Flexsim-X: A Method for the Detection of Molecules with Similar Biological Activity

Uta F. Lessel* and Hans Briem

Department of Lead Discovery, Boehringer Ingelheim Pharma KG, Binger Straβe 173, D-55216 Ingelheim, Germany

Received July 26, 1999

We describe the development of the method Flexsim-X, which can be used to detect molecules with similar biological activity. This procedure is based on comparing virtual affinity fingerprints made up from docking scores of the molecules with respect to a reference set of binding sites. Using a test data set consisting of ligands from five different activity classes and randomly chosen compounds, the reference panel of binding sites was optimized in terms of size and composition. Systematic approaches as well as genetic algorithm based (GA) optimization procedures have been evaluated. Additionally, the effectiveness of the method is illustrated.

INTRODUCTION

Measuring the similarity or diversity of molecules is an important task in pharmaceutical research. An appropriate diversity metric can, e.g., be used for data mining, i.e., searching databases for compounds similar to a given lead structure. Nearest neighbor lists obtained by similarity calculations are also required for clustering structures in chemical databases, for library design in high throughput organic synthesis, and for decisions in the acquisition process of external compounds in order to augment in-house compound pools.

Since results of similarity searches depend extremely on the criteria or molecular descriptors used, the important question is "What is similar?". In drug discovery one is mostly interested in finding molecules with similar biological activity. For this purpose numerous methods have been proposed. As similarity descriptors most of these approaches use properties which are directly derived from molecular structures. Although some of these methods have been proven to be very successful in classifying compounds which share the same biological activity,[1] most of them fail in finding hits belonging to structural classes other than the original query molecule.

To address this problem we developed the program Flexsim-X, which is based on the "affinity fingerprints" idea published by Kauvar et al.[2] They determined in vitro binding affinities to a reference set of receptors for comparing molecules. In Flexsim-X, however, virtual affinity fingerprints are computed by flexibly docking the ligands into a reference panel of proteins from the Brookhaven Protein Data Bank[3] using the program FlexX.[4] Flexsim-X is a further development of our earlier approach (called DOCKSIM in the following text), which used a rigid docking scheme instead.[5] A similar approach has recently been published by Ghuloum et al.[6] They used molecular hashkeys which represent molecular surface properties as a linear array of pairwise surface-based comparisons of target molecules with a common "basis set" of molecules. In their paper they also present results for variations of the hashkey size.

Here, we describe the efficiency of the Flexsim-X procedure for data mining. In addition, systematic and GA-based methods for the optimization of the binding site reference panel in terms of size and composition are introduced.

METHODS

Flexsim-X. To compare a small molecule with the compounds in a database and to determine the most similar structures (hereafter *nearest neighbors*), the query structure and each member of the database are flexibly docked with the program FlexX 1.65[4] into a reference set of binding sites. The docking results and scores depend on a number of parameters which can be set. Here, for all calculations we used the standard parameters which have been set by the developers of FlexX. None of these parameters was changed. The best solution of each docking process—that with the highest interaction energy—is selected. Finally, for each compound the docking scores for the reference binding sites are organized as linear vectors making up the virtual affinity fingerprints.

To compare two molecules A and B, the Euclidean distance between their affinity fingerprints consisting of docking scores for n binding sites is determined according to

$$\text{Euclidean distance(A,B)} = \sqrt{\frac{\sum_{i=1}^{n}(\text{FlexX score(A},i) - \text{FlexX score(B},i))^2}{n}} \quad (1)$$

If one or both of the ligands cannot be docked successfully into one or more of the reference binding sites, the score differences for these binding sites are arbitrarily set to 30 kJ/mol.

* To whom correspondence should be addressed. E-mail: lessel@ing.boehringer-ingelheim.com.

10.1021/ci990439e CCC: $19.00
Published on Web 02/05/2000

A *nearest neighbor list* for a query structure is generated by calculating the Euclidean distances of its affinity fingerprint to those of each database member and sorting the compounds by increasing distances.

The composition of a nearest neighbor list for a single query structure depends on the docking protocol used, because different docking scores give different affinity fingerprints and consequently diverging similarities. In our experience however, for a large data set, slight differences in the docking protocols are averaged out. What is more important is to use the same docking parameters for the whole set.

Optimization Process. One aim of our work was to determine a reference panel of binding sites for Flexsim-X which groups compounds best according to their biological activities. For this purpose we used a sample of protein binding sites, a ligand test set to assess the different panels, a success measure, and different optimization procedures which are described in the following section.

Protein Binding Sites. All protein binding sites considered as members of the reference panel are chosen from the Brookhaven Protein Data Bank (PDB). As a first choice those eight binding sites used for our original DOCKSIM approach,[5] generated by the SPHGEN module of DOCK (details of the DOCK algorithm are described in numerous papers, e.g., refs 7−9) were taken into account. All additional binding sites considered for the reference panel pool were taken from a set of binding sites created for the validation of FlexX.[10] None of the binding sites used is relevant to any of the five activity classes in the ligand test set (see below). At the final stage, a *reference panel pool* of 100 proteins was considered. A list of the respective PDB codes is included in the Supporting Information.

