


Computational Screening for Active Compounds Targeting Protein Sequences: Methodology and Experimental Validation

Fei Wang,[†] Dongxiang Liu,[†] Heyao Wang,[†] Cheng Luo,[†] Mingyue Zheng,[†] Hong Liu,[†] Weiliang Zhu,[†] Xiaomin Luo,[†] Jian Zhang,^{*,†} and Hualiang Jiang^{*,†,§}

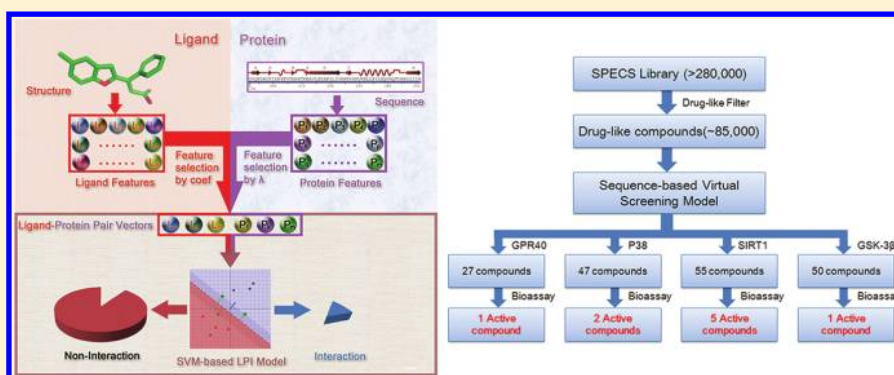
[†]Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 555 Zu Chong Zhi Road, Shanghai, 201203, China

[‡]Department of Pathophysiology, Key Laboratory of Cell Differentiation and Apoptosis of Chinese Ministry of Education, Shanghai Jiao-Tong University, School of Medicine, Shanghai, 200025, China

[§]School of Pharmacy, East China University of Science and Technology, Shanghai, 200237, China

 Supporting Information

ABSTRACT:



The three-dimensional (3D) structures of most protein targets have not been determined so far, with many of them not even having a known ligand, a truly general method to predict ligand–protein interactions in the absence of three-dimensional information would be of great potential value in drug discovery. Using the support vector machine (SVM) approach, we constructed a model for predicting ligand–protein interaction based only on the primary sequence of proteins and the structural features of small molecules. The model, trained by using 15 000 ligand–protein interactions between 626 proteins and over 10 000 active compounds, was successfully used in discovering nine novel active compounds for four pharmacologically important targets (i.e., GPR40, SIRT1, p38, and GSK-3 β). To our knowledge, this is the first example of a successful sequence-based virtual screening campaign, demonstrating that our approach has the potential to discover, with a single model, active ligands for any protein.

INTRODUCTION

Recent advances in the development of tools for virtual screening (VS) have demonstrated their efficiency in discovering potential lead compounds for drug development, and numerous targets have been deorphanized by using this technology.^{1,2} For all kinds of VS methods, the key challenge is to predict the ligand–protein interaction (LPI),^{3,4} which usually requires a knowledge of the three-dimensional (3D) structure of the target. However, the 3D structures of the overwhelming majority of drug targets, such as most of G protein-coupled receptors (GPCRs), ion channels, and other membrane proteins, are still unknown. Under these circumstances, ligand-based drug design (LBDD) approaches can be used involving pharmacophore modeling, molecular field analysis, 2D or 3D similarity assessment, provided that at least one bioactive structure or template is available as query reference. Unfortunately, many of the recently discovered

potential targets (e.g., most of the GPCRs and kinases) are orphan receptors, and it is therefore impossible to find corresponding bioactive reference compounds. The LBDD-VS approaches are not applicable for these systems.

To circumvent these issues, Laurent et al. proposed a new approach for LPI prediction designated *in silico* chemogenomics.^{5,6} This method attempts to predict ligands for a given target by leveraging binding information from other targets without considering 3D structural information for the given target. Yabuuchi et al. presented a chemical genomics-based virtual screening system to identify novel compounds binding to both G-protein-coupled receptors and protein kinases.⁷ The main limitation of this approach lies in the fact that it can only be applied to some

Received: June 14, 2011

Published: September 28, 2011

Table 1. Search Results of External Set

target	maximal identity ^a	known binders	decoy ligands	sampled binders ^b	sampled ligands ^c	EF ^d
CDC25B	<i>e</i>	830	45338	26	237	6.10
JSP-1	36%	218	45960	19	60	67.08
Glutamate (NMDA) receptor	<i>e</i>	1527	25000	55	90	10.62
Ornithine decarboxylase	<i>e</i>	194	25000	3	24	16.23
Muscarinic acetylcholine receptor	31%	1053	25000	22	233	2.34
α -1A adrenergic receptor	33%	525	25000	21	394	2.59
H ⁺ /K ⁺ ATPase	<i>e</i>	785	25000	6	156	1.26
Dopamine β -hydroxylase	<i>e</i>	100	25000	2	77	6.52
Prolyl endopeptidase	24%	315	25000	2	32	5.02

^a Maximal identity represents the highest sequence identity by Blastp with proteins used to train the LPI model. ^b Sampled binders denote true positive binders among the sampled ligands. ^c Sampled ligands denote ligands predicted to be positive binders. ^d EF denotes enrichment factor. EF = (sampled binders/sampled ligands)/(known binders/(known binders + decoy ligands)). ^e No significant hits found.

special subclasses of drug targets.^{5–9} Additionally, the reliability and practicability of in silico chemogenomics have not been experimentally tested. Therefore, a truly general model that could be used to predict ligand–protein interactions based solely on the sequence information of proteins and the structure of small molecules would provide a tool of great potential value in drug discovery.

