

Identification of G-protein coupled receptors and ligands interactions on a chemo-genomic scale

Wang Ting* and Duan Yong

(Genome Center and Bioinformatics Program and Department of Applied Science, 451 East Health Science Drive, University of California, Davis, CA 95616-8816, USA)

Abstract: G-protein coupled receptors (GPCR) represent a class of important therapeutic targets. Seeking novel ligands as potential drugs targeting GPCRs and identifying natural ligands for orphan GPCRs have been long-standing efforts of academic and pharmaceutical industrial research. To accelerate this effort, there is a critical need for methods capable of predicting GPCR-ligand interactions on a large scale. Such methods also may help to reveal cross-pharmacology of different GPCRs in order to alleviate side effects and toxicity of potential drugs. Here we report a support vector machine (SVM)-based method for predicting GPCR-ligand interactions on a chemo-genomic scale. In this method, GPCRs were characterized by the sequence information of the transmembrane segments and ligands were represented by their chemical structural information. The application of the method to a set of known GPCR-ligand interacting pairs that included GPCRs from 28 subfamilies of the A family led to a model of GPCR-ligand interaction network. The model was able to distinguish interacting pairs from non-interacting pairs with an average 86.9% true-positive rate and 99.97% true-negative rate. Moreover, the model correctly predicted the interactions of a number of new ligands and orphan GPCRs that were chemically and phylogenetically novel to the training data set. This method is expected to be applicable to *in silico* high-throughput GPCR-targeting drug discovery and ligand identification at the GPCRs with unknown functions.

Keywords: G-protein coupled receptors, ligands, network, machine learning, drug discovery

1 Introduction

The G-protein coupled receptors (GPCRs) participate in numerous cellular processes by transducing extracellular signals into cytoplasm through interaction with guanine nucleotide-binding (G) proteins. The extracellular signals include sensory signals of external origin such as odor and taste and endogenous signals such as biogenic amines, peptides and lipids. The GPCRs that require specific endogenous ligands to respond, referred to as endogenous GPCRs, are of particular importance as they are involved in the regulation of a wide range of physiological processes including metabolism, development and aging and can be effective drug targets against the associated diseases. In fact, ca. 50 % drugs on the market are the modulators of endogenous GPCRs and these drugs act as either agonists or antagonists. Recent genome sequencing projects revealed that approximately 370 sequences belong to the endogenous GPCRs class in the human genome^[1-4], of which ca. 270 have been paired with known ligands and 100 still remain as orphan GPCRs. Among the GPCRs with known ligands, however, only a small fraction (ca. 30 GPCRs) has been used as drug targets, leaving the majority of the GPCR family as the potential drug targets that have yet to be explored. Therefore, seeking for novel ligands as potential drugs targeting the known GPCRs and identifying natural ligands of orphan GPCRs have been long-standing efforts of academic and pharmaceutical industrial research. Success of the effort is not only crucial to enrich our understanding of the physiological and pathological significance of the GPCR family but also important in developing new therapeutic agents^[5]. To accelerate this process, there is a critical need for methods capable of predicting GPCR-ligand interactions on a large scale. Also importantly, such methods may help to reveal cross-pharmacology of different GPCRs in order to alleviate side effects and toxicity of potential drugs, which is an important step towards systems chemical biology^[6] involving GPCRs.

The GPCR proteins are believed to adopt a common structural framework comprising seven transmembrane (TM) helices, as shown in the recently determined crystal structures

of a few GPCRs, including the bovine and squid rhodopsins^[7-9], the turkey β_1 , human β_2 -adrenergic receptors^[10-12], and the human A_{2A} adenosine receptor^[13]. These are also the only available crystal structures of the GPCR super family. While these crystal structures confirmed the high structural similarity among GPCRs, they also show notable differences relevant to ligand-binding. For example, the ligand-binding pocket in the rhodopsin structures are deeply buried and completely isolated from the protein surface, which raised the question how a ligand goes in and out of the binding pocket^[14] whereas there is an obvious ligand entrance site in the β_2 -adrenergic receptor structure. Compared to the conservation of the structures, the amino acid sequences of GPCRs show considerable diversity with sequence identity typically lower than 30% between sequences of different subfamilies. The application of homology modeling is therefore limited to the GPCRs with sequences similar to the known crystal structures and it is difficult to use this technique on a large scale.

Recently, the methods that employed the amino acid sequence information of receptors in combination with the chemical information of the ligands have been developed to predict receptor-ligand interactions, such as Wikberg's proteochemometrics modeling techniques^[15-18] and Bock and Gough's approach^[19]. Our method, which will be described in this paper, follows the concept of proteochemometric modeling, but uses a novel strategy to represent protein sequence space.

Specifically, we take advantage of the unique feature of the transmembrane domains of GPCRs and developed a method capable of building a single model of GPCR-ligand interaction network for the GPCRs that interact with ligands at the transmembrane domains. These GPCRs include the entire A family that constitute over 80% of known GPCR genes and subsets of other families. In the method, GPCRs were characterized by the sequence information of the transmembrane segments and each of the sequence of varied length was converted into a numerical array of fixed length. The ligands, which interact with a single or multiple GPCRs were represented by their chemical structural information. A support vector machine (SVM) technique was used to derive the binary interaction relationships between the GPCRs and the ligands. The application of the method to a set of known GPCR-ligand interacting pairs, which included GPCRs from 28 subfamilies of the A family, led to a highly predictive model, able to distinguish

Received: 2009-3-5; Revised: 2009-4-18

*To whom correspondence should be addressed.
E-mail: twang@ucdavis.edu

interacting pairs from non-interacting pairs with an average 86.9% true-positive rate and 99.97% true-negative rate. Moreover, the model correctly predicted the interactions of a number of new ligands and orphan GPCRs that were chemically and phylogenetically novel to the training data set. With the model, it is straightforward to perform simultaneous multiple-receptor screening for small molecules, which would reveal the cross-pharmacology among different GPCRs. The method is expected to be applicable to *in silico* high-throughput GPCR-targeting drug discovery and ligand identification at the GPCRs with unknown functions.

2 Method and materials

We used the support vector machine (SVM) classification technique^[20] to derive binary interaction relationships between

GPCRs and ligands based on the amino acid sequence information of the GPCRs and the chemical structure information of the ligands. The SVM technique requires two data sets as input: interacting GPCR-ligand pairs and non-interacting pairs.

2.1 Data sets of GPCR-ligand pairs

2.1.1 Interacting pairs

Our data set of GPCR-ligand interacting pairs contain 1307 known GPCR-ligand pairs: 817 from the GLIDA database^[21] (release of March 24, 2006) and 490 from the International Union of Pharmacology (IUPHAR) GPCR database (<http://www.iuphar-db.org/GPCR/>). This dataset contained 106 human GPCR genes that spanned 28 subfamilies of the A family, and the corresponding mouse and rat genes were also included if they are not identical to the human ones, resulting 214 unique genes as listed in Table 1.

