Published in final edited form as:

Int J Comput Biol Drug Des. 2011 ; 4(3): 274–289. doi:10.1504/IJCBDD.2011.041415.

LigSeeSVM: Ligand-Based Virtual Screening Using Support Vector Machines and Information Fusion

Yen-Fu Chen¹, Kai-Cheng Hsu¹, Po-Tsun Lin¹, D. Frank Hsu^{3,*}, Bruce S. Kristal⁴, and Jinn-Moon Yang^{1,2,*}

¹Institute of Bioinformatics, National Chiao Tung University, Hsinchu, 30050, Taiwan

²Department of Biological Science and Technology, National Chiao Tung University Hsinchu, 30050, Taiwan

³Department of Computer and Information Science, Fordham University, New York, NY 10023, USA

⁴ Department of Neurosurgery, Brigham and Women's Hospital, Boston, MA ,02115, USA and Department of Surgery, Harvard Medical School, Boston, MA, 02115, USA

Abstract

Structure-based virtual screening of molecular compound libraries is a potentially powerful and inexpensive method for the discovery of novel lead compounds for drug development. That said, virtual screening is heavily dependent on detailed understanding of the tertiary or quaternary structure of the protein target of interest, including knowledge of the relevant binding pocket. Such information is often difficult or impossible to obtain. In contrast, ligand-based *in silico* drug screening can perform screening on drug targets whose three-dimensional structure is not yet determined. We previously demonstrated that the use of data fusion techniques strengthen virtual screening. Building on this data, we here describe LigSeeSVM, a ligand-based screening tool using data fusion and Support vector machines and termed. We combine atom pair (AP) structure descriptors and physicochemical (PC) descriptors to characterize compounds' features. We used SVM to generate SVM-AP model based on 825 AP descriptors and SVM-PC model based on 19 physicochemical descriptors. We combine SVM-AP and SVM-PC using rank-based information fusion to create LigSeeSVM model. LigSeeSVM was evaluated on five data sets, including thymidine kinase (TK) substrates, estrogen receptor (ER) antagonists, estrogen receptor agonists (ERA), GPCR and GABA_A ligands. Our results suggest that LigSeeSVM is useful for ligand-based virtual screening and offers competitive performance to other ligand-based screening approaches.

1. Introduction

Computational screening of compound databases recently has become increasingly popular in pharmaceutical research. The growing interest reflects the potential to reduce time and costs novel, potential inhibitors for diseases. The computational approaches used for virtual screening can be classified into two categories: structure-based virtual screening and ligand-based virtual screening. For ligand-based methods, the strategy is to use information provided by a compound or set of compounds that are known to bind to the desired target and to use this to identify other compounds in external databases with similar properties[12].

* Corresponding Authors moon@cc.nctu.edu.tw; hsu@cis.fordham.edu.

The applications of structure-based virtual screening approaches relying on a detailed three-dimensional model of the receptor binding pocket[15], but there are important drug targets whose three-dimensional structures are not sufficiently well characterized to permit structure-based virtual screening[7]. For example, membrane spanning G-protein-coupled receptors (GPCRs) or ion channels were the targets for nine of the top 20 selling prescription drugs worldwide in the year 2000, but 3D structures are unavailable for most GPCRs and ion channels[7,14]. Therefore, we sought to address this deficiency by building an entirely ligand-based approach to GPCRs and GABAA receptors.

A variety of molecular descriptors and methods have been developed and routinely used for describing physicochemical and structural properties of chemical agents[8,9]. These include both 2D and 3D methods. Most of the 2D methods are based upon structural indices. Although these structural indices represent different aspects of molecular structures, their physicochemical meaning is unclear, and they cannot distinguish stereoisomers[21]. A major benefit of 2D methods is that these methods do not require either conformational searches or structural alignment. Accordingly, 2D methods are easily automated and adapted to database searching, and/or virtual screening[16]. The major molecular descriptors used in this work are derived from 2D molecular topology (825 different atom pair descriptors)[21]. To complement this approach, and to help compensate for the potential weaknesses of 2D screening approaches, we also utilized a second algorithm that encompasses information from physicochemical descriptors derived from Accelrys Cerius2 QSAR module with 6 thermodynamic and 13 default descriptors[1].

