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

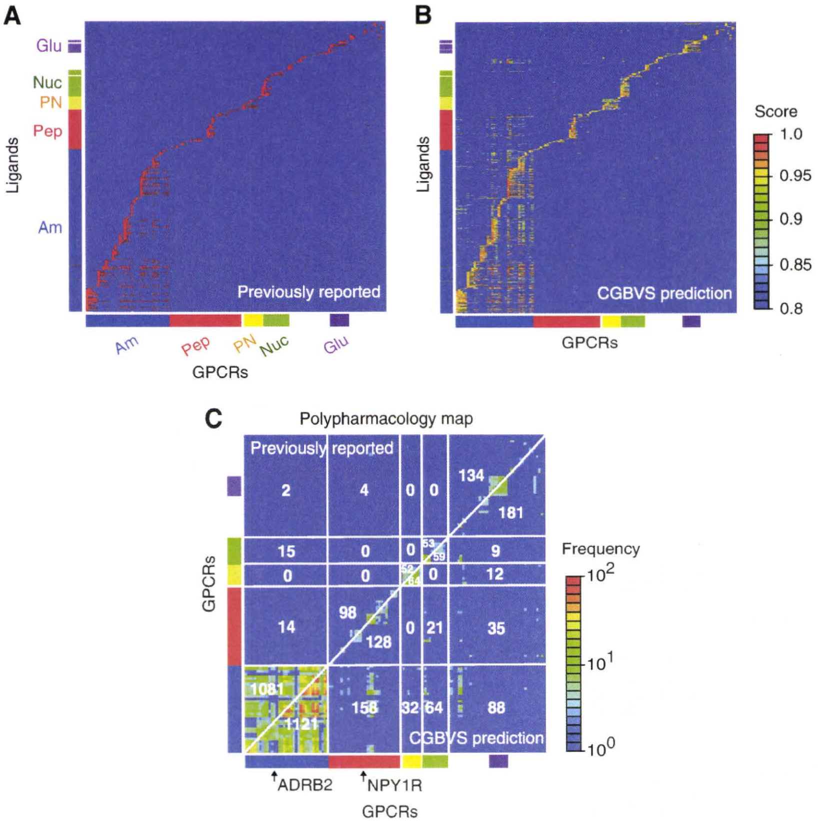


Figure 3 Comparative polypharmacology maps for GPCRs. (**A**) Map of previously reported compound–GPCR interactions. Vertical and horizontal axes represent the compounds and the GPCRs, respectively. The reported CPIs are depicted as red dots. CGBVS used the CPIs as training data. The colored bars along 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.

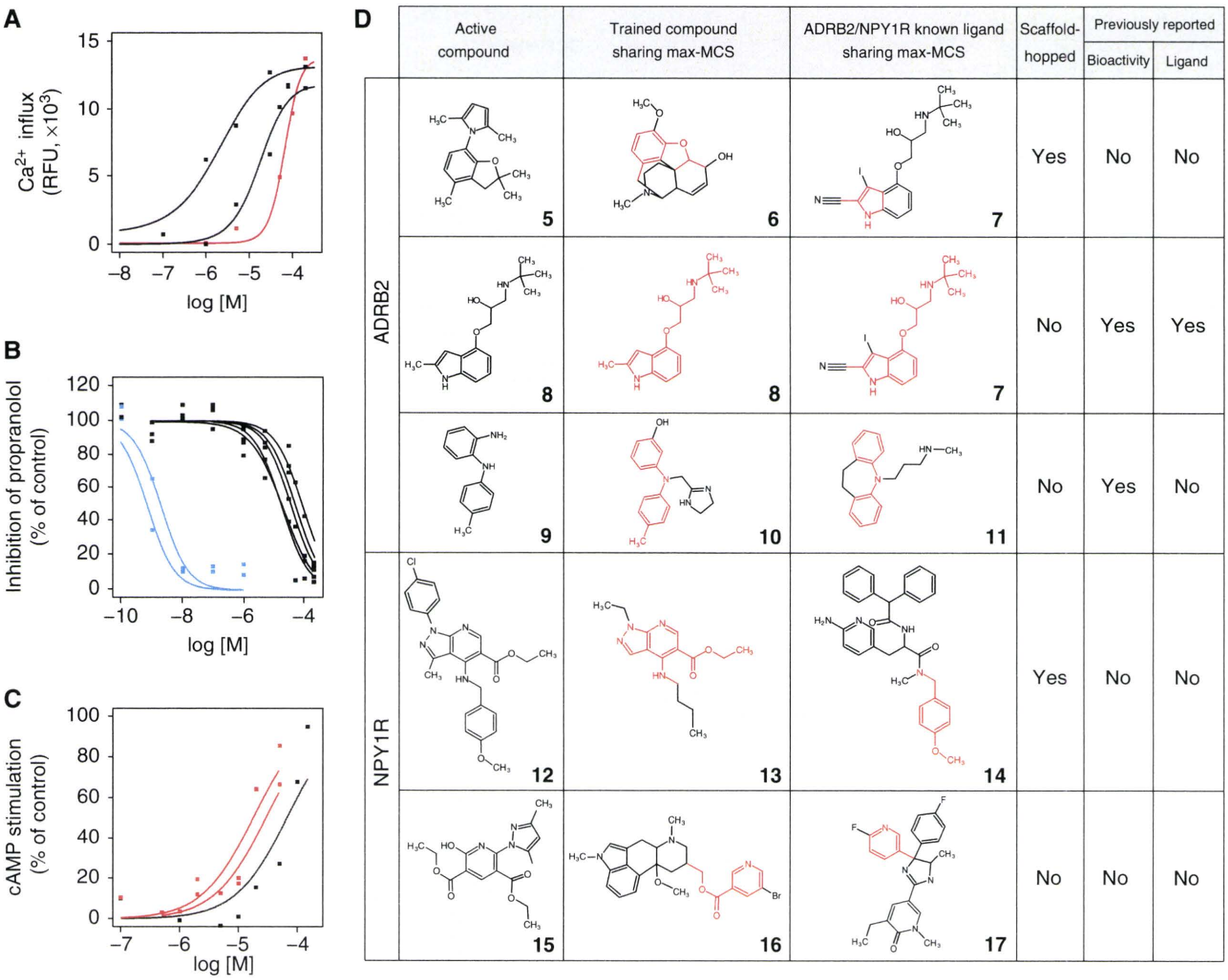
between distantly related GPCRs (Supplementary Figure S4). The cross-reactivity predictions provided by CGBVS also offer a promising approach for scaffold hopping in drug discovery. For example, many small ligands for non-peptidergic GPCRs were predicted to interact with peptidergic GPCRs as well, indicating 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



	Active compound	Trained compound sharing max-MCS	ADRB2/NPY1R known ligand sharing max-MCS	Scaffold-hopped	Previously reported	
					Bioactivity	Ligand
ADRB2	5	6	7	Yes	No	No
	8	8	7	No	Yes	Yes
	9	10	11	No	Yes	No
NPY1R	12	13	14	Yes	No	No
	15	16	17	No	No	No

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.

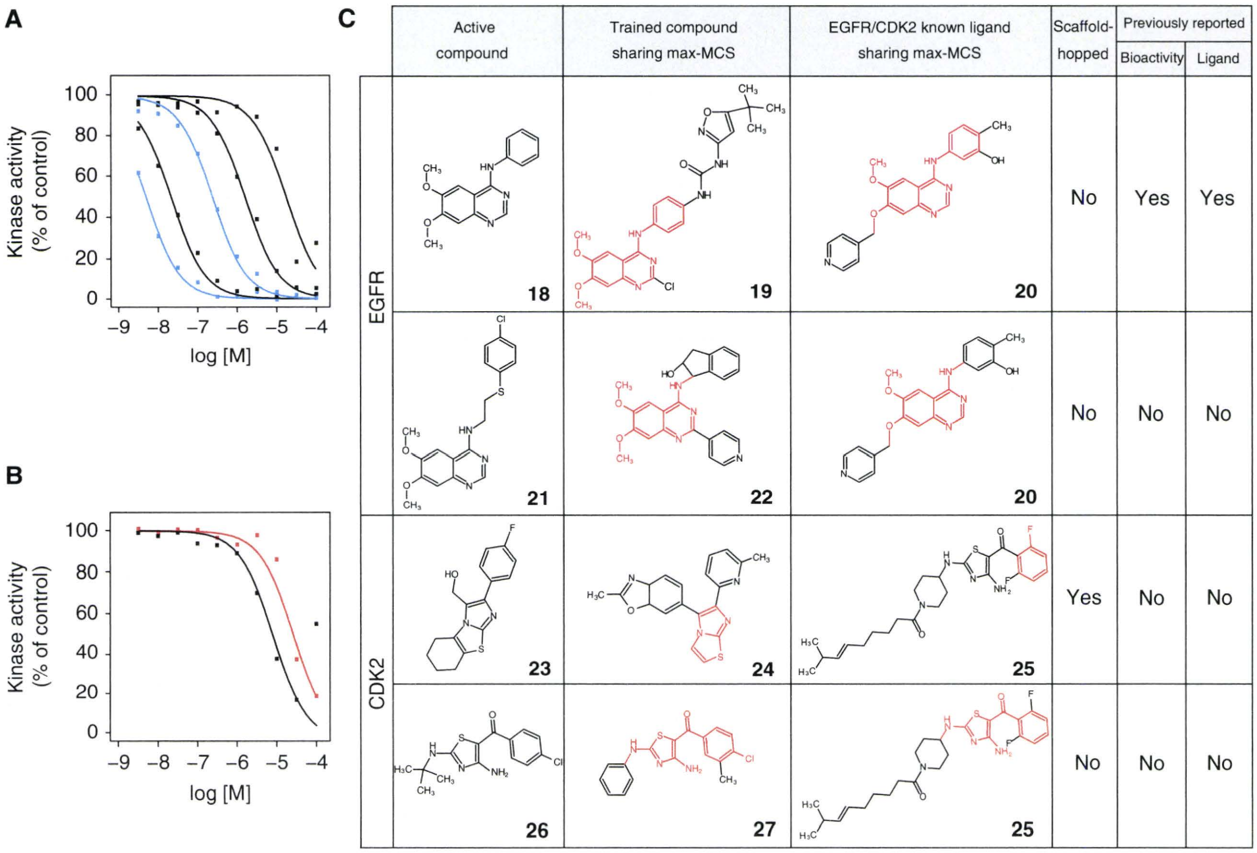
Application of CGBVS to kinase inhibitor screening

