Hidden Active Information in a Random Compound Library: Extraction Using a Pseudo-Structure−Activity Relationship Model

Hiroaki Fukunishi,*[,†] Reiji Teramoto,[‡] and Jiro Shimada[†]

Nano Electronics Research Laboratories, Central Research Laboratories, NEC Corporation, 34, Miyukigaoka, Tsukuba, Ibaraki 305-8501, Japan, and Bio-IT Center, Central Research Laboratories, NEC Corporation, 34, Miyukigaoka, Tsukuba, Ibaraki 305-8501, Japan

Received September 11, 2007

We propose a hypothesis that “a model of active compound can be provided by integrating information of compounds high-ranked by docking simulation of a random compound library”. In our hypothesis, the inclusion of true active compounds in the high-ranked compound is not necessary. We regard the high-ranked compounds as being pseudo-active compounds. As a method to embody our hypothesis, we introduce a pseudo-structure−activity relationship (PSAR) model. Although the PSAR model is the same as a quantitative structure activity relationship (QSAR) model, in terms of statistical methodology, the implications of the training data are different. Known active compounds (ligands) are used as training data in the QSAR model, whereas the pseudo-active compounds are used in the PSAR model. In this study, Random Forest was used as a machine-learning algorithm. From tests for four functionally different targets, estrogen receptor antagonist (ER), thymidine kinase (TK), thrombin, and acetylcholine esterase (AChE), using five scoring functions, we obtained three conclusions: (1) the PSAR models significantly gave higher percentages of known ligands found than random sampling, and these results are sufficient to support our hypothesis; (2) the PSAR models gave higher percentages of known ligands found than normal scoring by scoring function, and these results demonstrate the practical usefulness of the PSAR model; and (3) the PSAR model can assess compounds failed in the docking simulation. Note that PSAR and QSAR models are used in different situations; the advantage of the PSAR model emerges when no ligand is available as training data or when one wants to find novel types of ligands, whereas the QSAR model is effective for finding compounds similar to known ligands when the ligands are already known.

1. INTRODUCTION

A protein−ligand docking program has been used to efficiently discover lead compounds for a target protein from a huge compound database. Since the pioneering work by Kuntz et al.,[1] numerous docking programs have been developed.[2−15] Many reports on assessing the performance of the docking program also have been published.[16−29] Of particular interest is the recent critical assessment reported by Warren et al.[21] Docking programs are assessed from two standpoints: pose search or scoring. For the pose search, it is assessed whether the pose of crystal structure can be reproduced. It has been reported that the advanced docking programs achieved a high level of reproducibility. However, users often face difficulty in generating poses of known ligands. This is because the receptor is treated as a rigid body or the flexibility of only several residues are allowed in most docking programs. For the scoring, it is assessed whether the predicted score is correlated with experimentally measured binding affinity. Many reports generally concluded that the scoring function is less successful at correctly identifying the binding mode. Successful results are observed for specific combinations of the target and scoring functions.

Many users performing a docking screen of a random compound library sometimes have experienced the following issue. If no active compound is experimentally found in the compounds that are screened by the docking simulation, the docking simulation results become worthless, even though the docking simulation required a substantial computational cost. However, we think there might be a clue to active compounds in the results regarded as worthless. Compounds high-ranked by the docking simulation must partially interact with important interaction points of a receptor, such as electrostatic or hydrophobic points, even though the formed interactions are not complete as a whole. If the high-ranked compounds are docked at different interaction points, integrating these compounds will provide useful information to reproduce interactions with the receptor as a whole, as shown in Figure 1. Here, we propose a hypothesis that “A model of active compound can be provided by integrating information of compounds high-ranked by docking simulation of a random compound library”. Interestingly, in our hypothesis, the inclusion of true active compounds in the high-ranked compounds is not necessary. We regard the high-ranked compounds as being pseudo-active compounds.

As a method to embody our hypothesis, we introduce a pseudo-structure−activity relationship (PSAR) model. Although the PSAR model is the same as a quantitative structure activity relationship (QSAR) model, in terms of statistical methodology, the implications of the training data

[*] Author to whom correspondence should be addressed. Tel.: +81 298 856 6155. Fax: +81 298 856 6136. E-mail address: h-fukunishi@bu.jp.nec.com.

[†] Nano Electronics Research Laboratories, Central Research Laboratories, NEC Corporation.

[‡] Bio-IT Center, Central Research Laboratories, NEC Corporation.

10.1021/ci7003384 CCC: $40.75
Published on Web 02/16/2008

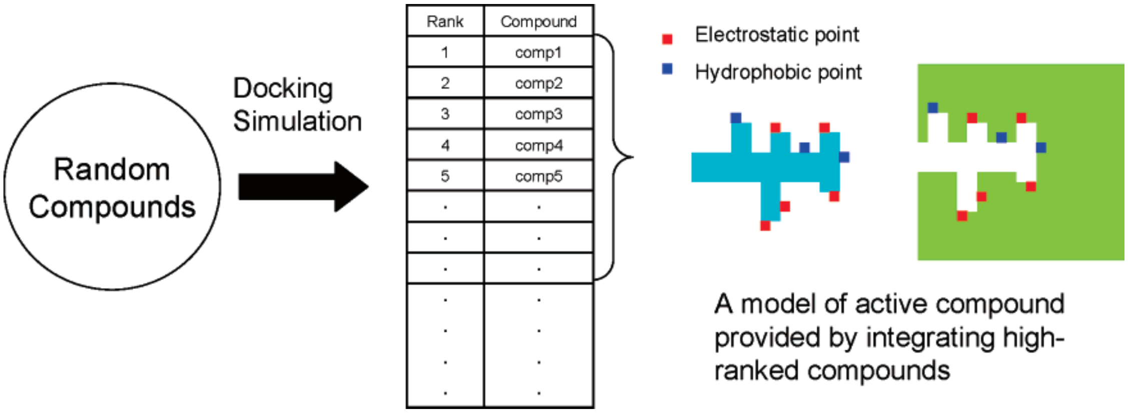

Figure 1. Illustration describing our hypothesis.

