

Virtual Screening System for Finding Structurally Diverse Hits by Active Learning

Yukiko Fujiwara,^{*,†} Yoshiko Yamashita,[‡] Tsutomu Osoda,[§] Minoru Asogawa,^{||} Chiaki Fukushima,[⊥] Masaaki Asao,[#] Hideshi Shimadzu,[⊥] Kazuya Nakao,[#] and Ryo Shimizu[#]

Service Platform Laboratories, NEC Corporation, 2-11-5, Shibaura, Minato-ku, Tokyo 108-8557, Japan, Business Innovation Center, NEC Corporation, 34, Miyukigaoka, Tsukuba, Ibaraki 305-8501, Japan, Business Innovation Center, NEC Corporation, 5-7-1, Shiba, Minato-ku, Tokyo 108-8001, Japan, Nanoelectronics Laboratories, NEC Corporation, 34, Miyukigaoka, Tsukuba, Ibaraki 305-8501, Japan, Tanabe Seiyaku Co., Ltd., 2-2-50, Kawagishi, Toda, Saitama 335-8505, Japan, and Tanabe Seiyaku Co., Ltd., 3-16-89, Kashima, Yodogawa-ku, Osaka 532-8505, Japan

Received March 4, 2007

Two virtual screening strategies, “query by bagging” (QBag) and “query by bagging with descriptor-sampling” (QBagDS), based on active learning were devised. The QBag strategy generates multiple structure–activity relationship rules by bagging and selects compounds to improve the rules. To find many structurally diverse hits, the QBagDS strategy generates rules by bagging with descriptor sampling. They can also use prior knowledge about hits to improve the efficiency at the beginning of screening. We performed simulation experiments and clustering analysis for several G-protein coupled receptors and showed that the QBag and QBagDS strategies outperform the conventional similarity-based strategy and that using both descriptor sampling and prior knowledge are effective for finding many hits. We applied the bagging with descriptor sampling strategy to novel hit finding, and 4 of the 10 selected compounds showed high inhibition.

INTRODUCTION

Screening a chemical library is to identify hits that show the desired bioactivities against the specific drug target in a collection of compounds. High-throughput screening (HTS) is usually performed in pharmaceutical companies. HTS tests the entire library by random screening in which a fraction of library is randomly selected and tested iteratively. However, HTS is technically difficult in some assay systems or costly for a large library. Without this approach, only a fraction of a library can be selected and tested. When the crystal structure of the target protein and the binding mode is known, structure-based virtual screening can be performed. However, in many cases, the structural information is not available. Consequently, the conventional common strategy without HTS is similarity-based selection. In regards to this strategy, compounds are represented by descriptors, and the similarity between two compounds is quantified by some index such as the Tanimoto coefficient.¹ This similarity-based strategy can be used to design the representative library or the focused library, screen the library to find the initial hits, and select compounds similar to the initial hits in the entire library. However, the representative library tends to contain few or no hits, and the focused library tends to contain structurally similar hits, so the selected compounds are not sufficient for sampling an entire chemical space; that is, the diversity of hits is insufficient.

To predict bioactives in compound library, several methods using support vector machines (SVMs)^{2,3} or random forests⁴ have been proposed previously. However, these methods are time-consuming and only applied to a small size library. In this paper, we propose new virtual screening strategies based on active learning. Active learning is one of several machine-learning methods, and differs from passive learning in that the learning algorithm is assumed to control the inputs. Active learning is probably more powerful than passive learning, giving better prediction performance for a fixed number of training data.⁵ Among active learning methods, we applied “query by bagging” (QBag)⁶ for its high performance and computational speed. QBag generates multiple structure–activity relationship (SAR) rules by data sampling, selects the input that predicted output splits most evenly, and queries its true output. The queried compound is considered very informative to improve the prediction performance. To find more structurally diverse hits, we propose a “query by bagging with descriptor-sampling” (QBagDS) strategy. This strategy generates multiple SAR rules by bagging with descriptor sampling. By using a part of the data and descriptors, we expect to be able to generate diverse SAR rules and select diverse compounds. Before active learning, we compare the prediction performance of bagging with descriptor sampling (BagDS), bagging, SVMs and random forests. The results show that the prediction performance of SVMs is inferior to that of the other methods. Also, BagDS outperforms random forests for severely unbalanced data.

In regards active learning, the identified hits are used to generate SAR rules and the generated rules are used to select compounds. The random screening should thus be repeated until the first hit is identified. However, if the ratio of hits in

* Corresponding author. Phone: +81-3-5476-4387. Fax: +81-3-5476-1083. E-mail: y-fujiwara@db.jp.nec.com.

[†] Service Platform Laboratories, NEC Corporation.

[‡] Business Innovation Center, NEC Corporation, Ibaraki.

[§] Business Innovation Center, NEC Corporation, Tokyo.

^{||} Nanoelectronics Laboratories, NEC Corporation.

[⊥] Tanabe Seiyaku Co., Ltd., Saitama.

[#] Tanabe Seiyaku Co., Ltd., Osaka.

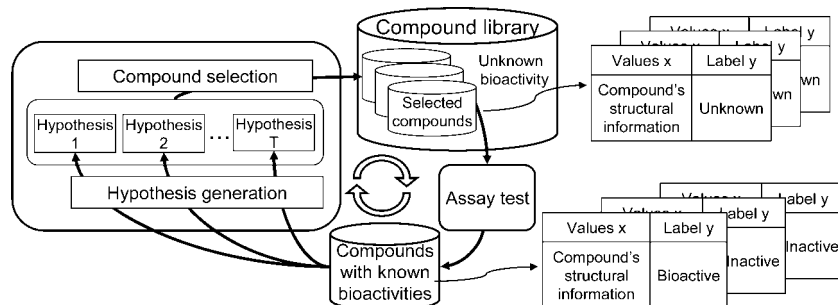


Figure 1. Virtual screening system based on active learning.

a library is extremely low, no hits are identified for a while in screening. To find the first hits efficiently at the beginning of virtual screening, we propose to use the chemical structures of ligands of biologically related proteins, namely, “pseudo-positives”, as prior knowledge about hits.

To evaluate the effectiveness of the proposed strategies, we performed simulation experiments of virtual screening for several G-protein coupled receptors (GPCRs), one of the most important targets in drug discovery.⁷ The library contains a practical number of compounds, namely, 240 000.

METHODS

System Overview. In this virtual screening system, compounds, called *examples* in machine learning, are represented by a collection of structural descriptors. The bioactivity of a compound is called a *label*. We set the label of a bioactive compound (a hit compound) to 1, and call a bioactive a *positive example*. We set the label of an inactive to be 0, and call it a *negative example*. A set of structure–activity relationship (SAR) rules, called *hypothesis* h , is equivalent to the function $h, y = h(x)$, which input is a collection of descriptor values (x) and output is bioactivity (y).

The system overview is shown in Figure 1. First, all compounds in a library are unlabeled, and the system randomly selects compounds from a library and queries their labels. Second, biochemical tests identify their labels, and the system generates multiple sets of SAR rules using the labeled compounds. Third, it selects compounds based on the SAR rules as shown below. These steps are iterated until a fixed number of compounds have been tested.

