Journal of Computer-Aided Molecular Design (2005) 19: 483–497
DOI 10.1007/s10822-005-9002-6

Surrogate docking: structure-based virtual screening at high throughput speed

Sukjoon Yoon[a,b], Andrew Smellie[a], David Hartsough[a] & Anton Filikov[a,*]
[a]*ArQule, Inc, 19 Presidential way, Woburn, MA, 01801, USA;* [b]*Department of Biological Sciences, Sookmyung Women's University, Hyochangwon-gil 52, Sookmyung, Yongsan-gu, Seoul, Republic of Korea*

Received 13 January 2005; accepted 6 July 2005

Key words: Bayesian categorization, docking, high throughput virtual screening, radial basis functions

Summary

Structure-based screening using fully flexible docking is still too slow for large molecular libraries. High quality docking of a million molecule library can take days even on a cluster with hundreds of CPUs. This performance issue prohibits the use of fully flexible docking in the design of large combinatorial libraries. We have developed a fast structure-based screening method, which utilizes docking of a limited number of compounds to build a 2D QSAR model used to rapidly score the rest of the database. We compare here a model based on radial basis functions and a Bayesian categorization model. The number of compounds that need to be actually docked depends on the number of docking hits found. In our case studies reasonable quality models are built after docking of the number of molecules containing ~50 docking hits. The rest of the library is screened by the QSAR model. Optionally a fraction of the QSAR-prioritized library can be docked in order to find the true docking hits. The quality of the model only depends on the training set size – not on the size of the library to be screened. Therefore, for larger libraries the method yields higher gain in speed no change in performance. Prioritizing a large library with these models provides a significant enrichment with docking hits: it attains the values of ~13 and ~35 at the beginning of the score-sorted libraries in our two case studies: screening of the NCI collection and a combinatorial libraries on CDK2 kinase structure. With such enrichments, only a fraction of the database must actually be docked to find many of the true hits. The throughput of the method allows its use in screening of large compound collections and in the design of large combinatorial libraries. The strategy proposed has an important effect on efficiency but does not affect retrieval of actives, the latter being determined by the quality of the docking method itself.

Introduction

In recent years virtual high throughput screening (VHTS) has become a very important technique. As compound collections continue to grow millions of compounds are available for biological testing. VHTS can be used as a method complementary to high throughput screening (HTS) in order to increase screening performance [1–3]. The methods currently used are limited to ligand-based virtual screening (LBVS), where only the properties of known binding ligands are considered. Receptor-based screening (where interactions of ligands with putative receptors are modeled) is slow and impractical or even impossible for large collections. LBVS methods used for VHTS are based on the analysis of molecular similarity [1], hence they depend on the knowledge of

*To whom correspondence should be addressed. Phone: +1-781-994-0647, Fax: +1-781-376-6019, E-mail: afilikov@arqule.com

experimentally active compounds. Molecular docking [4, 5] to a receptor structure does not require data on active compounds. Another method of virtual screening, 3D pharmacophore search [6–8] depends either on the knowledge of experimental actives or a receptor structure if pharmacophores are derived from the receptor. All three methods complement each other: LBVS is much faster, but docking or 3D pharmacophore search often can find molecules structurally more different from the known actives. Bringing receptor-based screening to a high throughput level would be an invaluable benefit for drug discovery.

High quality molecular docking (50,000–100,000 conformations/poses sampled) currently takes 30–60 s per molecule on a 3.0 GHz CPU. Docking of large libraries, such as collections of vendor compounds, can take many days even on computer clusters with multiple CPUs. 3D pharmacophore search run times are notoriously difficult to estimate, as they are so dependent on the query. Current DB search systems either precompute a set of conformations for a molecule prior to searching [9], or perform a flexible search from a limited number of conformations [10]. Either method is impractical for very large searching on the scale of millions of molecules.

The timing issue becomes insurmountable when it comes to receptor-focused design of large combinatorial libraries. In this problem one needs to score a large number of compounds ($\sim 10^4$–10^7) of an enumerated library in order to choose the best subsets of reagents out of the initial lists of candidates. These reagent subsets should give the best total score in the designed library. Typically such design is performed a number of times in order to compare different chemistries or scaffolds. In recent years tools have been developed that use molecular mechanics-based sampling to evaluate combinatorial libraries. None of the methods use docking of complete libraries of products, because this is too computationally expensive. They are based on a simplified approach: a conformational search and energy evaluation of substituents attached to a core moiety with the coordinates of the core atoms being fixed. The placement of the core is done in the following ways. It can be derived from docking of a representative sample of products [11], or coordinates of the core can be obtained by docking the core itself [12–14]. Finally, the core can be placed on the basis of known interactions of similar compounds [15]. The simplifications made in these algorithms in order to avoid docking of full combinatorial libraries made structure-based design of combinatorial libraries a practical approach. Simultaneously these simplifications restrict the utility of the methods since only one or few conformation/orientations of the core moiety are evaluated.

We have developed a fast screening method, which can be used for combinatorial or non-combinatorial compound sets. It consists of the following four steps:

1. Docking and scoring of a limited number of compounds ($\sim 10^3$–10^4) randomly chosen from the library to be screened or from any other library,
2. Building a model to predict the best docking score, using the values of various 2D descriptors of the compounds,
3. Using the predictive model to prioritize the molecules prior to (i) purchase or synthesis and experimental testing and/or (ii) true docking (optional).
4. (Optional). Docking a fraction of the prioritized molecules to find the true docking hits (i.e. compounds that dock with low scores).

The initial training set can be obtained by random or diverse selection from the library to be screened or from any other library. Ideally, this set must contain a significant number ($\sim 10^2$) of good binders, i.e. compounds scored better than the binder threshold when docked into the target of interest.

Any categorization or QSAR model can be used in the method. We chose a Laplacian-Modified Bayesian Classifier with Extended Connectivity Fingerprints (ECFP) and Functional Connectivity Fingerprints (FCFP) implemented in Pipe Line Pilot [16] because of technical ease of use. For comparison of model building methods we also used an in-house implementation of Radial Basis Functions (RBFs) [17] using the same fingerprints. Details are given in the Methods section.

In this work we optimize descriptors used in the method as well as the size of the training set and cutoff score for binders using the Bayesian classifier. We include validation of the method on two libraries: the NCI library and a CDK-focused combinatorial library. This validation gauges the performance of the method by its ability to retrieve docking hits. We also assess the ability of the method to retrieve actual experimental binders, although this ability primarily depends on the enrichment provided by the docking method itself. This is an important point: the fast screening method described here is *not* guaranteed to accelerate the recovery of true experimental binders. It will however speed up the recovery of compounds that dock well into the target of interest. True experimental binders will be recovered more quickly only if the docking method being used has the ability to enrich true experimental binders. This has been the subject of much study [18, 19], and is beyond the scope of this work.