Support vector machines (SVMs) have been applied to a wide range of pharmacological and biomedical problems including drug-likeness, drug blood-brain barrier penetration prediction and others[18,20]. Here, we used LibSVM 2.71 developed by Lin et al.[4], and the information fusion technique, called Combinatorial Fusion Analysis (CFA)[5], developed for virtual database screening, protein structure prediction, information retrieval and target tracking by Hsu et al.[5,6,10,13,19].

When LigSeeSVM obtained 100% for the recall, the false positive rates were 0.3% for TK, 0.6% for ER antagonists, and 0% for ERA. The ROC curves of GPCR and GABA_A screening sets shows that the performance of the LigSeeSVM is better than other ligand-based virtual screening approaches. The results of this study suggests that our approach, utilizing SVMs and methods of combination, can be explored as a general virtual screening and drug discovery tool and applied to a large variety of available datasets of biologically active compounds.

2. Material and Methods

We describe the data sets having used in our study and the features extracted from the data sets. Then we describe the prediction model LigSeeSVM. Figure 1 shows the framework and flowchart of our LigSeeSVM for ligand-based virtual screening.

2.1 Data sets

A. Thymidine kinase substrates, estrogen receptor antagonists and agonists

—Two screening sets designed for virtual screening against TK and ER α receptor of antagonist form were proposed by Bissantz *et al.* [2] in 2000 and retested by Jain in 2004 [7] (Fig. 2). A TK library contained 10 known ligands of TK and 990 randomly chosen molecules from the ACD; an ER α -antagonists library contained 11 known antagonists of ER α and 990 molecules used in the TK library. The files in the SYBYL mol2 format were converted to the MDL mol format with Corina 3.0 for running Accelrys Cerius² QSAR module. In addition, the screening set for ER α agonists included 10 known agonists [11] and

990 molecules used in the TK library. For each target, the active compounds in the box (Fig. 2) and 40 random compounds selected from 990 molecules are considered as a training set, and the other compounds are considered as an independent test set.

B. GPCR and GABA_A ligands—Ligands of serotonin, muscarinic, histamine, and GABA_A receptors (Fig. 3) [7] with 40 randomly chosen molecules from the 990 molecules were used for SVM training set. Compounds A and B act on serotonin receptor; Compounds C, D, and E act as muscarinic antagonists; Compounds F—H are histamine antagonists; and Molecules I—K are GABA_A agonists. In order to validate the SCM models, additional 85 GPCR ligands and 15 GABA_A agonists are included as testing sets and these molecules present widely varying structures (See Appendix 1) which proposed by Jain[7]. The 950 ACD used in the TK library are considered as the negative compounds for all cases.

2.2 Feature extraction

A. Atom pair descriptors—Atom pair descriptors (AP) were obtained with the AP generator program developed in our laboratory using the method proposed by Carhart *et al.* [3]. An AP descriptor is defined in terms of the atom types and the shortest graph distance between two atoms. The general form of an atom pair is defines as atom type i - (distance) - atom type j , where (distance) is the graph distance between atoms i and j .

In this study, we have clustered 23 SYBYL atom types into 10 AP atom types (See Appendix 2A). The total number of pairwise combinations of all 10 atom types is 55. Furthermore, 15 distance bins were defined in the interval of graph distance ranging from zero to 14. Thus, a total of 825 (55×15) atom pair descriptors were generated for each molecular structure.

B. Physicochemical descriptors—The physicochemical (PC) descriptors of a compound are generated based on the Accelrys Cerius² QSAR module. In this study, we used two functional families of descriptors, six descriptors in the thermodynamic family (See Appendix 2B) and 13 descriptors in the default family (See Appendix 2C).