Previous Strategies of Active Learning. One of the general strategies of active learning is the “query by committee” algorithm (QBC).⁸ QBC maintains a version space, namely, the set of hypotheses that is consistent with the labeled examples, and gradually shrinks the version space by querying the label of one selected example. First, in the case of QBC, multiple hypotheses are generated by the Gibbs algorithm,⁹ which randomly selects hypotheses from the version space. QBC then selects an unlabeled example that predicted output splits most evenly. QBC queries the label of this example, and then removes all hypotheses from the version space that are not consistent to the label. This reduces the version space to almost half.

Although QBC is the popular strategy, it requires an intractable amount of time in complex practical applications. To reduce computational time, “query by bagging” (QBag) was proposed.⁶ Instead of using the time-consuming Gibbs algorithm, QBag generates multiple hypotheses by “bag-

ging”,¹⁰ which takes repeated samples drawn randomly with replacement from the labeled set, and generates multiple hypotheses for the samples by using a training algorithm. QBag then selects an unlabeled example, x , according to the same criteria as QBC,

$$X = \arg \min \{t \leq T | h_t(X) = 1\} - |\{t \leq T | h_t(X) = 0\}| \quad (1)$$

where T is the number of hypotheses, and h_t is the t th hypothesis.

Modified QBag. The original QBag queries one example at a time. However, in the screening process for drug discovery, the practical size of a library is huge, and one-by-one biochemical testing is time-consuming. We thus modified QBag to query multiple unlabeled examples at a time. The number of queried examples at a time is called the “batch size”.

QBagDS. Although QBC uses any hypotheses in the version space, QBag uses hypotheses in the learnable part of the version space. In other words, some hypotheses in the version space are unused in the case of QBag. In virtual screening, these unused hypotheses prevent to identify more diverse hits. Consequently, to identify diverse hits, we propose to add “descriptor sampling” (DS) to QBag, forming QBagDS. The DS strategy repeatedly takes descriptors drawn randomly from the entire group of descriptors, and generates multiple hypotheses based on those descriptors.

The effectiveness of adding the DS strategy for the case of virtual screening is described next. Here, we suppose that one substructure has no contribution to the bioactivity, and is contained in all identified bioactives and few identified inactives. This substructure is not important in discriminating bioactives from inactives in the entire library, but it is important in discriminating identified bioactives from identified inactives. Consequently, QBag always generates hypotheses based on this substructure and predicts the compounds without this substructure as inactives in almost all the hypotheses. It is thus difficult to select the unidentified bioactives without this substructure. As all the identified bioactives have the same substructure, in the case of QBag, the hits are structurally similar. On the other hand, the DS strategy sometimes eliminates the information of this substructure. Consequently, QBagDS sometimes generates hypotheses not using this substructure and predicts the compounds without this substructure as bioactives. It is thus expected that QBagDS can select the unidentified bioactives without this substructure. As the bioactives with or without this substructure are expected to be identified, the hits tend to be structurally diverse in QBagDS.

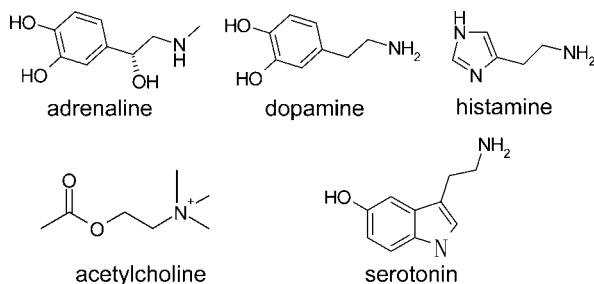


Figure 2. Biogenic amines.

SIMULATION OF VIRTUAL SCREENING

Data. To examine the efficiency of active learning in regards to the compound-screening process, we performed a computer experiment with the following three types of simulation.

(1) *Biogenic Amines (BA)*. Among GPCRs, the receptors of biogenic amines (BA) such as adrenaline, dopamine, histamine, acetylcholine, and serotonin are representative targets of drug discovery, and a lot of pharmacological knowledge on them has been accumulated (Figure 2). We therefore selected BA receptors as the targets of the first trial of application of active learning. In this simulation, a compound reported to have affinity to one or more of the BA receptors was treated as a positive example.

In this and following simulations, we did not discriminate agonists and antagonists. Since slight structural change of a compound, such as extension or branching of alkyl groups, causes the transformation of an agonistic effect to an antagonistic one, we thought that both agonists and antagonists would share some structural features and should be unified as “ligands” of the target GPCRs.

(2) *Histamine Receptors (HIST)*. To evaluate the effectiveness of active learning in the case of smaller numbers of positive examples, we performed a simulation to find hits for histamine receptors. Histamine is well-known as a chemical mediator, and its antagonists are widely utilized as therapeutic agents of allergies and ulcers.

(3) *Endothelin Receptors (ET)*. Although BA receptors have been the representative target family of drug discovery, the most attractive family of GPCR is receptors of peptide hormones. To assess the effectiveness of virtual screening for such kind of targets, we performed active-learning simulation in regards the screening process of endothelin receptor–ligands. Endothelin is a group of endogenous peptides, and their antagonists are thought to be promising as therapeutic agents for cardiovascular diseases.

It is well-known that compounds binding to proteins of the same family often share some structural features. For example, ATP-competitive inhibitors of protein kinases share structural characteristics such as heteroaromatic rings and hydrogen-bonding groups. This structural similarity of ligand compounds sometimes causes a problem of cross-reactivity among biologically related proteins.¹¹ When no active compounds are available as the starting material, a random screening has to be repeated until the first hit is identified. To increase the possibility of finding hits in this early stage of simulation, we examined the effectiveness of pseudo-positive compounds. The ligands of biologically related proteins are retrieved from the database and used as “sham” positives in the active learning until more than one “real”

positive is found. We expected that the structural features shared by ligands of the same protein family would give some guidance for choosing compounds from a library, enabling us to find hits earlier than the random screening can. Angiotensin (ANG) receptors and opioid (OPI) receptors belong to the same family of peptide GPCRs as ET receptors. We selected five pseudo-positive compounds from these two groups of peptide receptors, ANG and OPI, which are shown in Figure 3.

