Improving Structure-Based Virtual Screening by Multivariate Analysis of Scoring Data

Micael Jacobsson,*[,†,‡] Per Lidén,[§] Eva Stjernschantz,[†] Henrik Boström,[§,#] and Ulf Norinder[⊥]

Structural Chemistry, Biovitrum AB, SE-112 76 Stockholm, Sweden, Department of Medicinal Chemistry, Uppsala University, BMC, Box 574, SE-751 23, Uppsala, Sweden, Compumine AB, Österögatan 3, SE-164 40, Kista, Sweden, Department of Computer and Systems Sciences, Stockholm University and Royal Institute of Technology, Forum 100, SE-164 40 Kista, Sweden, and AstraZeneca R&D Södertälje, SE-151 85 Södertälje, Sweden

Received May 15, 2003

Three different multivariate statistical methods, PLS discriminant analysis, rule-based methods, and Bayesian classification, have been applied to multidimensional scoring data from four different target proteins: estrogen receptor α (ERα), matrix metalloprotease 3 (MMP3), factor Xa (fXa), and acetylcholine esterase (AChE). The purpose was to build classifiers able to discriminate between active and inactive compounds, given a structure-based virtual screen. Seven different scoring functions were used to generate the scoring matrices. The classifiers were compared to classical consensus scoring and single scoring functions. The classifiers show a superior performance, with rule-based methods being most effective. The precision of correctly predicting an active compound is about 90% for three of the targets and about 25% for acetylcholine esterase. On the basis of these results, a new two-stage approach is suggested for structure-based virtual screening where limited activity information is available.

Introduction

Docking and virtual screening are widely used for structure-based drug design and hit identification, as well as focused library design. Docking, in this context, means to predict the binding mode of a small molecule when binding to a target protein with a known 3D structure.[1] Flexible docking of small molecules to rigid protein structures using fast, approximate algorithms has been developed to a point where it is most often possible to reproduce binding modes of ligands, given a structure of a protein−ligand complex. It is often also possible to predict reasonable binding modes of ligands for which the correct binding mode is unknown,[2−5] and where no ligand−protein complex structure of a similar ligand is available. Sometimes it is even possible to dock ligands to a homology model.[5,6] Even though a lot remains to be done to solve problems such as induced fit and better handling of low quality protein structures, docking has become an invaluable tool in structure-based lead optimization.

There are also examples of successful structure-based virtual screens,[7] in which docking of large numbers of compounds followed by selection based on simple scoring functions has led to the identification of new binders or at least to significant enrichment of binders in the set of suggested compounds. Scoring functions are typically weighted sums of energy terms, in which the weights have been optimized to maximize correlation with binding affinity for a set of known binders, or to maximize the discrimination between a set of known actives and a set of known inactives.[5,8] However, it is clear that fast, approximate scoring functions still leave much room for improvement. More correct prediction of binding free energies still require more involved computational methods based on simulations,[9] and scoring functions used in virtual screening today should rather be regarded as filters for distinguishing binders from nonbinders than accurate predictors of binding free energies.[10]

Using multiple, rather than single, scoring functions has been shown to improve the discrimination between binders and nonbinders.[10−12] The individual scoring functions have been combined in different ways, e.g., strict intersection of lists of high-scoring compounds,[11] rank-sum, worst−best rank, or jury-based methods.[12] In this study, we have applied different multivariate statistical methods to multidimensional scoring data to investigate whether an improvement in discriminatory power can be achieved. For this purpose, we have docked 389 ligands with known activity and 999 random, diverse drug-like molecules from the MDL Drug Data Report (MDDR, http://www.mdl.com) database to four different target proteins. The sources of the actives are listed in Table 1.

The four target proteins, estrogen receptor α (ERα),[13] matrix metalloprotease 3 (MMP3),[14] acetylcholine esterase (AChE),[15] and factor Xa (fXa),[16] represent quite different active sites. All known ligands are derived from published work and are somewhat diverse, especially in the case of ERα (processed structures of ligands and proteins are available at http://www.compumine.com/research/scoring.html). The ligands have been prepared in an automated fashion, to simulate a typical virtual screen. The dockings have been performed using a single structure of each protein, without the inclusion of water molecules or modification of potentially flexible residues,

* To whom correspondence should be addressed. Phone: +46 8 6972551. Fax: +46 8 6972320. E-mail: micael.jacobsson@biovitrum.com.

[†] Biovitrum AB.

[‡] Uppsala University.

[§] Compumine AB.

[#] Stockholm University and Royal Institute of Technology.

[⊥] AstraZeneca R&D Södertälje.

10.1021/jm030896t CCC: $25.00
Published on Web 11/20/2003

Table 1. Ligand Sets and Crystal Structures Used in This Study

target	PDB file	ligand source	no. binders
ERα	1ere[13]	Sippl[40] (mimics)	36
ERα	1ere[13]	Shi et al.[41] (toxins)	110
AChE	1eve[15]	Contreras et al.[42]	54
MMP-3	1hy7[14]	Ha et al.[3]	60
fXa	1g2l[16]	Matter et al.[43]	129

for the same reason. One way to handle flexibility in the active site of the target protein is to dock to more than one static representation of the same protein. For large sets of ligands, this greatly increases the number of required dockings and the amount of data that needs to be processed. We want to see how well straightforward structure-based virtual screening, without any attempt to handle receptor flexibility, can fare. Three different docking programs, GOLD,[17] Glide version 2.0,[18,19] and ICM version 2.8[20,21] were used, but because the results obtained were comparable for all three programs, only results based on the ICM dockings are presented here.

Two scoring functions implemented in ICM and five scoring functions implemented in CScore from Tripos[12] were used. The resulting seven-dimensional scoring vectors were analyzed using both classical consensus scoring as implemented in CScore and three different statistical methods: PLS-DA,[22] Bayesian classification,[23] and rule-based methods.[24] The three different statistical methods all need a training set of known actives and inactives. For this purpose the known binders and potential nonbinders were partitioned into one training set and one validation set for each set of ligands and target protein, with a similar distribution of activities in the case of known actives and molecular weights in the case of potential inactives (for further description, see Methods). Discriminators were constructed and evaluated, using both internal cross-validation within the training set, and the external validation set. The three different statistical methods, single scoring function classifiers, and classical consensus scoring are compared below, the results are discussed, and a general, stepwise protocol for virtual screening utilizing information from small sets of known binders is proposed.

Results

The classification models were built using the training sets described in Methods below and subsequently evaluated using the external validation sets. ERα, being the only target for which we had two rather different sets of ligands, was treated in a more elaborate fashion.