2.3 Build LigSeeSVM prediction model

A. Divide data set into training set and testing set—Molecular similarity and cluster analysis are used to choose the positive examples from the known ligands. The similarity of compounds is measured by Euclidean distance using the above descriptors X . The distance $d(c_i, c_j)$ between any compound pair (c_i and c_j) in N -dimensional descriptor space was calculated as

$$d(c_i, c_j) = \sqrt{\sum_{n=1}^N (X_{ci}^n - X_{cj}^n)^2}$$

where X_{ci}^n and X_{cj}^n are the values of n^{th} descriptor for compounds c_i and c_j , respectively. Based on UPGMA algorithm[17], we created a rooted tree of the known ligands. Compounds with the maximum distance (most dissimilarity) are chosen as the positive examples in the training set. Figure 2 shows the known active ligands of TK and ER α , and the compounds in the box are selected as positive cases. In this training set, the ratio of the number of negative and positive data is about 20. For example, the SVM training set consists of two or three known ligands as positive examples and 40 randomly chosen compounds from 990 ACD compounds as negative examples. The remaining known ligands and 950 ACD molecules are used as the independent testing set.

B. Add –b and –wi parameters—We used the probability estimate of SVM as the score for ranking ligands. We added the “–b” parameter to allow LibSVM for training a model with probability estimates. Next, we transferred the probability score of a compound into Z-score which is used to sort ligands. As mention above, the ratio of negative and positive data is unbalanced. Therefore, we added “–wi” parameter to give the penalty 2 for the class “positive case” which is larger than the class “negative case.” Based on these parameters, we created SVM-AP and SVM-PC models using AP and PC descriptors, respectively.

C. Rank combination—In this study, we combined the results of SVM-AP and SVM-PC using a “combinatorial fusion analysis” method in which a new rank is generated by sorting the average rank of SVM-AP and SVM-PC[9,11]. Application of this hybrid method to each test set improves the performance of both SVM-AP and SVM-PC predicted models.

Let $D=\{d_1, d_2, \dots, d_n\}$ be the set of ligands. Let s be a function, called the score function, from D to R , the set of real numbers. Let r be the function, called the rank function, from D to $N=\{1, 2, \dots, n\}=[1, n]$. By sorting the values of $\{s(d_i)|i=1, 2, \dots, n\}$ into descending order and converting the function $s(d)$ into the function $r(d)$ using rank as its function value. The SVM-AP model and the SVM-PC model would generates two scoring systems AP and PC respectively, where each has a score function s_{AP} and s_{PC} sorting of these two systems s_{AP} and s_{PC} into descending order, two rank functions r_{AP} and r_{PC} are obtained. A new scoring system F with score function $s_F(d_i)=(r_{AP}(d_i)+r_{PC}(d_i))/2$ is obtained. By routing SF into ascending order, the rank function r_F for the combination F is obtained.

3. Results and Discussion

We evaluated LigSeeSVM on seven screening sets, including TK, ER-antagonists, ER-agonists, serotonin receptor, muscarinic receptor, histamine receptor, and GABA_A receptor. Each screening set includes several known active ligands and 950 ACD randomly selected compounds. Four common metrics are used to evaluate the screening quality including the true hit (the percentage of active ligands retrieved from the database), hit rate (the percentage of active ligands in the hit list), goodness-of-hit (GH), and false positive (FP) rate. The true hit and hit rate are defined as A_h/A (%) and A_h/T_h (%), respectively. The FP rate is defined as $(T_h - A_h)/(T - A)$ (%). A_h is the number of active ligands among the T_h highest ranking compounds, which is called the hit list, A is the total number of active ligands in the database, and T is the total number of compounds in the database. The GH score is defined as [6]

$$GH = \left(\frac{A_h (3A + T_h)}{4T_h A} \right) \left(1 - \frac{T_h - A_h}{T - A} \right)$$

In the case of TK, A and T are 8 and 950, respectively.