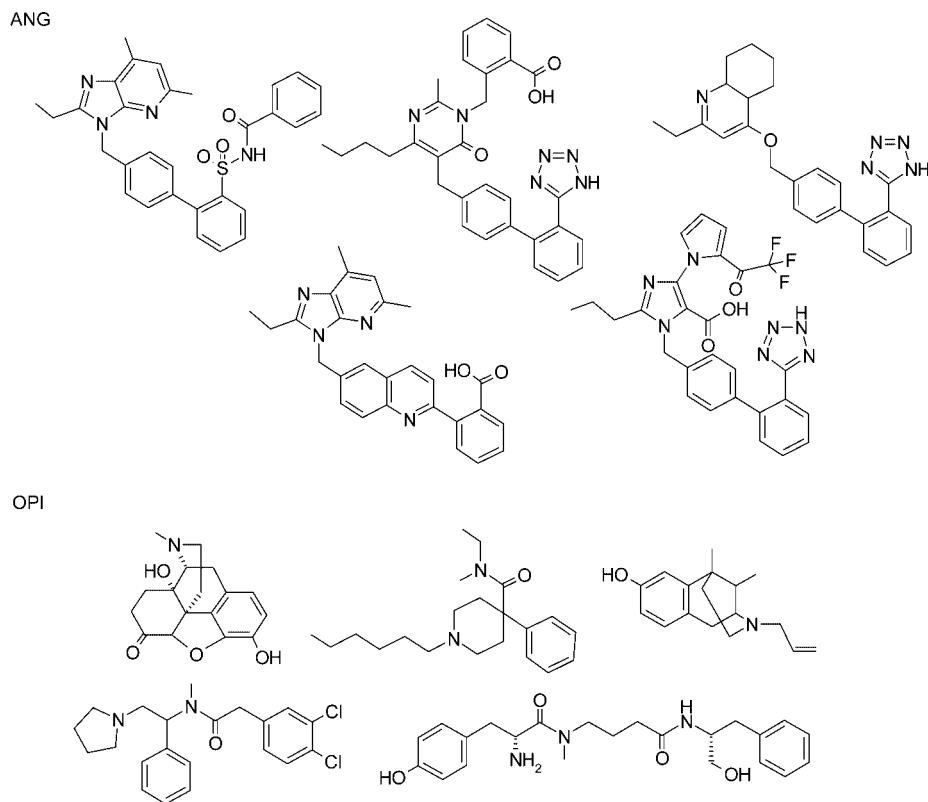
Data of all examples were taken from compound databases, desalted, and their molecular charge was corrected. Drug-like compounds were then extracted by applying the following “drug-likeness filters”: molecular weight between 100 and 1000, six and more heavy atoms; compounds consisting of usual elements as drugs such as carbon, hydrogen, nitrogen, oxygen, sulfur, phosphorus, fluorine, chlorine, bromine and iodine; no isotopes; no peptide structures larger than tetrapeptides; no reactive groups such as acid halides (see the Supporting Information).

The examples were taken from Pharmaprojects (version 2004.2),¹² which is a database of compounds under clinical and preclinical trials. After applying the drug-likeness filters, prodrugs of bioactive compounds were omitted because structural information of pro-moiety was thought to be noise in the learning process. Positive and negative examples were then marked off according to the description of the “mechanism of action” field.

Under the condition of real screening, the number of negative examples is much larger than that of positive examples. To set a similar condition in our simulations, we added a large number of drug-like compounds as probable negative examples from a reagent database, Available Chemical Directory (ACD; version 2005.1).¹³ From this database, 289 055 compounds were selected by using the drug-likeness filters. The numbers of both examples are listed in Table 1.

Molecular Descriptors. The chemical structures of all data were parametrized by using MDL Molskeys,¹⁴ the logarithm of partition coefficient in a 1-octanol/water system, ClogP, estimated by the CLOGP program,¹⁵ molecular weight, the numbers of hydrogen-donating and accepting groups¹⁶ and the number of rotatable bonds.¹⁶ The 166-bit string of MDL Molskeys was shown to be suitable to describe the chemical structures of common bioactivity by Brown and Martin,¹⁷ and widely used in diversity analysis of a chemical library and selection of similar compounds. The other five parameters are also popular in quantitative structure–activity relationship (QSAR) analysis of drugs.

SAR Rules. As mentioned above, support vector machines (SVMs)^{2,3} and random forests⁴ have previously been proposed to predict bioactives in compound library. To compare the prediction performance of bagging with descriptor sampling (BagDS), bagging, random forests¹⁸ and SVMs,¹⁹ 5-fold cross-validation (5-fold CV) was performed in which the positive and negative examples are taken from Pharmaprojects. In 5-fold CV, the compounds are randomly split into five mutually exclusive partitions (the folds) of approximately equal size. One of the 5-folds is set aside as the test set, and the remaining folds are used as the training set. This creates five sets of training and test sets. For each fold, the training set is used to generate the SAR rules and the

**Figure 3.** Pseudopositive compounds.**Table 1.** Number of Example Compounds

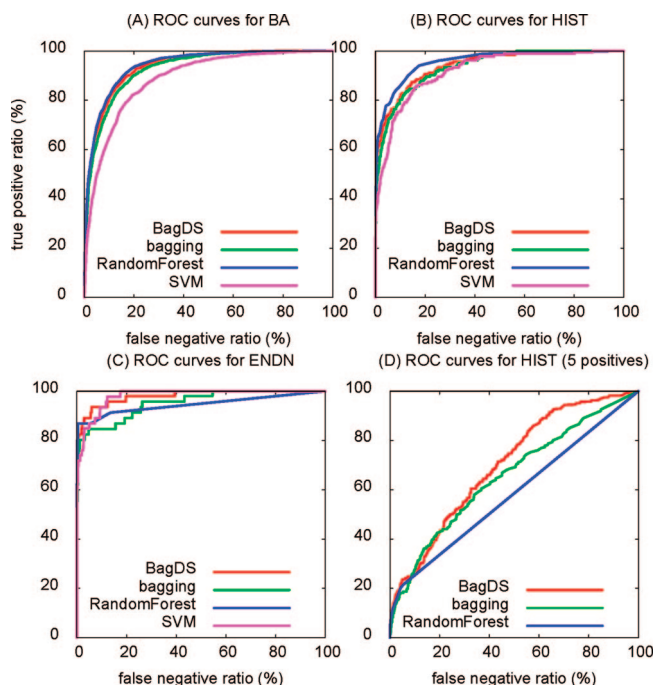
	positive	negative	
	Pharmaprojects	ACD	
biogenic amine receptors	1515	9072	289055
histamine receptors	182	10514	289055
endothelin receptors	46	10196	289055

test set is used to calculate the prediction performance. For all methods, all 171 descriptors are used before descriptor sampling.

In regard to SVMs, the prediction performance was very low using default linear kernel and default parameters (data not shown). Thus, the kernel function was set to the most commonly used radial basis function (RBF) and the combination of several parameters are tuned using the prediction performance of the test set. Finally, the tradeoff between training error and margin was set to 100 and RBF width was set to 0.00001. For BagDS or random forest, the number of trees is 100.

As regards prediction performance, the average receiver-operating-characteristic (ROC) curves of 5-fold CV are shown in Figure 4A–C. The ROC curve is a graphical plot of the true positive ratio vs the false positive ratio as its discrimination threshold is varied. The true positive ratio is the ratio of correctly predicting positive examples, and the false positive rate is the ratio of incorrectly predicting negative examples. As shown in the figures, BagDS, bagging, and random forests outperformed SVMs around in the top ranking scores. As we are interested in the top ranking hits in screening, we omitted SVMs for the future experiments.