The authors of the previously published works [18, 20–22] were the first to realize that one can benefit by combining molecular docking with Bayesian classification method. It was shown that the Bayesian classifier trained on docking scores of experimentally known binders can successfully be used for virtual screening [20]. It was also shown that rescoring by the Bayesian classifier trained on docking results can improve the docking enrichment [18, 21, 22]. In our work combining Bayesian classifier (or radial basis functions or any other 2D QSAR method) with docking serves a different purpose: we train a 2D model on docking results from a small diverse database (not necessarily derived from the database to screen) in order to replace docking with a much faster 2D model, i.e. we screen the large database with the 2D model rather then with docking. No knowledge of structure of experimental binders is necessary for the method.

If the Surrogate Docking Method is used with the Bayesian classifier and ECFP/FCFP fingerprints, it also gives the structural features most often encountered in binders and non-binders. This greatly facilitates interpreting the model for medicinal chemists. The usefulness of the structural features identified by the Bayesian model was previously discussed in [21, 23].

Methods

The ICM package (Molsoft) was used for docking and scoring. The details of ICM docking and scoring methods are described in detail elsewhere [24]. Briefly, the ICM docking algorithm applies Monte-Carlo-based torsional coordinate mechanics to dock flexible ligands into a grid representation of the receptor. During docking a protein is replaced by five energy grids representing hydrophobic, van der Waals for heavy atoms, van der Waals for hydrogen atoms, hydrogen bond and electrostatic energies of a predefined ligand binding site. The best energy docked conformation is subsequently evaluated using the scoring function optimized for discrimination of active ligands from random drug-like molecules [25]. The scoring function includes terms calculated on the energy grids and terms calculated with an all atom representation. The function takes into account the change in internal energy of the ligand upon binding, van der Waals, electrostatic and hydrogen bond interactions between the ligand and the receptor. It also accounts for ligand entropy loss, desolvation of hydrogen bond donors and acceptors, solvation electrostatic energy change upon binding (calculated by solving the Poisson equation by using the boundary element algorithm) and non-polar contribution to the free energy upon binding.

Coordinates of the structure of CDK2 (1G5S) and the estradiol receptor (1A52) were taken from the Protein Data Bank [26]. The docking was performed on the cavity occupied by the crystal ligand, defined as all the space within 5 Å of any ligand atom. Before docking, all ligands, water molecules and other atoms or molecules except the protein molecule were removed. All protein molecules were prepared for docking using the standard ICM receptor preparation procedure, which includes generation and optimization of hydrogen coordinates, optimization of histidine protomer form, optimization of proline conformation and optimization of χ^2 and χ^3 torsional angles for asparagine and glutamine residues.

NCI library

A total of 250, 251 compounds in SMILES format were downloaded from the NCI web site [27]. We first removed duplicate compounds and also applied filters based on Lipinski's rules and other criteria. Specifically, the filters were:

- number of atoms $> =$ 10
- nitrogen count + oxygen count $< =$ 10
- molecular weight $< =$ 500 and molecular

weight > = 100

- number of H donors < = 5
- AlogP < = 5
- number of fragments = 1 (entries with more than one molecule were removed)
- number of rotatable bonds < = 5.
- Compounds containing any atoms other than H, C, N, P, O, S, F, Cl, Br and I were removed.

We then generated 3D conformations of these filtered compounds using Concord 4.0.2 [28]. Thus, a total of 114,718 compounds were used in this study.

CDK-focused combinatorial library

In order to test the performance of the method for a combinatorial library, we generated a library of 2,6,9-trisubstituted purines according to the scheme shown in Figure 7. This chemistry scheme was used to synthesize a library, which has been shown to have many CDK2 inhibitors as well as many inhibitors of a number of other kinases [29]. The reagents were selected from the initial pool of all secondary amines from the ACD [30], a total of 45,188 reagents. Molecules with molecular weight of more than 150 were filtered out. Then all salts, bis-secondary amines, racemic mixtures, primary amines and amides, as well as reactive and toxic functionalities were removed [31]. This procedure yielded 360 reagents. A diverse subset of 63 amines was selected from this set using daylight fingerprint-based clustering [32]. The topological clustering algorithm of Butina et al. was used to select reagents [32]. Clusters were constructed with an intra-cluster radius of 0.7 (in Daylight fingerprint tanimoto units), which means that each member of a cluster is guaranteed to be within a similarity of 0.7 to its cluster centroid. A similar procedure was employed for primary alcohols to give 20 diverse reagents. The library was enumerated using MapMaker software [33] to give a total of 79,380 compounds. Smiles representations of these compounds were converted to 3D conformations using Concord 4.0.2.

Estradiol receptor compound set

To demonstrate and evaluate the utility of our approach, we tested this new surrogate docking procedure on the estrogen receptor α (ERα) system [34]. ERα is a good choice by virtue of the wealth of published experimental binding affinity data to this receptor for many compounds [35]. Given the diverse range of compounds that may bind to the ERα and exert an effect on human and animal health, there is considerable interest in understanding the details of ligand-ERα affinity and developing techniques to predict the affinity of compounds for ERα. For the test compounds, we obtained the SMILES representation of 205 compounds for which experimental binding affinities were taken from the published studies by Tong and co-workers [34, 35]. The experimental binding affinity log(RBA) was measured relative to the binding affinity, defined as the logarithm of the percent ratio of the IC_{50} between 17β-estradiol and a test compound. Thus the RBA of 17β-estradiol is 100, and log(RBA) of 17β-estradiol is 2. These workers measured log(RBA) as low as −5, defining compounds with log(RBA) < −5 as "nonbinders". Initial 3D structures were generated for these compounds using Concord 4.0.2.

Descriptors

2D chemical feature calculations were carried out using various descriptors available in Scitegic's Pipeline Pilot [16]. The selected descriptors were ECFP_6 (extended connectivity fingerprints) and FCFP_6 (functional-class fingerprints) topology fingerprints [36], 166 MDL topological keys [37] and various molecular properties (e.g. molecular weight, AlogP [38, 39], the number of H-bonding donors/acceptors and the number of rotatable bonds) The generation of an ECFP or FCFP fingerprint for a molecule begins with the assignment of an initial atom code for each heavy (non-hydrogen) atom in the molecule. Only differences in the initial atom code distinguish ECFPs and FCFPs; once the codes are assigned, both fingerprints are developed through the same process. For ECFPs, the initial atom code is derived from the following features: the number of connections to the atom, the element type, the charge, and the mass of the atom. Atoms that differ in any of these features generate a different and unique ECFP initial atom code. For FCFPs, the initial atom code is based on the quick estimate of the functional role the atom plays. The role is a combination of six properties: (1) being a hydrogen-bond acceptor; (2) being a hydrogen-bond donor; (3) being positively ionized or positively ionizable; (4) being negatively ionized or negatively ionizable; (5) being aromatic; or (6) being a halogen. Once the initial atom codes are determined, an algorithm is applied to (a) identify fragments of topology centered on each atom and radiating out symmetrically to a given depth and (b) map this topology uniquely to a bit in a long virtual bit string [40]. A different bit string is computed for each molecule and these can be compared using the Tanimoto coefficient to estimate topological similarity. For comparison, a hashed Daylight fingerprint [41] was also calculated for each compound using the Daylight toolkit.