Table 1 summarizes a brief comparison of the false positive rates at true hit rate 100% for our models and Surflex-Sim[1]. In six of seven testing cases, rank fusion model, LigSeeSVM has lower false positive rate than Surflex-Sim method. All of these 7 cases show that the rank combination strategy indeed improves the accuracies.

3.1 TK inhibitors

The accuracy of SVM models are further evaluated by true hit rate and GH score and the data is shown in Figure 4. The main objective of this study is to compare and evaluate the differences of single and combined models in virtual screening. Figures 4A and 4B show the screening accuracies of SVM-AP, SVM-PC, and LigSeeSVM on the TK dataset. The

screening accuracy is improved by rank combination of LigSeeSVM model. As shown in Figure 4A, the hit rates of predicted models are 50% (SVM-AP), 40% (SVM-PC), and 72.7% (LigSeeSVM) when the TP rate is 100%. In the case of LigSeeSVM, the GH score is 0.79 (Fig. 4B) and the FP rate is 0.3% (Table 1). Although the models with single descriptors have higher false positives than Surflex-Sim method on TK screening, the combined model still reduces the FP. The FP rates of Surflex-Sim, SVM-AP, and SVM-PC are 0.7%, 0.8%, and 1.3%, respectively.

3.2 ER α antagonists and agonists

Figure 4C shows that the hit rates on EST screening for SVM models are 23% (SVM-AP), 17.3% (SVM-PC), and 69.2% (LigSeeSVM) when the TP rate is 100%. The GH score and FP rates of LigSeeSVM are 0.77 (Fig. 4D) and 0.6% (Table 1), respectively. For ESA, the hit rates for SVM models are 15.9% (SVM-AP), 28% (SVM-PC), and 100% (LigSeeSVM) when the TP rate is 100% (Fig. 4E).

Table 2 shows top-ranking compounds of three SVM models for screening ER-agonist. LigSeeSVM is able to improve the accuracy of ligand screening. The AP descriptor and PC descriptor also show different behaviors. For example, among top-ranking 10 compounds, SVM-AP extract 7 sterol analogs and 5 of these are known actives. The others, MFCD00003689 (mestranol) and MFCD00010489 (2-methoxyestradiol), are also estrogen analogs. On the other hand, six of top 10 of SVM-PC are known actives, but the other four structures have no relationships to estrogen. When integrating the tendencies of SVM-AP and SVM-PC, LigSeeSVM model can improve the hits in top 10 and identify similar structures. Three unknown structures in top ten of LigSeeSVM are MFCD00010480, MFCD00003050, and MFCD00003678 which are sterol analogs and similar with estrogen on the scaffold.

3.3 GPCR and GABA $_A$

Figure 5 shows the ROC curve of muscarinic, histamine, GABA $_A$, and serotonin using three SVM models. The FP rates of LigSeeSVM are 0.21 (muscarinic), 0.14 (histamine), and 0.02 (GABA $_A$) when the TP rate is 1. For the serotonin case, which was built on just two ligands of vastly different structure than the test ligands, the FP rate of LigSeeSVM is 0.29 higher than other three cases when the TP rate is 1.

4. Conclusion

In summary, these results suggest that the SVMs are suitable for the ligand screening and yields a prediction accuracy superior to the earlier study derived from Surflex-Sim. The majority compounds with high Z-score often have structures similar to the known ligands. Some compounds with high Z-score have different structures compared with the known ligands, and these compounds have possibility to become the novel, potential lead compounds. Moreover, our study also shows that the strategy of rank combination has great improvement of the results of virtual ligand screening. In the future, our study will adapt feature selection approach to avoid a high variance for unseen molecules (overfitting) and to improve the prediction accuracy of SVM.

In our prediction model LigSeeSVM, we use rank combination to combine the scoring results from SVM-AP and SVM-PC. In future study, we will explore other methods of combination and search for the reason why one combination is better than the other. In particular, we will study the notion of “diversity” between different systems using the rank-score characteristic (RSC) graph studied earlier[19].

