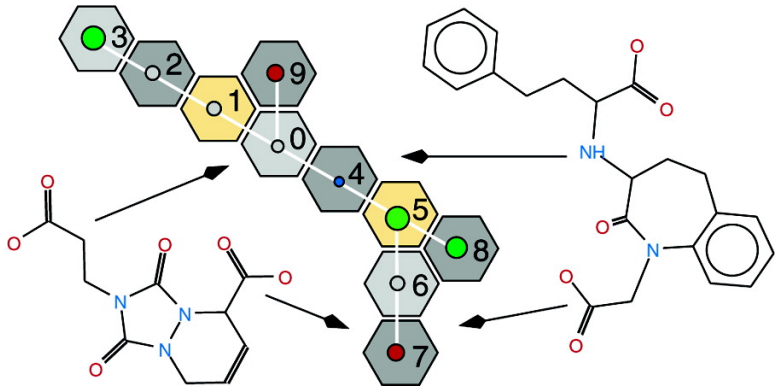
Journal of Medicinal Chemistry

Article

Multiple-Ligand-Based Virtual Screening: Methods and Applications of the MTree Approach

Gerhard Hessler, Marc Zimmermann, Hans Matter, Andreas Evers, Thorsten Naumann, Thomas Lengauer, and Matthias Rarey

J. Med. Chem., **2005**, 48 (21), 6575-6584• DOI: 10.1021/jm050078w • Publication Date (Web): 20 September 2005



More About This Article

Additional resources and features associated with this article are available within the HTML version:

- Supporting Information
- Links to the 9 articles that cite this article, as of the time of this article download
- Access to high resolution figures
- Links to articles and content related to this article
- Copyright permission to reproduce figures and/or text from this article

View the Full Text HTML

Journal of Medicinal Chemistry is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Multiple-Ligand-Based Virtual Screening: Methods and Applications of the MTree Approach

Gerhard Hessler,[†,‡] Marc Zimmermann,[†,§] Hans Matter,[‡] Andreas Evers,[‡] Thorsten Naumann,[‡] Thomas Lengauer,[∥] and Matthias Rarey*[,⊥]

Drug Design, Chemical Sciences, Sanofi-Aventis Deutschland GmbH, Frankfurt, Germany, Fraunhofer-Institute for Algorithms and Scientific Computing, St. Augustin, Germany, Max-Planck-Institute for Informatics, Saarbrücken, Germany, and Center for Bioinformatics, University of Hamburg, Hamburg, Germany

Received January 28, 2005

We present a novel approach for ligand-based virtual screening by combining query molecules into a multiple feature tree model called MTree. All molecules are described by the established feature tree descriptor, which is derived from a topological molecular graph. A new pairwise alignment algorithm leads to a consistent topological molecular alignment based on chemically reasonable matching of corresponding functional groups. These multiple feature tree models find application in ligand-based virtual screening to identify new lead structures for chemical optimization. Retrospective virtual screening with MTree models generated for angiotensin-converting enzyme and the α1a receptor on a large candidate database yielded enrichment factors up to 71 for the first 1% of the screened database. MTree models outperformed database searches using single feature trees in terms of hit rates and quality and additionally identified alternative molecular scaffolds not included in any of the query molecules. Furthermore, relevant molecular features, which are known to be important for affinity to the target, are identified by this new methodology.

1. Introduction

Virtual screening has become an established and important tool for lead identification as starting point for chemical optimization in drug discovery programs.[1−3] In the absence of any three-dimensional information of protein−ligand complexes from X-ray crystallography, virtual screening has to rely on one or multiple known active ligands of the biological target. Here, the similarity principle provides the conceptual basis for attempts to identify novel molecules related to known ligands by similarity searching.[4] To this end, standard topological descriptors such as MACCS substructure keys[5] or Unity topological fingerprints[6] are known to identify chemically related analogues,[7] while more advanced descriptors such as CATS[8] or feature trees,[9] are better suited for scaffold hopping.[10] Nevertheless, similarity searching is somewhat limited, since all features of a query molecule are equally important for ranking candidate molecules, regardless of any effect of these features on the biological activity at a particular target. Consequently, additional structural information on a small set of ligands for the same target allows for using approaches that first identify functional groups or features relevant to activity and second use this knowledge to rank-order novel molecules. Here, 3D pharmacophore-based methods are often used in ligand-based virtual screening to derive important features for biological activity to rank novel compounds. Those methods are often limited by the requirement of a reliable superposition plus the necessity to generate 3D conformations for all molecules in a candidate database for virtual screening.

For a timely and efficient follow-up on the huge amount of experimental data generated today in typical drug discovery programs in industrial settings, a reliable method to automatically extract pharmacophoric information for fast follow-up searching is required. This prompted us to develop a novel method that is intended to combine the advantages of similarity and pharmacophore searching on the basis of 2D structural representations only.

A selected set of query molecules is converted into a topological model, a so-called MTree model, based on chemically reasonable matching of corresponding functional groups. This matching is performed using a new alignment algorithm for feature trees,[11] called dynamic-match-search. The algorithm creates a topological mapping of the most similar fragments from a set of structurally diverse, but active, molecules. Conserved features are characterized by high similarity scores of the corresponding nodes in the MTree model. Such a model is conceptually similar to a 2D pharmacophore and highlights those chemical substructures that exhibit consistent protein−ligand interactions and can contribute significantly to biological affinity. In this paper, we present the basis of this approach, its validation with known X-ray structures or known pharmacophore models for ACE (angiotensin-converting enzyme), and the α1a receptor plus its efficiency in retrospective ligand-based virtual screening. In addition, for the α1a receptor, we compare the performance of the MTree model for virtual screening with a 3D-pharmacophore model generated with Catalyst.[12]

* Corresponding author. Address: Bundesstr. 43, 20146 Hamburg, Germany. Phone: +49(0)40 428387351. Fax: +49(0)40 428387351. E-mail: rarey@zbh.uni-hamburg.de.
† Both authors contributed equally to this work.
‡ Sanofi-Aventis Deutschland GmbH.
§ Fraunhofer-Institute for Algorithms and Scientific Computing.
∥ Max-Planck-Institute for Informatics.
⊥ University of Hamburg.