To test the applicability of this paradigm for virtual screening, we constructed a general model that can predict the interactions for any given protein and small molecule. Verified by cross-validation and the retrieval of compounds for external protein targets in the MDL Drug Data Report (MDDR) and the National Center for Drug Screening of China (NCFSC), this model was applied to active compound discovery for four typical targets, GPR40,¹⁰ SIRT1,¹¹ P38 kinase,¹² and glycogen synthase kinase-3 β (GSK-3 β).¹³ Experimental assays showed that nine novel compounds identified by this way are active against these four targets. This indicates the potential of this strategy in active compound discovery.

MATERIALS AND METHODS

Data Set Construction. LPI data used in the present study were collected from the Binding Database (<http://www.bindingdb.org>, accessed Jan 2009),¹⁴ which contains 26 225 active ligands and 562 protein targets from various species. For the rational model, we used the following steps to ensure the quality of the data set in our selection of ligands and proteins. The protein should be a “normal” macromolecule with known function and ligands. Considering that the original proteins in the Binding Database are categorized by functionalities and species, the proteins used to construct the LPI model were carefully evaluated. The following proteins were removed from the original protein set: (i) without bound small organic ligands, (ii) point mutants, (iii) with nonstandard amino acids, DNA and RNA, and (iv) sequence length less than 50 amino-acid residues. Then, proteins with a sequence similarity threshold above 95% were removed using a Perl script.¹⁵ Finally, 626 protein sequences belonging to 95 source organisms were retrieved as protein candidates for the LPI model. The ligand chosen for the model should be a “druglike” organic compound and bind to the receptor noncovalently. Therefore, covalently bound ligands, complex ligands (such as heme), nonorganic ligands, and very small ligands (molecular weight <100 Da) were excluded from the original set, retaining 25 820 ligands suitable for our purpose.

The refined protein and ligand data sets were first separated into training set for fitting the LPI model and test set for assessing the performance of the LPI model. After the evaluation, the training and test sets were combined to form a final model for prediction of external data sets and use in the sequence-based virtual screening.

For an objective assessment of the LPI model, diverse ligands should be assigned to the test set. The 25 820 refined ligands were clustered by structural similarity into 4000 classes, and the central ligand in each class was extracted to build a diverse collection of ligand-training candidates. The second round of clustering was then performed on the remaining 21 820 ligands in the same way, leading to a larger group containing 6000 ligands as test candidates. Due to a class-imbalance trend in the real LPI space, a multiple-sampling strategy in the partition of training and test sets was used in the present study, in order to evaluate dependence in the imbalanced data set. Three sets of training and test data were created by the same procedure: (i) the 626 proteins were clustered into 30 classes using the Blast Perl script¹⁵ mentioned above, and one protein was randomly extracted from each class to assemble a collection of protein test candidates, with the remaining 596 proteins as training candidates; (ii) the known ligand–protein pairs (LPPs) from the protein training candidates and ligand training candidates were randomly selected as the positive points for the training set; (iii) the same number of negative LPPs for training was created by mismatching the positive LPPs in the training set; (iv) the known LPPs from the protein test candidates and ligand test candidates were randomly selected as the positive points for the test set; (v) the same number of negative LPPs was created by mismatching the positive LPPs in the test set. Finally, about 15 000 LPPs in three training sets and 1500 LPPs in three test sets were constructed.

Data sets for external validation were collected from the MDL Drug Data Report database (MDDR, version 2007.1) and the National Center for Drug Screening database (NCDS, version 2008). The MDDR comprises over 180 000 biologically relevant compounds in about 700 activity classes, and the NCDS data set contains about 45 000 active compounds within the National Center. Targets from the two databases were carefully evaluated by two criteria: (i) each target should have more than 50 active compounds, (ii) each target has distantly related (<40%) or unrelated (<30%) sequence identities with the proteins used in the building of LPI model, as shown in Table 1. Therefore, two targets from the NCDS and seven targets from the MDDR with

their active ligands were subject to external data for validation. All sequences of the proteins used in the external validations are listed in Supporting Information Table S1. The ligand decoys used in the targets from the NCDS are true negative samples from high-throughput screening, and the negative ligands of the targets from the MDDR were 25 000 molecules from the Maybridge library (<http://www.maybridge.com>, accessed May 2009) constructed by the clustering method.

Two important drug targets unrelated to proteins (<30%) used in the final LPI model were selected to test the practicability of the method, viz, GPR40 and SIRT1. Since bioassays for p38 and GSK-3 β are available in our institute, these two related targets were also included in the evaluation of our method. The SPECS library was pretreated by “druglike” filters encoded in the Pipeline Pilot (Accelrys, Inc.) and 191 407 molecules were finally retained for the screening. Finally, positive ligands predicted by the LPI model (80 for GPR40, 165 for SIRT1, 137 for p38, and 150 for GSK-3 β) were purchased and experimentally tested.

Computation of Similarities. Similarities of LPPs were estimated by the average of ligand similarities and protein identities. The sequence identity of each protein was calculated by using stand-alone BLAST software (<http://blast.ncbi.nlm.nih.gov>) with an E-value cutoff of 0.01. The similarity of each ligand was calculated by using the FP2 method encoded in Openbabel (<http://openbabel.sourceforge.net>, version 2.3.0), which is a path-based fingerprint index on small molecule fragments. Furthermore, the similarity distribution of LPPs between a training set and a test set was calculated by the collection of the maximum similarities between each LPP in the test set and all LPPs in the training set.

Bioassay for GPR40 Ligand Screening. For screening compounds affecting GPR40 function, both agonistic and antagonistic activities were tested. A stable cell line expressing human GPR40 was generated using human embryonic kidney 293 cells (HEK293), which was cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum (FBS; Hyclone Laboratories) at 37 °C under 5% CO₂. HEK293/GPR40 cells were seeded into a Matrigel-coated 96-well black plate at a density of 30 000 cells (100 μ L/well) and cultured overnight. Subsequently, the cells were loaded with 1X HDB Calcium dye-loading solution (100 μ L/well) at 37 °C for 60 min. Compounds were screened for agonistic and antagonistic activities in tandem in a single experiment. First, for the agonist assay, a test agonist compound was diluted with 1X Calcium Assay Solution A buffer and added into the plate automatically using a Flex-Station II³⁸⁴ (Molecular Devices, USA). Meanwhile, Ca²⁺ influx was measured by the instrument at excitation wavelength of 485 nm and emission wavelength 525 nm. Subsequently, for the antagonist assay, 10 μ L/well loading dye was discarded and 10 μ L/well of 10 \times test antagonist compound was added into the plate, and incubated for 10 min before agonist addition. The GPR40 agonist linolenic acid (LA) was diluted with 1X Calcium Assay Solution A buffer and added into the plate automatically. Meanwhile, Ca²⁺ influx was measured by the instrument at excitation wavelength of 485 nm and emission wavelength 525 nm.