Ligand Test Set. All optimization procedures were performed using the same ligand test set. Its structures were extracted from the MACCS Drug Data Report (MDDR), provided by MDL Information Systems Inc. (San Leandro, CA). The test set contains 957 ligands and is a slightly modified version of the ligand set already used for the DOCKSIM development. It contains 134 PAF antagonists, 49 5-HT_3 antagonists, 49 TXA_2 antagonists, 40 ACE inhibitors, and 111 HMG-CoA reductase inhibitors. Additionally, 574 compounds not belonging to any of these five activity classes were randomly selected from the MDDR database. For creating the ligand data set, two-dimensional structures were extracted from MDDR. Next, three-dimensional structures were generated utilizing the CORINA program (version 2.12).[11] Hydrogens and partial charges, according to the Gasteiger−Marsili method,[12] were added by the DOCK utilities "sdf2mol2" and "sybdb".[13] A list of the MDDR ID codes is available in the Supporting Information.

Success Measure. The aim of the optimization process is to find a reference set of binding sites giving virtual affinity fingerprints which are appropriate to classify the compounds of the ligand test set best according to their biological activities. Consequently, the success of Flexsim-X is measured by determining the enrichment of molecules from the same biological activity class in the nearest neighbor lists of the ligand test set members. For this purpose the following procedure was used:

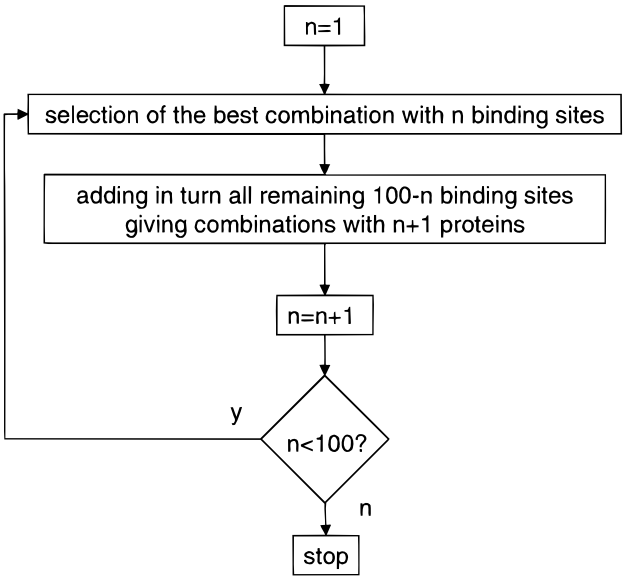

Figure 1. Flow chart showing the accumulation method.

Each of the 383 ligands from the five activity classes is taken in turn as a search query and is compared with the other 956 members of the test set on the basis of their Flexsim-X affinity fingerprints. In the next step, the number of molecules from the same activity class within the 10 nearest neighbors of the query molecules and the corresponding hit rates are determined (considering 10 nearest neighbors corresponds to a selection of about 1% from the whole database consisting of nearly 1000 compounds). Finally, the mean hit rate in the nearest neighbor lists of all 383 ligands characterizes the success of the similarity search. This *mean sample hit rate* can directly be used to compare the classification results of reference panels consisting of different protein binding sites.

Systematic Optimization Procedures. The basic idea behind these methods is to systematically build up reference panels with good classification potential by adding the best binding sites or eliminating the worst in turn.

Accumulation Method (see Figure 1). In a first step all binding sites are considered in turn to be the only member of the reference panel. The classification results for the ligand test set are determined for all these panels with only one protein. That panel yielding the highest mean sample hit rate is selected. Then each of the other 99 binding sites is added in turn to the selected protein and each of the resulting reference panels with two binding sites is used to classify the ligand test set. The best pair of binding sites in terms of mean sample hit rates is selected and is subsequently taken to form reference panels with three binding sites by the same procedure. The whole process is continued until all 100 binding sites of the pool are included in the reference panel.

Reduction Method. This method illustrated in Figure 2 is an inversion of the accumulation method.

In the first step the mean sample hit rates for all possible reference panels with 99 different binding sites are determined. The best combination is selected and the protein missing in this reference panel is excluded from all the following steps. (This is due to the fact that eliminating the binding site which is missing in the best combination of 99 proteins yields the highest gain or lowest damage to the mean sample hit rate.)

