

The following text is part of the book chapter by M. Stahl, M. Rarey, and G. Klebe, in *Bioinformatics: From Genomes to Drugs*, T. Lengauer Ed., VCH, Weinheim, 2001, pp. 229.
Screening of drug databases

Please note that the “Correlation Vectors” (CV) similarity measure referred to in this text corresponds to the CATS descriptor.

Ligand Similarity-based virtual screening

The experimental determination of protein structures is a difficult task and is not routinely possible. Especially membrane-bound proteins are difficult to crystallize [16]. Therefore, the situation that the target protein structure is unknown is common at an early stage of a pharmaceutical project. In such cases, only the structure of molecules already known to bind to the target protein, for example an endogenous ligand, can be used for virtual screening.

When a molecule binds to the protein, it adopts a conformation that sterically and electrostatically fits into the active site. In order to detect compounds possibly binding into the same active site, one can search for molecules that can adopt conformations that are similar to the given active molecule. This similarity is defined with respect to size, shape and spatial distribution of functional groups and can be regarded as “chemical similarity” [17-19]. In principle, however, we are interested in describing “biological similarity” in the sense that molecules exhibiting the same biological response will be considered as similar. Biological similarity, however, depends on many unknown factors such as details of the protein structure and, therefore, cannot be computed. Chemical similarity can be regarded as a necessary but insufficient criterion for similar biological response of compounds. Fortunately, searching for chemically similar compounds has been demonstrated to be highly successful in many applications [20-22].

Methods for similarity-based virtual screening can be classified by the kind of information used to define similarity. While topological descriptors are based on the 2D structure (connectivity, bond orders and atom types) only, three-dimensional descriptors take into account the spatial arrangement of atoms or their associated properties with respect to each other.

Topological (2D) descriptors

The 2D structure of a molecule can be used to compute bit strings describing the topology of a molecule in a piecewise manner. Originally introduced to solve the problem of fast substructure searching [19], today, bit strings are the most widely applied descriptors for fast similarity searching. Two approaches to create bit strings can be distinguished: structural keys and fingerprints. In a structural key, each bit represents a molecular fragment that is set to 1 if the fragment is present in a molecule [23]. Although this method performs very well in practice, it has the limitation that a reasonable set of fragments has to be predefined. In order to avoid this lack of generality, molecular fingerprints were introduced [24, 25]. The patterns for a molecule's fingerprint are generated from the molecule itself. Each bit represents a sequence of connected atoms having specified atom types, called a *path*. The fingerprinting algorithm generates a list of all paths up to a specified length. For each path, a hash function calculates the location of the bit that has to be set. This means that a particular bit represents not just one specific bond path, but several paths. The length of the fingerprint can be reduced by a process called folding. The shorter the length of the fingerprint, the more different paths are represented by the same bit and the less specific information is encoded. Examples for programs based on bit string descriptors are structural keys by MDL [26], Daylight fingerprints [25], and UNITY 2D screens by Tripos [27]. The latter two are used in the application study presented in this paper. Therefore they will be described in more detail.

Daylight fingerprints (DF) [25] are bit strings generated from bond paths of length zero to seven. The length of the fingerprints can be folded until a specified density of set bits is reached. Alternatively, the fingerprint length can be set to a fixed value. The similarity index used is the Tanimoto coefficient, which is the number of bit positions set to 1 in both strings divided by the number of bit positions set to 1 in at least one of the strings. If a set bit is considered as a feature present in the molecule, the Tanimoto coefficient is a measure of the number of common features in two molecules [28, 29].

UNITY 2D screens (UF) [27] are a mixture of fingerprints and structural keys. Unity allows the user to specify the individual components of the bit string, which can be paths within a specific length range or a variety of substructures. For this study, the default 988 bit strings were used. About 94% of the bits in the string are reserved for encoding paths of lengths 2 to 6 using atom type and bond type information. Another 3% of the bits in the string are reserved for atom counts of frequently occurring hetero atom types and the remaining 3% are used for important

fragments such as benzene rings and heterocycles. The Tanimoto coefficient is used as a similarity index.

Several other descriptors based on the topology of molecules were developed and have been successfully applied. Reviews on descriptor technology can be found in [19, 20, 22, 30]. The following discussion focuses on three recently developed descriptors used in the application study in section 3. The selection of descriptors was based on availability and speed of the software.

Correlation vectors (CV) [31] are created by first assigning a generalized atom type to each atom. In the implementation applied here [32], five atom types were used: hydrogen-bond acceptor (A), hydrogen-bond donor (D), positively charged (P), negatively charged (N) and lipophilic (L). In a second step, each pair of atoms in the molecule is classified by their assigned generalized atom types and the number of bonds by which they are separated. The correlation vector, a string of integers, is a histogram representation of how often each possible pair of atom types is found at each distance. For bond paths of up to n bonds this results in correlation vectors of n times 15 integers. In order to achieve a size-independent description, each integer is divided by the total number of non-hydrogen atoms in the molecule. The Euclidean distance between vectors is used as a similarity index. Values of $n=6$ through $n=10$ were used in this work and did not give significantly different results. Therefore, $n=6$ was chosen for comparison with other methods.

The methods described so far have in common that they create a one-dimensional representation of the molecule from its 2D structure, either a bit string or an integer vector. The advantage of this representation is that, in order to calculate a similarity index, there is no need to assign parts of two molecules or molecule representations to each other. This results in simple and extremely fast comparison algorithms. However, since molecules are three-dimensional objects, substantial information is neglected and, in consequence, some accuracy is sacrificed. The following two descriptors are not based on a one-dimensional representation and, therefore, need an assignment procedure for the similarity calculation. The first descriptor is based on a matrix representation, the second is based on a tree representation. The two methods also have in common that they read a single 3D conformation of each input molecule.

Topological pharmacophores (TP) [33] are based on data generated by the force field MAB [34, 35]. Hydrogen-bond strengths are assigned to atoms on a continuous scale. The atoms are then classified as hydrogen-bond donors, hydrogen-bond acceptors or hydrophobic. Hydrophobic

atoms are grouped together according to a fixed maximum inter-atom distance criterion. A molecule is described by a set of donor, acceptor and hydrophobic centers and a matrix containing the through-space distances between these centers. Topological pharmacophores thus depend on the input conformation to some extent. To arrive at a similarity index for two molecules, centers of the same type are matched onto each other, leading to several alternative pairings for all centers. For each of these pairings, a similarity score between zero and one is calculated based on hydrogen-bond strength and the differences in mutual distances of the matched centers. The final similarity index is the highest similarity score calculated for one of the pairings.

Feature trees (FT) [36] describe the overall topology of a molecule with a tree structure. They are independent from the input conformation of the molecule. Starting from the molecule graph, the corresponding feature tree is created by shrinking rings as well as terminal atoms with their neighbors to single nodes (see Figure 1 for an example). A steric and a chemical feature are assigned to each node, describing the part of the molecule represented by the node. The approximated van der Waals volume is used as a steric feature. The chemical feature is an interaction profile summarizing which interactions can be formed by the molecule part. In order to calculate a similarity index for two feature trees, nodes of the trees have to be assigned to each other. This is done by the so-called split-search algorithm, which divides and assigns subtrees in a recursive fashion. In this way, the algorithm calculates an assignment between nodes by optimizing the similarity index.