10.1021/jm050078w CCC: $30.25
Published on Web 09/20/2005

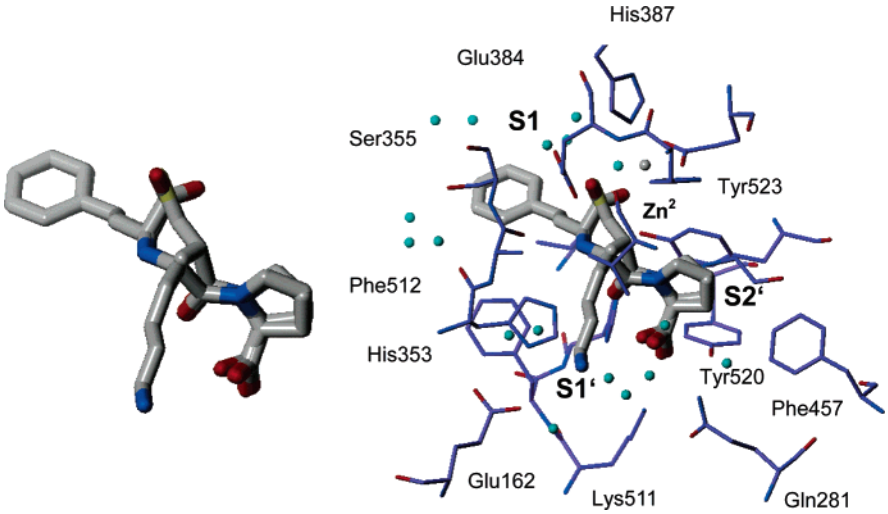


Figure 1. Structure-based alignment of ACE inhibitors with known experimental binding mode from comparison of X-ray structure analysis; key amino acids in the ACE binding site are labeled.

2. Results

To evaluate the performance of the MTree models we carried out retrospective virtual screenings for two different targets (ACE and the α1a receptor). For each target, we performed the following experiment. A small set of inhibitors was selected for model building. The MTree model was created automatically using the dynamic-match-search algorithm. Neither 3D information nor pharmacophore information was used. Then the models were applied to prioritize a database of 47 691 drug-like compounds. High-ranking molecules were then analyzed for activity against the corresponding target.

2.1. Angiotensin-Converting Enzyme. 2.1.1. X-ray Structures of ACE Inhibitor Complexes. Angiotensin-converting enzyme is a type I membrane-anchored dipeptidyl carboxypeptidase that is essential for blood pressure regulation and electrolyte homeostasis through the renin−angiotensin−aldosteron system.

Inhibitors of ACE, such as captopril, enalapril, and lisinopril, are widely described to control hypertension[13] and find application in other therapeutic areas, such as heart failure, myocardial infarction, and diabetic nephropathy. The knowledge on structural requirements for ACE inhibition has been derived from a multitude of SAR studies.[14,15] This led to the determination of a minimal set of functional groups and their common 3D geometry for ACE binding[16] as the basis for 3D-QSAR studies.[17]

Only recently, X-ray crystal structures of human ACE have been solved in complex with the inhibitors captopril,[18] enalapril,[18] and lisinopril[19] (PDB codes 1uze, 1uzf, and 1o86). These X-ray structures confirmed the proposed interactions of the ligands with the catalytic zinc at the active site encompassing the thiol group of captopril and the carboxylates of enalapril and lisinopril (Figure 1). Lisinopril furthermore occupies the S1′ pocket and interacts with two acidic amino acids via its primary amine, while enalapril only directs a methyl group into this pocket. The central carbonyl group of all ACE inhibitors is positioned by two strong hydrogen bonds to His513 and His353. In addition, one oxygen of the carboxylate from the inhibitor proline scaffold is engaged in interactions to Tyr520, Gln281, and Lys511, while the other oxygen is directed to the surrounding water molecules. Finally, a hydrophobic pocket with Phe512 and Val518 accommodates the amino terminal lipophilic inhibitor moieties.

2.1.2. Feature Tree Models of ACE Inhibitors. The three ACE inhibitors captopril, enalapril, and lisinopril, for which the details of structural alignment and molecular interactions are known from X-ray structures, were used to generate a MTree model.

Indeed, the resulting model (ACE), based on three individual feature trees, is characterized by the correct matching of related functional groups, in agreement with the superposition from X-ray structure analysis. The model and the individual feature trees are shown in Figure 2. The color codes of the hexagons indicate the chemical similarity of aligned functional groups. Red hexagons indicate identical groups; orange and yellow indicate similar groups, while green hexagons indicate no significant correspondence of aligned groups. Inspecting red, orange, or yellow hexagons thus identifies a set of features essential for binding, which is termed a *topological pharmacophore*.

In the ACE MTree model, the proline carboxylate is located in the upper right area, showing that all inhibitors with experimental binding mode are identical in this area interacting with Tyr520, Gln281, and Lys511. The model correctly aligns all compounds on the basis of corresponding pharmacophoric points. The proline scaffold including its carbonyl group are nearly identical in all inhibitors and mapped to corresponding hexagons in the final model. Furthermore, different zinc binding groups (thiol in captopril; carboxylate in enalapril, lisinopril) are mapped to the same hexagon in the lower left area, while the yellow color indicates a reduced but still significant chemical similarity of these aligned functional groups. Analysis of the corresponding interaction patterns, however, indicates favorable interactions of the receptor (Zn^{2+} ion) with acceptor functionalities in the ligands. More significant differences are observed for the alignment within other enzyme subpockets. Different parts of the inhibitors are correctly assigned to different subtrees of the final model, corresponding to substituents directed toward the S1′ pocket (the butylamine side chain of lisinopril

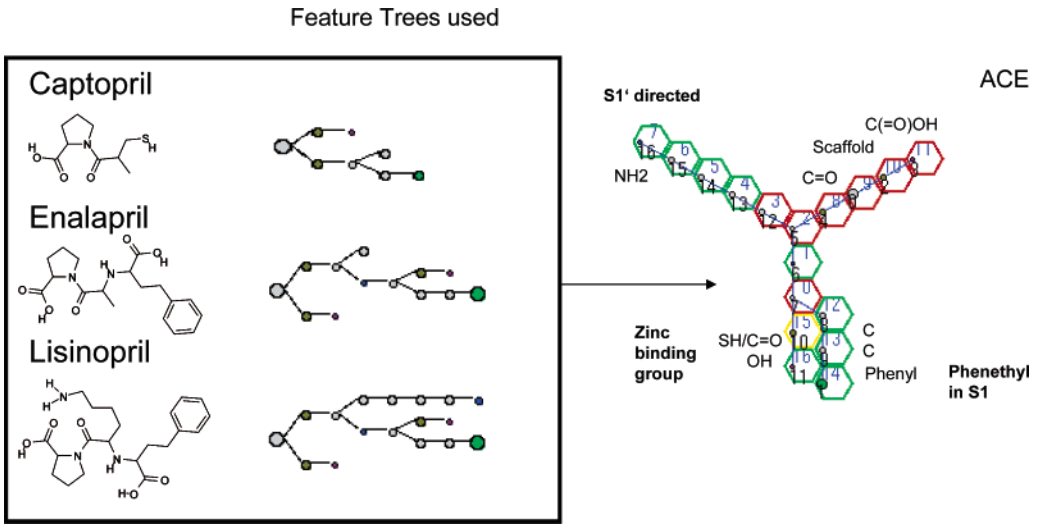


Figure 2. Chemical structures, individual feature trees for three ACE inhibitors and the derived ACE MTree model. The feature trees of each model are aligned on a hexagonal grid. Neighboring nodes or subtrees are placed in adjacent tiles. The nodes of whole subtree matches are grouped into a single tile. The color code of individual hexagons in the model indicates the chemical similarity of aligned functional groups. Red hexagons indicate identical groups; orange and yellow indicate similar groups, while green does not indicate any significant correspondence.

mapped to upper left hexagons) or the lipophilic S1 pocket (the phenethyl moiety of lisinopril and enalapril mapped to lower right hexagons).