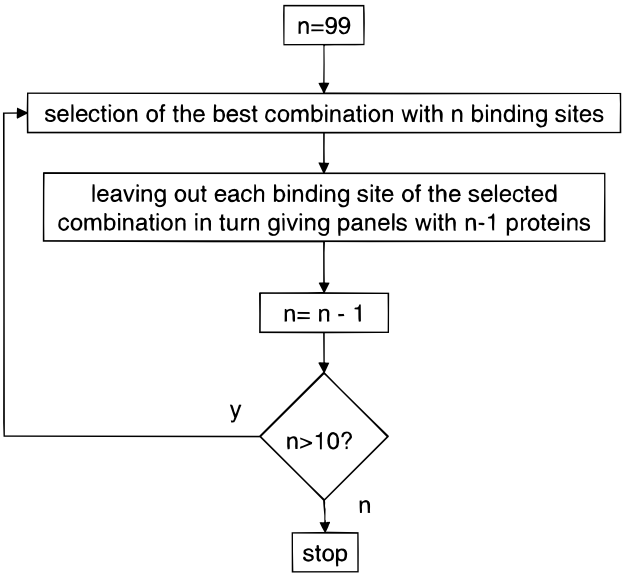

Figure 2. Flow chart showing the reduction method.

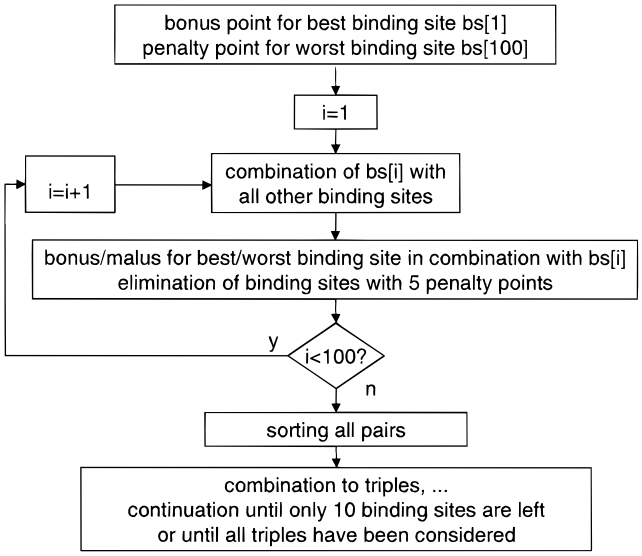

Figure 3. Flow chart showing the bonus−malus method.

Then all possible combinations of 98 out of the remaining 99 binding sites are used to reclassify the ligand test set. Again the best reference panel from this cycle is selected and the protein missing in this combination is eliminated. This procedure is continued until only 10 binding sites are left.

Bonus−Malus Method. Figure 3 shows a flow chart for this procedure.

In a first step each binding site is considered in turn to be the only member of the reference panel. These one-member reference panels are used to classify the ligand test set and are sorted by decreasing mean sample hit rates. The best binding site gets a bonus and the worst a penalty point. Next, the best binding site is combined in turn with all others. The new reference panels with two binding sites are used to reclassify the ligand test set and are assessed by the resulting mean sample hit rates. That protein which is the new partner in the best pair gets a bonus point and the binding site which adds to the worst pair gets a penalty point. In the following step the second best binding site is combined to pairs. Again the resulting reference panels with two binding sites are evaluated according to the mean sample hit rates. Bonus as well as penalty points are given to the best and worst partners. This procedure is continued for all other binding sites. After each step binding sites with at least five penalty points are eliminated. When all possible pairs have been built up, they are sorted according to decreasing mean sample hit rates. Analogous to the formation of pairs, now, triples are created and bonus and penalty points are given. The procedure is continued until only 10 binding sites are left or until all triples have been considered. In the latter case the remaining binding sites are ranked by their numbers of penalty points.

The limit of five penalty points was chosen empirically. Results are similar, if e.g. up to 10 penalty points are allowed.

Malus Method. This method shown in Figure 4 starts with classifying the ligand test set using all possible reference panels with two binding sites and sorting them according to increasing mean sample hit rates.

The worst pair of binding sites is considered first. The mean sample hit rates for both combinations with 99 proteins missing each of the members of the worst pair are determined. Additionally, the reference panel of 98 binding sites missing both members of the worst pair is used to classify the ligand test set. If the reference panel with 98 proteins yields the highest mean sample rate, both omitted binding sites get a penalty point, otherwise only that binding site which is missing in the best of the two 99 protein combinations. Each binding site with at least five penalty points is excluded from the following steps. The procedure is stopped if all pairs have been taken into account or if only 10 binding sites are left. Again, the limit of five penalty points was chosen empirically.

GA-Based Optimization. For the global optimization of the reference panel a string of 100 bits was used. If a bit is set, the corresponding binding site is included in the reference panel. To find the best combination of x binding sites, samples of x integer values ranging from 1 to 100 were optimized. Figure 5 gives an overview of the genetic algorithm applied. For this procedure the genetic algorithm package SUGAL 2.1[14] has been used.

Generally, in genetic algorithms, optimization procedures observed in evolutionary processes in nature are applied. At the beginning, a population of 200 different reference panels (chromosomes) is created (init). Each of these binding site combinations is assessed by the corresponding mean sample hit rate (fitness), which is achieved for the classification of the ligand test set. Next, members of the population are selected (selection) to produce children with the aid of mutation and crossover. Mutation means inverting single bits in the global optimization process or exchanging single binding sites in the reference panels while searching for the best combination of x binding sites. In crossover procedures parts of the chromosomes of two parents are exchanged.