Table 1 GPCR genes in the training set, including 106 human genes and 108 non-redundant corresponding mouse and rat genes belonging to 28 subfamilies of the A family.

amine			peptide			lipid			nucleotide-like		
family	gene	species	family	gene	species	family	gene	species	family	gene	species
acetylcholine	ACM1	H,M (10) ^c	angiotensin	AG2R	H (8)	bile acid	GPBAR	H (4)	adenosine	AA1R	H,M,R(11)
	ACM2	H,R (9)		AG22	H (4)		cannabinoid	CNR1		H,M (13)	AA2AR
	ACM3	H,R (10)	bombesin	GRPR	H,M,R (4)		CNR2	H,M (10)		AA2BR	H,M,R(10)
	ACM4	H,M,R (10)		NMBR	H,M,R (4)	cys.leukotriene ^a	CLTR1	H (2)		AA3R	H,M,R(6)
	ACM5	H,R (9)	bradykinin	BKRB1	H,M,R (8)		leukotriene	LT4R1		H (6)	P2Y
ADA1A	H,R (26)	BKRB2		H,M,R (11)		LT4R2	H (6)	P2RY2		H,M,R(5)	
adrenoceptor	ADA1B	H (23)	chemokine	CCR1	H,M (2)	free fatty acid	GPR40	H (1)		P2RY4	H,M,R(7)
	ADA1D	H,M (22)		cholecystokinin	CCKAR		H,R (3)			GPR41	H (1)
	ADA2A	H,M,R (16)		GASR	H,M,R (5)		GPR43	H (1)		P2Y11	H(11)
	ADA2B	H,M,R (16)	endothelin	EDNRA	H,M (3)	PAF ^b	PTAFR	H,M (8)		P2Y12	H(6)
	ADA2C	H,M (16)		EDNRB	H,M,R (2)		prostanoid	PD2R		H,M,R (20)	P2T13
	ADRB1	H,M,R (22)	melanocortin	MSHR	H,M (2)			PE2R1		H,M (19)	P2Y14
	ADRB2	H,M,R (23)		ACTHR	H (2)		PF2R2	H,M (15)			
	ADRB3	H,M,R (15)	MC3R	H,M,R (4)		PE2R3	H,M (18)				
	dopamine	DRD1	H (13)	MC4R	H (5)		PE2R4	H,M (14)	others		
		DRD2	H,R (20)	MC5R	H,M (3)		PF2R	H,M,R (14)	family	gene	species
DRD3		H,M,R (10)	neuropeptide Y	NPY1R	H,M (1)		PI2R	H (15)	melatonin	MTR1A	H,M(4)
DRD4		H,M,R (16)		NPY2R	H,M (2)		TA2R	H (20)		MTR1B	H(9)
histamine	HRH1	H,M,R (12)	NPY5R	H,M (1)	sphingolipid	EDG1	H (1)				
	HRH2	H,M,R (10)	neurotensin	NTR1		H,M,R (5)		EDG3	H(1)		
	HRH3	H (19)		NTR2	H,M,R (3)		EDG4	H(1)			
	HRH4	H,M,R (16)	opioid	OPRD	H,M (15)		EDG5	H(1)			
serotonin	5HT1A	H,M,R (10)		OPRK	H (8)		EDG6	H(1)			
	5HT1B	H,M (8)	OPRM	H,M (11)		EDG7	H(1)				
	5HT1D	H,M (5)	somatostatin	SSR1	H (4)		EDG8	H(1)			
	5HT1E	H (3)		SSR2	H (9)						
	5HT1F	H,R (2)	SSR3	H,M,R (5)							
	5HT2A	H,M (4)	SSR4	H,R (2)							
	5HT2B	H,R (4)	SSR5	H,M (3)							
	5HT2C	H,M (5)	urotensin II	UR2R	H,M,R (2)						
	5HT4R	H,M,R (3)									
	5HT5A	H (3)									
	5HT6R	H,R (1)									
	5HT7R	H(3)									

^a cysteinyl leukotriene.

^b platelet-activating receptor

^c number of distinct ligands.

The GPCRs in Table 1 can be clustered to five groups according to traditional classification of the A family on the basis of ligand similarity:

1) The amine group, including the acetylcholine, the adrenoceptor, the dopamine, the histamine and the serotonin subfamilies;

2) The peptide group, including the angiotensin, the bombesin, the bradykinin, the chemokine, the cholecystokinin, the endothelin, the melanocortin, the neuropeptide Y, the neurotensin, the opioid, the somatostatin and the urotensin II subfamilies;

3) The lipid group, including the bile acid, the cannabinoid, the leukotriene, the cysteinyl leukotriene, the prostanoid, the sphingolipid and the free fatty acid subfamilies;

4) The nucleotide-like group, including the adenosine and the P2Y subfamilies;

5) The melatonin group, including the melatonin subfamily.

There are 494 distinct ligands paired with the GPCRs, including both agonists and antagonists that are natural or synthetic small molecules, lipids and small peptides or peptide

mimics, of which, 209 were paired with the amine group, 105 with the peptide group, 110 with the lipid group, 60 with the nucleotide group and 10 with the melatonin group. No ligand was paired with GPCRs across the groups. The number of distinct ligands that interact with each of the GPCRs is also listed in Table 1. The sum of these numbers exceeds the total number of the distinct ligands (494) because some of the ligands interact with more than one GPCR, especially those belonging to a same subfamily. In addition, we can see that about half of the ligands are paired with the amine group GPCRs. This is not surprising as the amine group is the most studied GPCRs, but this bias may lead to that models derived based on this dataset perform better for the amine group than the other groups.

The original source of each pair included the amino acid sequence of the GPCR protein and the 2-dimensional structure of the corresponding ligand.

2.1.2 Non-interacting pairs

Non-interacting pairs were not readily available in public databases but are necessary for building a SVM model. A commonly used practice was to make an assumption that if a

compound is not listed as active for a target, the compound is not active for that target although it was known that this assumption could introduce false negatives^[22,23]. However, in the case of GPCRs, this assumption can be highly problematic because it is well known that a single drug may interact with multiple GPCRs. This cross-pharmacology is prevalent within a subfamily and has also been observed cross subfamilies, particularly in the amine group. Other known cases are that melatonin agonists were often found to be selective serotonin antagonists^[24] and the cysteinyl leukotriene receptor 1 was found to be a dual receptor that can be activated by leukotrienergic ligands and also a nucleotide, the ligand of P2Y receptors^[25,26].

Our strategy of constructing non-interacting pairs was to cross-combine the ligands and GPCRs in the interacting data set with three exceptions:

- 1) no cross-combination within a subfamily;
- 2) no cross-combination from different subfamilies in the amine group and the largest pair-wise phylogenetic distance in the amine group was used to define a distance cutoff;
- 3) no cross-combination if the phylogenetic distance of two sequences is smaller than the defined distance cutoff. To compute the phylogenetic distances, we first performed sequence alignment for all the human GPCRs studied in this work by using the ClustalW program^[27].