Model building: Bayesian classification

For Pipeline Pilot ECFP and FCFP fingerprints and Daylight fingerprints, a naïve Bayesian model [42] was used to identify bits in the fingerprint (i.e. sub-structural fragments) that are significant in separating compounds into "good" and "bad" classes from a set of training compounds. For this work, the good class was defined as the set of good virtual binders (i.e. those with a low docking score), and the bad class was defined as the set of poor binders.

Model building: RBFs

As a comparison we also built models based on RBF. These functions $s(x)$ attempt to interpolate a real-valued function $f(x)$ over a range of real values x. In this case, $f(x)$ is the docking score, and x are the descriptors used in the study. More formally:

$$f(x) \approx s(x) = p(x) + \sum_{i=1}^{N} w_i \Phi(r_i) \tag{1}$$

where $p(x)$ = a low order polynomial in x, N = number of training points used to construct the interpolation, w_i = weight accorded the ith training point in the interpolation, $\Phi(r_i)$ = a function that measures the distance from the ith training point to x.

For simplicity, we used the Pipeline Pilot fingerprints of the molecules as the descriptors in the study. It has been previously shown that the tanimoto distance between two bit vector fingerprints is a valid metric [43] (i.e. obeys the triangle inequality), so we used it as our distance function. $\Phi(r_i)$. We neglected the function $p(x)$ and did not use it in this study. The methods of training the model to find the weights $w(i)$ proceeds as follows: For a set of N training points, N equations of the form Equation. 1 are generated. Thus we have N equations where $f(x)$ is known (i.e. the docking scores), $\Phi(r_i)$ can be computed (from the tanimoto distance of the ECFP fingerprint from the ith training compound to all other compounds), and the N unknowns (the weights w_i) can be found by solving for these weights with conventional linear algebra methods [44]. To predict a new molecule using the RBF model, the descriptors x are computed for the new molecule and distance $\Phi(r_i)$ from the new molecule to all molecules in the training set is computed. These distances are substituted into Equation 1 to give an estimate of the predicted value (in this case, the docking score).

Data analysis

We analyzed the prediction results from the predictive models by using receiver operating characteristic (ROC) curves [45]. Ranking the molecules, with the predicted best binders coming first, according to the predictive model generates ROC curves. A ROC curve describes the tradeoff between sensitivity and specificity. Sensitivity is defined as the ability of the model to avoid "false negatives" (i.e. where a molecule classified as "bad" is really "good"), while specificity relates to its ability to avoid "false positives" (i.e. where a molecule classified as "good" is really "bad"). Thus, the area under the ROC curve (AUC) is a measure of the test accuracy. An AUC value of 0.5 represents a random prediction, while 1.0 represents a perfect prediction. In addition, the enrichment factor (EF) was calculated according to the following formula [46]:

$$\mathrm{EF} = \frac{N_{\mathrm{total}}}{N_{\mathrm{sampled}}} \times \frac{\mathrm{Hits}_{\mathrm{sampled}}}{\mathrm{Hits}_{\mathrm{total}}} \tag{2}$$

We used AUC and EF analyses for the comparison of performance of various chemical descriptors and predictive models on prioritizing good binders. Because of the rapid speed of construction of Bayesian models, these were used to find the "best mode" of model building in terms of descriptors used etc. Once the best mode was identified, it was applied to the RBF model building procedure.

Simple consensus scoring

How much overlap is there between the models? Will some form of consensus scoring improve the enrichments still further? To answer these questions, we implemented Borda [47] counting as a simple form of consensus scoring. An additional plot was added to both Figures 4 and 8 representing the enrichment obtained through this simple consensus model. If a molecule M has been ranked by T different models, it's rank according to the ith model is defined by R_{Mi}. Thus the consensus rank of the molecule according to the Borda method (R_M) is given by

$$R_M = \sum_{i=1}^{T} R_{M_i} \tag{3}$$

Thus, if each model ranks such that the "best" molecules are ranked first, the lower the Borda score, the "better" the molecule.

Results and discussion

Optimization of descriptors

Surrogate docking is used to pre-rank compounds prior to regular docking and is a fast screening based on supervised predictors developed from docking results of a limited set of compounds. This set must be diverse and must contain a number of docking hits (virtual binders) sufficient to build a model of a desired quality. It can be selected by a diverse or random selection from the library to be screened, or from any other library, which contains a significant number of virtual binders. The quality of the model depends on its type and also on the descriptors used. Any kind of categorization or QSAR method can be used in surrogate docking: (e.g. recursive partitioning [48], partial least-squares [49], neural networks [50], genetic algorithms [51], genetic function approximation etc). We used a Bayesian categorization model and a RBF model. The descriptors used to build a model should be capable of describing the molecular features responsible for binding, i.e. the features, which are more frequently encountered in the binders than in non-binders. In order to maximize the enrichment of docking hits we compared the performance of the method using several different descriptors and their combinations using the Bayesian model. The results are shown in Table 1. For the training set of 5000, 10,000 and 20,000 compounds, the ECFP descriptor outperforms all other single descriptors. Thus, this descriptor was used to build RBF models. Among combinations of several descriptors ECFP + Misc. outperforms the other combinations and

Table 1. Optimization of selection of descriptors using the NCI library. The prioritization of binders using different 2D descriptors is compared by AUC and enrichment factor analyses.

Size of the training set, total/actives	MDL[a]	ECFP[b]	FCFP[c]	Daylight[d]	Misc.[e]	ECFP + daylight	ECFP + misc.	ECFP + daylight + misc.
AUC analysis								
1000/11	0.66	0.67	0.60	**0.71**	0.61	0.71	0.68	**0.72**
5000/48	0.74	**0.79**	0.75	0.78	0.78	0.82	0.82	**0.83**
10,000/97	0.72	**0.79**	0.76	0.76	0.78	0.81	0.82	**0.83**
20,000/175	0.74	**0.84**	0.82	0.78	0.77	0.85	**0.86**	**0.86**
EF analysis								
1000/11	2.6	3.1	2.5	**3.4**	1.6	**3.5**	3.3	**3.5**
5000/48	4.3	**5.1**	4.6	3.8	3.2	4.9	**5.5**	**5.0**
10,000/97	3.2	**5.3**	4.8	3.4	3.0	4.8	**5.6**	4.9
20,000/175	4.0	**6.0**	5.6	3.7	2.8	5.4	**6.3**	5.5

EF is calculated for the top 10% of the test set. The bold font in each row represents the best results for individual descriptors and different combinations of descriptors, respectively. The test set includes 80,000 compounds.

[a] MDL keys.

[b] ECFPs (Pipeline Pilot).

[c] Functional Class Fingerprints (Pipeline Pilot).

[d] Daylight fingerprints.

[e] Misc.: logP, molecular weight, number of H-bond donors and acceptors, number of rotatable bonds.

slightly outperforms ECFP alone. This is the combination that was used in our further studies with the Bayesian classifier.

Why ECFP fingerprints perform slightly better in this particular learning task than MDL keys or Daylight fingerprints? Although we cannot state this conclusively, we can speculate that the properties of ECFPs, which are responsible for the better performance are the following two.