In the following step, a new generation of 200 reference panels is built up from the 200 parents and the 200 children (replacement) taking into account the fitness of the parents and the children. The whole cycle is controlled by different parameters. To explore the reference panel space as broadly as possible, we generally used the highest possible mutation rate of 1.0 and varied the selection and replacement parameters as well as the crossover rate in several different trials.

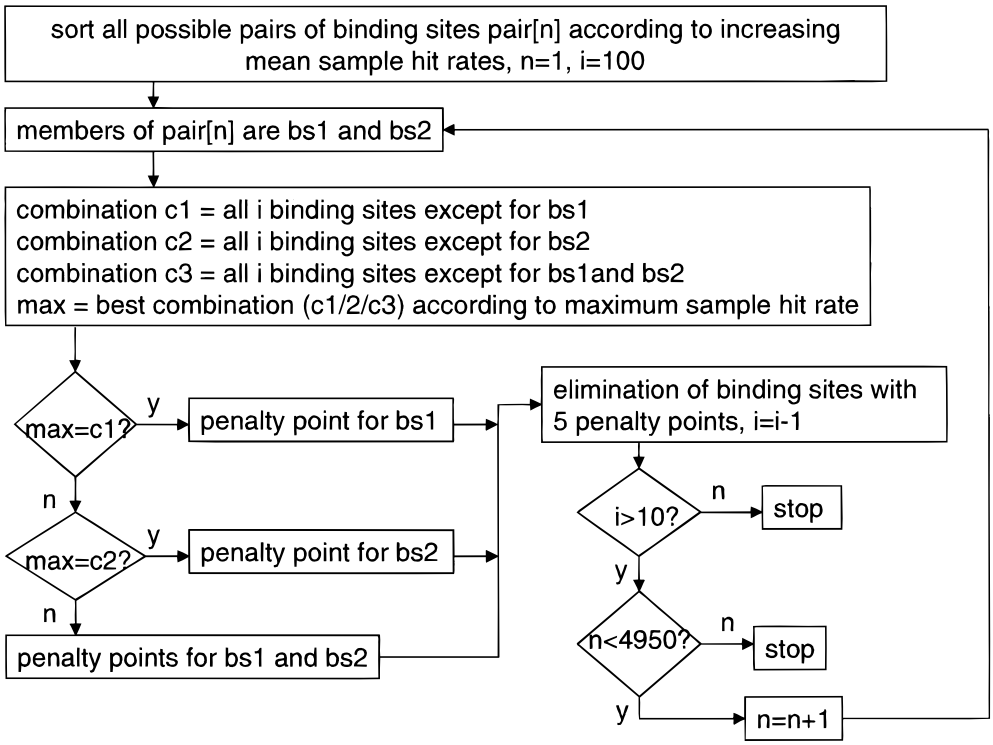

Figure 4. Flow chart showing the malus method.

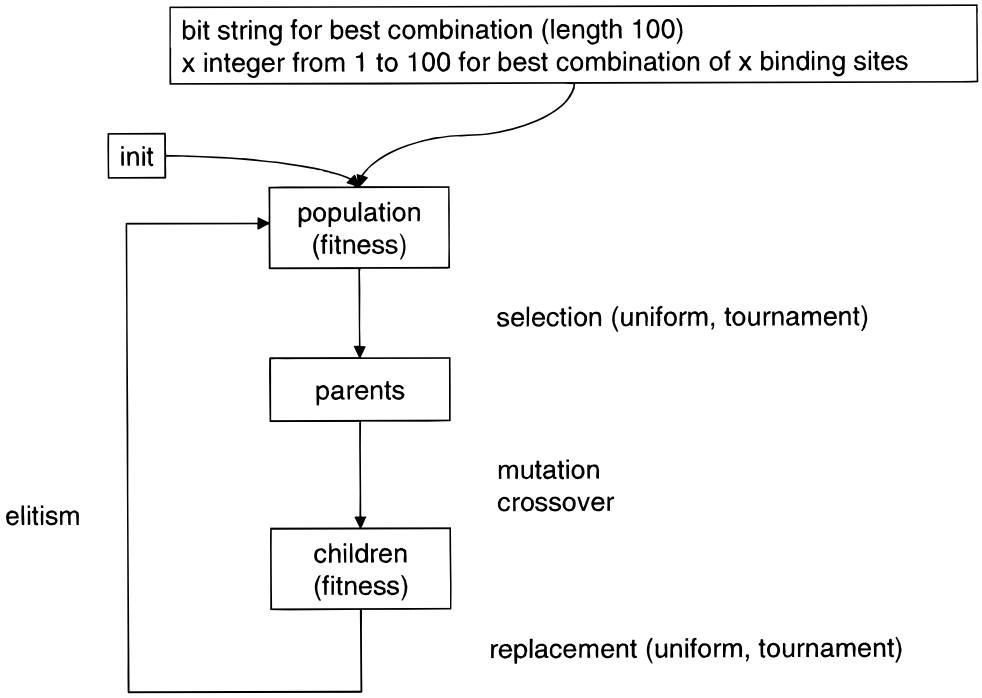

Figure 5. Flow chart showing the general course of the genetic algorithm.