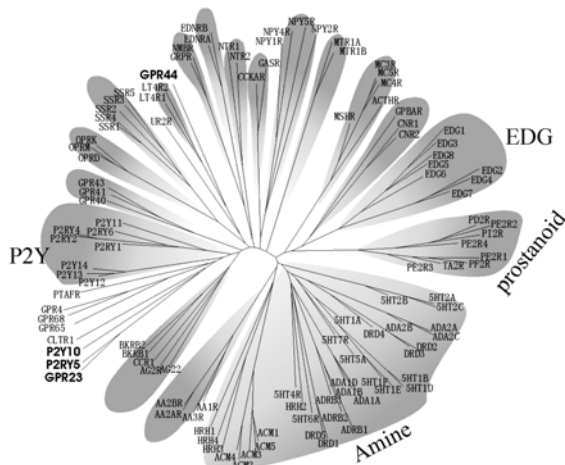


Fig.1 Phylogenetic tree of the GPCRs studied in this work. Sequences were aligned using the ClustalW program^[27] and the tree was built using the PHYLIP program^[28]. Proteins in a subfamily were placed in a shadow and all the subfamilies belonging to the amine group could be placed in a single shadow.

Then the PROTDIST program in the PHYLIP package^[28] was used to compute the distance matrix of the sequences. The distance matrix computed is provided in the Supporting Information. Here we present the phylogenetic tree of the sequences in Fig.1. The largest pair-wise distance is 2.57482 (between ACM3 and DRD4) in the amine group, 2.35351 (between MTR1A and 5HT7R) between the melatonin subfamily and the serotonin family, and 2.66446 between the cysteinyl leukotriene receptor 1 (CLTR1) and the P2Y receptors (P2Y11). We then defined 2.6 as the distance cutoff for the cross-combination between different subfamilies and excluded the combination between CLTR1 and P2Y11. As a result, 25230 non-interacting pairs were generated. It is worth noting that while the risk of introducing false negatives could be largely reduced, this strategy may result in the loss of some information on receptor selectivity, particularly for the receptors in a subfamily and in the amine group as no non-interacting pairs were made crossing these receptors.

2.2 Description of GPCR sequence and ligand structure

2.2.1 Protein sequences

For the family A GPCRs, it is believed that the ligand-binding pocket is located in the transmembrane region that is enclosed by seven transmembrane helices and connecting loops, as shown in the few solved crystal structures^[8,9,11,12]. We first detected the transmembrane region of each GPCR sequence by using the GPCRHMM web server (<http://gpcrhmm.cgb.ki.se/>)^[29]. This reduced the sequence length to between 246 and 475, and more importantly, the protein sequences were thus aligned to take similar spatial orientations. We then converted each of these sequences of varied lengths to a numerical array of fixed length by counting the types and frequencies of tripeptide composition in the transmembrane segment, starting from the second residue preceding the N-terminal of the first transmembrane helix. Tripeptide was used because it is the shortest peptide to reflect the local environment of residues in a sequence.

Since there are 20 amino acids, there could be $20 \times 20 \times 20 = 8000$ types of tripeptide composition, which is a huge number compared with only 214 protein sequences studied here (Table 1). In general, to obtain a statistically significant model, the number of variables should be smaller than one-third of the number of samples. Therefore, the 20 amino acids have to be grouped. It is well known that the transmembrane segments of GPCR sequences are dominated by neutral residues but ligand binding affinities are significantly determined by the relatively much fewer charged residues. Thus, we classified the 20 amino acids into three types on the basis of their charge properties:

- 1) neutral group, including A, C, F, G, H, I, L, M, N, P, Q, S, T, V, W, and Y;
- 2) positively charged group, including K and R;
- 3) negatively charged group, including D and E. We also explored other classifications and the results are discussed later.

This classification resulted in $3 \times 3 \times 3 = 27$ different types of tripeptide composition. As a result, each of the protein sequence was transformed to a 27-dimensional numerical vector. A schematic illustration on how to convert a sequence into such a vector is shown in Fig.2. The vectors of all 221 GPCR sequences (214 in the training set plus 7 in external tests) studied in this work are listed in a Microsoft Excel table in the Supporting information.

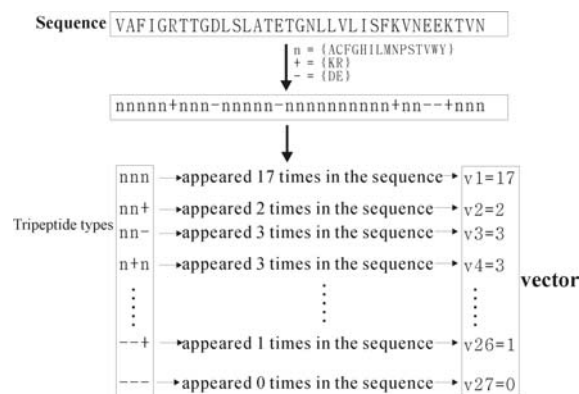


Fig.2 Schematic illustration on how to convert a sequence to a 27-dimensional vector that counts for the frequencies and types of tripeptide composition, where the 20 amino acids are classified to 3 types based on their charge properties.

To investigate the distribution of the proteins in the numerical space defined by the 27 variables, we performed a principal component analysis (PCA) for the 106 human GPCRs in Table 1. The score plot of the first four PCs (PC1 to PC4) is shown in Fig.3. We can see that PC1 distinguishes the amine group (black symbols) from the others (symbols in other

colors), with a few exceptions that are HRH2, 5HT4R, 5HT6R, ADRB3 in the amine group, CCKAR, GASR EDNRA, EDNRB and NPY5R in the peptide group, and PD2R in the lipid group. But no a single PC could separate any of the other groups from the rest. This is consistent with the phylogenetic profile of the proteins as shown in Fig.1. But the clustering of the subfamilies is less obvious in this numerical descriptor space.

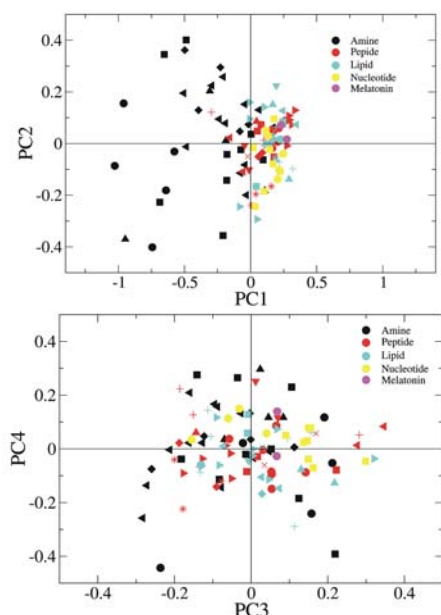


Fig.3 Score plots of the first four principal components (PC1 to PC4) for the 106 human GPCRs in Table 1 that were represented by 27-dimensional vectors. upper) PC1 and PC2; lower) PC3 and PC4. The proteins are represented by symbols with different styles representing different subfamilies. Black is for all the subfamilies belonging to the amine group, red is for the peptide group, cyan is for the lipid group, yellow is for the nucleotide group and purple is for the melatonin group.