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



	Active compound	Trained compound sharing max-MCS	EGFR/CDK2 known ligand sharing max-MCS	Scaffold-hopped	Previously reported	
					Bioactivity	Ligand
EGFR	18	19	20	No	Yes	Yes
	21	22	20	No	No	No
CDK2	23	24	25	Yes	No	No
	26	27	25	No	No	No

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_{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_{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

Cross-Target View to Feature Selection: Identification of Molecular Interaction Features in Ligand−Target Space

Satoshi Niijima,* Hiroaki Yabuuchi, and Yasushi Okuno

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

Received April 12, 2010

There is growing interest in computational chemogenomics, which aims to identify all possible ligands of all target families using in silico prediction models. In particular, kernel methods provide a means of integrating compounds and proteins in a principled manner and enable the exploration of ligand−target binding on a genomic scale. To better understand the link between ligands and targets, it is of fundamental interest to identify molecular interaction features that contribute to prediction of ligand−target binding. To this end, we describe a feature selection approach based on kernel dimensionality reduction (KDR) that works in a ligand−target space defined by kernels. We further propose an efficient algorithm to overcome a computational bottleneck and thereby provide a useful general approach to feature selection for chemogenomics. Our experiment on cytochrome P450 (CYP) enzymes has shown that the algorithm is capable of identifying predictive features, as well as prioritizing features that are indicative of ligand preference for a given target family. We further illustrate its applicability on the mutation data of HIV protease by identifying influential mutated positions within protease variants. These results suggest that our approach has the potential to uncover the molecular basis for ligand selectivity and off-target effects.

INTRODUCTION

The last several years have seen a paradigm shift in pharmaceutical research from traditional target-specific approaches to a cross-target approach, offering tremendous opportunities for establishing novel drug design strategies to accelerate the drug discovery process. Receptors are no longer viewed as single entities but grouped into sets of related proteins or protein classes that are explored in a systematic manner. Chemogenomics has emerged as an interdisciplinary field, aiming at comprehensive coverage of ligand−target interactions, that is, identifying all possible ligands of all target families.[1,2] Concomitantly, high-throughput data being accumulated at an ever-increasing rate have triggered the development of novel in silico methodologies to comprehensively predict ligand−target interactions and binding affinities.[3]

In particular, a ligand−target approach has recently received much attention.[1] This approach represents a single-step process to integrate compounds and proteins into pairs and predict ligand−target binding on a genomic scale using machine learning models (e.g., that of Bock and Gough[4] and Erhan et al.[5]). The advantage of the ligand−target approach lies in that it allows predictions of new interactions even when neither ligands for a specific target nor targets for a specific ligand are known. Moreover, the greatest impact can be expected for targets devoid of structural 3D data, because classical drug design strategies like structure-based virtual screening cannot be applied to such targets.[1] Importantly, the ligand−target approach also has the potential to reveal ligand selectivity and off-target effects by comprehensive analysis of cross-reactivity of ligands.

Previous studies on the ligand−target approach have devoted much effort to the development of prediction models. Although advanced statistical models often yield better performance, they are usually constructed in a black-box way, lacking transparency and interpretability. This significantly hinders our understanding of the molecular basis for ligand−target binding. In order to gain an in-depth understanding of the link between ligands and targets, a next step should then be directed toward the identification of structural and physicochemical features associated with the binding. A promising in silico approach to this problem is to apply feature selection techniques,[6] which are typically used for molecular descriptor selection in chemoinformatics (e.g., the work of Fröhlich et al.,[7] Byvatov and Schneider,[8] and Xue et al.[9]). However, existing techniques for the ligand-based approach only consider individual targets and perform feature selection in the ligand space of a specific target and, thus, have a major limitation in capturing cross-reactive patterns. Given the fact that a single compound exhibits different binding affinities against multiple targets, feature selection needs to be performed instead in a ligand−target space, into which compounds are mapped jointly with targets. Because the ligand−target approach is itself an emerging strategy, feature selection based on the cross-target view is entirely an unexplored topic of research, and to our knowledge, no method exists that enables feature selection in the ligand−target space.

Here we describe a feature selection approach that works in a kernel-induced feature space[10] representing a ligand−target space. In particular, we propose using kernel dimensionality reduction (KDR)[11,12] for feature selection with an efficient

* To whom correspondence should be addressed. E-mail: niijima@pharm.kyoto-u.ac.jp.

10.1021/ci1001394
Published on Web 12/10/2010

algorithm, in order to identify molecular interaction features that contribute to prediction of ligand−target binding affinities. The quality of a prediction model is known to highly depend on the selected features and, hence, potentially benefits from feature selection. Indeed, the prediction performance can be improved by using only informative features. Reducing the number of features also helps to avoid overfitting.[13] Most importantly, selected features often facilitate interpretation of the model. For example, selected features in the ligand−target binding affinity space can serve to characterize privileged structures—selected substructures able to provide high-affinity ligands for a set of receptors[14]—and thus have implications for lead compound optimization for drug design. Furthermore, feature selection based on the cross-target view may provide insights into the molecular basis for ligand selectivity and off-target effects and has the potential to uncover the complex mode of drug actions. In the present study, we apply the proposed algorithm to a data set on cytochrome P450 (CYP) enzymes and show its capability of selecting a small subset of predictive features, which are further found to be indicative of ligand preference for a set of targets. We also evaluate our algorithm on the mutation data of HIV protease and illustrate its applicability by identifying influential amino acid positions within mutated variants.

METHODS

Representation of Ligand−Target Space. A key element of the ligand−target approach is the construction of ligand−target pairs, which need to be integrated from heterogeneous data types of compounds and proteins.[15] For this purpose, unified pair descriptions have been proposed and applied to search for novel active pairs.[4,5,16−21] In particular, it has recently proven that kernel methods[10] provide a general framework for integrating compounds and proteins, regardless of how they are represented, respectively.[17] Here we describe how a ligand−target space can be constructed via kernels.

Let us denote compounds and proteins by c_i and p_i, respectively. The binding affinity prediction problem can be formulated as the following machine learning problem: given a set of n ligand−target pairs (c_1,p_1), ..., (c_n,p_n) with known affinity values, construct a model to make predictions of activities of candidate pairs. To apply standard regression models, we first consider representing each ligand−target pair by a vector. Formally, given a chemical vector $\Phi_\ell(c_i)$ and a protein vector $\Phi_t(p_i)$, we need to form a single vector $\Psi(c_i,p_i)$ using $\Phi_\ell(c_i)$ and $\Phi_t(p_i)$.

To capture interactions between features of the compound c_i and those of the protein p_i, previous studies[5,16,17] proposed to represent the pair (c_i,p_i) by

$$\mathbf{\Psi}(c_i, p_i) = \mathbf{\Phi}_\ell(c_i) \otimes \mathbf{\Phi}_t(p_i) \tag{1}$$

The tensor product operation ⊗ indicates that the pair is represented by all possible products (crossover) of the features of c_i and p_i, thereby seeking to fully encode correlations between them. The explicit computation of the products, however, demands expensive computation time and storage. Suppose that the number of features for c and p is d_c and d_p, respectively, then the pair is composed of $d_c \times d_p$ features. Fortunately, this computational bottleneck can be circumvented under the framework of kernel methods.[10]

Kernel methods are a class of algorithms that apply linear machine learning algorithms for classification or regression in a high-dimensional, possibly infinite-dimensional, feature space. Formally, the samples $x_i \in \mathrm{IR}^d$ are implicitly mapped into a feature space as $\Phi(x_i) \in \mathcal{F}$ such that the inner product between a pair of samples is given by a kernel function $k(x_i,x_j) = \Phi(x_i)\cdot\Phi(x_j)$, which measures the similarity between x_i and x_j. If the samples are expressed in terms of inner products only, the so-called kernel trick allows a variety of linear algorithms to work in the feature space constructed via a kernel function, without explicitly computing vectors comprising many features. In the case of binding affinity prediction, a ligand−target pair constitutes a single sample, and the kernel measures the similarity between ligand−target pairs. It can be shown that the kernel between pairs that are represented by eq 1 is decomposed as

$$\begin{aligned}\mathbf{\Psi}(c_i, p_i)^{\mathrm{T}}\mathbf{\Psi}(c_j, p_j) &= (\mathbf{\Phi}_\ell(c_i) \otimes \mathbf{\Phi}_t(p_i))^{\mathrm{T}}(\mathbf{\Phi}_\ell(c_j) \otimes \mathbf{\Phi}_t(p_j)) \\ &= \mathbf{\Phi}_\ell(c_i)^{\mathrm{T}}\mathbf{\Phi}_\ell(c_j) \times \mathbf{\Phi}_t(p_i)^{\mathrm{T}}\mathbf{\Phi}_t(p_j)\end{aligned} \tag{2}$$

It is readily seen that the similarity between two ligand−target pairs is simply the product of the similarity between the two compounds and the similarity between the two proteins. This indicates that eq 2, known as the tensor product kernel, can be computed easily once ligand and target kernels have been computed separately, avoiding the explicit computation of all the products of the features representing compounds and proteins. More generally, the kernels for compounds and proteins are not limited to the inner products of vectors and allow one to define the similarity based on nonvectorial data such as graphs for compounds, and amino acid sequences or 3D structures for proteins. Formally, denoting the kernels for compounds and proteins respectively by

$$k_\ell(c_i, c_j) = \mathbf{\Phi}_\ell(c_i)^{\mathrm{T}}\mathbf{\Phi}_\ell(c_j) \tag{3}$$

$$k_t(p_i, p_j) = \mathbf{\Phi}_t(p_i)^{\mathrm{T}}\mathbf{\Phi}_t(p_j) \tag{4}$$

the kernel between the two pairs is given by

$$k((c_i, p_i), (c_j, p_j)) = k_\ell(c_i, c_j) \times k_t(p_i, p_j)$$

In this way, the kernel-based approach allows versatile representation of the ligand−target space.