Hence, the model correctly superimposes all essential polar functional groups from three ACE inhibitors and rapidly detects essential features for biological affinity without taking the ligand 3D structure or an a priori assumption about pharmacophoric groups into account. This model consequently was used as query (ACE) for retrospective virtual screening.

2.1.3. Retrospective Virtual Screening for ACE Inhibitors. The ACE MTree model was subjected to a retrospective virtual screening study using the candidate database (see section 4.5).

The virtual screening run was carried out using the *best fit score* for comparing feature trees of database molecules to the MTree model. In addition, individual feature trees used for model building were subjected to similarity searching using the match search algorithm described previously.[9]

In Figure 3, the enrichment curve for the ACE model is shown in comparison to results from individual feature tree searches. On the *x*-axis the fraction of the ranked candidate database is plotted against the percent of labeled actives found at this fraction on the *y*-axis. As usual, the candidate database is ranked by similarity coefficients to the model or query used for screening. The MTree model outperforms individual feature tree similarity searches in terms of the percentage of actives [labeled as ACE inhibitors in the World Drug Index (WDI)] recovered at small percentages of the database screened. A detailed comparison performed on the level of 1%, 5%, and 10% of the candidate database screened is given in Table 1. The first percent of ranked compounds from the candidate database contains 62% of the known ACE inhibitors for the MTree model, corresponding to an enrichment factor of 70.9. Screening 5% of the candidate database would recover 85% actives, corresponding to an enrichment factor of 17.3, while the top-ranked 10% contains 89% actives, which corresponds to an enrichment rate of 9.7.

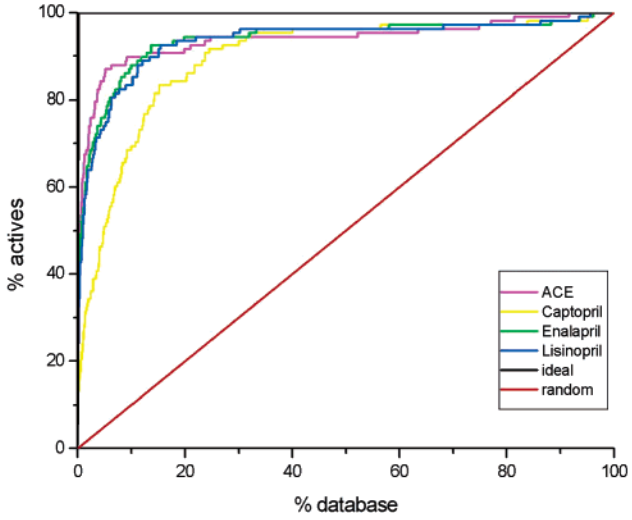


Figure 3. Enrichment curve for ACE in comparison to results from individual feature tree searches. The fraction of the ranked candidate database on the *x*-axis is plotted against the percent of labeled actives at this fraction on the *y*-axis. The candidate database is rank-ordered by similarity to the model or query.

Table 1. Percentage of Actives Identified and Enrichment Factors for MTree Models for Different Fractions of the Candidate Database Screened

	ACE		α1A receptor			
	MTree model		MTree model		Catalyst model	
fraction of database screened (%)	% act.	enrichment factor	% act.	enrichment factor	% act.	enrichment factor
1	62	70.9	17	16.3	7.6	7.3
5	85	17.3	45	9.1	32.1	6.5
10	89	9.7	52	5.2	36	3.6

Enalapril performed best as single feature tree for similarity searching in comparison to the other two ACE inhibitors captopril and lisinopril. This might be a consequence of the fact that this inhibitor is comparable in size to the actives in the candidate database. However, results from single feature tree searches are still

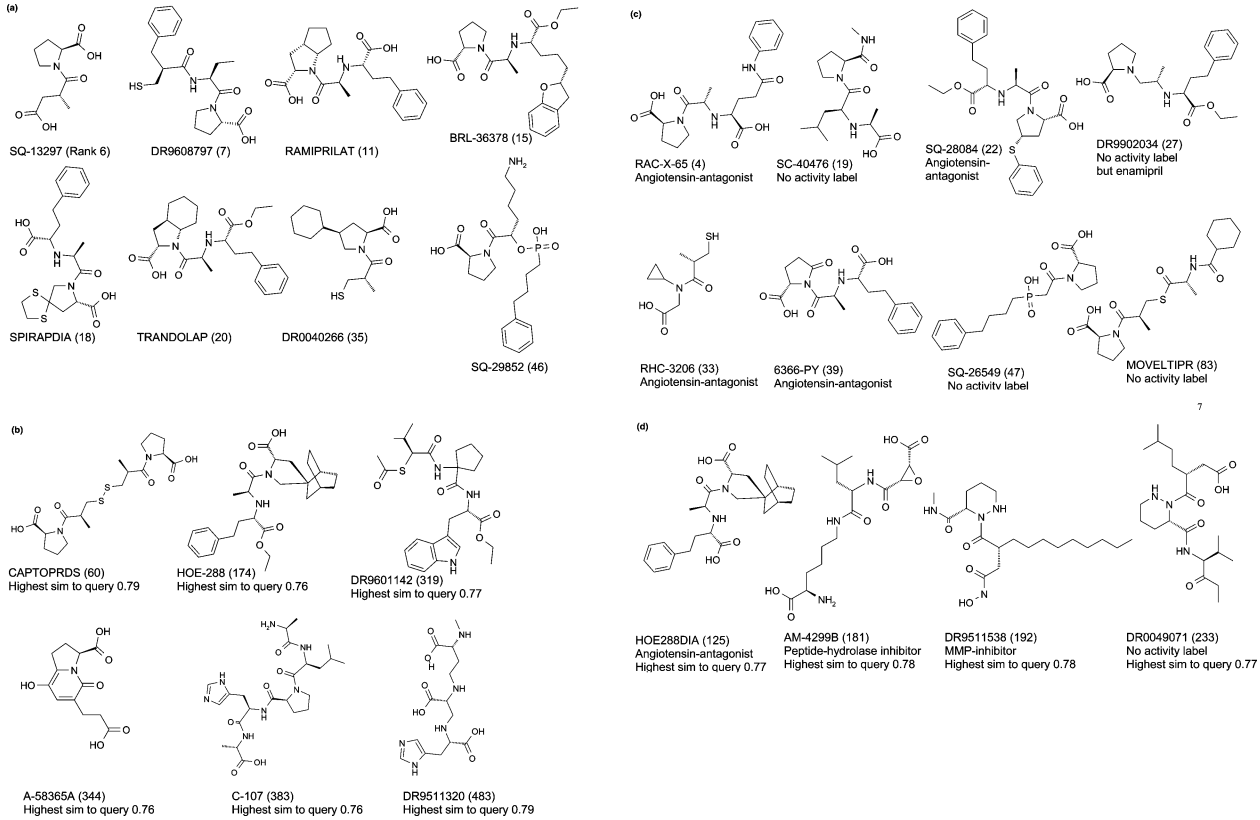


Figure 4. Selected chemical structures from retrospective virtual screening experiments for MTree models for ACE. (a) Top scoring inhibitors from the MTree model. For each entry, the WDI registry number is listed and its rank after virtual screening is indicated in brackets. (b) Molecules found by the MTree model, but not by individual feature tree searches. (c) Molecules found by the MTree model, but not labeled as actives. For each entry, the WDI registry number, its rank, and its activity are indicated. (d) Molecules found by the MTree model, not by individual feature tree searches, and not labeled as actives.

inferior to those obtained using the MTree model in terms of actives retrieved at only a few percent of the candidate database screened (cf. Figure 3 and Table 1). Figure 4 summarizes the chemical structures of some hits from retrospective virtual screening experiments based on the MTree model to illustrate their performance in identifying true ACE inhibitors and closely related database entries.