In HIST, random forests outperformed QBagDS and bagging. However, in virtual screening, the identified bioactives are biased, and the number of the identified bioactives is small. To

**Figure 4.** ROC curves (A) for biogenic amines, (B) for histamine receptors, (C) for endothelin receptors, and (D) for histamine receptors trained using 5 bioactives.

estimate the natural performance of virtual screening, we performed additional experiments in which bioactives are removed from the training set so that the number of bioactives is five. The average ROC curves of 5-fold CV are shown in Figure 4D. As shown in the figures, BagDS and bagging outperforms the random forest. According to the results, we developed a virtual screening system applicable to a practical

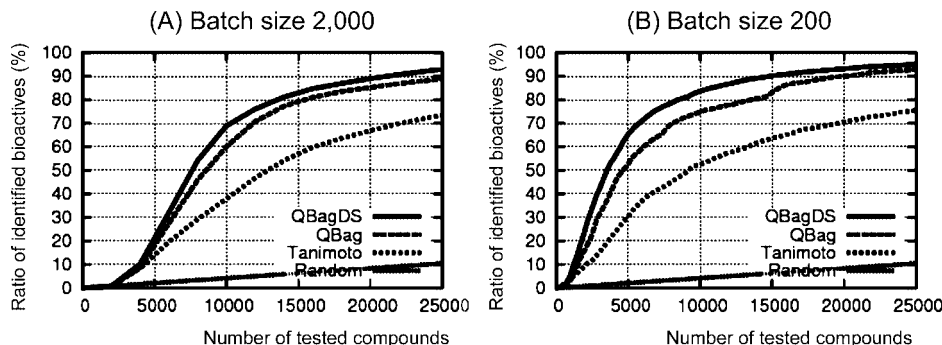


Figure 5. Learning efficiency for biogenic amines: batch size (A) 2000, (B) 200.

size library using QBagDS and QBag. Further considerations for random forest are given in discussion.

Virtual Screening Strategies. Two proposed strategies and two conventional strategies are compared in simulation experiments without assuming prior knowledge. The first proposed strategy is QBag. The second strategy is QBagDS. The compound selection of both strategies is performed by eq 1. The SAR rules are represented by decision trees²⁰ in this paper. The first conventional strategy is random screening (random). It randomly selects a set of compounds. The second conventional strategy is similarity-based selection using the Tanimoto coefficient (Tanimoto). There are several methods for calculating the similarity to a set of compounds. Among them, we used one nearest neighbor (1NN) method for conventional similarity-based selection, because the examination showed that 1NN is the most effective method to obtain the ligands for the same target.²¹ In 1NN, the similarity between compound x and a set of compounds R is the Tanimoto coefficient between x and its nearest neighbor in R . In this strategy, 166 Molskeys were used, because the Tanimoto coefficient is usually calculated using binary strings.

The other simulation experiments were performed to evaluate the effectiveness of pseudopositive knowledge. In these experiments, QBagDS was used to identify ligands of endothelin, because the number of bioactives for ET is the lowest among three ligands.

In all simulation experiments, the batch size (number of queried compounds at a time) is set to 2000 or 200. The batch size of 2000 was assumed for middle-throughput screening, and that of 200 was assumed for low-throughput manual screening.

Evaluation Methods. The purpose of screening is to find many hits and to generate high-performance SAR rules. To evaluate these purposes, we evaluated learning efficiency, that is, the ratio of the identified hits, and prediction performance, that is, the performance of the generated SAR rules. To test prediction performance, the compounds are randomly split into 5-folds. One of the 5-folds is set aside as the test set, and the remaining folds are used as the screening set. This creates five sets of screening and test sets. For each fold, the screening set is assumed to be a chemical library and is screened to calculate the learning efficiency

$$\text{learning efficiency} = \frac{\text{number of identified bioactives in screening set}}{\text{number of compounds in screening set}}$$

The test set is predicted according to SAR rules to calculate the prediction performance. As regards prediction perfor-

mance, one of the major measures is the area under the curve (AUC) of the ROC curve. However, in screening, we are interested in the top ranking hits. Consequently, we consider another measure, E_{1000} , for representing prediction performance:

$$E_{1000} = \frac{\text{number of hits in 1000 top ranking compounds in test set}}{\text{number of hits in test set}}$$

These two measures do not directly consider the diversity of hits. The diversity of hits is discussed later.

EXPERIMENTAL RESULTS

Results for Biogenic Amines (BA). Learning efficiency for BA until 10 % of screening set (25 000 compounds) had been tested and is shown in Figure 5. In this figure, the x -axis is the number of tested compounds in screening set, and the y -axis is the ratio of the identified hits in screening set. The average results in 5-folds for QBagDS, QBag, Tanimoto, and random strategies are described by the solid line, fine dashed line, solid dashed line, and dotted line, respectively. As shown in this figure, the proposed strategies, QBagDS and QBag, outperform the conventional strategies, Tanimoto and random. Moreover, QBagDS outperforms QBag. Table 2 lists the ratio of identified hits after 10 000 or 20 000 compounds had been tested. For example, after 20 000 compounds had been tested for batch size 2000, random, Tanimoto, QBag, and QBagDS identified 8.3%, 66.9%, 85.2%, and 89.0% of hits, respectively.

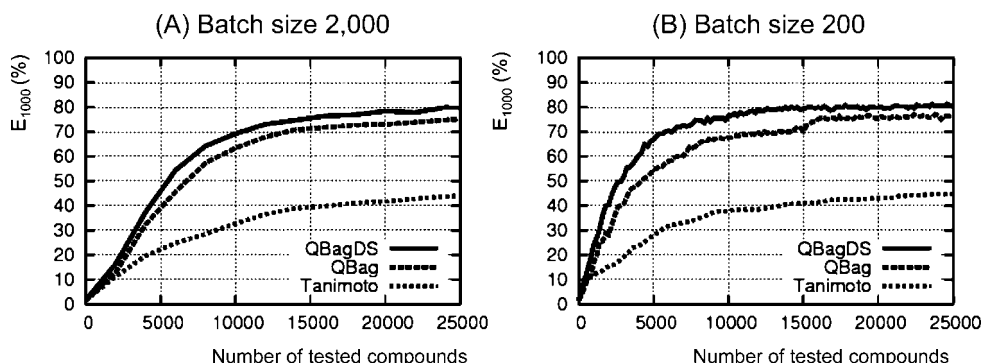
The average prediction performance for BA until 10 % of the screening set had been tested is shown in Figure 6. In this figure, the x -axis is the number of tested compounds in screening set, and the y -axis is E_{1000} . Note that the prediction of the random strategy is performed by random guess and E_{1000} is constant (1.7%), so random is omitted from this and the following figures. In this figure, the proposed strategies outperform the conventional strategies and QBagDS outperforms QBag. Table 3 lists the prediction performance E_{1000} after 10 000 or 20 000 compounds had been tested. For example, after 20 000 compounds had been tested for batch size 2000, E_{1000} values for random, Tanimoto, QBag, and QBagDS are 1.7%, 41.7%, 73.0%, and 78.3%, respectively.