SIRT1 Expression, Purification, and Screening Assay. DNA coding for the human SIRT1(156–664) fragment was cloned into the pET28a vector between NdeI and XhoI, and expressed in *E. coli* BL21(DE3) cells with a 6-histidine tag attached to the N-terminus. A single colony was cultured in LB medium containing 50 μ g/mL kanamycin at 37 °C until OD₆₀₀ reached 0.6–0.8, then induced with 0.5 mM isopropyl β -D-1-thiogalactopyranoside (IPTG) at 16 °C overnight. After centrifugation, the cell

pellet was suspended in lysis buffer (20 mM Tris, 200 mM NaCl, pH 8.0) and sonicated; the cell debris was removed and the supernatant loaded onto a Ni-NTA agarose column. The column was washed with 40 mM imidazole in lysis buffer and eluted with 500 mM imidazole; the eluate was centrifuged, dialyzed, and stored in 50 mM Tris, 100 mM NaCl, 1 mM dithiothreitol (DTT), and 10% glycerol, pH 8.0.

The 165 compounds were screened using a fluorescence-based method. The SIRT1 deacetylation reactions were performed in a black 96-well microplate (Cat. No. 655075, Greiner Bio-one) using 31.25 μ M acetyl-Arg-His-Lys-Lys(ϵ -acetyl)-AMC peptide and 750 μ M NAD⁺. The compounds were dissolved in DMSO; SIRT1 was added to compounds or DMSO wells in assay buffer (50 mM Tris, 137 mM NaCl, 2.7 mM KCl, and 1 mM MgCl₂, pH 8.0). The reactions were initiated by addition of the two substrates and run at room temperature for 45 min, then stopped with 10 mM nicotinamide and developed with 0.5 mg/mL trypsin at room temperature for 30 min. The fluorescence intensity was then read by a fluorometer using an excitation wavelength of 340 nm and an emission wavelength of 490 nm.

In-vitro Binding Assay by SPR for p38 and GSK-3 β . The binding affinities of compounds with p38 and GSK-3 β were assayed using a surface plasmon resonance-based biosensor instrument (Biacore 3000; Biacore AB, Uppsala, Sweden). The flow cell of the CM5 sensor chip (catalog no. BR-1000-14; Biacore AB) was activated by injecting a fresh mixture of equal volume of 0.2 M 1-ethyl-3,3-dimethylaminopropylcarbodiimide and 50 mM N-hydroxy-succinimide at 25 °C for 7 min. The p38 stock solution was diluted with 10 mM sodium acetate buffer (pH 4.13) and immobilized to the surface of the sensor chip. The GSK-3 β stock solution was diluted with 10 mM sodium acetate buffer (pH 3.72) and immobilized to the surface of the sensor chip. Excessive carboxyl groups were blocked by injecting 1 M ethanolamine hydrochloride at pH 8.5 for 7 min. Equilibration of the system was completed by continuous flow of HBS-EP buffer (10 mM HEPES, 150 mM NaCl, 3 mM EDTA, and 0.005% (v/v) Tween-20, pH 7.4) through the sensor chip overnight. For the screening assay, 10 mM stock solutions of compounds in DMSO were diluted in HBS-EP buffer to 10 μ M and flowed through the sensor chip at a speed of 30 μ L/min. To determine the affinity of compounds with the kinases, gradient concentrations of compounds were injected into the channel at a flow rate of 30 μ L/min for 120 s, followed by disassociation for 120 s. Binding affinities were calculated using the Biacore evaluation software, version 3.2 (GE Healthcare, Wisconsin, USA).

■ RESULT

Strategy of the Computational Approach. We began by constructing the LPI prediction model. Similar to our sequence-based prediction method of protein–protein interaction (PPI),¹⁶ the LPI prediction model was created based on a support vector machine (SVM) approach. The detailed procedure of SVM can be found in the original papers¹⁷ and in several recent applications.^{16,18–20} It is virtually axiomatic that “sequence specifies structure”, which gives rise to an assumption that knowledge of the amino-acid sequence alone might be sufficient to estimate the interacting propensity of a protein with small molecules.²¹ It might be possible to capture the LPI information for all proteins and a particular chemical space in a single model (ligand–protein interactome) by integrating sequence information of proteins,

