Molecular Systems Biology 7; Article number 472; doi:10.1038/msb.2011.5
**Citation: *Molecular Systems Biology* 7:472**

www.molecularsystemsbiology.com

molecular systems biology

# Analysis of multiple compound–protein interactions reveals novel bioactive molecules

**Hiroaki Yabuuchi[1,5], Satoshi Niijima[1,5], Hiromu Takematsu[2], Tomomi Ida[1], Takatsugu Hirokawa[3], Takafumi Hara[4], Teppei Ogawa[1], Yohsuke Minowa[1], Gozoh Tsujimoto[4] and Yasushi Okuno[1,*]**

[1] Department of Systems Biosciences for Drug Discovery, Graduate School of Pharmaceutical Sciences, Kyoto University, Kyoto, Japan, [2] Laboratory of Membrane Biochemistry and Biophysics, Graduate School of Biostudies, Kyoto University, Kyoto, Japan, [3] Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan and [4] Department of Genomic Drug Discovery Science, Graduate School of Pharmaceutical Sciences, Kyoto University, Kyoto, Japan
[5] These authors contributed equally to this work
* Corresponding author. Department of Systems Biosciences for Drug Discovery, Graduate School of Pharmaceutical Sciences, Kyoto University, 46-29 Yoshida-Shimo-Adachi-cho, Sakyo-ku, Kyoto 606-8501, Japan. Tel.: +81 75 753 4559; Fax: +81 75 753 4559; E-mail: okuno@pharm.kyoto-u.ac.jp

Received 21.7.10; accepted 20.1.11

**The discovery of novel bioactive molecules advances our systems-level understanding of biological processes and is crucial for innovation in drug development. For this purpose, the emerging field of chemical genomics is currently focused on accumulating large assay data sets describing compound–protein interactions (CPIs). Although new target proteins for known drugs have recently been identified through mining of CPI databases, using these resources to identify novel ligands remains unexplored. Herein, we demonstrate that machine learning of multiple CPIs can not only assess drug polypharmacology but can also efficiently identify novel bioactive scaffold-hopping compounds. Through a machine-learning technique that uses multiple CPIs, we have successfully identified novel lead compounds for two pharmaceutically important protein families, G-protein-coupled receptors and protein kinases. These novel compounds were not identified by existing computational ligand-screening methods in comparative studies. The results of this study indicate that data derived from chemical genomics can be highly useful for exploring chemical space, and this systems biology perspective could accelerate drug discovery processes.**
*Molecular Systems Biology* **7:** 472; published online 1 March 2011; doi:10.1038/msb.2011.5
*Subject Categories:* bioinformatics; computational methods
*Keywords:* chemical genomics; data mining; drug discovery; ligand screening; systems chemical biology



## Introduction

Experimental perturbations of biological systems, such as genetic mutation and chemical exposure, have been used as powerful approaches to deepen our systems-level understanding of biological processes and to discover unprecedented biological phenomena (Lehár *et al*, 2008). In particular, perturbations by chemical probes provide broader applications not only for analysis of complex systems but also for intentional manipulations of these systems, e.g., a medicine is a small molecule designed for the purpose of clinical therapy that can actively manipulate biological systems from disordered to well-ordered states. Unfortunately, the number of well-characterized chemical probes is highly limited, which has bottlenecked their wide range of application.

The set of all possible small organic molecules, referred to as chemical space, has been estimated to consist of more than $10^{60}$ compounds (Dobson, 2004). Chemical space is as vast as the diversity of biological systems, and the vastness of the two domains creates difficulty in comprehensive understanding of the interface between chemical space and biological systems (Lipinski and Hopkins, 2004; Renner *et al*, 2009). Recently, chemical genomics has emerged as a promising area of research applicable to exploration of novel bioactive molecules, and researchers are currently striving toward the identification of all possible ligands for all target protein families (Wang *et al*, 2009). Large-scale data sets of compound–protein interactions (CPIs) are being collected, and chemical genomics studies have shown that patterns of protein–ligand interactions are too diverse to be understood as simple one-to-one events. For example, multiple structurally different compounds have been shown to bind the same protein or express similar biological activities (Eckert and Bajorath, 2007; Young *et al*, 2008). In other cases, one drug has been shown to affect multiple targets from different protein families (MacDonald *et al*, 2006; Paolini *et al*, 2006). This phenomenon, termed polypharmacology, is thought to be one critical cause of adverse drug effects (Hopkins, 2008). There-

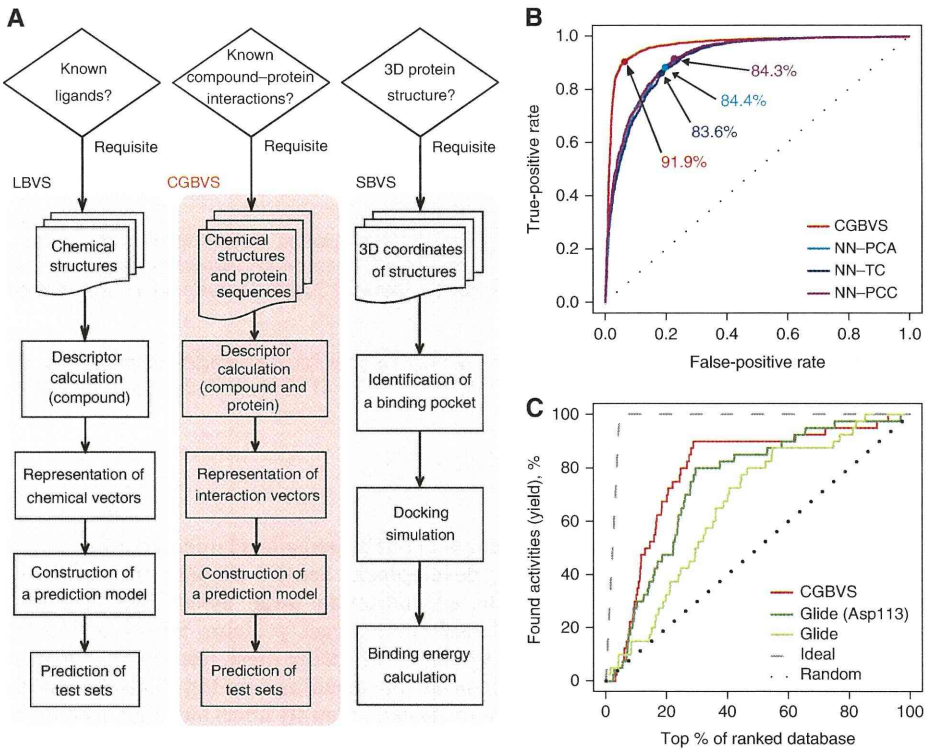


**Figure 1** Overview and performance of CGBVS. (**A**) Comparison of the strategies used for CGBVS, LBVS, and SBVS. CGBVS used multiple CPI data, represented the CPI in vector form, and used SVM for CPI pattern learning. (**B**) ROC curves obtained by fivefold cross-validation using compound–GPCR interactions for the CGBVS (red) and LBVS methods. The best accuracy rate for each method is also shown. NN–PCA (light blue), nearest neighbor (NN) method with the Pearson correlation coefficient in the constructed space using principal component analysis (PCA); NN–TC (navy), NN method with the Tanimoto coefficient (TC); and NN–PCC (purple), NN method with the Pearson correlation coefficient (PCC) in the original space. (**C**) Enrichment curves obtained by the CGBVS and SBVS methods. Glide was used both without constraints and with constrained hydrogen bonding between the compounds and Asp113, a residue known to be crucial for ligand binding to ADRB2 (Strader *et al*, 1987). Information regarding interactions between ADRB2 and its ligands was not used in test set for machine learning for CGBVS.

fore, an integrative understanding of multiple interactions among chemical and biological components beyond a one-compound/one-target simplification could open up new opportunities in drug development, but the need to develop appropriate data mining methods for characterizing and visualizing the full complexity of interactions between chemical space and biological systems is urgent (Oprea *et al*, 2007). Recently, mining of multiple CPI data sets has been used to identify new protein targets for known drugs and thereby predict unreported polypharmacology (Keiser *et al*, 2009). However, this approach only identifies additional targets for known drugs. No existing screening approach has so far succeeded in identifying novel bioactive compounds using multiple interactions among compounds and target proteins, and the potential application of analysis of multiple CPIs to identify novel bioactive molecules remains unknown.

High-throughput screening (HTS) and computational screening have greatly aided in the identification of early lead compounds for drug discovery. However, the large numbers of assays required for HTS to identify drugs that target multiple proteins render this process very costly and time-consuming. Therefore, interest in using *in silico* strategies for screening has been increasing. The most common computational approaches, ligand-based virtual screening (LBVS) and structure-based virtual screening (SBVS; Oprea and Matter, 2004; Muegge and Oloff, 2006; McInnes, 2007; Figure 1A), have been used for practical drug development. Unfortunately, these methods have important limitations. LBVS aims to identify molecules that are very similar to known active molecules and generally has difficulty identifying compounds with novel structural scaffolds that differ from reference molecules. Attempts to scaffold-hop using LBVS are prone to identification of increased numbers of false positives (Eckert and Bajorath, 2007). Therefore, the primary objective of virtual screening, reduction of the number of candidate compounds to be assayed, remains unachievable using this method. The other popular strategy, SBVS, is constrained by the number of three-dimensional crystallographic structures available and, more importantly, by the difficulty of accurately simulating molecular docking processes for targets, including membrane-spanning G-protein-coupled receptors (GPCRs). To circumvent these limitations, we have shown that a new computational screening strategy, chemical genomics-based virtual screening (CGBVS), has the potential to identify novel, scaffold-hopping compounds and assess their polypharmacology by using a machine-learning method to recognize conserved molecular patterns in comprehensive CPI data sets.

## Results

### Theoretical framework for CGBVS

The CGBVS strategy is made up of five steps: CPI data collection, descriptor calculation, representation of interaction vectors, predictive model construction using training data sets, and predictions from test data (Figure 1A and Supplementary Figure S1). Importantly, step 1, the construction of a data set of chemical structures and protein sequences for known CPIs, does not require the three-dimensional protein structures needed for SBVS. We chose GPCRs, important pharmaceutical targets (Hopkins and Groom, 2002), as our first target proteins for virtual ligand screening. In total, 5207 CPIs (including 317 GPCRs and 866 ligands) retrieved from the GLIDA database (Okuno *et al*, 2006) were used as experimental data (Supplementary Table S1). In step 2, compound structures and protein sequences were converted into numerical descriptors using 929-dimensional and 400-dimensional feature vectors, respectively. A wide variety of chemical descriptors was used to describe the substructures, as well as the physicochemical and molecular properties of the small molecules. Descriptors for protein sequences were created using a string kernel (see Materials and methods section and Supplementary information for details). These descriptors were used to construct chemical or biological spaces, in which decreasing distance between vectors corresponded to increasing similarity of compound structures or protein sequences. In step 3, we represented multiple CPI patterns by concatenating these chemical and protein descriptors (in $929+400$ dimensions). Using these interaction vectors, we could quantify the similarity of molecular interactions for compound–protein pairs, despite the fact that the ligand and protein similarity maps differed substantially (Keiser *et al*, 2007). In step 4, concatenated vectors for CPI pairs (positive samples) and non-interacting pairs (negative samples) were input into a support vector machine (SVM; Vapnik, 1995), an established machine-learning technique widely applied to pattern-recognition problems (Schölkopf *et al*, 2004; Shawe-Taylor and Cristianini, 2004). Using training sets, an SVM classifier was generated as a hyperplane dividing positive and negative samples into two distinct classes representing interaction and non-interaction. By mapping the samples into high-dimensional feature space using a nonlinear kernel function, samples that were linearly inseparable in the original input space could be linearly separated in the feature space. As a nonlinear SVM can extract patterns from data sets with nonlinear characteristics, non-intuitive interaction rules can be obtained from multiple CPI patterns, creating the potential to identify novel CPIs. In the final step, the SVM classifier constructed using training sets was applied to test data. Along with providing simple yes/no outputs, the calculated prediction scores also ranked all test compound–GPCR pairs in the order of binding probability.

### Computational evaluation of CGBVS

To evaluate the predictive value of CGBVS, we compared its performance with that of LBVS methodologies using respective data sets of 5207 interacting and non-interacting pairs. The performance of each method was tested by repeating fivefold cross-validations 20 times. CGBVS performed with a considerably higher accuracy ($91.9 \pm 0.3\%$) than LBVS ($84.4 \pm 0.3\%$, at best). We also recorded the number of true-positive interactions as a function of false positives and plotted receiver operating characteristic (ROC) curves (Hanley and McNeil, 1982), as shown in Figure 1B and Supplementary Table S2. ROC analysis revealed that CGBVS performed better than all LBVS methods in terms of the score ranking of CPI pairs.

Recently, the crystal structure of the β2-adrenergic receptor (ADRB2) has been determined (Cherezov *et al*, 2007; Rasmussen *et al*, 2007). Therefore, we were able to compare CGBVS and SBVS in a retrospective virtual screening based on the human ADRB2 using a representative docking program, Glide (Friesner *et al*, 2004). For CGBVS, we constructed a predictive model based on 5167 CPI pairs, excluding 40 known ADRB2-related CPIs to avoid any bias in favor of CGBVS during the machine-learning step. Using both methods, we predicted scores for the same 866 known GPCR ligands, including the 40 known ADRB2 ligands as positive controls. We asked whether the scores for these 40 known positive compounds were higher than scores for other compounds. Figure 1C and Supplementary Table S3 show that CGBVS provided higher enrichment factors (EFs) and hit rates than did SBVS. These results suggest that CGBVS is more successful than conventional approaches for prediction of CPIs.