Five different performance measures were used when evaluating the different classifiers. The basic measures are accuracy, precision, and recall. *Accuracy* is the overall classification accuracy of a prediction model, including both active and inactive compounds. It is defined by

$$\text{accuracy} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{fp} + \text{tn} + \text{fn}} \qquad (1)$$

where tp is the number of true positives, tn is the number of true negatives, fp is the number of false positives and fn is the number of false negatives. *Precision* is a measure of the accuracy of predicting a specific class. In this work, the precision of the active class is of particular interest and is thus the only class for which precision is reported. It is defined by

$$\text{precision} = \frac{\text{tp}}{\text{tp} + \text{fp}} \qquad (2)$$

Recall is a measure of the ability of a prediction model to select instances of a certain class from a data set. It is defined by the formula:

$$\text{recall} = \frac{\text{tp}}{\text{tp} + \text{fn}} \qquad (3)$$

None of these three numbers are absolute measures of classification performance by themselves, but should be seen together. It is important to have a high accuracy, but because the number of inactives is much larger than the number of actives, it is possible to get a good accuracy while predicting no actives. In the same way, it is possible to get a good precision by predicting only very few examples of the considered class, but being correct in those predictions. A high precision for the active class may be a good behavior of a classifier if the starting number of compounds to classify is very large. In large-scale structure-based virtual screening it might be acceptable to miss many binders, if only the compounds predicted to bind actually do so. However, in the tests created here, with relatively few compounds and a rather large ratio of actives to inactives, a classifier can only be seen as really successful if high precision is accompanied by high recall. The expectation values of the precision and recall for the active class of a random classifier is given by the ratio of actives in the entire set and the ratio between the number of predicted actives and the total number of compounds, respectively.

For each set of ligands and target protein, the *enrichment factor* (EF) was also calculated. This is the relative enrichment of active compounds in the set of instances predicted to be active in relation to the fraction of active compounds in the original data set:

$$\text{EF} = \frac{\text{precision}}{(\text{tp} + \text{fn})/(\text{tp} + \text{fp} + \text{tn} + \text{fn})} \qquad (4)$$

The enrichment factor of a random classifier is 1. Finally, the “E-value” of each classification result was calculated. This is defined as the probability to achieve the exact same classification or better by chance, and is calculated using the hypergeometric distribution. A better result is defined as classifying the same number of actives, but with a larger number of true actives. Since the probabilities are very low, the E-value is given as pE, i.e., the negative logarithm of the E-value. The results are shown in Table 2, and for the five ERα models also in Table 3. The different methods are summarized and compared in Figure 1.

Discussion

Enrichment is a good basis of comparison between different methods for the same ligand set and target protein, even though the actual values will depend on the number of actives included in the validation set. The basic result is that the rule-based RDS models have the best overall performance in all sets of ligands and target

Table 2. Performance of the Discriminators for the Different Ligand Sets and Target Proteins, Using the External Validation Set

data set	no. examples (no. actives)	method	accuracy	precision	recall	enrichment	pE
ERa-mimics[40]	345 (12)	C[a]	0.994	1.000	0.833	28.8	16.9
		Bayes, pact = 0.0348[b]	0.980	0.692	0.750	19.9	12.1
		PLS-DA	0.986	0.706	1.000	20.3	17.9
		PLS-DA, limit 0.14[c]	0.991	0.909	0.833	26.1	15.9
		ICM score, $T = -31.74$[d]	0.985	0.769	0.833	20.9	14.5
		consensus > 5[e]	0.568	0.067	0.833	1.8	2.3
ERa-toxins[41]	369 (36)	RP-B-10[f]	0.976	0.935	0.806	9.6	33.4
		Bayes, pact = 0.100[b]	0.905	0.508	0.861	5.2	22.4
		PLS-DA	0.881	0.446	0.917	4.6	22.4
		PLS-DA, limit 0.3[c]	0.959	0.784	0.806	8.0	28.6
		ICM score, $T = -33.64$[d]	0.940	0.927	0.464	8.9	17.8
		consensus > 5[e]	0.549	0.109	0.464	1.0	0.5
fXa[16]	376 (43)	RP−B-50s[g]	0.928	0.833	0.465	7.3	17.0
		Bayes, pact = 0.114[b]	0.944	0.750	0.767	6.56	28.4
		PLS-DA	0.910	0.567	0.884	5.0	27.6
		PLS-DA, limit 0.2[c]	0.918	0.630	0.674	5.5	20.4
		ICM score, $T = -22.0$[d]	0.615	0.252	0.853	1.8	10.7
		consensus > 4[e]	0.657	0.276	0.853	1.9	12.2
MMP3[14]	353 (20)	RP-B-10[f]	0.992	0.947	0.900	16.7	26.3
		Bayes, pact = 0.0567[b]	0.972	0.708	0.85	12.5	20.0
		PLS-DA	0.941	0.488	1.000	8.6	20.9
		PLS-DA, limit 0.3[c]	0.986	0.800	1.000	14.1	27.6
		ICM score, $T = -21.22$[d]	0.615	0.252	0.853	1.8	10.2
		consensus > 5[e]	0.657	0.276	0.853	1.9	10.8
AChE[15]	351 (18)	C-W20[h]	0.869	0.220	0.611	4.3	5.6
		Bayes, pact = 0.0513[b]	0.903	0.250	0.444	4.9	4.3
		PLS-DA	0.681	0.108	0.722	2.1	3.1
		PLS-DA, limit 0.2[c]	0.766	0.110	0.500	2.1	2.0
		ICM score, $T = -26.64$[d]	0.484	0.084	0.907	1.6	2.8
		consensus > 5[e]	0.496	0.086	0.907	1.7	2.9

[a] Covering. [b] Bayes, pact = p − Bayesian classification with a priori probability p for being active. [c] PLS-DA, limit R − partial least-squares discriminant analysis with y-value threshold R for actives. [d] ICM score, $T = t$ − single scoring function ICM to discern actives from inactives, threshold t. [e] Consensus > c − consensus scoring according to default settings in Tripos CScore, with threshold c. [f] Recursive-partitioning in combination with Bagging for 10 iterations. [g] Recursive-partitioning in combination with Bagging for 50 iterations, using random sampling of examples. [h] Covering with the active class up-weighted 20 times.

