JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

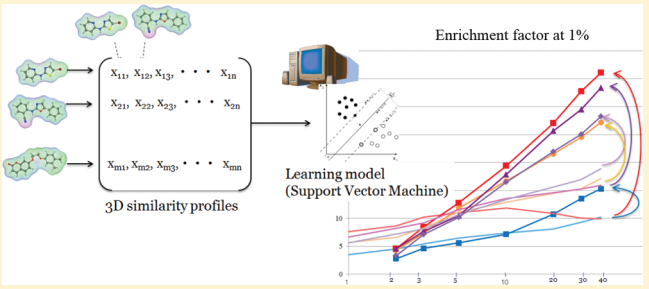
Application of Support Vector Machine to Three-Dimensional Shape-Based Virtual Screening Using Comprehensive Three-Dimensional Molecular Shape Overlay with Known Inhibitors

Tomohiro Sato, Hitomi Yuki, Daisuke Takaya, Shunta Sasaki, Akiko Tanaka, and Teruki Honma*

RIKEN Systems and Structural Biology Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan

Supporting Information

ABSTRACT: In this study, machine learning using support vector machine was combined with three-dimensional (3D) molecular shape overlay, to improve the screening efficiency. Since the 3D molecular shape overlay does not use fingerprints or descriptors to compare two compounds, unlike 2D similarity methods, the application of machine learning to a 3D shape-based method has not been extensively investigated. The 3D similarity profile of a compound is defined as the array of 3D shape similarities with multiple known active compounds of the target protein and is used as the explanatory variable of support vector machine. As the measures of 3D shape similarity for our new prediction models, the prediction performances of the 3D shape similarity metrics implemented in ROCS, such as ShapeTanimoto and ScaledColor, were validated, using the known inhibitors of 15 target proteins derived from the ChEMBL database. The learning models based on the 3D similarity profiles stably outperformed the original ROCS when more than 10 known inhibitors were available as the queries. The results demonstrated the advantages of combining machine learning with the 3D similarity profile to process the 3D shape information of plural active compounds.



INTRODUCTION

In recent years, virtual screening has become widely used in the early stages of drug discovery projects. Considering the expense of the experimental high throughput screening (HTS) of millions of compounds, virtual screening is frequently applied to reduce the number of compounds subjected to such screening.

The computational methods used for virtual screening can be roughly categorized into ligand-based methods and structure-based methods. When the X-ray structure of a target protein is not available, but the active compounds for the protein have been published, ligand-based two-dimensional (2D) and three-dimensionl (3D) similarity searches are often used. These ligand-based discovery methods are based on the principle that similar compounds tend to share similar biological activities. Typically, the 2D similarity is calculated by the similarity metrics, such as the Tanimoto coefficient (Tc), between two molecular descriptors or fingerprints, such as the MACCS structural keys[1] and ECFP4,[2] derived from the 2D structures of the query molecule and the database compounds.[3] The 2D similarity methods are reportedly able to detect active compounds sharing the same scaffold with considerable efficiency, as compared to other screening methods. The 2D similarity methods are useful for analyzing structure−activity relationships around a particular scaffold. On the other hand, the abilities of the 2D similarity methods to identify novel scaffolds are often lower than those of other methods.[4,5] In the drug discovery process, especially in the hit-to-lead stage, scaffold hopping is very important to improve the ADME and toxicity profiles of the hit compounds; however, it is difficult to discover novel scaffolds by only 2D similarity methods.

The 3D molecular shape overlay method is a ligand-based screening procedure that aligns the 3D structures of database compounds with known active compounds, and evaluates the similarities of the molecular shapes and/or geometries of the pharmacophore features.[6−8] As compared to the 2D similarity methods, the 3D molecular shape overlay does not depend on the 2D similarity of the molecular scaffolds and is expected to provide better scaffold hopping ability. In principle, virtual screening using docking simulation is expected to be a more powerful tool for scaffold hopping, because the docking methods can detect novel chemicals with different binding modes or binding sites. Various docking programs, as exemplified by DOCK,[9] AutoDock,[10] and GLIDE,[11] have previously been reported and provided numerous successful results.[12] However, the usefulness of the docking-based approaches is largely dependent on the properties and available information of the drug discovery targets. For example, the lack of experimental 3D structures and the issue of protein flexibility make docking-based approaches difficult to use for drug discovery.

According to previous studies, the ligand-based 2D and 3D similarity methods generally show better enrichments in virtual

Received: November 25, 2011
Published: March 17, 2012

dx.doi.org/10.1021/ci200562p | J. Chem. Inf. Model. 2012, 52, 1015−1026

screening benchmarks than the docking programs, with drastically shorter computation times.[13−15] In some cases, the scaffold hopping ability of the 3D similarity method is reportedly equivalent to those of the docking-based virtual screening methods.[13] Considering both hit rates, the enrichment factor and the scaffold hopping ability, the 3D similarity methods are well balanced, as compared with the 2D similarity and protein−ligand docking methods. As programs for 3D similarity searches, FlexS,[14] ROCS (Rapid Overlay of Chemical Structures)[17−19] (OpenEye, Inc.), and PHASE[20] have been developed and used for drug discovery.[21−30]

When multiple active compounds are available for a target, machine learning is another alternative, in addition to a simple similarity search. Successful applications of machine learning for various issues in virtual screening were recently reported.[31−45] Machine learning is a statistical approach to build discrimination/regression models, based on the given training examples. Recently developed machine learning techniques, such as support vector machine (SVM)[46] and random forest,[47] are known for their ability to create complex nonlinear learning models with good prediction performance. These methods are better at utilizing the information about plural known active compounds than the conventional methods, based on the maximum similarity value among the active compounds, and often show enhanced screening efficiencies. The 2D similarity methods can be easily combined with machine learning techniques, because the molecular fingerprints or descriptors for the 2D similarity calculations can be directly used as the explanatory variables for machine learning. Since the 3D molecular shape overlay methods do not directly provide such fingerprints or descriptors, there are fewer applications of machine learning to the 3D similarity methods, as compared to the 2D similarity methods.

In this study, SVM was applied to the 3D molecular shape overlay to improve the screening performance. To obtain the descriptors representing the 3D shape of a compound, alignment-free descriptors (such as alignment-free TAE descriptors,[48] property-encoded surface translation (PEST) descriptors,[49] and shape signature[50,51]) and overlay with reference shape (shape fingerprint[52]) were previously reported. These methods calculate the descriptors independently from known active compounds for a target protein. In this study, the descriptor named 3D similarity profile was developed for virtual screening. For this purpose, the idea of empirical kernel map[53−55] was applied to calculate the descriptors, using known active compounds as the reference structures, because their structural features should be important for binding with the target protein.

To calculate the input data for SVM, each compound in a compound database was superposed on all known active compounds using ROCS and was represented by a vector consisting of the resulting 3D shape similarities to the active compounds, named the 3D similarity profile. The efficiencies of the SVM models based on the 3D similarity profiles were tested in the virtual screenings of known inhibitors of 15 target proteins derived from the ChEMBL database[56] and were compared to the simple 3D shape similarity-based screening.

MATERIALS AND METHODS

Data Set. Inhibitors for various target proteins were collected from the ChEMBL database. ChEMBL is a medicinal chemistry database containing biological activity and/or binding affinity data between various compounds and proteins. From ChEMBL, 15 target proteins, for which at least 700 inhibitors were reported with IC_{50} values less than 10 μM, were selected as the screening targets to evaluate the screening methods (Table 1). These target proteins were selected, considering the number of inhibitors and the availability of X-ray structures. Therefore, this data set can be commonly used for the validation of both ligand-based and structure-based approaches. To avoid the bias of selecting too many targets from a certain protein family, the number of targets selected from a protein family was limited to 2. From the collected inhibitors, the unacceptable compounds for small molecule drugs, which violated the following rules and thus extended the rule of 5,[57] were excluded: 150 ≤ molecular weight ≤ 600, the number of nitrogens + the number of oxygens ≤ 12, the number of hydrogen bonding donors ≤ 6, and AlogP ≤ 6. When selecting the compounds for the data set for the validation of a machine learning-based method, it is necessary to avoid placing many similar compounds in the data set for training, to prevent the creation of an overestimating learning model that is overfitted to limited chemical structures. To obtain a structurally diverse data set, the inhibitors of each target protein were clustered into 100 clusters, according to their ECFP4 fingerprints. From each cluster, the center compound was chosen by maximum dissimilarity method,[58−60] resulting in a data set containing 100 active compounds for each of the 15 target proteins. Then, each data set was randomly divided into a training set for the SVM model building and an external test set for validation, so that both data sets contained 50 active compounds. As the decoys, 4950 compounds were randomly selected from the ZINC drug-like subset[61] for both the training and test sets. As a result, a training set and a test set containing 50 actives and 4950 decoys were created for all 15 target proteins. These active and decoy compounds were provided as the Supporting Information. As the 3D structures of the compounds for the ROCS calculation, the conformations of the compounds were generated using Omega.[62] The MMFF94s_NoEstat force field was used to build the model, and the default settings of Omega were used

Table 1. Data Set of Target Proteins and Their Inhibitors Derived from ChEMBL

protein name	number of assayed compounds	number of inhibitors	number of rule of 6 inhibitors
carbonic anhydrase 2	3402	2630	2099
MMP1	3706	1954	1784
integrin beta-3	2363	1777	1277
GluR1	3560	1866	1750
D2 dopamine receptor	5409	3362	3121
adenosine receptor A1	3322	1981	1836
VEGFR2	4545	3103	2858
CDK2	4053	2003	1919
estrogen receptor	3486	1459	1143
PPAR-gamma	2586	1234	667
factor Xa	6024	4353	3737
prothrombin	6529	3566	2806
cholinesterase	1670	788	521
cathepsin K	1777	1326	1147
HIV1 reverse transcriptase	3016	1782	1530

for the other parameters. Finally, the conformation with the lowest MMFF94s_NoEstat energy value was selected for each of the compounds.