Feature Selection in Ligand−Target Binding Affinity Space. To select informative features in the ligand−target space as defined above, we need to perform feature selection in a kernel-induced feature space. Despite a broad spectrum of existing methods for feature selection,[6] there are few techniques that can be applied to such a ligand−target space constructed via kernels.

Here we adapt a semiparametric dimensionality reduction approach, called kernel dimensionality reduction (KDR),[11,12] to feature selection for binding affinity prediction. In particular, we propose an efficient feature selection algorithm to identify molecular interaction features that contribute to prediction of ligand−target binding.

KDR is a statistically grounded approach to dimensionality reduction, which aims to represent new features in the form of linear combinations of original features. This is achieved by minimizing the independence (i.e., maximizing the dependence) between a set of features of samples and their labels. KDR can also be used to select a subset of original features that well captures the dependency. KDR enables us to measure the (in)dependence in a kernel-induced feature space, thereby providing a general means for dimensionality reduction and feature selection. In terms of statistics, KDR is based on the estimation and optimization of covariance operators on kernel-induced feature spaces, and the operators are used to provide a general characterization of conditional independence. KDR has the advantage that it imposes no strong assumptions either on the marginal distributions of samples and labels or on the conditional probability of labels given samples. This makes it applicable to diverse problems. Nevertheless, the application of KDR is still limited to typical machine learning problems[12] and yet to be seen in the chemoinformatics domain. Of note, this study is distinguished from others in that KDR is adapted to feature selection in a joint feature space of ligands and targets.

Among possible KDR objective functions, we employ the following simple function based on the trace of the empirical conditional covariance operator:[12]

$$\mathrm{Tr}[(\mathrm{HKH} + \lambda I_n)^{-1}\mathrm{HLH}] \qquad (5)$$

Here, Tr denotes the trace of a matrix, and $\mathrm{K}, \mathrm{L} \in \mathrm{IR}^{n\times n}$ are the kernel matrices for the samples x_i and the labels y_i, respectively. $I_n \in \mathrm{IR}^{n\times n}$ is the identity matrix, and λ denotes a regularization parameter. $\mathrm{H} = I_n - (1/n)ee^{\mathrm{T}}$ is a centering matrix, where $e = (1, ..., 1)^{\mathrm{T}}$ is an n-dimensional vector. It is worth noting that this objective function has a close relationship with the Hilbert–Schmidt independence criterion (HSIC),[22] and eq 5 can be derived from the objective function of kernel ridge regression.[23] It is interesting to note that sliced inverse regression (SIR),[24] which is well-known and closely related to KDR, has recently been extended to kernel SIR (KSIR)[25] to overcome some limitations of SIR, yet unlike KDR, KSIR is sensitive to the number of slices which needs to be set a priori.

In the context of binding affinity prediction, the kernel matrix K defines the similarities between pairs, $x_i = (c_i,p_i)$, and L is simply computed as $L_{ij} = y_iy_j$, where y_i represents the affinity value given to x_i. Further, if we use the tensor product kernel eq 2, K can be represented as

$$K = K_l \circ K_t$$

where the elements of K_l and K_t are calculated by eqs 3 and 4 and $\circ$ denotes the Hadamard product (elementwise product) operation.

As detailed in the work of Fukumizu et al.,[12] selection of relevant features exhibiting high dependence (i.e., low independence) on the labels reduces to the minimization of the objective function eq 5. Since exhaustive search is computationally prohibitive, we aim to achieve this with a backward elimination algorithm—the relevance of individual features is evaluated on a leave-one-out basis, and the least dependent feature maximizing the objective function is recursively eliminated from a full feature set. Alternative greedy algorithms such as forward search can also be used, but the backward elimination algorithm often yields better features, due to the evaluation of features in the presence of all others. To name but a few of this kind, SVM-RFE[26] and the BAHSIC algorithm[23] have indeed shown successful results.

Equation 5 can be computed independent of a particular classifier, yet the objective function involves the inverse of a sample-sized matrix. Thus, regardless of the search algorithm used, the computation of eq 5 based on leave-one-feature-out (LOFO) becomes intractable as the sample size and/or feature size increases.

Efficient Feature Selection Algorithm. To overcome this computational bottleneck, we propose an efficient algorithm for feature selection in the ligand–target binding affinity space constructed via the tensor product kernel. Specifically, we seek to improve the computational efficiency of

$$(\mathrm{H}(K_l \circ K_t)\mathrm{H} + \lambda I_n)^{-1}$$

in the LOFO process, i.e.,

$$\Delta^{(-i)} = (\mathrm{H}((K_l - f_l^{(i)}f_l^{(i)\mathrm{T}}) \circ K_t)\mathrm{H} + \lambda I_n)^{-1} \qquad (6)$$

when selecting chemical features of ligands, while keeping protein features of targets unchanged. Here, $f_l^{(i)}$ denotes a chemical feature to be left out. Note that the proposed algorithm is valid only when the linear kernel is used for ligands, but the targets can be represented by various features implicitly defined by kernels. First, we approximate the target kernel matrix K_t by a matrix G of lower-rank k as

$$K_t \approx \mathrm{GG}^{\mathrm{T}},\ \mathrm{G} = (g_1, ..., g_k) \in \mathrm{IR}^{n\times k} \qquad (7)$$

This low-rank approximation can be efficiently done using, e.g., incomplete Cholesky decomposition.[27] For simplicity, let us define

$$\mathrm{P} = \mathrm{H}(K_l \circ K_t)\mathrm{H} + \lambda I_n \in \mathrm{IR}^{n\times n}$$

$$\mathrm{Q} = \mathrm{H}(f_l^{(i)} \circ g_1, ... f_l^{(i)} \circ g_k) \in \mathrm{IR}^{n\times k}$$

From eq 6, we have

$$\Delta^{(-i)} = (\mathrm{P} - \mathrm{QQ}^{\mathrm{T}})^{-1}$$

Further, from the Sherman–Morrison–Woodbury formula,[28] we have

$$\Delta^{(-i)} = \mathrm{P}^{-1} + \mathrm{P}^{-1}\mathrm{Q}(I_k - \mathrm{Q}^{\mathrm{T}}\mathrm{P}^{-1}\mathrm{Q})^{-1}\mathrm{Q}^{\mathrm{T}}\mathrm{P}^{-1}$$

Therefore, if $k \ll n$, computing the matrix inversion of $I_k - \mathrm{Q}^{\mathrm{T}}\mathrm{P}^{-1}\mathrm{Q} \in \mathrm{IR}^{k\times k}$ is efficient, and hence, $\Delta^{(-i)}$. Equation 7 can be computed independent of the LOFO process, and P^{-1} can be updated consecutively. In the case of binding affinity prediction, k is upper-bounded by $\min(n_p,d_p)$, where n_p is the number of proteins. Because binding affinities are typically measured for a relatively small number of targets against a series of compounds in chemical libraries, $k \leq n_p \ll n$ usually holds, and the overall computation time can be saved by (n/k)-fold compared with a naive computation of eq 6. Of note, the proposed algorithm allows one to use different kernels for proteins when selecting chemical features. The algorithm can be summarized as follows:

```
Input: $\{((c_1,p_1),y_1),\ldots,((c_n,p_n),y_n)\}$,
        chemical feature set $\mathscr{S}$
Output: A ranking list $\mathscr{R}$ of chemical features
1: Compute $H, L, K_\ell$ and $K_t$;
2: Compute $G$, $Q$, and $P^{-1}$;
3: $\mathscr{R} \leftarrow \emptyset$;
Repeat 4–7 until $\mathscr{S} = \emptyset$
4: $j \leftarrow \arg\max_{i \in \mathscr{S}} \mathrm{Tr}\left[\Delta^{(-i)} HLH\right]$;
5: $P^{-1} \leftarrow \Delta^{(-j)}$;
6: $\mathscr{S} \leftarrow \mathscr{S} \setminus \{j\}$;
7: $\mathscr{R} \leftarrow \mathscr{R} \cup \{j\}$.
```

In the above algorithm, a single feature is recursively eliminated from $\mathscr{T}$ and added to the end of $\mathscr{R}$, in which the features toward the end of $\mathscr{R}$ have higher dependence on the labels in the presence of target information. Accordingly, the top-ranked features can be finally taken from the tail of $\mathscr{R}$.

Likewise, in the case of protein feature selection with chemical features unchanged, the same algorithm is applicable by simply replacing eqs 6 and 7 with

$$\Delta^{(-i)} = (\mathrm{H}((K_t - f_t^{(i)} f_t^{(i)\mathrm{T}}) \bigcirc K_l)\mathrm{H} + \lambda I_n)^{-1}$$

$$K_l \approx \mathrm{G}\mathrm{G}^{\mathrm{T}}$$

where $f_t^{(i)}$ denotes a protein feature.

EXPERIMENTS

Data Sets. The CYP data set was taken from the study of Kontijevskis et al.[29] The affinity values of 798 ligand−target pairs (consisting of 371 inhibitors and 14 CYP enzymes) were experimentally determined and thus available. Each pair has a $\mathrm{pIC}_{50} = -\log(\mathrm{IC}_{50})$ value, where IC_{50} represent half-maximal inhibitory concentrations. The pIC_{50} values range from 0.46 to 8.70, with a mean value of 4.39. The distributions of the pIC_{50} values are shown for each CYP enzyme in Figure 1.