For the selection process we applied two different modes: uniform and tournament. In a uniform selection, parents are randomly chosen which means for a replacement rate of 100% that each member of the population will produce exactly one child. During tournament selection, however, randomly chosen pairs of reference panels compete according to their fitness results for the chance to produce a child. The crossover rate was set to either 0.0 or 1.0.

Finally, uniform and tournament replacement were used. In uniform replacement the parents to be replaced by their children are randomly chosen, whereas in tournament replacement the worst of a randomly chosen pair of parents with respect to their fitness results is replaced by its child. In any case the best reference panel was taken over into the next generation (elitism).

Applying these parameters, we did not aim at and did not expect convergence of the whole population, but rather wanted to explore the huge reference panel space as broadly as possible. Therefore the process was controlled by the maximum number of cycles set. Generally, we created 200−1000 generations. If the maximum fitness did not improve during 100 cycles, the run was stopped.

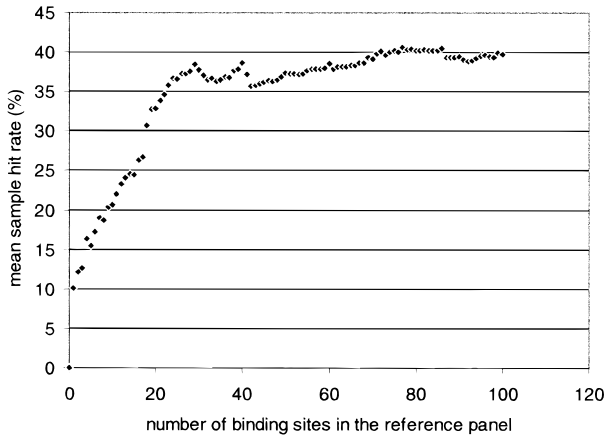

Figure 6. Mean sample hit rates (%) determined for stepwise random addition of binding sites to the reference panel.

Table 1: Mean Sample Hit Rates (%) for Randomly Chosen Combinations of *n* Binding Sites (25 000 Samples Each)

n	min	max	av	std dev
10	19.0	36.6	28.0	2.4
20	25.5	40.6	33.2	2.1
30	28.1	42.6	35.6	1.9
40	29.8	43.1	36.9	1.8
50	32.1	43.3	37.8	1.6
60	33.8	43.7	38.4	1.5
70	34.7	44.6	38.8	1.3
80	35.4	42.8	39.2	1.1
90	37.3	42.8	39.5	0.8

RESULTS AND DISCUSSION

To illustrate the efficiency of Flexsim-X, first of all, mean sample hit rates were calculated for the ligand test set using scrambled nearest neighbor lists or scrambled FlexX scores. In all of these trials mean sample hit rates of about 9% were achieved.

Next, Flexsim-X was used to determine mean sample hit rates for the ligand test set. For this purpose we started with one protein chosen arbitrarily from the reference panel pool and calculated the mean sample hit rate. Then the number of binding sites considered was systematically increased by randomly selected proteins. Figure 6 shows the mean sample hit rates achieved for up to 100 binding sites.

These results demonstrate that the random mean sample hit rates of 9% are already exceeded when two binding sites are included in the reference panel. Generally, the mean sample hit rates increase with rising number of binding sites, but the slope of the curve decreases. Moreover, it can be recognized that certain combinations of proteins give clearly higher mean sample hit rates than other sets with more binding sites. To analyze this phenomenon more precisely, in a second step, mean sample hit rates were determined for reference panels consisting of randomly selected combinations of 10−90 binding sites. The results are summarized in Table 1.

All mean sample hit rates determined show a normal distribution. In accordance with the results shown in Figure 6, the mean values increase using from 10 to 90 binding sites in the reference panel. For example, there is a very high probability to achieve about 39.5% mean sample hit rate using any combination of 90 binding sites. On the other hand, 39.5% mean sample hit rate can also be yielded with just 20 binding sites, but an appropriate combination has to be selected. For 10 binding sites in the reference panel, the difference between the minimum and maximum mean sample hit rates is large. The average mean sample hit rate shows a relatively low value with a high standard deviation. With growing panel sizes the minimum mean sample hit rates increase, whereas the differences between the minimum and maximum as well as the standard deviations decrease. In other words, the mean sample hit rates achieved with small reference panel sizes depend extremely on the binding sites selected. Therefore an optimal selection of reference panel proteins is required.

On the one hand, we were looking for a globally optimal reference panel in order to get an impression of which mean sample hit rates can be achieved by Flexsim-X. On the other hand, for a general application of the method, it has to be considered that docking large molecular databases can be very time-consuming. Therefore the reference panel should be limited to a manageable number of binding sites. Taking into account this tradeoff and the size of our in-house compound pool, we decided to search for an optimal combination of 10 binding sites additionally.