Evaluation of Screening Methods. The screening efficiencies of the ROCS similarity metrics and the SVM models were evaluated, using enrichment factor at 1% (EF1) and ROC score. EF is one of the most popular measures for evaluating the screening efficiency, and it indicates the ratio of the number of active compounds obtained by in silico screening against that generated by random selection at the predefined sampling percentage.[63] EF is often used to evaluate the early recognition property of screening methods at a certain range, such as 1%. ROC score evaluates the entire range (0−100% sampling) of the screening efficiency, while EF evaluates the screening efficiency of only a particular sampling percentage. ROC score is defined as the area under the receiver operation curve, which plots the ratio of true positive samples (detected active compounds) on the axis of false positive fractions and ranges from 0 (0%) to 1 (100%).[64] These measures were used to evaluate the screening efficiency of the methods in this study.

To assess the influence of the number of known active compounds used as the queries on the performance of the 3D shape similarity search, query subsets consisting of 1, 2, 5, 10, 20, 30, and 40 active compounds were created by random selection from the 50 training set active compounds for each of the 15 target proteins (Figure 1). In the cases of query subsets

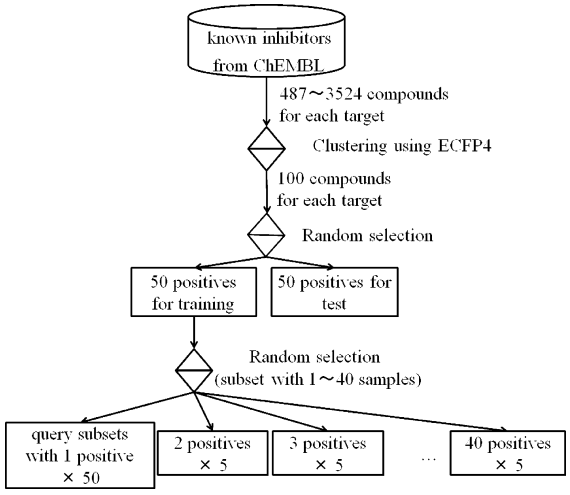


Figure 1. Selection procedure of the known inhibitors for training set and test set.

consisting of only one query compound, 50 query subsets were created using one of the 50 active compounds for training. As query subsets including 2−40 active compounds, 5 subsets with different members were created for each query size. The screening efficiencies using a certain query size were assessed by the averages of the EF1 values and the ROC scores of the results of an external test set, obtained using the 50 (single query) or five (multiple query) subsets of the training set, to avoid bias in the query selection. For specific cases, a paired *t* test was performed to validate the difference of the EF1 values and the ROC scores. The EF1 values (or ROC scores) obtained by two screening methods about each of the 15 targets were regarded as paired values, and the two sets of 15 EF1 values from the different screening methods were validated to assess the significance of the difference. The significance level of $p = 0.05$ was used in this study.

ROCS. The 3D molecular shape overlay procedure using ROCS consists of two steps: the alignment of two molecules and the evaluation of the similarity based on the Gaussian model of molecular shape. In this study, the following 3D shape similarity metrics were used to select the optimal alignment and to evaluate the similarity. (1) ShapeTanimoto, the ratio of the overlapping region of the two molecules to the whole volume, scaled from 0 to 1. (2) ScaledColor, the ratio of the matched region of color atoms (pharmacophore features) to all the color atoms in the query molecule, scaled from 0 to 1. Only the atoms that spatially overlap and have the same color (pharmacophore type) are regarded as the matched color atoms. (3) ColorTanimoto, the Tanimoto coefficient using the volume of the color atoms (pharmacophore features) of compound A, the volume of those of compound B, and the properly aligned volumes of compounds A and B. ColorTanimoto ranges from 0 to 1. (4) TanimotoCombo, a summation of ShapeTanimoto and ColorTanimoto. (5) ComboScore, a summation of ShapeTanimoto and ScaledColor. Using ROCS, the compounds of the screening source were overlaid onto a query compound and evaluated according to the metrics. When several active compounds were available as queries, the best score among those from the alignments between the database compound and the queries was employed to rank the compound.

Support Vector Machine. To utilize the results of the 3D molecular shape overlays for plural queries for in silico screening, machine learning using SVM was adopted in this study. As the explanatory variables, a compound was represented by the 3D similarity profile, as described in the next section. The SVM models were built by learning the descriptors of the training set active compounds as positive examples, and those of the training set decoys as negative examples. The SVM models were optimized by 5-fold cross validation and were evaluated using the test set. The SVM models using only two active compounds were created without cross-validation. The SVM models using three active compounds were optimized by 3-fold cross-validation.

SVM models nonlinearly discriminate two classes of compounds, by mapping the data vectors to a very high-dimensional descriptor space and finding the hyperplane that separates the two classes with the largest margin. The most significant difference between SVM and simple linear discrimination is the so-called "kernel trick". In this study, a radial basis function (RBF) kernel was used to obtain a complicated nonlinear separating hyperplane. The gamma for the RBF kernel and the "C" value of the constant for the slacks variant were optimized by 5-fold cross-validation. Other parameters were set to the default values of Pipeline Pilot.[65] The performance of the SVM models was compared to those of the conventional similarity search, which evaluates compounds by the maximum similarity value among the active compounds (MaxSim method), using the same training positive compounds as the queries.

SVM Models Using 3D Similarity Profile. To apply SVM to the 3D molecular shape overlay results, the idea of empirical kernel map[53−55] was utilized to develop the 3D similarity profile, which describes the properties of a compound by multiple 3D shape similarity values with known active compounds. First, a compound from the target database was superposed on all of the active compounds, and the similarity

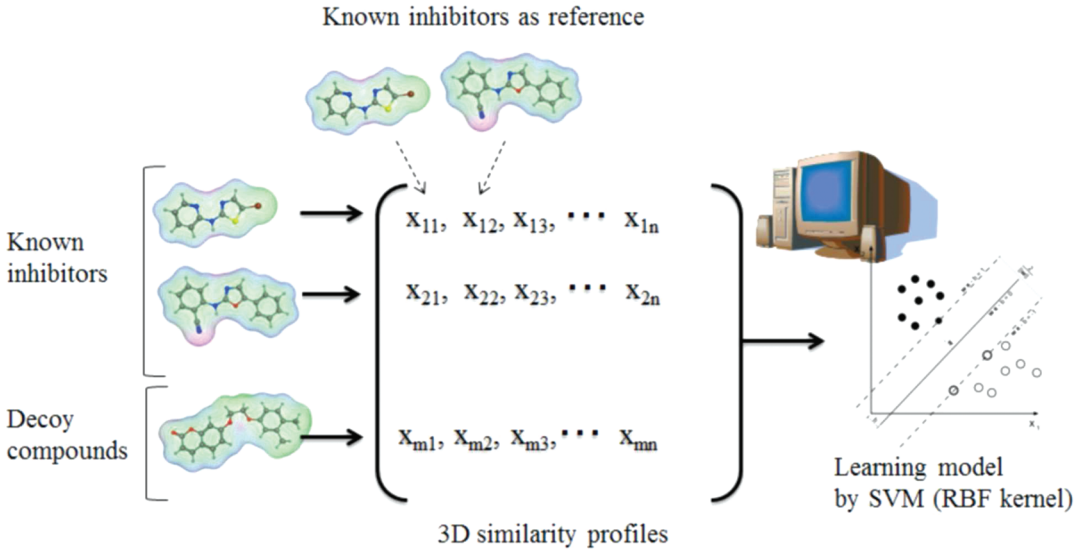


Figure 2. Scheme of SVM learning based on 3D similarity profiles. X_{ij} represents the 3D shape similarity (ShapeTanimoto, ScaledColor, ColorTanimoto, ComboScore, or TanimotoCombo) between the *i*th compound (the known inhibitor as a positive example, or the decoy compound as a negative example) and the *j*th compound (the known inhibitor used as a reference structure).

metrics, such as ShapeTanimoto or ScaledColor, with the active compounds were calculated using ROCS. Then, the 3D similarity profile was created, by arraying the calculated 3D shape similarity values (Figure 2). In a previous study, the SVM models generated by an analogous method, in which the compounds were represented by arrays of 2D similarities to reference compounds, showed successful results.[66] In this study, both the active and decoy compounds in the training set were aligned to the active compounds in the training set, and the resulting 3D similarity profiles of the training set were learned, to build the SVM models. To evaluate the screening efficiencies, both the active and decoy compounds in the test set were aligned to the active compounds in the training set, and the resulting 3D similarity profiles of the test set were input into the SVM models, to make the predictions. The screening efficiencies of the SVM models were assessed by the EF1 values and the ROC scores of the predictions of the test set compounds. The SVM models using 3D similarity profiles, consisting of ShapeTanimoto, ColorTanimoto, ScaledColor, ComboScore, and TanimotoCombo, were respectively evaluated and were compared to the screening results using Tc of the corresponding 3D shape similarity metrics. In addition, the SVM models based on the 3D similarity profile, using combinations of two different 3D shape similarity metrics, were assessed for further improvement of the SVM models based on 3D similarity profiling.