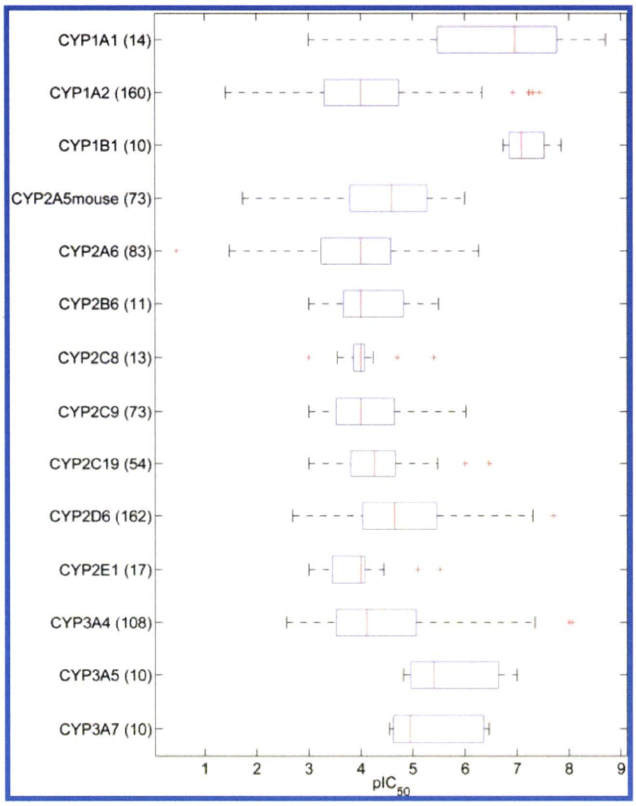


Figure 1. Boxplots of the pIC_{50} values for 14 CYP enzymes. Shown in parentheses are the numbers of inhibitors.

The mutation data of HIV protease were collected from the literature listed in the work of Lapins and Wikberg.[30] After carefully checking the literature sources, we chose to use a total of 389 ligand−target pairs with known $\mathrm{p}K_i$ values, where $\mathrm{p}K_i = -\log(K_i)$ and K_i represents inhibition constants. The ligand−target pairs consist of 21 ligands and 69 mutated protease variants as well as the wild-type, and the number of mutated positions amounts to 42 in the variants. The $\mathrm{p}K_i$ values range from 5.37 to 11.89, with a mean value of 8.75. The distributions of the $\mathrm{p}K_i$ values are shown for each ligand in Figure 2.

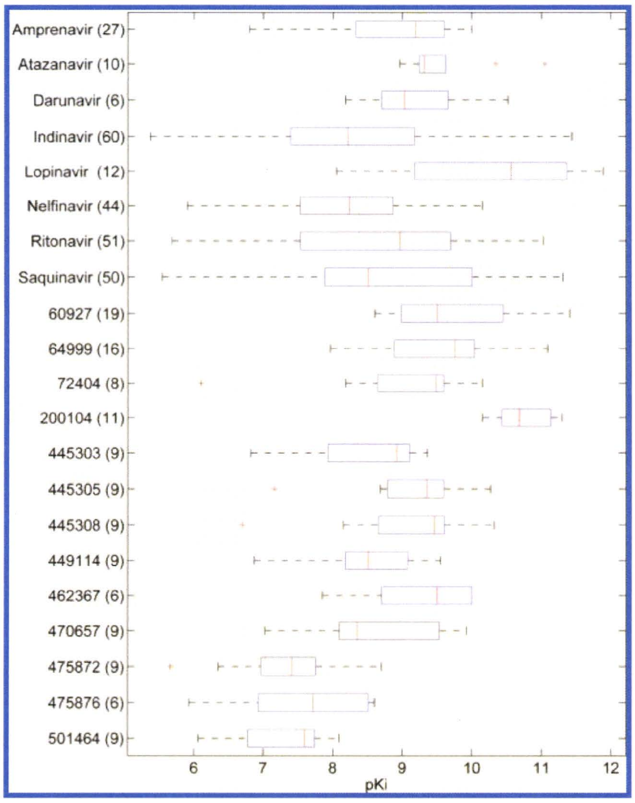


Figure 2. Boxplots of the $\mathrm{p}K_i$ values for 21 ligands. The ligand numbers indicate PubChem CIDs. Shown in parentheses are the numbers of mutated protease variants and the wild-type.

Kernels for Ligands and Targets. A wide variety of molecular features (descriptors) have been developed thus far to characterize the chemical structures and the physicochemical and molecular properties of compounds.[31] Here, we chose to use a total of 1664 descriptors calculated by version 1.4 of the Dragon software.[32] This descriptor set contains a range of 1D to 3D molecular features that fall into the following categories: constitutional descriptors, topological descriptors, walk and path counts, connectivity indices, information indices, 2D autocorrelations, edge adjacency indices, Burden eigenvalue descriptors, topological charge indices, eigenvalue-based indices, Randic molecular profiles, geometrical descriptors, RDF descriptors, 3D-MoRSE descriptors, WHIM descriptors, GETAWAY descriptors, functional group counts, atom-centered fragments, charge descriptors, and molecular properties. Before calculating these descriptors, MOE[33] was used to preprocess the raw macromolecular structures, including elimination of the crystallographic water molecules, removal of salts, addition of hydrogen atoms, and charge processing. A variation filter was then applied to eliminate the descriptors showing little variation across the compounds, resulting in 1397 descriptors for the CYP data set and 1378 descriptors for the HIV protease data set, and the values were scaled in the range of -1 to 1.

There exist several means of representing proteins or defining protein kernels. Among others, the sequence-based approach has proven effective when the availability of 3D structures is very limited. In view of this, we employed two different kernels in the experiment for CYPs: PROFEAT feature vectors[34] with RBF kernel (PROFEAT+RBF) and mismatch kernel.[35] These kernels can be computed from sequences alone and have shown good performance in protein classification and remote homology detection, as well as in ligand prediction.[17,20]

The PROFEAT feature vector provided by the PROFEAT Webserver[34] contains 1497 features representing, e.g., dipeptide composition and physicochemical properties of sequences. The RBF kernel was calculated using the feature vectors to represent the sequence similarity. The mismatch kernels are a class of string kernels, which can be computed as a dot product between two vectors consisting of frequencies of subsequences within the whole sequence. The mismatch kernels allow for mutations between the subsequences. Specifically, the mismatch kernel is calculated based on shared occurrences of (k,m)-patterns in the data, where the (k,m)-pattern consists of all k-length subsequences that differ from it by at most m mismatches. In our experiment, the typical choice of $k = 3$ and $m = 1$ was used in accordance with the work of Jacob and Vert.[17]

Whereas the chemical descriptors were subjected to feature selection for the CYP data set, feature selection was applied to protein features in the experiment for HIV protease, with the representation of ligands unchanged. Therefore, different kernels can be used for representing ligands, and we herein used the Dragon descriptors with RBF kernel. The targets were described using three z-scales, z_1, z_2, and z_3[36] following the work of Lapins and Wikberg.[30] The z-scales are the leading principal components obtained from 26 measured and computed physicochemical properties of amino acids and can be interpreted as hydrophobicity (z_1), steric properties (z_2), and polarity (z_3) of amino acids. As a result, the total number of protein features amounts to $42 \times 3 = 126$.

Performance Evaluation. The proposed algorithm selects features independent of a specific classifier used. It is therefore of interest to evaluate the predictive ability of the selected features using different kernel-based regression models. In the present study, we employed two representative models: kernel ridge regression (KRR)[37] and support vector regression (SVR).[38] The regularization parameter of KRR was fixed to the average eigenvalue of the kernel matrix. For SVR, the regularization parameter C was selected from $\{0.01, 0.1, ..., 100\}$, and the default value of 0.1 was used for ε of loss function. The γ parameter of RBF kernel for compounds (the CYP data set) and for proteins (the HIV protease data set) was set to α/(number of features), and α was selected from $\{2^{-4}, 2^{-3}, ..., 2^{4}\}$. The parameter λ of eq 5 was fixed to the average eigenvalue of HKH, which can easily be computed as Tr(HKH)/n. We used the LIBSVM library[39] for the implementation of SVR and in-house C codes for feature selection and KRR.

We used repeated random splitting for performance evaluation—the whole samples were partitioned randomly and repeatedly into training and test sets. The ratio of the training against test set was set to 6:1 for the CYP and HIV protease data sets, in accordance with previous studies.[29,30] Feature selection was performed using only the training set, and the q^2 value of the learnt regression model was obtained using the test set. Given n test samples, q^2 is defined as

$$q^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

where y_i and $\hat{y}_i$ are the true affinity value and estimated value of sample i, respectively, and $\bar{y}$ is the average value of y_is. Thus, the larger value of q^2 indicates the better performance. The random splitting was repeated 20 times, and the q^2 value averaged over the 20 runs and the corresponding standard deviations are reported here. Because the rank of features can vary depending on the training sets, we calculated scores as the average ranks of the 20 ranking lists.

The aim of the experiments was to evaluate how the prediction performance would be affected by eliminating possibly irrelevant features and whether our algorithm can identify a small set of informative features. To this end, the number of features was varied from all features to >50 by 10% decrements, and from 50 to 5 in decrements of 5. The predictive ability of the selected features was assessed by KRR and SVR as a function of the number of features. There exists no competing method that enables feature selection in a ligand−target space constructed via kernels, but it is worth making a comparison with random selection as a baseline to evaluate how well the proposed algorithm performs in practice. For this purpose, we randomly selected features from the whole feature set for each data set splitting and evaluated their prediction performance in the same way as for the selected features of the proposed algorithm.