Results for Histamine (HIST). The learning efficiency and prediction performance for HIST until ten percent of the screening set had been tested are shown in Figures 7 and 8, respectively. As shown in these figures, the proposed strategies outperform the conventional strategies in terms of learning efficiency and prediction performance. Moreover,

Table 2. Learning Efficiency (percent)

batch size	ligand	10000				20000			
		random	Tanimoto	QBag	QBagDS	random	Tanimoto	QBag	QBagDS
2000	BA	4.2	37.9	60.0	68.9	8.3	66.9	85.2	89.0
	HIST	4.2	41.8	47.0	55.0	8.3	67.3	80.5	84.2
	ET	4.2	33.6	44.6	57.0	8.3	42.4	85.8	91.3
200	BA	4.2	52.5	74.8	83.6	8.3	70.6	90.0	93.2
	HIST	4.2	45.1	75.1	78.8	8.3	71.4	87.0	90.8
	ET	4.2	34.7	70.1	82.1	8.3	42.9	88.1	95.6

^a Pseudo-positives (left ANG, right OPI) are used.

**Figure 6.** Prediction performance for biogenic amines: batch size (A) 2000, (B) 200.**Table 3.** Prediction Performance (percent)

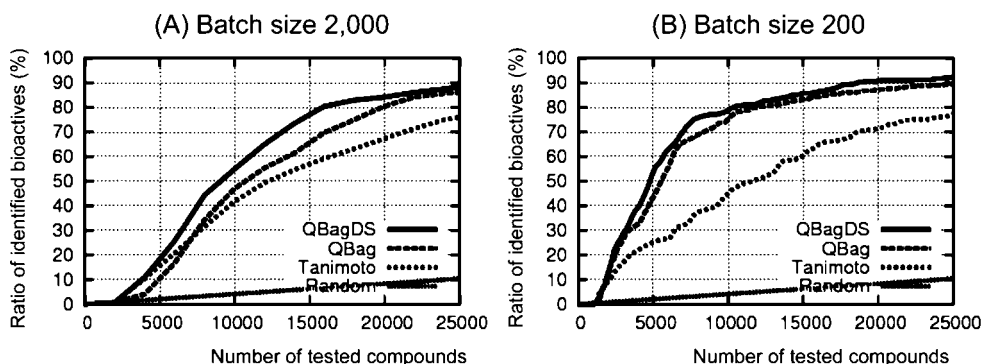
batch size	ligand	10000				20000			
		random	Tanimoto	QBag	QBagDS	random	Tanimoto	QBag	QBagDS
2000	BA	1.7	32.7	63.4	69.2	1.7	41.7	73.0	78.3
	HIST	1.7	39.5	50.5	61.0	1.7	54.5	76.4	76.4
	ET	1.7	32.7	71.3	73.6	1.7	36.9	73.6	84.4
200	BA	1.7	38.0	67.7	76.5	1.7	43.0	75.4	79.8
	HIST	1.7	40.8	72.0	75.3	1.7	56.1	74.8	80.8
	ET	1.7	32.9	60.0	80.2	1.7	36.9	66.9	78.0

QBagDS outperforms QBag in both measures. For example, in Table 2, after 20 000 compounds had been tested for batch size 2000, random, Tanimoto, QBag, and QBagDS identified 8.3%, 67.3%, 80.5%, and 84.2% of hits, respectively. As for prediction performance, E_{1000} values of random, Tanimoto, QBag, and QBagDS are 1.7%, 54.5%, 76.4%, and 76.4%, respectively.

Results for Endothelin (ET). The learning efficiency and prediction performance for ET until ten percent of the screening set had been tested are shown in Figures 9 and 10, respectively. These figures show that the proposed

strategies outperform the conventional strategies in terms of learning efficiency and prediction performance. Moreover, QBagDS outperforms QBag in both measures. For example, in Table 2, after 20 000 compounds had been tested for batch size 2000, random, Tanimoto, QBag, and QBagDS identified 8.3%, 42.4%, 85.8%, and 91.3%, respectively. Furthermore, E_{1000} values of random, Tanimoto, QBag, and QBagDS are 1.7%, 36.9%, 73.6%, and 84.4%, respectively.

Figure 11 shows the results of learning efficiency when the chemical structures of the ligands of angiotensin (ANG) and opioid (OPI) receptors were used as a pseudopositive

**Figure 7.** Learning efficiency for histamine receptors: batch size (A) 2000, (B) 200.

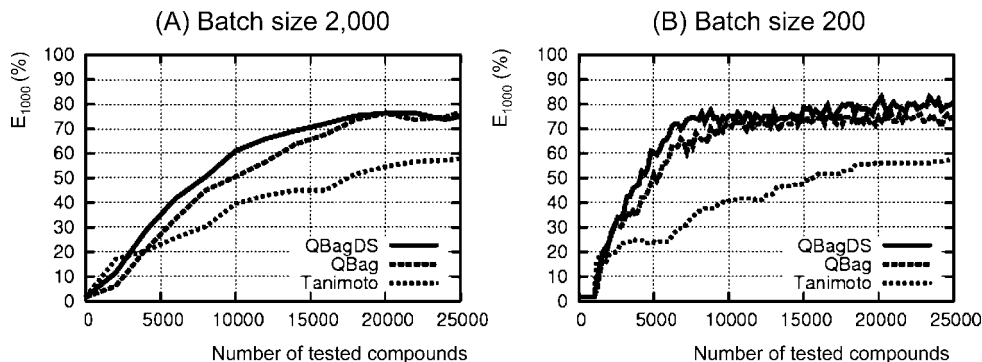


Figure 8. Prediction performance for histamine receptors: batch size (A) 2000, (B) 200.

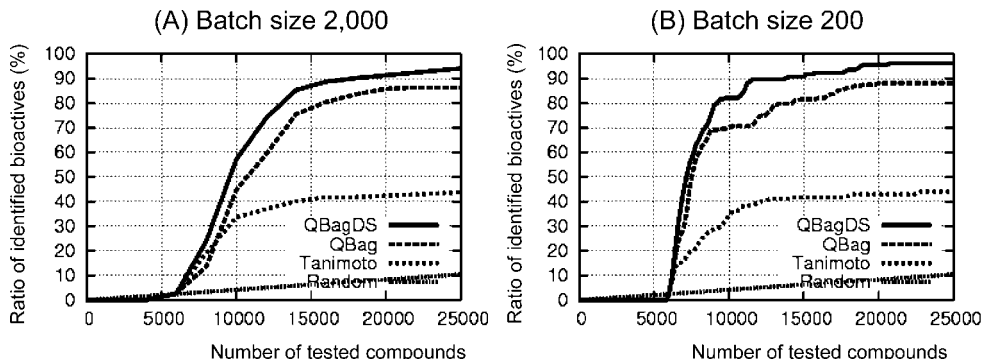


Figure 9. Learning efficiency for endothelin receptors: batch size (A) 2000, (B) 200.