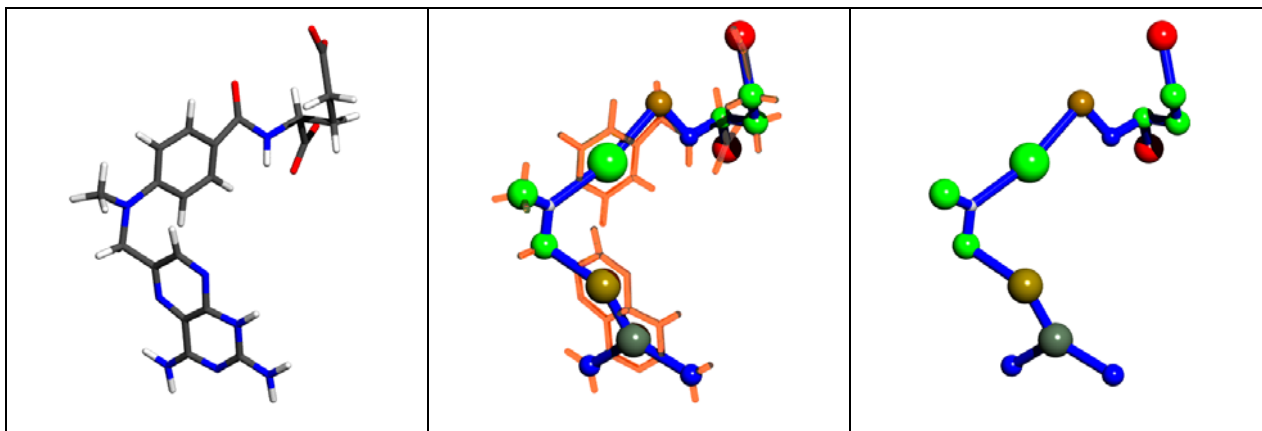


Figure 1. A molecule and its corresponding feature tree representation. Feature tree nodes are colored by the interaction profile. Red: hydrogen bond acceptor, blue: hydrogen bond donor, green: hydrophobic. Mixed colors correspond to mixed profiles.

Three-dimensional descriptors

While the previously discussed topological descriptors are based on the molecular graph only, a three-dimensional descriptor uses information about the spatial arrangement of the atoms or functional groups of a molecule. Since the process of complex formation between a receptor and a ligand is mainly driven by steric and electrostatic surface complementarity, methods based on 3D information seem to be more appropriate than topological features, at first sight. However, in practice it has often been observed that topological descriptors outperform 3D descriptors [20, 30, 37]. A general problem with 3D descriptors is handling conformational space. In principle, all low-energy conformers of a molecule would have to be taken into account, but conformational flexibility is often neglected and the descriptor is calculated for a single conformation only. On the other hand, it is likely that most data sets used for benchmarking similarity measures are biased towards methods based on 2D features, because inhibitors are often designed and synthesized in terms of series comprising common structural elements.

As in the 2D case, there exists a wide range of different descriptors (see [38-40] for reviews). Most of them are one-dimensional representations, i.e. they project 3D information into a bit- or integer string. In most cases not the analysis of the 3D structure is reduced to so-called pharmacophore elements or “hot spots” in the ligand structure that are most relevant for specificity and activity. Hydrogen bond donor, acceptor, and hydrophobic centers and their mutual distance are commonly used to address a bit in the string. This pair approach can be extended to triplets, where a characteristic value of the formed triangle (circumference, wing area, etc.) instead of the distance is used for addressing [41]. The approach can also be applied to conformation ensembles by disjunction of bit strings created from different conformations [42]. As a logical extension of such 3-point pharmacophores, 4-point pharmacophores have been used [43].

In contrast to the previous approaches, some other descriptors are based on superimposing 3D structures [44-46], leading to more time-consuming comparison algorithms. An example are so-called field descriptors. They are derived from the analysis of superimposed molecules on a rectangular grid, whose elements contain local properties of each molecule [47]. Other similarity calculations are based on fragment superposition [48, 49]. An interesting new approach for similarity calculations is the determination of affinity fingerprints [50, 51]. Here, the affinity to a reference panel of proteins is determined for each compound in the database is determined - either experimentally or by means of docking. Similarity indices are then calculated by comparing the affinity vectors. The underlying idea is that compounds showing similar binding

properties to the reference panel of proteins will also display similarity with respect to unknown targets.

Test Scenario: Application of Fast Similarity Searching Algorithms

In this test scenario, a single active compound per target is used to retrieve similar compounds from a database. Five fast algorithms are used for this purpose. Results are compared and possible ways to combine the results of searches with different algorithms are discussed.

Library Generation

Diverse sets of inhibitors were compiled manually for eleven well-established pharmaceutical targets. The main – subjective – selection criterion was to cover as many different compound classes as possible for each target. Compounds were retrieved from the WDI and from available complex crystal structures in the PDB [75]. Where patent literature was consulted in addition, references are given in the following list. The final data set consisted of 58 ACE inhibitors[76], 43 angiotensin II antagonists[77], 33 aldose reductase inhibitors[78, 79], 32 cyclooxygenase inhibitors[80-82], 55 H1 antihistamines[83], 41 HIV protease inhibitors[84], 21 HIV reverse transcriptase inhibitors[85], 19 monoamine oxidase A inhibitors[86, 87], 23 gelatinase A inhibitors[88], 36 thrombin inhibitors[89, 90] and 37 topoisomerase II inhibitors[91].

All compounds were stored as SMILES and converted to single 3D conformations in Sybyl mol2 format¹ by means of CORINA [13, 14]. A C routine was used to generate the molecular structure most likely to be the dominant species at neutral pH under physiological conditions. Hydrogen atoms were removed from acidic groups to generate the anionic form, and were added to aliphatic amines and acyclic amidine nitrogens to generate the charged cationic form.

A subset WDI database was prepared by stepwise removal of

- compounds containing one of the mechanism keywords “ACE”, “aldose-reductase”, “ampa”, “angiotensin”, “antihistamine”, “beta-lactamase”, “cyclooxygenase”, “hiv”, “mao”, “thrombin”, “topoisomerase” (in order to ensure complementarity to the individual sets of active compounds in this and other studies),

¹ Both SMILES[92] and mol2[93] are specific formats to store small molecule structures. SMILES encode the 2D structure of a molecule in a text string, whereas the mol2 format assigns a specific atom type class to each atom according to its environment and stores 3D coordinates.

- compounds with molecular weight greater than 800 or less than 200 (because drug-like molecules should fall in this molecular weight range)
- compounds with saturated carbon chains longer than 7 carbon atoms (in order to avoid non-drug-like alcohols and fatty acids)
- compounds without at least one oxygen (very few drugs contain no oxygen atom)
- compounds without at least one hydrogen (to exclude poly-halogenated compounds)
- compounds containing elements other than C, N, O, P, S and halogens (again to ensure the drug-likeness of all compounds)

The remaining WDI compounds were clustered according to Daylight fingerprint similarity by means of the nearest-neighbors clustering algorithm [94] implemented in the Daylight toolkit [95] considering the 14 nearest neighbors and requiring at least 8 cluster members per cluster. One compound per cluster was selected, resulting in a database of approximately 10,000 compounds. Further selection steps included the removal of macrocycles with larger than 12-membered rings, approximately 70% of the molecules with a steroid-type scaffold (because they were over-represented in the selected subset) and of molecules with more than 13 pharmacophore centers as determined by the TP software described in section 2.1. This step essentially serves to remove large molecules with many polar functional groups that are unlikely drug candidates. The final size of the database was 7528 compounds, which were stored as a SMILES list and converted to Sybyl mol2 format [96, 97] by means of CORINA. Protonation states were set as described above for the sets of active compounds. Although care has been taken to remove those compounds of the WDI that are inhibitors of one of the targets used in the following calculations, this removal is certainly not complete. Nevertheless, to a good approximation the WDI compounds can and will be regarded as inactive in all calculations presented in the following.