RESULTS AND DISCUSSION

Chemical Feature Selection for CYPs. CYPs constitute a superfamily of heme-containing enzymes, which are involved in the oxidative metabolism of a large number of structurally different compounds of both endogenous and exogenous origin. It is known that more than 90% of all pharmaceuticals are metabolized by CYPs; CYP1A2, CYP2C9, CYP2C19, CYP2D6, and CYP3A4 are predominant among others.[40] These enzymes are susceptible to inhibition due to their broad specificity, giving rise to unexpected drug−drug interactions and drug toxicity. This makes prediction of interactions between CYPs and drugs a challenging problem.

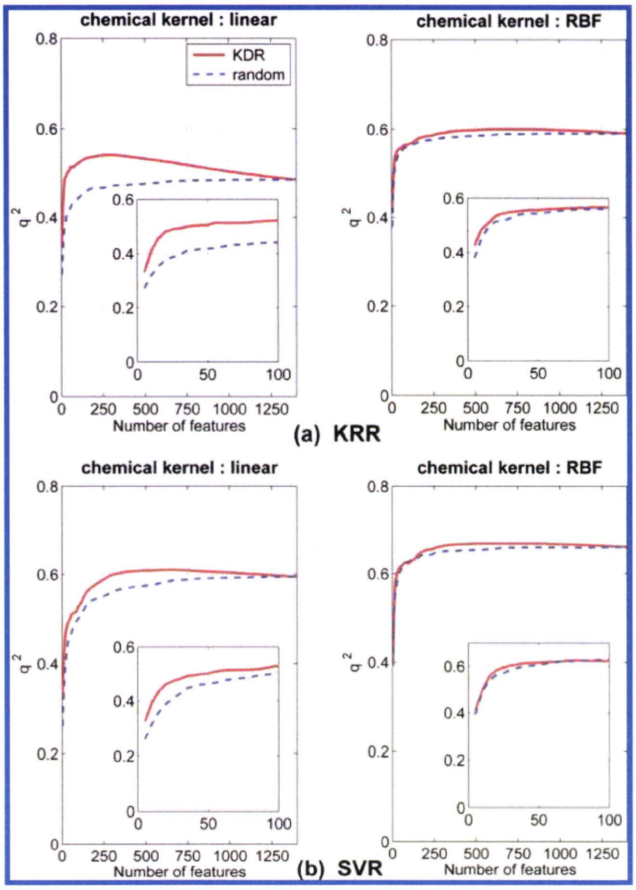


Figure 3. Average q^2 values as a function of the number of features. The PROFEAT+RBF kernel was used for CYPs. (a) KRR with the linear and RBF kernels for compounds. (b) SVR with the linear and RBF kernels for compounds.

Table 1. Performance Comparison for CYPs Using the PROFEAT+RBF kernel[a]

	KRR (chemical kernel: linear)		KRR (chemical kernel: RBF)	
no. of features	KDR	random	KDR	random
10	0.41 ± 0.08	0.32 ± 0.09	0.48 ± 0.07	0.45 ± 0.08
20	0.48 ± 0.07	0.38 ± 0.09	0.53 ± 0.06	0.51 ± 0.08
30	0.49 ± 0.07	0.39 ± 0.09	0.55 ± 0.07	0.53 ± 0.08
50	0.50 ± 0.07	0.42 ± 0.10	0.55 ± 0.07	0.54 ± 0.07
108	0.52 ± 0.08	0.44 ± 0.10	0.57 ± 0.06	0.56 ± 0.07
all (1397)	0.48 ± 0.08		0.59 ± 0.07	
	0.54 ± 0.07 (KDR, 185)		**0.60** ± 0.06 (KDR, 392)	
	SVR (chemical kernel: linear)		SVR (chemical kernel: RBF)	
no. of features	KDR	random	KDR	random
10	0.40 ± 0.09	0.32 ± 0.12	0.50 ± 0.10	0.49 ± 0.11
20	0.46 ± 0.08	0.39 ± 0.12	0.58 ± 0.08	0.56 ± 0.11
30	0.48 ± 0.10	0.43 ± 0.12	0.60 ± 0.08	0.58 ± 0.10
50	0.50 ± 0.09	0.46 ± 0.12	0.62 ± 0.07	0.60 ± 0.09
108	0.53 ± 0.09	0.51 ± 0.12	0.63 ± 0.07	0.63 ± 0.09
all (1397)	0.59 ± 0.09		0.66 ± 0.09	
	0.61 ± 0.09 (KDR, 392)		**0.67** ± 0.07 (KDR, 352)	

[a] Shown are the average q^2 values and standard deviations. The last row shows the best values in bold face (with method and the number of features in parentheses).

As shown in Figure 1, the distributions of the pIC_{50} values significantly overlap between the predominant CYPs and exhibit a wide range of inhibitory activities for most of the CYP enzymes, albeit biased for a fraction of them (e.g., CYP1B1). When simply estimating the value of a given inhibitor to be the mean value for the target CYP, we observed $q^2 = 0.18 \pm 0.07$, which clearly indicates the need for the use of both ligand and target information to make better predictions, and this can be achieved by the ligand−target approach.

We applied the proposed algorithm to the CYP data set. Figure 3 and Table 1 show the q^2 values for KRR and SVR with the linear and RBF kernels for compounds and the PROFEAT+RBF kernel for CYPs. Using all the chemical features, KRR and SVR with the RBF kernel yielded $q^2 =$ 0.59 and 0.66, respectively. The same data set was analyzed by Kontijevskis et al.[29] using a different ligand−target approach that is based on linear partial least-squares (PLS), and the PLS-based model yielded q^2 values of 0.61−0.66. Although a fair comparison of the performance with the present study cannot be made due to the difference in chemical descriptors used, $q^2 = 0.66$ obtained by our SVR is comparable to the reported values in the previous study. It should be noted that the q^2 value exceeding 0.60 can be considered highly predictive, compared with previous in silico models for predicting CYP inhibition.[29]

To evaluate whether our algorithm can narrow an abundance of features that possibly include irrelevant ones down to the most informative features, we compared the performance between the proposed algorithm based on KDR and random selection by varying the number of chemical features. As the chemical features were removed, the q^2 values for random selection dropped gradually, whereas our algorithm was able to reduce the number of features while maintaining the same level of performance. As seen for KRR with the linear kernel, the performance could even be improved by reducing the feature size (Table 1); however, this was not observed for random selection. This result suggests that only a small subset of features is sufficient for making accurate prediction, while most of the other features are likely irrelevant to the prediction. Indeed, it can be seen from Table 1 that the q^2 value for SVR with the RBF kernel decreased merely from 0.66 with all features to 0.62 with 50 features.

Overall, the proposed algorithm performs better than random selection, but the difference is less remarkable for the RBF kernel. This may be because the features were optimized in the ligand−target space with the linear kernel for compounds and, hence, are not necessarily optimal in the ligand−target space with the RBF kernel. In principle, KDR can be computed in the latter space as well, but the computational cost is so demanding that our efficient algorithm can be a compromise between the cost and performance.

Because our feature selection algorithm is amenable to various kernels for proteins when selecting chemical features, the mismatch kernel was also applied in the same way. As shown in Figure 4 and Table 2, a similar trend was observed as the features were removed. In particular, we confirmed again that a small subset of features was as predictive as the given full feature set.

In the context of binding affinity prediction, the minimization of KDR favors chemical features that exhibit high dependence on affinity values in the presence of protein information. It is therefore of interest to see whether the

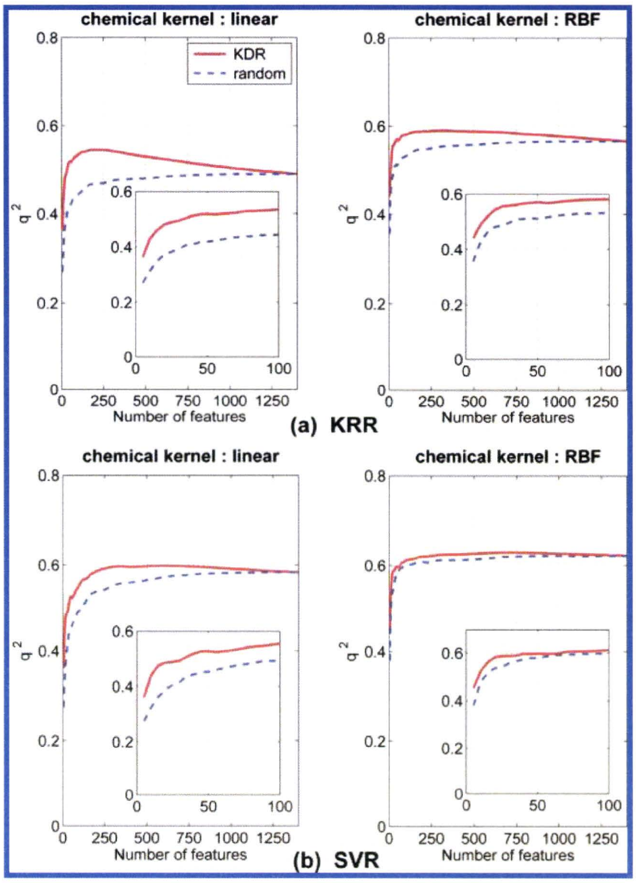


Figure 4. Average q^2 values as a function of the number of features. The mismatch kernel was used for CYPs. (a) KRR with the linear and RBF kernels for compounds. (b) SVR with the linear and RBF kernels for compounds.