### Polypharmacological interactions for ADRB2

We also evaluated the ability of the CGBVS method to predict the polypharmacology of ADRB2 by attempting to identify novel ADRB2 ligands from the above ligand data set. As an established, well-studied pharmaceutical target (Waldeck, 2002), we expected that novel ADRB2 ligands would be difficult to find, and that searching for novel scaffolds for such a well-known target would be a stringent assessment of the predictive ability of CGBVS. After training an SVM classifier using all 5207 CPIs, we ranked the prediction scores for the interactions of 826 reported GPCR ligands (excluding the 40 known ADRB2 ligands) with ADRB2, and then analyzed the 50 highest-ranked compounds in greater detail. To complement the less-than-comprehensive binding data available in the original GLIDA database, a literature search was performed using SciFinder (Wagner, 2006) and PubMed (http://www.ncbi.nlm.nih.gov/pubmed/). This search identified 15 of the top 50 compounds as known ADRB2 ligands. Importantly, these compounds were identified as ligands of other GPCRs, not of ADRB2, in training sets used in the machine-learning step. ADRB2 ligands already reported in the literature were excluded from analysis, but the remainder were tested in *in vitro* binding assays. From the remaining 35 ligands, 21 were commercially available. Of these 21, 11 were not previously reported, but were discovered to bind to ADRB2 (Figure 2A and Supplementary Table S5). These compounds included ligands for the acetylcholine, serotonin, dopamine, and neuropeptide Y receptors (Figure 2E), indicating the presence of potential polypharmacological interactions with ADRB2.

To substantiate the novelty of the ligands identified by CGBVS, we compared the predictive scores estimated by the three virtual screening approaches (CGBVS, LBVS, and SBVS).

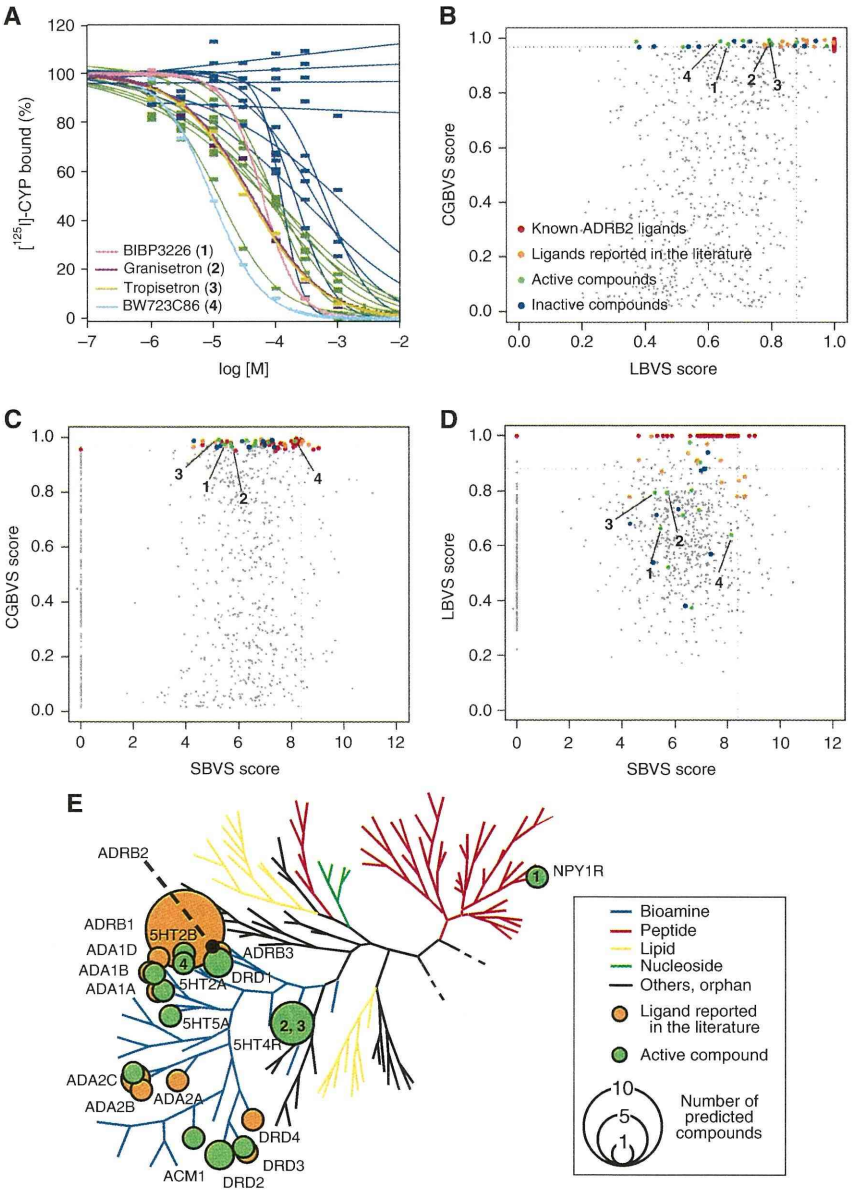


**Figure 2** Prediction of ADRB2 ligands. (**A**) Binding curves for 21 of the top 50 compounds ranked by CGBVS available commercially. Pink, BIBP3226 (**1**); purple, granisetron (**2**); dark yellow, tropisetron (**3**); light blue, BW723C86 (**4**) (Supplementary Table S5); green, 11 active compounds screened ($K_i < 100\ \mu M$); and blue, inactive compounds. (**B–D**) ADRB2 ligand prediction using CGBVS, LBVS (NN-PCA), and SBVS (GlideScore, Asp113). The NN–PCA score indicates the Pearson correlation coefficient between each compound and its nearest ligands in principal component space. Red, ADRB2 ligands; orange, ligands reported in the literature; green, screened active compounds; and blue, inactive compounds. The numbers **1–4** are corresponding to BIBP3226 (**1**), granisetron (**2**), tropisetron (**3**), BW723C86 (**4**), respectively. Dotted lines show the score of the fiftieth-ranked compound using each method. (**E**) Polypharmacological relationships of newly identified ADRB2 ligands. Newly identified (green) and known (orange) ADRB2 ligands were mapped on a human GPCR phylogenetic tree (Fredriksson *et al*, 2003). The size of the circles indicates the number of compounds reported to bind to each GPCR on the tree. Target GPCRs of the four compounds **1–4** are shown in the circles. 5HT2A, 5-hydroxytryptamine receptor 2A; 5HT2B, 5-hydroxytryptamine receptor 2B; 5HT4R, 5-hydroxytryptamine receptor 4; 5HT5A, 5-hydroxytryptamine receptor 5A; ACM1, acetylcholine receptor M1; ADA1A, alpha 1-adrenergic receptor type A; ADA1B, alpha 1-adrenergic receptor type B; ADA1D, alpha 1-adrenergic receptor type D; ADA2A, alpha 2-adrenergic receptor type A; ADA2B, alpha 2-adrenergic receptor type B; ADA2C, alpha 2-adrenergic receptor type C; ADRB1, beta 1-adrenergic receptor; ADRB3, beta 3-adrenergic receptor; DRD1, dopamine receptor D1; DRD2, dopamine receptor D2; DRD3, dopamine receptor D3; DRD4, dopamine receptor D4.

We used 26 (orange and green dots in Figure 2B–D) of the top 50 compounds predicted by CGBVS that had ligand activity, but that had not been used in the training data set, to evaluate the other methods. As shown in Figure 2B and D, the known ADRB2 ligands (orange dots) clustered with the highest scores on the LBVS axis, whereas most of the other active compounds identified (green dots) fell outside the top 50 scores. Active compounds (red, orange, and green) were widely scattered along the SBVS axis (Figure 2C and D), and only six were found in the top 50 SBVS scores. This was consistent with the known

**Table I** Compound IDs and names (see Supplementary Table S4 for the chemical structures)

| Compound | GLIDA ID | Bionet ID | Compound name |
|---|---|---|---|
| **1** | L000117 | | BIBP3226 |
| **2** | L003700 | | Granisetron |
| **3** | L002023 | | Tropisetron |
| **4** | L000152 | | BW723C86 |
| **5** | | MS-2742 | 2,5-Dimethyl-1-(2,2,4-trimethyl-2,3-dihydro-1-benzofuran-7-yl)-1H-pyrrole |
| **6** | L001048 | | Codeine |
| **7** | L000315 | | Iodocyanopindolol |
| **8** | L001311 | 12L-933 | 1-(*Tert*-butylamino)-3-[(2-methyl-1H-indol-4-yl)oxy]-2-propanol |
| **9** | | MS-2807 | (2-Aminophenyl)(4-methylphenyl)amine |
| **10** | L013420 | | Phentolamine |
| **11** | L001089 | | Desipramine |
| **12** | | 7W-0360 | Ethyl 1-(4-chlorophenyl)-4-[(4-methoxybenzyl)amino]-3-methyl-1H-pyrazolo[3,4-b]pyridine-5-carboxylate |
| **13** | L001167 | | Cartazolate |
| **14** | | | 3-(6-Aminopyridin-3-yl)-2-(diphenylacetamido)-N-(4-methoxybenzyl)-*N*-methylpropionamide |
| **15** | | 3H-950 | Diethyl 2-(3,5-dimethyl-1H-pyrazol-1-yl)-6-hydroxy-3,5-pyridinedicarboxylate |
| **16** | L000717 | | Nicergoline |
| **17** | | | 3-Ethyl-5-[4-(4-fluorophenyl)-4-(6-fluoropyridin-3-yl)- 5-methyl-4,5-dihydro-1H-imidazol-2-yl]-1-methylpyridin-2(1H)-one |
| **18** | | 11N-058 | 6,7-Dimethoxy-*N*-phenyl-4-quinazolinamine |
| **19** | | | 1-(5-*Tert*-butyl-isoxazol-3-yl)-3-[4-(2-chloro-6,7-dimethoxy-quinazolin-4-ylamino)-phenyl]-urea |
| **20** | | | 5-[6-Methoxy-7-(pyridin-4-ylmethoxy)-quinazolin-4-ylamino]-2-methyl-phenol |
| **21** | | 12N-063 | *N*-{2-[(4-chlorophenyl)sulfanyl]ethyl}-6,7-dimethoxy-4-quinazolinamine |
| **22** | | | 1-(6,7-Dimethoxy-2-pyridin-4-yl-quinazolin-4-ylamino)-indan-2-ol |
| **23** | | MS-2894 | [2-(4-Fluorophenyl)-5,6,7,8-tetrahydroimidazo[2,1-b][1,3]benzothiazol-3-yl]methanol |
| **24** | | | 2-Methyl-6-[6-(6-methyl-pyridin-2-yl)-imidazo[2,1-b]thiazol-5-yl]-3a,7a-dihydro-benzooxazole |
| **25** | | | 1-{4-[4-Amino-5-(2,6-difluoro-benzoyl)-thiazol-2-ylamino]-piperidin-1-yl}-8-methyl-non-6-en-1-one |
| **26** | | 7N-773 | [4-Amino-2-(*tert*-butylamino)-1,3-thiazol-5-yl](4-chlorophenyl)methanone |
| **27** | | | (4-Amino-2-phenylamino-thiazol-5-yl)-(4-chloro-3-methyl-phenyl)-methanone |
| **28** | | 9X-0942 | 2-[2,5-Dimethyl-4-(morpholinomethyl)phenoxy]acetamide |
| **29** | | 2W-0814 | *N*-(*tert*-butyl)-*N'*-(4-methoxybenzyl)thiourea |
| **30** | | MS-0062 | 2-Ethyl-2-{[(2-fluorobenzyl)oxy]methyl}-5,5-dimethyltetrahydrofuran |
| **31** | | MS-3556 | 2-(3-Isopropoxyphenyl)-1-ethanamine |
| **32** | | 3F-004 | 2-Morpholino-2-oxoacetohydrazide |
| **33** | | 1M-918 | 1-[(3-Methoxypropyl)amino]-3-[(2-methyl-1H-indol-4-yl)oxy]-2-propanol |
| **34** | | 6W-0328 | Ethyl 4-chloro-1-(4-chlorophenyl)-3-methyl-1H-pyrazolo[3,4-b]pyridine-5-carboxylate |
| **35** | | 10N-835 | 7-Chloro-*N*-(3-methoxybenzyl)-4-quinazolinamine |
| **36** | | 12N-055 | 6,7-Dimethoxy-*N*-(2-thienylmethyl)-4-quinazolinamine |
| **37** | | 4X-0854 | 2-{[4-(2-Chloroacetyl)-1H-pyrrol-2-yl]methylene}malononitrile |

limitations of LBVS in identification of novel structures and of SBVS in accurate scoring. Figure 2B–D shows that use of CGBVS resulted in the identification of the majority of the novel active compounds (green dots), few of which were identified by LBVS or SBVS. Four of these compounds (**1**–**4** in Table I and Supplementary Table S4) contained novel scaffolds compared with known ADRB2 agonists (catecholamine or isoprenaline derivatives) or ADRB2 antagonists (arylalkylamine derivatives). Notably, these compounds included a neuropeptide Y-type 1 receptor (NPY1R) antagonist (**1**). This observation suggests that only CGBVS could identify this unexpected cross-reaction for a ligand developed as a target to a peptidergic receptor that has low protein homology to ADRB2 (Figure 2E).

## Polypharmacology map of the GPCR family

To identify possible polypharmacological relationships among GPCRs, we constructed polypharmacology maps, first based on multiple interactions between GPCRs and their ligands predicted by CGBVS, and second based on previously reported interactions (Figure 3). CGBVS predicted many unexpected multiple interactions between GPCRs and ligands, including, interestingly, interactions shared by members of distantly related subfamilies. (See Supplementary Figure S2 for a correlation map of ligands and orphan GPCRs with no known ligands.) To better understand the propensity for ligand promiscuity, we extracted chemical substructures characteristic of the putatively promiscuous ligands (Supplementary Figure S3), as described in the Supplementary information. This analysis has shown that tertiary amine and sulfur-containing heterocycles are recurring substructures in the promiscuous ligands when compared with selective ligands (Supplementary Table S6). For example, these substructures are typically seen in antidepressants used to treat depression and anxiety disorders, which interact promiscuously with a range of dopamine and serotonin receptors (Roth *et al*, 2004a). This observation suggests that the ligands containing such substructures can be non-selective.