are different. Known active compounds are used as training data in the QSAR model, whereas the pseudo-active compounds are used in the PSAR model. The purpose of this study is to test our hypothesis and to assess the practical usefulness of the PSAR model based on our hypothesis for compound screening.

Yoon et al.[30] and Cherkasov et al.[31] have published methods similar to the PSAR model. However, their and our motivations of study are quite different. Their motivation is to replace docking simulation with a QSAR model constructed using the results of the docking simulation,to accelerate the screening. Yoon et al.[30] concluded that their method does not improve or degrade the quality of the docking simulation: it only helps to find them faster. On the other hand, our motivation is to extract active information from docking simulation results regarded as worthless, and note that the PSAR model is one of the means to demonstrate our motivation. We also show that the PSAR model gives different performance from normal use of docking simulation, contrary to the conclusion determined by Yoon et al.[30] There are two more differences between Yoon et al.[30] and us. First, they used a naïve Bayesian model as a machine-learning method while we used a random forest (RF). Second, they used CDK2 and an estrogen receptor to test their method, whereas we used four targets unrelated in terms of biological functions: estrogen receptor antagonist (ER), thymidine kinase (TK), thrombin, and acetylcholine esterase (AChE).

2. METHODS

2.1. Protocol for Testing the PSAR Model. The testing protocol involves four steps:

Step 1. Compounds for training and test sets are prepared and descriptors based on physicochemical and structural factors are assigned for all compounds. The training set includes only random compounds and the test set includes random compounds and known ligands.

Step 2. A docking simulation of the compounds in the training set is performed. For each compound, the best score is selected from the scores of all docking poses generated, and then compounds are ranked based on their best score. High-ranked and low-ranked compounds are classified as pseudo-active and inactive, respectively.

Step 3. The PSAR model is constructed by machine-learning of the pseudo active and inactive compound descriptors.

Table 1. Number of Compounds in the Training and Test Sets

		Number of Compounds	
target	PDB ID	training set	test set[a]
ER	3ert	700	767 (39)
TK	1kim	400	508 (22)
thrombin	1ba8	1200	1312 (72)
AChE	1eve	1900	2055 (107)

[a] The number of ligands is given in parentheses.

Step 4. The activity of the compounds in the test set is predicted using the PSAR model.

In this study, four PSAR models were constructed, based on the percentages of high-ranked and low-ranked compounds: (1) top 5%/bottom 95%, (2) top 10%/bottom 90%, (3) top 20%/bottom 80%, and (4) top 30%/bottom 70%.

2.1.1. Pose Generation Program and Scoring Functions. FlexSIS[15] in the Sybyl7.1J[32] module was used to generate docking poses. Scoring functions used are FlexX (F-score), DOCK (D-score), GOLD (G-score), PMF (P-score), and ChemScore (C-score). The last four scoring functions are included in CScore of the Sybyl module.

2.1.2. Target Proteins, Known Ligands, and Random Compounds. The PSAR models were tested using four targets, which are unrelated, in terms of biological functions: estrogen receptor antagonist (ER), thymidine kinase (TK), thrombin, and acetylcholine esterase (AChE). Known ligands and decoy compounds for each target were retrieved from benchmark sets for molecular docking published by Huang et al. (http://blaster.docking.org/dud/).[33] We treated these decoys as random compounds. Their decoy compounds are chemically distinct from the ligands and they are inactives. The benchmark sets were built to overcome the critical problem that if the distributions of known ligands and decoys differ greatly, in terms of physical properties, they can be simply separated and the enrichment factor obtained is significantly biased. To obtain a more meaningful enrichment factor, the physical property distributions of decoy compounds and ligands should resemble each other. For their benchmark sets, physical property distributions considered are those of druglike descriptors: molecular weight, number of hydrogen bond acceptor, number of hydrogen bond acceptor, number of rotatable bonds and logP. We divided the decoys for each target into training and test sets. Isomers of a compound were counted as different compounds. For each target, Table 1 lists the number of compounds, i.e., total number of known ligands and random compounds. Because compounds failed in the docking simulation were eliminated, the number of compounds used in this study is slightly different from the number reported by Huang et al.

