for mutations between the subsequences. Specifically, the mismatch kernel is calculated based on shared occurrences of (k,m)-patterns in the data, where the (k,m)-patterns consist of all k-length subsequences that differ from a fixed k-length sequence pattern by at most m mismatches. In the present study, the typical choice of $k = 3$ and $m = 1$ was used in accordance with a previous study.[56] For the mismatch-full kernel, the 3−1 mismatch subsequence vectors of full length amino acid sequences were calculated, and then, these descriptor vectors were scaled to the range −1 to 1 for each descriptor and input into the RBF kernel.

(b) Mismatch Kernel with TM Sequences (Mismatch-TM Kernel). Gap-free, multiply aligned TM sequences were extracted using GPCRalign,[7,57] which uses position-specific

score matrices (PSSM) to generate the alignments. By using TM sequences generated this way, the mismatch-TM kernel implicitly contains information of multiple alignments among class A GPCRs.

The mismatch kernel with TM sequences was defined as a product of mismatch kernels computed for each sequence of the 7TM regions. This product is then postprocessed using RBF kernelization.

(c) Mismatch Kernel with Loop Sequences (Mismatch-Loop Kernel). The loop sequences consist of the remaining sequences after the removal of TM regions calculated by GPCRalign. As the mismatch-TM kernel calculation, we computed the mismatch kernel for each of the eight loop sequences and normalized the product as above.

(d) Local Alignment (LA) Kernel. For full length sequences, the local alignment kernel[58] was also applied. The LA kernel is another class of string kernels, which measures the sequence similarity by summing up local alignment scores allowing for gaps. The summation formulation gives the LA kernel mathematical properties that better suit it to kernel-type machine learning compared to using a single alignment and is consequently more effective in remote homology detection. We used the default parameters as suggested by the previous study.[58]

(e) PSSM-TM Kernel. As the 7TM sequences are gap-free, the amino acid sequences can be directly converted to numerical vectors based on PSSM scores for class A GPCRs[6] used by GPCRalign. As each of the 189 TM residues was directly substituted by the single PSSM score, the descriptor vector became 189-dimensional. Then, these descriptor vectors were scaled to the range −1 to 1 for each descriptor and input into the RBF kernel.

(f) z-Scale-TM Kernel. We also performed the 7TM sequence conversion to the z-scales.[59] The z-scales (Supporting Information Table S1) are the leading principal components obtained from 26 measured and computed physicochemical properties of amino acids and can be interpreted as hydrophobicity (z1), steric properties (z2), and polarity (z3) of amino acids. 7TM sequences were directly substituted by the z-scale vector. As 3-dimensional vectors (z1−z3) for each of the 189 TM residues were concatenated to a single vector, the dimensionality of the descriptor vector became 567. Scaling and normalization was done as above.

(g) Bit PROFEAT-TM Kernel. We also performed the substitution via the bit representation of PROFEAT.[60] The bit representation of PROFEAT also encodes the residues on the basis of their physicochemical properties, but in a different way than the z-scales. The properties used were hydrophobicity, normalized van der Waals volume, polarity, polarizability, charge, secondary structure, and solvent accessibility (Supporting Information Table S2). Each of these seven properties for amino acids is divided into three classes. Therefore, each amino acid was substituted by a 21-bit length vector. As a result, TM sequences were converted to 3969-dimensional bit vectors. Scaling and normalization was done as above.

Ligand−Target Kernels and SVM Predictions. We integrated a chemical kernel and a protein kernel for use as a ligand-target kernel. Specifically, the similarity between two ligand-target pairs is simply the product of the similarity between the two compounds and the similarity between the two proteins. The ligand-target kernels were incorporated into support vector machines (SVMs)[61] for constructing prediction models. For SVM calculations, we employed the LIBSVM library.[62] The parameters of the SVM regularization and RBF kernel were optimized using a grid search.

Co-occurrence Score for Residue−Fragment Pairs. To evaluate the relevance of each residue and chemical fragment pair for its contribution in ligand−target binding, we defined a co-occurrence score for residue−fragment pairs. Co-occurrence scores are based on measuring the dependency on co-occurring frequency for a pair of variables.[63] The frequencies used herein are the PROFEAT bits representing GPCR physicochemical properties and the chemical fragment ECFP bits.

Let r represent a specific bit in the bit representation of a protein, and let f represent a specific bit in the representation of a compound. Next we define the following probabilities:

1. $\Pr(r)$ represents the occurrence probability of a particular protein feature bit r present among the entire training set;
2. $\Pr(f)$ analogously is the probability for a specific feature bit f representing a chemical substructure among the entire training set;
3. Then, $\Pr(r,f)$ represents the joint probability of r and f appearing among the entire training set.

Our co-occurrence score is then defined as

$$\text{co-occurrence\_score}(r, f) = \log \frac{\Pr(r, f)}{\Pr(r)\Pr(f)}$$

which quantifies their relevance of a specific pair of protein and compound features such that higher scores indicate stronger relevance to ligand−target interactions.

Co-occurrence Score for GPCR−Ligand Pairs. An intermediate problem that must be resolved is how to derive a two-class model in the presence of “positive only” data. To select plausible negative samples from unlabeled data, we define a co-occurrence score for GPCR−ligand pairs (p, c) as

$$\begin{aligned}&\text{co-occurrence\_score}(p,\ c)\\&\quad = \max_{(r,f)} \text{co-occurrence\_score}(r, f)\end{aligned}$$

where (r, f) runs over all combinations of residue−fragment pairs occurring in (p, c). The selection criterion using this score is given below.

Internal Validation. To compare the performance between the proposed kernels with the conventional kernels[56,58] (Figure 1), we randomly selected from the GVK Bioscience database 4000 ligand−GPCR pairs for each of the respective CPI training and test pairs. The training and test pairs for noninteraction were generated by combining compounds and proteins at random. To avoid the effect of difference in the numbers of ligands available for each GPCR,[46] we selected the same number of CPI pairs for each GPCR used in training and test data.

External Validation. To validate our negative sampling strategy, we randomly selected 15 000 CPI pairs as positive training samples from the GVK Bioscience database. The same number of noninteraction pairs (negative training samples) were generated by two methods:

(1) Generate pairs by randomly combining compounds and proteins that are not included in CPI pairs (Table 1; model 1);

Table 1. External Validation Data Sets and Methods[a]

model number	positive training samples	negative training samples
model 1	GVK	GVK nonpositive random CPI, no selection filter
model 2	GVK	GVK nonpositive random CPI, filter by low co-occurrence score
model 3	GLL	GLL nonpositive random CPI, no selection filter
model 4	GLL	GLL nonpositive random CPI, filter by low co-occurrence score
model 5	GLL	GDD

[a] Data sources and compound−protein interaction (CPI) data sets used in tests for model construction with external validation are listed. Negative CPIs were generated by combining, at random, one compound and one protein from the positive training sample database such that the new pair was not an existing positive training sample. In all models, the number of positive and negative CPI pairs sampled or generated was fixed at 15 000 unique samples each. (GVK: GVK Biosciences Database, GLL: GPCR Ligand Library, GDD: GPCR Decoy Data set).