2.2.2 Ligand structures

For the ligands, we used 57 classical quantitative structure-activity relationship (QSAR) descriptors. The descriptors mainly account for the molecular physical properties, molecular surface, volume and charge properties that were calculated based on the 2D and 3D structures of the ligands in the MOE software (Chemical Computing Group, Inc. See the Supporting information in Table S1 for the list of the descriptors). Prior to the calculations of these descriptors, the 2D structures of the ligands were converted to 3D and energy minimized by using the default parameters in the MOE software. It is important to properly treat the protonation states of the ligands as this can affect the net charge and charge distribution of a ligand and charge-charge interactions have been found important in GPCR-ligand binding. The protonation state is mainly dependent on the local chemical environment and the pH value. In this work, as the structure of the receptor is not considered, we determined the protonation states of the ligands based on their own chemical structures at physiological pH. As a result, the protonation states were assigned by setting amino and guanidino groups to be protonated and carboxylic acid and phosphate acid groups deprotonated.

Thus, a GPCR-ligand pair was characterized by concatenating the vectors of the protein and the ligand and represented by an 84-dimensional vector (27 for the protein sequence plus 57 for the ligand). In addition, a label of +1 or -1 was added to indicate the interacting and non-interacting pairs.

2.3 Support vector machine

The support vector machine (SVM) is a binary classification technique that attempts to define a hyperplane to separate two

classes of data: interacting pairs and non-interacting pairs. The hyperplane is generated by computing supporting vectors based on the descriptors of the data. The accuracy of a model generated by SVM relies on the robustness of the algorithm and the relevance of the descriptors to the problem. Since first introduced in 1995^[20], the SVM method has increasingly been used in the chemoinformatics^[22,23,30,31] and bioinformatics^[32-34] to solve classification problems and achieved high precision in many applications.

In this work, we used the SVM algorithm implemented in the libsvm 2.84 package (<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>) and the radial basis function (RBF) kernel for feature space transformation. The RBF kernel is also called Gaussian function kernel that non-linearly transforms the original data into a higher dimensional space. The RBF kernel is set to be the default kernel in the libsvm 2.84 package among three other kernels: linear, polynomial and sigmoid functions because it can better handle non-linear problems with less numerical difficulty. The RBF kernel has two user-adjustable parameters: the error-tolerance parameter C and the kernel width parameter γ , as well as a weighting parameter for unbalanced data in which the numbers of the data of the two classes are not equal. The value of the C parameter is usually between 1 and 1000. Although there is no established guidance as to what is an appreciate value for a classification problem, in general, a larger C would lead to better fitting but may introduce a problem of over fitting. The γ parameter is usually a positive number smaller than 1.0 and we used the default value of 0.5. The weighting parameter (-wi option) can be set to be a number larger than 1 to penalize the possible misclassification of the class with much fewer number of data, i.e. the interacting pairs in this study. This will lead to more pairs to be classified as interacting. But we decided not to use any penalty for two reasons:

- 1) we were concerned about both true positive rates and true negative rates.
- 2) a positive decision made by a model generated without such a penalty for a data set with significantly more negative data may exhibit higher confidence, compared to using the penalty.

In other words, an interacting pair predicted by the model has a higher possibility to be a true positive, compared to a prediction using a penalty. This confidence is particularly important in drug discovery and also in the ligand identification of orphan GPCRs. In addition, before model generation, the data set was subject to scaling, which is to normalize the values of the descriptors to be between -1.0 and +1.0.

In addition to the libsvm 2.84 package for the SVM classification task, a variety of C language programs and Linux C-shell scripts were used to prepare the data and validate the models.

3 Results

3.1 Model generation and validation

The data set consisting of 1307 interacting pairs and 25230 non-interacting pairs was trained to generate SVM models with default parameters in the libsvm 2.84 package, except the error tolerance parameter C , which was tested between 10.0 and 20.0 with an increment of 2.0 by using a 10-fold cross-validation test (-v option in the libsvm 2.84 package). The test results showed that the accuracies (fraction of pairs correctly predicted) of models improved slightly from 99.12% to 99.23%. We then chose 20.0 for C to build the final model.

The model correctly recognized 1251 of the 1307 interacting pairs and all non-interacting pairs, yielding a sensitivity of 95.7% (fraction of interacting pairs predicted as such), a specificity of

100% (fraction of non-interacting pairs predicted as such) and an accuracy of 99.79% (fraction of pairs correctly predicted). In the 56 interacting pairs that were not recognized by the model, two were paired with 5HT6R of the amine group and the rest 54 were paired with 27 GPCRs in the other four groups (the peptide, the lipid, the nucleotide-like, and the melatonin groups). Besides the 10-fold cross-validation provided in the libsvm 2.84 package, we performed another 10 tests. In each test, 100 interacting pairs (ca. one-tenth of the total interacting pairs) and 2000 non-interacting pairs (ca. one-tenth of the total non-interacting pairs) were randomly selected from the training set and predicted by the models generated by the remaining data. The results are shown in Fig.4. In the ten cross-validations, the average sensitivity was 86.9%, the specificity was 99.97% indicating that the high quality of the model was not due to over fitting and the model thus has strong predictive power.

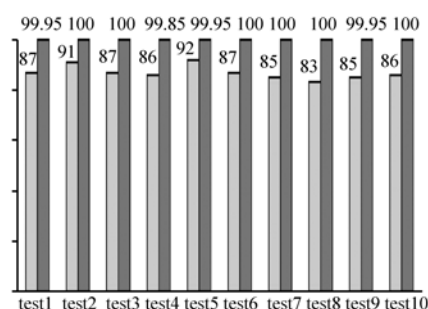


Fig.4 Performance of 10 cross-validation tests of the SVM model. Sensitivity (%; light grey column) and specificity (%; dark grey column) are listed for each of the 10 tests.

As we can see in Fig.4, the validation models showed a common feature that specificities or true negative rates were always much higher than sensitivities or true positive rates. The main reason could be the over-abundance of the non-interacting pairs compared to the interacting pairs: a ratio of ca. 20:1 in the datasets. This led to that the algorithm knew more about what promoted a non-interacting pair and therefore identified the non-interacting pairs better. And on the other hand, this could give a high confidence to a predicted interacting pair, which would be a desired feature for drug discovery.

It is worth to mention that although high quality models were generated by SVM, it is difficult to interpret the models, especially to figure out the most important descriptors. There are advances^[35,36] towards feature selection and ranking, but these methods are still not commonly integrated into available software. So, it is not clear here that which of the 27 receptor descriptors and the 57 ligand descriptors contributed to the models most.