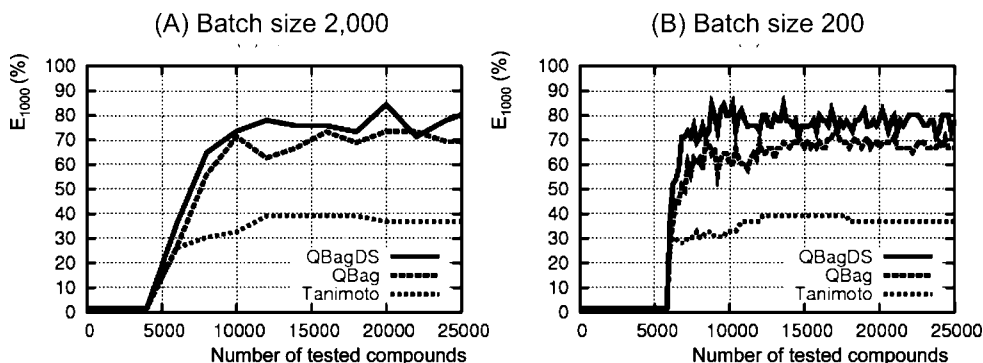


Figure 10. Prediction performance for endothelin receptors: batch size (A) 2000, (B) 200.

example. In all simulations, learning pseudopositives led to quicker finding of ET ligands, which proved the effectiveness of using knowledge concerning ligands of biologically related GPCRs in the early stages.

FINDING OF NOVEL BIOACTIVES

To examine the predictability of the active-learning strategy, we evaluated whether active hits could be selected from 52 657 compounds available from Maybridge Corp. We estimated the affinity of these Maybridge compounds to BA receptors according to the SAR rule obtained by learning the results on 20 000 compounds. Among 50 high-scored compounds using BagDS, we manually selected and purchased 10 compounds. Each score is calculated by the average score of 100 trees. Their antagonistic effects against α_1 - and α_2 -adrenarine, muscarinic acetylcholine, and serotonin receptors were studied at 1×10^{-6} M at Cerep.²² As listed in Table 4, four compounds exhibited inhibitory potency of more than 50%. This percentage proves that active learning is effective for finding novel bioactive compounds.

DISCUSSION

We developed a virtual screening system based on active learning and evaluated it in the case of a practical size library. The screening set contains 240 000 compounds. The results show that the proposed strategies outperformed the conventional similarity-based strategies in terms of both learning efficiency and prediction performance. For these three ligands, after 20 000 compounds had been tested for batch size 2000, the number of identified hits of the proposed QBagDS was 10 times higher than those of the random strategy and 1.3–2.1 times higher than those of the Tanimoto strategy. Using SAR rules, after 20 000 compounds had been tested for batch size 2000, the number of top-1000 ligand-like compounds in QBagDS was 32 to 50 times higher than those of random, 1.4–2.3 times higher than those of Tanimoto. Moreover, the proposed DS strategy is effective. For any of these three ligands, the learning efficiency and prediction performance with descriptor sampling are superior to those without descriptor sampling. For these three ligands, after 20 000 compounds had been tested for batch size 2000,

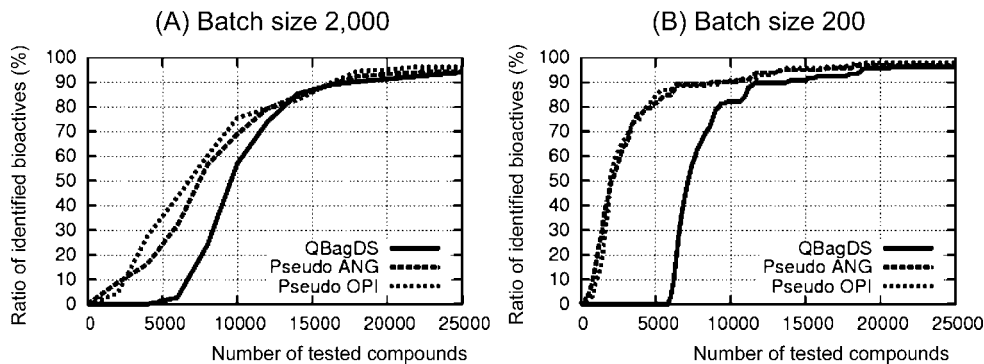
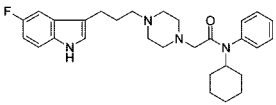
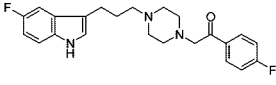
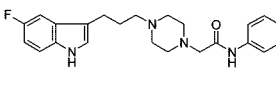
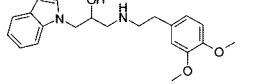


Figure 11. Effect of pseudopositives: batch size (A) 2000, (B) 200.

Table 4. Bioactivity of Novel Active Compounds (percent inhibition at 1×10^{-6} M)

Compound	α_1 adrenergic	α_2 adrenergic	Muscarinic	Serotonin
	47	3	64	55
	96	21	23	60
	52	0	21	52
	70	0	0	26

QBagDS identified 3.7–5.5% more compounds than QBag, and the number of top-1000 ligand-like compounds in the case of QBagDS was up to 10.8% more than QBag.

To compare the diversity of hits, all bioactives in a screening set were clustered by Ward's method based on structural similarity.²³ Figure 12 shows one dendrogram and identified hits in 5-fold cross-validation for ET with batch size of 2000. In this figure, the bioactives are classified into two clusters characterized by the presence of carboxyl and sulfonamide groups. The conventional Tanimoto strategy starts with the carboxyl derivative **13**, and failed to identify sulfonamides after 14 000 compounds had been tested. On the other hand, the proposed strategies start with the same compound **13**, and successfully identified sulfonamide derivatives after 8000 compounds had been tested.

After 14 000 compounds had been tested, the random strategy selects the members of both clusters with a half-probability. However, in all five trials, the Tanimoto strategy identified only the members of one cluster. This results shows that the identified hits of similarity-based strategy are structurally similar to each other. On the other hand, in all five trials, after only 10 000 compounds had been tested, the proposed strategies, QBag and QBagDS, identified the compounds in both clusters. These results show the effectiveness of our strategies to overcome the "local minimum" problem in regards to chemical space and to find diverse hits in screening.

As for QBag and QBagDS, as two clusters could not reveal the difference, five clusters were considered. As shown in Figure 12, after 14 000 compounds had been tested, QBagDS identified the members of all five clusters, while QBag identified the members of only four clusters. The same results for QBagDS were obtained in the other 4-folds; that is, QBagDS identified the members of all five clusters after 14 000 compounds had been tested. These results show that the descriptor sampling strategy is effective in finding diverse hits in screening. The representative compounds in each cluster are shown in Figure 13.

To examine more details, we experimented whether the generated 100 trees correctly predict the compounds in cluster 2. A training set was 2000 compounds, which contains one positive of the carboxyl derivative **13** and 1999 negatives. Then, 100 trees are generated by bagging or bagging with descriptor sampling (BagDS) using this same training set, respectively. The results showed that the number of trees which predict the compound **32** as a positive was zero for bagging, six for BagDS. As for descriptors, f28 ("QCH2Q") contained in many trees generated by bagging and disturbed the correct prediction of the compounds in cluster 2. However, in BagDS, 35 trees were generated without f28 by descriptor sampling, and six out of 35 trees correctly discriminated the compound **32** as a positive. In one of the six trees, the compound **32** was predicted correctly because the tree interprets a molecule having the substructure of "C%N" (f65), "QHQH" (f68), "OCO" (f123) and not having the substructure of "A\$A!A\$A" (f62) as a positive. Thus, some descriptors, such as f28, have marginal contribution to the bioactivities but are important in discriminating given training compounds by chance. Therefore it is desired to generate a path from root to leaf of a tree without some descriptors, which can achieve high prediction performance.