4.2. Computational Details

The 398 molecules in the 11 inhibitor classes were successively selected as query molecules. In each run, the complete database of 7925 compounds (7528 WDI compounds plus the remaining 397 active compounds) was ordered from high to low similarity with respect to the query molecule according to the five different algorithms described in section 2: Daylight fingerprints, UNITY 2D screens, correlation vectors, topological pharmacophores and feature trees. This leads to five different ranking lists per molecule that were subsequently analyzed with respect to the distribution of compounds belonging to the same class as the query molecule (“active com-

pounds”). The Daylight software was used with four different settings regarding the length of the fingerprints (default, 512 bit, 1 kbit and 2 kbit). In general, we have observed that a performance increase for longer fingerprints becomes visible at low similarity levels only (Tanimoto coefficient < 0.9). Fingerprint length is a critical issue in the assembled database. Standard fingerprints (whose size varies from molecule to molecule until a fixed density of set bits is achieved) performed significantly worse than those of constant length, as shown in Figure 2 for five selected targets. Fingerprints of 1 kbit length were chosen for comparison with the other methods, because increasing the size to 2 kbit does not improve performance. Standard Unity 2D screens also have a length of approximately 1 kbit.

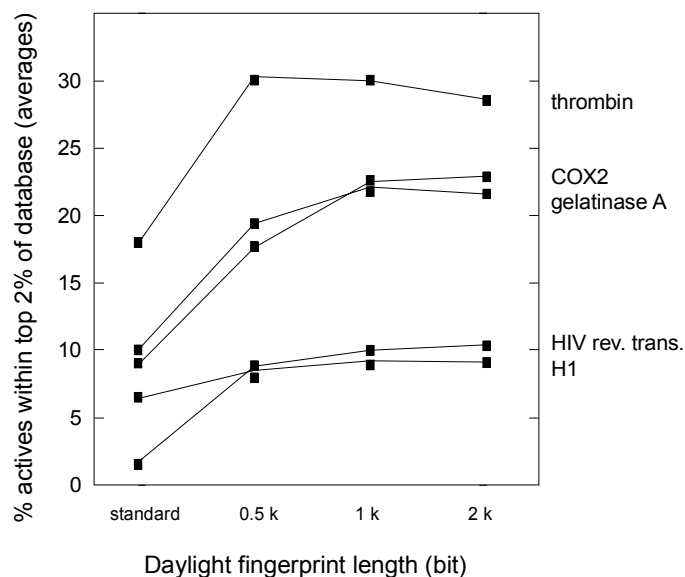
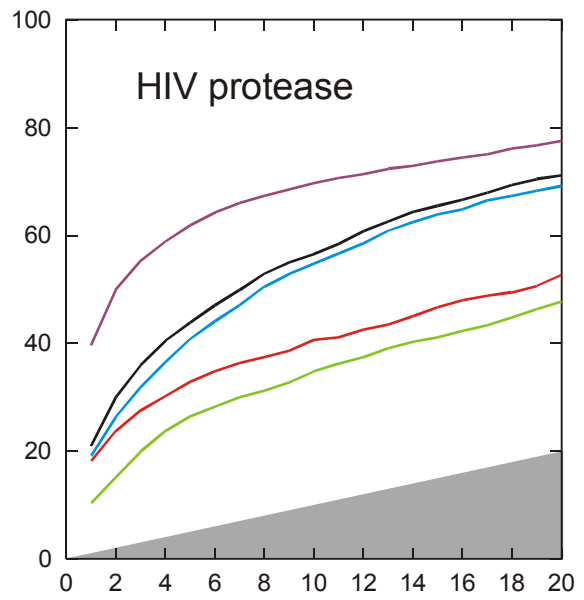
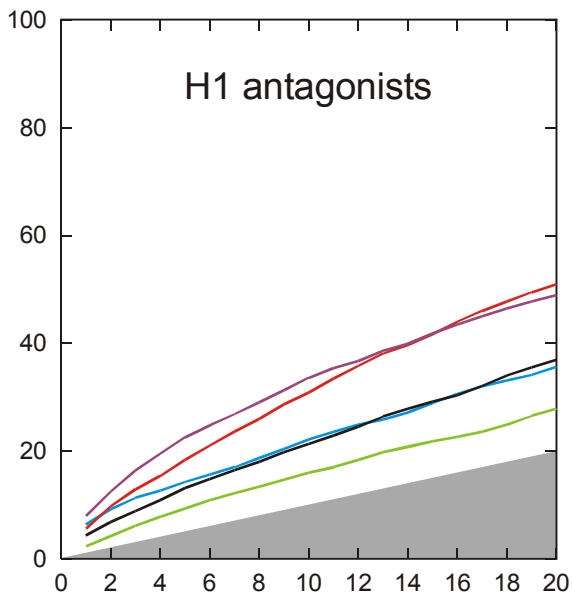
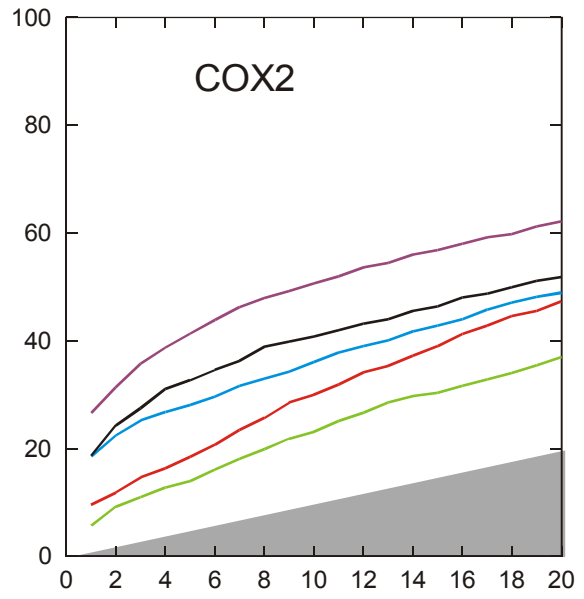
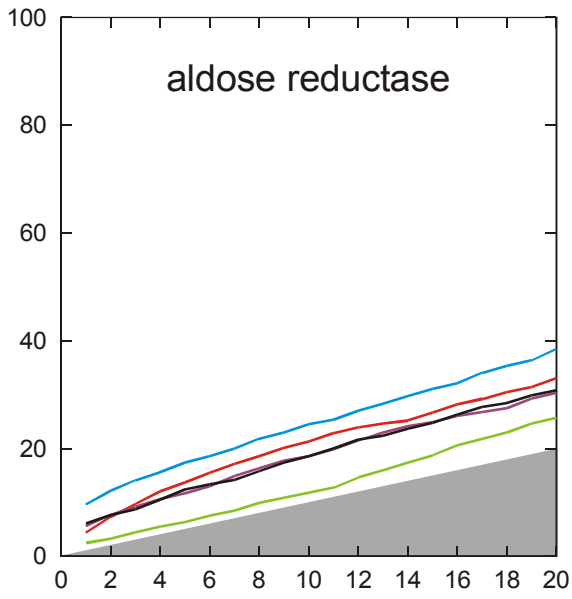
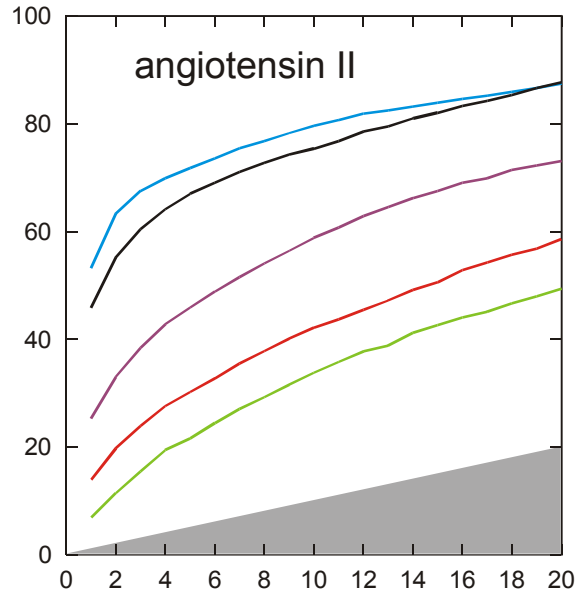
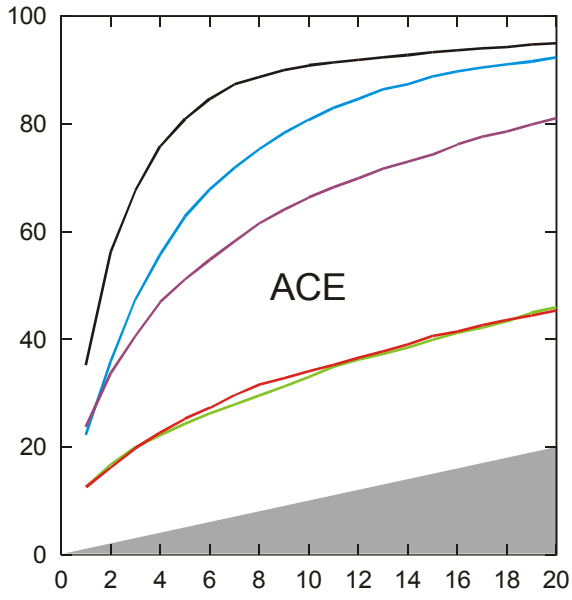
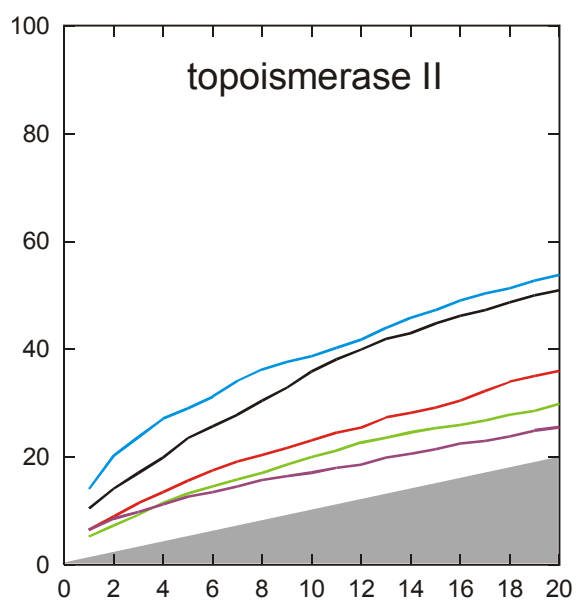
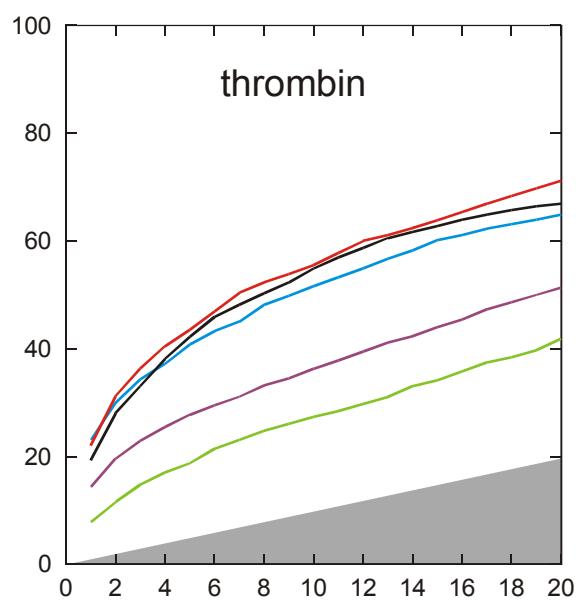
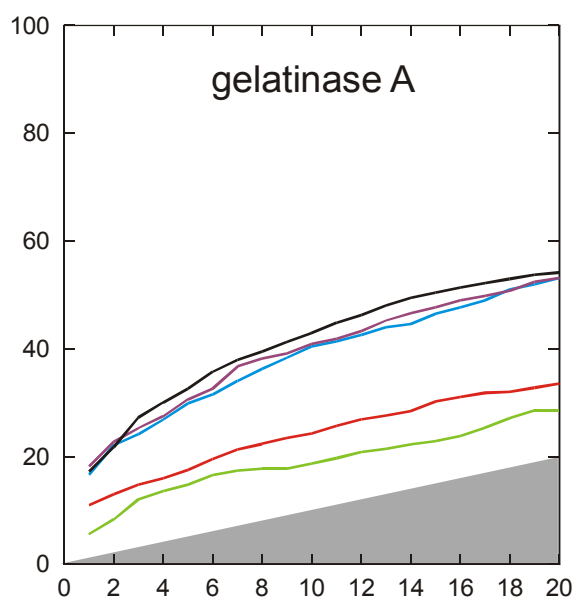
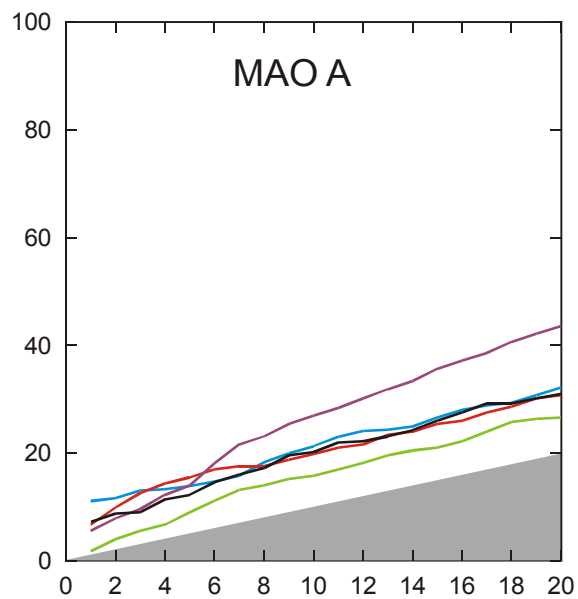
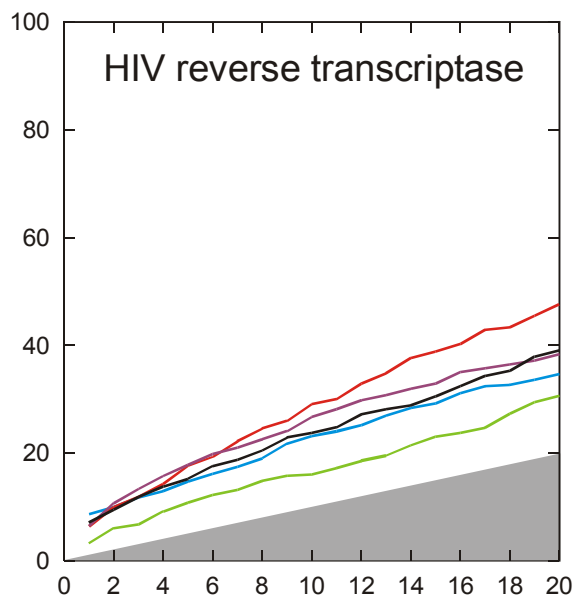