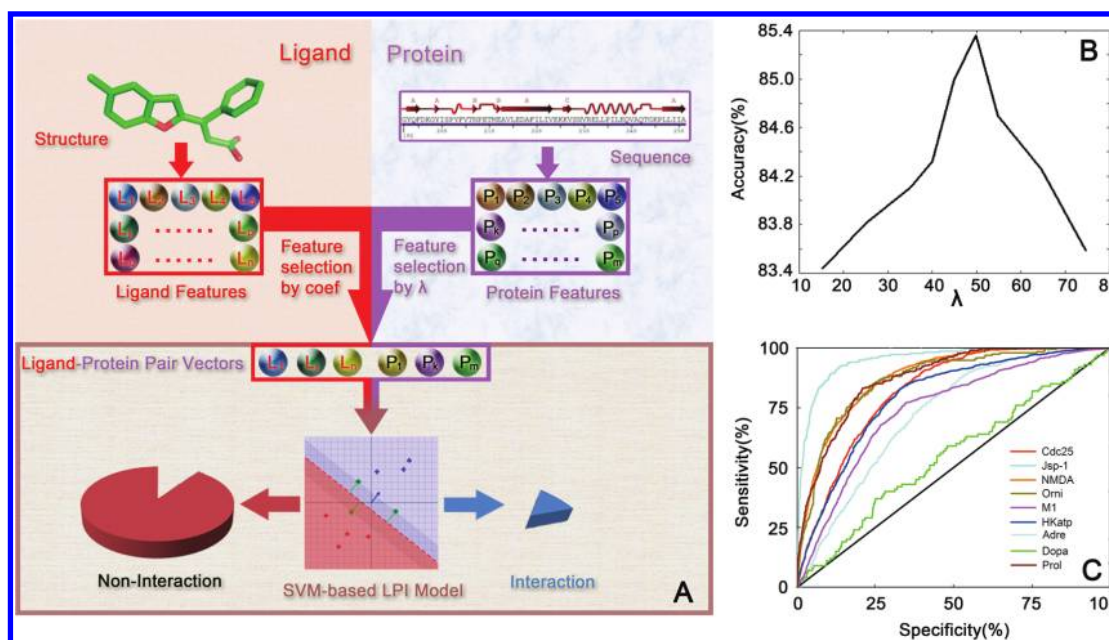


Figure 1. (A) Flowchart showing the derivation of LPP vectors and the SVM-based LPI model. (B) Plot showing the change of prediction accuracy of 5-fold cross-validation versus λ values. (C) ROC curves of LPI predictions for nine protein targets outside the training set.

structural information of small molecules, and LPI data (Figure 1A).

In principle, the properties of ligands and proteins are described by using molecular and sequence descriptors, respectively. Here, the descriptors of ligands were generated by Discovery Studio suites 2.1 (Accelrys, Inc.),²² including the 2D descriptors (e.g., Estate keys, Molecular properties, Molecular Property Counts, and Topological Descriptors) and the 3D descriptors (e.g., Dipole, Jurs Descriptors, Molecular properties, Principal Moments of Inertia, Shadow Indices, Surface Area, and Volume). Non-numerical and constant descriptors and the descriptors with high correlation coefficients (>0.95) were removed. Totally, 110 descriptors were finally selected to represent the ligands.

The protein descriptors were devised based on the pseudo amino-acid composition (PseAAC)²³ with T-scale²⁴ properties (described in Supporting Information Table S2). The PseAAC is a combination of a set of discrete sequence correlation factors and the content of the 20 conventional amino acids. The sequence order-correlated factor (θ_λ) of a protein with length of L residues, $R_1R_2R_3R_4R_5R_6R_7 \dots R_L$, is defined as

$$\theta_\lambda = \frac{1}{L-\lambda} \sum_{i=1}^{L-\lambda} \Theta(R_i, R_{i+\lambda}) \quad (\lambda < L) \quad (1)$$

The correlation function $\Theta(R_i, R_{i+\lambda})$ is given by

$$\Theta(R_i, R_{i+\lambda}) = \frac{\sum_{j=1}^N [P_j(R_i) - P_j(R_{i+\lambda})]^2}{N} \quad (2)$$

where λ is the length factor between two correlative residues for controlling the number of sequence descriptors, $P_j(R_i)$ is the j th property of residue R_i , and θ_λ is the λ th tier correlation factor that reflects the sequence order correlation between all the continuous residues along a protein sequence. The T-scale properties

defined by the components of principal component analysis (PCA) from 67 structure and topological variables²⁴ were used to represent the properties of the residues, which have been successfully introduced in the prediction of peptide modeling.²⁵ To calculate the content descriptors of the 20 amino acids (AA_i), the normalized occurrence frequency of each amino acid (PA_i) was defined as follows:

$$PA_i = \frac{AA_i - AA_{\min}}{AA_{\max} - AA_{\min}} \quad (3)$$

where AA_i is the occurrence frequency of A_i , and AA_{\max} and AA_{\min} represent the largest and the smallest occurrence frequency of each amino acid in the sequence, respectively.

Next, we formulated an LPI vector. The LPP descriptor was created to depict the interacting features between ligand and protein. LPP is a one-dimensional vector space concatenated by a variety of ligand features and protein sequence-derived features (Figure 1A). Different from the tensor product between target and ligand pairs used in other investigations,^{6,7,26} we represent the single interactive ligand-protein pairs (D_{lp}) by concatenating the vector spaces of the ligands (D_l) and proteins (D_p):

$$(D_{lp}) = (D_l) \oplus (D_p) \quad (4)$$

Thus, a 180-dimensional vector (110 from a ligand plus 70 from a protein sequence) was constructed to represent an LPI pair.

Finally, a prediction model was constructed by using LIBSVM software²⁷ based on a training data set. Basically, for a given data set $x_i \in R^n$ ($i = 1, \dots, N$) with corresponding labels y_i ($y_i = 1$ or 0, representing the two classes of interactive or noninteractive LPIs), SVM presents a decision function (classifier):

$$f(x) = \text{sgn}\left(\sum_{i=1}^N \gamma_i \alpha_i K(x, x_i) + b\right) \quad (5)$$

Table 2. Prediction Results of 10-cv and Test Sets

model	training set	test set			
	10-cv (%)	accuracy (%)	sensitivity (%) ^a	specificity (%) ^b	AUC
1	89.06	82.60	81.23	83.97	0.920
2	90.20	82.14	84.07	80.11	0.919
3	88.09	82.82	80.98	84.78	0.919

^a Sensitivity = true positive/(true positive + false negative). ^b Specificity = true negative/(true negative + false positive).

where α_i is the coefficient to be learned and K is a kernel function. Parameter α_i is trained through maximizing the Lagrangian expression given below:

$$\max_{\alpha_i} \left[\sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j \gamma_i \gamma_j K(x_i, x_j) \right] \quad (6)$$

Subject to $0 \leq \alpha_i \leq C$ ($i = 1, \dots, N$) and $\sum_{i=1}^N \alpha_i \gamma_i = 0$.