(2) Generate pairs in the same way as step 1, and use only pairs with co-occurrence scores smaller than 1 (Table 1; model 2).

We further used the GPCR Ligand Library (GLL) data set[64] as positive training samples. The same number of noninteracting pairs were generated in the same way as (1) (Table 1; model 3) and (2) (Table 1; model 4). In addition, we also tested the decoy pairs provided by the GPCR Decoy Data set (GDD)[64] as negative training samples (Table 1; model 5).

As an external test set for all models trained, we used interaction and noninteraction pairs in the GPCR SARfari database,[65] which provides valuable quantitative binding affinity of compound−protein pairs. We extracted more than 100 000 pairs with K_i values, consisting of over 50 000 compounds against 158 class A GPCRs. We dichotomized pairs into active and inactive classes on the basis of the values of K_i. Since the threshold for labeling a compound as "active" is usually set at 1 or 10 μM, GPCR-compound pairs with $K_i < 1$ uM were used as a positive set, and those with $K_i > 100$ uM were used as a negative set. As a result, 43 160 active pairs and 17 453 inactive pairs were extracted.

From these pairs, any pairs containing compound components such that the compound ECFP fingerprint representation contains more than five fragments not also present in the training set were further filtered out, in order to account for the applicability domain.[66] The resulting numbers of pairs available were 1472 CPIs and 1217 non-CPIs.

■ RESULTS AND DISCUSSIONS

The design of kernels significantly affects the performance of CPI predictions. Several kernels have been proposed to encode protein sequence information (Figure 1). The spectrum and mismatch kernels[67] for full length amino acid sequences are among the most popular protein kernels because of their general applicability. However, these kernels have difficulty in the interpretation between physical interaction of amino acids and chemical fragments. To address this problem, we have developed TM-based kernels using the aligned sequences of TM regions. In contrast to the popular protein kernels, the fixed length TM sequences and their 7TM alignment-based descriptions are available not only for kernel construction but also for structural and physical interpretation of compound−protein interactions. Using the TM-based descriptions, we compared CPI prediction performance and evaluated the relevance of residue−fragment pairs.

Comparison of 7TM-Based Kernels with Conventional Kernels in CPI Prediction. We compared the prediction performance of different GPCR kernels for internal validation. In the experiments, we randomly selected 4000 ligand−GPCR pairs for training and test sets, respectively, repeated this process 20 times, and evaluated the performance in terms of average accuracy and metrics calculated from rate of change (ROC) curves.[68] Note that for internal validation, the negative pairs were randomly selected from unknown interaction pairs in the same way as in a previous study.[46]

First, we examined the isolated loop and TM regions in comparison to full length sequences. As shown in Figure 2A−C, the mismatch kernel for full-length sequences combined with compound ECFP fragments achieved an accuracy of 83.32 ± 0.36%. Compared to the full-length mismatch kernel, the 7TM sequence mismatch kernel yielded substantially higher

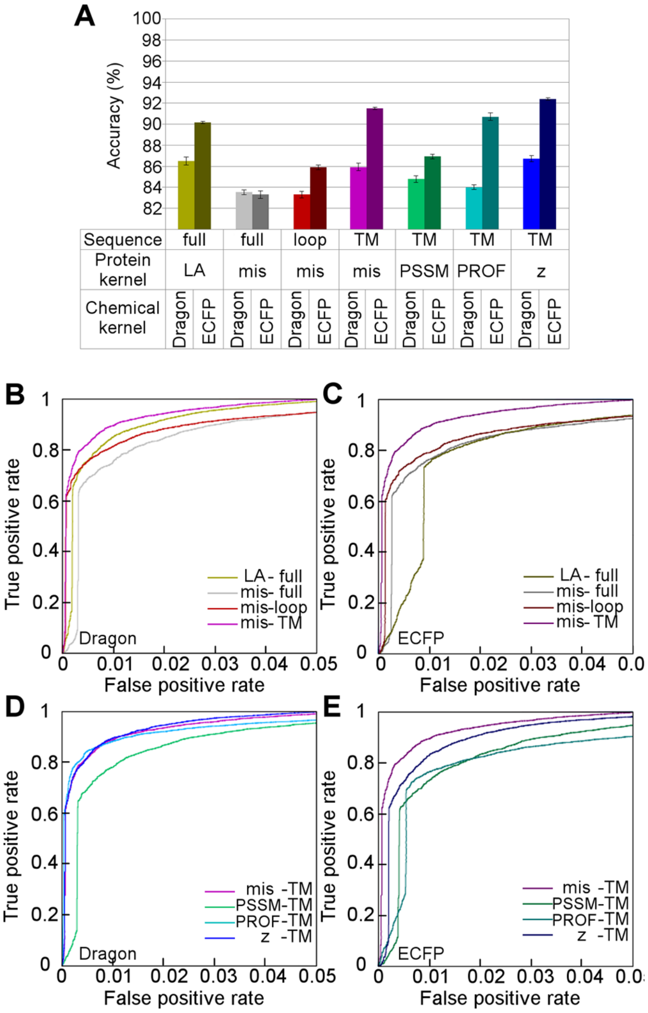


Figure 2. Performance comparison between GPCR kernels encoding different protein information. (A) Average accuracy and standard deviation of internal validation for each kernel. (B) ROC curves of the models based on Dragon descriptors for the first four GPCR kernels. (C) ROC curves of the models based on ECFPs for the first four GPCR kernels. (D) ROC curves of the models based on Dragon descriptors for the second four GPCR kernels. (E) ROC curves of the models based on ECFPs for the second four GPCR kernels.

performance (91.52 ± 0.11% with ECFP). In contrast, the loop sequence mismatch kernel showed no significant difference (85.91 ± 0.21% with ECFP). This result is consistent with the fact that the TM regions are involved mainly in ligand recognition for class A GPCRs, although allosteric binding sites[69,70] as well as some interactions in which loop regions play a relevant role cannot be recognized due to discarding the loop region. Moreover, the TM-based kernel compares favorably to the full length LA kernel (90.18 ± 0.11% with ECFP), which encodes local pairwise alignment information. Taken together, the performance comparison suggests the TM regions of class A GPCRs are crucial for ligand binding.

Next, we focused on TM-specific encoding methods of amino acid residues. Specifically, bit PROFEAT-TM, PSSM-TM, and z-scale-TM kernels were constructed by directly replacing amino acid residues in TM regions with numerical vectors. As shown in Figures 2A, D, and E, the bit PROFEAT-TM and z-scale-TM kernels showed higher prediction performance (90.71 ± 0.20% and 92.40 ± 0.04%, with ECFP respectively) than the PSSM-TM kernel (86.93 ± 0.31% with ECFP). Both kernels encode amino acids using physicochemical properties, whereas the PSSM-TM kernel does not encode that information. As such, the multiple alignments for the TM regions and physicochemical properties of the aligned residues are crucial for predicting its interacting molecules.