- They represent a much larger set of features than 960 features in the MDL private keys and these features are not "pre-selected", but are generated directly from the molecules.
- ECFPs represent information about tertiary and quaternary centers, which is not the case for path-based fingerprints such as Daylight fingerprints. One can see from Figures 6 and 9 – our two case studies – that the features identified by ECFPs as important for binders contain many tertiary and quaternary centers.The finding that FCFPs (functional-class fingerprints) perform worse than ECFPs is more surprising: evidently the atom typing based on the initial atom code used in ECFPs works better than the functional typing used in FCFPs.

Performance of the method on test sets of different size

When a Bayesian model is built using docking results of the initial set of compounds, the model can be used for screening any library, and performance of the method should not depend on the size of this library: this size is not related to the size of the initial set of compounds, which determines the quality of the model. Hence, the larger the library to screen, the bigger gain in speed one gets by switching to surrogate docking. In order to illustrate this point we tested the performance of the method on two test sets of different sizes (10,000 and 80,000 compounds). The results are shown in Table 2. As expected, the performance does not deteriorate as the size of the test set gets larger.

Table 2. AUC analysis of the performance of the method on test sets of different sizes.

Size of the training set, total/actives	Size of the test set	
	10,000	80,000
1000/11	0.64	0.68
5000/48	0.83	0.82
10,000/97	0.84	0.82
20,000/175	0.87	0.86

The NCI library is used. The set of descriptors used is ECFP + Misc.

Optimization of the size of the training set

What size training set should be used to build a model? It depends on the following three factors: (i) diversity of the set, (ii) number of docking hits in the set, and (iii) desired quality of the model. We investigated the dependence of the Bayesian method performance on the size of the training set obtained by random selection from the NCI library. The results are shown in Figure 1: increase in the size of the training set above 5000/48 (total/actives) does not provide a significant increase in the performance as judged by both AUC and EF analysis. These numbers cannot be considered as a general guidance when the method is applied to other libraries: the number of docking hits in the set and its diversity can be very different in different libraries. The minimal size of the training set can be derived from the cross-validated enrichment factor for the training set. This analysis can be performed simultaneously with docking, so that docking is automatically stopped, when a desired value for

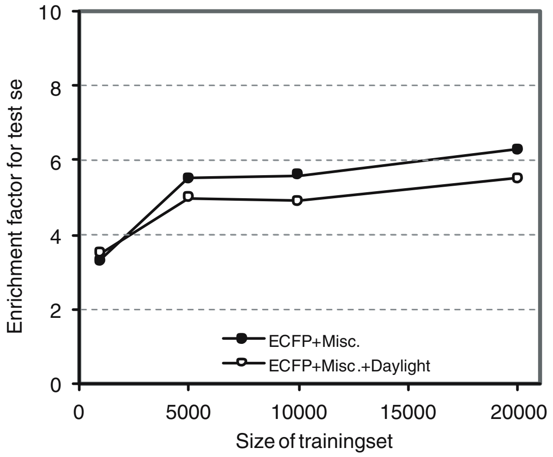

Figure 1. The effect of the training set size on the method performance for the NCI library. The test set includes 80,000 compounds. EF is calculated for the top 10% of the test set. Consequently, its theoretical maximum value is 10.

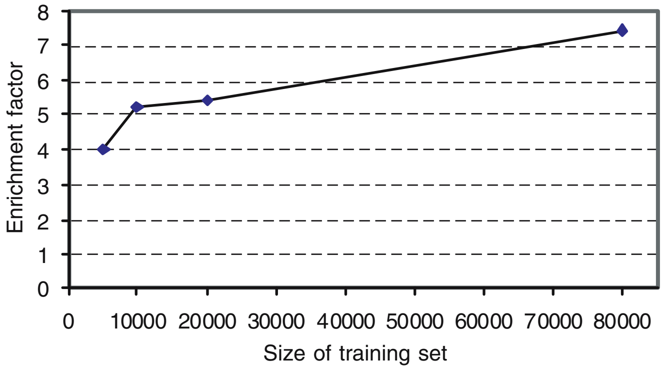

Figure 2. The effect of the training set size on the method performance for the NCI library: EF (5-fold cross-validated) on the training set. ECFP and Misc. descriptors were used for training. EF is calculated for the top 10% of the training set. Consequently, its maximum theoretical value is 10.

the factor is reached. This can be implemented on any platform including PipeLine Pilot. An example of such calculation is shown in Figure 2. This method is fairly robust at least in this case: it can be seen that the performance of the method as measured by cross-validated EF on the training set (Figure 2) is fairly consistent with its performance on the test set (Figure 1).

Optimization of the cutoff score for binders

The cutoff score (i.e. the score that divides "good" virtual binder from "bad" virtual binders) can be derived from different considerations. If the receptor structure is in complex with a ligand (cognate ligand), the docking score of this ligand can be used to derive the cutoff score for binders. Alternatively, the mean docking score of a number of known ligands can be used. The best way is to derive the cutoff by optimizing the performance of the method. The result of such optimization is shown in Figure 3. In this graph, AUC for the test set is plotted against different cutoff scores that differentiate good virtual binders from poor ones. One can see that a cutoff of −31 to −35 is optimal for this receptor/library combination. This result is easy to rationalize: a low cutoff score means the training set has very few good virtual binders to learn from; a high cutoff score means that compounds with very similar structural features appear in both the good and bad classes making model building more difficult. We use the cutoff of −35 for all studies reported here. In our case the ICM docking score of the cognate ligand is −46.2.

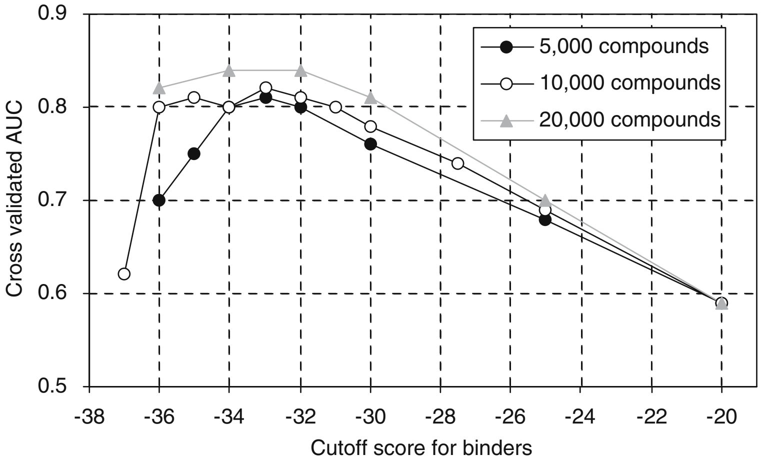

Figure 3. Dependence of cross-validated AUC on the cutoff score for binders for the NCI library.

The use of such an aggressive cutoff would result in a dramatic deterioration of the performance due to a large reduction in the number of virtual binders in the training set.