In regard to the difference of BagDS and random forest, the descriptors are sampled once in BagDS for generating a tree, while the descriptors are sampled many times in random forest. Random forest randomly samples the descriptors on each node during generating a tree, thus it tends to remain some obstructive descriptors. For example, the similar experiment was performed using the same training set and the results showed that the number of trees which predict the compound **32** correctly was zero for random forest. In QBagDS, the descriptors are randomly sampled once before generating a tree, thus sometimes it is easy to delete many obstructive descriptors as shown above.

Unlike random forest, BagDS can be applied to SVMs. We experimented the BagDS using SVMs on the previously

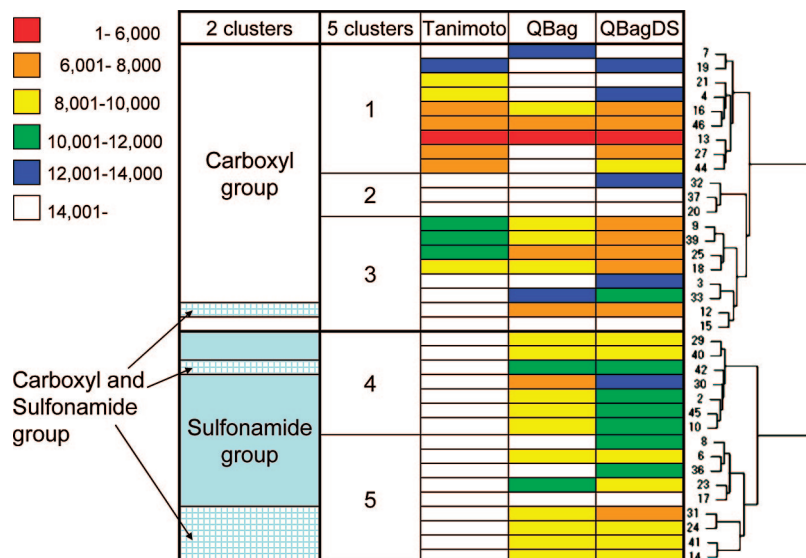


Figure 12. Identified order of bioactives.

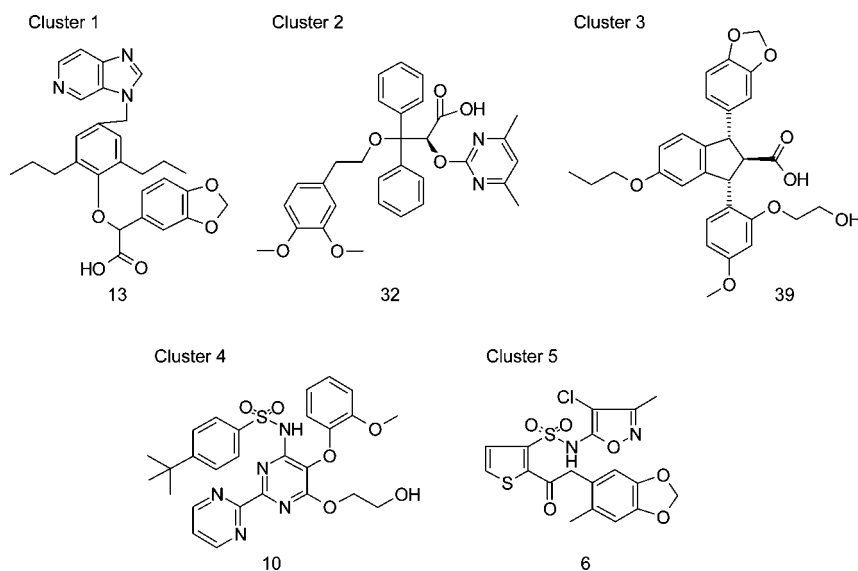


Figure 13. Representative compounds of ET receptor-ligands in each cluster.

adjustment parameters (data not shown). However, the prediction performance of BagDS using SVMs was the similar to that of the single SVM. We considered that the parameters must be adjusted for BagDS using SVMs to achieve higher prediction performance. In the practical screening, this parameter adjustment is hard at the beginning, because the number of identified hits is very small. To apply SVMs in the screening system is thus our future works.

The purpose of the proposed strategies is to achieve high prediction performance. Usually, learning efficiency and prediction performance cannot always improve simultaneously. For example, in active learning with SVMs, most ligand-like-compound-selection strategies produced the most bioactives early on.² However, in our simulation experiments, a very small number of bioactives is contained in the screening set, and tends to improve prediction performance. Consequently, these two measures improve simultaneously.

As for batch size, the number of identified hits and the prediction performance for 200 compounds is superior to those for 2000, especially at the beginning of the active learning. To identify 50% of bioactives, QBagDS required

four percent of biochemical tests for a batch size of 2000 and required three percent of tests for batch size of 200. To identify 90% of bioactives, QBagDS required 10% of tests for batch size of 2000, and 9% of tests for batch size of 200. These results show that a small batch size is superior to a large batch size in reducing the number of tested compounds in virtual screening.

In regard to similarity-based strategies, it is suspected that a minority of the bioactives tends to dominate the selections, which disturbed the identification of structurally diverse bioactives. It is also suspected that the negatives from the generic ACD library are structurally different from the negatives from Pharmprojects, which influenced the experimental results. Thus, additional experiments are performed using the other similarity-based strategy. The positives and negative were taken from Pharmprojects. For the other similarity-based strategy, forming TanimotoEach, the closest compound for each bioactive in the identified compounds are selected from the remaining screening set and the process loops until the number of selected compounds is the batch size. The results of 5-fold CV for ET

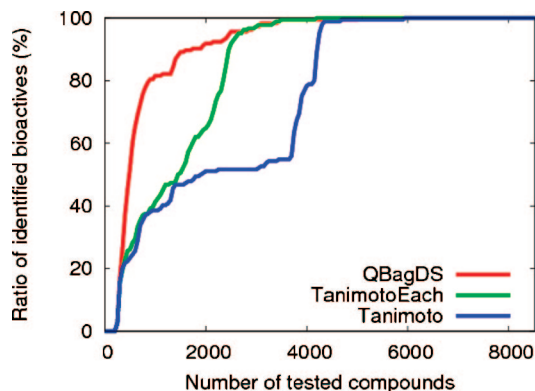


Figure 14. Learning efficiency for endothelin receptors in Pharmaprojects (batch size 50).

are shown in Figure 14. As shown in the figure, the TanimotoEach strategy outperforms the Tanimoto strategy. However, QBagDS outperforms the TanimotoEach for learning efficiency. As for structural diversity, in the average, Tanimoto, TanimotoEach or QBagDS identified one of the compounds in cluster 2, after 49%, 22% or 9% of compounds had been tested, respectively.