In Figure 4a, a representative selection of eight top scoring ACE inhibitors retrieved by the ACE model is shown. For each entry, the WDI registry number is listed and its rank after virtual screening is indicated in brackets. Although the three ACE inhibitor query molecules are similar and based on proline to orient essential pharmacophoric groups in the metalloproteinase binding site, the hits are characterized by a more diverse range of structural elements at the central scaffold and the pharmacophoric groups. This suggests that feature trees and the MTree model approach go beyond a simple 2D-substructure-based similarity approach.

Each of the molecules in Figure 4a was also retrieved by different individual feature tree searches using individual members of the MTree model. However, our interest was to investigate to what extent the MTree model is able to find molecules that are not picked by feature tree searches based on individual molecules. On the basis of experience from internal projects, we have set a cutoff value between 0.8 and 0.85 (feature tree similarity based on match search algorithm) to count a molecule as a hit in individual feature tree searches. There are several examples of ACE inhibitors found only using the MTree model. Six of them shown in Figure 4b were less similar than 0.79 to any of the three query molecules, but they were still picked by the MTree model. They all exhibit a larger structural variation compared to individual query molecules.

A representative selection of eight molecules retrieved by the MTree model, but not labeled as "ACE-inhibitors", is shown in Figure 4c. A closer look reveals that four of the most similar compounds are labeled as "angiotensin antagonists". This pharmacological family encompasses ACE inhibitors, among other compounds, which suggests true ACE inhibitory activity for some hits. One molecule, DR9902034 (enamipril), has no activity label, while it corresponds to enalapril (see Figure 1) with a reduced alanine carbonyl group.[20] Three other molecules were not labeled as ACE inhibitors, although they share remarkable similarities with typical ACE inhibitors in terms of scaffold and pharmacophoric groups (cf. Figures 1 and 4a).

While those molecules from Figure 4c were retrieved by feature tree searches using different individual queries, our interest was to identify those molecules that are only identified by the MTree model and not using any individual feature tree query. Four representative examples are shown in Figure 4d. One entry is labeled as "angiotensin antagonist", suggesting some activity as ACE inhibitor. Two others are labeled as peptide-hydrolase inhibitors and MMP inhibitors and thus fall into the same target family. For the final entry, no activity label is present, while again the chemical similarity to the query molecules is obvious. These

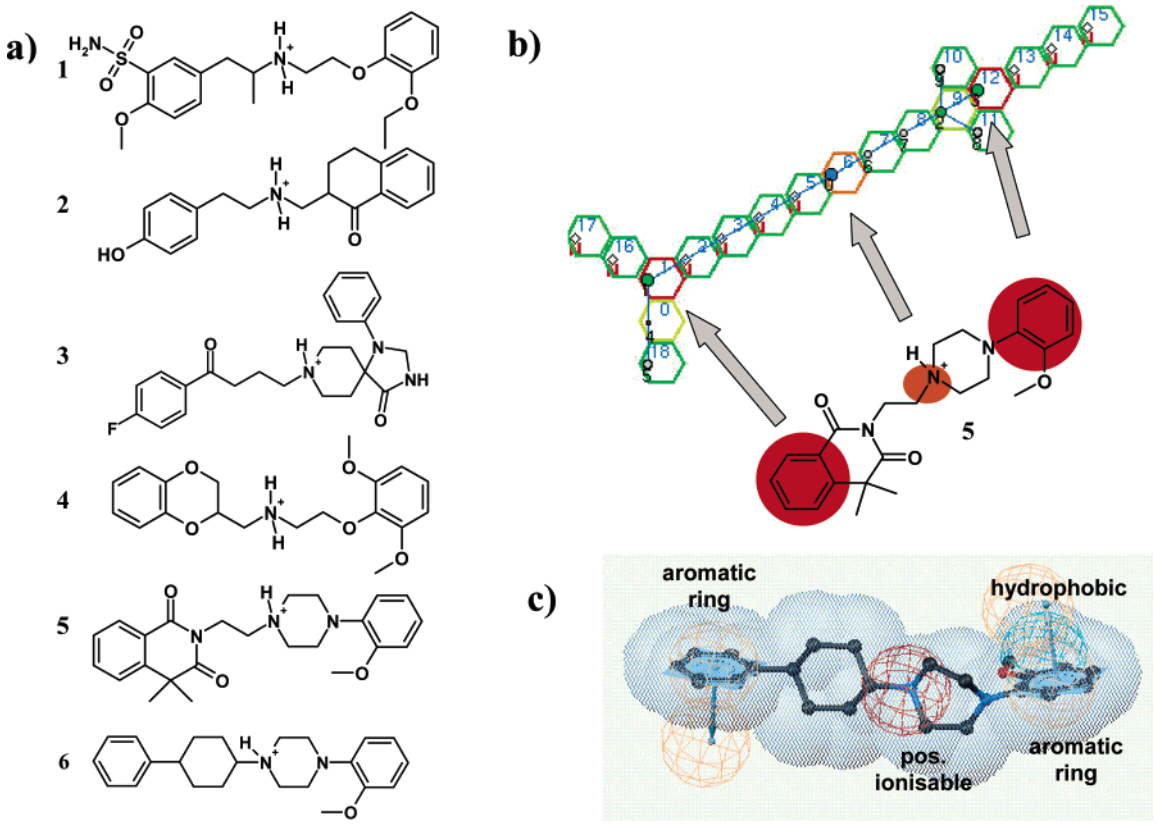


Figure 5. (a) 2D structures of known α1A antagonists used for the derivation of the MTree model of α1A receptor. (b) Derived MTree model of α1A receptor and compound **5**. The conserved functional groups (i.e., the positively ionizable nitrogen and two aromatic moieties) are indicated. (c) Derived 3D-pharmacophore model generated with Catalyst for the α1A receptor. Mapping of compound **6** onto this model is shown.

analyses illustrate that MTree models retrieve not only obvious ACE inhibitors, which might be found by subsequent individual searches, but also entries that are structurally less related to any of the query molecules and thus would have been overlooked in subsequent single feature tree searches.