To compare how different chemical descriptors affect the performance in combination with different GPCR kernels, we compared the well-known Dragon chemical descriptors[54] to ECFPs.[53] Overall, ECFPs consistently performed better than the Dragon descriptors, albeit to varying degrees depending on the GPCR kernels used.

The bit PROFEAT-TM and z-scale-TM kernels with ECFPs can not only yield high prediction performance but also make both chemical and biological interpretation easier. In contrast, although the mismatch-TM kernel also shows better performance with ECFPs, the mismatch kernel loses information about the position of amino acid residues. In this regard, because it is a bit-type fingerprint, the PROFEAT-TM kernel is suitable for the interpretation of interaction and further allows for the selection of more plausible noninteraction pairs as shown below.

Distribution of CPI Pairs. In the previous section, we have shown that the bit PROFEAT-TM kernel with ECFPs resulted in high performance. Here, we compared the distributions of co-occurrence scores for CPI pairs and those of pairs whose interaction is unknown. The co-occurrence score is calculated for each residue–fragment pair with the bit PROFEAT kernel and ECFPs. As a result, both scores showed extreme value distributions with different peaks (Figure 3). As can be seen, a substantial portion of the distribution of unknown interaction pairs overlaps with the distribution for known CPI pairs. In most of the previous studies, compound–protein pairs without known interaction have been used as noninteraction pairs.[44−49] However, as Figure 3 demonstrates, this strategy has a limitation in that such pairs might contain positive CPI pairs and, hence, deteriorates the prediction performance. To alleviate this problem, we selected unknown pairs with low co-occurrence scores as negative noninteraction pairs, as elaborated next.

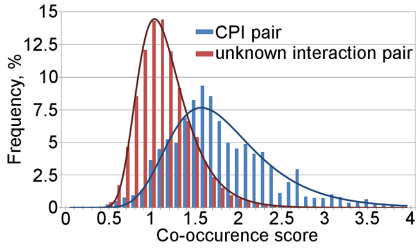


Figure 3. Co-occurrence score distributions for CPI pairs and unknown interaction pairs.

Selection of Negative Pairs with Low Co-occurrence Scores Improves the Prediction Performance. To validate our strategy for selecting the negative pairs, we compared a conventional prediction model based on negative pairs that were randomly generated from unknown pairs (Table 1, model 1) with a model constructed from pairs selected by our proposed strategy (Table 1, model 2), based on the assumption that negative pairs should have low co-occurrence scores. We employed the 1472 positive pairs and 1217 negative pairs from the GPCR SARfari database for external validation.

As can be seen in Figure 4, the area under the ROC curve (AUC) of the proposed model was 0.75, whereas the AUC of the conventional model was 0.69, showing that the prediction performance could be improved by our proposed strategy. Given the promiscuous nature of GPCRs,[46,71] the negative training pairs generated from known ligands potentially contain pairs that may actually be false negatives, and this likely made a difference in the performance. This result indicates that the performance can be further improved by generating negative training pairs that are more dissimilar to known ligand–GPCR pairs.

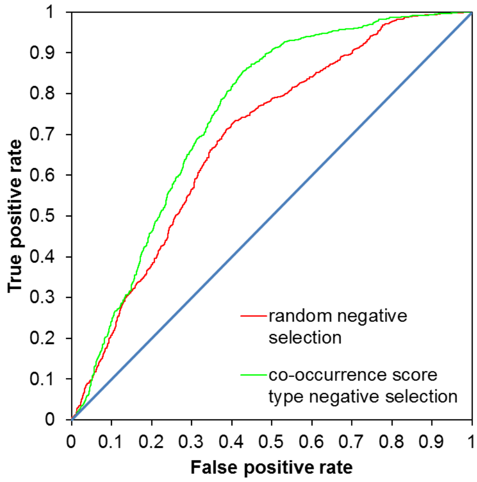


Figure 4. Performance comparison between conventional and proposed models using the GVK database. ROC curves for a conventional model based on negative pairs randomly generated from unknown pairs (red, model 1) and for our proposed model based on the negative pairs with low co-occurrence scores (green, model 2).

For more validation, we compared performance using our generated negative samples with the performance obtained by using available decoy interactions. For the comparison, we used the GLL/GDD databases. 10000 pairs from GLL were employed as positive training data. The same number of negative training data was collected from GDD (Table 1, model 5) or generated in the same ways as the previous section (Table 1, models 3 and 4).

As can be seen in Figure 5, AUC of the proposed model (model 4) was 0.75, whereas the AUC of the conventional model (model 3) was 0.70 and decoy model (model 5) was 0.69, showing that the prediction performance could be improved by our proposed strategy even if compared using a decoy database. The GDD or other decoy database compounds are generated based on the active compound to a single target. On the other hand, our method can be regarded as a decoy generation method based on both protein and compound information. The reason for higher prediction performance using the proposed selection technique can be partly attributed to decoy generation that considers not only ligand information, but rather the combination of both protein information and ligand information, which taken as a pair may be more informative than ligand dissimilarity alone.

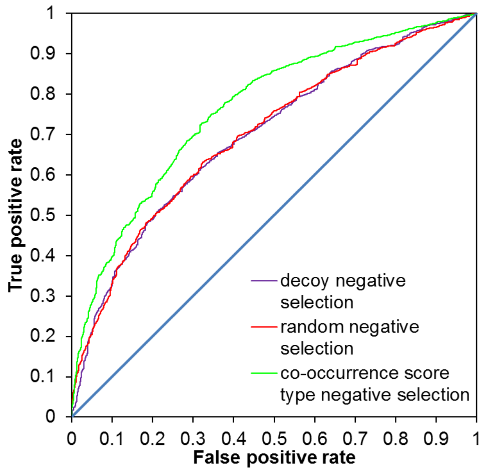


Figure 5. Performance comparison between conventional and proposed models using the GLL database. ROC curves for a conventional model based on negative pairs randomly generated from unknown pairs (red, model 3), our proposed model based on the negative pairs with low co-occurrence scores (green, model 4), and negative pairs randomly selected from the GDD decoy pairs (purple, model 5).

The prediction performance for the external data set is not as high as for internal validation. One of the reasons can be the different distributions of similarity between positive samples and negative samples. Figure 6 shows the distribution of kernel values between negative samples and their top 0.1% most similar positive samples. Notice for the GPCR SARfari external data set that there are a considerable number of noninteraction pairs with high similarity to interacting pairs in the training set, making the test set more difficult to discriminate. With a distribution of negative samples considerably different from the training set negative CPI distribution, we can understand why prediction on the external set was less successful than for internal validation.