Unlike CGBVS, SBVS cannot predict CPIs for multiple GPCRs, because only limited three-dimensional structural information is available. LBVS is applicable only to targets with known reference ligands and is therefore unsuitable for identifying polypharmacological interactions, particularly

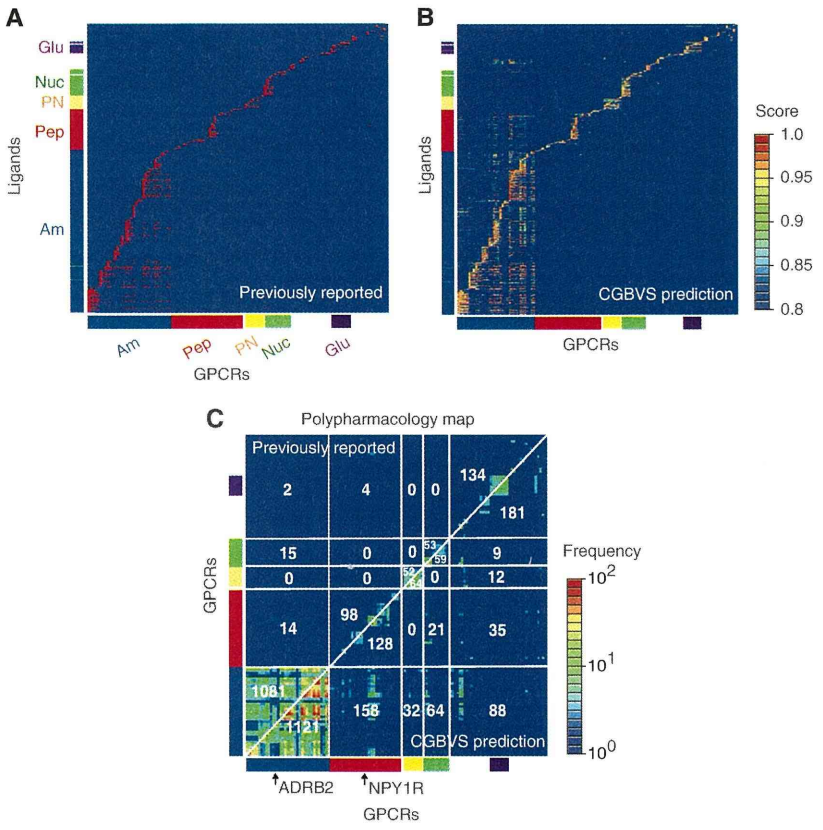


**Figure 3** Comparative polypharmacology maps for GPCRs. (**A**) Map of previously reported compound–GPCR interactions. Vertical and horizontal axes represent the compounds and the GPCRs, respectively. The reported CPIs are depicted as red dots. CGBVS used the CPIs as training data. The colored bars along each axis indicate the classes to which the compounds and GPCRs belong. Am, amines; Pep, peptides; PN, prostanoids; Nuc, nucleotides; and Glu, glutamates. (**B**) Map of predicted compound–GPCR interactions based on CGBVS. The CPIs are plotted with colors ranging from blue (low) to red (high), according to prediction scores. (**C**) Comparative polypharmacology map of GPCRs showing the number of shared compounds within a receptor family. The polypharmacology map was constructed as described by Paolini *et al* (2006) by plotting the numbers of common ligands for two given receptors. Previously reported and CGBVS-predicted interactions are shown in the upper-left and the lower-right diagonal halves, respectively. Each value indicates the number of common ligands for each GPCR subfamily. For example, 1081 compounds were reported to be ligands for amine receptors that cross-reacted with other amine receptors, and 14 amine receptor ligands were reported to cross-react with peptide receptors.

between distantly related GPCRs (Supplementary Figure S4). The cross-reactivity predictions provided by CGBVS also offer a promising approach for scaffold hopping in drug discovery. For example, many small ligands for non-peptidergic GPCRs were predicted to interact with peptidergic GPCRs as well, indicating that CGBVS has further potential in the discovery of novel non-peptidergic compounds for peptidergic receptors by using these small ligands as reference molecules.

## GPCR ligand screening

Although preliminary results indicated that CGBVS was useful for identifying polypharmacological relationships among ligands for the GPCR family, all of the analyzed compounds were known GPCR ligands and, therefore, represent a very limited number of examples within the vastness of chemical space. The true value of CGBVS in lead discovery must be tested by assessing whether this method can identify scaffold-hopping lead compounds from a set of compounds that is structurally more diverse. To assess this ability, we analyzed 11 500 compounds from the Bionet chemical library (Key Organics Ltd, Cornwall, UK) to predict compounds likely to bind to two GPCRs from different subfamilies, ADRB2 and NPY1R (Supplementary Table S7).

The 30 highest-scoring compounds for ADRB2 were tested in calcium mobilization assays, in which nine compounds (hit rate=30%) exhibited either half-maximal effective concentrations ($EC_{50}$) or half-maximal inhibitory concentrations ($IC_{50}$) between 0.7 nM and 65 μM. These results suggest that CGBVS is highly capable of mining of general chemical libraries (Figure 4A and B, Supplementary Figure S5A and B, and Supplementary Table S8A). For NPY1R, the 20 highest-scoring compounds were tested in cAMP assays. Of these compounds, three (hit rate=15%) exhibited agonist activity with $EC_{50}$ values of 16, 16, and 63 μM (Figure 4C, Supplementary Figure S5C and D, and Supplementary Table S8B).

Despite the fact that these $EC_{50}$ values were in the micromolar range, CGBVS could prove highly useful for lead screening in drug development, as the lead-screening stage is distinct from the optimization stage. For lead screening, it is

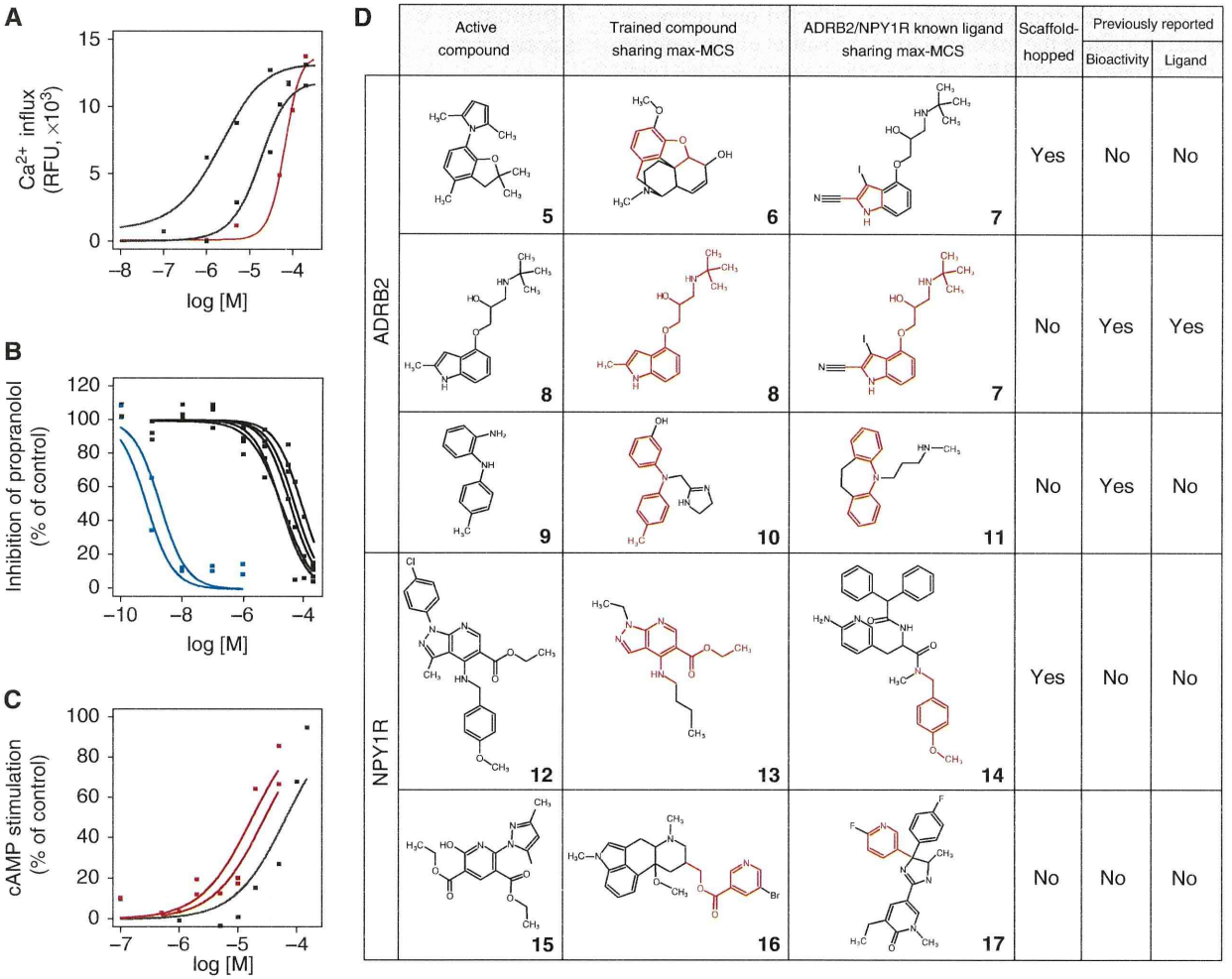


**Figure 4** Experimental confirmation of *in vitro* GPCR activity of compounds and their scaffolds screened from a chemical library. Dose–response curves of the top-ranked compounds from the Bionet chemical library for (**A**) ADRB2 agonists, (**B**) ADRB2 antagonists, and (**C**) NPY1R agonists. Inactive compounds (cutoff of 100 μM in $EC_{50}/IC_{50}$ value) are not shown. Red lines indicate results for compounds exhibiting scaffold hopping based on the criteria explained in the Results section. Blue lines indicate results from compounds with almost completely overlapping structures. Compounds **29** (*N*-(*tert*-butyl)-*N*′-(4-methoxybenzyl)thiourea), **30** (2-ethyl-2-{[(2-fluorobenzyl)oxy]methyl}-5,5-dimethyltetrahydrofuran), and **5** are corresponding to the curves in A from left to right. Compounds **8**, **33**, **30**, **28** (2-[2,5-dimethyl-4-(morpholinomethyl)phenoxy]acetamide), **31** (2-(3-isopropoxyphenyl)-1-ethanamine), **9** ((2-aminophenyl)(4-methylphenyl)amine), and **32** (2-morpholino-2-oxoacetohydrazide) are corresponding to the curves in B from left to right. Compounds **12**, **34**, and **15** (diethyl 2-(3,5-dimethyl-1H-pyrazol-1-yl)-6-hydroxy-3,5-pyridinedicarboxylate) are corresponding to the curves in C from left to right. (**D**) Max-MCSs between identified active compounds (left) and the most relevant compounds found within the entire training compound data set (center) or within the ligand set of each target protein (right) that exhibited scaffold hopping. The max-MCSs between compounds are indicated in red. The columns 'bioactivity' and 'ligand' indicate the existence of publications regarding the active compound: 'bioactivity' indicates whether a publication has already described that the compound is bioactive; 'ligand' indicates whether a publication has uncovered the ADBR2/NPY1R ligand activity, having known the compound is bioactive. All identified active compounds are shown in Supplementary Figure S9. See Table I for compound names of the numbered compounds.

important to identify bioactive compounds with diverse, novel structures, rather than compounds with extremely high activities in the nanomolar range, because lead candidates are subsequently structurally optimized to generate higher activity in the lead-optimization process.

## Evaluation of ligand scaffold hopping

We next wanted to evaluate the extent of scaffold hopping achieved in the identification of these novel ligands. However, so far no explicit definition of scaffold hopping exists. Therefore, we began by establishing definitive criteria for scaffold hopping through analysis of the structural relationships between pairs of newly identified active compounds and known ligands in the training data set by calculating their maximum common substructures (MCSs). The number of constituent atoms and bonds in the MCS is typically used as a measure of structural similarity between two molecules. We first calculated MCSs for each Bionet active compound against all of the GPCR ligands in the training data set. Because known GPCR ligands have diverse molecular scaffolds, we selected a single ligand with the largest MCS value (max-MCS) among all the calculated MCSs as the most relevant structure for each active compound (shown in the middle column of

Figure 4D). For comparison, we also selected one reference ligand exhibiting the max-MCS from the subset of the training data specific for ADRB2 and NPY1R (shown in the right column of Figure 4D). When the two max-MCSs (shown as the red colored substructures of Figure 4D) contained in these two selected ligands did not overlap, the newly identified active compound in the pair was deemed to have undergone scaffold hopping. This can be a useful criterion for screening lead compounds.