2.2. α1A Antagonists. The α1 adrenergic receptors belong to the family of G-protein coupled receptors (GPCRs). They are involved in blood pressure maintenance, modulating vascular muscle tone. They are subdivided into the α1A, α1B and α1D adrenoreceptor subtypes.[21] Antagonists of the α1 adrenergic receptors such as indoramin and prazosin are employed as antihypertensive agents. In addition, α1A antagonists such as alfuzosin and prazosin are thought to be effective in the management of benign prostatic hypertrophy.

2.2.1. Molecular Recognition at the α1A Receptor. Due to the fact that GPCRs are membrane-bound proteins, their expression, purification, crystallization, and structure determination remain a major enterprise. Although crystal structures of protein−ligand complexes for the α1A receptor are not available, details of molecular recognition were derived from experimental data, for example, through mutational studies and comparative affinity determinations based on ligand binding.[22,23] It is generally accepted for all biogenic amine binding GPCRs that Asp3.32 (according to the Ballesteros−Weinstein nomenclature) located in the transmembrane helix TM3 is involved in binding the biogenic amine group contained in all α1A ligands. It was suggested by Jacoby et al. that Asp3.32 is the central anchor point for α1A ligands flanked by different individual hydrophobic subpockets.[24]

2.2.2. Feature Tree and Catalyst Pharmacophore Models of the α1A Receptor. In a previous study, Klabunde et al.[25] generated 3D-pharmacophore models (using Catalyst, Accelrys Inc., San Diego, CA) for the α1A and further biogenic amine receptors. These models describe the key chemical features present within these biogenic amine antagonists and rationalize the binding of compounds at the referring receptors. In a virtual screening experiment, an α1A pharmacophore model (Figure 5c) was shown to recognize a large fraction of known α1A antagonists. To allow for a comparison of the performance of the MTree methodology to a well-established technique in the field, we generated an MTree model based on the same set of ligands that was used for the generation of the above-mentioned Catalyst pharmacophore model. Six α1A antagonists were used for the generation of these models (see Figure 5a). Again, the chemical similarities of aligned groups in the MTree model (Figure 5b) are indicated by the color codes of the hexagons, with red hexagons representing identical groups and the orange code indicating highly similar nodes. The resulting topological pharmacophore is in good agreement with the 3D-pharmacophore model generated with Catalyst and our general knowledge about molecular recognition at the α1A receptor. A central positively ionizable group (represented by the orange hexagon) is interacting with Asp3.32, and two hydrophobic subpockets are addressed by the red-colored hydrophobic/aromatic groups. These

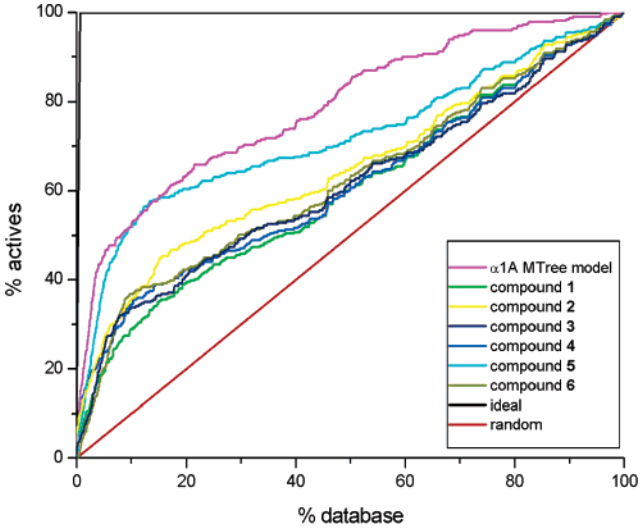


Figure 6. Enrichment curves for the MTree model of the α1A receptor and the individual feature tree searches.

groups are connected to the central positively ionizable group by linkers of variable length.

2.2.3. Retrospective Virtual Screening for α1A Antagonists. The α1A MTree model was used for virtual screening of the filtered version of the candidate database. In addition, individual feature tree searches were performed. Furthermore, the 3D-pharmacophore model generated with Catalyst was used for virtual screening of the same candidate database. Those compounds from the screening set, which were in agreement with the pharmacophore hypothesis, were ranked according to their *fit values*. For further details of Catalyst model generation and conformer generation of the screening data set, the reader is referred to ref 25. As mentioned above, the central biogenic amine group of α1A ligands establishes a hydrogen bond with Asp3.32 of the α1A receptor. Since tertiary nitrogens are not recognized as hydrogen-bond donors by the feature tree, we converted all ligands of the training set (as depicted in Figure 5) and the candidate database into their protonated form.

Enrichment curves for the α1A MTree model and the searches of the individual feature trees are shown in Figure 6. The number of active compounds and enrichment factors for the MTree model, the Catalyst pharmacophore model, and the individual feature trees at 1%, 5%, and 10% of the screened database are given in Table 1. Here, the MTree model shows the best enrichment factors. It outperforms the individual feature tree searches and the 3D-pharmacophore-based screening. It should be mentioned that several compounds retrieved among the top hits are labeled as ligands targeting other biogenic amine binding GPCRs than the α1A receptor. Thus (cross-) activity toward the α1A receptor is likely suggesting even higher enrichments than calculated. Using the α1A MTree model as query, we retrieved 167 α1A antagonists with a similarity score of at least 0.80, whereas the search based on the best individual reference tree (derived from compound **5**) only yields 123 compounds. Figure 7 shows four representative examples retrieved with the α1A MTree model but not retrieved by any of the six individual query molecules. In particular, S-9874 (see Figure 7) is interesting. Here, the hydrogen-bond donor supposed to

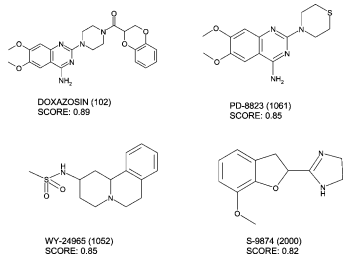


Figure 7. Examples of molecules that were identified as a hit (with similarity scores > 0.80) by the MTree model of α1A receptor but by none of the individual feature tree queries.

interact with Asp3.32 is linked to a 4,5-dihydroimidazole nitrogen, which does not occur in any of the six individual query molecules. Furthermore, whereas each individual query molecule comprises two aromatic groups (addressing two different subpockets), the model is able to identify α1A binders that only address one of these putative subpockets.

3. Conclusions and Outlook

Searching for new or alternative lead structures based on a set of known actives is a very important task in lead discovery and optimization. Nevertheless, the variety of available methods is limited. While single compound similarity searching neglects important information on features common to multiple active molecules, traditional QSAR approaches do not always allow for extrapolation to novel structural classes. On the other hand, ligand-based pharmacophore searching typically contains many manual steps for model building and requires 3D structure information for larger databases.