Appendix

Appendix 1

100 positive compounds of GPCR and GABA_A studied in this paper, with annotation of known targets.

Molecule name	GPCR			GABA _A
	Serotonin receptor ligand	Muscarinic receptor ligand	Histamine receptor ligand	GABA _A receptor ligand
abecarnil				x
aloseptron	x			
alpidem				x
amitriptyline	x	x	x	
amoxapine		x	x	
astemizole			x	
atropine		x		
bcce ^a				x
benztropine		x	x	
bethanechol		x		
bromolysergide	x			
brompheniramine		x	x	
bupropion			x	
carbachol		x		
carbinoxamine			x	
cetirizine			x	
chlorpheniramine		x	x	
chlorpromazine	x	x	x	
clemastine			x	
clobazam				x
clomipramine		x	x	
clozapine	x	x	x	
cocaine		x		
cypheptadine		x	x	
darifenacin		x		
desipramine		x	x	
dicyclomine		x		
dolasetron	x			
dotarizine	x			
doxepin		x	x	
estazolam				x
fluphenazine		x	x	
flutoprazepam				x
granisteron	x			
halazepam				x
haloperidol		x		
hydroxyzine			x	

Molecule name	GPCR			GABA _A
	Serotonin receptor ligand	Muscarinic receptor ligand	Histamine receptor ligand	GABA _A receptor ligand
iloperidone	x			
imipramine		x	x	
itasetron	x			
ketanserin	x			
levocabastine			x	
lidocaine		x		
loratadine			x	
loxapine		x	x	
maprotiline		x	x	
meclizine			x	
mesoridazine		x	x	
metergoline	x			
methotrimeprazine			x	
methscopolamine		x		
methysergide	x			
metitepine	x		x	
mianserin	x	x	x	
midazolam				x
molindone		x		
nefazodone	x			
nortriptyline	x	x	x	
olanzapine	x			
ondansetron	x			
oxetorone	x			
oxybutynin		x		
perospirone	x			
perphenazine		x	x	
phenindamine			x	
phenoxybenzamine			x	
pilocarpine		x		
pimozide		x		
pindolol	x			
pizotyline	x			
prazepam				x
prochlorperazine		x	x	
procyclidine		x		
promethazine		x	x	
protriptyline		x	x	
quazepam				x
ramosteron	x			
risperidone	x		x	

Molecule name	GPCR			GABA _A
	Serotonin receptor ligand	Muscarinic receptor ligand	Histamine receptor ligand	GABA _A receptor ligand
ritanserin	x		x	
sertindole	x	x	x	
suriclone				x
telenzipine		x		
terfenadine			x	
tetrazepam				x
thiethylperazine			x	
thioridazine		x	x	
thiothixene		x	x	
tiotropium		x		
trazodone			x	
triazolam				x
trifluoperazine		x	x	
triflupromazine		x	x	
trimipramine		x	x	
tripelennamine			x	
triprolidine			x	
tropisetron	x			
zaleplon				x
ziprasidone	x			
zolpidem				x
zotepine	x			

^aNote: bcce Is Ethyl β-Carboline-3-carboxylate

Appendix 2A

Ten atom types used for atom pair descriptors.

	Atom type	Description	Atom type mol2 format
1	C.ar	aromatic carbons	C.ar
2	C.na	nonaromatic carbons	C.3, C.2, C.1, and C.cat
3	N.ar	aromatic nitrogen	N.ar
4	N.na	nonaromatic nitrogen	N.3, N.2, N.1, N.am, N.4, and N.p1
5	O.3	oxygen atoms in the sp ³ hybridization state	O.3
6	O.2	oxygen atoms in the sp ² hybridization state	O.2
7	S	all sulfur atoms	S.3, S.2, S.O, and S.O2
8	P.3	phosphorus atoms	P.3
9	X	halogen atoms	F, Cl, Br, and I
10	Other atoms		The other atom types