Table 3. Performance of Classifiers Constructed and Tested Using Different Combinations of Training and Validation Sets from the Two Available ERα Sets

data set combination	no. examples (no. actives)	method	accuracy	precision	recall	enrichment	pE
actives[a]	381 (48)	RP-B-50	0.979	0.955	0.875	7.6	46.2
		Bayes, pact = 0.126	0.895	0.551	0.896	4.4	28.9
		PLS-DA	0.869	0.489	0.938	3.9	28.2
		PLS-DA, limit 0.3	0.961	0.811	0.896	6.4	40.7
mimics[b]	345 (12)	C	0.994	1.000	0.833	28.8	23.0
		Bayes, pact = 0.0348	0.980	0.692	0.750	19.9	16.5
		PLS-DA	0.986	0.706	1.000	20.3	24.1
		PLS-DA, limit 0.14	0.991	0.909	0.833	26.1	20.9
toxins[c]	369 (36)	RP-B-10	0.976	0.935	0.806	9.6	33.4
		Bayes, pact = 0.100	0.905	0.508	0.861	5.2	22.4
		PLS-DA	0.881	0.446	0.917	4.6	22.4
		PLS-DA, limit 0.3	0.959	0.784	0.806	8.0	28.6
mimics to toxins[d]	443 (110)	RP-B-50	0.887	0.969	0.564	3.9	41.8
		Bayes, pact = 0.248	0.935	0.945	0.782	3.8	61.8
		PLS-DA	0.932	0.851	0.882	3.4	63.8
		PLS-DA, limit 0.2	0.932	0.955	0.764	3.8	60.8
toxins to mimics[e]	369 (36)	C-B-10s	0.986	0.919	0.944	9.4	41.4
		Bayes, pact = 0.0976	0.911	0.522	1	5.4	30.4
		PLS-DA	0.894	0.480	1.000	4.9	28.6
		PLS-DA, limit 0.2	0.965	0.745	0.972	7.6	36.8

[a] Training set is from both Sippl[40] and Shi et al.,[41] validation set is from both Sippl[40] and Shi et al.[41] [b] Both training set and validation sets are from Sippl.[40] [c] Both training set and validation set are from Shi et al.[41] [d] Training set is from Sippl,[40] validation set from Shi et al.[41] [e] Training set is from Shi et al.,[41] validation set from Sippl.[40]

proteins (see Figure 1 and Table 2). PLS-DA where the threshold for being part of the active class has been set using the training set often performs slightly better than Bayesian classification. The a priori probability of being active, a parameter used in the Bayesian classifier, can be changed to alter the behavior of that particular type of classifier. By decreasing this probability, the precision increases and the recall decreases, while the number of predicted actives decreases, and the enrichment factor increases (data not shown). This can be utilized to get a useful number of predicted actives from a structure-based virtual screen, i.e., if the number of starting compounds is large, the a priori probability of being active can be decreased to get a manageable number of predicted actives, without losing precision but probably missing a few true binders.

Of the seven single scoring functions used here, the energy-based ICM scoring function shows best discrimi-

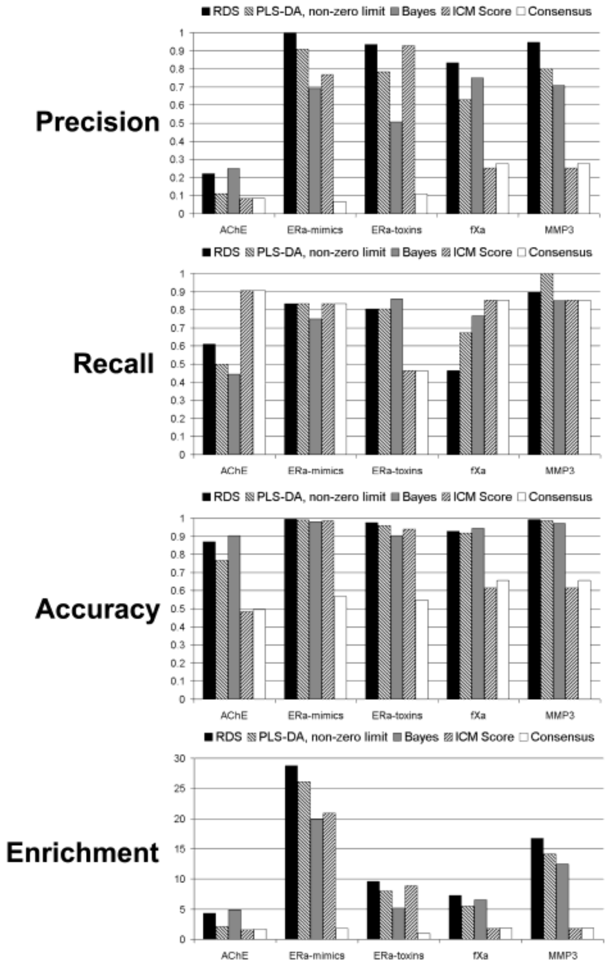

Figure 1. Accuracy, precision, recall, and enrichment of all methods for the different target proteins and ligand sets.

natory power. As can be seen for ERα, it is very good at scoring hydrophobic interactions. Using the ICM scoring function only, with a threshold based on the results for the training set, is the second or third best overall method for discriminating actives when docking to the mainly hydrophobic binding site of ERα. However, no single scoring function is good for all target proteins. For example, the Glide scoring function showed similarly good results for MMP-3 (data not shown), due to the explicit metal−ion interaction term in that scoring function, but performed much worse on ERα. Looking at the result for classical consensus scoring, it can be observed that it performs slightly better than any single scoring function if no scoring function stands out as superior, i.e., consensus scoring as implemented here can improve the results if the scoring functions are comparable in performance, but if one scoring function is much more correct than the others this will be lost in the consensus approach, as can be seen in the ERα results. On the contrary, the multivariate methods do capture such information.

AChE stands out as a target being particularly hard to dock to. Looking closer at the individual docking modes of the known actives, it is apparent that many of the suggested binding modes in fact are wrong. The AChE binding cavity is large with many water molecules and more than one clear binding region in the pocket has been identified.[2] The ligand−protein interactions observed in the crystal structure of the AChE complex used for docking mainly consist of van der Waals and hydrophobic interactions, with only one positively charged ligand atom involved in electrostatic interactions. No direct hydrogen bonds between the ligand and the protein have been observed, only water-bridged hydrogen bonds. The known actives used here are rather symmetrical, with aromatic rings involved in $\pi-\pi$ interactions with the protein in both ends of the molecule. $\pi-\pi$ stacking is not modeled by the force field employed in ICM. It has also been observed that hydrogen bonds are particularly important for obtaining correct docking modes for this particular docking program. The symmetry of the molecules, the lack of modeling of $\pi-\pi$ interactions, and falsely predicted hydrogen bonds result in a large number of improbable docking poses. To obtain more consistently correct docking results for this particular target one would probably need to include specific water molecules and visually inspect more than one suggested docking pose per ligand. However, even for AChE, the best classifier finds about 60% of the actives, with an average accuracy of about 90% for a single classification and with a precision of about 25% (Table 2).