RESULTS AND DISCUSSION

Comparison of 3D Shape Similarity Metrics in a Simple Similarity Search. The average EF1 value and ROC score of each 3D shape similarity metrics for the 15 targets are plotted in Figure 3. All of the average values were calculated from 50 runs of similarity search, using each one of the 50 known active compounds as a single query. Tables 2 and 3 show the averages and standard deviations of the EF1 values and the ROC scores, respectively. ScaledColor generated the best EF1 value (7.64) and ROC score (0.603) among the five 3D shape similarity metrics, in the average of the screening for the 15 target proteins (Figure 3). After ScaledColor, Combo-

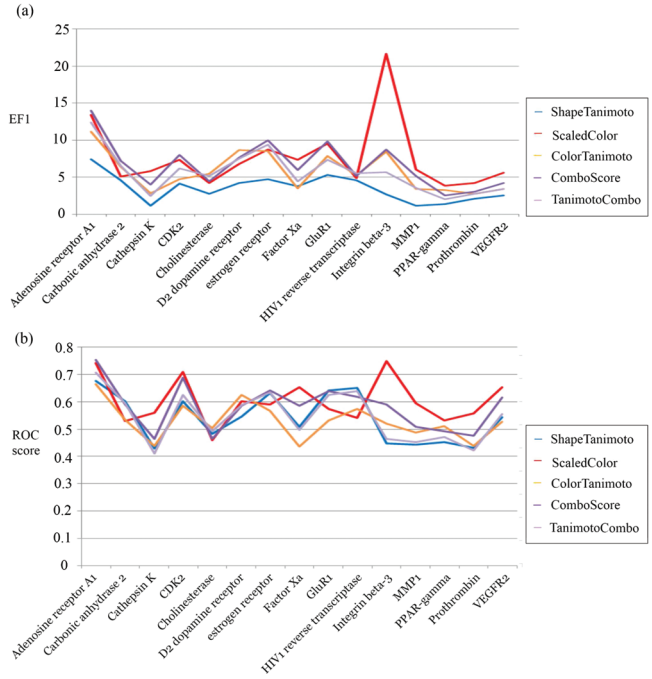


Figure 3. Results of ROCS, using a single known inhibitor as a query. (a) EF1 values of about 15 target proteins (mean of all of the results using each of the queries). (b) ROC scores of about 15 target proteins (mean of all of the results using each of the queries).

Score (EF1 6.67, ROC score 0.581), ColorTanimoto (5.70, 0.530), TanimotoCombo (5.62, 0.545), and ShapeTanimoto (3.51, 0.539) followed, in this order. To assess the statistical significance of the EF1 and ROC scores of ScaledColor, paired *t*-tests between ScaledColor and other metrics were performed. The *p*-values of the EF1 values and the ROC scores of ScaledColor and ComboScore were 0.151 and 0.114, respectively. The *p*-values of the EF1 values and the ROC scores of ScaledColor and the other metrics were less than 0.041 and 0.026, respectively. Thus, the EF1 value and the ROC score of ScaledColor were significantly higher than those

Table 2. Average EF1 Values (Standard Deviation) of Single Query ROCS[a]

	CT	CS	SC	ST	TC
adenosine receptor A1	11.1 (7.1)	14.0 (8.7)	13.4 (7.3)	7.4 (7.2)	12.4 (8.4)
carbonic anhydrase 2	6.4 (5.6)	7.2 (6.2)	5.1 (4.9)	4.6 (4.5)	6.5 (5.2)
cathepsin K	2.8 (2.7)	4.0 (3.6)	5.8 (4.3)	1.1 (1.5)	2.5 (2.6)
CDK2	4.7 (4.1)	8.0 (4.8)	7.4 (5.1)	4.2 (4.0)	6.2 (4.3)
cholinesterase	5.4 (4.5)	4.4 (3.6)	4.2 (3.8)	2.8 (2.3)	5.2 (4.6)
D2 dopamine receptor	8.7 (6.4)	7.6 (6.2)	6.8 (4.8)	4.2 (4.0)	7.5 (5.9)
estrogen receptor	8.5 (10.3)	10.0 (11.0)	8.8 (10.3)	4.8 (5.5)	9.4 (10.5)
factor Xa	3.5 (3.8)	6.0 (5.4)	7.4 (6.1)	3.8 (3.8)	4.4 (4.6)
GluR1	7.8 (5.5)	9.8 (7.4)	9.5 (6.4)	5.3 (4.6)	7.4 (6.0)
HIV1 reverse transcriptase	5.2 (4.1)	5.2 (4.8)	4.8 (3.8)	4.6 (5.4)	5.6 (5.0)
integrin beta-3	8.5 (8.4)	8.7 (8.6)	21.7 (12.9)	2.7 (3.0)	5.7 (6.9)
MMP1	3.4 (2.8)	5.2 (4.0)	6.0 (4.6)	1.2 (1.6)	3.6 (3.1)
PPAR-gamma	3.2 (2.9)	2.5 (2.7)	3.9 (4.3)	1.4 (1.5)	2.0 (2.3)
prothrombin	2.8 (2.5)	3.0 (2.7)	4.2 (3.5)	2.1 (2.2)	2.8 (2.3)
VEGFR2	3.4 (2.6)	4.2 (3.2)	5.6 (3.1)	2.6 (2.6)	3.4 (2.7)
15 targets	5.7 (6.0)	6.7 (6.7)	7.6 (7.7)	3.5 (4.3)	5.6 (6.1)

[a]CT ColorTanimoto, CS ComboScore, SC ScaledColor, ST ShapeTanimoto, TC TanimotoCombo.

Table 3. Average ROC Score (Standard Deviation) of Single Query ROCS[a]

	CT	CS	SC	ST	TC
adenosine receptor A1	0.665 (0.101)	0.753 (0.118)	0.741 (0.092)	0.677 (0.123)	0.706 (0.119)
carbonic anhydrase 2	0.535 (0.092)	0.590 (0.144)	0.530 (0.086)	0.601 (0.145)	0.594 (0.142)
cathepsin K	0.439 (0.077)	0.464 (0.091)	0.559 (0.077)	0.429 (0.068)	0.410 (0.077)
CDK2	0.585 (0.092)	0.687 (0.079)	0.710 (0.082)	0.602 (0.061)	0.626 (0.077)
cholinesterase	0.503 (0.063)	0.464 (0.041)	0.459 (0.068)	0.482 (0.034)	0.491 (0.029)
D2 dopamine receptor	0.625 (0.124)	0.586 (0.090)	0.601 (0.131)	0.545 (0.051)	0.588 (0.054)
estrogen receptor	0.567 (0.122)	0.640 (0.109)	0.590 (0.113)	0.633 (0.093)	0.633 (0.106)
factor Xa	0.436 (0.062)	0.584 (0.075)	0.653 (0.050)	0.508 (0.086)	0.497 (0.087)
GluR1	0.533 (0.099)	0.639 (0.160)	0.575 (0.097)	0.642 (0.168)	0.626 (0.164)
HIV1 reverse transcriptase	0.573 (0.079)	0.618 (0.132)	0.541 (0.077)	0.651 (0.136)	0.639 (0.130)
integrin beta-3	0.520 (0.149)	0.591 (0.123)	0.749 (0.101)	0.447 (0.104)	0.464 (0.138)
MMP1	0.487 (0.067)	0.509 (0.081)	0.594 (0.064)	0.443 (0.066)	0.453 (0.075)
PPAR-gamma	0.512 (0.070)	0.491 (0.061)	0.532 (0.071)	0.452 (0.063)	0.471 (0.060)
prothrombin	0.437 (0.057)	0.476 (0.070)	0.558 (0.049)	0.430 (0.074)	0.423 (0.070)
VEGFR2	0.526 (0.061)	0.616 (0.043)	0.653 (0.053)	0.544 (0.038)	0.554 (0.045)
15 targets	0.530 (0.112)	0.581 (0.130)	0.603 (0.116)	0.539 (0.129)	0.545 (0.133)

[a]CT ColorTanimoto, CS ComboScore, SC ScaledColor, ST ShapeTanimoto, TC TanimotoCombo.

of the other metrics, except for ComboScore. Considering that ComboScore was calculated as the summation of ScaledColor and ShapeTanimoto, the screening efficiency of ComboScore partially depends on ScaledColor, and it is reasonable that the *p*-values between ComboScore and ScaledColor are relatively large. The importance of pharmacophore-based 3D similarity metrics was also shown in the previous study by McGaughey.[17] The reason why the pharmacophore-based 3D similarity metrics outperformed the shape-based metrics in our case could be the use of shape information for the filtering of the initial alignments. Before the final evaluation of the molecular overlay using ShapeTanimoto, ScaledColor, etc., ROCS selects the initial alignments of the compounds by the shape complementarity, to ensure reasonable shape matching. The alignments showing extremely low shape complementarity, despite reasonable pharmacophore matching, could be omitted by the initial selection. Thus, the use of pharmacophore-based metrics as the decisive parameter could be the implicit combination of shape-based metrics and pharmacophore-based metrics in this study.