Case study. NCI database

The surrogate docking method using Bayesian classification and RBFs was applied to screen the NCI library on the structure of CDK2 (1G5S). The optimized values were used for the size of the training set (5000) and for the cutoff score for binders (−35). In other words the compounds with docking score ≤−35 were considered the binders, all the others are the non-binders. 80,000 compounds were used for the test set. Figure 4 shows the results. One can see that both methods provide significant enrichment with docking hits: an

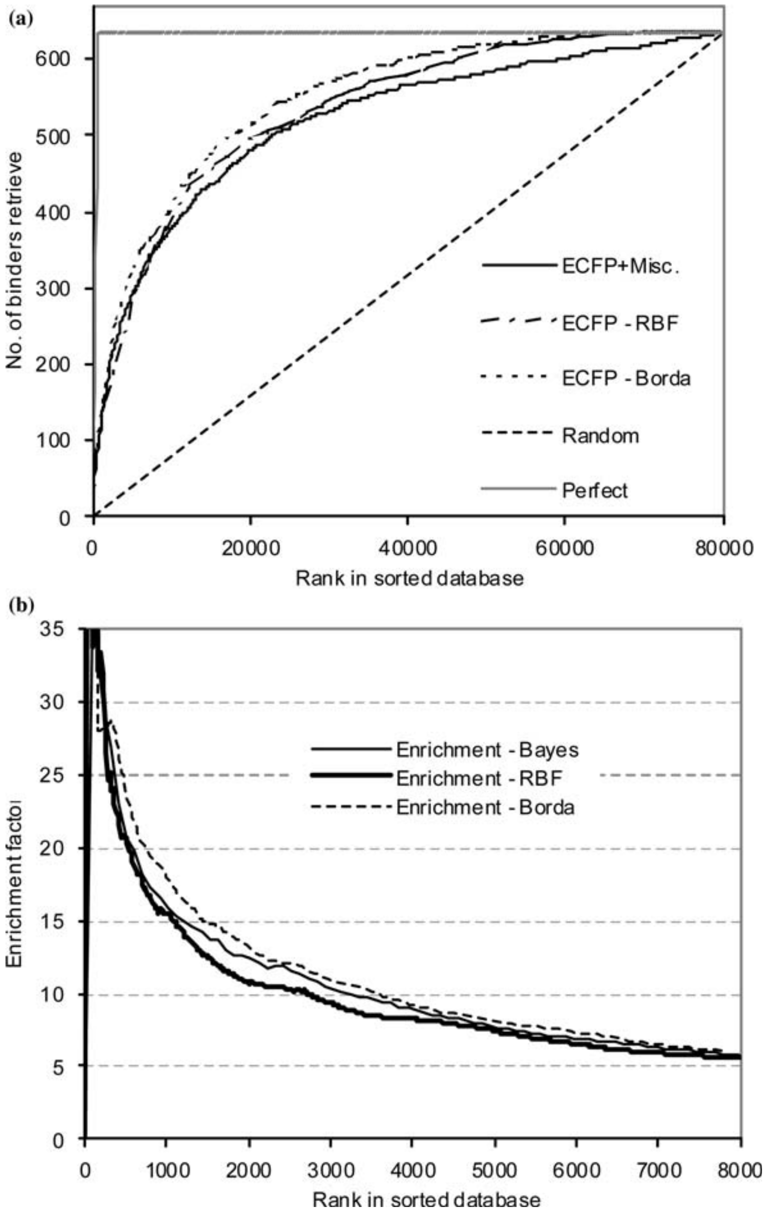

Figure 4. Case study for the NCI library (screening on CDK2 structure). 5000 and 80,000 compounds were used for training and test sets, respectively. (a) Number of compounds tested vs. number of binders retrieved for both Bayesian and RBF models, (b) EF for a point x is calculated for the fraction of the sorted database between 0 and x.

enrichment factor of ~35 is achieved at the beginning of the library sorted by the models. There is very little difference in the performance of Bayesian and RBF models, or using the simple Borda consensus.

Figure 5 shows histograms of ICM docking scores for the whole NCI library and two its subsets: 1000 randomly selected compounds and the best 1000 compounds selected by the Bayesian classifier. One can see a significant shift toward better scores for the compounds selected by the model. Based on the results from Figure 4, a similar result is expected using the RBF models. The average scores are −30.6 kcal/mol for the classifier compounds, −23.0 kcal/mol for the randomly selected compounds, and −22.8 kcal/mol

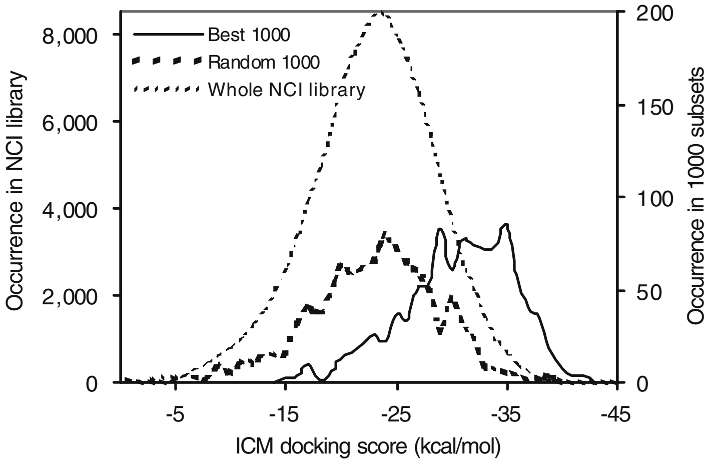

Figure 5. Histograms of ICM docking scores for the whole NCI library and two of its subsets: 1000 randomly selected compounds and the best 1000 compound selected by the Bayesian classifier (ECFP, Misc.). The average scores are: −30.6 kcal/mol for the Bayesian classifier compounds, −23.0 kcal/mol for the randomly selected compounds, and −22.8 kcal/mol for the whole NCI library (114,718 compounds).

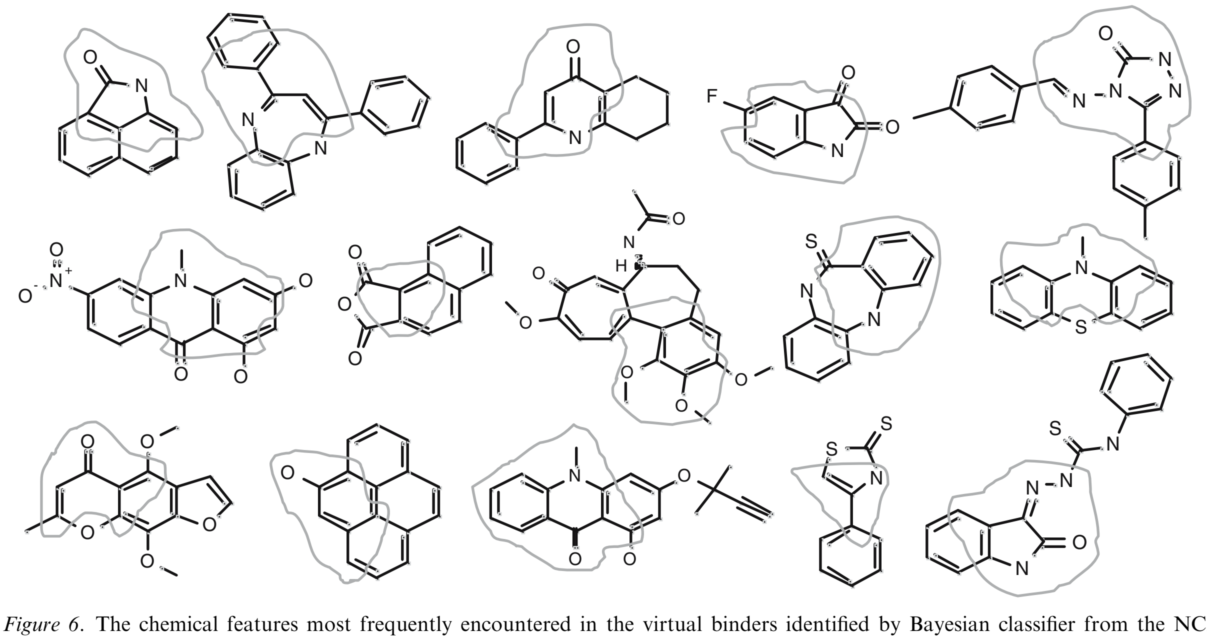

Figure 6. The chemical features most frequently encountered in the virtual binders identified by Bayesian classifier from the NCI library using ECFP + Misc. descriptors.