The precision is the ratio of compounds classified as active actually being active. Hence, a precision of about 90%, while retaining well above half of all actives, as was obtained for three of the datasets, is very useful for focusing a given library toward a given target, and a precision of 25%, with a recall of about 50%, is still not useless. The precision of a random classifier is given by the ratio between the number of actives and the total number of examples (given in Tables 2 and 3 for the tests constructed here), so for AChE it is 5% and for fXa, the target with the highest number of actives included, it is 11%. The pE-values, the negative logarithms of the probabilities of achieving the exact same results or better by random selection of compounds, are an alternative way of measuring the significance of the results, when compared to a random classifier. The pE-values of the different classifiers are given in Table 2, and these values also indicate the usefulness of multivariate analysis of scoring data, as compared to using a single scoring function or classical consensus scoring.

Taken together, these results show that multivariate statistical methods greatly increase the discriminative power of consensus scoring, perhaps mostly so for the rule-based methods implemented in RDS, but both PLS-DA and Bayesian classification can be used to build useful classifiers able to discriminate between actives and nonactives, given a virtual screen.

However, we have failed to build good regression models from our data, i.e., we cannot find a straightforward correlation between docking scores and true activity enabling us to reliably predict the quantitative activity given a set of docking scores, using PLS or other methods. Therefore, we think the best approach when using structure-based virtual screening for finding new binders is to use the docking scores for removing as many nonbinders as possible without losing the true binders. This means that a classifier with high precision and high recall for the active class is needed. This will decrease the number of compounds necessary to take into consideration, enabling more thorough analysis of the remaining compounds, by looking closer on the dockings and/or applying other, more computationally expensive methods or even go ahead and test every remaining compound experimentally. We show here that different multivariate statistical methods are well suited for this type of filtering, specifically much better than using classical consensus scoring or single scoring functions.

All methods described here, apart from the consensus scoring approach implemented in CScore, require a training set. Our results imply that accurate quantitative activity information for that set is not necessary. Also, the results for the different ERα sets, especially the mimics to toxins results, where the 36 known actives in the mimics set of ligands is used to train and the resulting classifier is evaluated using the 110 known actives in the toxins set of ligands as an external validation set, imply that the training set can be structurally diverse compared to the evaluation set and still be used to construct models that can identify binders. This suggests that the docking scores and their correlation patterns mostly describe the binding site and how well the ligands fit this, and not the ligands themselves. Hence, results from a limited experimental screen, identifying a few binders, probably with confirmation assays run for the hits, could be used as a training set. Virtual screening followed by scoring using multiple functions and proper statistical analysis as described here could then serve as a method to identify more hits with alternative scaffolds, in a way similar to similarity searching[25,26] or pharmacophore searching,[27] but requiring less ligand activity information which also can be less precise, and probably enabling the identification of scaffolds more different from the known binders. Our proposed combination of structure-based virtual screening and supervised multivariate classifiers is illustrated in Figure 2.

Using RDS, PLS-DA, or Bayesian classification to construct classifiers from a training set of known binders and nonbinders gives a much better classification than using a previously presented consensus scoring method. We think this is both because a training set is used, making the classifier applicable to the target being studied, and the ability of our methods to take quantitative measures of correlations between different scoring function into account. The approach to include inactives in the training set and create classifiers instead of regression models based only on known actives is probably also important. A regression model can have a good Q^2 when predicting the activity of active compounds, but fail totally when confronted with a nonbinder, where the docking scores have been calculated from a docking mode of a compound, which does not actually bind.

Terp et al.[28] also use two multivariate methods, PCA and PLS, to analyze scoring data. They dock a set of known binders to three different matrix metallo proteases, score 10 poses per ligand using 8 different scoring functions and perform a PCA on the resulting set of scores. The first principal component is then used to re-rank poses when performing docking. Since the activities are known, a PLS model is also constructed, and is used to rescore docked ligands. They do not use the resulting scoring function to do structure-based virtual screening, but evaluate it using known binders and try to quantitatively predict the binding affinity of these compounds. There are no results indicating how well the scoring function fares when confronted with scoring data from compounds which actually are nonbinders. Hence, the most important differences between the work presented here and the work done by Terp et al. is the inclusion of inactives and construction of classifiers, for use in structure-based hit identification, as well as the difference in the employed multivariate methods.

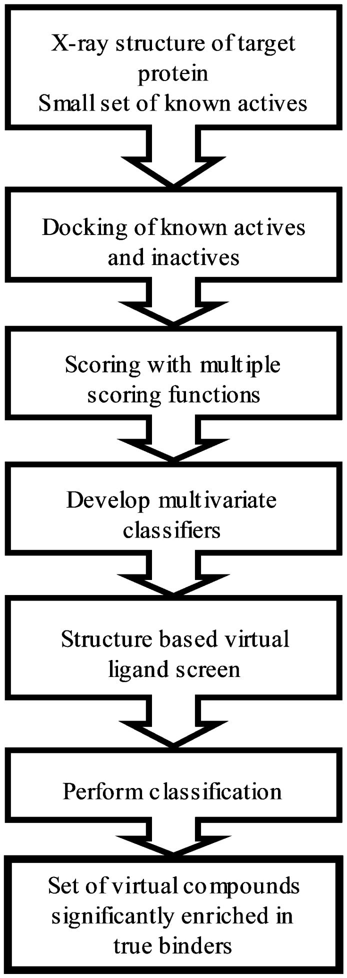
X-ray structure of target protein
Small set of known actives
Docking of known actives and inactives
Scoring with multiple scoring functions
Develop multivariate classifiers
Structure based virtual ligand screen
Perform classification
Set of virtual compounds significantly enriched in true binders

Figure 2. Schema of our proposed use of virtual screening for identifying binders from large sets of virtual compounds.