Figure 2. Performance of Daylight fingerprints of different length in retrieving active compounds related to a given query molecule. Fingerprints of 1kbit length are a reasonable choice for the present test scenario.

4.3. Discussion of Screening Results

The performance of the five methods was compared by averaging, over all members of the class, the percentage of active compounds contained in each percentile of the ranked database. The results, graphically depicted in Figure 3, vary strongly depending on the inhibitor class. For some inhibitor classes several algorithms generate very satisfying results, while for others all algorithms perform poorly.





- Feature trees (FT)
- Daylight fingerprints (DF)
- Topol. pharmacophores (TP)
- Correlation vectors (CV)
- UNITY 2D screens (UF)

Figure 3. (preceding 2 pages) Average percentage of active compounds versus percentage of the ranked database plotted for eleven targets and five different fast similarity algorithms.

On average, many active compounds receive top ranks in the ACE, angiotensin II and HIV protease classes, while enrichment is very poor for aldose reductase, H1 antagonists, HIV reverse transcriptase and MAO A. Results of intermediate quality are obtained for COX2, gelatinase A and thrombin. The two fingerprint-based methods (DF and UF) and the feature trees software (FT) outperform correlation vectors (CV) and topological pharmacophores (TP) in database ranking. The CV method performs worst in all cases. However, there is no single method that performs best for all eleven inhibitor classes. The FT method seems to be superior to the other methods if correct matching of hydrophobic groups connected by single bonds is important. For example, this is the case with COX2 inhibitors, the majority of which consists of three planar ring moieties whose centers span an angle of about 60°. As a point in case, consider the three structurally related COX2 inhibitors 1-3 in Figure 4. The differences in the central ring lead to different sets of bond paths and thus to different fingerprints whereas, for the FT software, this difference is of minor importance, since the relative orientation and polarity remains the same for all three compounds. The comparison of fused aromatic and heterocyclic compounds is an inherent shortcoming of the FT method, as is obvious from the topoisomerase II results. To impart an impression of topoisomerase II inhibitors, three representatives of this class are shown in Figure 5. Finally, TP emphasizes matches of hydrogen-bond donors and acceptors more than the other methods do. For this reason, the TP method performs well for sets of polar molecules like thrombin inhibitors.

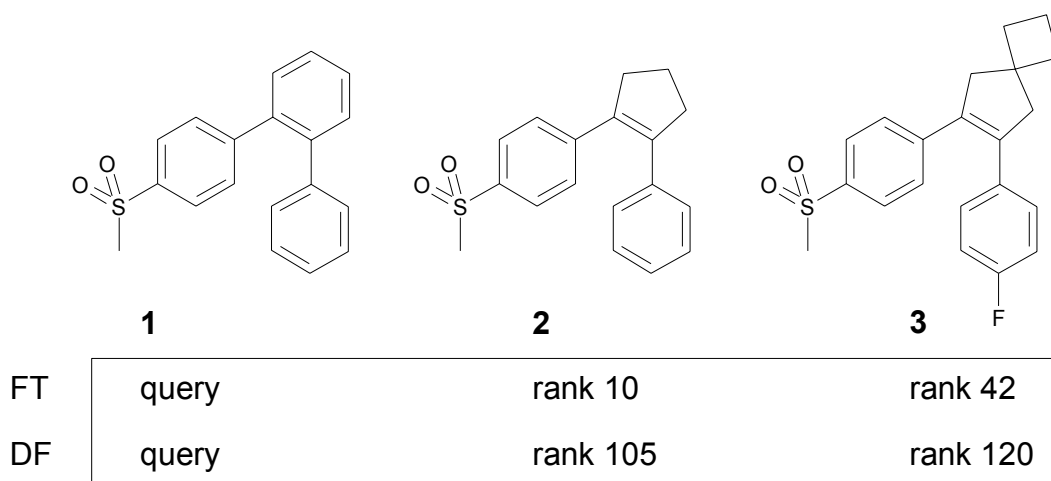


Figure 4. Three structurally related COX2 inhibitors. Ranks for compounds 2 and 3 were calculated using the FT and DF methods with compound 1 as a query molecule.

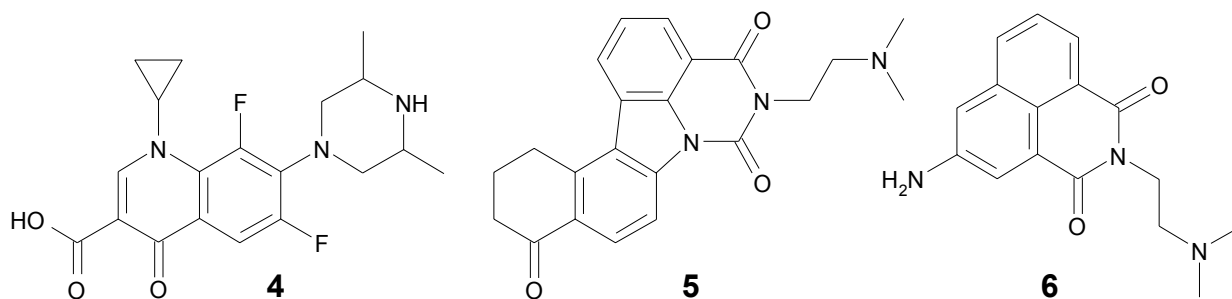


Figure 5. Three topoisomerase II inhibitors

Plots of average enrichment, for example as shown in Figure 3, can be misleading, because they can be influenced easily by outliers. They do not give information on the variation of enrichment achieved for different members of a target class. This variation is high for all methods and target classes: A search with a single active molecule can retrieve many related active compounds or no actives at all. Small structural changes in the query molecule can significantly change the database ranking.

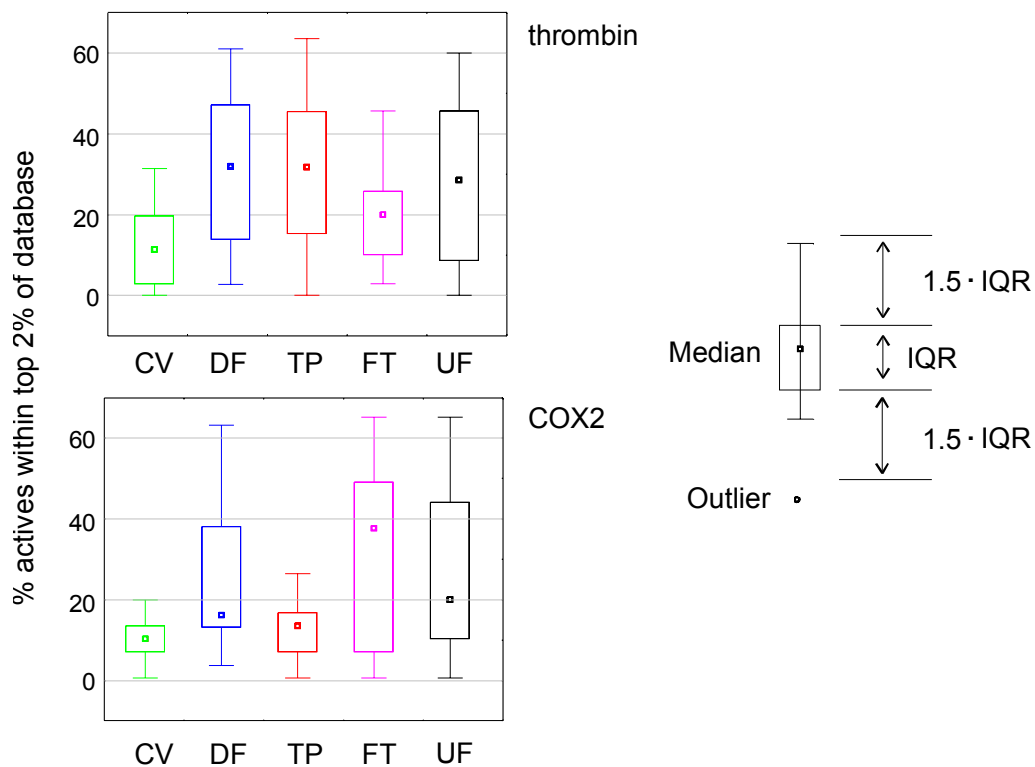


Figure 6. Box plots showing the variation of the percentage of active compounds contained in the top 2% of the ranked database. Results are shown for the COX2 and thrombin libraries only. In the legend on the right, IQR is the inter-quartile range, i.e. the interval of a data set that is centered around the median and contains 50% of the data points.