Table 2. Performance Comparison for CYPs Using the Mismatch Kernel[a]

	KRR (chemical kernel: linear)		KRR (chemical kernel: RBF)	
no. of features	KDR	random	KDR	random
10	0.43 ± 0.08	0.31 ± 0.09	0.49 ± 0.08	0.43 ± 0.08
20	0.48 ± 0.06	0.37 ± 0.08	0.54 ± 0.07	0.48 ± 0.08
30	0.49 ± 0.07	0.39 ± 0.08	0.56 ± 0.07	0.49 ± 0.07
50	0.52 ± 0.08	0.42 ± 0.08	0.57 ± 0.07	0.51 ± 0.08
108	0.54 ± 0.07	0.45 ± 0.08	0.58 ± 0.07	0.53 ± 0.07
all (1397)	0.49 ± 0.07		0.57 ± 0.07	
	0.54 ± 0.07 (KDR, 108)		**0.59** ± 0.07 (KDR, 134)	

	SVR (chemical kernel: linear)		SVR (chemical kernel: RBF)	
no. of features	KDR	random	KDR	random
10	0.44 ± 0.08	0.32 ± 0.11	0.52 ± 0.10	0.48 ± 0.09
20	0.48 ± 0.07	0.39 ± 0.10	0.58 ± 0.08	0.54 ± 0.10
30	0.49 ± 0.08	0.42 ± 0.10	0.59 ± 0.08	0.56 ± 0.09
50	0.53 ± 0.09	0.45 ± 0.10	0.60 ± 0.08	0.58 ± 0.08
108	0.56 ± 0.09	0.50 ± 0.10	0.61 ± 0.08	0.60 ± 0.08
all (1397)	0.58 ± 0.08		0.62 ± 0.08	
	0.60 ± 0.09 (KDR, 316)		**0.63** ± 0.08 (KDR, 352)	

[a] Shown are the average q^2 values and standard deviations. The last row shows the best values in bold face (with method and the number of features in parentheses).

selected features can give some explanation about the characteristics of CYP inhibitors. Table 3 lists the top 30 features selected from a total of 1397 chemical features. It can be seen that two features representing the octanol−water partition coefficient (logP) received relatively high ranks (6th and 21st). Given that the ligand−target pairs for CYP1A2 and CYP3A4 account for more than 30% of the data set, this is in line with the fact that inhibitors of CYP1A2 and CYP3A4 are known to show high lipophilicity.[41,42] ARR (10th) and nBnz (23rd) are likely to reflect that CYP1A2 inhibitors have high aromaticity (number of aromatic carbons).[41] Also, aromatic groups such as pyridines, imidazoles, and phenols have been reported to characterize CYP3A4 inhibitors.[43,44] In addition, charge descriptors, qnmax (9th), qpmax (14th), RPCG (16th), and RNCG (17th), are indicative of the involvement of polarizability in CYP3A4 inhibitors.[42] These observations suggest that increasing lipophilicity, aromaticity, and polarizability would enhance inhibitory activity.

Taken together, our approach is capable of identifying predictive features, as well as prioritizing features that are characteristic of CYP inhibition. Since the feature set used contains many features that are not easily interpretable, it is difficult to fully explain the relevance of the selected features. Nevertheless, predictive features may serve as markers for triaging compounds with desired affinities. In light of interpretability, more elaborate description of structural features of compounds, such as extended connectivity fingerprints[45] may be preferred to the Dragon descriptors. To explore the predictive ability of such fingerprints, we also tested ECFP6 and ECFC6 (as calculated by Pipeline Pilot[46]) for the CYP data set. However, we found that ECFP and ECFC were less predictive than the Dragon descriptors and that physicochemical and molecular properties of compounds are better suited to predict the binding affinities of CYP inhibitors.

Protein Feature Selection for Mutated HIV Protease Variants. The proposed algorithm was also evaluated on the mutation data of HIV protease, a major target for highly active antiretroviral therapy. The ability of the HIV virus to mutate and develop drug resistance by accumulating mutations severely hinders the treatment of HIV. To guide the design of new inhibitors that surmount the resistance, it is of great value to understand the mutational determinants involved in the interactions between inhibitors and HIV protease variants. The composite effects of distantly located mutations and the phenomenon of cross-resistance further motivate us to explore the mutational space of the protease in a comprehensive manner.[47]

As shown in Figure 2, the distributions of the pK_i values are wide-ranging to varying degrees and heavily overlap among one another. This suggests that simply estimating the value of a given protease variant to be the mean value for the ligand of interest is unsatisfactory ($q^2 = 0.21 \pm 0.09$) and that both target and ligand information is needed for accurate predictions.

Figure 5 and Table 4 show the q^2 values for KRR and SVR with the RBF kernel for compounds and the linear and RBF kernels for mutated HIV protease variants. Using all the protein features, KRR and SVR with the RBF kernel yielded $q^2 = 0.70$ and 0.78, respectively. A q^2 value of 0.78 is quite consistent with the best q^2 values of 0.78−0.83 reported in the study of Lapins andWikberg,[30] despite some differences in the data set and descriptors used.

Table 3. Top-Ranked Chemical Features of CYP Inhibitors[a]

rank	symbol	description	score
1	piPC09	molecular multiple path count of order 09	1.70
2	MATS1v	Moran autocorrelation−lag 1/weighted by atomic van der Waals volumes	5.15
3	Hypertens-80	Ghose−Viswanadhan−Wendoloski antihypertensive-like index at 80%	5.90
4	BIC0	bond information content (neighborhood symmetry of 0-order)	6.05
5	Infective-80	Ghose−Viswanadhan−Wendoloski antiinfective-like index at 80%	8.45
6	ALOGP2	Squared Ghose−Crippen octanol−water partition coeff (logP^2)	9.20
7	RARS	R matrix average row sum	9.60
8	G3s	third component symmetry directional WHIM index/weighted by atomic electrotopological states	13.25
9	qnmax	maximum negative charge	13.40
10	ARR	aromatic ratio	14.60
11	GATS3v	Geary autocorrelation−lag 3/weighted by atomic van der Waals volumes	15.35
12	MATS1p	Moran autocorrelation−lag 1/weighted by atomic polarizabilities	16.40
13	BEHp6	highest eigenvalue *n*. 6 of Burden matrix/weighted by atomic polarizabilities	16.80
14	qpmax	maximum positive charge	19.00
15	C-015	$=CH_2$	20.00
16	RPCG	relative positive charge	20.70
17	RNCG	relative negative charge	21.15
18	REIG	first eigenvalue of the R matrix	21.50
19	GATS1m	Geary autocorrelation−lag 1/weighted by atomic masses	22.35
20	R3e+	R maximal autocorrelation of lag 3/weighted by atomic Sanderson electronegativities	24.95
21	ALOGP	Ghose−Crippen octanol−water partition coeff (logP)	27.00
22	BLTF96	Verhaar model of Fish baseline toxicity for Fish (96 h) from MLOGP (mmol/L)	27.15
23	nBnz	number of benzene-like rings	28.55
24	Mor23v	3D-MoRSE−signal 23/weighted by atomic van der Waals volumes	28.95
25	SIC0	structural information content (neighborhood symmetry of 0-order)	30.05
26	BEHm7	highest eigenvalue *n*. 7 of Burden matrix/weighted by atomic masses	31.80
27	BEHv6	highest eigenvalue *n*. 6 of Burden matrix/weighted by atomic van der Waals volumes	32.70
28	GATS3p	Geary autocorrelation−lag 3/weighted by atomic polarizabilities	32.85
29	BLTA96	Verhaar model of Algae baseline toxicity for Algae (96 h) from MLOGP (mmol/L)	33.60
30	G3e	third component symmetry directional WHIM index/weighted by atomic Sanderson electronegativities	34.25

[a] The PROFEAT+RBF kernel was used for CYPs.

We then compared the performance between the proposed algorithm based on KDR and random selection with varying numbers of protein features. As the protein features were removed, the q^2 values for random selection dropped markedly. In contrast, the proposed algorithm successfully reduced the number of features to less than 30 while maintaining high q^2 values, clearly outperforming random selection. Indeed, in the case of SVR with the RBF kernel, the performance slightly drops from $q^2 = 0.78$ to 0.73 using the top 30 features of KDR, yet this value is significantly higher than $q^2 = 0.60$ obtained by random selection.

The selected protein features are those exhibiting high dependence on affinity values in the presence of chemical information, and the performance curve in Figure 5 suggests the biological relevance of the top 20−30 features. Table 5 lists the 20 top-ranked mutated positions with amino acid properties (z-scales). This list indicates that the most influential positions are 36, 48, 50, 63, 82, 84, and 90. Indeed, positions 48, 50, 82, and 84 are known as the active site of the protease.[47] It is thus likely that mutating these positions has a great effect on decreasing inhibitory activity of a group of inhibitors. The proposed algorithm is a multivariate approach and hence can detect composite effects of multimutations. Consistent with this, positions 48, 82, 84, and 90 have been identified as being interrelated with each other.[30] Interestingly, position 90 is located in the dimerization region of the protease, but such a distantly located mutation has been shown to prevent ligands from binding the protease by changing the geometry of the active site.[48] On the other hand, positions 36 or 63 are prone to natural genetic variations and may not by themselves confer resistance to inhibitors.[49] Our analysis identified them as informative, and this may be due to composite effects with other mutated positions, an observation that has also been suggested in the previous study.[30] For most of the top-ranked mutated positions, all the three amino acid properties seem to be relevant, but this is not the case for position 82. Specifically, 82 (z_1) and 82 (z_3) representing hydrophobicity and polarity were ranked the 14th and 9th, respectively, whereas 82 (z_2) representing steric properties was ranked the 48th and hence less relevant. This is also in good agreement with the previous study,[30] but the relevance of other top-ranked mutations such as 37 (z_2) and 71 (z_3) remains elusive. Overall, these results illustrate that the proposed feature selection approach serves as a useful tool not only for identifying informative chemical