for the whole NCI library (114,718 compounds). An important feature of the Bayesian classifier that uses ECFP fingerprints is its ability to directly identify molecular fragments responsible for activity. Many other QSAR models as well as the docking method itself lack this ability, which results in the problem of interpreting the model in order to formulate specific recommendations for medicinal chemists. The binder features identified by the classifier for the NCI library are shown in Figure 6.

Case study. CDK-focused combinatorial library

The Bayesian and RBF methods were applied to screen the CDK-focused combinatorial library, generated from the chemical scheme depicted in Figure 7. The same optimized values were used for the size of the training set (5000) and for the cutoff score for binders (−35). 40,000 compounds were used for the test set. Figure 8 shows the results. The methods both provide significant enrichment with docking hits: they reach a value of ~16 at the beginning of the library sorted by the model score. The RBFs give a marginally better enrichment and AUC than the Bayesian model, and the consensus results are marginally better still. Figure 9 shows the binder features identified by the classifier.

Validation. estradiol receptor screening

Our case studies on the NCI and CDK-focused libraries showed that the surrogate docking can significantly enrich the library with docking hits. Figure 4a shows that the first ~5000 compounds out of 80,000 ranked by the Bayesian classifier contain roughly 50% of all docking hits. In a practical application one can dock these 5000 compounds to find the docking hits, or just screen experimentally all 5000 compounds. The decision will depend on the availability of resources. In the NCI and CDK-focused library validations the quality of docking is not considered at all. Any docking software, procedure or protocol can be used regardless of its ability to find experimental binders. The surrogate docking method does not improve or degrade the quality of the docking hits – it only helps to find them faster, at the expense of losing some hits. This clearly illustrates that surrogate docking is a practical compromise between speed and accuracy**.** From a practical point of view it would be useful to evaluate how much the surrogate docking dilutes the true experimental hits with false positives. We performed such an evaluation on a set of 205 compounds with known affinities to the estradiol receptor [34, 35]. A leave-one-out cross-validation procedure was used using the Bayesian classifier. Only the Bayesian method was used here because of the speed of construction of the models. We used the cutoff affinity for binders of −1 (logarithm of relative to estradiol binding affinity). This cutoff yields 42 active and 163 inactive compounds. The cutoff score for binders was −35 kcal/mol, which yields 48 active and 157 inactive compounds.

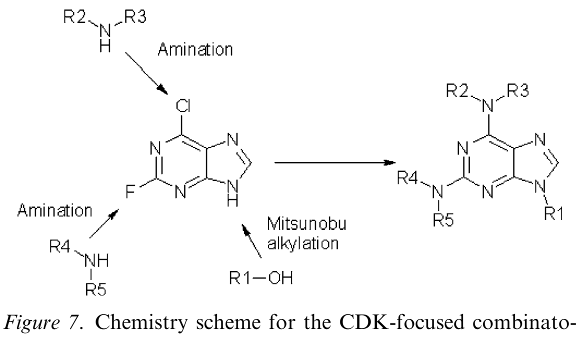

Figure 7. Chemistry scheme for the CDK-focused combinatorial library.

The results are shown in Figure 10. It can be seen that the enrichment yielded by the surrogate docking and the docking itself are fairly similar in this test case: the enrichment factor for the first 10% of the ranked database equals 3.7 and 2.6 for the docking itself and surrogate docking correspondently. Note that the maximum enrichment calculated for the first 10% of a database equals 10. The AUC analysis gives even closer results: 0.72 and 0.71 for the docking itself and surrogate docking respectively.

Practical recommendations

The surrogate docking method can be applied to a particular screening problem in a number of ways. We recommend the following protocol.

1. Dock and score a limited number of compounds using a preferred docking program.

Source of the compounds. These compounds can be a random or diverse selection from the library

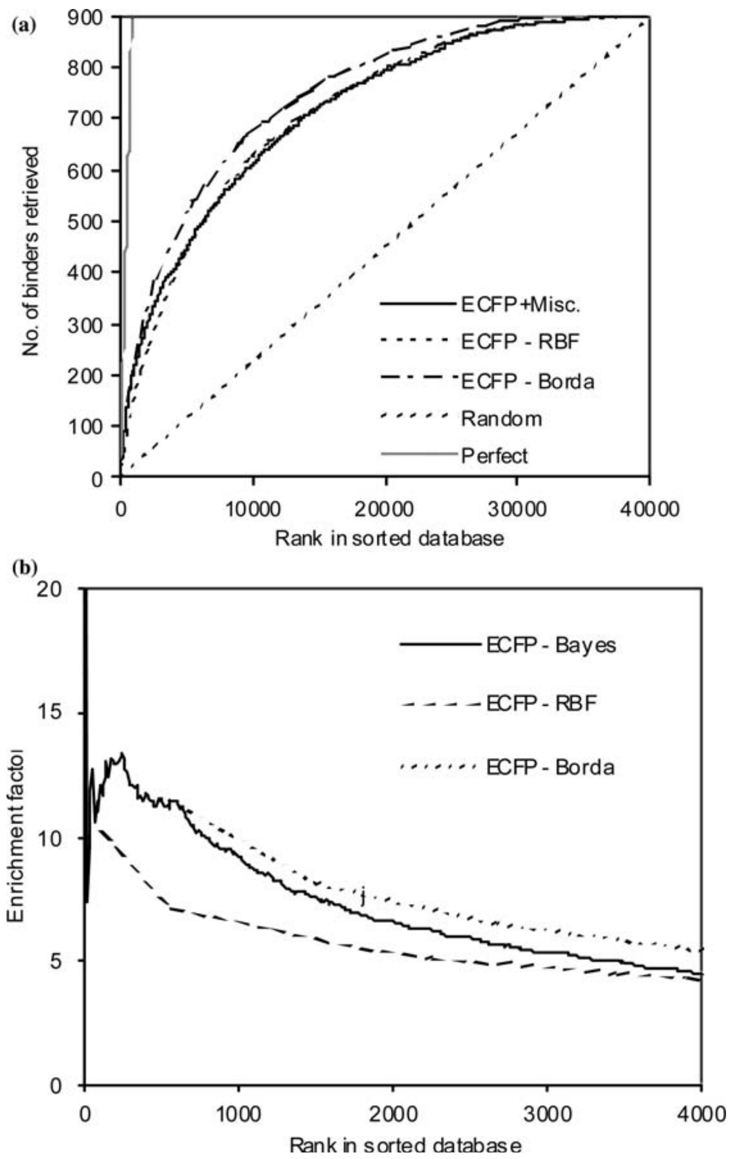

Figure 8. Case study. screening of the CDK-focused combinatorial library on CDK2 structure. 5000 and 40,000 compounds were used for training and test sets, respectively. (a) Number of compounds tested vs. number of binders retrieved, (b) EF curve. EF for a point *x* is calculated for the fraction of the sorted database between 0 and *x*. Note that only the first 10% of the sorted library is shown. For the Bayesian model: AUC = 0.81, EF of top 10% = 4.5. For the RBF model: AUC = 0.79, EF of top 10% = 4.32.

to be screened or from any other library, which preferably should be diverse and enriched with docking hits.