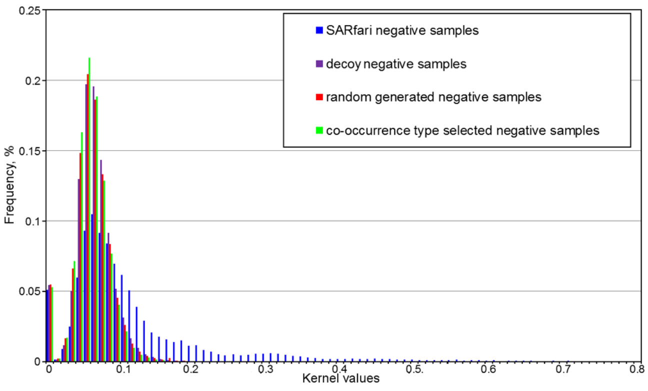


Figure 6. CPI similarity distribution. The distribution of kernel values between negative samples and their top 0.1% most similar positive samples.

Relevance of Top-Ranking Residue–Fragment Pairs to Known Important Residues. In the previous section, we showed that the co-occurrence score is of great benefit for selecting negative training samples from unknown interaction pairs. This in turn suggests that the top-ranking residue–fragment pairs might have relevance to physical interaction. To illustrate this, we compared the co-occurrence score and physical interaction for each residue–fragment pair by examining two recently resolved crystal structures: beta-2 adrenergic receptor (ADRB2)[25] and C-X-C chemokine receptor type 4 (CXCR4),[41] both of which are complexed with their ligands (PDB ID: 2RH1 for ADRB2 and 3ODU for CXCR4).

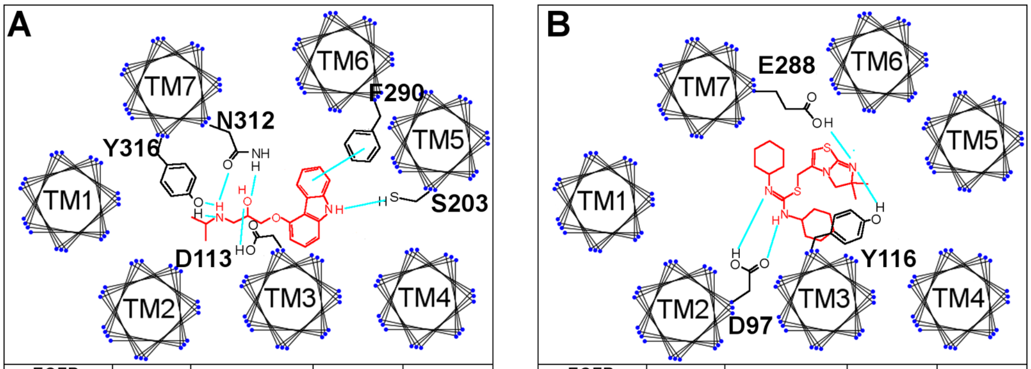
Carazolol is known to contact with D113, S203, F290, N312, and Y316 of ADRB2[25] (Figure 7A). We found that all of the fragment–residue pairs being in contact had high co-occurrence scores compared with the other pairs. In particular, the pair of N312 bounded to a hydroxyl fragment appears multiple times in the top ranking residue–fragment pairs. This is in accordance with the conserved property of amine recognition via N312 as reported in a previous study[72] and consistent with the observation that N312 plays an essential role in ligand binding for ADRB2.[73]

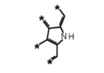
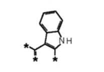
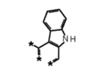
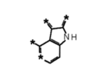
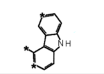
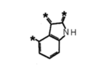
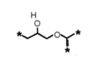
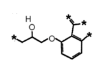
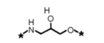
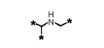
We next examined the peptide receptor CXCR4 complexed with a nonpeptide ligand, isothiourea-1t (Figure 7B). The three residue–fragment pairs involved in hydrogen bonding have high scores for CXCR4 as well as for ADRB2. The CXCR4 ligands are known to bind to one of several typical GPCR binding pockets[7] and to lie in between TM helices 2, 3, and 7.[41] The high co-occurrence scores of D97 and Y116 with one end of the ligand and E288 with the other end likely reflect the known bridge-like binding mode.[7,41]

For the GPCR superfamily, this bridge-like binding mode results in two critical ligand-binding pockets in which the ligand must attach.[7] Consistent with this, our analysis suggests the importance of physical interaction within TM3-6 for ADRB2 and TM7-2 for CXCR4. This observation shows that the residue–fragment pairs with high co-occurrence scores are indicative of the conserved binding modes across GPCR families.

In earlier studies,[8,9,18] amino acid residues relevant to ligand binding were detected by mutational analysis. Yet, these analyses have a limitation in that which chemical substructures prefer the relevant residues remains unclear. In contrast, our approach is capable of extracting the association between chemical substructures and residues, thus allowing more in-depth analysis about residue–fragment pairs that are responsible for ligand binding, and importantly, further enabling the analysis of GPCR–compound pairs without requiring cocrystallization.

About 5.4% of residue–fragment pairs possessed positive co-occurrence scores, with the top 1% possessing co-occurrence scores exceeding 2.0. We evaluated whether these pairs are associated with known residues relevant to physical interaction as reported in the literature.[8] As a result, 37/47 pairs possessed



ECFP fragment	Residue	Property	Score	Distance
	S203	middle hydrophilic	1.15	3.32
	F290	high hydrophobic	0.81	3.67
	S203	middle hydrophilic	2.00	3.32
	F290	high hydrophobic	1.53	3.67
	S203	middle hydrophilic	2.12	3.32
	F290	high hydrophobic	1.65	3.67
	S203	middle hydrophilic	1.59	3.32
	F290	high hydrophobic	1.22	3.62
	S203	middle hydrophilic	1.21	3.32
	F290	high hydrophobic	1.66	3.62
	S203	middle hydrophilic	1.21	3.32
	F290	high hydrophobic	0.84	3.52
	D113	negative charge	1.54	2.61
	N312	high solubility	4.64	2.77
	D113	negative charge	1.57	2.61
	N312	high solubility	5.09	2.77
	N312	high solubility	1.43	2.77
	Y316	middle solubility	1.73	3.43

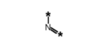
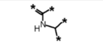
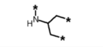
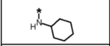
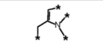
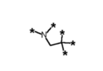
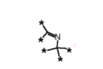
ECFP fragment	Residue	Property	Score	Distance
	D97	negative charge	1.03	4.44
	D97	negative charge	1.08	2.92
	D97	negative charge	1.03	2.92
	D97	negative charge	3.04	2.92
	E288	0.2-0.4 polarizability	0.68	4.84
	Y116	middle hydrophillic	3.16	4.22
	E288	negative charge	0.22	3.86
	Y116	high volume	1.46	4.22
	E288	0.2-0.4 polarizability	0.63	2.84