How strongly the enrichment varies depending on the query molecule can be illustrated by means of Box plots [98]. As illustrated in Figure 6 for thrombin and COX2, the percentage of active compounds within the top 2% of the database varies between zero and very high values. For thrombin, the average values shown in Figure 6 seem to indicate that the order of performance for the five methods is TP, DF, UF > FT > CV, but from the Box plots it is apparent that the difference between CV and FT is actually very small. In fact, it is insignificant at an error level of 5%, as can be deduced from statistical analyses of the data by means of the Tukey test [99]. Similarly, the plots for COX2 in Figure 3 suggest the order FT > UF, DF > TP > CV. Consulting the Box plots one must conclude that the median percentage of actives retrieved is almost as low for DF as it is for CV and TP. According to the Tukey test, at an error level of 5% the only robust statements about the COX2 data are that FT performs better than CV and TP.

From the above it can be concluded that it is difficult to predict which of the methods will work best for a particular query molecule. To increase the success rate, it is therefore advisable to perform searches with two or more different methods. But how should the results be combined? Two possible strategies to select compounds from two searches employing methods A and B are:

1. Selection of all molecules that receive high ranks with either method A *or* method B. For each compound occurring in both, the A and B rank lists, the position with the higher rank is selected and the combined list is sorted again according to rank. In cases of conflict, compounds are placed in arbitrary order. For this *union strategy* to be successful, methods A and B should emphasize different aspects of similarity and should therefore retrieve different active molecules on top ranks. It is then likely that more hits will be found in the union rank list than in any individual rank list.
2. Selection of all molecules that receive high ranks with method A *and* method B. Such a procedure is successful if a common set of active molecules is ranked high by both methods A and B, while the top-ranking inactive compounds differ. This *intersection strategy* will be followed if it is important to reduce the number of false positives – even at the cost of reducing the number of true positives.

For the present test libraries, the union strategy was tested by counting the number of actives among the top 2% of the individual rank lists in comparison to those among the top 2% of the union rank list. On average, the union strategy leads to significantly more hits than the weaker individual method, but it is rarely superior to the better performing method. Table 1 illustrates this observation for the ACE sublibrary. The union strategy thus does not retrieve more actives, but, since it is difficult to estimate in advance which method will perform better, it leads to more

robust results. A generally valid recommendation is to combine one of the fingerprint-based methods with any of the other three algorithms.

Table 1. Results of the union strategy for the set of ACE inhibitors: Table entries are the average numbers of active compounds contained in the top 2% (158 compounds) of the union rank lists. Values on the diagonal (combinations of each method with itself) are the average number of actives in the top 2% of the individual rank lists.

	FT	TP	CV	DF	UF
FT	19				
TP	17	9			
CV	18	12	10		
DF	22	17	18	21	
UF	27	23	24	28	32

An example for the application of the intersection strategy is given in Table 2 with the antihistamine 7 as a query molecule. C_{tot} is the number of molecules occurring in the top 2% of both rank lists; C_{act} is the number of active molecules contained therein. From Table 2 it can be seen that the proportion of active molecules in the intersection lists is relatively high compared to the top 2% of the individual rank lists (values on the diagonal of Table 1). This comparison is not fair, however, since most of the active molecules should be found on the highest ranks of the individual rank lists. Considering only a single rank file, one would select only very few top-ranked compounds in order to keep the number of false positives to a minimum. Therefore, C_{act} is best compared to the number of actives among the C_{tot} top ranked compounds of each individual ranking list. This number is given as I_{act} in Table 2.

In the example in Table 2, C_{act} is larger than I_{act} for most combinations, which means that the intersection strategy identifies more active molecules than any individual algorithm. The efficiency of the intersection strategy is smallest for the combination of the two fingerprint methods DF and UF, because the two rank lists overlap significantly. For all other combinations, between 1/4 and 1/3 of C_{tot} are active compounds. Figure 7 shows results of the combination of UF and FT. Seven active compounds (8-14) are contained within the intersection list. Compounds 13 and 14 are among the top 16 compounds of both the UF and the FT rank lists. In addition, the top 16 compounds of the FT rank list also contain compounds 12, 15 and 16. The latter two compounds are not on the intersection list, since they are on rank positions beyond 158 on the UF rank list. The example presented in Table 2 is representative for the data used in this study. On average, the intersection strategy significantly reduces the number of false positives. Best results are achieved if two methods are combined that perform well separately. The combination of FT and one of the fingerprint methods can be generally recommended.

Table 2. Results of the intersection strategy for query molecule 7 in Figure 7: C_{tot} is the total number of common molecules in the top 2% of both rank lists, C_{act} is the number of H1 antagonists contained therein. I_{act} is the number of active molecules contained in the C_{tot} top ranking molecules of the library ranked according to the method given in the leftmost column. Diagonal elements of this table give the number of active molecules within the top 2% of the ranked database for each method (where $C_{\text{act}} = I_{\text{act}}$ and $C_{\text{tot}} = 158$).

	FT			TP			CV			DF			UF		
	C_{tot}	C_{act}	I_{act}	C_{tot}	C_{act}	I_{act}	C_{tot}	C_{act}	I_{act}	C_{tot}	C_{act}	I_{act}	C_{tot}	C_{act}	I_{act}
FT	158	13	13	19	4	5	8	3	3	12	6	4	16	7	5
TP	19	4	2	158	8	8	3	1	0	9	4	1	12	3	1
CV	8	3	0	3	1	0	158	5	5	11	3	0	13	3	0
DF	12	6	2	9	4	1	11	3	2	158	8	8	90	7	6
UF	16	7	2	12	3	2	13	3	2	90	7	5	158	8	8

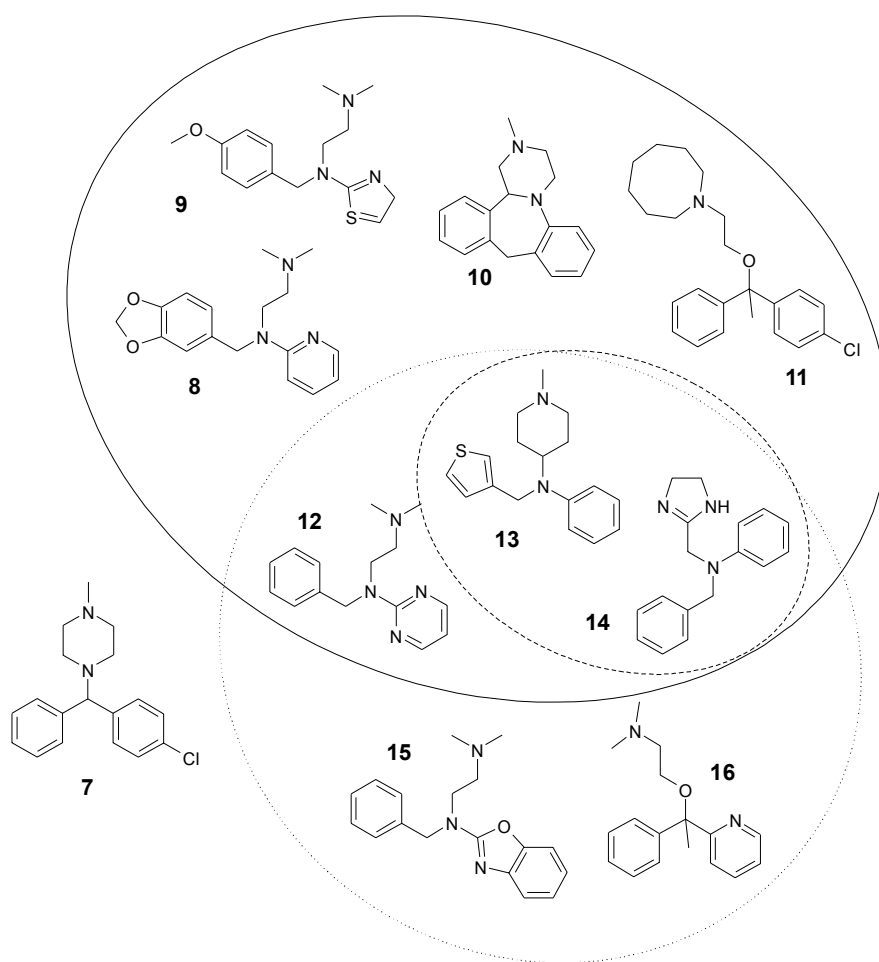


Figure 7. Results of the intersection strategy for the H1 receptor antagonist 7 as a query molecule. Compounds 8-14 are contained within the 16-membered intersection list of the DF and FT rank lists. Compounds 13 and 14, encircled by a dashed line, are within the top 16 molecules of the FT and DF rank lists, meaning that they are regarded as highly similar to the query molecule by both methods. The top 16 molecules of the FT list also include 12, 15 and 16.