Considering the results for each target protein, the integrin beta-3 inhibitors showed a significant preference for ScaledColor over the other metrics. The integrin beta-3 inhibitors bind to the RGD peptide binding site by mimicking the common pharmacophore features of extracellular matrix proteins, including fibrinogen, fibronectin, osteopontin, thrombospondin, and vitronectin.[67] Figure 4 shows the complex structure of integrin beta-3 and its inhibitor, as an example. The average number of rotatable bonds in the integrin inhibitors was 9.25, which is the highest number among those for the 15 target proteins in this study (5.52 in average), indicating that a precise 3D alignment would be difficult to obtain for the integrin beta-3 inhibitors. Comparing the shape-based metrics (ShapeTanimoto) and the pharmacophore-based metrics (ScaledColor and ColorTanimoto), the alignment of the entire shape of such long integrin beta-3 inhibitors would be more difficult than the alignment of several neighboring pharmacophore features of integrin beta-3 inhibitors, which mimic the hydrogen donors/acceptors of the binding proteins. ScaledColor and ColorTanimoto differ in terms of the dependence of the scaling method on the number of pharmacophores in the

dx.doi.org/10.1021/ci200562p | J. Chem. Inf. Model. 2012, 52, 1015−1026

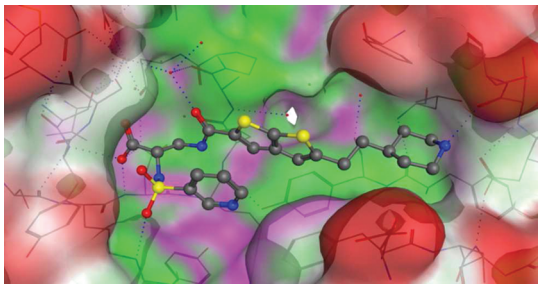


Figure 4. X-ray structure of the complex between integrin beta-3 and the inhibitor, 2-(*S*)-[*N*-(3-pyridylsulfonyl)amino]-3-[[2-carbonyl-5-[2-(piperidin-4-yl)ethyl]-thieno[2,3-*b*]thiopheneyl]amino]-propionic acid (PDB: 2VC2). The hydrophobic surface areas of the protein are shown in green, the polar surface areas are shown in purple, and the exposed areas are shown in red.

screening target compound. ScaledColor sums up the number of matched pharmacophore features with different weights and tolerances defined in the color force field file and is scaled by the perfect coverage of the query pharmacophore features, while ColorTanimoto calculates the ratio of matched pharmacophore features to the total number of pharmacophore features in both the query and target molecules. When the query molecules are large and contain many pharmacophore features, as in the integrin beta-3 inhibitors, the ColorTanimoto values of such large active compounds tend to have a lower distribution than those of the small decoy compounds, because it is difficult to align large compounds properly, and the unmatched pharmacophores in such compounds often generate a larger denominator in ColorTanimoto. In contrast, since ScaledColor evaluates the matched pharmacophore features in the query compound, the unmatched regions of the large active compounds do not cause a deduction, unlike the case of ColorTanimoto, and therefore, large active compounds have an advantage, as compared with small decoy compounds. Thus, ScaledColor detected integrin beta-3 inhibitors more efficiently than ColorTanimoto in this study.

The worst result in the test was recorded for PPAR-gamma. The EF1 values by all of the 3D similarity metrics were only 1.4−3.2, as compared to those of the 15 targets (3.5−7.6 on average). The reason why the EF1 values of all the 3D metrics fell below 4.0 would be the structural features of the PPAR-gamma inhibitors used in this study. First, the PPAR-gamma inhibitors in the data set have relatively large structural diversity (average 2D structural similarity using MACCS fingerprint is 0.391), as compared with the inhibitors of the other 14 targets (0.457 on average for the 14 targets). Second, among all of the compounds in this study, the PPAR-gamma inhibitor set was the only compound set that has both more rotatable bonds (6.57) and fewer hydrogen bond donors/acceptors (5.51) than the average of the 15 targets (5.52 and 6.53, respectively). The increased number of rotatable bonds makes the appropriate molecular overlay difficult, and the fewer hydrogen donors/acceptors make the evaluation using pharmacophore-based metrics difficult. A similar situation was found with cathepsin K. The cathepsin K inhibitors have 8.15 rotatable bonds and 6.72 hydrogen bond donors/acceptors, on average. The average EF1 for cathepsin K among the 3D similarity metrics was also very low (3.24). The structural features mentioned above would make the screening test using PPAR-gamma inhibitors the most difficult, among the 15 target proteins.

In addition to the integrin beta-3 inhibitors, ScaledColor also showed superior performance in detecting other inhibitors that bind to the protein/peptide binding sites, such as cathepsin K, factor Xa, MMP-1, and prothrombin. Scaling the EF1 value of each method by ScaledColor revealed that the inhibitors binding to the protein/peptide binding site showed a significant preference for ScaledColor over the other similarity metrics (Figure 5). The results suggested that ScaledColor generally

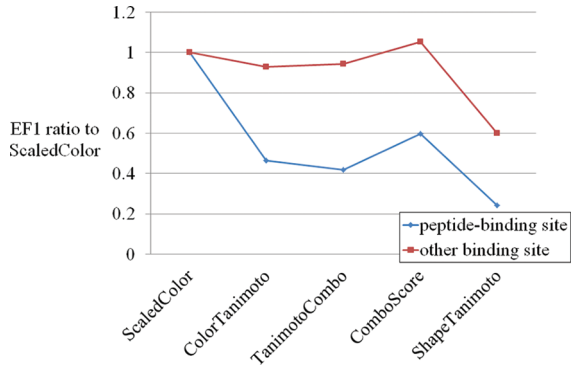


Figure 5. Comparison of the ratio of the EF1 values of the 3D shape similarity metrics to ScaledColor, for the inhibitors targeting peptide-binding sites (cathepsin K, factor Xa, integrin beta-3, MMP-1, and prothrombin) and the others (carbonic anhydrase 2, GluR1, D2 dopamine receptor, adenosine receptor A1, VEGFR2, CDK2, estrogen receptor, PPAR-gamma, cholinesterase, and HIV1 reverse transcriptase).

outperforms the other 3D shape similarity metrics and is especially effective for the detection of large and flexible compounds, such as protease inhibitors or protein−protein interaction inhibitors. For in silico screening using 3D molecular shape overlay, a careful inspection of the known active compounds would facilitate the selection of 3D shape similarity metrics to evaluate target compounds.

To assess the differences of the scaling methods, the difference between the EF1 values of ScaledColor and ColorTanimoto was compared to the number of rotatable bonds and the molecular weights of the inhibitors. Figure 6 shows the scatter plot of the ratio of the EF1 values obtained by ScaledColor and ColorTanimoto (EF1(ScaledColor)/EF1-(ColorTanimoto)) and (a) the number of rotatable bonds and (b) the average molecular weight of the inhibitors for each of the 15 targets. On the vertical axis of the plot, a value greater than 1.0 means that the EF1 value of ScaledColor is larger than that of ColorTanimoto. The square correlation coefficient between the ratio and the number of rotatable bonds was 0.5564, supporting the indication that ScaledColor efficiently detects flexible compounds. The square correlation coefficient between the ratio and the average molecular weight was 0.628, indicating that ScaledColor is especially efficient in detecting larger inhibitors. With an average molecular weight less than 350, ColorTanimoto generated better EF1 values than those by ScaledColor, on average. The results suggested that ColorTanimoto could be useful to detect fragment- or scaffold-size inhibitors.

SVM Models Using 3D Similarity Profiles. The EF1 values and the ROC scores of the conventional MaxSim searches using ROCS and SVM models, based on the 3D similarity profiles using the training set with each size of active compounds, are plotted in Figure 7. Tables 4 and 5 show the

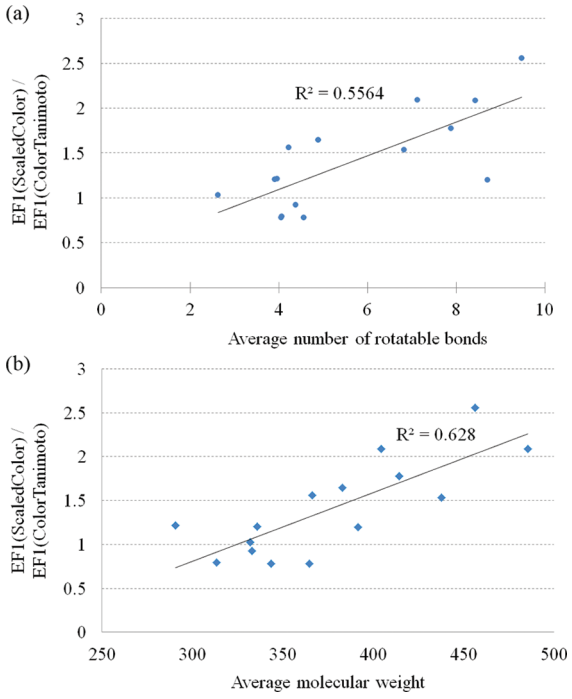


Figure 6. Comparison of the EF1 values by ScaledColor and ColorTanimoto to (a) the average number of rotatable bonds and (b) the average molecular weights, of the inhibitors for each of the 15 targets.