How many compound to dock? If it is reasonable to assume that the library to be screened is similar to the NCI library in terms of the number of docking hits and diversity, then this number can be chosen from Table 1 and Figure 1, given the desired enrichment factor or AUC. If this assumption is not valid, then the best procedure is to start docking compounds while simultaneously building the QSAR models and evaluating their quality by cross-validation (see Figure 2). The docking can stop when the desired enrichment factor or AUC is attained.

2. Build a QSAR or classification model using ECFP or ECFP + Misc descriptors.
3. Screen the rest of the library with the model built.

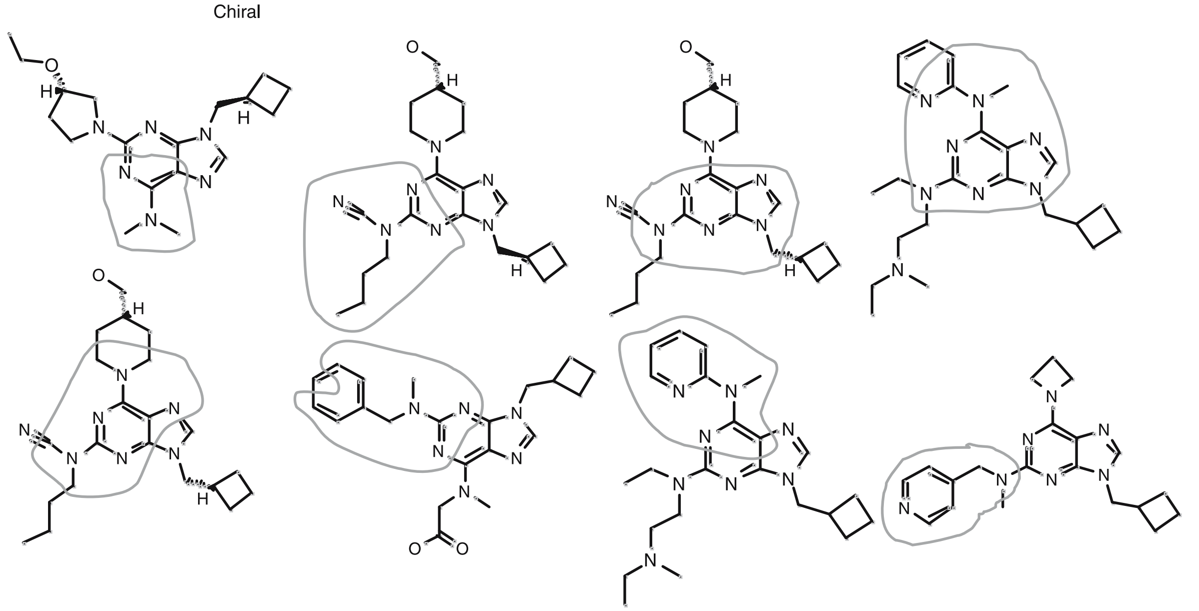

Figure 9. The chemical features most frequently encountered in the virtual binders identified by Bayesian classifier for the CDK focused library using ECFP + Misc. descriptors.

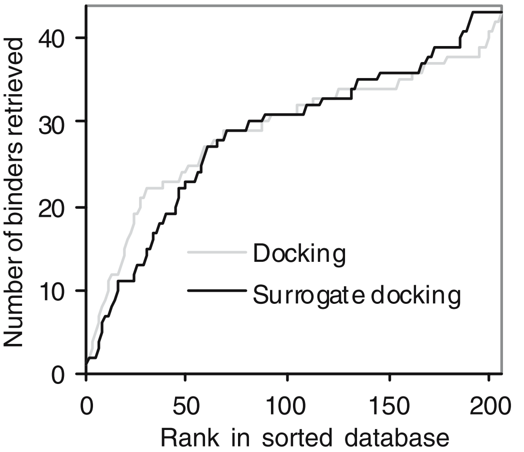

Figure 10. Accumulation curve for the set of 205 compounds with known affinities for estradiol receptor. We used the cutoff affinity for binders of −1 (logarithm of relative to estradiol binding affinity). This cutoff yields 42 active and 163 inactive compounds. The cutoff score for binders was chosen to roughly match the proportion of actives in the set: it was −35 kcal/mol, which yields 48 active and 157 inactive compounds.

4. (Optional). Dock a fraction of the prioritized molecules to find the true docking hits (i.e. compounds that dock with a low score). The size of this fraction is determined by the resources available as well as by the number of docking hits in the database.

The surrogate docking method combines true docking with screening by a QSAR model built on the results of limited docking. The method yields some scaffold-hopping, which is desirable in many drug discovery projects. It primarily results from the true docking part of the method.

Conclusions

We have developed a fast structure-based screening method, which uses a 2D QSAR model built on a training set obtained by docking of a limited number (~10^3–10^4) of diverse compounds. The molecules can be selected from the library to be screened or from another library, which preferably should be diverse and enriched with docking hits.

We have compared here two methods: a model based on RBFs and a Bayesian categorization model. Both methods provide significant enrichment of docking hits. In our two case studies the beginning of the libraries sorted by the classifier score yields 13- and 35-fold enrichment respectively. The Bayesian categorization method is more practical, since it is much faster. It also identifies the structural features, which are more frequently encountered in the binders than in non-binders. This can be helpful in interpretation of modeling results in order to formulate specific recommendations for medicinal chemists.

When a Bayesian or redial basis functions model is built using docking results of the initial set of compounds, the model can be used for screening any library, and performance of the method does not depend on the size of this library. The performance is only determined by the quality of the model, which, in turn, is only determined by the size of the initial set of compounds as well as by its diversity and the number of binders it contains. Therefore, the performance of the model does not depend on the size of the library, i.e. the performance gain is higher for larger libraries.

The analysis of the ability to enrich a database with experimental binders showed that the performance of the surrogate docking is comparable to the one of docking itself.

The throughput of the method allows its use in the screening of large compound collections and in the design of large combinatorial libraries.

Electronic supplementary material is available at http://dx.doi.org/10.1007/s10822-005-9002-6.