In our virtual screening setup, there are many sources of errors in the scoring vectors, apart from the inherent inaccuracy of the scoring functions themselves. The docking mode used as input for the scoring function must be correct, both in terms of ligand and protein conformation and in terms of ligand placement, for the scoring function to be able to produce a correct result. We have not assured, using visual inspection, that every docking mode used for scoring is plausible, since this would not have been possible if the number of compounds had been in the 100 000 range. Since the results acquired for AChE differed significantly from those acquired for the other targets, we performed a closer visual inspection of these docking modes. As was stated above, this showed a large number of improbable docking poses, which of course will have a negative effect on the resulting classifiers. We also did not refine our binding sites to any larger extent, meaning that we

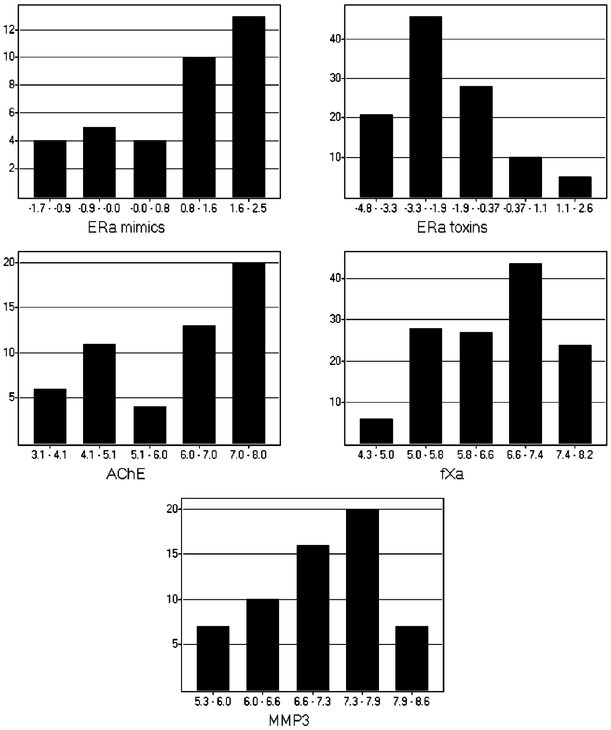

Figure 3. Activity histograms of each of the five sets of known binders. For the two ERα sets activities are given as log(RBA), i.e., logarithm of 100 × [17β-estradiol]/[tested compound] at 50% decrease of receptor bound 17β-estradiol; for AChE and MMP3 the activities are given as pIC50 and for fXa the activities are given as pK_i.

have not tried to include explicit waters or refined the conformations of seemingly flexible side-chains to ensure consistently correct docking poses of known binders. It would probably be beneficial for the docking of known actives with similar structures if the binding site would be relaxed around such a bound ligand with a known binding mode (i.e., from a complex structure). However, this would also bias the binding site toward ligands with that specific scaffold and possibly make the resulting classifier, comprised of a docking setup and a multivariate classifier, less general. The main focus of our study is the statistical properties of scoring data and identification of appropriate methods to capture the information present in such data. More correct protein conformations and more elaborate methods for ensuring the correctness of the individual binding modes can only improve the results.

The tests constructed for this study, i.e., the combinations of target structures and sets of ligands, are not actual HTS results, but a combination of published, known binders and random, drug-like molecules. The known binders are to large extent analogues of each other, or are at least made up of a small number of compound series, with the exception of the ERα toxins set. However, the activity spans are quite large (see Figure 3). The internal diversity of the sets of known actives and the cross-diversity between actives and inactives, and between the two different ERα sets, are given in Table 4. Comparing the classifiers performances (Tables 2 and 3 and Figure 1) to the diversity of the sets, it looks like it is much more important to dock the compounds correctly than to use similar compounds in the test and training sets, since the AChE set of ligands are quite similar, but the difficulties in docking and scoring them correctly makes the resulting classifiers fare much worse than for example the ERα toxins set or the mimics to toxins set. This implies that the methods evaluated here are applicable to structure-based virtual screening of large sets of diverse compounds using screening data from a much smaller number of diverse compounds. We have also used RDS and the methodology presented here in actual in-house projects, creating classifiers using screening data from diverse compounds and successfully sorting out actives from inactives after docking and scoring of diverse, virtual compounds.

To be able to construct good regression models, or to improve the performance of the classifiers for hard cases such as AChE, one might combine the scoring matrices with molecular descriptors. By adding ligand-specific descriptors to the score matrices before performing the multivariate analysis, it might be possible to construct QSAR models with higher significance than when using only docking scores. However, the resulting models will inevitably be more local, i.e., more dependent on the structures in the training set. If a rather homogeneous set of known actives is used, and molecular descriptors are calculated, the resulting model will not be able to predict structurally different ligands. This might be a rewarding extension of our methodology in some cases, but probably not as useful for filtering virtual screening results from large sets of diverse virtual molecules.

Conclusions

In conclusion, we have generated scoring matrices for known actives and potential inactives for four different target proteins, using docking followed by scoring with seven different scoring functions. We used these matrices to construct multivariate classifiers, evaluated these with external test sets, and compared them to classical consensus scoring and single scoring functions. We found that proper multivariate analysis of scoring data is very rewarding in terms of recall of known actives and enrichment of true actives in the set of predicted actives. Rule-based methods implemented in RDS show the best performance, but also PLS-DA and simple Bayesian classification perform very well.

On the basis of this we propose a new methodology for the use of virtual screening to identify novel binders, requiring prior knowledge of a small set of actives (Figure 2). We think it is clear that the imprecise nature of docking and scoring makes blind virtual screening of large number of compounds without any information about true actives or known experimental complex structures a risky exercise. Limited experimental information and proper multivariate statistical treatment of the scoring data dramatically increase the value of these kinds of computations.

Methods

Docking and Scoring. To evaluate the multivariate analysis methods on scoring data, sets of binders with known affinity for the four target proteins were collected from the literature. The number of ligands and their sources are listed in Table 1. In addition to the known binders, we used 999 diverse drug-like ligands extracted from MDDR, selected using 2D fingerprint-based clustering in ChemEnlighten.[29] 3D representations of the ligands were generated using CORINA

Table 4. Diversity of the Ligand Sets Used in the Study, as Measured by dbcmpr from Tripos,[a] Which Uses UNITY 2D Fingerprints to Calculate Tanimoto Similarities between Each Compound in the Reference and Test Sets

ref set	test set	Tanimoto similarity < 0.85[b]	Tanimoto similarity mean	Tanimoto similarity standard deviation
AChE	AChE	26%	0.88	0.13
fXa	fXa	7.1%[c]	0.95	0.05
MMP3	MMP3	16%[d]	0.93	0.07
ERα mimics	ERα mimics	42%[e]	0.83	0.18
ERα toxins	ERα toxins	34%	0.86	0.15
ERα mimics	ERα toxins	83%[f]	0.63	0.20
ERα toxins	ERα mimics	50%[g]	0.84	0.15
Inactives	Inactives	100%, nearest 0.82	0.58	0.09
Inactives	AChE	93%	0.68	0.12
Inactives	fXa	100%, nearest 0.70	0.56	0.03
Inactives	MMP3	96%	0.64	0.09
Inactives	ERα mimics	100%, nearest 0.68	0.56	0.06
Inactives	ERα toxins	100%, nearest 0.77	0.50	0.11