2.1.3. Machine-Learning Algorithm. Random Forest (RF),[34] implemented in statistical software R,[35] was used as a machine-learning algorithm. This algorithm is one of the ensemble learning algorithms, and it is a learning machine that is constructed by combining several low-performance machines. When a large proportion of the data is missing, RF includes an effective method for estimating the missing data and maintains accuracy. Sub-datasets are generated in RF by randomly sampling with replacement; these are called bootstrap samples. A decision tree is constructed using each sub-dataset. The prediction value for each compound is calculated as the percentage of the decision trees judging the compound as active, and the compounds can be ranked based on their prediction values. RF has the following features. It does not take much time to learn descriptors, even though all of the descriptors are huge. This is because a small number of descriptors M selected randomly from all descriptors $N_{\mathrm{descriptor}}$ are used for the machine-learning of each sub-dataset. In this study, a value of $M = \sqrt{N_{\mathrm{descriptor}}}$ was used.[34] Recently, RF was applied to QSAR,[36−41] QSPR,[42] or virtual screening.[43] Svetnik et al.[38] and Plewczynski et al.[39] showed that RF is one of the best performing machine-learning algorithms for the QSAR model.

2.1.4. Descriptor. Native values and pharmacophore fingerprints in JOELib were used as the descriptors. Details of the descriptors can be found at the following website: http://www-ra.informatik.uni-tuebingen.de/software/joelib/tutorial/descriptors/descriptors.html.

3. RESULTS

Results were compared from two standpoints: (1) PSAR model versus random sampling, and (2) PSAR model versus normal scoring by the scoring function. If the PSAR model significantly outperforms random sampling, the result supports our hypothesis. Note that if our hypothesis is irrelevant, the PSAR model will give the performance equivalent to random sampling. If the PSAR model significantly outperforms normal scoring, the PSAR model proves to be practically useful for compound screening.

PSAR models 1−4 described in this section of the method are represented as A5%−I95%, A10%−I90%, A20%−I80%, and A30%−I70%. (Here, A denotes active and I denotes inactive.) Random sampling and normal scoring based on the scoring function are also represented as “Random” and “Normal”.

3.1. Enrichment Curve. Figure 2 shows the enrichment curves of the four PSAR models and the normal scoring in the case of various scoring functions for each target. We refer to the known ligands found and the ranked database as “Hit” and “RankedDB”, respectively.

3.1.1. ER. First, the A5%−I95% and A10%−I90% models with each scoring function gave higher percentages of Hit than Random throughout RankedDB. The A20%−I80% and A30%−I70% models also gave similar results, except for the F-score. Second, the A5%−95% model with each scoring function, which was the best of the four PSAR models, gave a higher percentage of Hit than Normal in the range of more than ∼10% of RankedDB, although their performances were equivalent in the range of less than that percentage. Note that the slope of the curve for the PSAR models hardly declines as the percentage of RankedDB increases, until the percentage of Hit attains a value of 100%. Table 2 lists the percentage of RankedDB at 100% of Hit for ER. The best percentage is 19.7% for the A5%−I95% model with the C-score, which is quite low, in comparison with 83.7% given by the normal scoring. The percentage is somewhat large, but it is still less than 50% for the A5%−I95% models with the F-, D-, G-, and P-scores.

3.1.2. TK. From the poor results of the normal scorings, this target is observed to be difficult for the docking simulation. First, A5%−I95% models with the D-, G-, and P-scores gave higher percentages of Hit than Random throughout RankedDB, whereas hardly any difference was observed between Random and the A5%−I95% models with the F-score, whose graph was the worst in all of the 20 graphs in Figure 2. Second, the A5%−I95% models with each scoring function gave a higher percentage of Hit than Normal throughout RankedDB, although the attained percentage of Hit was still poor.

3.1.3. Thrombin. First, four PSAR models (A5%−I95% to A30%−I70%) with each scoring function gave higher percentages of Hit than Random over most ranges of the RankedDB. Second, the A5%−I95% and A10%−I90% models with the D-, G-, P-, and C-scores gave higher percentages of Hit than Normal over most ranges of the RankedDB. Especially, a clear difference was observed for the PSAR models with the P-score. The A5%−I95% models with the F-score gave a higher percentage of Hit than Normal in the range of more than ∼10% of RankedDB.

3.1.4. AChE. From the poor results of the normal scorings, especially the F-, G-, and P-scores, this target proved to be difficult for the docking simulation. First, the A5%−I95% and A10%−I90% models for each scoring function gave higher percentages of Hit than Random throughout RankedDB. Second, the A5%−I95% and A10%−I90% models with the F-, C-, and P-score gave higher percentages of Hit than Normal throughout RankedDB. The A5%−I95% models with the D- and C-scores also gave higher percentages of Hit than Normal in the range of more than ∼10% and ∼20%, respectively.

3.2. The Area under the Receiver Operating Characteristic Curve. The area under the receiver operating characteristic curve (AUC) is a convenient indicator for estimating the performance of compound screening. The receiver operating characteristic curve is expressed as the proportion of actives recovered versus the proportion of inactives recovered. The area under the curve is 1.0 for perfect enrichment and 0.5 for random performance, and the area directly measures the probability that a randomly chosen active will be ranked higher than a randomly chosen inactive. Table 3 lists the AUC data for the PSAR models and normal scoring. The PSAR models give higher AUC data than does normal scoring. Among the thresholds examined, the PSAR model has the following tendency: the lower the percentage of pseudo-active compounds, the higher the AUC (i.e., the higher the performance). The best performance is obtained for the A5%−I95% models.