3.1.1 Prediction of novel GPCR-ligand pairs

The predictive ability of the model was evaluated by external tests in which either the ligands were chemically and structurally distinct from those in the training set or the GPCR sequences were novel.

3.1.2 Novel ligands

The first test was for two novel compounds that were not present in the training set, lysergic acid diethylamide (LSD) and salvinorin A (SA) (see Table 2 for their chemical structures). Both compounds are known hallucinogens. LSD is a synthetic compound that mainly acts as an antagonist of serotonin receptors^[37]. SA is the main active ingredient of a psychoactive plant called *Salvia divinorum* and was recently identified as the first naturally occurring non-nitrogenous κ -opioid selective agonist^[38]. LSD was in the GLID database but not collected in our data set and two structurally related compounds lisuride and pergolide were also excluded. Fingerprint-based similarity

searches for LSD and SA were carried out with the bit-packed MACCS structure keys^[39] that characterize molecules by using a set of small substructures in the MOE software. With the Tanimoto coefficient cutoff of 0.65, no hit was found for LSD and one hit, ginkgolide A was found for SA, which is an antagonist of the platelet-activating receptor (PTAFR). With a lower Tanimoto coefficient cutoff of 0.60, 12 hits were found for LSD, which are the ligands of ACM1, ACM4, DRD1, DRD2, DRD4, MTR1A, AA1R, AA2BR, EDNR1, and OPRD, respectively. 3 hits, ginkgolide A, B and J were found for SA, which are antagonists of the platelet-activating receptor (PTAFR).

Table 2 Prediction of interactions between compounds lysergic acid diethylamide (LSD) and salvinorin A (SA) and 30 GPCR proteins, which experimental results were reported in reference^[38].

GPCR	ligand			
	LSD		SA	
	Exp. ^a	Pred. ^b	Exp.	Pred.
5HT1A_HUMAN	√ ^c	√	x ^d	x
5HT1B_HUMAN	√	√	x	x
5HT1D_HUMAN	√	√	x	x
5HT1E_HUMAN	√	√	x	x
5HT2A_HUMAN	√	x	x	x
5HT2B_HUMAN	√	√	x	x
5HT2C_HUMAN	√	√	x	x
5HT5A_HUMAN	√	√	x	x
5HT7R_HUMAN	√	√	x	x
DRD1_HUMAN	√	√	x	x
DRD2_HUMAN	√	√	x	x
DRD3_HUMAN	√	√	x	x
DRD4_HUMAN	√	√	x	√
DRD5_HUMAN	√	√	x	√
OPRK_HUMAN	x	x	√	x
OPRD_HUMAN	x	x	x	x
OPRM_HUMAN	x	x	x	x
ACM1_HUMAN	x	√	x	x
ACM2_HUMAN	x	x	x	x
ACM3_HUMAN	x	√	x	x
ACM4_HUMAN	x	√	x	x
ACM5_HUMAN	x	x	x	x
ADRB1_RAT	√	√	x	x
ADRB2_RAT	√	√	x	x
ADA1A_HUMAN	√	√	x	x
ADA1B_HUMAN	x	√	x	x
HRH1_RAT	√	√	x	x
ADA2A_HUMAN	√	√	x	x
ADA2B_HUMAN	√	√	x	x
ADA2C_HUMAN	√	√	x	x

^a experimental results from reference^[38]. ^b predicted results. ^c interacting.

^d non-interacting.

We made predictions for the interactions between these two compounds (LSD and SA) and all 106 human GPCRs in the training data set. This can be regarded as simultaneous multiple-receptor screening. The evaluation of the prediction results was based on a single literature resource, an experimental report by Roth and coworkers^[38]. In that work, Roth and coworkers measured the inhibitory activities of LSD and SA at 10 μ M concentrations against a large number of GPCRs, of which 30 belonged to the A family and were in the serotonin, the dopamine, the opioid, the acetylcholine, the adrenoceptor and the histamine subfamilies. We defined that >50% inhibition indicates interactions between the protein and ligand and converted the quantitative experimental results into

binary results, accordingly. The predictions from our model as well as the experimental results are listed in Table 2. We can see that for LSD, the model correctly predicted 25 of the 30 GPCR-ligand pairs measured in the experiments^[38], leaving the rest 4 as false positives and 1 as false negative. Importantly, the model correctly predicted the major biological activity of LSD, the interactions with serotonin receptors. In addition, the model made negative decisions for 5HT6R and HRH2 of the amine group and all the proteins in the peptide group, the lipid group, the nucleotide-like group and the melatonin receptors. For SA, the model made only three positive decisions, which were for two dopamine receptors DRD4 and DRD5 and a serotonin receptor 5HT4R. Although these are false positives and the only true positive was missed, the high selectivity of SA for receptors was obvious in the prediction and more importantly, the model correctly predicted the inactivity of SA at serotonin receptors, which is an important feature distinguishing SA from other classical hallucinogens.

3.2 Orphan GPCRs

The second external test was for four human GPCRs that were not present in the training set, GPR44, GPR23, P2RY5 and P2Y10. These four GPCRs are recently de-orphanized lipid mediator GPCRs, in which GPR44 is a prostanoid receptor^[40,41], GPR23 and P2RY5 are lysophosphatidic acid (LPA) receptors^[42,43] and P2Y10 was reported to be a dual receptor of lysophosphatidic acid (LPA) and sphingosine 1-phosphate (S1P)^[44,45]. LPA is the ligand of EDG2, EDG4 and EDG7 of the EDG subfamily and S1P is the ligand of the rest member of the EDG subfamily. The phylogenetic relationships between these four novel GPCRs with other training GPCRs can be seen in Fig.1. None of them can be easily assigned to the identified ligands as they are phylogenetically located far from the GPCRs known to respond to such ligands, i.e. the prostanoid subfamily and the EDG subfamily, respectively. The sequence identity between GPR44 and the members in the prostanoid subfamily is lower than 20%. Instead, GPR44 is more closely related to leukotriene subfamily (LT4R1 and LT4T2) with sequence identity of 32% and the urotensin II receptor (UR2R) with sequence identity of 27%. GPR23, P2RY5 and P2Y10 fall into the cluster of the P2Y receptors, the platelet-activating receptor (PTAFR) and the cysteinyl leukotriene receptor 1 (CLTR1). Based on the assumption that neighboring GPCRs in the phylogenetic tree may share the same ligands, we constructed the putative interacting pairs by combining the four GPCRs with the representative ligands of their close neighbors as well as with their identified ligands. The results of prediction are shown in Fig.5.

For GPR44, the model predicted that it had interactions with five prostanoid compounds and had no interaction with any of the six ligands of the leukotriene receptors or the agonist of the urotensin II receptor. As the five prostanoid compounds were experimentally verified agonists of GPR44^[40-42], this prediction correctly identified the cognate ligands of GPR44 and assigned it as a new member of the prostanoid receptor subfamily although it is dissimilar to other members in sequence.