[a] htpp://www.tripos.com. [b] The percentage of compounds in the test set for which the Tanimoto similarity to the nearest neighbor in the reference set is less than or equal to 0.85. [c] Of the 129 fXa compounds, two have the exact same 2D fingerprints as other compounds, i.e., the percentage refers to a total of 127 compounds. [d] 5 of 60 compounds have 2D fingerprints identical to other compounds in the set. [e] 10 of 36 compounds have 2D fingerprints identical to other compounds in the set. [f] Of the 17% which have a Tanimoto similarity > 0.85 to its nearest neighbor in the ERα mimics set, 9.1% (10 compounds) have a Tanimoto similarity of 1, i.e., the 2D fingerprints are identical. [g] Ten compounds in the ERα mimics set have 2D fingerprints identical to its nearest neighbor's in the ERα toxins set.

version 2.4 (Molecular Networks GmbH, http://www.molecular-networks.de),[30] and it was assumed that the stereochemistry implied in the drawings in the original publications was the correct for binding. That is, only the drawn stereoisomer was used, not all possible. Ionization states for the ligands were set using a SYBYL Programming Language (SPL) script, in which SLNs[31] are used to find substructures (e.g., carboxylic acids and various amines) that are substituted for their ionized counterparts. Only one ionization state per ligand, the fully ionized one, was used. For the MDDR ligands, a single, randomly selected stereoisomer was used for chiral compounds. This automated pretreatment of ligands is typical for a large-scale virtual screening study.

The diversity of the different ligand sets was analyzed using dbcmpr from Tripos (http://www.tripos.com). dbcmpr is a utility which calculates the Tanimoto similarities, using UNITY 2D fingerprints, between all compounds in a set and their nearest neighbors in another set. The results are shown in Table 4. The percentage of compounds in the test set having a Tanimoto similarity less than or equal to 0.85 to its nearest neighbor in the reference set is given, as well as the means and standard deviations of the Tanimoto similarities between all compounds in the test set and their nearest neighbors. The comparison is done by constructing UNITY databases of each set. When the databases are constructed only one entry is created for each unique 2D fingerprint. In the fXa, MMP3, and ERα mimics ligand sets, there are 2, 5, and 10 compounds with identical 2D fingerprints, respectively. This should be taken into account when the results in Table 4 are interpreted.

The target protein structures (PDB files are listed in Table 1) were preprocessed using ICM. First, hydrogens were added with random orientation, and then all polar hydrogens were oriented, one at a time, by performing a systematic search of the relevant torsion angle, while keeping the rest of the structure fixed. Arginines and lysines were set to be positively charged, and aspartates and glutamates were set to be negatively charged. All other side-chains were treated as neutral. Finally, all histidines and side-chain amides (glutamines and asparagines) were processed. For each histidine, the two possible side-chain tautomers were tested and the one with the lowest energy was kept. The side-chain terminal amides were tested both in their original orientation and rotated by 180° and the orientation with the lowest energy was kept. Because we wanted to dock diverse ligands with large variations in size and possible interactions, all waters in the active sites were removed, as to not bias the docking to one particular binding mode. The idea is that a smaller ligand, binding to the binding site together with a number of water molecules, will still be able to dock correctly if no waters are included, but a larger ligand occupying parts of space occupied by water molecules when the smaller ligand binds, cannot dock correctly if these waters are kept. All the processed structures (SD and PDB files) are available for downloading at http://www.compumine.com/research/scoring.html, with the ligands in their respective docking poses.

Dockings were performed using ICM version 2.8.[20] It is implemented as a Monte Carlo minimization of the total energy of the ligand (both internal energy and interaction energy), where a set of protein-derived grids is used to model the interaction energy of the protein and ligand. The Monte Carlo procedure produces a stack of possible docking modes for each ligand, sorted by energy. In this study we kept only the highest ranking mode, again to simulate a virtual screening setup with its demand for throughput and automation. During lead optimization, when trying to dock a manageable number of known binders, it can be very rewarding to analyze a number of high-ranking docking poses, and choose the most probable, to get a good binding hypothesis for proposing chemical modifications to increase the affinity of the known binder. In structure-based virtual screening, the goal is to propose a small set of potential actives given a large number of compounds, and it is not possible to manually review the docking stacks of each compound, but one has to let the scoring function rank the different poses suggested by the docking program. For the same reason, the docking modes were not manually analyzed to remove obviously faulty dockings, but all ligands were scored. This will further increase the noise in the resulting scoring data and make the task of building a classifier harder. Two other docking programs, GOLD[17] and Glide,[19] were also evaluated, but since the statistical analysis gave similar results for all three programs and for clarity and brevity, only the results for ICM are described further.

Two scoring functions are implemented in ICM: one energy-based, calculated as a weighted sum of energy terms,[20] and one calculated using MolSofts implementation of potential mean forces (PMF), derived from a set of known protein−ligand complexes, similar to Muegge et al.[32] In addition to these two scoring functions, the five scoring functions implemented in Tripos CScore were used,[12] resulting in a total of seven score values per docked ligand. The CScore scoring functions are Tripos implementations of FlexX-score,[33] DOCK score,[34] PMF score,[32] GOLD score,[35] and ChemScore.[18]

The ligands were partitioned into four different groups for each set of ligands and target proteins. Approximately two-thirds of the actives and inactives were used as training sets and one-third were used as external test sets. The partitioning into training and validation sets was done by sorting the ligands according to activity (known actives) and molecular weight (inactives), treating each category separately, and setting apart every third compound as a validation compound. This scheme results in four matrices of scores for each combination of ligand set and target protein, with the dimensions 7× (number of ligands in set), namely, scores of actives in training set, scores of actives in validation set, scores of inactives in training set, and scores of inactives in validation set. The quantitative activity values of the known actives were not used explicitly in this study even though they span a rather wide activity range (the activity histograms are given in Figure 3). The same set of 999 MDDR compounds were used as inactives in all five sets of ligands and target proteins.