In this paper, we presented and validated a novel approach for ligand-based virtual screening capable of dealing with a set of known actives simultaneously. The actives are used to create a so-called MTree model, which is conceptually similar to a topological pharmacophore. Due to the low dependence on chemical substructures, we believe that the MTree model is especially useful for the identification of alternative novel molecular scaffolds or chemotypes.[10] The resulting model is conformation-independent and recognizes common features and functional groups among actives. While distances between these features are implicitly modeled to the bond paths, angular relationships as well as stereochemistry are not considered in model generation.

Two examples (ACE and the α1A receptor) were selected for validation by retrospective virtual screening on the basis of MTree models versus similarity searches. The resulting models are in very good agreement with the available X-ray structural information and known pharmacophores, underscoring the quality and possibility for chemical interpretation in these models. Enrichment factors between 71 and 16 were obtained after investigating the first percent of the ranked candidate database. In addition, the MTree model outperformed individual feature tree similarity searches, which have been reported to result in higher enrichments than frequently used linear descriptors.[9] Additionally, for the α1A receptor, the MTree approach showed slightly better enrichments than a 3D-pharmacophore-based virtual screening with Catalyst. Therefore, the auto-

matic detection of common features in MTree models opens the road for prospective virtual screening.

Like all other descriptors used in ligand-based molecular design, feature trees have their strengths and weaknesses. Clearly, the modular, nonlinear construction of the descriptor belongs to its strengths, allowing for an alignment-based comparison of molecules as well as model building from multiple active compounds, as shown in this paper. Due to the tree structure, feature trees are inappropriate if the active compounds contain macrocycles or highly bridged ring systems. So far, stereochemistry cannot be considered and neither can a detailed arrangement of functional groups in heterocycles. While macrocycles are incompatible with feature trees in general, the latter two aspects are not.

Due to the automatic model building and the consideration of multiple compounds, MTree models are especially interesting for secondary screening and high-throughput screening data analysis. The technology allows for introducing weights of individual nodes of the MTree model, giving differing importance to those features. Such weights could be derived on the basis of experimental affinity data. Hence, a new view from multiple active compounds to unravel relevant functional groups emerges from these models.

4. Computational Details

Due to the frequent application of similarity-based methods in molecular design, a large variety of molecular descriptors has been suggested (for recent reviews, see refs 26, 27). Most of them are alignment-free, i.e., molecules are compared without calculating an assignment of parts of one molecule to parts of the other. These methods allow for very time-efficient comparisons and are therefore well-suited for virtual screening. It is difficult, however, to create a model by combining information from several known active compounds. At the other extreme, three-dimensional descriptors require a three-dimensional alignment, which involves dealing with the conformational space of the compounds, making the alignment problem difficult to solve. To create models from multiple known actives efficiently and reliably, we developed the feature tree descriptor,[9] a compromise between the classical two-dimensional and three-dimensional descriptors. The comparison with feature trees does not depend on molecular conformations, making it easier to compute. However, the descriptor is based on an alignment and is therefore suited for creating models from multiple compounds.

4.1. Feature Trees. A feature tree[9] is created from a molecular graph by shrinking rings and single nodes. Also, end-standing atoms are merged with their neighbors. Each node is labeled with a chemical and steric feature computed from the molecular fragment the node represents. The steric feature is the van der Waals volume and the number of ring closures, and the chemical feature is a profile of molecular interactions the fragment can form with surrounding molecules. Figure 8 shows an example of a molecule and its feature tree.

From a computational point of view, a feature tree is a node-labeled, unrooted tree. To compare two such trees, an alignment or matching algorithm is necessary to decide which part of the first tree should be matched to which part of the second one. Obviously, this matching of trees should obey the tree-topology, i.e., the set of matches should have the same relative arrangement in both trees. Tree matching is a well-studied problem in computer science,[21,28–30] however, all known algorithms perform a node-to-node mapping. This is inappropriate for feature trees, since a matching has to be balanced with respect to the size of the molecular parts. Since a feature tree node may correspond to a full ring system or a chain atom only, it is necessary to match sets of nodes—so-called *subtrees*—to each other.

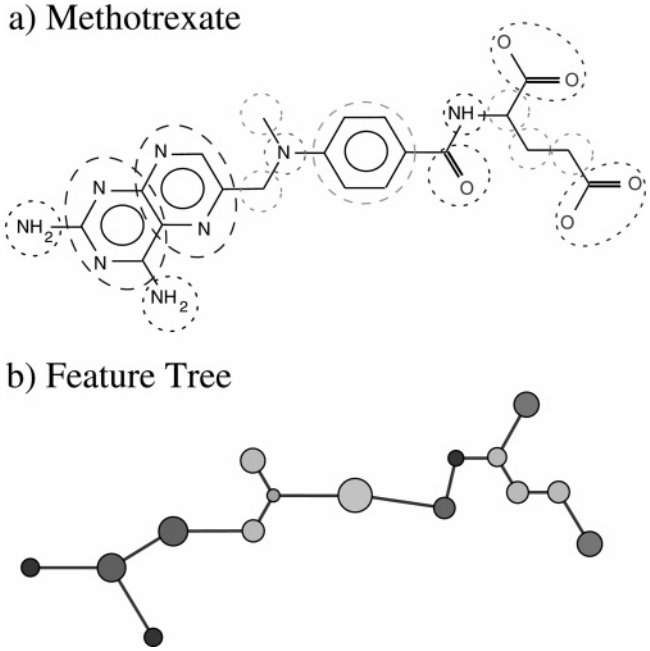


Figure 8. Generation of feature trees. Small molecular fragments correspond to a single node. Rings are merged into one node. Edges connect neighboring fragments (sharing atoms or bonds). The feature tree and the molecule are shown in the same orientation. Hydrophobic fragments are light-gray, hydrogen-bond donors are dark-gray, and hydrogen-bond acceptors are gray. Bonds connecting neighboring fragments are represented by edges.

For the feature tree software, two algorithms for subtree matching have been developed. The split-search algorithm is a divide & conquer scheme, dividing the molecules in a hierarchical, top-down fashion. The match-search algorithm is a dynamic programming approach creating a matching in an incremental fashion. For a detailed description of the algorithms, we refer to ref 9. Both algorithms are well-suited for pairwise comparisons of molecules; however, both are problematic for constructing multiple compound models. The split-search algorithm allows for unmatched parts between the matched ones (so-called inner-NIL matches) but is unable to score them (a large unmatched part results in the same similarity value as a small unmatched part). The match-search algorithm is not able to create inner NIL-matches at all. Although this is unproblematic for pairwise comparisons, handling inner-NIL matches is of importance for model building. Inner-NIL matches allow for the correct handling of two subsets of molecules, in which molecules within one subset share a common feature that molecules in the other subset do not have.

In the following, we will briefly describe the enhancements made in order to allow for the creation of multiple feature tree models (MTree models) for virtual screening.

4.2. Comparing Two Feature Trees under Consideration of Inner-NIL Matches. The *dynamic-match-search algorithm* is a novel pairwise comparison algorithm allowing creating and scoring inner-NIL matches. Here, only a rough outline of the algorithm is given. A detailed description can be found in ref 11.