We performed scaffold-hopping analysis after having defined the criterion. For example, compound **5** (2,5-dimethyl-1-(2,2,4-trimethyl-2,3-dihydro-1-benzofuran-7-yl)-1H-pyrrole), which showed weak ADRB2 agonist activity (Figure 4A), did not exhibit overlapping substructure between the max-MCSs of codeine (**6**) or iodocyanopindolol (**7**; Figure 4D), which were selected from all the GPCRs and ADRB2 ligand sets, respectively. Therefore, compound **5** was categorized as representative of scaffold hopping. Indeed, the seven active compounds (**5**, **9**, **28**–**32**), including scaffold-hopped compound **5**, identified as ADRB2 ligands did not contain an oxypropanolamine moiety, an established constituent of β-adrenergic blockers (Supplementary Tables S9 and S10A). No biological activities have been reported for four (**5**, **28**, **29**, and **30**) of these compounds, whereas ADRB2 activities of the rest compounds (**9**, **31**, and **32**) have not been reported previously (see Supplementary information for details). In contrast, compounds **8** (Sandoz-21-009) and **33** (1-[(3-methoxypropyl)amino]-3-[(2-methyl-1H-indol-4-yl)oxy]-2-propanol) both showed strong ADRB2 antagonist activity and had max-MCSs that overlapped with that of iodocyanopindolol (**7**), a known ADRB2 ligand (Figure 4D and Supplementary Figure S9A). These compounds were, therefore, categorized as non-hopping, although these were originally reported as serotonin receptor ligands and were thus not included in the training data set for ADRB2. Indeed, this max-MCS contained a representative moiety of β-adrenergic blockers (Supplementary Table S9). Compounds, such as this example with heavily overlapping MCSs, could likely be identified using LBVS. Nevertheless, max-MCS profiling analysis confirmed the reliability of our criteria for scaffold hopping, the accuracy of predictions, and the reliability of the *in vitro* assays. Furthermore, we identified the three novel active compounds for NPY1R that have not previously been known to exhibit biological activity. Of these compounds, compounds **12** (ethyl 1-(4-chlorophenyl)-4-[(4-methoxybenzyl)amino]-3-methyl-1H-pyrazolo[3,4-b]pyridine-5-carboxylate) and **34** (ethyl 4-chloro-1-(4-chlorophenyl)-3-methyl-1H-pyrazolo[3,4-b]pyridine-5-carboxylate) included examples of scaffold hopping (Figure 4D and Supplementary Figure S9B).

Overall, CGBVS identified compounds for both GPCRs analyzed that exhibited scaffold hopping, indicating that CGBVS can use this characteristic to rationally predict novel lead compounds, a crucial and very difficult step in drug discovery. This feature of CGBVS is critically different from existing predictive methods, such as LBVS, which depend on similarities between test and reference ligands, and focus on a single protein or highly homologous proteins. In particular, CGBVS is useful for targets with undefined ligands, because this method can use CPIs with target proteins that exhibit lower levels of homology.

## Application of CGBVS to kinase inhibitor screening

Having demonstrated that CGBVS is a valuable strategy for predicting CPIs for GPCRs, we also wanted to show the general utility of this method for other target proteins. Therefore, we selected the protein kinase family, another popular chemotherapeutic target (Manning *et al*, 2002), for the application of CGBVS. A CGBVS model for the kinase family was constructed using a training data set of 15 616 CPI samples (including 143 kinases and their 8830 inhibitors) from the GVK Biosciences Pvt Ltd., (Hyderabad, India) kinase inhibitor database (Supplementary Table S11). Similar to the GPCR results, polypharmacological predictions for the kinases indicated many possible multiple interactions between kinases and their ligands (Supplementary Figure S6). The analysis of ligand promiscuity has shown that iodophenyl and polycyclic aromatic groups (containing five-membered heterocycles) are characteristic of the putatively promiscuous ligands (Supplementary Figure S7 and Supplementary Table S12). In particular, polycyclic aromatic compounds are likely to interact across kinase subfamilies in a manner reminiscent of staurosporine, a well-known promiscuous inhibitor (Karaman *et al*, 2008).

We focused on two protein kinases, the epidermal growth factor receptor (EGFR) tyrosine kinase and the cyclin-dependent kinase 2 (CDK2) serine/threonine kinase. We first compared CGBVS with LBVS and SBVS by making predictions using a validation data set that was designed for evaluation of docking programs (Huang *et al*, 2006). For both kinases, CGBVS was able to identify true inhibitors within the top-ranked compounds more effectively than the LBVS and SBVS methods (Supplementary Figure S8).

We then made prospective predictions for EGFR and CDK2 from the 11 500 Bionet compounds and selected the 20 highest-scoring compounds for experimental verification (Supplementary Table S7). For EGFR, the off-chip mobility shift assay revealed that 5 of the 20 compounds (hit rate=25%) ranked by CGBVS were inhibitors, with $IC_{50}$ values between 0.014 and 13 μM (Figure 5A, Supplementary Figure S5E and Supplementary Table S13A). However, MCS analysis suggested that these compounds did not exhibit scaffold hopping (Figure 5C and Supplementary Figure S9C). Indeed, the max-MCSs of four of the active compounds (**18**, **21**, **35**, and **36**) for EGFR were quinazoline derivatives (Supplementary Tables S9 and S10B), which are well-characterized EGFR inhibitors that include the antitumor agent gefitinib. Compound **18** (6,7-dimethoxy-*N*-phenyl-4-quinazolinamine) was shown to act as an EGFR inhibitor. Although compounds **35** (7-chloro-*N*-(3-methoxybenzyl)-4-quinazolinamine) and **36** (6,7-dimethoxy-*N*-(2-thienylmethyl)-4-quinazolinamine) were known to inhibit other proteins such as NOD1 and STAT, their inhibitory activities for EGFR have not been reported. No biological activities have been reported for the remaining two compounds **21** (*N*-{2-[(4-chlorophenyl)sulfanyl]ethyl}-6,7-dimethoxy-4-quinazolinamine) and **37** (2-{[4-(2-chloroacetyl)-1H-pyrrol-2-yl]methylene}malononitrile).

For CDK2, 2 of the 20 compounds (hit rate=10%) identified had $IC_{50}$ values of 4.9 and 19 μM in the off-chip mobility shift assay (Figure 5B, Supplementary Figure S5F and

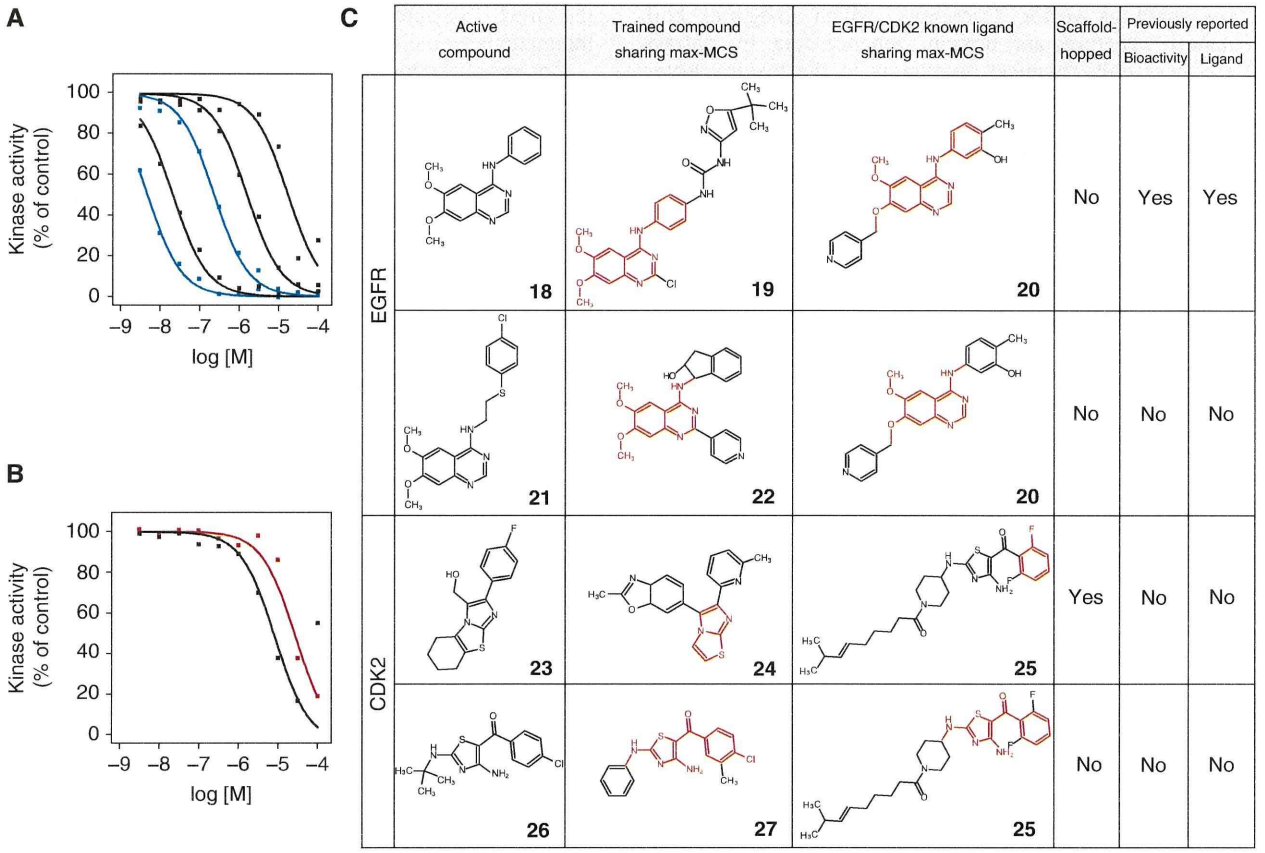


**Figure 5** Experimental confirmation of the *in vitro* kinase activity of compounds and their scaffolds screened from a chemical library. Dose–response curves of the top-ranked compounds from the Bionet chemical library for EGFR tyrosine kinase inhibitor activity (**A**) and CDK2 inhibitor activity (**B**). Inactive compounds (cutoff of 100 μM in $IC_{50}$ value) are not shown. Color of lines in A and B is the same as in Figure 4. Compounds **18**, **36**, **21**, **35**, and **37** are corresponding to the curves in A from left to right. Compounds **26** ([4-amino-2-(*tert*-butylamino)-1,3-thiazol-5-yl](4-chlorophenyl)methanone) and **23** are corresponding to the curves in B from left to right. (**C**) Max-MCSs and publication status are similar to Figure 4 for the EGFR and CDK2. See Table I for compound names of the numbered compounds.

Supplementary Table S13B), and bioactivity of these two compounds also has not been reported previously. One active compound, **23** ([2-(4-fluorophenyl)-5,6,7,8-tetrahydroimidazo[2,1-b][1,3]benzothiazol-3-yl]methanol), was an example of scaffold hopping (Figure 5C). This structure shared the max-MCS of the imidazothiazole moiety with compound **24** (2-methyl-6-[6-(6-methyl-pyridin-2-yl)-imidazo[2,1-b]thiazol-5-yl]-3a,7a-dihydro-benzooxazole), a known inhibitor of transforming growth factor-β receptor type 1 tyrosine kinase, unlike CDK2 serine/threonine kinase (Supplementary Table S10B).

To assess prospective prediction performance of CGBVS versus LBVS and SBVS, we have performed additional experimental validation of the prediction results from LBVS and SBVS for both EGFR and CDK2. Along with the validation protocol for CGBVS, the off-chip mobility shift assays confirmed the bioactivities of the 20 highest-scoring Bionet compounds that were selected by LBVS and SBVS. Consequently, inhibitors were identified by LBVS for neither EGFR nor CDK2 (Supplementary Figure S10A and B, and Supplementary Table S14). SBVS identified one EGFR inhibitor ($IC_{50}$=0.73 μM) and one CDK2 inhibitor ($IC_{50}$=26 μM), but neither exhibited scaffold hopping (Supplementary Figure S10C–G and Supplementary Table S15). The hit rate of SBVS was 5% (1 hit out of 20 at 10 μM), consistent with the hit rate of SBVS previously reported (Shoichet, 2004). As CGBVS exhibited 25 and 10% hit rates for EGFR and CDK2, respectively, the prediction performance of CGBVS was superior to those of existing methods (LBVS and SBVS) for the kinase family as well. These results indicate that CGBVS not only achieves higher hit rates but also predicts ligands with scaffolds different from known ligands in the case of protein kinases, suggesting that CGBVS is applicable to the identification of novel bioactive compounds for multiple protein families.

## Discussion

Whereas a critical first step in drug development is the identification of compounds with novel scaffolds, the next crucial step is assessment of selectivity. Information regarding the novelty and selectivity of lead candidates obtained by virtual screening can accelerate the subsequent lead-optimization stage of drug development. A paradigmatic advantage of CGBVS is the incorporation of multiple CPIs, numerically

represented as vector descriptors, which integrates both chemical structures and protein sequence data. In contrast, LBVS uses only chemical descriptors in the feature vector. This difference appears to provide CGBVS with a relatively high ability to predict ligand binding to multiple proteins (a measure of selectivity), while allowing scaffold hopping through the use of CPIs. In fact, we observed a concomitant loss of predictive performance when the number of elements for protein vectors was reduced (Supplementary Figure S11). The difference in the selectivity predictions of CGBVS and LBVS can be explained by the absence of vector descriptors for proteins in LBVS and the related lack of CPI data reflecting ligand recognition. Machine learning with protein data sets also enabled CGBVS to identify compounds that exhibited scaffold hopping because CPIs could be subdivided into chemical substructures and amino acid interactions, on which SBVS relies. In CGBVS, CPIs were described as amino acid versus chemical structure-derived feature vectors. CGBVS predictions are based on extraction of conserved patterns from subdivided interaction vectors involving both proteins and their corresponding ligands. Our successful identification of novel, scaffold-hopping ligands indicates that these conserved patterns included as yet undetermined signatures in the multiple CPIs captured.

Recently, computational approaches conceptually similar to our CGBVS approach have been proposed (Faulon *et al*, 2008; Jacob and Vert, 2008; Wassermann *et al*, 2009). However, these studies were limited in scope by the fact that they focused on computational validation through retrospective prediction and lacked experimental verification of the concept. Therefore, the practical utility and general applicability of these methods, specifically aimed at novel lead identification, are questionable.