For one target, ERα, we had two different sets of ligands with known activity, which were used to build discriminators to test the applicability of our methodology to more diverse sets of ligands. In addition to building discriminators using the two different training sets and evaluating them with the corresponding external validation sets (denoted mimics and toxins in Table 3), we built one model with both training sets, evaluating it with both validation sets (denoted Actives in Table 3), and two “cross-models” (denoted mimics to toxins and toxins to mimics in Table 3), where the models were built using all active compounds from one set as training set and all active compounds from the other set as validation set. As can be seen in Table 4, the toxins set cover the mimics set quite well, and there are actually 10 compounds in the ERα toxins set having 2D fingerprints identical to a compound in the ERα mimics set. Hence, the toxins to mimics results in Table 3 are not too surprising, even though they show that the training set can be diverse in itself. However, the mimics set cover the toxins set to a much lower degree, and the toxins set is rather diverse in itself, but still the precision and recall values for mimics to toxins are comparable to the other results.

The number of actives and the total number of compounds in each validation set are shown in Tables 2 and 3.

Rule-Based Methods. The data mining system rule discovery system (RDS)[24] was used to create rule-based prediction models based on the scoring matrices. Rule-based models are sets of if−then rules, in which each rule has conditions for one or more of the attributes of the examples, in this case the numerical values of the individual scoring functions. These rules can be derived from data either for the purpose of making categorical predictions (classification) or numerical predictions (regression). In this study, classification models were created only. Models were induced in a number of ways. Two basic rule-induction strategies, namely, recursive-partitioning (RP),[36] which produces decision trees, and covering (C),[37] which produces unordered sets of rules, were used both individually and in combination with the ensemble learning scheme bagging.[38] Bagging generates ensemble models consisting of a preselected number of basic models by generating repeated bootstrap replicates of the training examples. A large number of models were generated and compared. The models were evaluated using cross-validation with the training set, and the best model for each set of ligands and target protein was put forward for final evaluation with the external validation set.

PLS and PLS-DA. The relationships between the dependent values (active = 1 or inactive = −1) for each data set and the computed scoring values for each compound were determined using the PLS (partial least squares projections to latent structures) method,[22] employing an in-house program implemented at AstraZeneca, with core algorithms very similar to those implemented in SIMCA from Umetrics AB (http://www.umetrics.com). The number of significant components was determined using a 4-fold cross-validation procedure.[39] The difference between ordinary PLS analysis and PLS discriminant analysis (PLS-DA) is that in the former analysis the dependent variable is a continuous variable while in the latter type of analysis there are only two levels, in this case 1 or −1, for the dependent variable, related to the two classes under investigation. The standard classification of the results from PLS-DA is as follows: positive predictions from the model are regarded as belonging to active class (1), while negative predictions from the model are regarded as belonging to inactive class (−1). However, if the classes are unevenly populated, which is the case in the investigations presented here, the cross-validation procedure may indicate that there is a need to shift the cutoff limit for assessing class membership from zero (0) to a number that ensures a more balanced prediction of false positives and false negatives.

Bayesian Classification. MatLab (MathWorks Inc., http://www.mathworks.com) was used to implement a simple Bayesian two-class classifier, assuming normally distributed docking scores. A Bayesian classifier is constructed using Bayes theorem, assuming a distribution for the data and a priori probabilities for the classes (eq 5).

Set example E to class:

$$\arg\max_{i} P(C_i|E),\ P(C_i|E) = \frac{P(C_i)P(E|C_i)}{P(E)} \qquad (5)$$

Here E is an example being classified and C_i is a class (e.g., active). $P(E|C_i)$ can be calculated given a distribution. As is common, the normal distribution was used, which requires estimating the mean value vectors and covariance matrices of our two classes. The multivariate normal distribution is assumed for computational simplicity, not because of any inherent property of multidimensional scoring data. The multivariate normal distribution is given by eq 6.

$$f(\bar{X}) = \frac{1}{(2\pi)^{3/2}|C|^{1/2}} e^{-1/2(\bar{X}-\bar{\mu})^T C^{-1}(\bar{X}-\bar{\mu})} \qquad (6)$$

Here C is the covariance matrix, and μ is the expectation value vector. The actual classifier is implemented by reformulating eq 5 using probability density functions and inserting eq 6. Using the natural logarithm results in the discriminant function eq 7.

$$g_i(\bar{x}) = P(i)\left(-\frac{1}{2}\ln\left|C_i\right| + \bar{\mu}_i^T C_i^{-1}\bar{x} - \frac{1}{2}\bar{\mu}_i^T C_i^{-1}\bar{\mu}_i - \frac{1}{2}\bar{x}^T C_i^{-1}\bar{x}\right) \qquad (7)$$

$P(i)$ is the a priori probability for class i. Classifiers were built using different a priori probabilities of a given example to belong to the active class. Varying the a priori probability can be seen as weighting the different classes, and we have tested different levels to estimate the impact of doing this. In a real case, the a priori probability of being active, i.e., the ratio of actives in the total set being classified (e.g., a large set of commercial compounds or the in house compound collection of a pharmaceutical company) is often not known. In the constructed evaluation presented here, it is known that the ratio of actives in the evaluation set is similar to that of the training set. Therefore, we show the results from using the actual priors, as calculated from the ratio of inactives and actives in the training set, even though in an actual case the a priori probability of being active can be varied to achieve a satisfying number of predicted actives.

$P(E)$ is simply a normalization factor and can be disregarded when comparing a posteriori probabilities of different classes. The training set is used to estimate the means and covariance matrices of the two classes. An example vector is classified to belong to the active class if the a posteriori probability, $P(C_i|E)$, for that class is larger than the probability of the inactive class, which is equivalent to the discriminant function $g_i(x)$ being larger for the active class.

A Bayesian classifier is optimal if the dimensions are independent, which is most likely not the case for multidimensional scoring data, but given its simplicity it often works remarkably well also for dependent data.[23]

Classical Consensus Scoring. As a comparison a simple classical consensus scoring classifier was built. The implementation is identical to the default behavior of the classifier in Tripos CScore.[12] For each dimension (i.e., scoring function), the range is calculated and halved, giving a threshold value. The minimum score in the training set for each scoring function is subtracted from each individual score. If the resulting number is higher than the threshold (calculated as half the range) for a specific scoring vector, that vector is given a score of one for that particular scoring function. Hence, in our case the maximum consensus score is 7. A classifier is built by setting a threshold for being active in "consensus scores", typically, 5−6 in our case.

Clark et al.[12] compared four different consensus scoring approaches, of which CScore is one, and got comparable results. Therefore, we have only implemented CScore as an example of classical consensus scoring.