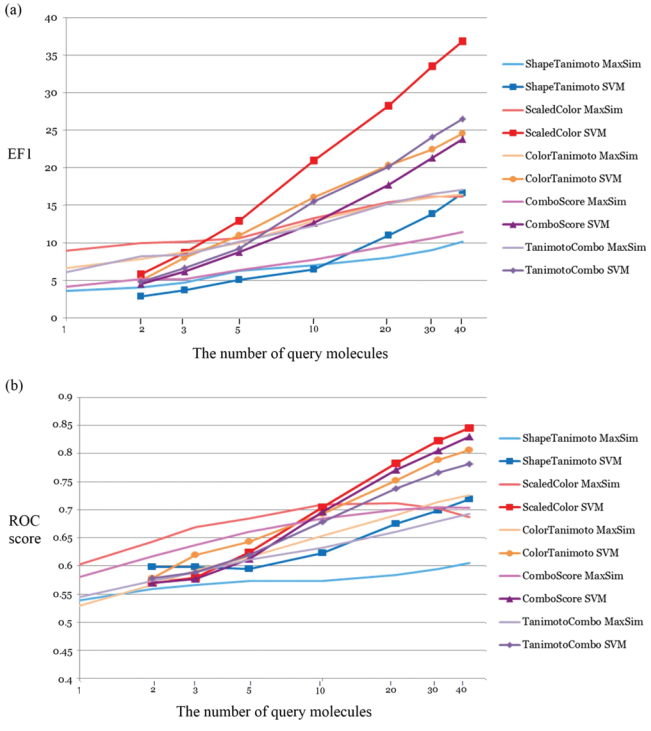


Figure 7. Screening efficiencies of SVM models and conventional 3D shape similarity searches. The horizontal axis represents the number of known inhibitors used as the queries with logarithmic scaling. The vertical axis represents the EF1 (a) and the ROC score (b).

values and standard deviations of the EF1 and ROC scores, respectively. The respective results about each of the 15 target proteins are provided as Supporting Information, in Table S1. The EF1 values and the ROC scores of the SVM models significantly improved when more active compounds were available as the queries. For most of the 3D shape similarity metrics, the SVM models recorded superior results to the conventional MaxSim selection, when more than 10 active compounds were available as queries. In most cases, the SVM models based on ScaledColor showed the best results, among the SVM models and MaxSim selections. After learning 10 active compounds, the SVM models using ScaledColor generated an average EF1 value of 19.4 (improved by 63.4% from that of ScaledColor itself (11.7)). Along with ScaledColor, the ComboScore models produced an average EF1 value of 17.8 (improved by 31.7% from that of ComboScore (13.5)); the ColorTanimoto models generated an average EF1 value of 16.7 (improved by 29.3% from that of ColorTanimoto (12.9)); and the TanimotoCombo models produced an average EF1 value of 16.5 (improved by 23.4% from that of TanimotoCombo (13.4)). The only exception was that the SVM models using ShapeTanimoto generated an average EF1 value of 7.15, while ShapeTanimoto itself produced an average EF1 value of 7.36. As the statistical assessment of the EF1 values using 10 query compounds, paired t tests of the EF1 values from SVM models and MaxSim method, using each of the 3D similarity metrics, resulted in p-values of 5.43×10^{-5} for ScaledColor, 4.06×10^{-3} for ComboScore, 2.13×10^{-4} for ColorTanimoto, and 2.13×10^{-2} for TanimotoCombo. These very low p-values suggested that the SVM models using the above metrics significantly improve the EF1, as compared to those of the conventional similarity searches in the case of using 10 query compounds. Figure 8 shows a plot of the average ratio of the EF1 values obtained by the SVM models and the conventional similarity searches with the corresponding metrics at each number of available queries. When more active compounds were available as queries, the advantage of the SVM models increased. This result suggested that SVM could utilize the information of plural active compounds better than the traditional MaxSim similarity search method. Generally, the use of SVM is recommended when at least 10 active compounds are available as queries. Especially, an improvement of more than 70% in EF1 could be expected when more than 20 active compounds are available.

SVM Models Using 3D Shape Profiles Generated by Two Shape Similarity Metrics. TanimotoCombo and ComboScore are 3D shape similarity metrics that consider both shape complementarity and pharmacophore matching. These metrics were calculated by adding ShapeTanimoto and ColorTanimoto (TanimotoCombo) or ShapeTanimoto and ScaledColor (ComboScore). In the results shown in Figure 7a and b, ScaledColor outperformed both TanimotoCombo and ComboScore in most cases, suggesting that the simple summation of shape-based metrics and pharmacophore-based metrics cannot utilize both information sources properly during machine learning. To create SVM models that can simultaneously consider different types of 3D shape similarity metrics, the arrays of two 3D similarity profiles generated by different 3D shape similarity metrics were used as the input for building SVM models. To improve the screening efficiency of the SVM models using ScaledColor, which displayed the best screening efficiencies, SVM models using the combination of ScaledColor and ShapeTanimoto (pharmacophore-based metrics and shape-based metrics), and the combination of ScaledColor and ColorTanimoto (two pharmacophore-based metrics using different scaling methods) were tested.

dx.doi.org/10.1021/ci200562p | J. Chem. Inf. Model. 2012, 52, 1015−1026

Table 4. Average (Standard Deviation) of EF1 Values about 15 Targets Using Each Size of the Query Active Compounds[a]

	no. of queries	2	3	5	10	20	30	40
ST	MaxSim	4.5 (4.8)	5.4 (5.0)	6.5 (5.1)	7.4 (5.1)	8.1 (5.6)	9.3 (5.8)	10.2 (5.9)
	SVM	2.8 (2.9)	4.6 (4.2)	5.6 (3.9)	7.2 (4.3)	10.8 (5.8)	13.6 (7.1)	15.3 (7.5)
SC	MaxSim	8.6 (7.5)	10.2 (7.8)	11.0 (9.1)	11.9 (9.5)	10.9 (9.8)	10.1 (9.1)	9.9 (9.7)
	SVM	4.6 (4.9)	8.5 (6.7)	12.8 (8.9)	19.4 (11.3)	27.1 (13.0)	32.8 (13.8)	36.1 (14.7)
CT	MaxSim	6.6 (5.5)	8.3 (6.2)	11.1 (8.3)	12.9 (8.9)	14.5 (8.3)	15.3 (8.8)	17.1 (9.7)
	SVM	4.3 (3.7)	7.4 (5.9)	11.8 (8.3)	16.7 (10.1)	21.6 (11.5)	24.5 (12.6)	27.2 (12.2)
CS	MaxSim	8.2 (6.4)	9.2 (6.5)	11.6 (7.9)	13.6 (8.9)	14.6 (8.4)	15.3 (9.0)	15.9 (9.8)
	SVM	4.6 (4.6)	7.7 (5.2)	10.5 (7.2)	17.8 (8.8)	25.7 (10.4)	29.5 (11.3)	33.4 (10.1)
TC	MaxSim	7.2 (5.7)	8.1 (6.0)	10.6 (7.0)	13.4 (8.3)	15.6 (8.3)	17.0 (9.0)	18.9 (10.0)
	SVM	3.3 (3.2)	7.1 (5.0)	10.2 (6.4)	16.5 (7.8)	22.0 (8.5)	25.1 (9.7)	28.2 (9.7)

[a]CT ColorTanimoto, CS ComboScore, SC ScaledColor, ST ShapeTanimoto, TC TanimotoCombo.

Table 5. Average (Standard Deviation) of ROC Scores about 15 Targets Using Each Size of the Query Active Compounds

	no. of queries	2	3	5	10	20	30	40
ST	MaxSim	0.560 (0.133)	0.566 (0.142)	0.573 (0.137)	0.573 (0.139)	0.583 (0.132)	0.595 (0.125)	0.605 (0.122)
	SVM	0.598 (0.071)	0.598 (0.065)	0.594 (0.058)	0.623 (0.058)	0.675 (0.056)	0.699 (0.059)	0.719 (0.059)
SC	MaxSim	0.643 (0.109)	0.669 (0.113)	0.685 (0.111)	0.710 (0.107)	0.712 (0.107)	0.702 (0.128)	0.687 (0.149)
	SVM	0.569 (0.057)	0.579 (0.066)	0.624 (0.086)	0.705 (0.098)	0.783 (0.083)	0.822 (0.074)	0.845 (0.068)
CT	MaxSim	0.566 (0.114)	0.592 (0.120)	0.616 (0.120)	0.654 (0.114)	0.690 (0.100)	0.714 (0.093)	0.726 (0.088)
	SVM	0.577 (0.058)	0.619 (0.075)	0.643 (0.076)	0.693 (0.087)	0.752 (0.076)	0.788 (0.071)	0.806 (0.069)
CS	MaxSim	0.617 (0.125)	0.637 (0.124)	0.661 (0.123)	0.684 (0.112)	0.700 (0.104)	0.705 (0.106)	0.704 (0.112)
	SVM	0.570 (0.050)	0.576 (0.056)	0.612 (0.073)	0.697 (0.073)	0.771 (0.063)	0.805 (0.060)	0.830 (0.061)
TC	MaxSim	0.573 (0.135)	0.589 (0.139)	0.611 (0.133)	0.632 (0.123)	0.660 (0.113)	0.680 (0.106)	0.693 (0.102)
	SVM	0.578 (0.064)	0.589 (0.060)	0.619 (0.076)	0.679 (0.070)	0.738 (0.058)	0.766 (0.058)	0.781 (0.058)

[a]CT ColorTanimoto, CS ComboScore, SC ScaledColor, ST ShapeTanimoto, TC TanimotoCombo.