The safest way to find the optimal reference panels would be an exhaustive exploration of all possible binding site combinations, but unfortunately this is computationally not feasible. Taking into account 3 CPU seconds per combination, the calculation of mean sample hit rates for all possible 1.3×10^{30} reference panel compositions would take about 10^{23} years of CPU time. Even all 1.7×10^{13} combinations of 10 binding sites would take 1.6 million years of CPU time. Thus, other optimization procedures have to be applied.

The basic idea behind the systematic optimization procedures is to accumulate highly competitive proteins in the reference panel by adding the best binding sites or eliminating the worst.

Thus, the accumulation method starts with the best binding site. This is combined with all other proteins to find the best pair, which is used to form the best triple and so on. In this way the "best combination of 10 binding sites" is detected. A further increase of the reference panel up to all 100 binding sites passes the "overall best combination". The reduction method, vice versa, starts with all 100 binding sites and systematically eliminates the worst proteins. In the course of this procedure the "best combination" that gives the maximum mean sample hit rate is passed and the method stops at the "best combination of 10 binding sites". The bonus−malus and the malus methods intend to eliminate binding sites only if they have performed poorly several times.

In addition, we looked for an efficient reference panel with genetic algorithm procedures which use optimization mechanisms observed in evolutionary processes in nature. In this way best overall combinations and best reference panels of 10 proteins were determined for different parameter sets (selection, replacement, crossover rate).

The results of all systematic optimization procedures and the different GA trials are shown in Table 2.

The best mean sample hit rates achieved by both the bonus−malus and the malus methods are worse than those obtained by the other procedures, which all give similar maximum mean sample hit rates of about 50% for the overall

Table 2: Mean Sample Hit Rates (%) Achieved with Different Reference Panel Optimization Procedures[a]

(a) Systematic Optimization Procedures

method	best 10 out of 100	overall best out of 100
accumulation	39.2	*49.2 (51)*
reduction	*40.1*	48.3 (48)
bonus−malus	37.1	44.5 (57)
malus	34.9	45.5 (64)

(b) GA-Based Optimization Procedures

selection	replacement	crossover rate	best 10 out of 100	overall best out of 100
uniform	tournament	0.0	*41.0*	49.9 (41)
uniform	tournament	1.0	40.8	49.4 (48)
tournament	tournament	0.0	40.3	49.2 (48)
tournament	tournament	1.0	40.3	49.3 (54)
tournament	uniform	0.0	40.8	*50.3 (41)*
tournament	uniform	1.0	40.3	50.2 (47)

[a] Number of binding sites for overall best in parentheses; best results achieved in italics.

best and about 40% for the best combination of 10 binding sites, respectively.

At this point it is important to know whether we have found the global optimum, or whether there are other combinations giving even higher mean sample hit rates. To answer this question, the optimization procedures have to be put to the test. As mentioned before, an exhaustive exploration of all possible protein combinations is computationally not feasible. Therefore we chose the following procedure to support our results.

Exemplarily we selected the GA procedure and repeated the reference panel optimizations for smaller panel pool sizes. The mean sample hit rates for the best 10 and the overall best 14 out of the first 18 binding sites of the whole pool were confirmed by exhaustive calculations of all possible combinations.

Furthermore, the maximum mean sample hit rates for the overall best combinations and the best combinations of 10 binding sites out of the first 28, 36, 50, 66, 76, and 90 proteins have been determined by several GA trials. The best results of these optimizations (maximum mean sample hit rates for overall best and best 10 out of *x*) are shown in Figure 7.

Analogous to the curve showing the mean sample hit rates achieved for using all binding sites in the reference panel pool for different reference panel pool sizes, the success rates for the optimized reference panels (overall best and best with 10 binding sites) follow an asymptotic course. These curves suggest that for each panel pool size the global optimum has at least almost been achieved, which shows the feasibility of the optimization procedures. Moreover, a further addition of binding sites to the reference panel pool may have only little effect on the mean sample hit rates. Finally, it is remarkable that the results for the best combinations of 10 binding sites are more or less equal to those achieved with all proteins from the pool.

Taken together, these results show that nearest neighbor lists determined by Flexsim-X give clearly increased hit rates with respect to random samples and an optimization of the reference panel yields even more enriched hit lists.