Figure 7. Association of co-occurrence scores with known crystal structures for GPCRs. Carazolol interacts with ADRB2 at 5 TM residues in the crystal structure (PDB ID: 2RH1) as shown in part A. Isothiourea-1t interacts with CXCR4 at 3 TM residues in the crystal structure (PDB ID: 3ODU) as shown in part B. These residues and their associated chemical substructures possessed high co-occurrence scores. The column labeled "distance" indicates the distance in angstroms in the PDB structure between the row's residue and chemical fragment.

co-occurrence scores satisfying this criterion (Figure 8 and Supporting Information Figure S1). For example, the binding conformation of the receptor P2Y13 has already been predicted by homology modeling and docking.[74] The high co-occurrence scores for the adenosine-like fragment with residues 6.55 and 7.35 are in accordance with the reported binding mode (Figure 8). Further, all of the residues relevant to ligand recognition of peptide receptors, e.g. CCR5, GNRHR, and C5AR (Figure 8 and Figure S1) with their ligands, had high scores, each with distinct peptide structures. This result implies that varied residues in TM regions may be involved in ligand specificity and selectivity.

Taken together, these results suggest that the co-occurrence score is an indicator of great use for dissecting the physical interaction of ligands. Of note, unlike previous studies,[8,9,18] our approach is capable of extracting important residues together with their chemical substructures, even in the case where cocrystal structure is unavailable.

In the early process of drug discovery, lead compounds are subjected to modification to improve activity, selectivity, or properties such as drug-likeness. Co-crystal structures with the target protein give us a hint as to how to optimize the compounds without losing bioactivity by focusing on the binding conformation. The paucity of cocrystal structures, however, leads to the lack of clues for optimization. Our proposed co-occurrence scores inform us of the key interacting sites without relying on structural information, providing an alternative guide for lead optimization.

■ CONCLUSION

The comparison of different kernels revealed that 7TM-based kernels are effective for GPCR compound–protein interaction prediction. In contrast to the mismatch and LA kernels, the bit PROFEAT-TM kernel provides a simple yet useful approach for identifying and interpreting the residue–fragment pairs associated with ligand binding, while maintaining high prediction performance. Notably, our statistical approach can be used for selecting more plausible negative noninteraction pairs to construct more reliable models. Indeed, we have shown

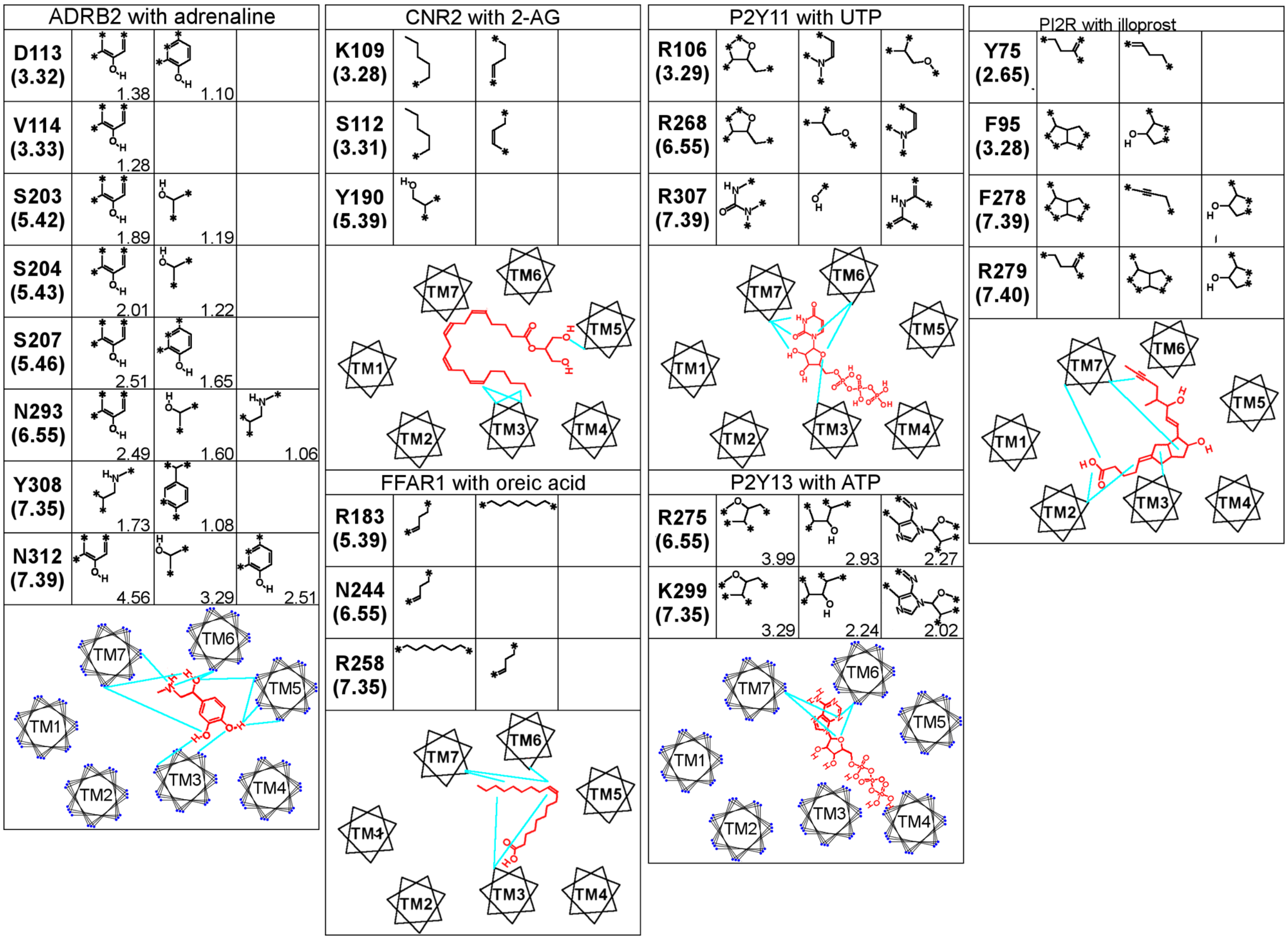


Figure 8. Co-occurrence scores for pairs of chemical fragments and known interacting residues. We focused on known interacting residues in nonorphan GPCRs[8,9,18] which are not crystallized with native ligands yet. These residues and their natural ligand substructures showed high co-occurrence scores. Combining the individual residue−fragment pairs with high co-occurrence scores, we can postulate an approximate interaction conformation.

that the prediction performance could be improved by selecting negative pairs with low co-occurrence scores when compared with model construction using all possible randomly generated negative pairs. Further, we have illustrated that the residue−fragment pairs with higher co-occurrence scores are more likely associated with ligand binding.