Our present study has demonstrated that chemical genomics data are of immense practical use for lead discovery. Importantly, in further comparative analyses of the virtual screening of 11 500 Bionet compounds, the novel compounds that we identified using CGBVS were not in the high-scoring range using LBVS or SBVS (Supplementary Figure S12). Combining CGBVS with conventional methods, such as SBVS and LBVS, can significantly enhance the power of *in silico* strategies.

In the present form, as a learning machine, we used a SVM, which models the two class patterns of interacting pairs and non-interacting pairs by using proteins and their ligands. Therefore, the quality of these two types of training data has much effect on the prediction performance. In this study, we used manually curated protein–ligand interaction data sets from the GLIDA and GVK Biosciences databases. However, even curated data sets are likely to contain some factual errors, which tend to reduce the effectiveness of machine-learning methods. Therefore, improvement in the quality and quantity of the training data resource could enhance the prediction accuracy. A frequent hurdle to overcome when using CPI data is the acquisition of reliable data representing non-interacting pairs of ligands and their targets. Our strategy was to generate the same quantity of negative data from unknown interactions as that available for known interactions. The potential drawback is the possible introduction of a small number of false-negative examples in which the ligand does in fact bind to the target. The publication of experimentally confirmed non-interactions would benefit the CGBVS strategy greatly.

Moreover, it is not easy to comprehensively retrieve enough reliable activity information (that is, $IC_{50}$, $EC_{50}$, $K_i$ value, etc.) about ligands because our available CPI databases consist of heterogeneous experimental results from many researchers who screened different compound sets for their targets of interest using their original bioassay systems. If we could obtain sufficient and non-biased quantitative affinity data and generate a regressor model such as support vector regression, the prediction performance might be further improved.

Although the traditional drug design process focused on designing a single ligand specifically for a single receptor molecule, our results suggest that a systems biology-based 'integrationist mindset' (Peterson, 2008) is more appropriate for understanding and computing complex systems in some or all of their entirety. Although the integrationist view is a relatively recent approach that drug discovery research has not embraced completely, the view is beginning to receive attention for such research. Recently, drugs that target multiple proteins have been attracting interest for the development of novel effective therapeutics (Roth *et al*, 2004a; Fliri *et al*, 2005; Morphy and Rankovic, 2007; Apsel *et al*, 2008). As a predictive model, CGBVS could provide an important step in the discovery of such multi-target drugs by identifying the group of proteins targeted by a particular ligand, leading to innovation in pharmaceutical researches.

## Materials and methods

### CPI data

Data for 5207 ligand–GPCR pairs (including 317 GPCRs and their 866 ligands) with known CPIs were collected from the GLIDA database (Okuno *et al*, 2006), and 15 616 inhibitor–kinase pairs (including 143 kinases and their 8830 inhibitors) were collected from the GVK Biosciences kinase inhibitor database. The GLIDA database was constructed from several reliable resources, including IUPHAR-RD (Foord *et al*, 2005), PubMed (http://www.ncbi.nlm.nih.gov/pubmed/), PubChem (Wang *et al*, 2009), DrugBank (Wishart *et al*, 2006), the Ki Database (Roth *et al*, 2004b), and MDL ISIS/Base 2.5. Then, we carefully checked each compound against the primary literature to ensure that the chemical structure, target protein name, and binding and activity information were correct. Although the number of GPCR ligands was relatively small, the CPI pairs represent a credible data set because only interactions with relatively high affinities ($K_i$, $EC_{50}$, and $IC_{50} < 1\ \mu M$) are deposited in the GLIDA database.

### Chemical and protein descriptors

Chemical descriptors were calculated using the DRAGONX program (version 1.2; Talete S.r.l., Milan, Italy). Protein descriptors were calculated from the sequences alone. Specifically, dipeptide composition-based description (a mismatch-allowed spectrum kernel) was used to represent GPCRs, providing 400 dimensions (Leslie *et al*, 2004). For kinases, we used descriptors consisting of 1497 features provided by the PROFEAT Webserver (Li *et al*, 2006). Calculations of these descriptors were applied to the kinase domain, not to full-length sequences. Finally, these descriptor vectors were separately scaled to the range −1 to 1.

### SVM calculations

SVM calculations for CGBVS used a portion of the LIBSVM suite of programs (http://www.csie.ntu.edu.tw/~cjlin/libsvm). The

parameters of the SVM with the radial basis function kernel were optimized using a grid search.

### Retrospective virtual screening for ADRB2 by CGBVS

A predictive model for ADRB2 was constructed using a 5207-CPI pair data set, but excluding 40 ADRB2-related CPIs, leaving 5167 pairs. The number of predicted compounds was 866 (including the 40 known ligands). We then combined the chemical descriptors for these compounds and the protein descriptors for ADRB2, and predicted the probability of interactions. The ligand prediction was repeated 20 times with different negative sample sets, and the prediction score was set to the maximum probability.

### Retrospective virtual screening for ADRB2 by SBVS

We used the recently published crystal structure of ADRB2 (Cherezov *et al*, 2007) as a starting model. The ADRB2 structure and the 866 known ligands were prepared for docking simulations using the Protein Preparation Wizard and LigPrep script within Maestro (Schrödinger Inc, Portland, OR), respectively. This protein preparation procedure involved optimizing contacts by changing hydroxyl group orientations, flipping Asn and Gln side chains, and selecting His tautomeric states, followed by refining energy constraints using the OPLS-AA force field. Glide (SP mode) (Friesner *et al*, 2004) was used for grid generation and rigid receptor docking of the ligands. During the simulations, five docking models for each ligand were predicted, and the model with the minimum GlideScore was chosen as the final docking structure. SBVS was performed under two conditions: (1) without constraints, and (2) with constrained hydrogen bonding between the compounds and Asp113, a residue previously shown to be crucial for ligand binding (Strader *et al*, 1987). The different screening approaches were evaluated in terms of the hit rate and the EF (Supplementary Table S3) using the following equations:

$$\text{Hit rate} = 100\times(\text{Hits}_{\text{sampled}}/N_{\text{sampled}}),$$

where $N_{\text{sampled}}$ represents the total number of high-scoring compounds and $\text{Hits}_{\text{sampled}}$ represents the number of active compounds, and

$$\text{EF} = (\text{Hits}_{\text{sampled}}/N_{\text{sampled}})/(\text{Hits}_{\text{total}}/N_{\text{total}})$$

where $N_{\text{total}}$ represents the total number of compounds in the complete database and $\text{Hits}_{\text{total}}$ represents the number of active compounds therein. These values were calculated based on the assumption that all compounds reported to interact with ADRB2 were truly active compounds and that those with unknown activity for the target were inactive.

### Retrospective virtual screening for kinases

EGFR and CDK2 were chosen for model validation. In total, 15 616 kinase–inhibitor pairs were used to construct a CGBVS model. First, validation data sets from the DUD website (http://dud.docking.org/) were used (Huang *et al*, 2006). Compounds duplicated in the training and test data sets were removed from the test data. For comparison, binding free energy data, calculated by DOCK (Makino and Kuntz, 1997), was downloaded from the DUD site. Grid generation and rigid receptor ligand docking was performed using another SBVS method, GOLD (Jones *et al*, 1997). During simulations, three docking models for each ligand were predicted, and the model with the minimum ChemScore (or Astex Statistical Potential) was chosen as a final docking structure. Prospective predictions were generated for EGFR and CDK2 using 11 500 Bionet compounds (Key Organics Ltd), and the 20 highest-scoring compounds were selected for experimental verification (See Supplementary Table S7 for compound ID, names, and scores, and Supplementary Data Set 1 for chemical structures of the Bionet compounds).

### Polypharmacological prediction and prospective virtual screening by CGBVS

A prediction model for ADRB2 was constructed for CGBVS as described for retrospective screening, with the exception that all interaction data were included in the training data set. A prediction model for NPY1R was constructed using the 5207 CPI samples and an additional 3106 CPI samples of peptidergic GPCRs from the Integrity database (Prous Science S.A., Tokyo, Japan). Similarly, a prediction model for kinases was constructed using 15 616 CPI samples from the GVK Biosciences kinase inhibitor database. In each case, 11 500 compounds from the Bionet compound set were screened by CGBVS.

### MCS identification for active compounds

MCSs for each active compound and every GPCR ligand (or kinase inhibitor) in the training data set were calculated using the LibMCS program in the JChem module (Csizmadia, 2000). A single known ligand with the highest MCS value (max-MCS) was selected as the most relevant structure for each active compound. For comparison, we also selected a compound with the max-MCS compared with known ligands of the test receptor from the training data. Scaffold hopping was defined as the absence of overlaps between these max-MCSs.

### Experiments and the other calculations

A detailed description of the applied experimental and other computational techniques is given in the Supplementary information.

### Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

## Acknowledgements

This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology, Japan; the Japan Society for the Promotion of Science (JSPS); and the Targeted Proteins Research Program. This work was also supported, in part, by the New Energy and Industrial Technology Development Organization (NEDO) and the National Institute of Information and Communications Technology (NICT) of Japan. Financial support from Ono Pharmaceutical Co Ltd, the National Institute of Biomedical Innovation, The Suntory Institute for Bioorganic Research, and the Takeda Science Foundation is also gratefully acknowledged.

*Author contributions*: YO conceived this study. HY and SN performed *in silico* screening. HT, TI, TH, and GT carried out *in vitro* binding assays. TH performed docking simulations. SN, HY, TO, YM, and YO analyzed the data. HY, SN, and YO wrote the manuscript.

## Conflict of interest

The authors declare that they have no conflict of interest.

## References

Apsel B, Blair JA, Gonzalez B, Nazif TM, Feldman ME, Aizenstein B, Hoffman R, Williams RL, Shokat KM, Knight ZA (2008) Targeted polypharmacology: discovery of dual inhibitors of tyrosine and phosphoinositide kinases. *Nat Chem Biol* **4:** 691–699

Cherezov V, Rosenbaum DM, Hanson MA, Rasmussen SG, Thian FS, Kobilka TS, Choi HJ, Kuhn P, Weis WI, Kobilka BK, Stevens RC (2007) High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science* **318:** 1258–1265

Csizmadia F (2000) JChem: Java applets and modules supporting chemical database handling from web browsers. *J Chem Inf Model* **40:** 323–324

Dobson CM (2004) Chemical space and biology. *Nature* **432:** 824–828

Eckert H, Bajorath J (2007) Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov Today* **12:** 225–233

Faulon JL, Misra M, Martin S, Sale K, Sapra R (2008) Genome scale enzyme-metabolite and drug-target interaction predictions using the signature molecular descriptor. *Bioinformatics* **24:** 225–233

Fliri AF, Loging WT, Thadeio PF, Volkmann RA (2005) Analysis of drug-induced effect patterns to link structure and side effects of medicines. *Nat Chem Biol* **1:** 389–397

Foord SM, Bonner TI, Neubig RR, Rosser EM, Pin JP, Davenport AP, Spedding M, Harmar AJ (2005) International Union of Pharmacology. XLVI. G protein-coupled receptor list. *Pharmacol Rev* **57:** 279–288

Fredriksson R, Lagerström MC, Lundin LG, Schiöth HB (2003) The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol Pharmacol* **63:** 1256–1272

Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem* **47:** 1739–1749

Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143:** 29–36

Hopkins AL (2008) Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol* **4:** 682–690

Hopkins AL, Groom CR (2002) The druggable genome. *Nat Rev Drug Discov* **1:** 727–730

Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking sets for molecular docking. *J Med Chem* **49:** 6789–6801

Jacob L, Vert JP (2008) Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics* **24:** 2149–2156

Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* **267:** 727–748

Karaman MW, Herrgard S, Treiber DK, Gallant P, Atteridge CE, Campbell BT, Chan KW, Ciceri P, Davis MI, Edeen PT, Faraoni R, Floyd M, Hunt JP, Lockhart DJ, Milanov ZV, Morrison MJ, Pallares G, Patel HK, Pritchard S, Wodicka LM *et al* (2008) A quantitative analysis of kinase inhibitor selectivity. *Nat Biotechnol* **26:** 127–132

Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK (2007) Relating protein pharmacology by ligand chemistry. *Nat Biotechnol* **25:** 197–206

Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH, Kuijer MB, Matos RC, Tran TB, Whaley R, Glennon RA, Hert J, Thomas KL, Edwards DD, Shoichet BK, Roth BL (2009) Predicting new molecular targets for known drugs. *Nature* **462:** 175–181

Lehár J, Stockwell BR, Giaever G, Nislow C (2008) Combination chemical genetics. *Nat Chem Biol* **4:** 674–681

Leslie CS, Eskin E, Cohen A, Weston J, Noble WS (2004) Mismatch string kernels for discriminative protein classification. *Bioinformatics* **20:** 467–476

Li ZR, Lin HH, Han LY, Jiang L, Chen X, Chen YZ (2006) PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res* **34:** W32–W37

Lipinski C, Hopkins A (2004) Navigating chemical space for biology and medicine. *Nature* **432:** 855–861

MacDonald ML, Lamerdin J, Owens S, Keon BH, Bilter GK, Shang Z, Huang Z, Yu H, Dias J, Minami T, Michnick SW, Westwick JK (2006) Identifying off-target effects and hidden phenotypes of drugs in human cells. *Nat Chem Biol* **2:** 329–337

Makino S, Kuntz ID (1997) Automated flexible ligand docking method and its application for database search. *J Comput Chem* **18:** 1812–1825

Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002) The protein kinase complement of the human genome. *Science* **298:** 1912–1934

McInnes C (2007) Virtual screening strategies in drug discovery. *Curr Opin Chem Biol* **11:** 494–502

Morphy R, Rankovic Z (2007) Fragments, network biology and designing multiple ligands. *Drug Discov Today* **12:** 156–160