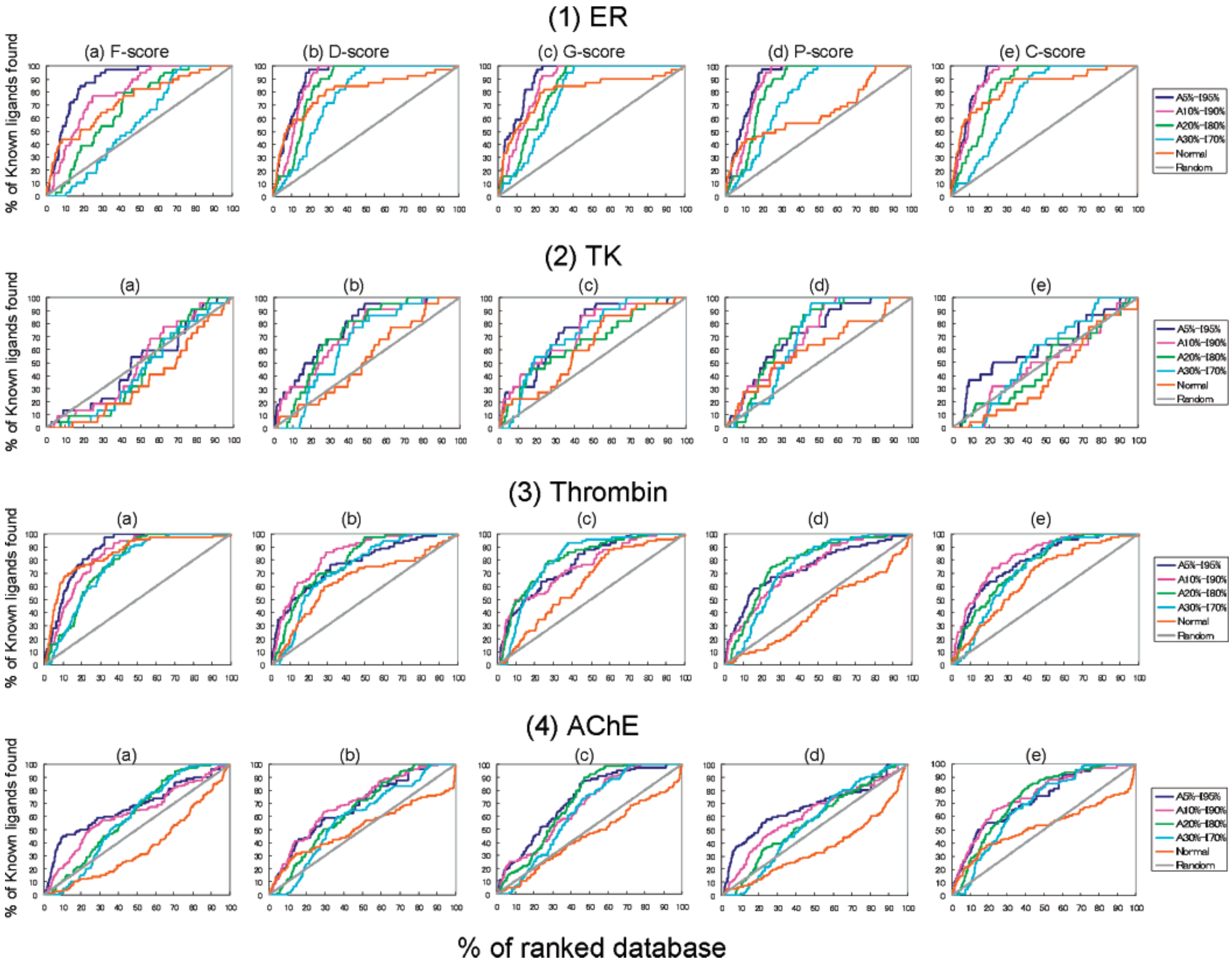

Figure 2. Enrichment curves. "A(*x*)%−I(*y*)%", "Normal", and "Random" represent the PSAR model, normal scoring, and random sampling, respectively.

Table 2. Percentage of Ranked Database When the Known Ligands Found Reached 100% for ER

model	Percentage				
	F-score	D-score	G-score	P-score	C-score
A5%−I95%	49.4	30.2	23.9	30.2	19.7
A10%−I90%	56.3	24.6	31.9	24.6	25.9
A20%−I80%	76.7	32.9	36.6	32.9	36.1
A30%−I70%	72.2	49.4	40.7	49.4	52.7
normal	88.5	97.7	98.6	80.6	83.7

3.3. Enrichment Curve Averaged over Four Targets. Figure 3 shows the enrichment curves averaged over four targets for the indicated scoring functions, which are plotted every 5%. General results that are not dependent on scoring functions were obtained. Interestingly, as the percentage of pseudo-active compounds increased from A5%−I95% to A30%−I70%, the shape changed from parabola-like to sigmoid-like.