To justify the limitation of the reference panel to 10 binding sites, we chose the GA optimization procedure to

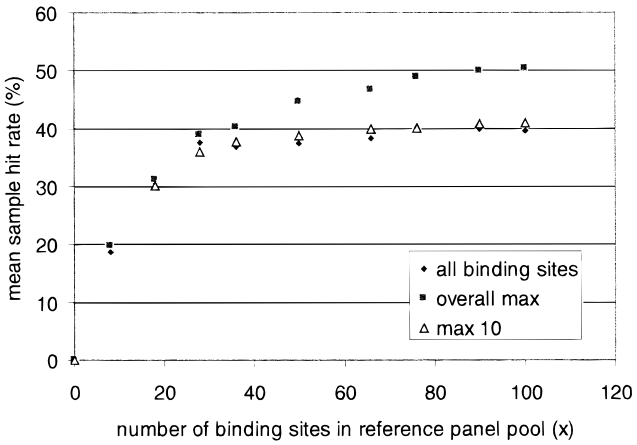

Figure 7. GA optimization results for different reference panel pool sizes (mean sample hit rates achieved by all *x*, the overall best combination out of *x*, and the best 10 out of *x* binding sites in the reference panel pool).

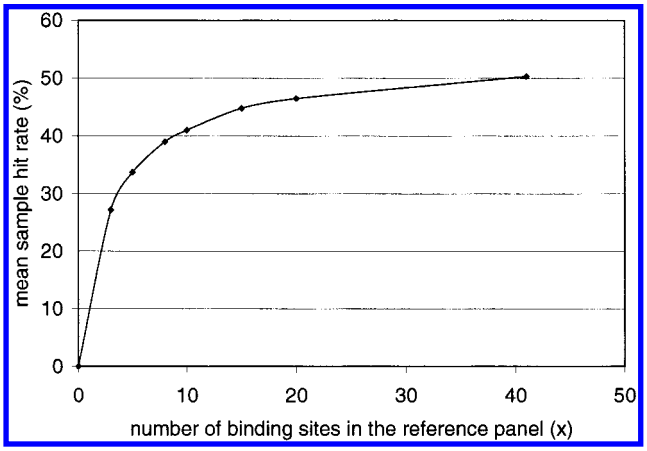

Figure 8. Mean sample hit rates (%) for the best *x* out of 100 binding sites in the reference panel pool determined by GA optimization.

determine the best combinations of 3, 5, 8, 15, and 20 binding sites. The results are shown in Figure 8.

The curve shows a decreasing slope up to the maximum of 50.3% mean sample hit rate with 41 binding sites. Especially for small reference panels the mean sample hit rates increase with each additional binding site. The limitation to 10 binding sites represents a compromise on classification quality and computer time needed for the docking process. This optimized panel consists of a diverse subset of protein binding sites with completely different functions.

Moreover, the method is capable of finding molecules with similar biological activity from different structural classes. This is illustrated in Figures 9 and 10, which show different PAF and $5HT_3$ antagonists detected by Flexsim-X. Whereas molecules with the same basic scaffold as the query can also easily be found using for example two-dimensional fingerprints, the other nearest neighbors shown are more difficult to detect. All examples are found among the 10 nearest neighbors of the respective queries.

CONCLUSIONS

We clearly demonstrated that Flexsim-X can be used to detect molecules with similar biological activities from different structural classes without prior knowledge of the target protein structures.

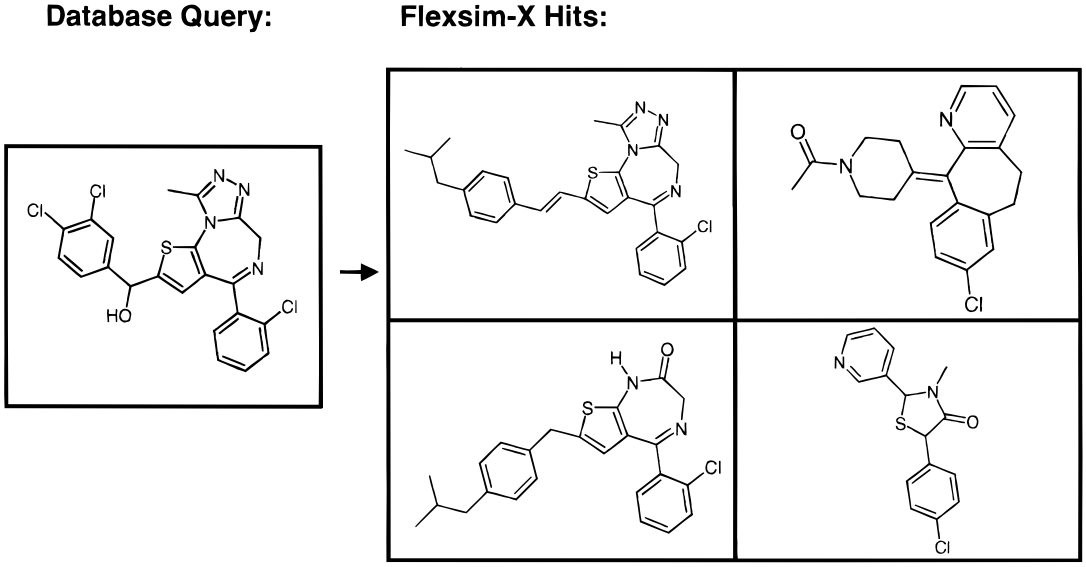

Figure 9. Example for PAF antagonists found by Flexsim-X.