Muegge I, Oloff S (2006) Advances in virtual screening. *Drug Discov Today Technol* **3:** 405–411

Okuno Y, Yang J, Taneishi K, Yabuuchi H, Tsujimoto G (2006) GLIDA: GPCR–ligand database for chemical genomic drug discovery. *Nucleic Acids Res* **34:** D673–D677

Oprea TI, Matter H (2004) Integrating virtual screening in lead discovery. *Curr Opin Chem Biol* **8:** 349–358

Oprea TI, Tropsha A, Faulon JL, Rintoul MD (2007) Systems chemical biology. *Nat Chem Biol* **3:** 447–450

Paolini GV, Shapland RH, van Hoorn WP, Mason JS, Hopkins AL (2006) Global mapping of pharmacological space. *Nat Biotechnol* **24:** 805–815

Peterson R (2008) Chemical biology and limits of reductionism. *Nat Chem Biol* **4:** 635–638

Rasmussen SG, Choi HJ, Rosenbaum DM, Kobilka TS, Thian FS, Edwards PC, Burghammer M, Ratnala VR, Sanishvili R, Fischetti RF, Schertler GF, Weis WI, Kobilka BK (2007) Crystal structure of the human beta2 adrenergic G-protein-coupled receptor. *Nature* **450:** 383–387

Renner S, van Otterlo WA, Dominguez Seoane M, Möcklinghoff S, Hofmann B, Wetzel S, Schuffenhauer A, Ertl P, Oprea TI, Steinhilber D, Brunsveld L, Rauh D, Waldmann H (2009) Bioactivity-guided mapping and navigation of chemical space. *Nat Chem Biol* **5:** 585–592

Roth BL, Lopez E, Beischel S, Westkaemper RB, Evans JM (2004b) Screening the receptorome to discover the molecular targets for plant-derived psychoactive compounds: a novel approach for CNS drug discovery. *Pharmacol Ther* **102:** 99–110

Roth BL, Sheffler DJ, Kroeze WK (2004a) Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat Rev Drug Discov* **3:** 353–359

Schölkopf B, Tsuda K, Vert JP (2004) *Kernel Methods in Computational Biology*. Cambridge, Massachusetts, USA: MIT Press

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

Shoichet BK (2004) Virtual screening of chemical libraries. *Nature* **432:** 862–865

Strader CD, Sigal IS, Register RB, Candelore MR, Rands E, Dixon RA (1987) Identification of residues required for ligand binding to the beta-adrenergic receptor. *Proc Natl Acad Sci USA* **84:** 4384–4388

Vapnik VN (1995) *The Nature of Statistical Learning Theory*. New York, USA: Springer

Wagner AB (2006) SciFinder Scholar 2006: an empirical analysis of research topic query processing. *J Chem Inf Model* **46:** 767–774

Waldeck B (2002) Beta-adrenoceptor agonists and asthma—100 years of development. *Eur J Pharmacol* **445:** 1–12

Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res* **37:** W623–W633

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

Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* **34:** D668–D672

Young DW, Bender A, Hoyt J, McWhinnie E, Chirn GW, Tao CY, Tallarico JA, Labow M, Jenkins JL, Mitchison TJ, Feng Y (2008) Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nat Chem Biol* **4:** 59–68

Theor Chem Acc (2011) 130:1131–1143
DOI 10.1007/s00214-011-1034-5

REGULAR ARTICLE

# Computational analysis of ligand recognition mechanisms by prostaglandin $E_2$ (subtype 2) and $D_2$ receptors

**Hiromi Daiyasu · Takatsugu Hirokawa · Narutoshi Kamiya · Hiroyuki Toh**

Received: 13 April 2011 / Accepted: 29 August 2011 / Published online: 20 September 2011


**Abstract** The eight members of the prostanoid receptor family belong to the class A G protein-coupled receptors. We investigated the evolutionary relationship of the eight members by a molecular phylogenetic analysis and found that prostaglandin $E_2$ receptor subtype 2 ($EP_2$) and prostaglandin $D_2$ receptor (DP) were closely related. The structures of the ligands for the two receptors are similar to each other but are distinguished by the exchanged locations of the carbonyl oxygen and the hydroxy group in the cyclopentane ring. The ligand recognition mechanisms of the receptors were examined by an integrated approach using several computational methods, such as amino acid sequence comparison, homology modeling, docking simulation, and molecular dynamics simulation. The results revealed the similar location of the ligand between the two receptors. The common carboxy group of the ligands interacts with the Arg residue on the seventh transmembrane (TM) helix, which is invariant among the prostanoid receptors. $EP_2$ uses a Ser on TM1 to recognize the carbonyl oxygen in the cyclopentane ring of the ligand. The Ser is specifically conserved within $EP_2$. On the other hand, DP uses a Lys on TM2 to recognize the hydroxy group of the $\omega$ chain of the ligand. The Lys is also specifically conserved within DP. The interaction network between the D(E)RY motif and TM6 was found in $EP_2$. However, DP lacked this network, due to the mutation in the D(E)RY motif. Based on these observations and the previously published mutational studies on the motif, the possibility of another activation mechanism that does not involve the interaction between the D(E)RY motif and TM6 will be discussed.

**Keywords** Prostanoid receptor · Prostaglandin $E_2$ · Prostaglandin $D_2$ · Ligand specificity · Molecular evolution · Molecular dynamics simulation

Dedicated to Professor Akira Imamura on the occasion of his 77th birthday and published as part of the Imamura Festschrift Issue.

**Electronic supplementary material** The online version of this article (doi:10.1007/s00214-011-1034-5) contains supplementary material, which is available to authorized users.

H. Daiyasu (✉)
The Center for Medical Engineering and Informatics, Osaka University, 1-5, Yamadaoka, Suita Osaka 565-0871, Japan
e-mail: daiyasu@ist.osaka-u.ac.jp

T. Hirokawa · H. Toh
Computational Biology Research Center, Advanced Industrial Science and Technology, 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan
e-mail: t-hirokawa@aist.go.jp

H. Toh
e-mail: toh-hiroyuki@aist.go.jp

N. Kamiya
Institute for Protein Research, Osaka University, 6-2-3, Furuedai, Suita Osaka 565-0874, Japan

H. Toh
Medical Institute of Bioregulation, Kyushu University, 3-1-1 Maidashi, Higashi-ku, Fukuoka 812-8582, Japan

## 1 Introduction

Stimulation of cells with proinflammatory factors activates phospholipase $A_2$, which releases arachidonic acid from the membrane phospholipids to the cytoplasm [1]. From arachidonic acid, a wide variety of bioactive compounds, such as prostaglandins (PGs), thromboxanes (TXs), leukotrienes, and lipoxins, are generated by several enzymes. The first step in the generation of PGs and TXs, which are collectively termed prostanoids, is catalyzed by

cyclooxygenase. Therefore, the prostanoid generation pathway is called the "cyclooxygenase pathway" (see Fig. 1). The prostanoids, which include $PGD_2$, $PGE_2$, $PGF_{2\alpha}$, $PGI_2$ (prostacyclin), and $TXA_2$, are involved in essential biological functions, such as inflammation, blood clotting, and sleep induction [1]. Most prostanoids function as autacoids and exert their physiological functions through binding to their own receptors on the cells. The prostanoid synthases and receptors are being extensively investigated as targets for medicines and therapeutics [2–6].

The prostanoid receptors (PNRs) belong to the class A G protein-coupled receptor (GPCR) family [6]. GPCRs are membrane proteins characterized by seven transmembrane helices (TM1–7), which are connected by three intracellular loops (ICL1–3) and three extracellular loops (ECL1–3). The alteration of the relative locations of the helices caused by ligand binding is considered to induce signal transduction. There are four $PGE_2$ receptor subtypes, which are referred to as $EP_1$–$EP_4$. In contrast, $PGD_2$, $PGF_{2\alpha}$, $PGI_2$, and $TXA_2$ uniquely correspond to their receptors, DP, FP, IP, and TP, respectively. In addition, a chemoattractant receptor-homologous molecule expressed on Th2 cells (CRTH2) was recently reported to function as a $PGD_2$ receptor [7]. Previous studies suggested that DP, $EP_1$–$EP_4$, FP, IP, and TP are closely related and form a subfamily in the class A GPCRs. However, CRTH2 is distantly related to the other PNRs, suggesting that CRTH2 may have evolved independently from the other PNRs [8].

The ligand recognition mechanism by the receptors is an important and interesting subject and is presently being analyzed experimentally and computationally. For example, the residues involved in the ligand recognition by PNRs have been experimentally investigated by amino acid substitution studies [9–13]. In addition, computational methods, such as molecular dynamics simulation [14] and homology modeling [11], have also been applied to the analysis of the ligand recognition. In this report, we describe a synergistic approach integrating several computational methods for the analysis of the ligand recognition specificity of the PNRs. At first the targets of this study were selected by a molecular phylogenetic analysis. In the phylogenetic tree $EP_2$ and DP are closely related to each other. In addition, the structures of their ligands, $PGE_2$ and $PGD_2$, are similar (see Fig. 1). The only difference between $PGE_2$ and $PGD_2$ is that the carbonyl oxygen and the hydroxy group are exchanged at positions 9 and 11 of the cyclopentane ring. Therefore, $EP_2$ and DP were selected as the targets in this analysis. Although amino acid sequence data for the receptors are abundantly available, no PNR structure has been determined yet. However, coordinate data for several class A GPCRs are available. We built the model structures of $EP_2$ and DP by homology modeling, using the structure of squid rhodopsin as a template. Sequence comparisons and model structure studies indicated the candidate residues that may be involved in the ligand specificity. Finally, molecular dynamics (MD) simulations of the ligand-bound models in

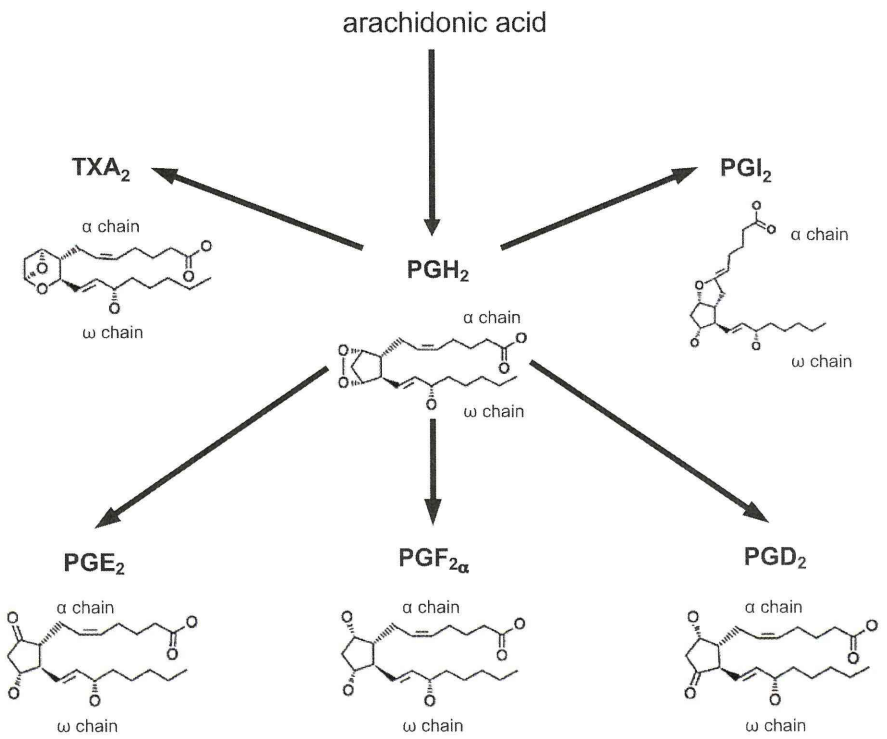


**Fig. 1** Cyclooxygenase pathway

explicit lipid and water were performed. Based on the computational results, the ligand specificities of $EP_2$ and DP will be discussed.

We will also discuss the D(E)RY motif of the PNRs. The D(E)RY motif is present at the cytosolic end of TM3, which regulates the receptor activation and the interaction with G proteins [15]. The first acidic residue, Asp or Glu, of the motif is involved in receptor activation and in coupling agonist binding to the activation of G protein signaling [16–18]. The second residue (Arg) of the motif is highly conserved and is essential for the intramolecular interactions that constrain receptors in either the inactive or activated conformation [19]. In contrast, the third residue of the motif is not highly conserved, although the Tyr residue participates in intramolecular contacts in some GPCRs that are important for stabilizing the receptor protein in a functional conformation [20]. The intramolecular interaction between the D(E)RY motif and TM6 was observed in $EP_2$, but not in DP, due to the mutation in the D(E)RY motif. Based on these observations, we will discuss the possibility that the activation mechanism through the D(E)RY motif has diverged during the course of the molecular evolution of PNRs.

## 2 Methods

### 2.1 Sequence data and representation of amino acids

The amino acid sequences of PNR homologues were collected by searching the non-redundant protein sequence database at the NCBI site (http://www.ncbi.nlm.nih.gov/BLAST/) with BLAST version 2.2.23 [21], using the human DP sequence as a query. Searches were also performed against several genome databases. Among them, one PNR homologue was detected from the genome database of *Lottia gigantea* (http://genome.jgi-psf.org/Lotgi1/Logti1.home.html). Since most mammals have all eight PNR members, the human, mouse, dog, and bovine homologues were selected as the representative mammalian PNRs in this study.

The one-letter and three-letter representations of amino acids used in this manuscript are as follows: A, Ala—alanine; C, Cys—cysteine; D, Asp—aspartic acid; E, Glu—glutamic acid; F, Phe—phenylalanine; G, Gly—glycine; H, His—histidine; I, Ile—isoleucine; K, Lys—lysine; L, Leu—leucine; M, Met—methionine; N, Asn—asparagine; P, Pro—proline; Q, Gln—glutamine; R, Arg—arginine; S, Ser—serine; T, Thr—threonine; V, Val—valine; W, Trp—tryptophan; and Y, Tyr—tyrosine.