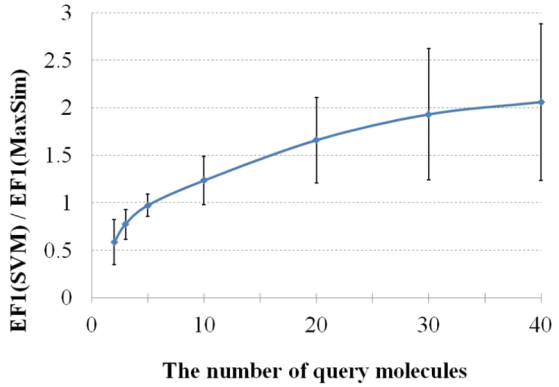


Figure 8. Comparison of the EF1 values between SVM models and 3D shape similarity searches (MaxSim). The values of the vertical axis were calculated by dividing the EF1 values of the SVM models by the EF1 values of the similarity searches, using the corresponding 3D shape similarity metrics. The graph shows the average for all of the metrics and target proteins. The error bars represent the standard deviations. An EF1 ratio of 1, shown by the horizontal dotted line, means that the two methods showed equal efficiency. An EF1 ratio greater than 1 means that the SVM models showed better performance than the conventional 3D shape similarity searches.

SVM Models Using a Combination of ScaledColor and ShapeTanimoto. Figure 9 shows the EF1 and ROC score values of the SVM models built using ScaledColor, ShapeTanimoto, ComboScore (calculated by summing ScaledColor and ShapeTanimoto), and both ScaledColor and ShapeTanimoto as the independent descriptors for the 3D similarity profiles. The 3D similarity profile using both ScaledColor and ShapeTanimoto was described as (SC_1, SC_2, ..., SC_n, ST_1, ST_2, ..., ST_n), where SC_i and ST_i represent the ScaledColor and the ShapeTanimoto between the target compound and the *i*th query compound. In the case of the combination of ScaledColor and ShapeTanimoto, the length of the vector in the 3D similarity profile is doubled, as compared to the other three metrics. The EF1 and ROC score values in Figure 9 are the average values of about 15 targets, using each number of queries. The SVM models based on the combination of ScaledColor and ShapeTanimoto recorded the best EF1 values and ROC scores with any numbers of query molecules as positive examples. The paired *t*-test confirmed that the EF1 values and ROC scores of the SVM models using both ScaledColor and ShapeTanimoto were significantly higher than those using other similarity metrics, with a significance level of $p = 0.05$ at query sizes from 5 to 30. In the case of the EF1 values of the SVM models using both ScaledColor and ShapeTanimoto with 40 query compounds, the *p*-values of the EF1 and ROC scores were 0.06 and 0.009, respectively. Using 40 query compounds, the EF1 values of at least one of the SVM models exceeded the value of 40 for adenosine receptor A1, cathepsin K, cholinesterase, D2 dopamine receptor, glutamate receptor 1, and integrin beta-3. In such cases, the EF1 values of both SVM models showed little difference. Since the combination of ScaledColor and ShapeTanimoto outperformed the individual metrics and their summation (ComboScore), SVM could adequately handle both shape complementarity and pharmacophore matching, thus improving the screening efficiencies.

SVM Models Using a Combination of ScaledColor and ColorTanimoto. Figure 10 shows the results of SVM models obtained using ScaledColor, ColorTanimoto, and both ScaledColor and ColorTanimoto as the independent descriptors. The screening efficiencies of the SVM models using both ScaledColor and ColorTanimoto for the 3D similarity profiles were

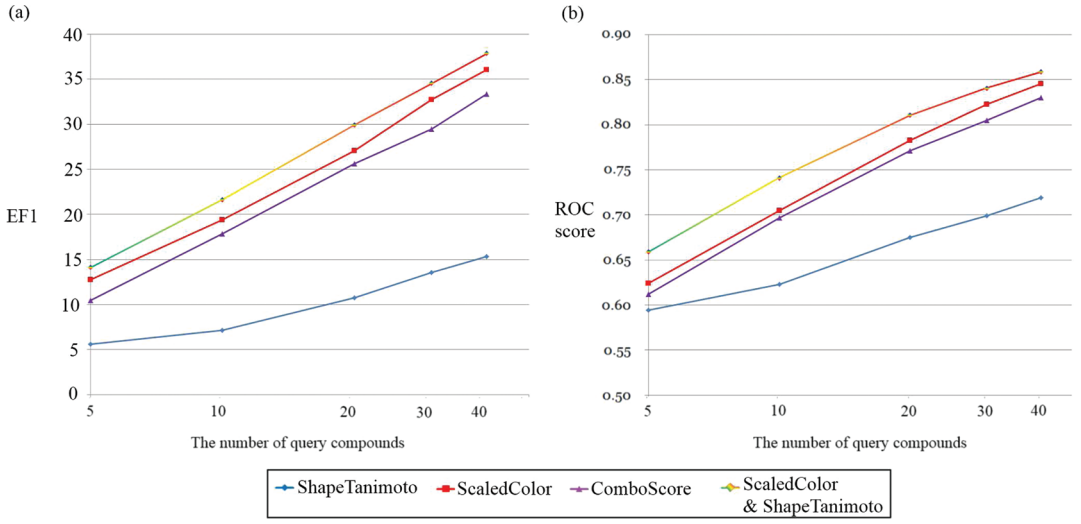


Figure 9. EF1 (a) and the ROC score (b) of SVM models using both ScaledColor and ShapeTanimoto, compared to SVM models using single 3D shape similarity metrics.

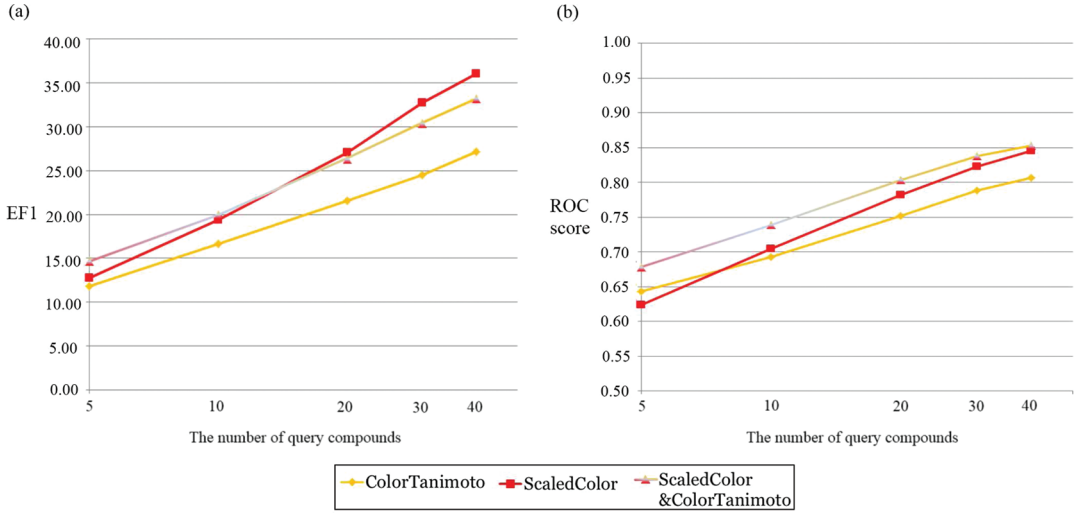


Figure 10. EF1 (a) and the ROC score (b) of SVM models generated using both ScaledColor and ColorTanimoto, compared to SVM models obtained using single 3D shape similarity metrics.

improved, as compared to those of the individual metrics, especially when relatively fewer active compounds were available as the queries. A paired *t*-test with a significance level of $p = 0.05$ showed that the EF1 values of SVM models using both ScaledColor and ColorTanimoto were significantly higher than those of the SVM models using single 3D similarity metrics when 5 query compounds were available, and the ROC scores were significantly higher when fewer than 30 active compounds were available. The difference between ScaledColor and ColorTanimoto lies in the scaling methods. ScaledColor is scaled using the number of pharmacophores only in the query compound, while ColorTanimoto is scaled to the number of pharmacophores in both the query and target compounds. Figure 6 indicates that the two metrics are suitable for detecting compounds with different sizes. The SVM models using both metrics could weight these two metrics properly for each target protein, to improve the screening efficiencies.

Comparing the SVM models using the two 3D similarity metrics, the combination of ScaledColor and ShapeTanimoto recorded higher EF1 values and ROC scores than the combination of ScaledColor and ColorTanimoto. Generally, the combination of highly correlated variables is not useful for building prediction/regression models. It is reasonable that the combination of shape-based metrics and pharmacophore-based metrics would provide more information than the combination of two pharmacophore-based metrics with different scaling methods.