In terms of the first standpoint, the A5%−I95% and A10%−I90% models gave higher percentages of Hit than Random throughout RankedDB in all scoring functions. These results were sufficient to support our hypothesis. The A20%−I80% models gave similar results, except for the F-score, which gave a lower percentage only at 5% of RankedDB. The A30%−I70% models also gave lower

Table 3. Area under the Receiver Operating Characteristic Curve

scoring function	Area				
	A5%−I95%	A10%−I90%	A20%−I80%	A30%−I70%	normal
		ER			
F-score	0.902	0.810	0.690	0.562	0.745
D-score	0.928	0.904	0.864	0.779	0.814
G-score	0.931	0.888	0.831	0.790	0.812
P-score	0.928	0.904	0.864	0.779	0.658
C-score	0.937	0.925	0.861	0.758	0.854
		TK			
F-score	0.494	0.499	0.453	0.439	0.360
D-score	0.784	0.733	0.736	0.657	0.532
G-score	0.760	0.740	0.667	0.730	0.628
P-score	0.729	0.712	0.742	0.698	0.615
C-score	0.621	0.491	0.481	0.566	0.404
		Thrombin			
F-score	0.894	0.850	0.789	0.774	0.864
D-score	0.789	0.856	0.806	0.750	0.649
G-score	0.788	0.762	0.823	0.820	0.634
P-score	0.758	0.750	0.791	0.752	0.420
C-score	0.799	0.843	0.774	0.740	0.677
		AChE			
F-score	0.679	0.626	0.616	0.596	0.356
D-score	0.690	0.715	0.655	0.596	0.515
G-score	0.736	0.686	0.720	0.642	0.443
P-score	0.681	0.621	0.568	0.553	0.351
C-score	0.742	0.767	0.752	0.697	0.526

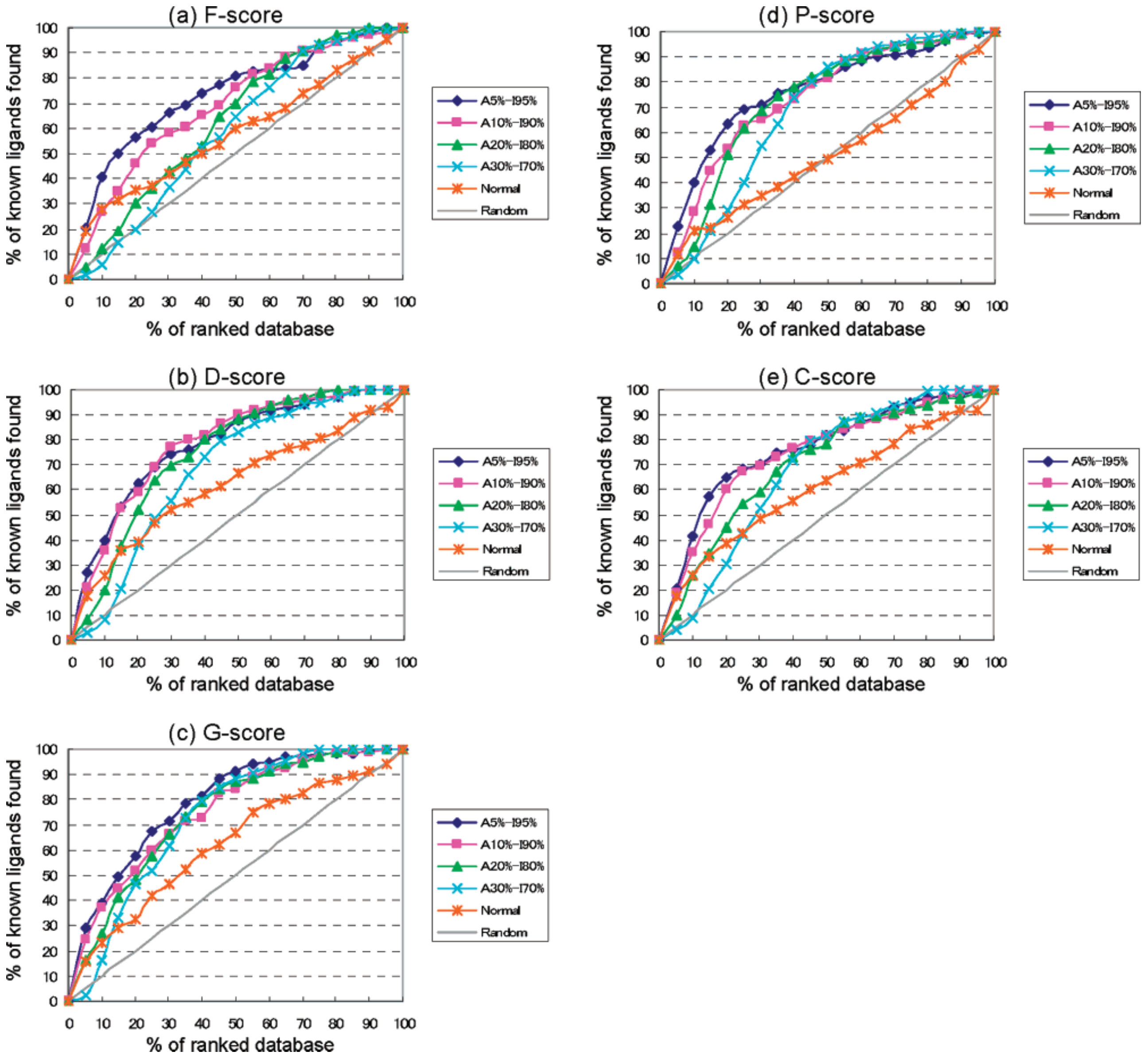

Figure 3. Enrichment curves averaged over four targets. "A(*x*)%–I(*y*)%", "Normal", and "Random" represent the PSAR model, normal scoring, and random sampling, respectively.

percentages of Hit than Random only in the range of <20% of RankedDB at the worst case with F-score.