Method Training and Optimization. To avoid data dependency, a large data set including more than 15 000 diverse LPI pairs from the BindingDB database¹⁴ was refined to train the model. The sequence order-correlated factor θ_λ is essential in the model construction. This factor is dependent on the correlation distance λ and dominates the number of PseAAC descriptors (eq 1). To obtain the optimal value of θ_λ , the independent λ was investigated in the model training process. The statistical plot of prediction accuracy of 5-fold cross-validation versus λ values is shown in Figure 1B. The prediction accuracy profile reaches a maximum when λ is 50, which was therefore used to define the sequence descriptors in the PseAAC.

The performances of the SVM for classification are dependent on the combination of parameters: the capacity parameter C and kernel type K . C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing the training error. The RBF kernel, which has been successfully used in both bioinformatics and chemoinformatics,^{28,29} was selected to build up the LPI model of SVM in the present study. For the RBF kernel, γ dominates the generalization ability of SVM by regulating the amplitude of the kernel function (eq 5). Accordingly, both C and γ need to be optimized. The parameter optimization was performed by using a grid search approach within a limited range and the optimal values of C and γ for constructing SVM models are 8 and 0.25, respectively.

Method Validation. After having obtained the optimal λ for PseAAC and C and γ for SVM, we constructed an LPI prediction model based on the data set. To assess the predictive power of our model, we performed three validations: (i) cross-validation; (ii) retrieval of active ligands for external targets from external databases; (iii) active compounds discovery against four important drug targets.

First, to circumvent the problem of overfitting and data set dependency of the model, cross-validation tests were performed to evaluate its ability of extrapolation to other proteins and ligands. LPI prediction is a class-imbalance problem. Considering 100 positive LPIs, there are $100 \times 100 - 100 = 9900$ negative LPIs. Although many strategies have been tentatively proposed to improve the prediction accuracy of imbalanced data sets,^{30,31} there is no universally recognized representation scheme for this kind of problem. Combined with the mismatch approach of positive LPIs for negative LPIs, a multiple sampling strategy to partition the training/test sets could alleviate data dependence.

Therefore, three separate pairs of training/test sets were prepared by sampling methods described in Materials and Methods. Next, three predictive models were built based on three training sets for performance analysis. The results are listed in Table 2. The accuracies of the 10-fold cross-validation for three training sets are all >85%. Using the models derived from training data, the prediction accuracies on test sets are all >82%, and both the sensitivities and specificities are >80%. In addition, the fitting ability of the models can be reflected by the receiver operation characteristic (ROC) curves, in which the areas under the curves (AUC) are 0.920, 0.919, and 0.919 for three test tests, respectively (Table 2). To analyze data redundancy between training and test sets, the similarity distribution of LPPs between a training set and a test set was defined as the collection of the maximal similarities between each LPP in the test set and all LPPs in the training set. The distribution of LPPs in three sets of training and test data has been calculated and the result is shown in Supporting Information Figure S5. Although some redundancy has been found between the training sets and the test sets, almost half of LPPs actually have maximal similarity of less than 0.60 with those in corresponding training sets, indicating the moderately overlapping distribution of LPPs between the training and test sets. Furthermore, the extrapolation ability of our LPI model also has been revealed in subsequent tests (discussed below). Therefore, the high value of AUC in the test sets may derive not only from some redundancy of the data but from the extrapolation ability of our LPI model. Overall, our method may provide significant power in the prediction of LPI despite the potential errors of gathering a large amount of LPI data from different sources and experimental errors.

Second, our method was tested by retrieving active ligands for external targets from external small molecule databases (Table 1). The first training/test sets were combined to construct the final LPI model, and nine simulated orphan protein targets including enzymes, ion channels, and GPCRs were screened according to the protocol described in Materials and Methods. Among the targets, the low pairwise identities between 4.4% and 8.8% indicate that the external test set covers high structural diversity (described in Supporting Information Table S3). The predicted results are shown in Figure 1C, in which the ROC curves reveal better prediction abilities on all external targets than random guess represented by the diagonal line. Meanwhile, the enrichment factors (EF) of active ligands corresponding to their targets were also calculated (Table 1). EFs of active ligands in three targets are larger than 10 and those in six targets are larger than 5, indicating that our LPI method is capable of digging out active ligands for a protein only according to the information encoded in the protein sequence.

Experimental Validation in Active Compound Discovery. The most solid validation for a methodology is to test its applicability in concrete experiments. To our knowledge, LPI prediction methods based solely on protein sequences have not been reported in the application for discovering real active compounds. Thus, we finally applied our method to discover active compounds in four important targets, in order to test our model through experimental practice. The tested targets represent the diversity of target categories, including a G-protein-coupled receptor (GPR40), a deacetylase (SIRT1), and two kinases (p38 and GSK-3 β). These targets are also associated with a variety of diseases such as diabetes and Alzheimer's disease.

GPR40 is activated by saturated and unsaturated long- or medium-chain fatty acids in pancreatic β -cells,³² which may