For GPR23, P2RY5 and P2Y10, although they are the most related sequences as shown in Fig.1 and they share identical sequences of 50% between GPR23 and P2RY5 and 30% between P2RY5 and P2Y10, our model placed P2RY5 and P2Y10 in two distinct subfamilies: the P2Y subfamily for P2RY5 and the sphingolipid subfamily for P2Y10 while no ligand was predicted for GPR23. In detail, the model predicted interactions for P2RY5 with four of six typical agonists of the P2Y receptors and no interaction with the ligands from other subfamilies or with LPA. Although the very recent evidence showed that P2RY5 is a LPA receptor, P2RY5 was previously

reported to response to ATP^[46]. For P2Y10, the model correctly predicted S1P as the ligand although the protein has low sequence identity (< 18%) with the EDG subfamily that was the only proteins interacting with S1P in the training data set.

In addition, we applied the model to three other orphan GPCRs, GPR4, GPR65, and GPR68, which were previously reported to be activated by sphingosylphosphorylcholine (SPC) and lysophosphatidylcholine (LPC) and formed a novel family of receptors for lipid messengers. Our model made negative decisions for all the pairs between these three GPCRs and SPC or LPC, in agreement with the recent experimental results showing these three GPCRs are proton-sensing receptors (see review^[47] for the history of the de-orphanized of these three GPCRs).

The results of these external tests showed that the model was able to predict the GPCR-ligand pairs coming from a different chemo-genomic space that were not learnt by the model. Moreover, the first external test demonstrated that the method could be applied to *in silico* high-throughput screening for the discovery of new chemicals that bind to different GPCRs, namely simultaneous multiple-receptor screening. Furthermore, the predictions for the orphan GPCRs in the second external test, particularly for GPR44 and P2Y10 indicated that the method was able to uncover the similarity of dissimilar protein sequences in terms of ligand binding, which distinguishes it from sequence-homology-based methods.

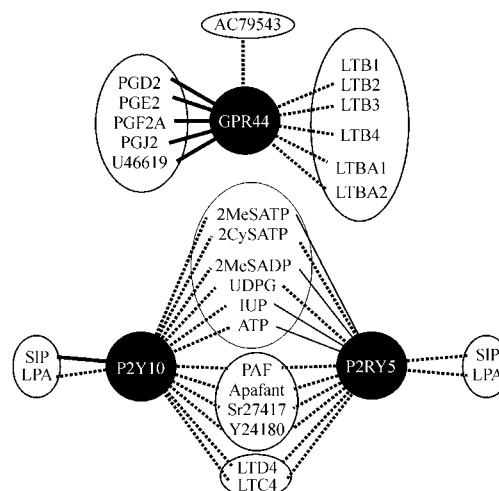


Fig.5 Putative interaction pairs of GPR44, GPR23, P2RY5 and P2Y10 and their prediction results. The ligands are the agonists or antagonists of the GPCRs that are the closest neighbors of the tested GPCRs in the phylogenetic tree as shown in Fig.1 (see text for the detail). The ligands of the same GPCR subfamily are circled. Bold solid lines represent predicted and experimentally verified interactions. Thin solid lines represent predicted interactions without experimental confirmation. Dash lines represent predicted non-interactions.

4 Discussion and conclusion

GPCRs and their ligands comprised a complex receptor-ligand interaction network. We have developed a method for predicting such a network by building a binary classification model to distinguish interacting GPCR-ligand pairs from non-interacting pairs. The good predictive performance of the model obtained indicates that the chemo-genomic information-based descriptors of GPCR-ligand pairs have captured the fundamental properties of GPCR-ligand interactions. In particular, this reflects the importance of the charge properties of the amino acids in the transmembrane segment for ligand binding, as a minimum 3-type classification (neutral residues, positively charged residues and negatively charged residues) has

been able to generate a predictive model. In principle, such a 3-type classification does not necessarily represent the only suitable classification of the amino acids. For example, the neutral residues could be further classified. However, such classification may be the most practical one for GPCRs when considering the limited number of the proteins.

We did tests on the 7-type classification proposed by Shen and coworkers^[33] based on the dipoles and volumes of the amino acids, in which the neutral residues were classified to additional 5 types. This classification generated a $7 \times 7 \times 7 = 343$ -dimensional vector for a protein sequence. Our test of using this 7-type classification generated a model with similar quality to that of the 3-type classification, having a sensitivity of 98.0%. However, the model did not make any positive decision for the putative GPCR-ligand pairs in the external tests. One possible reason could be over fitting because the number of the descriptors of a receptor far exceeded the number of the receptors in the data set. And its complete failure in predicting any interacting pair for the novel orphan GPCRs could suggest another possible reason that the similarities between protein sequences were overly diminished by the sparse and high dimensional protein descriptors. In principle, however, if feature selection^[35,36] that is similar to variable selection in classical QSAR methods could be integrated into the available SVM software, high dimensional descriptors could be used, for example, all 20 amino acids could be used to generate a $20 \times 20 \times 20 = 8000$ -dimensional vector for a protein sequence and then subject to feature selection. We speculate that this would very likely highlight charged residues as the most important descriptors as revealed in this work.

Furthermore, as demonstrated in the external tests, the model was able to correctly predict the interactions of a number of GPCR-ligand pairs in which the ligands and the GPCRs were chemically and phylogenetically novel to the training data set. In particular, the predictions for the orphan GPCRs demonstrated that the method is superior to sequence-homology-based methods. Moreover, the method provided a straightforward way to investigate the cross-pharmacology among different GPCRs by enabling simultaneous multiple-receptor screening.

Supporting information available:

- 1) list of the 57 classical QSAR descriptors used to represent ligands (Table S1);
- 2) list of the 27-dimensional vectors of all 221 GPCR sequences (214 in the training set plus 7 in the external tests) studied in this work (a Microsoft Excel table);
- 3) distance matrix of 113 human GPCRs computed by the PROTDIST program in the PHYLIP package.

And the C language programs and Linux C-shell scripts that were used to prepare the data and test the models are available upon request.