Taken together, the key interaction sites estimated by our approach have significant implications for ligand screening of large compound libraries as well as for lead optimization with the aim of improving activity or selectivity. Therefore the proposed methods can serve as a complementary approach to the structure-based analyses.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Figure S1. Co-occurrence scores for pairs of chemical fragments and known interacting residues. We focused on known interacting residues in nonorphan GPCRs which are not crystallized with native ligands yet. These residues and their natural ligand substructures showed high co-occurrence scores. Combining the individual residue−fragment pairs with high co-occurrence scores, we can postulate an approximate interaction conformation. Table S1. z-Scales for the standard 20 amino acids. The z-scales are the leading principal components obtained from 26 measured and computed physicochemical properties of amino acids. Table S2. Properties of bit-type PROFEAT. In the PROFEAT calculations, seven amino acid physicochemical features with three categories are defined (first column for features, and second column for categories). These categories are represented as a 21-bit code for each amino acid. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: okuno@pharm.kyoto-u.ac.jp.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by Grant-in-Aid for JSPS Fellows 23-3990; the Funding Program for Next Generation World-Leading Researchers; a Grant-in-Aid for Scientific Research on Innovative Areas “HD Physiology” of the Ministry of Education, Culture, Sports, Science and Technology, Japan; and the New Energy and Industrial Technology Development Organization (NEDO) under the Ministry of Economy Trade and Industry of Japan. Financial support from Ono Pharmaceutical Co., Ltd., is also gratefully acknowledged.

REFERENCES

(1) Vroling, B.; Sanders, M.; Baakman, C.; Borrmann, A.; Verhoeven, S.; Klomp, J.; Oliveira, L.; de Vlieg, J.; Vriend, G. GPCRDB: information system for G protein-coupled receptors. *Nucleic Acids Res.* **2011**, *39*, D309−19.

(2) Kristiansen, K. Molecular mechanisms of ligand binding, signaling, and regulation within the superfamily of G-protein-coupled receptors: molecular modeling and mutagenesis approaches to receptor structure and function. *Pharmacol. Ther.* **2004**, *103*, 21−80.

(3) Jacoby, E.; Bouhelal, R.; Gerspacher, M.; Seuwen, K. The 7 TM G-protein-coupled receptor target family. *ChemMedChem.* **2006**, *1*, 761−82.

(4) Klabunde, T.; Hessler, G. Drug design strategies for targeting G-protein-coupled receptors. *ChemBioChem.* **2002**, *3*, 928−44.

(5) Gether, U. Uncovering molecular mechanisms involved in activation of G protein-coupled receptors. *Endocr. Rev.* **2000**, *21*, 90−113.

(6) Attwood, T. K.; Findlay, J. B. Fingerprinting G-protein-coupled receptors. *Protein Eng.* **1994**, *7*, 195−203.

(7) Surgand, J. S.; Rodrigo, J.; Kellenberger, E.; Rognan, D. A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors. *Proteins* **2006**, *62*, 509−38.

(8) Sanders, M. P.; Fleuren, W. W.; Verhoeven, S.; van den Beld, S.; Alkema, W.; de Vlieg, J.; Klomp, J. P. ss-TEA: Entropy based identification of receptor specific ligand binding residues from a multiple sequence alignment of class A GPCRs. *BMC Bioinf.* **2011**, *12*, 332.

(9) Ye, K.; Lameijer, E. W.; Beukers, M. W.; Ijzerman, A. P. A two-entropies analysis to identify functional positions in the transmembrane region of class A G protein-coupled receptors. *Proteins* **2006**, *63*, 1018−30.

(10) Davies, M. N.; Gloriam, D. E.; Secker, A.; Freitas, A. A.; Timmis, J.; Flower, D. R. Present perspectives on the automated classification of the G-protein coupled receptors (GPCRs) at the protein sequence level. *Curr. Top. Med. Chem.* **2011**, *11*, 1994−2009.

(11) Mulder, N. J.; Apweiler, R.; Attwood, T. K.; Bairoch, A.; Bateman, A.; Binns, D.; Bork, P.; Buillard, V.; Cerutti, L.; Copley, R.; Courcelle, E.; Das, U.; Daugherty, L.; Dibley, M.; Finn, R.; Fleischmann, W.; Gough, J.; Haft, D.; Hulo, N.; Hunter, S.; Kahn, D.; Kanapin, A.; Kejariwal, A.; Labarga, A.; Langendijk-Genevaux, P. S.; Lonsdale, D.; Lopez, R.; Letunic, I.; Madera, M.; Maslen, J.; McAnulla, C.; McDowall, J.; Mistry, J.; Mitchell, A.; Nikolskaya, A. N.; Orchard, S.; Orengo, C.; Petryszak, R.; Selengut, J. D.; Sigrist, C. J.; Thomas, P. D.; Valentin, F.; Wilson, D.; Wu, C. H.; Yeats, C. New developments in the InterPro database. *Nucleic Acids Res.* **2007**, *35*, D224−8.

(12) Mouillac, B.; Chini, B.; Balestre, M. N.; Elands, J.; Trumpp-Kallmeyer, S.; Hoflack, J.; Hibert, M.; Jard, S.; Barberis, C. The binding site of neuropeptide vasopressin V1a receptor. Evidence for a major localization within transmembrane regions. *J. Biol. Chem.* **1995**, *270*, 25771−7.

(13) An, J.; Totrov, M.; Abagyan, R. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol. Cell Proteomics* **2005**, *4*, 752−61.

(14) Cavasotto, C. N.; Orry, A. J.; Abagyan, R. A. Structure-based identification of binding sites, native ligands and potential inhibitors for G-protein coupled receptors. *Proteins* **2003**, *51*, 423−33.

(15) Frimurer, T. M.; Ulven, T.; Elling, C. E.; Gerlach, L. O.; Kostenis, E.; Högberg, T. A physicogenetic method to assign ligand-binding relationships between 7TM receptors. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3707−12.

(16) Kratochwil, N. A.; Malherbe, P.; Lindemann, L.; Ebeling, M.; Hoener, M. C.; Mühlemann, A.; Porter, R. H.; Stahl, M.; Gerber, P. R. An automated system for the analysis of G protein-coupled receptor transmembrane binding pockets: alignment, receptor-based pharmacophores, and their application. *J. Chem. Inf. Model.* **2005**, *45*, 1324−36.

(17) Bywater, R. P. Location and nature of the residues important for ligand recognition in G-protein coupled receptors. *J. Mol. Recognit.* **2005**, *18*, 60−72.

(18) Wichard, J. D.; Ter Laak, A.; Krause, G.; Heinrich, N.; Kühne, R.; Kleinau, G. Chemogenomic analysis of G-protein coupled receptors and their ligands deciphers locks and keys governing diverse aspects of signalling. *PLoS One* **2011**, *6*, e16811.

(19) Reynolds, K. A.; Katritch, V.; Abagyan, R. Identifying conformational changes of the beta(2) adrenoceptor that enable accurate prediction of ligand/receptor interactions and screening for GPCR modulators. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 273−88.