In terms of the second standpoint, the A5%–I95% model gave higher percentages of Hit than Normal throughout RankedDB in all scoring functions. This result was sufficient to demonstrate the practical usefulness of the PSAR model. The A10%–I90%, A20%–I80%, and A30%–I70% models gave lower percentages of Hit than Normal only in the range of <15%, <30%, and <40% of the RankedDB, respectively, at their worst cases.

As the amount of pseudo-active compounds increased (from A5%–I95% to A30%–I70%), the performance of the PSAR models became poorer in the low range of RankedDB. We consider this observation to be due to an increase of randomness contained in information extracted from learning uncertain positives. However, because the extracted information was not completely random, the moderate percentage of Hit appeared in the middle range of RankedDB. It is difficult to determine the optimal threshold of the pseudo-active compounds. If the percentage of the pseudo-active compounds is too low, the extracted information would not be sufficient to construct the model of the active compound. We considered that a value of 5% of the pseudo-active compounds was an appropriate percentage in this study.

Table 4. Number of Compounds in the Training and Test Sets Used To Compare the PSAR and QSAR Models

	Experimental Known Active/Inactive Compound Descriptors			
set	ER	TK	Thrombin	AChE
training set	19/700	11/400	36/1200	53/1900
test set	20/728	11/486	36/1240	54/1947

4. DISCUSSION

4.1. Comparison of PSAR and QSAR Models. As stated previously, the PSAR and QSAR models are used in different situations: the PSAR model is used when no ligand is available, whereas the QSAR model is used when several ligands are already known. Thus, it is not appropriate to compare their performance on the same grounds. However,

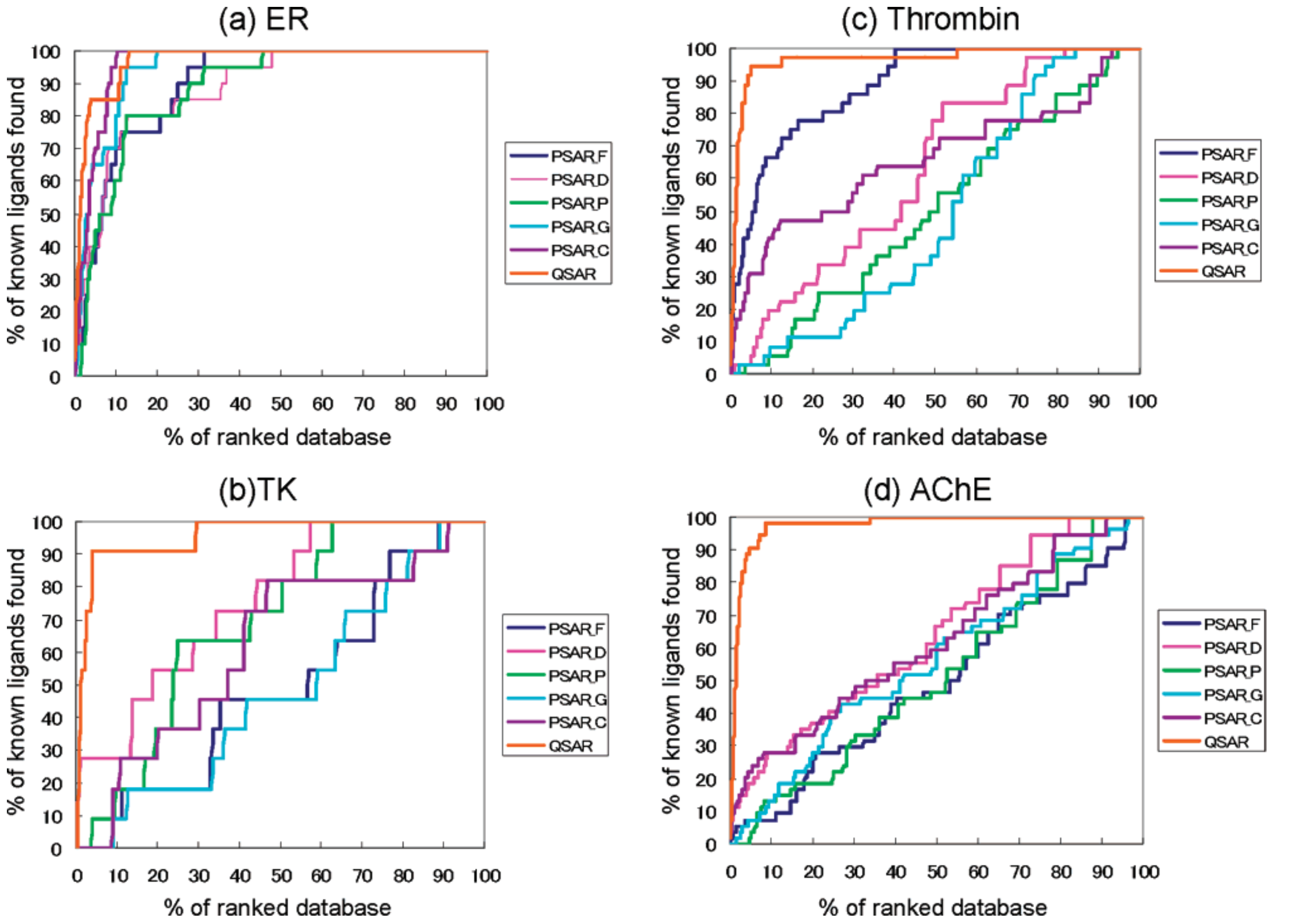

Figure 4. Enrichment curves for comparing PSAR and QSAR models. PSAR_F, PSAR_D, PSAR_P, PSAR_G, and PSAR_C represent the PSAR model of a A5%−I95% model based on the F-, D-, P-, G-, and C-score, respectively. QSAR represents a model obtained using known active and inactive compounds.

it is still interesting to compare them to clarify the significance of the PSAR model.