References

- 1 H. Kubinyi, *Pharmazie* **1995**, *50*, 647-662.
- 2 J. Kuhlmann, *Int. J. Clin. Pharmacol. Ther.* **1997**, *35*, 541-552.
- 3 R. Lahana, *Drug Discovery Today* **1999**, *4*, 447-448.
- 4 J. H. Van Drie, M. S. Lajiness, *Drug Discovery Today* **1998**, *3*, 274-283.
- 5 W. P. Walters, M. T. Stahl, M. A. Murcko, *Drug Discovery Today* **1998**, *3*, 160-178.
- 6 Available Chemicals Directory, MDL Information Systems Inc., San Leandro, California, USA.
- 7 World Drug Index, 2/96, Derwent Information **1996**.
- 8 MDDR, MACCS Drug Data Report, MDL Information Systems Inc., San Leandro, California, USA.
- 9 Beilstein Database, Beilstein Informationssysteme GmbH, Frankfurt, Germany.
- 10 J. Cadwell, I. Gardner, N. Swales, *Toxicol. Pathol.* **1995**, *23*, 102-114.
- 11 C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, *Adv. Drug Delivery Rev.* **1997**, *23*, 3-25.
- 12 D. E. Clark, S. D. Pickett, *Drug Discovery Today* **2000**, *5*, 49-58.
- 13 J. Sadowski, C. Rudolph, J. Gasteiger, *Tetrahedron Comput. Methodol.* **1990**, *3*, 537-547.
- 14 Corina 2.1, Molecular Networks GmbH Computerchemie, Erlangen **1998**.
- 15 Concord, Tripos Associates Inc., St. Louis, Missouri, USA.
- 16 E. M. Landau, *Biomed. Health Res.* **1998**, *20*, 23-38.
- 17 M. A. Johnson, G. M. Maggiora, *Concepts and Applications of Molecular Similarity*, John Wiley & Sons, New York **1990**.
- 18 P. M. Dean, *Molecular similarity in drug design*, Chapman & Hall, London **1995**.
- 19 G. M. Downs, P. Willett, in K. B. Lipkowitz, D. B. Boyd (Eds.): *Reviews in Computational Chemistry*, Vol. 7, VCH Publishers, New York **1996**, p. 1-57.
- 20 R. D. Brown, Y. C. Martin, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572-584.
- 21 T. Pötter, H. Matter, *J. Med. Chem.* **1998**, *41*, 478-488.
- 22 H. Matter, M. Rarey, in G. Jung (Ed.): *Combinatorial Organic Chemistry*, Wiley-VCH, New York **1999**, p. 409-439.
- 23 M. F. Lynch, *Screening large chemical data files*, Ellis Horwood, Chichester **1975**.
- 24 R. Nilakantan, N. Bauman, J. S. Dixon, R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82-85.
- 25 Daylight Software Manual, Daylight Inc., Mission Viejo, California, USA.
- 26 MACCS II, MDL Information Systems Inc., San Leandro, California, USA.

- 27 Unity Chemical Information Software, Version 4.0, Tripos Associates Inc., St. Louis, Missouri, USA .
- 28 D. R. Flower, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379-386.
- 29 P. Willett, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983-996.
- 30 R. D. Brown, Y. C. Martin, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1-9.
- 31 G. Moreau, P. Broto, *Nouv. J. Chimie* **1980**, *4*, 359-360.
- 32 G. Schneider, W. Neidhart, T. Giller, G. Schmid, *Angew. Chem. Int. Ed.* **1999**, *38*, 2894-2896.
- 33 P. R. Gerber, *Proceedings of IV Girona Seminar on Molecular Similarity, Girona, Spain, July 5-7, 1999*.
- 34 P. R. Gerber, K. Müller, *J. Comput.-Aided Mol. Design* **1994**, *9*, 251-268.
- 35 P. R. Gerber, *J. Comput.-Aided Mol. Design* **1998**, *12*, 37-51.
- 36 M. Rarey, J. S. Dixon, *J. Comput.-Aided Mol. Design* **1998**, *12*, 471-490.
- 37 Y. C. Martin, M. G. Bures, R. D. Brown, *Pharm. Pharmacol. Commun.* **1998**, *4*, 147-152.
- 38 P. A. Bath, A. R. Poirette, P. Willett, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 141-147.
- 39 Y. C. Martin, M. G. Bures, P. Willett, in K. B. Lipkowitz, D. B. Boyd (Eds.): *Reviews in Computational Chemistry, Vol. 1*, VCH Publishers, New York **1990**, p. 213-263.
- 40 A. C. Good, J. S. Mason, in K. B. Lipkowitz, D. B. Boyd (Eds.): *Reviews in Computational Chemistry, Vol. 7*, VCH Publishers, New York **1996**, p. 67-117.
- 41 A. C. Good, T. J. A. Ewing, D. A. Gschwend, I. D. Kuntz, *J. Comput.-Aided Mol. Design* **1995**, *9*, 1-12.
- 42 S. D. Pickett, J. S. Mason, I. M. McLay, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1214-1223.
- 43 J. S. Mason, I. Morize, P. Menard, D. L. Cheney, D. L. Hulme, R. F. Labaudiniere, *J. Med. Chem.* **1999**, *42*, 3251-3264.
- 44 S. K. Kearsley, G. M. Smith, *Tetrahedron Comput. Methodol.* **1990**, *3*, 615-630.
- 45 M. D. Miller, R. P. Serhidan, S. K. Kearsley, *J. Med. Chem.* **1999**, *42*, 1505-1514.
- 46 C. Lemmen, T. Lengauer, G. Klebe, *J. Med. Chem.* **1998**, *41*, 4502-4520.
- 47 D. A. Thorner, D. J. Wild, P. Willett, P. M. Wright, in H. Kubinyi, G. Folkers, Y. C. McMartin (Eds.): *3D QSAR in drug design: ligand protein interactions and molecular similarity, Vol. 9/10/11*, Kluwer/Escom, Dordrecht **1998**, p. 301-320.
- 48 C. Lemmen, T. Lengauer, in K. Gundertoft, F. S. Jorgensen (Eds.): *Molecular modelling and prediction of bioactivity, Proceedings of the 12th European symposium on quantitative structure-activity relationships*, Plenum Press, New York **1999**.
- 49 C. Lemmen, T. Lengauer, *J. Comput.-Aided Mol. Design* **2000**, *14*, 215-232.
- 50 L. M. Kauvar, D. L. Higgins, H. O. Villar, J. R. Sportsman, A. Engvist-Goldstein, R. Bukar, K. E. Bauer, H. Dilley, D. M. Rocke, *Chem. Biol.* **1995**, *2*, 107-118.