The residues in the transmembrane region are numbered according to the Ballesteros–Weinstein nomenclature in superscript [22]. The first digit indicates the number of the TM helix on which the residue is located, and the second digit is the position counted from the most conserved site in each TM, to which number 50 is assigned. As for the residue in the loop region, the abbreviated name of the loop is shown in parentheses.

### 2.2 Phylogenetic analysis

A multiple amino acid sequence alignment was performed with the alignment software MAFFT version 6.808a [23, 24]. Based on the alignment, an unrooted molecular phylogenetic tree was constructed by the neighbor-joining (NJ) method [25]. The genetic distance between every pair of aligned sequences was calculated as a maximum likelihood estimate [26] under the JTT model [27] for the amino acid substitutions. The sites including gaps in the alignment were excluded from the calculation. The statistical significance of the NJ tree topology was evaluated by a bootstrap analysis [28]. The procedure for the bootstrap analysis is as follows: (1) Prepare a register for each node of the original tree, which is set to 0. (2) Sample the alignment sites randomly without replacement from the alignment used for the construction of the original tree. The number of collected sites should be the same as that of the original alignment. (3) Construct a phylogenetic tree with the collected alignment sites. (4) Examine each node of the tree. If the set of proteins under the node is the same as that of a node of the original tree, then one is added to the register of the node. (5) Repeat steps (3)–(4) by a given number of times. In this study, 1,000 iterations were done for the bootstrap analysis. (6) Divide the number in each register by the given number, which provides the bootstrap probability of the corresponding node. Two software packages, PHYLIP 3.5c [29] and MOLPHY 2.3b3 [30], were used for the phylogenetic analyses.

### 2.3 Construction of the model structures of $EP_2$ and DP

The amino acid sequences of proteins with available coordinates were collected by a BLAST search through the Protein Data Bank (PDB). The amino acid sequence of DP was used as a query for this search. The database search yielded squid rhodopsin, bovine rhodopsin, turkey $\beta_1$ adrenergic receptor, human $\beta_2$ adrenergic receptor, and human $A_{2A}$ adenosine receptor. The squid rhodopsin showed the lowest *E*-value in the output list (see Supplement 1). The sequence identities of the query to the five structures were only about 20%. Among them, the sequence identity between the query and the squid rhodopsin was ranked third. However, the sequence positive percent between the query and the squid rhodopsin was ranked second. Amino acid substitutions and/or a domain fusion were introduced into turkey $\beta_1$ adrenergic receptor and human $A_{2A}$ adenosine receptor, to determine the

structures. There are many unidentified regions in human $\beta_2$ adrenergic receptor due to the low-resolution analysis. However, there are no such artificial modifications in the squid and bovine rhodopsins. Considering the results of the database search, the situation of the artificial modification, and the resolution of the structure, the coordinates of the squid rhodopsin structure (pdb ID: 2Z73) were used as the template to build the model structures of $EP_2$ and DP.

A multiple alignment of DP, $EP_2$, squid rhodopsin, and the remaining four structures was created by MAFFT [23, 24]. The alignment was modified by visual inspection, so the gaps were not aligned to the secondary structure regions. In the operation, the structural information of the four structures, as well as that of the squid rhodopsin, was also considered. In general, the N- and C-terminal regions of GPCRs are divergent from each other in both length and residue composition. Therefore, the model structure was built from Q9 to L330 for $EP_2$ and from P4 to I338 for DP, respectively.

The model structures for $EP_2$ and DP were built by homology modeling with MOE (Molecular Operating Environment, version 2008.1002, Chemical Computing Group, Montreal, Canada). Energy minimization was performed with the MMFF94x force field [31], under an implicit solvent model using the generalized Born/volume integral (GB/VI) model [32]. In the model building procedure, hydrogen atoms were added to the template structure, partial charges were assigned to all of the atoms of the structure, and the dielectric constant was set to 1.0. Energy minimization was performed by the successive application of three methods, the steepest descent, the conjugate gradient, and the truncated Newton. The root mean square deviation (RMSD) gradient was set to 1,000 during the steepest descent stage and 100 in the conjugate gradient stage and was finally fixed at 0.001 in the truncated Newton stage. Then, 25 intermediate model structures were generated by the Boltzmann-weighted randomized modeling procedure, and minimizations were performed until an RMS gradient of 1 kcal/mol/Å was attained. Among them, the final model was selected based on the GB/VI scoring, which was subsequently minimized until an RMS gradient of 0.5 kcal/mol/Å was attained.

The model structures were validated by a Ramachandran plot and ProSA. The stereochemical qualities of the models were evaluated using a Ramachandran plot in MOE. The ProSA tool was used to check the overall and local structures for potential errors [33, 34]. As the baseline, validation data for the template structure were also assessed.

### 2.4 Construction of the initial ligand-bound model structures of $EP_2$ and DP

The coordinates of $PGD_2$ were obtained from the PDB (1RY0), while the 3-D structure of $PGE_2$ was constructed by modifying $PGD_2$ with MOE. Plural candidate docking models for the receptor and the ligand were generated by MOE ASEDock [35]. The ASEDock docking was performed in this study as follows: (1) generation of 1,000 ligand conformations by the stochastic conformation search, (2) concavity search of a target protein, (3) characterization of the surface and the surrounding region of the detected concavity to generate a model for ligand binding at the selected concavity, which is called an ASE model, (4) rigid body alignment of 3,000 poses for each ligand conformation to the ASE model and selection of 200 poses with the best scores to evaluate the fitness of the poses to the ASE model, and (5) stepwise energy minimization of the 200 posed conformations of the ligand in the concavity. In step (5), the ligand was free to move and the backbone atoms of the receptor were constrained by using the tether weights of the default values throughout all of the minimization procedures. In the first rough minimization, a cutoff distance of 4.5 Å and a convergence criterion for the RMS gradient of 10 kcal/mol/Å were used. The interaction energies were then calculated with a cutoff distance to 8 Å, and ten structures with the minimum interaction energies were selected. Likewise, ten structures with the maximum ASE scores were selected, although the structures that were also detected by the minimum interaction energy criterion were neglected. The structures thus selected were further refined by setting the cutoff distance to 10 Å and the RMS gradient to 0.1 kcal/mol/Å. The protein–ligand interaction energy $U_{\text{total}}$ is expressed as follows: $U_{\text{total}} = U_{\text{ele}} + U_{\text{vdw}} + U_{\text{solv}} + U_{\text{ligand}}$, in which $U_{\text{ele}}$, $U_{\text{vdw}}$, and $U_{\text{solv}}$ indicate the coulombic, van der Waals, and generalized Born/solvent accessible solvation interaction energies, respectively. $U_{\text{ligand}}$ indicates the internal conformation energy of the ligand. The obtained docking models for MD were classified based on the similarity in ligand conformation by using the cluster analysis function of ASEDock, although the models with the positive $U_{\text{total}}$ were neglected from the subsequent procedure. The members in which the carboxy group of the ligand was close to the conserved $Arg^{7.40}$ on TM7 were then collected from each cluster, since the residue is involved in the recognition of the carboxy group of the prostanoids [9–13]. Among them, the model with the lowest $U_{\text{total}}$ was selected as the representative of the cluster. Finally, the model structure with the lowest $U_{\text{total}}$ among the representatives of the clusters was selected as the initial ligand-bound model structure for the following MD simulations.

### 2.5 MD simulations

Molecular dynamics simulations of the ligand-bound forms, built as described above, were performed for both

$EP_2$ and DP using Desmond version 2.2 [36]. The OPLS2005 force field [37] was used for the simulations. Initial ligand-bound model structures were placed into a large equilibrated dipalmitoylphosphatidylcholine (DPPC at 325 K) bilayer and TIP3P water molecules solvated with 0.15 M NaCl. The resulting system for the ligand-bound $EP_2$ had 179 lipid molecules, 48 sodium ions, 48 chloride ions, and 17,396 water molecules, for a total of 80,754 atoms, and the boundary condition measured $\approx 89 \times 83 \times 118$ Å$^3$. The resulting system for the ligand-bound DP had 171 lipid molecules, 55 sodium ions, 55 chloride ions, and 19,618 water molecules, for a total of 86,582 atoms, and the boundary condition measured $\approx 90 \times 82 \times 127$ Å$^3$. All system setups were performed using Maestro (Schrödinger LLC, New York NY). No artificial constraint was introduced in the simulation. After minimization, heating, and equilibration, the production MD phase was performed for 10 ns in the isothermal-isobaric (NPT) ensemble at 310 K and 1 bar using the Berendsen coupling scheme. The covalent bonds involving hydrogen atoms were constrained by the M-SHAKE method [38]. The time step of the MD was 2 fs. The van der Waals and short-range electrostatic interactions were truncated with a 9-Å-cutoff. Long-range electrostatic interactions were computed using the Particle Mesh Ewald method [39].

The equilibrium of the MD simulation was evaluated by the RMSDs between the Cα atoms of the model structure at each time step of a run and those of the reference structure. Here, the first structure (0 ns) in the production MD phase was used as the reference structure. The total potential energy was also used to evaluate the equilibrium, which was calculated by summing up the energies of bond stretching, angle bending, and torsional potential, their cross terms, and the non-bonded interaction energies (van der Waals and coulomb) for all systems including a protein with ligand, lipids, ions, and water molecules. The fluctuation of each Cα atom of the model structures in the equilibrium state was evaluated with the root mean square fluctuation (RMSF). The trace in a time window of 10 ns in the equilibrium state was obtained from each run of the simulation, and then the average coordinates over the window were calculated. The squared deviation of each Cα atom between the model structure and the average coordinates was averaged over the 10 ns. The square root of the averaged value of an atom was the RMSF for the atom. The interaction energy for a ligand–receptor complex was calculated by the Prime MM-GBSA [40] in Maestro using 10 MD conformations, corresponding to each 0.1-ns step of the final 1-ns trajectories.

## 3 Results

### 3.1 Molecular evolution of PNRs

#### *3.1.1 Phylogenetic analysis of PNRs*

To examine the evolutionary relationships of PNRs, the sequences of PNR homologues and several class A GPCR proteins were collected and aligned (Supplement 2). A phylogenetic tree of these proteins is shown in Fig. 2. The tree is divided into two subtrees, which are rooted at nodes X and Y, respectively. One of them consists of four PNR clusters, corresponding to $EP_4$, IP, $EP_2$, and DP, whereas the remaining four PNR clusters, $EP_3$, TP, FP, and $EP_1$, constitute the other subtree. Both subtrees included different subtypes of $PGE_2$ receptors. The distribution of PNRs in the tree suggests that the current receptors originated from an ancestral receptor for $PGE_2$, consistent with previous reports [8, 41]. Each PNR cluster includes the putative vertebrate orthologues, which suggests that the genome of an ancestral organism of vertebrates already encoded the eight PNRs. The amphioxus (*Branchiostoma floridae*) and the ascidian (*Ciona intestinalis*), which are close relatives of vertebrates, had one or two copies of PNR homologues. The vertebrates, the amphioxus, and the ascidian constitute the deuterostome. PNR homologues were also found in the mollusk (*L. gigantea*), which belongs to the protostome. Currently, the genome data for several protostomes, such as insects and nematoda, are available. However, PNR homologues were not detected in the genome data of these organisms. On the other hand, a PNR homologue was also detected from the cnidarian (*Nematostella vectensis*). The cnidarian is considered to have branched out before the divergence of the protostome and the deuterostome. The vertebrate PNRs have a conserved $Arg^{7.40}$ on TM7, which is believed to be essential to interact with the carboxy group of prostanoids [9–13]. This residue does not exist in the other class A GPCRs. The $Arg^{7.40}$ residue was also conserved in the invertebrate PNR homologues, except for those from *C. intestinalis* (gi number: 198418121) and *B. floridae* (gi number: 260823840) (see Supplement 2). The tree topology, together with the conservation of the Arg residue, suggests the possibility that the date for the divergence between the two subtrees would trace back before the divergence between the protostome and the deuterostome. Further accumulation of genome data will reveal the cause of this bias in the distribution of PNR homologues.