The overall goal of the algorithm is the calculation of a matching between subtrees that maximizes the similarity value calculated as the volume-weighted sum of the similarity values of all matches:

$$S(T_1,T_2) = S(M) = \frac{\sum_{m_i \in M} \mathrm{sim}(m_i)\,\mathrm{size}(m_i)}{(\omega \min\{\mathrm{size}(T_1), \mathrm{size}(T_2)\} + (1-\omega)\max\{\mathrm{size}(T_1), \mathrm{size}(T_2)\})} \quad (1)$$

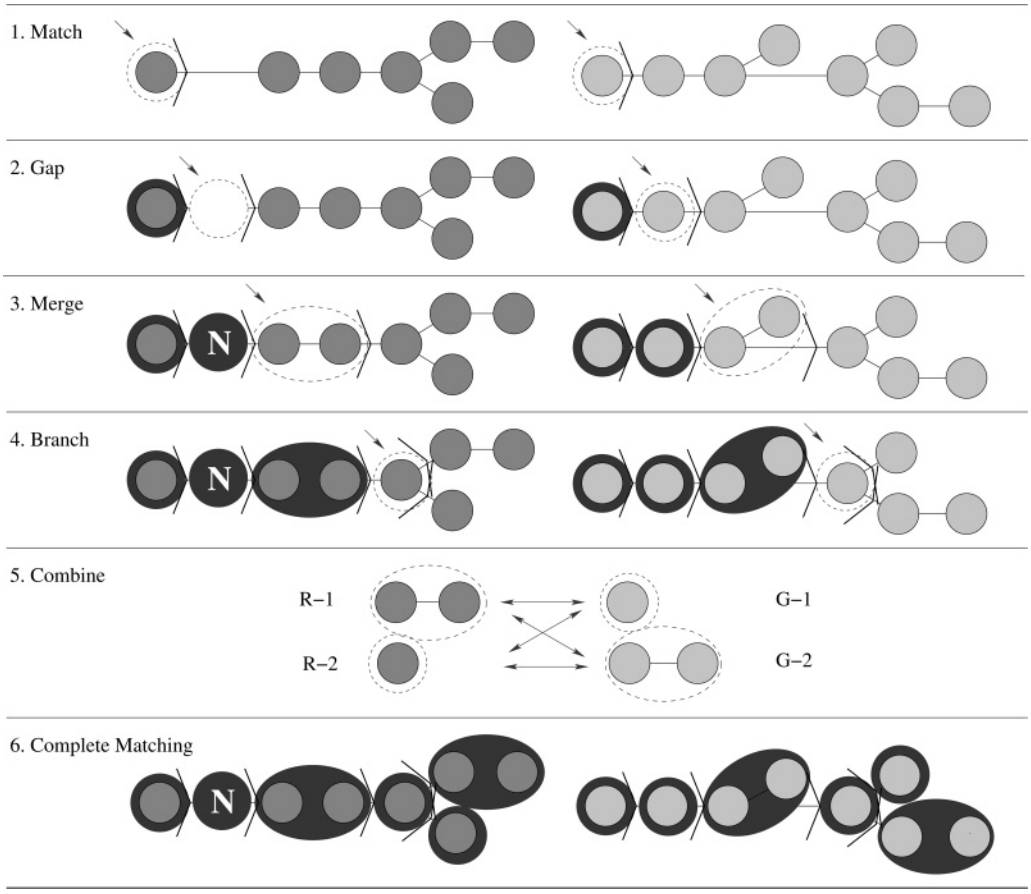


Figure 9. Comparing two feature trees by mapping nodes and subtrees onto each other. The matching is constructed from the left (leaves) to the right (root) by adding a new match in each step.

The function sim calculates the similarity and the function size calculates the average volume of the matched subtrees. The parameter ω allows for tuning the similarity measure toward local similarity (ω close to 1) and global similarity (ω close to 0). Note that the above formula can easily be extended to calculate similarity values of multiple feature trees.

Finding the optimal matching M with the highest similarity value is a combinatorial optimization problem that is efficiently solvable with dynamic programming.[31] Dynamic programming breaks a problem into smaller subproblems, here the matching of smaller subtrees of the two feature trees. The optimal matching of two larger portions of the feature trees can be calculated by combining the results achieved for smaller portions. Since only a limited number of possible subtree matches can occur, similarity values for these matches can be stored in a matrix, the so-called *dynamic programming matrix*, for reuse during the calculation. The calculation itself proceeds bottom-up. First, all combinations of leaves of the feature trees are compared. Then, combinations of slightly larger subtrees are calculated. This process is iterated until the similarity value for the whole trees can be accessed.

The dynamic-match-search algorithm is best described by the operations performed in order to extend a matching (see also Figure 9). Let us assume that two trees have been compared already up to a certain pair of edges (one in each tree), called a *split*. The matching can now be extended as follows:

(1) A *match operation* forms a new match between the nodes adjacent to the previous match.

(2) A *gap operation* skips some part of one tree forming an inner NIL-match.

(3) A *merge operation* forms a new match between two subtrees adjacent to the previous match. The topologies of the subtrees do have to fit. The operation therefore introduces a certain degree of fuzziness into the matching process. The maximal size of the subtrees considered in a merge operation is limited.

Due to the bottom-up order of calculation, the algorithm can access the relevant information inside the dynamic programming matrix and therefore directly select the operation that results in the highest similarity value. In the case of a branching node, the dynamic-match-search algorithm has to decide which outgoing edge of one tree is to be mapped to which outgoing edge of the other tree. For all combinations of outgoing edges, the similarity value can be derived from the dynamic programming matrix. Finding the best assignment of edges is related to a well-studied problem in computer science (maximum weighted bipartite matching)[32] and can be done efficiently.

4.3. Forming Multiple Feature Tree Models. With an algorithm for comparing feature trees in hand, we can now describe how to generate multiple feature tree (MTree) models. On the basis of the matching calculated for the comparison of two trees, a new tree combining the information from both input feature trees can be created. The nodes represent the matches containing the features of the mapped subtrees. The edges are formed by following the topologies of the input feature trees. The resulting tree is called an MTree model. Since it has the same structure as a feature tree, it can be compared with other MTree models or feature trees using the same algorithm as for feature trees, i.e., the dynamic-match-search algorithm.

To generate an MTree model from more than two feature trees, the dynamic-match-search algorithm can be applied in a hierarchical manner. We developed an efficient heuristic for this task.

The strategy is to incrementally add single molecules to the model (starting with two molecules). This step is iterated. In each step the molecule that is the most similar molecule to the model is chosen. Thus, there are $n - i$ comparisons in the ith step and $n - 1$ steps are needed altogether until a single MTree model remains.

MTree models constructed this way can be used for virtual screening purposes. By merging the information of the underlying trees into an MTree model, virtual screening can be done by simple pairwise comparisons. Thus, we are applying local data fusion to each match and do not have to compare the ranks of several runs comparing individual query molecules to each database molecule. A further advantage is that the matches can be weighted by the local similarity of the corresponding fragments.