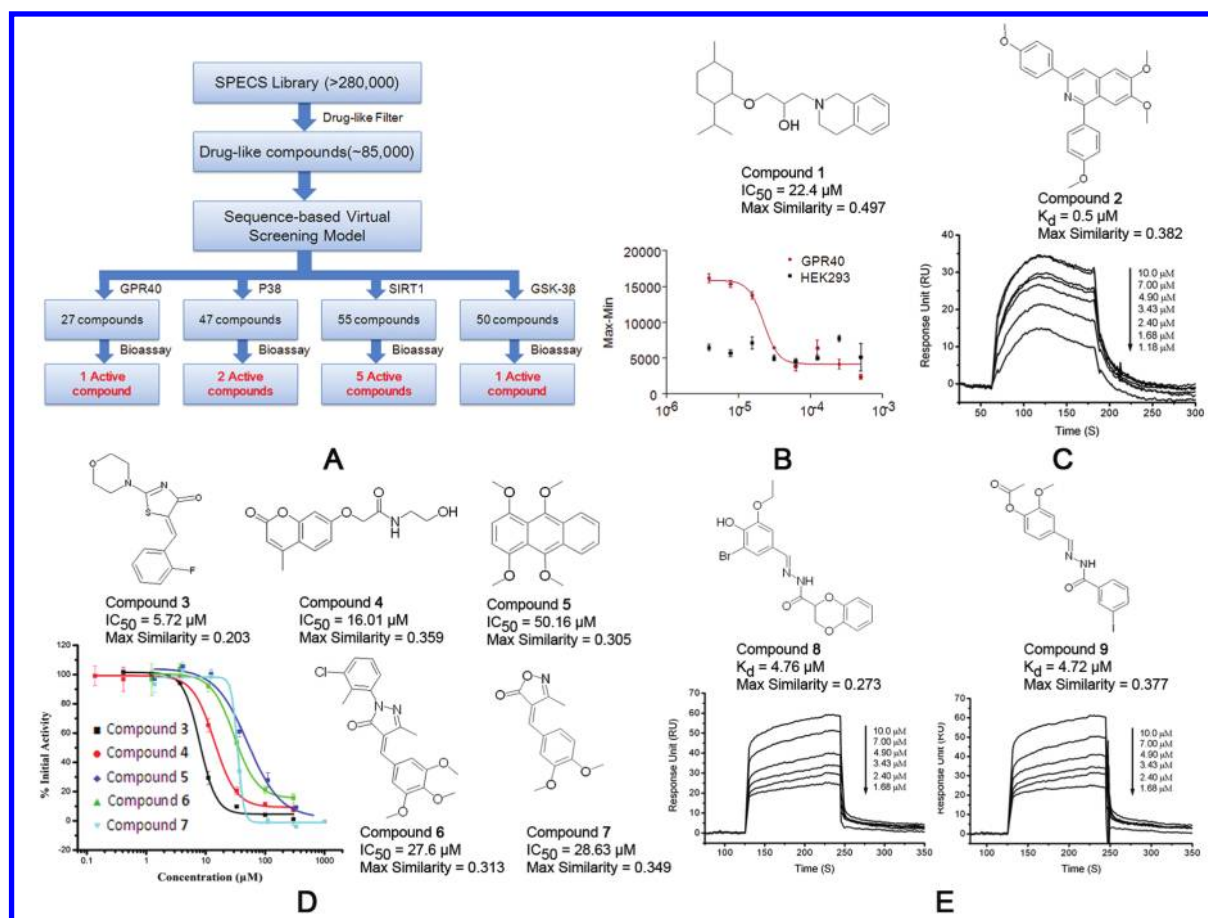


Figure 2. (A) Diagram showing the usages of the sequence-based LPI model in VS. (B–E) Nine active compounds that have been discovered for four pharmacologically important targets: (B) one GPR40 antagonist, (C) five SIRT1 inhibitors, (D) two p38 binders, and (E) one GSK-3 β binder.

mediate both acute and chronic effects of the free fatty acids on insulin secretion and has been recognized as a potential target in both obesity and diabetes.¹⁰ SIRT1 functions as a NAD⁺-dependent deacetylase;¹¹ it mediates the transfer of the acetyl group of an acetylated protein substrate and is involved in the expression of genes involved in cell survival, metabolism, proliferation, and differentiation through deacetylation of histones, transcription factors, and transcriptional cofactors.³³ p38 is a member of the mitogen-activated protein kinase family¹² and can phosphorylate transcription regulators ATF2, MEF2C, and MAX, cell cycle regulator CDC25B, and tumor suppressor p53.³⁴ GSK-3 β is a serine/threonine kinase that regulates cell proliferation, apoptosis, and stem-cell maintenance by the phosphorylation of a broad range of substrates,¹³ making GSK-3 β a candidate therapeutic target in leukemia and inflammatory diseases.³⁵ Actually, all these targets were just deorphanized a few years ago, and drug discovery on the targets is still at the beginning (inhibitors or activators). In addition, the 3D structures of GPR40 and SIRT1 are unknown. Accordingly, discovering new ligands for these targets is a challenge for current computational drug design methods.

SPECS (<http://www.specs.net>, accessed Jan 2010) contains structural information for ~280 000 chemical compounds. The SPECS database was pretreated by a “rule-of-5” drug-likeness filter, and 191 407 molecules were finally retained for virtual screening (Figure 2A). Targeting the four proteins, the treated SPECS database was screened. The LPI model selected 80, 165,

Table 3. Search Results of Experimental Set

target	maximal identity ^a	novel binders	maximal similarity ^b
GPR40	26%	1	0.68
SIRT1	c	5	0.59
p38	76%	2	0.59
GSK-3 β	86%	1	0.69

^a Maximal identity represents the highest sequence identity by Blastp with proteins used to train the LPI model. ^b Maximal similarity denotes the highest structural similarity with ligands used to train the final LPI model by the FP2 method encoded in Openbabel. ^c No significant hits found.

137, and 150 positive ligands for GPR40, SIRT1, p38, and GSK-3 β , respectively (Figure 2A). We purchased all these compounds and their activities were determined in our in-house experimental screening assays. Among the 522 compounds, nine active compounds have been identified, including one GPR40 antagonist with an EC_{50} value of $22.4 \mu M$, five SIRT1 inhibitors with IC_{50} values ranging from 5 to $50 \mu M$, two p38 binders with dissociation constant (K_d) at the μM level, and one GSK-3 β binder with K_d of $0.5 \mu M$ (Figure 2B–E). These validation results show that our LPI model is applicable in discovering active compounds.

To comprehensively evaluate the predictive ability of our method, we searched the literature and patent databases and found active compounds published for these four targets.