Appendix 2B

Six thermodynamic descriptors available in Accelrys Cerius² QSAR+ module

Symbol	Description
AlogP	Log of the partition coefficient
AlogP98	Log of the partition coefficient, atom-type value
Fh2o	Desolvation free energy for water
Foct	Desolvation free energy for octanol
Hf	Heat of formation
MolRef	Molar refractivity

Appendix 2C

13 default descriptors available in Accelrys Cerius² QSAR+ module

Symbol	Description
Charge	Sum of partial charges
Apol	Sum of atomic polarizabilities
Dipole-mag	The strength and orientation behavior of a molecular in an electrostatic field
RadOfGyration	The radius of gyration
Area	Molecular surface area
MW	Molecular weight
Vm	Molecular volume
Density	Density
PMI-mag	Principal moments of inertia
Rotlbonds	Number of rotatable
Hbond acceptor	Number of hydrogen-bond acceptors
Hbond donor	Number of hydrogen-bond donors
AlogP98	Log of the partition coefficient, atom-type value

References

1. Accelrys, "Cerius² 4.8," 2003.
2. Bissantz C, Folkers G, Rognan D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *Journal of Medicinal Chemistry*. 2000; 43:4759–4767. [PubMed: 11123984]
3. Carhart RE, Smith DH, Venkataraghavan R. Atom pairs as molecular features in structure-active studies: definition and applications. *Journal of Chemical Information and Computer Sciences*. 1985; 25:64–73.
4. Chang CC, Lin CJ. Training nu-support vector classifiers: theory and algorithms. *Neural Computation*. 2001; 13:2119–2147. [PubMed: 11516360]
5. Hsu DF, Taksa I. Comparing rank and score combination methods for data fusion in information retrieval. *Information Retrieval*. 2005; 8:449–480.
6. Hsu, DF.; Chung, Y-S.; Kristal, BS. Chapter 3 in *Advanced Data Mining Technologies in Bioinformatics*, edited by Hui-Huang Hsu. Idea Group Inc.; 2006. *Combinatorial Fusion Analysis: Methods and Practices of Combining Multiple Scoring Systems*; p. 32-62.
7. Jain AN. Ligand-based structural hypotheses for virtual screening. *Journal of Medicinal Chemistry*. 2004; 47:947–961. [PubMed: 14761196]

8. Katritzky AR, Gordeeva EV. Traditional topological indices vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research. *Journal of Chemical Information and Computer Sciences*. 1993; 33:835–857. [PubMed: 8113335]
9. Karelson M, Lobanov VS, Katritzky AR. Quantum-chemical descriptors in QSAR/QSPR studies. *Chemical Reviews*. 1996; 96:1027–1044. [PubMed: 11848779]
10. Lin KL, Lin CY, Huang CD, Chang HM, Yang CY, Tang CT, Hsu DF. Feature Combination Criteria for improving accuracy in protein structure prediction. *IEEE Trans. on NanoBioscience*; V.6. 2007; (2):186–196.
11. van Lipzig MM, ter Laak AM, Jongejan A, Vermeulen NP, Wamelink M, Geerke D, Meerman JH. Prediction of ligand binding affinity and orientation of xenoestrogens to the estrogen receptor by molecular dynamics simulations and the linear interaction energy method. *Journal of Medicinal Chemistry*. 2004; 47:1018–1030. [PubMed: 14761204]
12. Lyne PD. Structure-based virtual screening: an overview. *Drug Discovery Today*. 2002; 7:1047–1055. [PubMed: 12546894]
13. Lyons DM, Hsu DF. Combining Multiple scoring systems for target tracking using rank-score characteristics. *Information Fusion*. 2009; 10(2):124–136.
14. Renfrey S, Featherstone J. Structural proteomics. *Nature Reviews. Drug Discovery*. 2002; 1:175–176.
15. Schneider G, Nettekoven M. Ligand-based combinatorial design of selective purinergic receptor (A_{2A}) antagonists using self-organizing maps. *Journal of Combinatorial Chemistry*. 2003; 5:233–237. [PubMed: 12739938]
16. Shen M, LeTiran A, Xiao Y, Golbraikh A, Kohn H, Tropsha A. Quantitative structure-activity relationship analysis of functionalized amino acid anticonvulsant agents using k nearest neighbor and simulated annealing PLS methods. *Journal of Medicinal Chemistry*. 2002; 45:2811–2823. [PubMed: 12061883]
17. Sneath PHA, Snokal RR. Unweighted pair-group method using arithmetic averages. *Numerical Taxonomy*. 1973:230–234.
18. Xue Y, Li ZR, Yap CW, Sun LZ, Chen X, Chen YZ. Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. *Journal of Chemical Information and Computer Sciences*. 2004; 44:1630–1638. [PubMed: 15446820]
19. Yang J-M, Chen Y-F, Shen T-W, Kristal BS, Hsu DF. Consensus scoring criteria for improving enrichment in virtual screening. *Journal of Chemical Information and Modeling*. 2005; 8:449–480.
20. Zernov VV, Balakin KV, Ivaschenko AA, Savchuk NP, Pletnev IV. Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *Journal of Chemical Information and Computer Sciences*. 2003; 43:2048–2056. [PubMed: 14632457]
21. Zheng W, Tropsha A. Novel variable selection quantitative structure--property relationship approach based on the k-nearest-neighbor principle. *Journal of Chemical Information and Computer Sciences*. 2000; 40:185–194. [PubMed: 10661566]