- 51 H. Briem, I. D. Kuntz, *J. Med. Chem.* **1996**, *39*, 3401-3408.
- 52 J. M. Blaney, J. S. Dixon, *Perspectives in Drug Discovery and Design, Vol. 1*, Kluwer/Escom, Dordrecht **1993**, p. 301-319.
- 53 I. D. Kuntz, E. C. Meng, B. K. Shoichet, *Acc. Chem. Res.* **1994**, *27*, 117-123.
- 54 P. M. Colman, *Curr. Op. Struct. Biol.* **1994**, *4*, 868-874.
- 55 T. P. Lybrand, *Curr. Op. Struct. Biol.* **1995**, *5*, 224-228.
- 56 G. Jones, P. Willett, *Curr. Op. Biotechnology* **1995**, *6*, 652-656.
- 57 T. Lengauer, M. Rarey, *Curr. Op. Struct. Biol.* **1996**, *6*, 402-406.
- 58 H. Kubinyi, *Curr. Op. Drug Disc. Dev.* **1998**, *1*, 16-27.
- 59 M. Rarey, S. Wefing, T. Lengauer, *J. Comput.-Aided Mol. Design* **1996**, *10*, 41-54.
- 60 M. Rarey, B. Kramer, T. Lengauer, G. Klebe, *J. Mol. Biol.* **1996**, *261*, 470-489.
- 61 M. Rarey, B. Kramer, T. Lengauer, *J. Comput.-Aided Mol. Design* **1997**, *11*, 369-384.
- 62 B. Kramer, M. Rarey, T. Lengauer, *Proteins* **1997**, *Suppl. 1*, 221-225.
- 63 M. Rarey, B. Kramer, T. Lengauer, *Proteins* **1999**, *34*, 17-28.
- 64 M. Rarey, B. Kramer, T. Lengauer, *Bioinformatics* **1999**, *15*, 243-250.
- 65 B. Kramer, M. Rarey, T. Lengauer, *Proteins* **1999**, *37*, 228-241.
- 66 G. Klebe, F. Dullweber, H.-J. Böhm, in R. B. Raffa (Ed.): *Thermodynamics of the Drug-Receptor Interaction*, John Wiley & Sons, Inc., New York, in press.
- 67 Ajay, M. A. Murcko, *J. Med. Chem.* **1995**, *38*, 4953-4967.
- 68 T. I. Oprea, G. R. Marshall, in H. Kubinyi, G. Folkers, Y. C. McMartin (Eds.): *3D QSAR in drug design: ligand protein interactions and molecular similarity, Vol. 9/10/11*, Kluwer/Escom, Dordrecht **1998**, p. 3-17.
- 69 R. M. A. Knegtel, P. D. J. Grootenhuis, in H. Kubinyi, G. Folkers, Y. C. McMartin (Eds.): *3D QSAR in drug design: ligand protein interactions and molecular similarity, Vol. 9/10/11*, Kluwer/Escom, Dordrecht **1998**, p. 99-114.
- 70 J. D. Hirst, *Curr. Op. Drug Disc. Dev.* **1998**, *1*, 28-33.
- 71 H.-J. Böhm, M. Stahl, *Med. Chem. Res.* **1999**, *9*, 445-462.
- 72 J. R. H. Tame, *J. Comput.-Aided Mol. Design* **1999**, *13*, 99-108.
- 73 H.-J. Böhm, *J. Comput.-Aided Mol. Design* **1994**, *8*, 243-256.
- 74 H. Gohlke, M. Hendlich, G. Klebe, *J. Mol. Biol.* **2000**, *295*, 337-356.
- 75 F. C. Bernstein, T. E. Koetzle, G. J. B. Williams, J. Meyer, E. F., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, T. M., *J. Mol. Biol.* **1977**, *112*, 535-542.
- 76 C. A. Fink, *Exp. Opin. Ther. Patents* **1996**, *6*, 1147-1164.
- 77 P. K. Chakravarty, *Exp. Opin. Ther. Patents* **1995**, *5*, 431-458.
- 78 L. Costantino, G. Rastelli, G. Cignarella, P. Vianello, D. Barlocco, *Exp. Opin. Ther. Patents* **1997**, *7*, 843-858.

- 79 N. Ashizawa, T. Aotsuka, *Drugs of the Future* **1998**, *5*, 521-529.
- 80 J. S. Carter, *Exp. Opin. Ther. Patents* **1997**, *8*, 21-29.
- 81 R. W. Friesen, C. Brideau, C. C. Chan, S. Charleson, D. Deschenes, D. Dubé, D. Ethier, R. Fortin, J. Y. Gauthier, Y. Girard, R. Gordon, G. M. Greig, D. Riendau, C. Savoie, Z. Wang, E. Wong, D. Visco, L. J. Xu, R. N. Young, *Bioorg. Med. Chem. Lett.* **1998**, *8*, 2777-2782.
- 82 A. S. Kalgutkar, *Exp. Opin. Ther. Patents* **1999**, *9*, 831-849.
- 83 R. Aslanian, J. J. Piwinski, *Exp. Opin. Ther. Patents* **1997**, *7*, 201-207.
- 84 R. A. Crusciel, K. R. Romines, *Exp. Opin. Ther. Patents* **1997**, *7*, 111-121.
- 85 J. R. Proudfoot, *Exp. Opin. Ther. Patents* **1998**, *8*, 971-982.
- 86 C. H. Gleiter, H.-P. Volz, *Exp. Opin. Invest. Drugs* **1996**, *5*, 409-419.
- 87 S. Jegham, P. George, *Exp. Opin. Ther. Patents* **1998**, *8*, 1143-1150.
- 88 R. P. Beckett, M. Whittaker, *Exp. Opin. Ther. Patents* **1998**, *8*, 259-282.
- 89 M. R. Wiley, M. J. Fisher, *Exp. Opin. Ther. Patents* **1997**, *7*, 1265-1282.
- 90 P. E. J. Sanderson, A. M. Naylor-Olsen, *Curr. Med. Chem.* **1998**, *5*, 289-304.
- 91 A. K. Chakraborty, H. K. Majumder, C. P. Hodgson, *Exp. Opin. Ther. Patents* **1994**, *4*, 655-668.
- 92 D. Weininger, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31-36.
- 93 M. Clark, R. D. Cramer III, N. Van Opdenbosch, *J. Comp. Chem.* **1989**, *10*, 982-1012.
- 94 R. A. Jarvis, E. A. Patrick, *IEEE Transactions on Computers* **1973**, *C22*, 1025-1034.
- 95 Daylight Chemical Information Systems, San Leandro, California, USA.
- 96 M. Clark, R. D. Cramer III, N. Van Opdenbosch, *J. Comp. Chem.* **1989**, *10*, 982-1012.
- 97 Sybyl molecular modeling software, Version 6.2, Tripos Associates Inc., St. Louis, Missouri, USA.
- 98 R. R. Sokal, F. J. Rohlf, *Biometry: The principles and practice of statistics in biological research*, W. H. Freeman and Company, New York **1995**.
- 99 J. C. Hsu, *Multiple comparisons, theory and methods*, Chapman & Hall, London **1996**.
- 100 G. J. Hanson, *Exp. Opin. Ther. Patents* **1997**, *7*, 729-733.
- 101 <http://www.protherics.com/crunch/>
- 102 P. R. Gerber, K. Müller, *J. Comput.-Aided Mol. Design* **1995**, *9*, 251-268.
- 103 M. D. Elridge, C. W. Murray, T. R. Auton, G. V. Paolini, R. P. Mee, *J. Comput.-Aided Mol. Design* **1997**, *11*, 425-445.
- 104 C. A. Baxter, C. W. Murray, D. E. Clark, D. R. Westhead, M. D. Eldridge, *Proteins* **1998**, *33*, 367-382.
- 105 I. Muegge, Y. C. Martin, *J. Med. Chem.* **1999**, *42*, 791-804.

- 106 J. B. O. Mitchell, R. A. Laskowski, A. Alex, J. M. Thornton, *J. Comp. Chem.* **1999**, *20*, 1165-1177.
- 107 M. Stahl, H.-J. Böhm, *J. Mol. Graphics Mod.* **1998**, *16*, 121-132.
- 108 P. S. Charifson, J. J. Corkery, M. A. Murcko, W. P. Walters, *J. Med. Chem.* **1999**, *42*, 5100-5109.