Abbreviations:

PGD2	prostaglandin D2
PGE2	prostaglandin E2
PGF2A	prostaglandin F2a
PGJ2	prostaglandin J2
U46619	15-Hydroxy-11 α ,9 α -(epoxymethano) prosta-5,13-dienoic acid
LTB1	12-oxoleukotriene B4
LTB2	12R-HETE; LTB3, 20-carboxy-leukotriene B4
LTBA1	N-[4-oxo-2-(2H-tetrazol-5-yl)chromen-7-yl]-4-(4-phenylbutoxy)ben-zamide
LTBA2	6-(6-(3-Hydroxy-1,5-undecadien-1-yl)-2-pyridinyl)-1, 5-hexanediol
LPA	lysophosphatidic acid
S1P	sphingosine 1-phosphate

2CySATP	2-cyanoethylthioATP
2MeSADP	2-methylthio-ADP
2MeATP	2-methylthio-ATP
IUP	inosine triphosphate
ATP	adenosine triphosphate
UDPG	UDP-glucuronic acid
PAF	Platelet activating factor
SR27417	N,N-dimethyl-N'-(pyridin-3-ylmethyl)-N'-[4-(2,4,6-tripropyl-2-ylphenyl)-1,3-thiazol-2-yl]ethane-1,2-diamine
LTC4	Leukotriene C4
LTD4	Leukotriene D4
AC79543	(4-chlorophenyl)-3-(2-(dimethylamino)ethyl)isochroman-1-one hydrochloride (agonist of UR2R)

References

- 1 Fredriksson R and Schiöth HB. The repertoire of G-protein-coupled receptors in fully sequenced genomes. *Mol Pharmacol*, 2005, 67: 1414-1425.
- 2 Vassiliatis DK, Hohmann JG, Zeng H, Li F, Ranchalis JE, Mortrud MT, Brown A, Rodriguez SS, Weller JR, Wright AC, Bergmann JE and Gaitanaris GA. The G protein-coupled receptor repertoires of human and mouse. *Proc Natl Acad Sci USA*, 2003, 100: 4903-4908.
- 3 Fredriksson R, Lagerstrom MC, Lundin LG and Schiöth HB. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups and fingerprints. *Mol Pharmacol*, 2003, 63: 1256-1272.
- 4 Chung S, Funakoshi T and Civelli O. Orphan GPCR research. *Br J Pharmacol*. 2008, 153: S339-S346.
- 5 Civelli O, Saito Y, Wang Z, Nothacker HP and Reinscheid RK. Orphan GPCRs and their ligands. *Pharmacol Ther*, 2006, 110: 525-532.
- 6 Oprea TI, Tropsha A, Faulon JL and Rintoul MD. Systems chemical biology. *Nat Chem Biol*, 2007, 3: 447-450.
- 7 Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, Le Trong I, Teller DC, Okada T, Stenkamp RE, Yamamoto M and Miyano M. Crystal structure of rhodopsin: A G-protein-coupled receptor. *Science*, 2000, 289: 739-745.
- 8 Okada T, Sugihara M, Bondar AN, Elstner M, Entel P and Buss V. The retinal conformation and its environment in rhodopsin in light of a new 2.2 Å crystal structure. *J Mol Biol*, 2004, 342: 571-583.
- 9 Murakami M and Kouyama T. Crystal structure of squid rhodopsin. *Nature*, 2008, 453: 363-367.
- 10 Warne T, Serrano-Vega MJ, Baker JG, Moukhametzianov R, Edwards PC, Henderson R, Leslie AG, Tate CG and Schertler GF. Structure of a beta1-adrenergic G-protein-coupled receptor. *Nature*, 2008, 454: 486-491.
- 11 Cherezov V, Rosenbaum DM, Hanson MA, Rasmussen SG, Thian FS, Kobilka TS, Choi HJ, Kuhn P, Weis WI, Kobilka BK and Stevens RC. High-resolution crystal structure of an engineered human {beta}2-adrenergic G protein coupled receptor. *Science*, 2007, 318: 1258-1265.
- 12 Rosenbaum DM, Cherezov V, Hanson MA, Rasmussen SG, Thian FS, Kobilka TS, Choi HJ, Yao XJ, Weis WI, Stevens RC and Kobilka BK. GPCR engineering yields high-resolution structural insights into {beta}2 adrenergic receptor function. *Science*, 2007, 318: 1266-1273.
- 13 Jaakola VP, Griffith MT, Hanson MA, Cherezov V, Chien YET, Lane JR, Ijzerman AP and Stevens RC. The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. *Science*, 2008, 322: 1211-1217.
- 14 Wang T and Duan Y. Chromophore channeling in the G-protein coupled receptor rhodopsin. *J Am Chem Soc*, 2007, 129: 6970-6971.
- 15 Kontijevskis A, Petrovska R, Mutule I, Uhlen S, Komorowski J, Prusis P and Wikberg JE. Proteochemometric analysis of small cyclic peptides' interaction with wild-type and chimeric melanocortin receptors. *Proteins*, 2007, 69: 83-96.
- 16 Lapinsh M, Prusis P, Uhlen S and Wikberg JE. Improved approach for proteochemometrics modeling: application to organic compound-amine G protein-coupled receptor interactions. *Bioinformatics*, 2005, 21: 4289-4296.
- 17 Lapinsh M, Prusis P, Petrovska R, Uhlen S, Mutule I, Veiksina S and Wikberg JE. Proteochemometric modeling reveals the interaction site for Trp9 modified alpha-MSH peptides in melanocortin receptors. *Proteins*, 2007, 67: 653-660.
- 18 Strombergsson H, Prusis P, Midelfart H, Lapinsh M, Wikberg JE and Komorowski J. Rough set-based proteochemometrics modeling of G-protein-coupled receptor-ligand interactions. *Proteins*, 2006, 63: 24-34.
- 19 Bock JR and Gough DA. Virtual screen for ligands of orphan G protein-coupled receptors. *J Chem Inf Model*, 2005, 45: 1402-1414.