(20) Kenakin, T. Ligand-selective receptor conformations revisited: the promise and the problem. *Trends. Pharmacol. Sci.* **2003**, *24*, 346−54.

(21) Kobilka, B. K.; Deupi, X. Conformational complexity of G-protein-coupled receptors. *Trends. Pharmacol. Sci.* **2007**, *28*, 397−406.

(22) Schwartz, T. W.; Frimurer, T. M.; Holst, B.; Rosenkilde, M. M.; Elling, C. E. Molecular mechanism of 7TM receptor activation—a global toggle switch model. *Annu. Rev. Pharmacol. Toxicol.* **2006**, *46*, 481−519.

(23) Urban, J. D.; Clarke, W. P.; von Zastrow, M.; Nichols, D. E.; Kobilka, B.; Weinstein, H.; Javitch, J. A.; Roth, B. L.; Christopoulos, A.; Sexton, P. M.; Miller, K. J.; Spedding, M.; Mailman, R. B. Functional selectivity and classical concepts of quantitative pharmacology. *J. Pharmacol. Exp. Ther.* **2007**, *320*, 1−13.

(24) Palczewski, K.; Kumasaka, T.; Hori, T.; Behnke, C. A.; Motoshima, H.; Fox, B. A.; Le Trong, I.; Teller, D. C.; Okada, T.; Stenkamp, R. E.; Yamamoto, M.; Miyano, M. Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* **2000**, *289*, 739−45.

(25) 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. High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science* **2007**, *318*, 1258−65.

(26) Jaakola, V. P.; Griffith, M. T.; Hanson, M. A.; Cherezov, V.; Chien, E. Y.; Lane, J. R.; Ijzerman, A. P.; Stevens, R. C. The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. *Science* **2008**, *322*, 1211−7.

(27) Hanson, M. A.; Roth, C. B.; Jo, E.; Griffith, M. T.; Scott, F. L.; Reinhart, G.; Desale, H.; Clemons, B.; Cahalan, S. M.; Schuerer, S. C.; Sanna, M. G.; Han, G. W.; Kuhn, P.; Rosen, H.; Stevens, R. C. Crystal structure of a lipid G protein-coupled receptor. *Science* **2012**, *335*, 851−5.

(28) Zhang, C.; Srinivasan, Y.; Arlow, D. H.; Fung, J. J.; Palmer, D.; Zheng, Y.; Green, H. F.; Pandey, A.; Dror, R. O.; Shaw, D. E.; Weis, W. I.; Coughlin, S. R.; Kobilka, B. K. High-resolution crystal structure of human protease-activated receptor 1. *Nature* **2012**, *492*, 387−92.

(29) Warne, T.; Serrano-Vega, M. J.; Baker, J. G.; Moukhametzianov, R.; Edwards, P. C.; Henderson, R.; Leslie, A. G.; Tate, C. G.; Schertler, G. F. Structure of a beta1-adrenergic G-protein-coupled receptor. *Nature* **2008**, *454*, 486−91.

(30) Haga, K.; Kruse, A. C.; Asada, H.; Yurugi-Kobayashi, T.; Shiroishi, M.; Zhang, C.; Weis, W. I.; Okada, T.; Kobilka, B. K.; Haga, T.; Kobayashi, T. Structure of the human M2 muscarinic acetylcholine receptor bound to an antagonist. *Nature* **2012**, *482*, 547−51.

(31) Kruse, A. C.; Hu, J.; Pan, A. C.; Arlow, D. H.; Rosenbaum, D. M.; Rosemond, E.; Green, H. F.; Liu, T.; Chae, P. S.; Dror, R. O.; Shaw, D. E.; Weis, W. I.; Wess, J.; Kobilka, B. K. Structure and dynamics of the M3 muscarinic acetylcholine receptor. *Nature* **2012**, *482*, 552−6.

(32) Chien, E. Y.; Liu, W.; Zhao, Q.; Katritch, V.; Han, G. W.; Hanson, M. A.; Shi, L.; Newman, A. H.; Javitch, J. A.; Cherezov, V.; Stevens, R. C. Structure of the human dopamine D3 receptor in complex with a D2/D3 selective antagonist. *Science* **2010**, *330*, 1091−5.

(33) Shimamura, T.; Shiroishi, M.; Weyand, S.; Tsujimoto, H.; Winter, G.; Katritch, V.; Abagyan, R.; Cherezov, V.; Liu, W.; Han, G. W.; Kobayashi, T.; Stevens, R. C.; Iwata, S. Structure of the human

histamine H1 receptor complex with doxepin. *Nature* **2011**, *475*, 65−70.

(34) Wu, H.; Wacker, D.; Mileni, M.; Katritch, V.; Han, G. W.; Vardy, E.; Liu, W.; Thompson, A. A.; Huang, X. P.; Carroll, F. I.; Mascarella, S. W.; Westkaemper, R. B.; Mosier, P. D.; Roth, B. L.; Cherezov, V.; Stevens, R. C. Structure of the human *κ*-opioid receptor in complex with JDTic. *Nature* **2012**, *485*, 327−32.

(35) Granier, S.; Manglik, A.; Kruse, A. C.; Kobilka, T. S.; Thian, F. S.; Weis, W. I.; Kobilka, B. K. Structure of the *δ*-opioid receptor bound to naltrindole. *Nature* **2012**, *485*, 400−4.

(36) Manglik, A.; Kruse, A. C.; Kobilka, T. S.; Thian, F. S.; Mathiesen, J. M.; Sunahara, R. K.; Pardo, L.; Weis, W. I.; Kobilka, B. K.; Granier, S. Crystal structure of the *μ*-opioid receptor bound to a morphinan antagonist. *Nature* **2012**, *485*, 321−6.

(37) Thompson, A. A.; Liu, W.; Chun, E.; Katritch, V.; Wu, H.; Vardy, E.; Huang, X. P.; Trapella, C.; Guerrini, R.; Calo, G.; Roth, B. L.; Cherezov, V.; Stevens, R. C. Structure of the nociceptin/orphanin FQ receptor in complex with a peptide mimetic. *Nature* **2012**, *485*, 395−9.

(38) White, J. F.; Noinaj, N.; Shibata, Y.; Love, J.; Kloss, B.; Xu, F.; Gvozdenovic-Jeremic, J.; Shah, P.; Shiloach, J.; Tate, C. G.; Grisshammer, R. Structure of the agonist-bound neurotensin receptor. *Nature* **2012**, *490*, 508−13.

(39) Gayen, A.; Goswami, S. K.; Mukhopadhyay, C. NMR evidence of GM1-induced conformational change of Substance P using isotropic bicelles. *Biochim. Biophys. Acta* **2011**, *1808*, 127−39.

(40) Park, S. H.; Das, B. B.; Casagrande, F.; Tian, Y.; Nothnagel, H. J.; Chu, M.; Kiefer, H.; Maier, K.; De Angelis, A. A.; Marassi, F. M.; Opella, S. J. Structure of the chemokine receptor CXCR1 in phospholipid bilayers. *Nature* **2012**, *491*, 779−83.