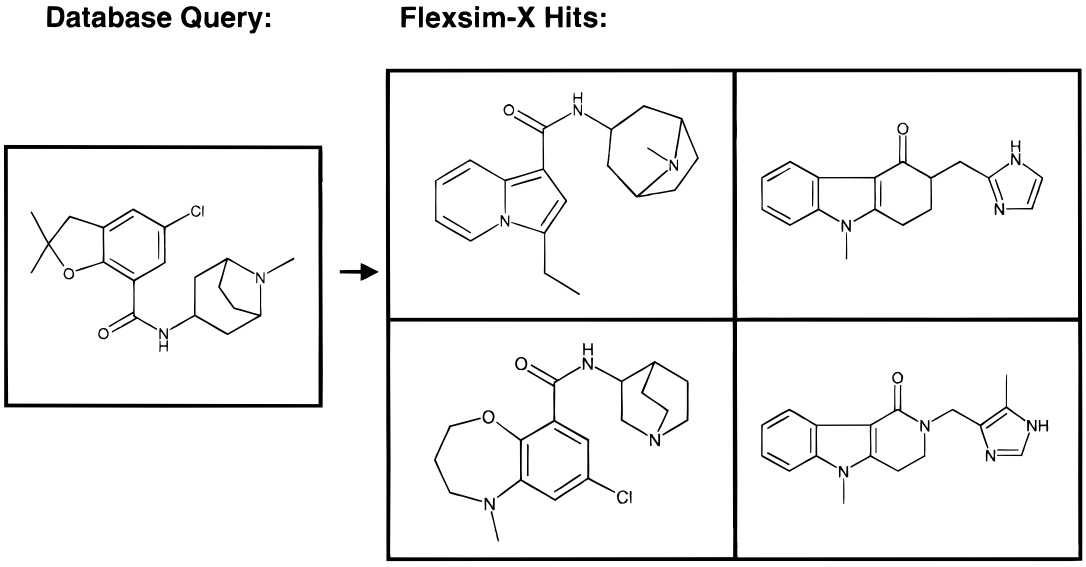

Figure 10. Example for $5HT_3$ antagonists found by Flexsim-X.

Both systematic as well as GA-based procedures for the optimization of the binding site reference panel in terms of size and composition were developed, as a full combinatorial exploration of all possible reference panel compositions is computationally not feasible. Almost all optimization procedures yield similar maxima.

Variations in the panel pool size show that the optimization procedures work and suggest that further addition of binding sites may have only little effect on the mean sample hit rates achieved for the classification of the ligand test set.

The optimization procedures have yielded mean sample hit rates of about 50%. Additionally, about 41% mean sample hit rate can be achieved with a reference panel consisting of only 10 highly competitive binding sites. This promises higher enrichment of molecules with similar biological activity during data mining processes.

ACKNOWLEDGMENT

This work was performed as part of the RELIMO project, which is funded by the German Federal Ministry for Education, Science, Research, and Technology (BMBF) under Grant No. 0311623.

Supporting Information Available: List of proteins considered during the reference panel optimization and list of MDDR ID codes used in the ligand test set. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Brown, R. D.; Martin, Y. C. Use of Structure−Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(2) Kauvar, L. M.; Higgins, D. L.; Villar, H. O.; Sportsman, J. R.; Engqvist-Goldstein, A.; Bukar, R.; Bauer, K. E.; Dilley, H.; Rocke, D. M. Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.* **1995**, *2*, 107−118.

(3) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures. *J. Mol. Biol.* **1977**, *112*, 535−542.

(4) 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.

(5) Briem, H.; Kuntz, I. D. Molecular Similarity Based on DOCK-Generated Fingerprints. *J. Med. Chem.* **1996**, *39*, 3401−3408.
(6) Ghuloum, A. M.; Sage, C. R.; Jain, A. N. Molecular Hashkeys: A Novel Method for Molecular Characterization and Its Application for Predicting Important Pharmaceutical Properties of Molecules. *J. Med. Chem.* **1999**, *42*, 1739−1748.
(7) 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.
(8) Shoichet, B. K.; Bodian, D. L.; Kuntz, I. D. Molecular Docking Using Shape Descriptors. *J. Comput. Chem.* **1992**, *13*, 380−397.
(9) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. Automated Docking with Grid-Based Energy Evaluation. *J. Comput. Chem.* **1992**, *13*, 505−524.
(10) Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FlexX Incremental Construction Algorithm for Protein−Ligand Docking. *PROTEINS: Struct. Funct. Genet.* **1999**, *37*, 228−241.
(11) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to Three-dimensional Atomic Coordinates: Automatic Model Builders. *Chem. Rev.* **1993**, *93*, 2567−2581.
(12) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity: a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3228.
(13) Ewing, T. *DOCK Manual*, version 4.0; University of California at San Francisco, 1997.
(14) Hunter, A. *SUGAL Genetic Algorithm package*, version 2.1; University of Sunderland, England.

CI990439E