Their typical structures are shown in Supporting Information Figures S1–S4. The structural similarities of the nine active compounds with the respective existing active compounds have been calculated. The result shows that our compounds share low max-similarities (<0.5) with the existing active compounds (Figure 2). In addition, these nine active compounds have also been analyzed for their structural similarities to compounds used to train the final LPI model, resulting in max-similarities between 0.59 and 0.69 (Table 3). These results indicate that our model could be used to find novel biological ligands chemically quite distinct from both existing active ligands and training compounds.

For comparison, docking-based and pharmacophore-based virtual screening was also performed respectively on GPR40 and SIRT1 in parallel. Unfortunately, pharmacophore-based screening on SIRT1 failed to discover active compounds. For GPR40, a 3D structural model constructed by using homology modeling was targeted by screening the SPECS database employing docking approach. Through the experimental bioassay, one sulfonamide antagonist (compound 34 in Supporting Information Figure S1) was discovered from the 218 candidates predicted by the docking calculation.³¹ From these two special examples, we might conclude that the sequence-based virtual screening approaches indeed have advantages in case of targets without enough available information on ligands (e.g., SIRT1) or without accurate 3D structures (e.g., GPR40). In addition, in comparison with the docking approach, which requires homology modeling, docking simulation, and sometimes molecular dynamics simulation for taking into account the flexibility of targets, our method is computationally effective.

DISCUSSION

In summary, we developed a statistical model that predicts ligand–protein interactions with reasonable accuracy based on primary sequences of proteins and structural characteristics of small ligands. Remarkably, we discovered novel active compounds against four important potential drug targets by using our model. To our knowledge, this is the first example for the sequence-based virtual screening approach to be applied in discovering real active compounds.

Unrelated or distantly related proteins are necessary to validate the generality of the model. Actually, we did carefully select both external and experimental sets on the validation. Among 13 systems, 11 targets were distantly related (<36%) and 6 targets were unrelated (<20%) to proteins used to train the final LPI model by the evaluation of Blastp (Tables 1 and 3). Furthermore, unrelated/distantly related LPPs in the external validation have been shown in Supporting Information Figure S6, in which more than 90% positive LPPs in the external and experimental sets share less than 0.45 maximum similarity with LPPs used to train the final LPI model. Overall, both of the assessments indicate the generality of our LPI model. As to the applicable types of proteins, successful retrieval of active compounds and discovery of novel compounds have been achieved in a spectrum of proteins including enzymes (CDC25B, JSP-1, dopamine β -hydroxylase, and prolyl endopeptidase), ion channels (H^+/K^+ ATPase and NMDA receptor), and GPCRs (Muscarinic acetylcholine receptor, α -1A adrenergic receptor, and GPR40), suggesting our method is universally applicable to all proteins as long as they comprise more than 50 residues. Moreover, this approach can be applied in various fields. First, this method is able to predict active ligands for targets without 3D structures and/or known ligands

(orphan receptors), which applies for most targets. Second, the computational efficiency of the method provides a convenient way for screening very large databases simultaneously targeting numerous proteins or even whole genomes. Third, because of the second advantage, this method could be used to identify possible binding proteins for an active compound or existing drug, which is an important issue for target identification.^{36–38}

ASSOCIATED CONTENT

S Supporting Information. Protein sequences for external and experimental tests, existing active compounds for experimental targets, and figures describing the similarity distributions of LPPs. This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*Phone: +86-21-50807188 Fax: +86-21-50807708 E-mail: hljiang@mail.shcnc.ac.cn (H.J.). Phone: +86-21-63846590-776922. Fax: +86-21-64154900. E-mail: jian.zhang@sjtu.edu.cn (J.Z.).

ACKNOWLEDGMENT

We thank Prof. Rolf Hilgenfeld at University of Lübeck for the polish language assistance. We gratefully acknowledge financial support from the National Natural Science Foundation of China (21021063, 91029704, 21002062, and 207201020402), the State Key Program of Basic Research of China (2009CB918502), the Shanghai Pujiang Program (10PJ406800), and the National S&T Major Project (2009ZX09501-001 and 2009ZX09301-001)

REFERENCES

- (1) Ripphausen, P.; Nisius, B.; Bajorath, J. State-of-the-art in ligand-based virtual screening. *Drug Discovery Today* **2011**, *16*, 372–376.
- (2) Ripphausen, P.; Nisius, B.; Peltason, L.; Bajorath, J. Quo vadis, virtual screening? A comprehensive survey of prospective applications. *J. Med. Chem.* **2010**, *53*, 8461–8467.
- (3) Joseph-McCarthy, D. Computational approaches to structure-based ligand design. *Pharmacol. Ther.* **1999**, *84*, 179–191.
- (4) Leach, A. R.; Gillet, V. J.; Lewis, R. A.; Taylor, R. Three-Dimensional Pharmacophore Methods in Drug Discovery. *J. Med. Chem.* **2010**, *53*, 539–558.
- (5) Jacob, L.; Vert, J.-P. Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics* **2008**, *24*, 2149–2156.
- (6) Jacob, L.; Hoffmann, B.; Stoven, V.; Vert, J.-P. Virtual screening of GPCRs: An in silico chemogenomics approach. *BMC Bioinformatics* **2008**, *9*, 363.
- (7) Yabuuchi, H.; Nijima, S.; Takematsu, H.; Ida, T.; Hirokawa, T.; Hara, T.; Ogawa, T.; Minowa, Y.; Tsujimoto, G.; Okuno, Y. Analysis of multiple compound-protein interactions reveals novel bioactive molecules. *Mol. Syst. Biol.* **2011**, *7*, 472.
- (8) Geppert, H.; Humrich, J.; Stumpfe, D.; Gärtner, T.; Bajorath, J. Ligand prediction from protein sequence and small molecule information using Support Vector Machines and fingerprint descriptors. *J. Chem. Inf. Model.* **2009**, *49*, 767–779.
- (9) 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.
- (10) Itoh, Y.; Kawamata, Y.; Harada, M.; Kobayashi, M.; Fujii, R.; Fukusumi, S.; Ogi, K.; Hosoya, M.; Tanaka, Y.; Uejima, H.; Tanaka, H.; Maruyama, M.; Satoh, R.; Okubo, S.; Kizawa, H.; Komatsu, H.; Matsumura,

F.; Noguchi, Y.; Shinohara, T.; Hinuma, S.; Fujisawa, Y.; Fujino, M. Free fatty acids regulate insulin secretion from pancreatic β cells through GPR40. *Nature* **2003**, *422*, 173–176.