(41) Wu, B.; Chien, E. Y.; Mol, C. D.; Fenalti, G.; Liu, W.; Katritch, V.; Abagyan, R.; Brooun, A.; Wells, P.; Bi, F. C.; Hamel, D. J.; Kuhn, P.; Handel, T. M.; Cherezov, V.; Stevens, R. C. Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. *Science* **2010**, *330*, 1066−71.

(42) Phatak, S. S.; Gatica, E. A.; Cavasotto, C. N. Ligand-steered modeling and docking: A benchmarking study in class A G-protein-coupled receptors. *J. Chem. Inf. Model.* **2010**, *50*, 2119−28.

(43) Eckert, H.; Bajorath, J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discovery Today* **2007**, *12*, 225−33.

(44) Klabunde, T. Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. *Br. J. Pharmacol.* **2007**, *152*, 5−7.

(45) Weill, N.; Rognan, D. Development and validation of a novel protein-ligand fingerprint to mine chemogenomic space: application to G protein-coupled receptors and their ligands. *J. Chem. Inf. Model.* **2009**, *49*, 1049−62.

(46) Yabuuchi, H.; Niijima, S.; Takematsu, H.; Ida, T.; Hirokawa, T.; Hara, T.; Ogawa, T.; Minowa, Y.; Tsujimoto, G.; Okuno, Y. Analysis of multiple compound-protein interactions reveals novel bioactive molecules. *Mol. Syst. Biol.* **2011**, *7*, 472.

(47) Jacob, L.; Vert, J. P. Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics* **2008**, *24*, 2149−56.

(48) van der Horst, E.; Okuno, Y.; Bender, A.; IJzerman, A. P. Substructure mining of GPCR ligands reveals activity-class specific functional groups in an unbiased manner. *J. Chem. Inf. Model.* **2009**, *49*, 348−60.

(49) Wassermann, A. M.; Geppert, H.; Bajorath, J. Ligand prediction for orphan targets using support vector machines and various target-ligand kernels is dominated by nearest neighbor effects. *J. Chem. Inf. Model.* **2009**, *49*, 2155−2167.

(50) Schiöth, H. B.; Fredriksson, R. The GRAFS classification system of G-protein coupled receptors in comparative perspective. *Gen. Comp. Endrocrinol.* **2005**, *142*, 94−101.

(51) *GVK Bio Target inhibitor databases*; GVK Biosciences Private Limited: Hyderabad, India, 2007.

(52) Shawe-Taylor, J.; Cristianini, N. *Kernel Methods for Pattern Analysis*; Cambridge University Press: Cambridge, 2004.

(53) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−54.

(54) *DRAGON*, version 5.5; Talete srl: Milano, Italy, 2011.

(55) *Pipeline Pilot*, version 6.5.1; Accelrys, Inc.: San Diego, CA, 2007.

(56) El-Manzalawy, Y.; Dobbs, D.; Honavar, V. Predicting linear B-cell epitopes using string kernels. *J. Mol. Recognit.* **2008**, *21*, 243−55.

(57) Bissantz, C.; Logean, A.; Rognan, D. High-throughput modeling of human G-protein coupled receptors: amino acid sequence alignment, three-dimensional model building, and receptor library screening. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1162−76.

(58) Saigo, H.; Vert, J. P.; Akutsu, T. Optimizing amino acid substitution matrices with a local alignment kernel. *BMC Bioinf.* **2006**, *7*, 246.

(59) Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J. Med. Chem.* **1998**, *41*, 2481−91.

(60) Rao, H. B.; Zhu, F.; Yang, G. B.; Li, Z. R.; Chen, Y. Z. Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res.* **2011**, *39*, W385−90.

(61) Vapnik, V. N. *Statistical Learning Theory*; John Wiley & Sons, Inc.: New York, 1998.

(62) Chang, C.-C.; Lin, C.-J. LIBSVM: a library for support vector machines. *ACM Trans. Intell. Systems Technol.* **2011**, *2*, 27.

(63) Shackelford, G.; Karplus, K. Contact prediction using mutual information and neural nets. *Proteins* **2007**, *69* (Suppl 8), 159−64.

(64) Gatica, E. A.; Cavasotto, C. N. Ligand and decoy sets for docking to G protein-coupled receptors. *J. Chem. Inf. Model.* **2012**, *52*, 1−6.

(65) *GPCR SARfari*, version 2.0; European Bioinformatics Institute: Hinxton, U.K., 2011.

(66) Weaver, S.; Gleeson, M. P. The importance of the domain of applicability in QSAR modeling. *J. Mol. Graphics Modell.* **2008**, *26*, 1315−26.

(67) Leslie, C. S.; Eskin, E.; Cohen, A.; Weston, J.; Noble, W. S. Mismatch string kernels for discriminative protein classification. *Bioinformatics* **2004**, *20*, 467−76.

(68) Hanley, J. A.; McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29−36.

(69) Christopoulos, A. Allosteric binding sites on cell-surface receptors: novel targets for drug discovery. *Nat. Rev. Drug Discovery* **2002**, *1*, 198−210.

(70) van Koppen, C. J.; Zaman, G. J.; Timmers, C. M.; Kelder, J.; Mosselman, S.; van de Lagemaat, R.; Smit, M. J.; Hanssen, R. G. A signaling-selective, nanomolar potent allosteric low molecular weight agonist for the human luteinizing hormone receptor. *Naunyn Schmiedebergs Arch. Pharmacol.* **2008**, *378*, 503−14.

(71) Keiser, M. J.; Setola, V.; Irwin, J. J.; Laggner, C.; Abbas, A. I.; Hufeisen, S. J.; Jensen, N. H.; Kuijer, M. B.; Matos, R. C.; Tran, T. B.; Whaley, R.; Glennon, R. A.; Hert, J.; Thomas, K. L.; Edwards, D. D.; Shoichet, B. K.; Roth, B. L. Predicting new molecular targets for known drugs. *Nature* **2009**, *462*, 175−81.

(72) Suryanarayana, S.; Kobilka, B. K. Amino acid substitutions at position 312 in the seventh hydrophobic segment of the beta 2-adrenergic receptor modify ligand-binding specificity. *Mol. Pharmacol.* **1993**, *44*, 111−4.

(73) Katritch, V.; Reynolds, K. A.; Cherezov, V.; Hanson, M. A.; Roth, C. B.; Yeager, M.; Abagyan, R. Analysis of full and partial agonists binding to beta2-adrenergic receptor suggests a role of transmembrane helix V in agonist-specific conformational changes. *J. Mol. Recognit.* **2009**, *22*, 307−18.

(74) Ivanov, A. A.; Costanzi, S.; Jacobson, K. A. Defining the nucleotide binding sites of P2Y receptors using rhodopsin-based homology modeling. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 417−26.