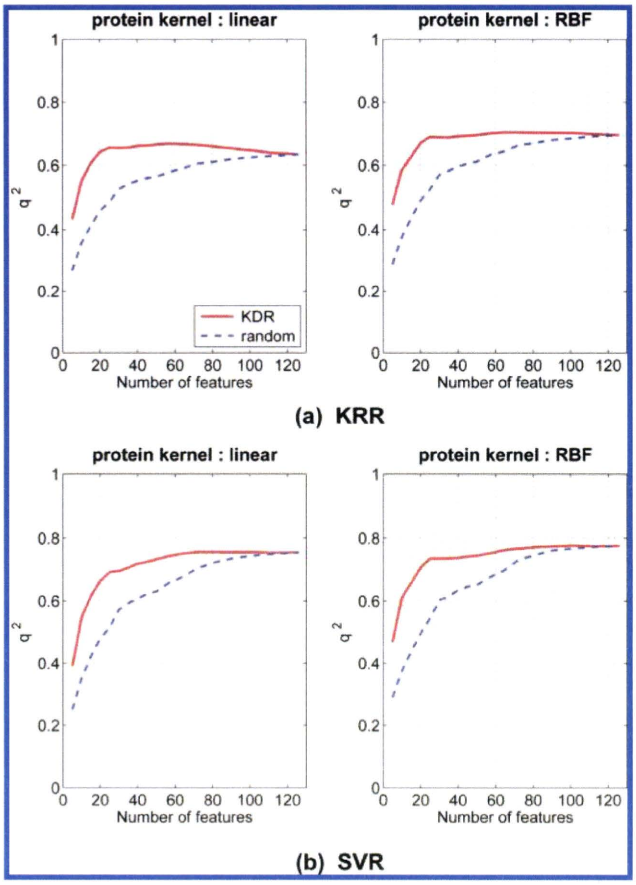


Figure 5. Average q^2 values as a function of the number of features. The RBF kernel was used for mutated HIV protease variants. (a) KRR with the linear and RBF kernels for compounds. (b) SVR with the linear and RBF kernels for compounds.

Table 4. Performance Comparison for Mutated HIV Protease Variants[a]

	KRR (protein kernel: linear)		KRR (protein kernel: RBF)	
no. of features	KDR	random	KDR	random
10	0.55 ± 0.08	0.36 ± 0.12	0.58 ± 0.08	0.38 ± 0.13
20	0.64 ± 0.07	0.46 ± 0.12	0.67 ± 0.07	0.49 ± 0.12
30	0.65 ± 0.06	0.53 ± 0.09	0.69 ± 0.06	0.57 ± 0.10
all (126)	0.63 ± 0.06		**0.70** ± 0.05	
	0.67 ± 0.06 (KDR, 50)		**0.70** ± 0.06 (KDR, 50)	
	SVR (protein kernel: linear)		SVR (protein kernel: RBF)	
no. of features	KDR	random	KDR	random
10	0.55 ± 0.11	0.35 ± 0.14	0.61 ± 0.08	0.37 ± 0.15
20	0.66 ± 0.08	0.48 ± 0.13	0.71 ± 0.07	0.50 ± 0.13
30	0.70 ± 0.07	0.57 ± 0.12	0.73 ± 0.06	0.60 ± 0.12
all (126)	0.75 ± 0.04		**0.78** ± 0.04	
	0.76 ± 0.05 (KDR, 72)		**0.78** ± 0.04 (KDR, 101)	

[a] Shown are the average q^2 values and standard deviations. The last row shows the best values in bold face (with method and the number of features in parentheses).

features but also for analyzing the effect of multimutations on their affinity to a series of inhibitors. Importantly, the selected positions and properties have implications for engineering new mutations at the same positions.

Table 5. Top-Ranked Protein Features of Mutated HIV Protease Variants

rank	mutated position (z-scale)	score	rank	mutated position (z-scale)	score
1	90 (z_3)	1.00	11	71 (z_3)	10.15
2	36 (z_3)	2.00	12	50 (z_2)	11.95
3	84 (z_1)	4.60	13	54 (z_3)	12.40
4	63 (z_3)	4.85	14	82 (z_1)	13.20
5	84 (z_3)	6.90	15	90 (z_2)	15.20
6	50 (z_3)	8.20	16	50 (z_1)	15.50
7	36 (z_2)	8.35	17	30 (z_3)	18.15
8	48 (z_3)	8.55	18	84 (z_2)	19.20
9	82 (z_3)	8.95	19	37 (z_3)	19.35
10	37 (z_2)	9.75	20	46 (z_2)	19.75

CONCLUSION

We have proposed an efficient feature selection algorithm based on KDR to identify molecular interaction features that contribute to prediction of ligand−target binding. Unlike existing feature selection techniques for chemoinformatics, our approach performs chemical (protein) feature selection coupled with protein (compound) information. In particular, the proposed algorithm works in a ligand−target space defined by kernels, allowing one to use various kernels for proteins (compounds) in selecting chemical (protein) features and, thus, provides a useful general approach to feature selection for chemogenomics.

The experiment on CYPs has shown that the algorithm is capable of identifying predictive features, as well as prioritizing features that are indicative of ligand preference for a given target family. Notably, using only the relevant features can lead to an improved performance. We have further illustrated the applicability on the mutation data of HIV protease by identifying influential amino acid positions within mutated variants. These results suggest that our feature selection approach based on the cross-target view can not only aid in drug design but also provide clues as to the molecular basis for ligand selectivity and off-target effects. We envision that this study will encourage further research in computational chemogenomics and contribute to a better understanding of the mechanism of molecular recognition.

ACKNOWLEDGMENT

This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology, Japan; the Targeted Proteins Research Program; and the New Energy and Industrial Technology Development Organization (NEDO) under the Ministry of Economy Trade and Industry of Japan. Financial support from Ono Pharmaceutical Co., Ltd., is also gratefully acknowledged.

REFERENCES AND NOTES

(1) Klabunde, T. Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. *Br. J. Pharmacol.* **2007**, *152*, 5–7.
(2) Rognan, D. Chemogenomic approaches to rational drug design. *Br. J. Pharmacol.* **2007**, *152*, 38–52.
(3) Bajorath, J. Computational analysis of ligand relationships within target families. *Curr. Opin. Chem. Biol.* **2008**, *12*, 352–358.
(4) Bock, J. R.; Gough, D. A. Virtual screen for ligands of orphan G protein-coupled receptors. *J. Chem. Inf. Model* **2005**, *45*, 1402–1414.
(5) Erhan, D.; L'Heureux, P.-J.; Yue, S. Y.; Bengio, Y. Collaborative filtering on a family of biological targets. *J. Chem. Inf. Model.* **2006**, *46*, 626–635.

(6) Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.

(7) Fröhlich, H.; Wegner, J. K.; Zell, A. Towards optimal descriptor subset selection with support vector machines in classification and regression. *QSAR Comb. Sci.* **2004**, *23*, 311–318.

(8) Byvatov, E.; Schneider, G. SVM-based feature selection for characterization of focused com- pound collections. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 993–999.

(9) Xue, Y.; Li, Z. R.; Yap, C. W.; Sun, L. Z.; Chen, X.; Chen, Y. Z. Effect of molecular de-scriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1630–1638.

(10) Schölkopf, B.; Smola, A. J. *Learning with Kernels: Support Vector Machines, Regulariza-tion, Optimization, and Beyond*; MIT Press: Cambridge, MA, 2002.

(11) Fukumizu, K.; Bach, F. R.; Jordan, M. I. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *J. Mach. Learn. Res.* **2004**, *5*, 73–99.

(12) Fukumizu, K.; Bach, F. R.; Jordan, M. I. Kernel dimensionality reduction in regression. *Ann. Stat.* **2009**, *37*, 1871–1905.

(13) Hawkins, D. M. The problem of overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1–12.

(14) Schnur, D. M.; Hermsmeier, M. A.; Tebben, A. J. Are target-family-privileged substructures truly privileged. *J. Med. Chem.* **2006**, *49*, 2000–2009.

(15) Schuffenhauer, A.; Jacoby, E. Annotating and mining the ligand–target chemogenomics knowledge space. *Drug Discovery Today* **2004**, *2*, 190–200.

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

(17) Jacob, L.; Vert, J.-P. Protein-ligand interaction prediction: an improved chemogenomics ap-proach. *Bioinformatics* **2008**, *24*, 2149–2156.

(18) Nagamine, N.; Sakakibara, Y. Statistical prediction of protein-chemical interactions based on chemical structure and mass spectrometry data. *Bioinformatics* **2007**, *23*, 2004–2012.

(19) Strömbergsson, H.; Daniluk, P.; Kryshtafovych, A.; Fidelis, K.; Wikberg, J. E.; Kley-wegt, G. J.; Hvidsten, T. R. Interaction model based on local protein substructures generalizes to the entire structural enzyme-ligand space. *J. Chem. Inf. Model* **2008**, *48*, 2278–88.

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

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

(22) Gretton, A.; Bousquet, O.; Smola, A. J.; Schölkopf, B. *Proceedings of the Sixteenth International Conference on Algorithmic Learning Theory*; Singapore, Oct 8–11; Springer: Berlin/Heidelberg, 2005; pp 63–78.

(23) Song, L.; Bedo, J.; Borgwardt, K. M.; Gretton, A.; Smola, A. Gene selection via the BAHSIC family of algorithms. *Bioinformatics* **2007**, *23*, i490–i498.

(24) Li, K.-C. Sliced inverse regression for dimension reduction. *J. Am. Stat. Assoc.* **1991**, *86*, 316–327.

(25) Yeh, Y.-R.; Huang, S.-Y.; Lee, Y.-J. Nonlinear dimension reduction with kernel sliced inverse regression. *IEEE Trans. Knowledge Data Eng.* **2009**, *21*, 1590–1603.

(26) Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **2002**, *46*, 389–422.

(27) Bach, F. R.; Jordan, M. I. Kernel independent component analysis. *J. Mach. Learn. Res.* **2002**, *3*, 1–48.

(28) Golub, G. H.; Loan, C. F. V. *Matrix Computations*, 3rd ed.; Johns Hopkins University Press: Baltimore, 1996.

(29) Kontijevskis, A.; Komorowski, J.; Wikberg, J. E. S. Generalized proteochemometric model of multiple cytochrome P450 enzymes and their inhibitors. *J. Chem. Inf. Model* **2008**, *48*, 1840–1850.