Table 4 lists the number of compounds in training and test sets for comparing PSAR and QSAR models. Known active and inactive compounds were randomly divided into training and test sets. The PSAR models (with A5%−I95% models) were constructed from the training set. Pseudo-active compounds were obtained after ranking compounds including both known active and inactive compounds. Figure 4 shows the enrichment curve obtained for every target. As expected, the QSAR model provides higher performances than the PSAR models, because positive data is not contaminated by false positives. It is important to note that this fact does not decrease the value of our PSAR model. Rather, this result underscores how important it is to find the first several hits in screening. As a further extreme example, we included only one known ligand in the training set. The obtained QSAR enrichment curve indicates that compounds dissimilar to the ligand (i.e, of different chemical entity) cannot be readily found, whereas compounds similar to the ligand can be found very rapidly (see Figure 5). Thus, we consider the PSAR model that has been developed here to provide an alternative method to find the first hits of various chemical entities, especially for difficult cases in which the docking simulation alone fails to find a hit.

4.2. The PSAR Model Can Access All Compounds. In practice, the docking simulation run sometimes fails for various reasons (file format, unreasonable initial structure, etc.). The PSAR model can remedy this weak point of the docking simulation. For TK and thrombin, our docking simulation runs of some known ligands failed and their docking scores could not be obtained (i.e., these ligands were out of assessment for Normal). On the other hand, the PSAR model can access all ligands.

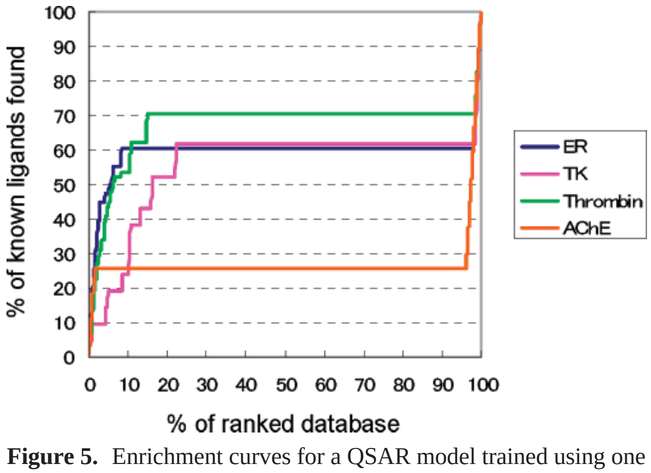

Figure 5. Enrichment curves for a QSAR model trained using one known ligand.

5. CONCLUSION

We hypothesized that “A model of active compound can be provided by integrating information of compounds high-ranked by docking simulation of a random compound library". Interestingly, in our hypothesis, true active compounds do not need to be included in the high-ranked compound. From tests for four functionally different targets, we obtained three conclusions:

(1) The pseudo-structure−activity relationship (PSAR) models significantly gave higher percentages of known ligands found than random sampling, and these results are sufficient to support our hypothesis.

(2) The PSAR models gave higher percentages of known ligands found than normal scoring, and these results demonstrate the practical usefulness of the PSAR model.

(3) The PSAR model can assess compounds that failed in the docking simulation.

Note that the PSAR and quantative structure−activity relationship (QSAR) models are used in different situations: the advantage of the PSAR model emerges when no ligand is available as training data or when one wants to find novel type of ligands, whereas the QSAR model is effective for finding compounds that are similar to known ligands when the ligands are already known.

Our computational result indicated that the screening using our novel combination of docking + PSAR is more efficient, in most cases, than the screening using docking simulation alone. Thus, typical screening using the PSAR model may be performed in the following steps.

(1) Conventional docking simulation of compound library against a target receptor is performed. The compound library is desirably broad in molecular diversity and chemical entities. We can provide no specific recommendation of docking software.

(2) The top 5% and bottom 95% of ranked compounds are classified as pseudo-active and inactive, respectively, and the PSAR model is constructed based on them.

(3) Compounds high-ranked by the PSAR model are subject to an assay.

If real active compounds are detected, one may switch to a QSAR model to accelerate the discovery of closely related analogues. However, note that some active compounds, especially different chemical entities, are likely to be missed by the QSAR alone.

The PSAR model will be more sophisticated by addressing the following issues: (i) choosing an optimal descriptor set and (ii) establishing a method that logically determines the optimal threshold of pseudo-active compounds.

ACKNOWLEDGMENT

This work is supported in part by New Energy and Industrial Technology Development Organization, Japan (NEDO) under the project of Development of Basic Technologies for Advanced Production Methods Using Microorganism Functions.

REFERENCES AND NOTES

(1) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Sheridan, R. P.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol*. **1982**, *161*, 269−288.

(2) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.

(3) Welch, W.; Ruppert, J.; Jain, A. N. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. *Chem. Biol.* **1996**, *3*, 449−462.

(4) Jones, G.; Wilett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.

(5) McMartin, C.; Bohacek, R. QXP: Powerful, rapid computer algorithms for structure-based drug design. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 333−344.