- 20 Vapnik V. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995.
- 21 Okuno Y, Yang J, Taneishi K, Yabuuchi H and Tsujimoto G. GLIDA: GPCR-ligand database for chemical genomic drug discovery. *Nucleic Acids Res.* 2006, 34: D673-677.
- 22 Saeh JC, Lyne PD, Takasaki BK and Cosgrove DA. Lead hopping using SVM and 3D pharmacophore fingerprints. *J Chem Inf Model*, 2005, 45: 1122-1133.
- 23 Jorissen RN and Gilson MK. Virtual screening of molecular databases using a support vector machine. *J Chem Inf Model*, 2005, 45: 549-561.
- 24 Millan MJ, Gobert A, Lejeune F, Dekeyne A, Newman-Tancredi A, Pasteau V, Rivet JM and Cussac D. The novel melatonin agonist agomelatine (S20098) is an antagonist at 5-hydroxytryptamine_{2C} receptors, blockade of which enhances the activity of frontocortical dopaminergic and adrenergic pathways. *J Pharmacol Exp Ther*, 2003, 306: 954-964.
- 25 Mellor EA, Maekawa A, Austen KF and Boyce JA. Cysteinyl leukotriene receptor 1 is also a pyrimidinergic receptor and is expressed by human mast cells. *Proc Natl Acad Sci U S A*, 2001, 98: 7964-7969.
- 26 Mamedova L, Capra V, Accomazzo MR, Gao ZG, Ferrario S, Fumagalli M, Abbraccio MP, Rovati GE and Jacobson KA. CysLT1 leukotriene receptor antagonists inhibit the effects of nucleotides acting at P2Y receptors. *Biochem Pharmacol*, 2005, 71: 115-125.
- 27 Thompson JD, Higgins DG and Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 1994, 22: 4673-4680.
- 28 Felsenstein J. PHYLIP- phylogeny inference package (version 3.2). *Cladistics*, 1989, 5: 164-166.
- 29 Wistrand M, Kall L and Sonnhammer EL. A general model of G protein-coupled receptor sequences and its application to detect remote homologs. *Protein Sci*, 2006, 15: 509-521.
- 30 Glick M, Jenkins JL, Nettles JH, Hitchings H and Davies JW. Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and laplacian-modified naive bayesian classifiers. *J Chem Inf Model*, 2006, 46: 193-200.
- 31 Lepp Z, Kinoshita T and Chuman H. Screening for new antidepressant leads of multiple activities by support vector machines. *J Chem Inf Model*, 2006, 46: 158-167.
- 32 Bock JR and Gough DA. Predicting protein-protein interactions from primary structure. *Bioinformatics*, 2001, 17: 455-460.
- 33 Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y and Jiang H. Predicting protein-protein interactions based only on sequences information. *Proc Natl Acad Sci USA*, 2007, 104: 4337-4341.
- 34 Bhasin M and Raghava GP. GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. *Nucleic Acids Res*, 2004, 32: W383-389.
- 35 Weston J, Mukherjee S, Chapelle O, Pontil M, Poggio T and Vapnik V. In *Advances in Neural Information Processing Systems*; Todd K. Leen, T. G. D., Volker Tresp, Ed. 2000, 13: 668-674.
- 36 Zhang X, Lu X, Shi Q, Xu XQ, Leung HC, Harris LN, Iglehart JD, Miron A, Liu JS and Wong WH. Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics*, 2006, 7: 197.
- 37 Green AR. Gaddum and LSD: the birth and growth of experimental and clinical neuropharmacology research on 5-HT in the UK. *Br J Pharmacol*, 2008, 154: 1583-1599.
- 38 Roth BL, Baner K, Westkaemper R, Siebert D, Rice KC, Steinberg S, Ernsberger P and Rothman RB. Salvinorin A: a potent naturally occurring nonnitrogenous kappa opioid selective agonist. *Proc Natl Acad Sci U S A*, 2002, 99: 11934-11939.
- 39 Durant JL, Leland BA, Henry DR and Nourse JG. Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci*, 2002, 42: 1273-1280.
- 40 Sawyer N, Cauchon E, Chateaufort A, Cruz RP, Nicholson DW, Metters KM, O'Neill GP and Gervais FG. Molecular pharmacology of the human prostaglandin D2 receptor, CRTH2. *Br J Pharmacol*, 2002, 137: 1163-1172.
- 41 Sugimoto H, Shichijo M, Okano M and Bacon KB. CRTH2-specific binding characteristics of [3H]ramatroban and its effects on PGD₂-, 15-deoxy-Delta12, 14-PGJ₂- and indomethacin-induced agonist responses. *Eur J Pharmacol*, 2005, 524: 30-37.
- 42 Noguchi K, Ishii S and Shimizu T. Identification of p2y9/GPR23 as a novel G protein-coupled receptor for lysophosphatidic acid, structurally distant from the Edg family. *J Biol Chem*, 2003, 278: 25600-25606.
- 43 Pasternack SM, von Kugelgen I, Aboud KA, Lee YA, Ruschendorf F, Voss K, Hillmer, AM, Molderings GJ, Franz T, Ramirez A, Nurnberg P, Nothen MM and Betz RC. G protein-coupled receptor P2Y5 and its ligand LPA are involved in maintenance of human hair growth. *Nat Genet*, 2008, 40: 329-334.
- 44 Murakami M, Shiraishi A, Tabata K and Fujita N. Identification of the orphan GPCR, P2Y(10) receptor as the sphingosine-1-phosphate and lysophosphatidic acid receptor. *Biochem Biophys Res Commun*, 2008, 371: 707-712.
- 45 Briscoe CP, Tadayyon M, Andrews JL, Benson WG, Chambers JK, Eilert MM, Ellis C, Elshourbagy NA, Goetz AS, Minnick DT, Murdock PR, Sauls HR, Jr Shabon U, Spinage LD, Strum JC, Szekeres PG, Tan KB, Way JM, Ignar DM, Wilson S and Muir AI. The orphan G protein-coupled receptor GPR40 is activated by medium and long chain fatty acids. *J Biol Chem*, 2003, 278: 11303-11311.
- 46 King BF and Townsend-Nicholson A. Recombinant P2Y receptors: the UCL experience. *J Auton Nerv Syst*, 2000, 81: 164-170.
- 47 Seuwen K, Ludwig MG and Wolf RM. Receptors for protons or lipid messengers or both? *J Recept Signal Transduct Res*, 2006, 26: 599-610.

在基因化学的规模上识别 G 蛋白偶合受体(GPCR)与其配体的相互作用

王亭*, 段勇

(加利福尼亚大学, 戴维斯分校, 基因中心, CA95616-8816, 美国)

摘要: G 蛋白偶合受体(GPCR)不仅是一类重要的生物膜蛋白, 而且代表着一类重要的治疗疾病的生物靶标。长期以来, 学术研究界和制药工业界都在努力寻找能与这些蛋白发生相互作用的配体分子以期成为潜在药物, 其中包括对那些生物功能还未知的 GPCR 的配体的寻找。一个能对 GPCR 以及可能配体的相互作用关系作出准确预报和筛选的方法显然可以加速这一过程, 尤其是这个方法具有进行大规模预报的能力, 即可以同时处理多个 GPCR 和大量小分子配体。这样的预报结果可以揭示出多个不同 GPCR 和小分子的交叉反应关系, 有助于减少药物副作用和毒性。本文提出这样一种使用支持向量机器学习法在基因化学的规模上识别 GPCR-配体的相互作用的方法。其基本思想是以其跨膜部分的氨基酸序列表达每个 GPCR 蛋白和以化学分子结构表达每个配体。我们将此方法应用于一套 GPCR-配体相互作用数据集, 其中的 GPCR 包含 A 类中的 28 个子类。建立的模型能够正确预报 86.9% 的相互作用配对和 99.97% 的非相互作用配对。该模型的预报能力进一步在新的数据集上进行了验证, 其中的小分子配体具有全新化学结构, GPCR 的氨基酸序列也是全新的, 并且在进化上远离训练集中的 GPCR。基于以上结果, 我们认为该方法可以应用于大规模筛选 GPCR 小分子配体, 以及识别生物功能还未知的 GPCR 的配体, 这些都有助于加速针对 GPCR 的药物设计。

关键词: G 蛋白偶合受体(GPCR); 配体; 氨基酸序列; 机器学习法; 药物发现

中图分类号:

文献标识码: A

文章编号: 1001-4160(2009)06-689-696

收稿日期: 2009-3-5; **修回日期:** 2009-4-18

联系人: 王亭. E-mail: twang@ucdavis.edu