Acknowledgment. We gratefully acknowledge Dr. Wolfgang Sippl for supplying the 3D structures of the ERα "mimics" set of known actives and Dr. Jerk Vallgårda for selecting the set of diverse drug-like molecules from MDDR used as inactives. Dr. Evert Homan and Dr. Anders Karlén are acknowledged for critical review of the manuscript.

References

(1) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **2002**, *47*, 409−443.

(2) Sippl, W.; Contreras, J. M.; Parrot, I.; Rival, Y. M.; Wermuth, C. G. Structure-based 3D QSAR and design of novel acetylcholinesterase inhibitors. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 395−410.

(3) Ha, S.; Andreani, R.; Robbins, A.; Muegge, I. Evaluation of docking/scoring approaches: a comparative study based on MMP3 inhibitors. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 435−448.

(4) Abagyan, R.; Totrov, M. High-throughput docking for lead generation. *Curr. Opin. Chem. Biol.* **2001**, *5*, 375−382.

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

(6) Schapira, M.; Raaka, B. M.; Samuels, H. H.; Abagyan, R. Rational discovery of novel nuclear hormone receptor antagonists. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 1008−1013.

(7) Filikov, A. V.; Mohan, V.; Vickers, T. A.; Griffey, R. H.; Cook, P. D. et al. Identification of ligands for RNA targets via structure-based virtual screening: HIV-1 TAR. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 593−610.

(8) Schapira, M.; Totrov, M.; Abagyan, R. Prediction of the binding energy for small molecules, peptides and proteins. *J. Mol. Recognit.* **1999**, *12*, 177−190.

(9) Åqvist, J.; Marelius, J. The linear interaction energy method for predicting ligand binding free energies. *Comb. Chem. High Throughput Screening* **2001**, *4*, 613−626.

(10) 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.

(11) 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.

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

(13) Brzozowski, A. M.; Pike, A. C.; Dauter, Z.; Hubbard, R. E.; Bonn, T.; et al. Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature* **1997**, *389*, 753−758.

(14) Natchus, M. G.; Bookland, R. G.; Laufersweiler, M. J.; Pikul, S.; Almstead, N. G. et al. Development of new carboxylic acid-based MMP inhibitors derived from functionalized propargylglycines. *J. Med. Chem.* **2001**, *44*, 1060−1071.

(15) Kryger, G.; Silman, I.; Sussman, J. L. Structure of acetylcholinesterase complexed with E2020 (Aricept): implications for the design of new anti-Alzheimer drugs. *Struct. Fold Des.* **1999**, *7*, 297−307.

(16) Nar, H.; Bauer, M.; Schmid, A.; Stassen, J. M.; Wienen, W.; et al. Structural basis for inhibition promiscuity of dual specific thrombin and factor Xa blood coagulation inhibitors. *Structure (Cambr.)* **2001**, *9*, 29−37.

(17) Jones, G.; Willett, P.; Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, *245*, 43−53.

(18) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425−445.

(19) Schrödinger *FirstDiscovery 2.5 Operating Manual*; Schrödinger Press: 2003.

(20) Abagyan, R.; Totrov, M. *ICM* online manual, http://www.molsoft.com/.

(21) Abagyan, R.; Totrov, M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.* **1994**, *235*, 983−1002.

(22) Wold, S.; Johansson, E.; Cocchi, M. PLS−Partial least-squares projections to latent structures. In *3D QSAR in Drug Design*; *ESCOM*: Leiden, 1993; pp 523−550.

(23) Domingos, P.; Pazzani, M. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning* **1997**, *29*, 103−130.

(24) *Rule Discovery System (RDS) 0.8*, http://www.compumine.com; Compumine AB.

(25) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.

(26) Wilton, D.; Willet, P. Comparison of Ranking Methods for Virtual Screening in Lead-Discovery Programs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 469−474.

(27) Patel, Y.; Gillet, V. J.; Bravi, G.; Leach, A. R. A comparison of the pharmacophore identification programs: Catalyst, DISCO and GASP. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 653−681.

(28) Terp, G. E.; Johansen, B. N.; Christensen, I. T.; Jørgensen, F. S. A New Concept for Multidimensional Selection of Ligand Conformations (MultiSelect) and Multidimensional Scoring (MultiScore) of Protein−Ligand Binding Affinities. *J. Med. Chem.* **2001**, *44*, 2333−2343.

(29) Clark, R. D. OptiSim: An Extended Dissimilarity Selection Method for Finding Diverse Representative Subsets. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1181−1188.

(30) Sadowski, J.; Gasteiger, J. From atoms and bonds to three-dimensional atomic coordinates: Automatic model builders. *Chem. Rev.* **1993**, *93*, 2567−2581.

(31) Ash, S.; Cline, M. A.; Homer, R. W.; Hurst, T.; Smith, G. B. SYBYL line notation (SLN): A versatile language for chemical structure representation. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 71−79.

(32) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein−ligand interactions: a simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791−804.

(33) 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.

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

(35) Jones, G.; Willett, 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.

(36) Quinlan, J. R. Induction of decision trees. *Machine Learning* **1986**, *1*, 81−106.

(37) Boström, H. Covering vs Divide-and-Conquer for Top-Down Induction of Logic Programs. *Fourteenth International Joint Conference on Artificial Intelligence*; Morgan Kaufmann: San Mateo, California, 1995; pp 1194−1200.

(38) Breiman, L. Bagging Predictors. *Machine Learning* **1996**, *24*, 123−140.

(39) Wold, S. Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics* **1979**, *20*, 379−405.

(40) Sippl, W. Receptor-based 3D QSAR analysis of estrogen receptor ligands−merging the accuracy of receptor-based alignments with the computational efficiency of ligand-based methods. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 559−572.

(41) Shi, L. M.; Fang, H.; Tong, W.; Wu, J.; Perkins, R.; et al. QSAR models using a large diverse set of estrogens. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 186−195.

(42) Contreras, J. M.; Parrot, I.; Sippl, W.; Rival, Y. M.; Wermuth, C. G. Design, synthesis, and structure−activity relationships of a series of 3-[2-(1-benzylpiperidin-4-yl)ethylamino]pyridazine derivatives as acetylcholinesterase inhibitors. *J. Med. Chem.* **2001**, *44*, 2707−2718.

(43) Matter, H.; Defossa, E.; Heinelt, U.; Blohm, P. M.; Schneider, D. et al. Design and quantitative structure−activity relationship of 3- amidinobenzyl-1H-indole-2-carboxamides as potent, nonchiral, and selective inhibitors of blood coagulation factor Xa. *J. Med. Chem.* **2002**, *45*, 2749−2769.

JM030896T