As previously mentioned, the employment of pseudopositive examples was effective in improving learning efficiency when no “real” positives are available. Schuffenhauer et al. also reported that similarity searching can be extended to identify not only ligands to the same target but also those of other homologous targets without initially known actives.²¹ Again, the smaller batch size is superior, as expected. When active learning is applied in order to select compounds in stepwise manner, the number of compounds to be tested at once should be determined by not only the efficiency of learning but also the efficiency of the screening experiment. As shown in Table 1, learning efficiency attained using pseudopositives is always higher than QBagDS without them, especially in the earlier process of learning. For example, 80% of positives could be identified by using pseudopositives in the case of a smaller batch size of 200, where no active compounds were found in the same number of tested compounds without pseudopositives. This result suggests that the iteration procedure of compound selection, bioassay and active learning with pseudopositives is very effective in a low-throughput assay system for a very novel target.

In this paper, we reported the results of finding ET ligands according to the knowledge of chemical structure of related angiotensins and opioid receptor–ligands. Recently, the concept of chemical genomics prompts researchers to construct the 2D-matrices of target proteins and chemicals to affect those proteins.²⁴ The accumulation of such kind of knowledge might reveal an effective method for selecting informative pseudopositives. In addition, Schuffenhauer et al. showed that the other methods, such as average of the Tanimoto coefficient, might also be effective in identifying the ligands of homologous targets. The use of the most suitable method for pseudopositives is our future work.

The results of finding novel bioactive compounds with high probability confirm the simulation results. For the other GPCR ligands, we applied the system to 1.3 million compounds in a library, and 62 of the 173 selected compounds showed high inhibitions at concentration of 10 μ M. The details of this application will be published elsewhere.

CONCLUSION

The proposed strategies were effective for finding GPCR ligands and, simultaneously, the generated SAR rules achieved high prediction performance compared to that of conventional similarity-based strategies. Especially, the descriptor-sampling strategy was effective for virtual screening. Furthermore, the clustering analysis reveals that the proposed strategies found the structurally diverse hits that could not be found using the conventional similarity-based strategy. The application of pseudopositive compounds was effective when no active compounds were known.

With the QBagDS strategy, only popular structural descriptors led to high learning efficiency and prediction performance. In the future, to improve performance, we would like to consider the other descriptors that present compounds precisely, and use appropriate descriptors depending on a screening purpose. Moreover, using more precise descriptors, we would like to predict not only bioactive/inactive classification but also the strength of bioactivity.

As for SVMs, their prediction performance is little lower than that of BagDS. Apply SVMs in the virtual screening system is our future work.

ACKNOWLEDGMENT

The authors thank Dr. Hiroshi Mamitsuka of Kyoto University, Dr. Shun Doi and Dr. Kenichi Kamijo of NEC Corporation, and Dr. Hiroki Shirai, Dr. Masataka Kuroda, Dr. Kazuteru Wada, Emi Kushiya, and Takanori Ogaru of Tanabe Seiyaku Co., Ltd. for their useful comments, and we also thank Dr. Kenji Yamanishi of NEC Corp. for providing the sample programs. The authors also thank PBJ Publications Ltd., Shiryō Kenkyūjo Co., Ltd. and Elsevier MDL for providing the database.

Supporting Information Available: Reactive group used in the “drug-likeness filter” (Figure S1), the ROC curves with error bars for Figure 4 (Figure S2), and a list of 50 high-scored compounds in finding novel bioactives (Table S1). This material is available free of charge via the Internet at <http://pubs.acs.org>.

REFERENCES AND NOTES

- (1) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
- (2) Warmuth, M. K.; Liao, J.; Ratch, G.; Mathieson, M.; Putta, S.; Lemmen, C. Active Learning with Support Vector Machines in the Drug Discovery Process. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 667–673.
- (3) Liu, Y. Active Learning with Support Vector Machine Applied to Gene Expression Data for Cancer Classification. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1936–1941.
- (4) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958.
- (5) Cohn, D.; Atlas, L.; Ladner, R. Improving Generalization with Active Learning. *Machine Learning* **1994**, *15* (2), 201–221.
- (6) Abe, N.; Mamitsuka, H. Query Learning Strategies using Boosting and Bagging. *Proceedings of the 15th International Conference on Machine Learning (ICML)*; Morgan Kaufmann: San Francisco, CA, 1998; pp 1–9.
- (7) Horn, F.; Weare, J.; Beukers, M. W.; Horsch, S.; Bairoch, A.; Chen, W.; Fdwarden, E.; Champagne, F.; Vriend, G. GPCRDB: an Information System for G protein-Coupled Receptors. *Nucleic Acids Res.* **1998**, *26* (1), 275–279.

- (8) Seung, H. S.; Oppen, M.; Sompolinsky, H. In Query by Committee. *Proceedings of the 5th Annual Workshop on Computer Learning Theory (COLT)*; 1992; Vol. 28, pp 7–294.
- (9) Haussler, D.; Kearns, M.; Schapire, R. Bounds on the Sample Complexity of Bayesian Learning using Information Theory and the VC Dimension. *Machine Learning* **1994**, *14*, 83–113.
- (10) Beiman, L. Bagging Predictors. *Machine Learning* **1996**, *24* (2), 123–140.
- (11) Vieth, M.; Higgs, R. E.; Robertson, D. H.; Shapiro, M.; Gragg, E. A.; Hemmerle, H. Kinomics—Structural Biology and Chemogenomics of Kinase Inhibitors and Targets. *Biochim. Biophys. Acta* **2004**, *1697*, 243–257.
- (12) *Pharmaprojects*, version 2004.2; PJB Publications Ltd.: UK, 2004; <http://www.pjbpubs.com/>.
- (13) *Available Chemical Directory*, ver. 2005.1; Elsevier MDL: San Leandro, CA, 2005; <http://www.mdli.com/>.
- (14) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1273–1280.
- (15) *Daylight*, ver. 4.82; Daylight Chemical Information Systems, Inc.: CA, USA, 2005; <http://www.daylight.com>.
- (16) *Cerius2*, ver. 4.8; Accelrys Inc.: San Diego, CA, 2005; <http://www.accelrys.com/>.
- (17) Brown, R.; Martin, Y. C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1–9.
- (18) Witten, I. H.; Frank, E. *Data Mining: Practical Machine Learning Tools and Technique*, 2nd ed.; Morgan Kaufmann: San Francisco, CA, 2005.
- (19) Collobert, R.; Sinz, F.; Weston, J.; Bottou, L. Large Scale Transductive SVMs. *J. Machine Learning Research* **2006**, *7*, 1687–1712.
- (20) Quinlan, J. R. *C4.5: Programs for Machine Learning*, Morgan Kaufmann: San Mateo, CA, 1993.
- (21) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity Metrics for Ligands Reflecting the Similarity of the Target Proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391–405.
- (22) <http://www.cerep.fr>, Cerep SA: Paris, France.
- (23) Brown, R. D.; Martin, Y. C. Use of Structure-Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572–584.
- (24) Okuno, Y.; Yang, J.; Teneishi, K.; Yabuchi, H.; Tsujimoto, G. GLIDA: GPCR-ligand database for Chemical Genomic Drug Discovery. *Nucleic Acids Res.* **2006**, *34*, D673–D677.

CI700085Q