Acknowledgement

The authors gratefully acknowledge Dr Weida Tong (National Center for Toxicological Research, U.S. Food & Drug Administration, Jefferson, AR) for providing us the experimental data set and 2D structures of Estrogen Receptor compounds used in this study.

References

1. Stahura, F.L. and Bajorath, J., Comb. Chem. High Throughput Screen, 7 (2004) 259.
2. Bajorath, J., Nat. Re.v Drug Discov., 1 (2002) 882.
3. Engels, M.F. and Venkatarangan, P., Curr. Opin. Drug Discov. Develop., 4 (2001) 275.
4. Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. and Ferrin, T.E., J. Mol. Biol., 161 (1982) 269.
5. Tatsumi, R., Fukunishi, Y. and Nakamura, H., J. Comput. Chem., 25 (2004) 1995.
6. Holtje, H.D., Arch. Pharm. (Weinheim), 307 (1974) 969.
7. Steindl, T. and Langer, T., J. Chem. Inf. Comput. Sci., 44 (2004) 1849.
8. Eksterowicz, J.E., Evensen, E., Lemmen, C., Brady, G.P., Lanctot, J.K., Bradley, E.K., Saiah, E., Robinson, L.A., Grootenhuis, P.D. and Blaney, J.M., J. Mol. Graph. Model., 20 (2002) 469.
9. Smellie, A., Kahn, S.D. and Teig, S., J. Chem. Inf. Comput. Sci., 35 (1995) 285.
10. Hurst, T., J. Chem. Inf. Comput. Sci., 34 (1994) 190.
11. Sprous, D.G., Lowis, D.R., Leonard, J.M., Heritage, T., Burkett, S.N., Baker, D.S. and Clark, R.D., J. Comb. Chem., 6 (2004) 530.
12. Makino, S., Ewing, T.J. and Kuntz, I.D., J. Comput. Aided Mol. Des., 13 (1999) 513.
13. Sun, Y., Ewing, T.J., Skillman, A.G. and Kuntz, I.D., J. Comput. Aided Mol. Des., 12 (1998) 597.
14. Lamb, M.L., Burdick, K.W., Toba, S., Young, M.M., Skillman, A.G., Zou, X., Arnold, J.R. and Kuntz, I.D., Proteins, 42 (2001) 296.
15. Kick, E.K., Roe, D.C., Skillman, A.G., Liu, G., Ewing, T.J., Sun, Y., Kuntz, I.D. and Ellman, J.A., Chem. Biol., 4 (1997) 297.
16. Pipeline Pilot V 3.5, Scitegic Inc., (2004) San Diego.
17. Buhmann, M.D., Radial Basis Functions: Theory and Implementations, Cambridge University Press, 2003.
18. Klon, A.E., Glick, M., Thoma, M., Acklin, P. and Davies, J.W., J. Med. Chem., 47 (2004) 2743.
19. Kellenberger, E., Rodrigo, J., Muller, P. and Rognan, D., Proteins, 57 (2004) 225.
20. Jacobsson, M., Liden, P., Stjernschantz, E., Bostrom, H. and Norinder, U., J. Med. Chem., 46 (2003) 5781.
21. Klon, A.E., Glick, M. and Davies, J.W, J. Chem. Inf. Comput. Sci., 44 (2004) 2216.
22. Klon, A.E., Glick, M. and Davies, J.W., J. Med. Chem., 47 (2004) 4356.
23. Bender, A., Mussa, H.Y., Glen, R.C. and Reiling, S., J. Chem. Inf. Comput. Sci., 44 (2004) 170.
24. Schapira, M., Abagyan, R. and Totrov, M., J. Med. Chem., 46 (2003) 3045.
25. Abagyan, R. and Orry, A. ICM User's Guide. MolSoft, L.L.C, La Jolla, 2004.
26. http://www.rcsb.org/pdb/.
27. http :// dtp.nci.nih.gov/docs/3d_database/structural_information/structural_data.html.
28. Pearlman, R.S. and Kubinyi H. (Eds.), 3D Molecular Structures: Generation and Use in 3D-Searching, ESCOM Science Publishers Leiden, 1993, p. 21.
29. Chang, Y.T., Gray, N.S., Rosania, G.R., Sutherlin, D.P., Kwon, S., Norman, T.C., Sarohia, R., Leost, M., Meijer, L. and Schultz, P.G., Chem. Biol., 6 (1999) 361.
30. Available Chemicals Directory, Elsevier MDL, San Leandro, 2004.
31. Hann, M., Hudson, B., Lewell, X., Lifely, R., Miller, L. and Ramsden, N., J. Chem. Inf. Comput. Sci., 39 (1999) 897.
32. Butina, D., J. Chem. Inf. Comput. Sci., 39 (1999) 747.
33. Li, D., MapMaker: an integrated compound library design tool, Philadelphia, 2004, August 22–26.
34. Blair, R.M., Fang, H., Branham, W.S., Hass, B.S., Dial, S.L., Moland, C.L., Tong, W., Shi, L., Perkins, R. and Sheehan, D.M., Toxicol. Sci., 54 (2000) 138.
35. Fang, H., Tong, W., Shi, L.M., Blair, R., Perkins, R., Branham, W., Hass, B.S., Xie, Q., Dial, S.L., Moland, C.L. and Sheehan, D.M., Chem. Res. Toxicol., 14 (2001) 280.
36. Rogers, D., Multicriteria Modeling: The Next Stage in Handling Large Data Sets, Anaheim, 2004, March 27–April 1.
37. MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577.

38. Ghose, A.K. and Crippen, G.M., J Comput. Chem., 7 (1986) 565.
39. Ghose, A.K., Pritchett, A. and Crippen, G.M., J. Chem. Inf. Comput. Sci., 9 (1988) 80.
40. *Pipeline Pilot V 3.5. User Manual; section "Extended Connectivity Fingerprints"*, Scitegic Inc., San Diego, 2004.
41. *Daylight Theory User Manual; section "Fingerprints - Screening and Similarity"*, Daylight Chemical Information Systems, Inc., Mission Viejo, 2004.
42. Bayes, T., Biometrika, 45 (1958) 296.
43. Xu, H. and Agrafiotis, D.K., J. Chem. Inf. Comput. Sci., 43 (2003) 1933.
44. Dongarra, J.J., LINPACK, http://www.netlib.org/linpack/, (1988) .
45. Hand, D., Mannila, H. and Smyth, P. Principles of Data Mining. The MIT Press, Cambridge, Massachsetts, 2001.
46. Pearlman, D.A. and Charifson, P.S., J. Med. Chem., 44 (2001) 502.
47. De Borda, J. Memoire sur les elections au scrutin. historie de l'academie royale des sciences, Paris, 1781 .
48. Breiman, Freidman, Olshen and Stone, 1984. Classification and Regression Trees, Wadsworth.
49. Wold, H. and Gani, J. (Ed.), The PLS Approach, in Perspectives in Probability and Statistics, Academic Press London, 1975.
50. Aleksander, I. and Morton, H., 1995. An introduction to Neural Computing, Chapman and Hall.
51. Back, T. Evolutionary Algorithms in Theory and Practice – Evolution Strategies, Evolutionary Programming, Genetic Algorithms. Oxford University Press, New York, Oxford, 1996.