One of the interesting points of the tree was the close evolutionary relationship between $EP_2$ and DP. The bootstrap probability for the clustering of the two receptors was 89.9%, and the sequence identity between the two

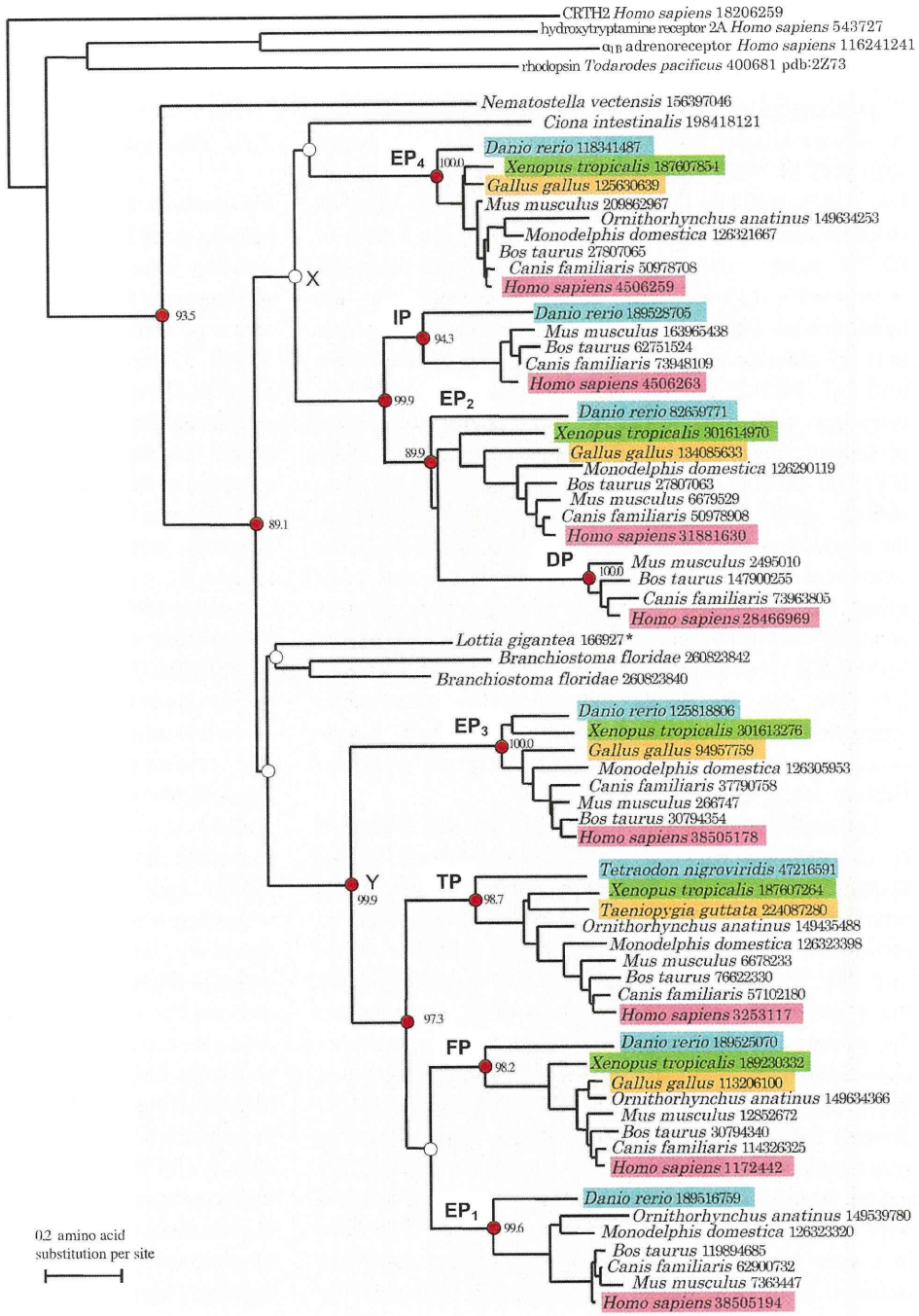


**Fig. 2** Phylogenetic tree of PNRs. PGR homologues are within the *gray rectangle*. The other class A GPCRs outside the *rectangle* are introduced as the outgroup. The source organism and the NCBI gi number for each sequence are shown at the terminal node. *Asterisk* indicates the sequence derived from *L. gigantea* genome data (see Sect. 2). The nodes that are considered to be critical for the evolution of the PNRs are indicated by *circles*. Critical nodes with bootstrap probabilities >80% are colored *red*, with the bootstrap probability shown near the node

receptors was about 40%. As described above, the structures of the ligands for the two receptors are quite similar to each other. The only difference is that the positions of the carbonyl oxygen and the hydroxy group are exchanged in the cyclopentane ring (Fig. 1). Thus, the structural similarity of the ligands seems to reflect the evolutionary closeness of the receptors. The medium sequence divergence of the receptors and the resemblance of the ligand structures made it challenging to investigate the relationship between the differences in the amino acid residues and the ligand recognition specificity of the receptors. In this study, we have analyzed the mechanism of the ligand recognition specificity by focusing on $EP_2$ and DP.

### 3.1.2 *Conservation pattern of amino acid residues in PNRs*

At first, we examined the residue conservation by constructing a multiple alignment of PNRs (Supplement 2).

We obtained 24 invariant sites among all of the PNR homologues, which are shown in the alignment of $EP_2$ and DP (Fig. 3). The invariant $Arg^{7.40}$ was also included in these sites. Some of the sites may also be involved in the recognition of the common structure of prostanoids, such as $Arg^{7.40}$. The conserved sites also included the NPxxY motif on TM7, although the motif was found as DPWxF in the PNRs. This motif contributes to the internalization and the signal transduction of class A GPCRs [42]. The cholesterol-binding motif on TM4 [43] was not found in the PNRs.

Next, we examined the amino acid conservation pattern between $EP_2$ and DP (Fig. 3). We assumed that the amino acid sites specifically conserved in $EP_2$ or DP are involved in the specific ligand recognition by the receptors. Thus, the sites that are invariant in $EP_2$, but are not occupied by the same residue in DP, were identified from the alignment at first. The DP-specific invariant sites were identified in the same manner. We identified 21 sites for $EP_2$ and 76 for DP. Hereafter, these residues will be referred to as "specifically conserved sites". The abundance of the specifically conserved residues in DP, in comparison with $EP_2$, was considered to reflect the difference in the evolutionary closeness of the sources, since the amino acid sequences of DP were only from mammals, whereas the sequences of $EP_2$ used in this study were collected from vertebrates. As described below, the candidate sites involved in the ligand recognition were further selected from these specifically conserved sites, considering the structural information of the model structures.

In addition to the specifically conserved sites, we identified 164 alignment sites that are occupied by physicochemically similar residues between $EP_2$ and DP. These sites will simply be called "conserved sites". Naturally, the 24 invariant sites within PNRs described above were included in the sites. Most of the 164 conserved sites were found in the seven TM helices, based on the model structures, as described below. Some of the conserved sites were found in ICL2 and ECL2. The observation suggests that the conservation may reflect the common structural and/or functional constraints on the receptors. Therefore, the residues corresponding to the conserved sites were also regarded as the candidate residues involved in the ligand recognition.

## 3.2 Molecular mechanisms of ligand recognition by $EP_2$ and DP

### 3.2.1 *Model construction of $EP_2$ and DP*

The ligand-free models for $EP_2$ and DP were constructed by homology modeling, as described in Sect. 2. To assess

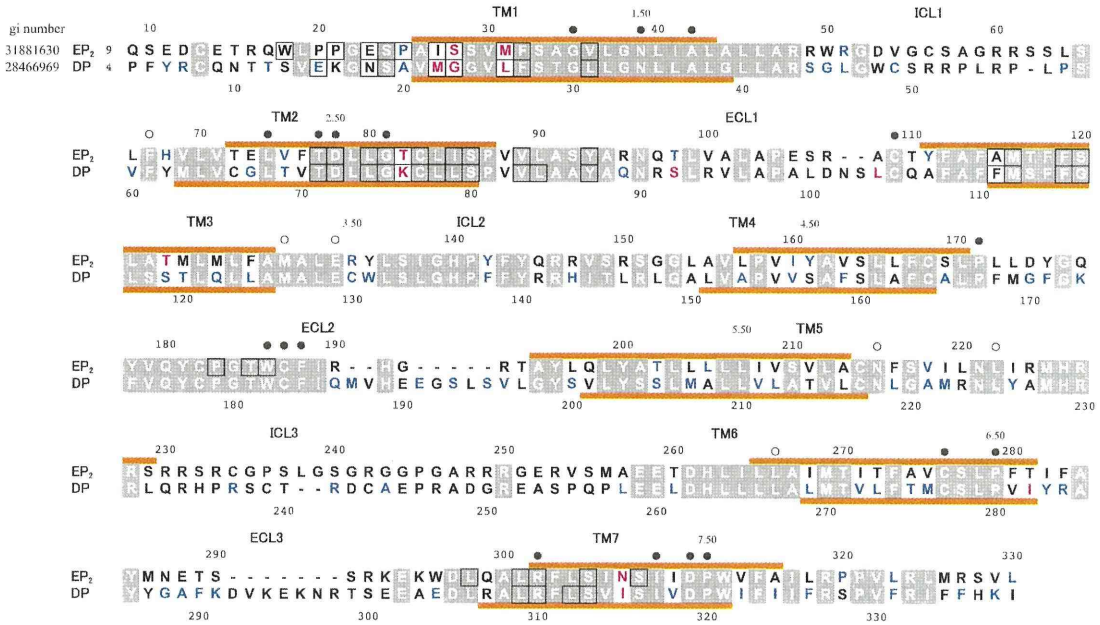


**Fig. 3** Amino acid sequence alignment of human $EP_2$ and DP. Among the sites specifically conserved in each subfamily, those located in the pocket surface regions detected by CASTp [44] are colored *red*, whereas the remaining sites are colored *blue*. The conserved sites between $EP_2$ and DP are represented by *white* characters with a *gray* background. The invariant sites among the PNR homologues are indicated by the *circles* over the alignment (*black*: located in the ligand-binding pocket, *white*: near the D(E)RY motif). The *orange bar* shows the TM regions predicted by TMDET [52]. For the prediction, the model structure with the lowest MM-GBSA energies was used. The residues located within 5 Å from the ligand were obtained from the selected model structures, which are *encircled* by *black squares* in the alignment

the quality of the structures, the generated models were examined by the Ramachandran plots and the ProSA program (Supplement 3). One residue for $EP_2$ and eight for DP were identified as outliers in their phi and psi backbone dihedral angles by the Ramachandran plot evaluation. All residues thus identified were located in loop regions and were far from the pocket and the D(E)RY motif. The ratios of the residues of the model structures in the core region were 93.4% for $EP_2$ and 87.7% for DP. The ProSA program evaluates a model structure by two methods. One of them is the overall model quality in comparison with the experimentally determined structural data in the PDB database, and the goodness of the model by this evaluation is given by the *Z*-score. The *Z*-score for the $EP_2$ model structure was −5.84 and that for the DP model structure was −6.13, whereas the *Z*-score of the template structure was −4.94 (see Supplement 3). In the other method, the local model quality is expressed as a function of the amino acid sequence position. In general, an amino acid segment with positive values indicates erroneous parts of the given structure. There were far fewer such regions in the $EP_2$ and DP model structures than in the template. Thus, both evaluations suggested that the model structures were built with good quality. Therefore, we used the model structures to build the models of the ligand-bound forms of $EP_2$ and DP.

The candidate residues involved in the ligand recognition were selected from the residues corresponding to the specifically conserved sites described above. At first, the cavity of each model structure was identified by CASTp [44], a program to identify the surface accessible pockets of a given tertiary structure. Then, the residues corresponding to the specifically conserved sites, which were located on the surface of the detected cavity, were selected as the candidates. Five residues were selected for $EP_2$, while eight specifically conserved residues were found in the cavity of DP. These residues are shown in Fig. 3.

ASEDock generated 16 ligand-bound models for $EP_2$ and 20 ligand-bound models for DP (see Supplement 4). Among the 16 $EP_2$ models, 10 models with negative $U_{total}$ were selected and classified into five groups by the cluster analysis. According to the procedure described in Sect. 2, the representative model of group 1 with $U_{total}$ of −35.5 kcal/mol was selected as an initial model structure for the MD simulation. The distance between the oxygen in the carboxy group of the ligand and the hydrogen in the guanidino group of $R302^{7.40}$ for $EP_2$ was 2.0 Å. Likewise, 18 ligand-bound models for DP with negative $U_{total}$ were selected from the 20 models and classified into seven groups. The selection procedure chose the representative model of group 2 with $U_{total}$ of −59.5 kcal/mol as the initial structure for the MD simulation. The distance between the ligand and $R310^{7.40}$ of DP was 1.7 Å. To identify the residues that can stably interact with the ligand, the ligand-bound models for $EP_2$ and DP were subjected to the MD simulation.

### 3.2.2 *MD simulations of $EP_2$ and DP*

Molecular dynamics simulations were monitored with the RMSD and the potential energy, as described in Sect. 2. Supplement 5 shows a plot of the RMSD between each model structure during the MD trajectory and the initial model structure. In each model system, the RMSD reached equilibrium and oscillated around the average value after 4–5 ns. The potential energies for the entire systems during the MD trajectory are shown in Supplement 6. The energies rapidly reached equilibrium for both systems with the $EP_2$ and DP models.

As described above, the conserved $Arg^{7.40}$ of PNRs is considered to interact with the carboxy group of prostanoids [9–13]. The length of the salt bridge between the carboxy group and the Arg was monitored during the simulation (Supplement 7) and suggested that the interaction was stable in ligand-bound models of both $EP_2$ and DP.

To evaluate the fluctuation of each residue of the model systems during MD, the RMSF was calculated for the Cα atom. The RMSF is plotted as a function of the residue position in Supplement 8. As shown in the plots, the fluctuations of the TM helices in both the $EP_2$ and DP model structures were small. In contrast to those regions, the fluctuations of the loop regions were large. Especially, the N-terminal loop, ICL1, and ICL3 of the $EP_2$ model structure greatly fluctuated, whereas the fluctuations of the N-terminal loop and ICL3 of the DP model structure were large. ICL3 is considered to be intrinsically disordered and involved in interactions with G proteins [45]. Thus, the fluctuation in ICL3 may be related to the signal transduction by PNRs. The N- and the C-terminal loops were also suggested to be intrinsically disordered [45]. However, the fluctuations of the C-terminal loops of both PNRs were smaller than those of the other intrinsically disordered loops.

### 3.2.3 *Differences in ligand recognition mechanisms between $EP_2$ and DP*

*3.2.3.1 Interaction between $EP_2$ and $PGE_2$* As described in Sect. 2, 10 structures were sampled from the last 1 ns of the simulation of the ligand-bound model of $EP_2$. The MM-GBSA ligand interaction energies of these structures are shown in Supplement 9. The whole structure and the vicinity of the ligand in the model with the lowest MM-GBSA energy are depicted in Fig. 4. As shown in Fig. 3, 32 amino acid residues of $EP_2$ were found within 5 Å from