(30) Lapins, M.; Wikberg, J. E. S. Proteochemometric modeling of drug resistance over the mu-tational space for multiple HIV protease variants and multiple protease inhibitors. *J. Chem. Inf. Model* **2009**, *49*, 1202–1210.

(31) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.

(32) *DragonX*, version 1.2; Milano Chemometrics and QSAR Research Group: Milan, 2007.

(33) *MOE*, version 2008.10; Chemical Computing Group Inc.: Montreal, Canada, 2008.

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

(35) Leslie, C. S.; Eskin, E.; Cohen, A.; Weston, J.; Noble, W. S. Mismatch string kernels for discriminative protein classification. *Bioinformatics* **2004**, *20*, 467–476.

(36) Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J. Med. Chem.* **1998**, *41*, 2481–2491.

(37) Saunders, C.; Gammerman, A.; Vovk, V. *Proceedings of the Fifteenth International Conference on Machine Learning*; Wisconsin, July 24–27; Morgan Kaufmann Publishers, Inc.: San Francisco, 1998; pp 515–521.

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

(39) Chang, C.-C.; Lin, C.-J. LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

(40) de Groot, M. J. Designing better drugs: predicting cytochrome P450 metabolism. *Drug Discovery Today* **2006**, *11*, 601–606.

(41) Chohan, K. K.; Paine, S. W.; Mistry, J.; Barton, P.; Davis, A. M. A rapid computational filter for cytochrome P450 1A2 inhibition potential of compound libraries. *J. Med. Chem.* **2005**, *48*, 5154–5161.

(42) Kriegl, J. M.; Arnhold, T.; Beck, B.; Fox, T. Prediction of human cytochrome P450 inhibition using support vector machines. *QSAR Comb. Sci.* **2005**, *24*, 491–502.

(43) Jensen, B. F.; Vind, C.; Padkjær, S. B.; Brockhoff, P. B.; Refsgaard, H. H. F. In silico predic-tion of cytochrome P450 2D6 and 3A4 inhibition using Gaussian kernel weighted k-nearest neighbor and extended connectivity fingerprints, including structural fragment analysis of inhibitors versus noninhibitors. *J. Med. Chem.* **2007**, *50*, 501–511.

(44) Riley, R. J.; Parker, A. J.; Trigg, S.; Manners, C. N. Development of a generalized, quanti-tative physicochemical model of CYP3A4 inhibition for use in early drug discovery. *Pharm. Res.* **2001**, *18*, 652–655.

(45) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model* **2010**, *50*, 742–754.

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

(47) Lapins, M.; Eklund, M.; Spjuth, O.; Prusis, P.; Wikberg, J. E. S. Proteochemometric modeling of HIV protease susceptibility. *BMC Bioinformatics* **2008**, *9*, 181.

(48) Muzammil, S.; Ross, P.; Freire, E. A major role for a set of non-active site mutations in the development of HIV-1 protease drug resistance. *Biochemistry* **2003**, *42*, 631–638.

(49) Rhee, S.-Y.; Fessel, W. J.; Zolopa, A. R.; Hurley, L.; Liu, T.; Taylor, J.; Nguyen, D. P.; Slome, S.; Klein, D.; Horberg, M.; Flamm, J.; Follansbee, S.; Schapiro, J. M.; Shafer, R. HIV-1 protease and reverse-transcriptase mutations: correlations with antiretroviral therapy in subtype B isolates and implications for drug-resistance surveillance. *J. Infect. Dis.* **2005**, *192*, 456–465.

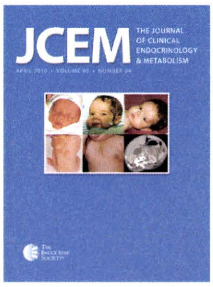
CI1001394

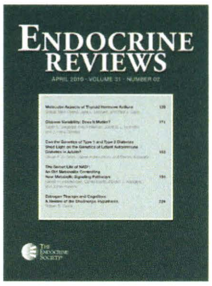
Endocrinology

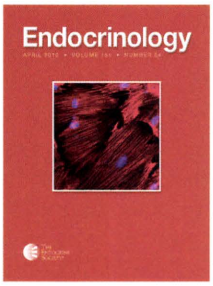
Abnormal Epithelial Cell Polarity and Ectopic Epidermal Growth Factor Receptor (EGFR) Expression Induced in Emx2 KO Embryonic Gonads

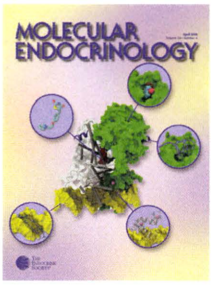
Masatomo Kusaka, Yuko Katoh-Fukui, Hidesato Ogawa, Kanako Miyabayashi, Takashi Baba, Yuichi Shima, Noriyuki Sugiyama, Yukihiko Sugimoto, Yasushi Okuno, Ryuji Kodama, Akiko Iizuka-Kogo, Takao Senda, Toshikuni Sasaoka, Kunio Kitamura, Shinichi Aizawa and Ken-ichirou Morohashi

Endocrinology 2010 151:5893-5904 originally published online Oct 20, 2010; , doi: 10.1210/en.2010-0915









Abnormal Epithelial Cell Polarity and Ectopic Epidermal Growth Factor Receptor (EGFR) Expression Induced in Emx2 KO Embryonic Gonads

Masatomo Kusaka, Yuko Katoh-Fukui, Hidesato Ogawa, Kanako Miyabayashi, Takashi Baba, Yuichi Shima, Noriyuki Sugiyama, Yukihiko Sugimoto, Yasushi Okuno, Ryuji Kodama, Akiko Iizuka-Kogo, Takao Senda, Toshikuni Sasaoka, Kunio Kitamura, Shinichi Aizawa, and Ken-ichirou Morohashi

Division for Sex Differentiation (M.K., Y.K.-F., H.O., K.M., T.B., Y.Sh., N.S., Y.Su., K.-i.M.), and Laboratory of Neurochemistry (T.Sa.), Center for Transgenic Animals and Plants, National institute for Basic Biology, National Institutes of Natural Sciences, Okazaki 444-8787, Japan; Department of Molecular Biology (K.M., T.B., Y.Sh., K.-i.M.), Graduate School of Medical Sciences, Kyushu University, Fukuoka 812-8582, Japan; Departments of Physiological Chemistry (Y.Su.) and Systems Biosciences for Drug Discovery (Y.O.), Graduate School of Pharmaceutical Sciences, Kyoto University, Kyoto 606-8501, Japan; Laboratory of Morphodiversity (R.K.), National Institute for Basic Biology, National Institutes of Natural Sciences, Okazaki 444-8585, Japan; Department of Anatomy I (A.I.K., T.Se.), School of Medicine, Fujita Health University, Toyoake 470-1192, Japan; Department of Mental Retardation and Birth Defect Research (K.K.), National Institute of Neuroscience, National Center of Neurology and Psychiatry, Kodaira, Tokyo 187-8502, Japan; and Laboratory for Vertebrate Body Plan (S.A.), Center for Developmental Biology, RIKEN, Kobe 650-0047, Japan

The gonadal primordium first emerges as a thickening of the embryonic coelomic epithelium, which has been thought to migrate mediodorsally to form the primitive gonad. However, the early gonadal development remains poorly understood. Mice lacking the paired-like homeobox gene *Emx2* display gonadal dysgenesis. Interestingly, the knockout (KO) embryonic gonads develop an unusual surface accompanied by aberrant tight junction assembly. Morphological and *in vitro* cell fate mapping studies showed an apparent decrease in the number of the gonadal epithelial cells migrated to mesenchymal compartment in the KO, suggesting that polarized cell division and subsequent cell migration are affected. Microarray analyses of the epithelial cells revealed significant up-regulation of *Egfr* in the KO, indicating that *Emx2* suppresses *Egfr* gene expression. This genetic correlation between the two genes was reproduced with cultured M15 cells derived from mesonephric epithelial cells. Epidermal growth factor receptor signaling was recently shown to regulate tight junction assembly through sarcoma viral oncogene homolog tyrosine phosphorylation. We show through *Emx2* KO analyses that sarcoma viral oncogene homolog tyrosine phosphorylation, epidermal growth factor receptor tyrosine phosphorylation, and *Egfr* expression are up-regulated in the embryonic gonad. Our results strongly suggest that *Emx2* is required for regulation of tight junction assembly and allowing migration of the gonadal epithelia to the mesenchyme, which are possibly mediated by suppression of *Egfr* expression. (*Endocrinology* 151: 5893–5904, 2010)

ISSN Print 0013-7227 ISSN Online 1945-7170
Printed in U.S.A.

doi: 10.1210/en.2010-0915 Received August 13, 2010. Accepted September 23, 2010.
First Published Online October 20, 2010

Abbreviations: BrdU, Bromodeoxyuridine; CCFSE, 5-(and-6)-carboxy-2′,7′-dichlorofuluorescein diacetate, succinimidyl ester; c-Src, sarcoma viral oncogene homolog; c-Yes, Yamaguchi sarcoma viral oncogene homolog 1; E, embryonic day; EGF, epidermal growth factor; EGFR, EGF receptor; Emx2, empty spiracles homeobox 2; Emx2(HA), HA-tagged EMX2; GATA, globin transcription factor; GATA4, GATA binding protein-4; HA, hemaglutinin; KO, knockout; *Lhx9*, *LIM homeobox gene 9*; PAR, partitioning-defective protein; PI, propidium iodide; S, sense; SEM, scanning electron microscopy; SF-1, steroidogenic factor-1; siRNA, short interfering RNA; *Sry*, sex-determining region Y chromosome; TUNEL, terminal deoxynucleotidyltransferase-mediated 2′-deoxyuridine 5′-triphosphate nick end labeling; ts, tail somite; WT1, suppressor gene for Wilms' tumor; ZO, zonula occludens.