Figure 1. Overview of LigSeeSVM for ligand-based virtual screening.

Figure 2. Thymidine kinase inhibitors, estrogen receptor antagonists and agonists for ligand-based screening. (A), (C) and (E) (ligands in the boxes) are the positive cases of the training sets. (B), (D) and (F) are the ligands for independent tests.

Figure 3.

The positive compounds in training sets of GPCR and GABA_A. A and B act on serotonin receptor. C, D, and E are muscarinic antagonists. Molecules F—H are H1 receptor antagonists. Molecules I—K are GABA_A receptor agonists. All of positive compounds are with nanomolar binding affinities.

Figure 4.

The accuracy of models for screening. The SVM models are evaluated by true hit rate and GH score. (A) and (B) screening on TK testing set. (C) and (D) screening on ER antagonist set. (E) and (F) screening on ER agonist set..

Figure 5. ROC curves of GPCR and GABA_A. Screening results of 44, 48, 30, and 15 active compounds for (A) muscarinic, (B) histamine, (C) serotonin, (D) GABA_A receptor, respectively, with 950 negative compounds selected from ACD.

Table 1

Comparison LigSeeSVM with other methods using false positive rates on seven sets when the recall is 100%.

Target	False positive rate (%)			
	Surflex-Sim	SVM-AP	SVM-PC	LigSeeSVM
TK inhibitor	0.7	0.8	1.3	0.3
ER antagonist	13.4	3.2	4.5	0.6
ER agonist	NA ^a	3.9	1.9	0
Muscarinic	28	44	27	21
Histamine	21	18	37	14
Serotonin	22	35	50	29
GABA _A	18	20	0.3	2

^aData not available.

Table 2

Comparison of top-ranking 10 compounds of three SVM-based models for screening ER-agonist data set

SVM-AP		SVM-PC		LigSeeSVM	
ID	Structure	ID	Structure	ID	Structure
ESA05		ESA06		ESA02	
ESA02		ESA07		ESA06	
MFCD00003689		ESA02		ESA07	
ESA06		ESA10		ESA10	
MFCD00010489		ESA03		ESA05	
ESA07		ESA08		ESA08	
ESA10		MFCD00000439		ESA03	
MFCD00002206		MFCD00011526		MFCD00010480	

SVM-AP		SVM-PC		LigSeeSVM	
ID	Structure	ID	Structure	ID	Structure
MFCD00012180		MFCD00009902		MFCD00003650
MFCD00005048		MFCD00003747		MFCD00003678