(11) Zschoernig, B.; Mahlknecht, U. SIRTUIN 1: Regulating the regulator. *Biochem. Biophys. Res. Commun.* **2008**, *376*, 251–255.

(12) Yong, H. Y.; Koh, M. S.; Moon, A. The p38 MAPK inhibitors for the treatment of inflammatory diseases and cancer. *Exp. Opin. Invest. Drugs*. **2009**, *18*, 1893–1905.

(13) Phukan, S.; Babu, V. S.; Kannoji, A.; Hariharan, R.; Balaji, V. N. GSK3 β : role in therapeutic landscape and development of modulators. *Br. J. Pharmacol.* **2010**, *160*, 1–19.

(14) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198–201.

(15) Holm, L.; Sander, C. Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics* **1998**, *14*, 423–429.

(16) Shen, J.; Zhang, J.; Luo, X.; Zhu, W.; Yu, K.; Chen, K.; Li, Y.; Jiang, H. Predicting protein–protein interactions based only on sequences information. *Proc. Nat. Acad. Sci.* **2007**, *104*, 4337–4341.

(17) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: New York, 1999.

(18) Zheng, M.; Liu, Z.; Xue, C.; Zhu, W.; Chen, K.; Luo, X.; Jiang, H. Mutagenic probability estimation of chemical compounds by a novel molecular electrophilicity vector and support vector machine. *Bioinformatics* **2006**, *22*, 2099–2106.

(19) Zheng, M.; Luo, X.; Shen, Q.; Wang, Y.; Du, Y.; Zhu, W.; Jiang, H. Site of Metabolism (SOM) Prediction for six biotransformations mediated by cytochromes P450. *Bioinformatics* **2009**, *25*, 1251–1258.

(20) Chen, X.; Jeong, J. C. Sequence-based prediction of protein interaction sites with an integrative method. *Bioinformatics* **2009**, *25*, 585–591.

(21) Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **1973**, *81*, 223–230.

(22) *Accelrys Discovery Studio 2.1*; Accelrys, San Diego, CA, 2006.

(23) Chou, K. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins: Struct., Funct., Genet.* **2001**, *43*, 246–255.

(24) Tian, F.; Zhou, P.; Li, Z. T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides. *J. Mol. Struct.* **2007**, *830*, 106–115.

(25) Zhang, G.; Li, H.; Gao, J.; Fang, B. Prediction of lipases types by different scale pseudo-amino acid composition. *Chin. J. Biotechnol.* **2008**, *24*, 1968–1974.

(26) 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.

(27) Chang, C. C.; Lin, C. J. LIBSVM: a library for support vector machine, version 2.8; <http://www.csie.ntu.edu.tw/~cjlin/libsvm>, 2009.

(28) Ben-Hur, A.; Ong, C. S.; Sonnenburg, S.; Schölkopf, B.; Rätsch, G. Support Vector Machines and kernels for computational biology. *PLoS Comput. Biol.* **2008**, *4*, No. e1000173.

(29) Ivanciuc, O. *Reviews in Computational Chemistry*; Lipkowitz, K. B., Cundary, T. R., Eds.; Wiley-Vch, John Wiley & Sons, Inc: Weinheim, 2007; Vol. 23; p 110.

(30) Dohkan, S.; Koike, A.; et al. Improving the performance of an SVM-based method for predicting protein-protein interactions. *In Silico Biol.* **2006**, *6*, 515–529.

(31) Tasadduq, I.; Kai, T.; Joarder, K. z-SVM: An SVM for improved classification of imbalanced data. *Proceedings of the 19th Australian Joint Conference on Artificial Intelligence (AJCAI 2006)*, Hobart, Australia; Springer Press, 2006; 264–273.

(32) Hu, H.; He, L. Y.; Gong, Z.; Li, N.; Lu, Y. N.; Zhai, Q. W.; Liu, H.; Jiang, H. L.; Zhu, W. L.; Wang, H. Y. A novel class of antagonists for the FFAs receptor GPR40. *Biochem. Biophys. Res. Commun.* **2009**, *390*, 557–563.

(33) Guarani, V.; Potente, M. SIRT1 - a metabolic sensor that controls blood vessel growth. *Curr. Opin. Pharmacol.* **2010**, *10*, 139–145.

(34) Coulthard, L. R.; White, D. E.; Jones, D. L.; McDermott, M. F.; Burchill, S. A. p38(MAPK): stress responses from molecular mechanisms to therapeutics. *Trends Mol. Med.* **2009**, *15*, 369–379.

(35) Song, E. Y.; Palladinetti, P.; Klammer, G.; Ko, K. H.; Lindeman, R.; O'Brien, T. A.; Dolnikov, A. Glycogen synthase kinase-3 β inhibitors suppress leukemia cell growth. *Exp. Hematol.* **2010**, *38*, 908–921.

(36) Chen, Y.; Ung, C. Y. Prediction of potential toxicity and side effect protein targets of a small molecule by a ligand-protein inverse docking approach. *Mol. Graph. Mod.* **2001**, *20*, 199–218.

(37) Li, H.; Gao, Z.; Kang, L.; Zhang, H.; Yang, K.; Yu, K.; Luo, X.; Zhu, W.; Chen, K.; Shen, J.; Wang, X.; Jiang, H. TarFisDock: a web server for identifying drug targets with docking approach. *Nucleic Acids Res.* **2006**, *34*, 219–224.

(38) Campillos, M.; Kuhn, M.; Gavin, A. C.; Jensen, L. J.; Bork, P. Drug target identification using side-effect similarity. *Science* **2008**, *321*, 263–266.