CONCLUSION

In this study, the 3D similarity metrics implemented in ROCS and the use of machine learning theory to 3D molecular shape overlay were comprehensively assessed. Among the 3D similarity metrics, ScaledColor displayed the best screening efficiencies, in terms of both the EF1 value and ROC score, on average for the 15 target proteins. The pharmacophore-based metrics, such as ScaledColor and ColorTanimoto, generally showed better results than the shape-based metrics. ScaledColor was especially efficient in detecting the compounds with long chains, such as integrin beta-3 inhibitors. The source of this advantage of ScaledColor over the other metrics seemed to be the scaling method of the scoring. The scaling method of

dx.doi.org/10.1021/ci200562p | J. Chem. Inf. Model. 2012, 52, 1015−1026

ScaledColor uses only the number of pharmacophores in the query molecules, while the other metrics using the Tanimoto coefficient consider those of both the query and target molecules. When long and flexible compounds are used as the queries, the alignments of a query to the flexible active compounds tend to be more difficult than those to smaller decoy compounds. Thus, the Tanimoto coefficient-based metrics often preferred smaller compounds to larger compounds for flexible queries, resulting in poorer screening efficiencies in such tests, as with the integrin beta-3 inhibitors. Conversely, the EF1 values of ColorTanimoto exceeded those of ScaledColor when the average molecular weight of the known inhibitors was less than 350, indicating that ColorTanimoto would be efficient to detect fragment- or scaffold-sized inhibitors. For further study, the effect of a method that samples the 3D conformation should be assessed. In this study, only the lowest energy conformation calculated by MMFF94s_NoEstat was used for each compound, to reduce the calculation time. This setting certainly made the alignment of large and flexible compounds difficult, as compared to using multiple conformations. A more exhaustive alignment considering multiple conformations could improve the results, especially with the metrics using the Tanimoto coefficient, according to the comparison between ScaledColor and ColorTanimoto in the case of long and flexible inhibitors.

The application of SVM to 3D molecular shape overlay significantly improved the screening efficiency (EF1), by 23−63%, when more than 10 known inhibitors were available as queries. The SVM models using ScaledColor for the 3D similarity profiles displayed the best results in both the SVM models and conventional similarity searches. The additional advantage of machine learning was the use of multiple 3D shape similarity metrics for explanatory variables. In the standard ROCS procedure, ShapeTanimoto and ScaledColor were considered by simply summing both metrics (ComboScore). In this study, ComboScore did not show better screening efficiencies than ScaledColor, suggesting that the simple summation of shape-based metrics and pharmacophore-based metrics could not effectively utilize both information sources. In contrast, the SVM models using both ShapeTanimoto and ScaledColor separately for the 3D similarity profiles showed higher EF1 and ROC score values than those using ShapeTanimoto, ScaledColor, or ComboScore. The results suggested that the machine learning techniques could handle plural information sources properly.

For future developments, the methods utilized for the selection of optimal query compounds and 3D shape similarity metrics should be further refined. In this study, known inhibitors for the test set and the training set were selected by using structural clustering, to provide a structurally nonredundant compound set. In actual drug discovery projects, the known inhibitors would have some redundancy, and using all of the compounds as the queries for the 3D alignment may not be fruitful. To achieve the optimal balance of screening efficiency and calculation time, the selection methods of query compounds for the 3D alignment should be considered. Recently, some studies about optimal query compounds and their 3D conformations were reported.[60−70]

The combination of 3D similarity metrics and other molecular descriptors for SVM could improve the screening efficiency. Generally, the addition of new explanatory variables, which have little correlation to the previously used explanatory variables, could be effective to improve the accuracy of prediction models. Thus, the combination of 3D similarity profiles using ScaledColor, ColorTanimoto, and ShapeTanimoto, presented in this study, and conventional molecular descriptors, such as the molecular weight, the number of hydrogen bond donors/acceptors, or the molecular fingerprints, such as MACCS structural keys and ECFP4, could be expected to show an improvement over the current prediction models.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Excel data and sdf files. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Telephone: +81-45-503-9551. Fax: +81-45-503-9432. E-mail: honma@gsc.riken.jp.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This research was performed as part of the RIKEN Program for Drug Discovery and Medical Technology Platforms. The authors thank Dr. Hirofumi Watanabe for fruitful discussions about the 3D molecular overlay algorithm using ROCS.

■ ABBREVIATIONS

SVM, support vector machine; RBF, radial basis function; ROC, receiver operating characteristics