(6) Morris, G. M.; Goodsell, D. S.; Halliday, R.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.

(7) Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D. Flexible docking using Tabu search and an empirical estimate of binding affinity. *Proteins* **1998**, *33*, 367−382.

(8) Hou, T.; J., W.; Chen, L.; Xu, X. Automated docking of peptides and proteins by using a genetic algorithm combined with a tabu search. *Protein Eng.* **1999**, *12*, 639−647.

(9) Liu, M.; Wang, S. MCDOCK: a Monte Carlo simulation approach to the molecular docking problem. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 435−451.

(10) Perola, E.; Xu, K.; Kollmeyer, T. M.; Kaufmann, S. H.; Prendergast, F. G.; Pang, Y. P. Successful virtual screening of a chemical database for farnesyltransferase inhibitor leads. *J. Med. Chem.* **2000**, *43*, 401−408.

(11) Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided. Mol. Des.* **2001**, *15*, 411−428.

(12) Zavodszky, M. I.; Sanschagrin, P. C.; Korde, R. S.; Kuhn, L. A. Distilling the essential features of a protein surface for improving protein. ligand docking, scoring, and virtual screening. *J. Comput. Aided. Mol. Des.* **2002**, *16*, 883−902.

(13) Jain, A. N. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46*, 499−511.

(14) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.

(15) *FlexSIS*. BioSolveIT GmbH: Sankt Augustin, Germany.

(16) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **2004**, *57*, 225−242.

(17) Perola, E.; Walters, W. P.; Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins* **2004**, *56*, 235−249.

(18) Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of docking performance: Comparative data on docking algorithms. *J. Med. Chem.* **2004**, *56*, 235−249.

(19) Kroemer, R. T.; Vulpetti, A.; McDonald, J. J.; Rohrer, D. C.; Trosset, J. Y.; Giordanetto, F.; Cotesta, S.; McMartin, C.; Kihlen, M.; Stouten, P. F. W. Assessment of docking poses: Interactions-based accuracy classification (IBAC) versus crystal structure deviations. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 871−881.

(20) Kontoyianni, M.; Sokol, G. S.; McClellan, L. M. Evaluation of library ranking efficacy in virtual screening. *J. Comput. Chem.* **2005**, *26*, 11−22.

(21) Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912−5931.

(22) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **1999**, *42*, 5100−5109.

(23) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases: 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.

(24) Stahl, M.; Rarey, M.; Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035−1042.

(25) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus scoring for ligand/protein interactions. *J. Mol. Graph. Model.* **2002**, *20*, 281−295.

(26) Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46*, 2287−2303.

(27) Jacobsson, M.; Liden, P.; Stjernschantz, E.; Bostrom, H.; Norinder, U. Improvement structure-based virtual screening by multivariate analysis of scoring data. *J. Med. Chem.* **2003**, *46*, 5781−5789.

(28) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T. M.; Murray, C. W.; Taylor, T. D.; Watson, P. Virtual screening using protein−ligand docking: Avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 793−806.

(29) Yang, J. M.; Chen, Y. F.; Shen, T. W.; Kristal, B. S.; Hsu, D. F. Consensus scoring criteria for improving enrichment in virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 1134−1146.

(30) Yoon, S.; Smellie, A.; Hartsough, D.; Filikov, A. Surrogate docking: structure-based virtual screening at high throughput speed. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 483−497.

(31) Cherkasov, A.; Ban, F.; Li, Y.; Fallahi, M.; Hammond, G. L. Progress Docking: A Hybrid QSAR/Docking Approach for Accelerating In Silico High Throughput Screening. *J. Med. Chem.* **2006**, *49*, 7466−7478.

(32) Tripos; St. Louis, MO.

(33) Huang, N.; Schoichet, K. B.; Irwin, J. J.; Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.

(34) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.

(35) R Development Core Team. R: *A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2005.

(36) Svetnki, 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. Model.* **2003**, *43*, 1947−1958.

(37) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to Molecules in the Training Set Is a Good Discriminator for Prediction Accuracy in QSAR. *J. Chem. Inf. Model.* **2004**, *44*, 1912−1928.

(38) Svetnik, V.; Wang, T.; Tong, C.; Liaw, A.; Sheridan, R. P.; Song, Q. Boosting: An Ensemble Learning Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Model.* **2005**, *45*, 786−799.

(39) Plewczynski, D.; Spieser, S. A. H.; Koch, U. Assessing Difference Classification Methods for Virtual Screening. *J. Chem. Inf. Model.* **2006**, *46*, 1098−1106.

(40) Bruce, C. L.; Melville, J. L.; Pickett, S. D.; Hirst, J. D. Contemporary QSAR Classifiers Compared. *J. Chem. Inf. Model.* **2007**, *47*, 219−227.

(41) Ehrman, T. M.; Barlow, D. J.; Hylands, P. J. Virtual Screening of Chinese Herbs with Random Forest. *J. Chem. Inf. Model.* **2007**, *47*, 264−278.

(42) Palmer, D. S.; O'Boyle, N. M.; Glen, R. C.; Mitchell, J. B. O. Random Forest Models To Predict Aqueous Solubility. *J. Chem. Inf. Model.* **2006**, *47*, 150−158.

(43) Teramoto, R.; Fukunishi, H. Supervised consensus scoring for docking and virtual screening. *J. Chem. Inf. Model.* **2007**, *47*, 526−534.

CI7003384