■ REFERENCES

(1) *MACCS Structural keys*; Accelrys: San Diego, CA, 2011.
(2) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model* **2010**, *50*, 742−754.
(3) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046−1053.
(4) Moffat, K.; Gillet, V. J.; Whittle, M.; Bravi, G.; Leach, A. R. A comparison of field-based similarity searching methods: CatShape, FBSS, and ROCS. *J. Chem. Inf. Model.* **2008**, *48*, 719−729.
(5) Simoes, C. J.; Mukherjee, T.; Brito, R. M.; Jackson, R. M. Toward the discovery of functional transthyretin amyloid inhibitors: application of virtual screening methods. *J. Chem. Inf. Model.* **2010**, *50*, 1806−1820.
(6) Kortagere, S.; Krasowski, M. D.; Ekins, S. The importance of discerning shape in molecular pharmacology. *Trends Pharmacol. Sci.* **2009**, *30*, 138−147.
(7) Nicholls, A.; McGaughey, G. B.; Sheridan, R. P.; Good, A. C.; Warren, G.; Mathieu, M.; Muchmore, S. W.; Brown, S. P.; Grant, J. A.; Haigh, J. A.; Nevins, N.; Jain, A. N.; Kelley, B. Molecular shape and medicinal chemistry: a perspective. *J. Med. Chem.* **2010**, *53*, 3862−3886.
(8) Ebalunode, J. O.; Zheng, W. Molecular shape technologies in drug discovery: methods and applications. *Curr. Top. Med. Chem.* **2010**, *10*, 669−679.
(9) 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.
(10) Goodsell, D. S.; Lauble, H.; Stout, C. D.; Olson, A. J. Automated docking in crystallography: analysis of the substrates of aconitase. *Proteins* **1993**, *17*, 1−10.
(11) 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.
(12) Cavasotto, C. N.; Orry, A. J. Ligand docking and structure-based virtual screening in drug discovery. *Curr. Top. Med. Chem.* **2007**, *7*, 1006−1014.
(13) Giganti, D.; Guillemain, H.; Spadoni, J. L.; Nilges, M.; Zagury, J. F.; Montes, M. Comparative evaluation of 3D virtual ligand screening methods: impact of the molecular alignment on enrichment. *J. Chem. Inf. Model.* **2010**, *50*, 992−1004.
(14) Hawkins, P. C.; Skillman, A. G.; Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* **2007**, *50*, 74−82.
(15) Sheridan, R. P.; McGaughey, G. B.; Cornell, W. D. Multiple protein structures and multiple ligands: effects on the apparent goodness of virtual screening results. *J. Comput.-Aided. Mol. Des.* **2008**, *22*, 257−265.
(16) Lemmen, C.; Lengauer, T.; Klebe, G. FLEXS: a method for fast flexible ligand superposition. *J. Med. Chem.* **1998**, *41*, 4502−4520.
(17) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J. F.; Cornell, W. D. Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.* **2007**, *47*, 1504−1519.
(18) *ROCS*; OpenEye Scientific Software, Inc., Santa Fe, NM, USA, 2008.
(19) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *J. Comput. Chem.* **1996**, *17*, 1653−1666.
(20) Dixon, S. L.; Smondyrev, A. M.; Knoll, E. H.; Rao, S. N.; Shaw, D. E.; Friesner, R. A. PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. *J. Comput.-Aided. Mol. Des.* **2006**, *20*, 647−671.
(21) Tresadern, G.; Bemporad, D.; Howe, T. A comparison of ligand based virtual screening methods and application to corticotropin releasing factor 1 receptor. *J. Mol. Graphics Modell.* **2009**, *27*, 860−870.
(22) Naylor, E.; Arredouani, A.; Vasudevan, S. R.; Lewis, A. M.; Parkesh, R.; Mizote, A.; Rosen, D.; Thomas, J. M.; Izumi, M.; Ganesan, A.; Galione, A.; Churchill, G. C. Identification of a chemical probe for NAADP by virtual screening. *Nat. Chem. Biol.* **2009**, *5*, 220−226.
(23) Freitas, R. F.; Oprea, T. I.; Montanari, C. A. 2D QSAR and similarity studies on cruzain inhibitors aimed at improving selectivity over cathepsin L. *Bioorg. Med. Chem.* **2008**, *16*, 838−853.
(24) Perez-Nueno, V. I.; Ritchie, D. W.; Rabal, O.; Pascual, R.; Borrell, J. I.; Teixido, J. Comparison of ligand-based and receptor-based virtual screening of HIV entry inhibitors for the CXCR4 and CCR5 receptors using 3D ligand shape matching and ligand-receptor docking. *J. Chem. Inf. Model.* **2008**, *48*, 509−533.
(25) Bostrom, J.; Berggren, K.; Elebring, T.; Greasley, P. J.; Wilstermann, M. Scaffold hopping, synthesis and structure-activity relationships of 5,6-diaryl-pyrazine-2-amide derivatives: a novel series of CB1 receptor antagonists. *Bioorg. Med. Chem.* **2007**, *15*, 4077−4084.
(26) Jacobsson, M.; Garedal, M.; Schultz, J.; Karlen, A. Identification of Plasmodium falciparum spermidine synthase active site binders through structure-based virtual screening. *J. Med. Chem.* **2008**, *51*, 2777−2786.
(27) Kumar, R. J.; Chebib, M.; Hibbs, D. E.; Kim, H. L.; Johnston, G. A.; Salam, N. K.; Hanrahan, J. R. Novel gamma-aminobutyric acid rho1 receptor antagonists; synthesis, pharmacological activity and structure-activity relationships. *J. Med. Chem.* **2008**, *51*, 3825−3840.
(28) Hall, M. D.; Salam, N. K.; Hellawell, J. L.; Fales, H. M.; Kensler, C. B.; Ludwig, J. A.; Szakacs, G.; Hibbs, D. E.; Gottesman, M. M. Synthesis, activity, and pharmacophore development for isatin-beta-thiosemicarbazones with selective activity toward multidrug-resistant cells. *J. Med. Chem.* **2009**, *52*, 3191−3204.
(29) Broer, B. M.; Gurrath, M.; Holtje, H. D. Molecular modelling studies on the ORL1-receptor and ORL1-agonists. *J. Comput.-Aided. Mol. Des.* **2003**, *17*, 739−754.
(30) Abate, C.; Mosier, P. D.; Berardi, F.; Glennon, R. A. A structure-affinity and comparative molecular field analysis of sigma-2 (sigma2) receptor ligands. *Cent. Nerv. Syst. Agents Med. Chem.* **2009**, *9*, 246−257.
(31) Schneider, G.; Wrede, P. Artificial neural networks for computer-based molecular design. *Prog. Biophys. Mol. Biol.* **1998**, *70*, 175−222.
(32) Agrafiotis, D. K.; Cedeno, W.; Lobanov, V. S. On the use of neural network ensembles in QSAR and QSPR. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 903−911.
(33) Kauffman, G. W.; Jurs, P. C. QSAR and k-nearest neighbor classification analysis of selective cyclooxygenase-2 inhibitors using topologically-based numerical descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1553−1560.
(34) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. 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.
(35) Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1882−1889.
(36) Winkler, D. A. Neural networks as robust tools in drug lead discovery and development. *Mol. Biotechnol.* **2004**, *27*, 139−168.
(37) Guha, R.; Jurs, P. C. Interpreting computational neural network QSAR models: a measure of descriptor importance. *J. Chem. Inf. Model.* **2005**, *45*, 800−806.
(38) Plewczynski, D.; Spieser, S. A.; Koch, U. Assessing different classification methods for virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 1098−1106.
(39) Chen, B.; Harrison, R. F.; Papadatos, G.; Willett, P.; Wood, D. J.; Lewell, X. Q.; Greenidge, P.; Stiefl, N. Evaluation of machine-learning methods for ligand-based virtual screening. *J. Comput.-Aided. Mol. Des.* **2007**, *21*, 53−62.
(40) 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.
(41) Kawai, K.; Fujishima, S.; Takahashi, Y. Predictive activity profiling of drugs by topological-fragment-spectra-based support vector machines. *J. Chem. Inf. Model.* **2008**, *48*, 1152−1160.
(42) Fukunishi, Y. Structure-based drug screening and ligand-based drug screening with machine learning. *Comb. Chem. High Throughput Screen.* **2009**, *12*, 397−408.
(43) Sato, T.; Honma, T.; Yokoyama, S. Combining machine learning and pharmacophore-based interaction fingerprint for in silico screening. *J. Chem. Inf. Model.* **2010**, *50*, 170−185.
(44) Kinnings, S. L.; Liu, N.; Tonge, P. J.; Jackson, R. M.; Xie, L.; Bourne, P. E. A machine learning-based method to improve docking scoring functions and its application to drug repurposing. *J. Chem. Inf. Model.* **2011**, *51*, 408−419.
(45) Yabuuchi, H.; Niijima, S.; Takematsu, H.; Ida, T.; Hirokawa, T.; Hara, T.; Ogawa, T.; Minowa, Y.; Tsujimoto, G.; Okuno, Y. Analysis of multiple compound-protein interactions reveals novel bioactive molecules. *Mol. Syst. Biol.* **2011**, *7*, 472.
(46) Vapnik, V. *The nature of statistical learning theory*; Springer-Verlag: New York, 1995.
(47) Breiman, L. Random forests. *Machine Learn.* **2001**, *45*, 5−32.
(48) Whitehead, C. E.; Breneman, C. M.; Sukumar, N.; Ryan, M. D. Transferable atom equivalent multicentered multipole expansion method. *J. Comput. Chem.* **2003**, *24*, 512−529.
(49) Breneman, C. M.; Sundling, C. M.; Sukumar, N.; Shen, L.; Katt, W. P.; Embrechts, M. J. New developments in PEST shape/property hybrid descriptors. *J. Comput.-Aided. Mol. Des.* **2003**, *17*, 231−240.
(50) Chekmarev, D. S.; Kholodovych, V.; Balakin, K. V.; Ivanenkov, Y.; Ekins, S.; Welsh, W. J. Shape signatures: new descriptors for predicting cardiotoxicity in silico. *Chem. Res. Toxicol.* **2008**, *21*, 1304−1314.
(51) Chekmarev, D.; Kholodovych, V.; Kortagere, S.; Welsh, W. J.; Ekins, S. Predicting inhibitors of acetylcholinesterase by regression and

classification machine learning approaches with combinations of molecular descriptors. *Pharm. Res.* **2009**, *26*, 2216−2224.
(52) Haigh, J. A.; Pickup, B. T.; Grant, J. A.; Nicholls, A. Small molecule shape-fingerprints. *J. Chem. Inf. Model.* **2005**, *45*, 673−684.
(53) Scholkopf, B.; Mika, S.; Burges, C. C.; Knirsch, P.; Muller, K. R.; Ratsch, G.; Smola, A. J. Input space versus feature space in kernel-based methods. *IEEE Trans. Neural Netw.* **1999**, *10*, 1000−1017.
(54) Tsuda, K. Support vector classifier with asymmetric kernel function. In *Proceedings ESANN*; Verleysen, M., Ed.; Brussels, 1999; pp 183−188.
(55) Scholkopf, B. Smola, A. J. *Learning with Kernels*; MIT Press: Cambridge, MA, 2002.
(56) Warr, W. A. ChEMBL An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). *J. Comput.-Aided. Mol. Des.* **2009**, *23*, 195−198.
(57) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.
(58) Lajiness, M. S. Molecular similarity-based methods for selecting compounds for screening; In *Computational chemical graph theory*; Rouvray, D. H.; Nova Science Publishers, Inc.: Commack, NY, USA, 1990; pp 299−316.
(59) Bawden, D. 1993. Molecular dissimilarity in chemical information systems, In *Chemical Structures 2. The International Language of Chemistry*; Warr, W. A., Ed.; Springer-Verlag, Heidelberg, 1993; pp 383−388.
(60) Willett, P. Dissimilarity-based algorithms for selecting structurally diverse sets of compounds. *J. Comput. Biol.* **1999**, *6*, 447−457.
(61) Irwin, J. J.; Shoichet, B. K. ZINC--a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.
(62) Boström, J.; Greenwood, J. R.; Gottfries, J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J. Mol. Graphics Modell.* **2003**, *21*, 449−462.
(63) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004**, *47*, 1750−1759.
(64) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J. P.; Bertrand, H. O. Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* **2005**, *48*, 2534−2547.
(65) *Pipeline Pilot*; Accelrys Software Inc.: San Diego California, USA, 2007.
(66) Sato, T.; Matsuo, Y.; Honma, T.; Yokoyama, S. In silico functional profiling of small molecules and its applications. *J. Med. Chem.* **2008**, *51*, 7705−7716.
(67) Dayam, R.; Aiello, F.; Deng, J.; Wu, Y.; Garofalo, A.; Chen, X.; Neamati, N. Discovery of small molecule integrin alphavbeta3 antagonists as novel anticancer agents. *J. Med. Chem.* **2006**, *49*, 4526−4534.
(68) Perez-Nueno, V. I.; Ritchie, D. W.; Borrell, J. I.; Teixido, J. Clustering and classifying diverse HIV entry inhibitors using a novel consensus shape-based virtual screening approach: further evidence for multiple binding sites within the CCR5 extracellular pocket. *J. Chem. Inf. Model.* **2008**, *48*, 2146−2165.
(69) Kirchmair, J.; Distinto, S.; Markt, P.; Schuster, D.; Spitzer, G. M.; Liedl, K. R.; Wolber, G. How to optimize shape-based virtual screening: choosing the right query and including chemical information. *J. Chem. Inf. Model.* **2009**, *49*, 678−692.
(70) Perez-Nueno, V. I.; Ritchie, D. W. Using Consensus-Shape Clustering To Identify Promiscuous Ligands and Protein Targets and To Choose the Right Query for Shape-Based Virtual Screening. *J. Chem. Inf. Model.* **2011**, *51*, 1233−1248.