4.4. Multiple Feature Tree Scoring Schemes. There are two applications of MTree models that require an extension of the pairwise scoring scheme sim. First of all this is the construction of MTree models and the comparison of a model with compounds for virtual screening purposes. There are several ways to compare a single subtree t_q (from a query T_q) to a set of n subtrees (t_1, ..., t_n), a match of the MTree model. The easiest one is to compare the query subtree to each subtree of the set (i.e. the model) and take the highest score of each match, which is called *best fit scoring*. The total score $S(T_q, T_1, ..., T_n)$ is again a combination of the local similarities, as shown in formula 1:

$$\text{best fit score: } \mathrm{sim}(t_q,t_1,...,t_n) = \max_i \{\mathrm{sim}(t_q,t_i)\} \quad (2)$$

Alternatively an average score is defined as the mean of the pairwise similarity scores:

$$\text{average score: } \mathrm{sim}(t_q,t_1,...,t_n) = \frac{1}{n}\sum_i \mathrm{sim}(t_q,t_i) \quad (3)$$

The *average score* is suited to emphasize a common scaffold of a model. The best fit score allows for finding new molecules that contain combinations of fragments of the molecules in the model (i.e. having a new scaffold).

4.5. Data Set for Model Generation and Virtual Screening. The candidate database was extracted as a drug-like subset from the WDI.[33] To reject compounds with undesirable properties for oral delivery, limits on essential physicochemical properties have been applied. Only compounds with up to nine rotatable bonds, a molecular mass of less than 600 Da, up to 8 donor atoms, and up to 15 acceptor atoms were considered. In addition, compounds with a low (<4) or high number (>25) of feature tree nodes were rejected, since multiple feature model searches perform best if the number of nodes in the query and the candidate molecules are of comparable sizes (no data given). The filtered candidate database contained 47 691 compounds, comprising 331 α1a and 108 ACE inhibitors.

It should be noted that the activity labels in the candidate database derived from the WDI might be incomplete or only roughly indicating the molecule's true mode of action. Furthermore, not all compounds have been tested for ACE, or α1A receptor inhibition. It cannot be ruled out that compounds labeled as active for a different target also show cross-activity against one of the targets considered in this study. Hence, there always is some uncertainty about the real activity of virtual hits.

For model generation, the training molecules were combined into an MTree model by the incremental build-up strategy using the dynamic-match-search algorithm. The resulting models for each data set were used as queries for virtual screening in the candidate database. The scores describing the similarity of each data molecule to either individual query molecules or the MTree model were computed using the best fit score. The dynamic-match-search algorithm was used for virtual screening with the MTree model, while the match search algorithm was used for individual feature tree queries.

Acknowledgment. This work has been done in the context of the LeadID project, which aimed at developing computer methods for high-throughput screening data analysis. The LeadID project is a collaboration among Aventis Pharma Deutschland GmbH, BioSolveIT, and FhI SCAI. We thank Thomas Klabunde for providing the Catalyst pharmacophore model of the α1A receptor.

References

(1) Böhm, H.-J.; Schneider, G. Virtual Screening for Bioactive Molecules; Wiley-VCH: Weinheim, 2000.
(2) Oprea, T. I.; Matter, H. Integrating Virtual Screening in Lead Discovery. *Curr. Opin. Chem. Biol.* **2004**, *8*, 349−358.
(3) Lengauer, T.; Lemmen, C.; Rarey, M.; Zimmermann, M. Novel technologies for virtual screening. *Drug Discovery Today* **2004**, *9*, 27−34.
(4) Johnson, M. A., Maggiora, G. M., Eds. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York, 1990.
(5) Grethe, G.; Moock, T. E. Similarity searching in REACCS. A new tool for the synthetic chemist. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 511−520.
(6) *UNITY Chemical Information Software Version 4.0*; Tripos Inc., St. Louis, MO, 1994.
(7) Baringhaus, K.-H.; Hessler, G. Fast Similarity Searching and Screening Hit Analysis. *Drug Discovery Today: Technol.* **2004**, *1* (3), 197−202.
(8) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem. Int. Ed.* **1999**, *38*, 2894−2896.
(9) Rarey, M.; Dixon, J. S. Feature trees: A new molecular similarity measure based on tree matching. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 471−490.
(10) Good, A. C.; Hermsmeier, M. A.; Hindle, S. A. Measuring CAMD technique performance: A virtual screening case study in the design of validation experiments. *J. Comput.-Aided Mol. Des.***2004**, *18*, 529−536.
(11) Zimmermann, M. Rechnerunterstütze Analyse von HTS-Daten. In *Mathematisch-Naturwissenschaftliche Fakultät*; Rheinische Friedrich-Wilhelms-Universität Bonn: Bonn, 2003.
(12) Sprague, P. W. Automated chemical hypothesis generation and database searching with Catalyst. *Perspect. Drug Discovery Des.* **1995**, *3*, 21−33.
(13) Cushman, D. W.; Ondetti, M. A. Design of angiotensin converting enzyme inhibitors. *Nat. Med.* **1999**, *5*, 1110−1112.
(14) Hangauer, D. G. Computer-Aided Design and Evaluation of Angiotensin Converting Enzyme Inhibitors. *Computer-Aided Drug Design: Methods and Applications*; Marcel Dekker: New York, 1989; pp 253−531.
(15) Wyvratt, M. J.; Patchett, A. A. Recent Developments in the Design of Angiotensin-Converting Enzyme Inhibitors. *Med. Res. Rev.* **1985**, *5*, 483−531.
(16) Meyer, D.; Naylor, C. B.; Motoc, I.; Marshall, G. R. A unique geometry of the active site of angiotensin-converting enzyme consistent with structure−activity studies. *J. Comput.-Aided Mol. Des.* **1987**, *1*, 3−16.
(17) DePriest, S. A.; Mayer, D.; Naylor, C. B.; Marshall, G. R. 3D-QSAR of Angiotensin-Converting Enzyme and thermolysin Inhibitors: A Comparison of CoMFA Models Based on Deduced and Experimentally Determined Active Site Geometries. *J. Am. Chem. Soc.* **1993**, *115*, 5372−5384.
(18) Natesh, R.; Schwager, S. L. U.; Evans, H. R.; Sturrock, E. D.; Acharya, K. R. Structural Details on the Binding of Antihypertensive Drugs Captopril and Enalaprilat to Human Testicular Angiotensin I-Converting Enzyme. *Nature* **2004**, *421*, 551−554.
(19) Natesh, R.; Schwager, S. L. U.; Sturrock, E. D.; Acharya, K. R. Crystal structure of the human angiotensin-converting enzyme-lisinopril complex. *Nature* **2003**, *421*, 551−554.
(20) Schoenmakers, R. G.; Stehouwer, M. C.; Tukker, J. J. Structure-transport relationship for the intestinal small-peptide carrier: Is the carbonyl group of the peptide bond relevant for transport? *Pharm. Res.* **1999**, *16*, 62−68.
(21) Zhang, K.; Shasha, D. Simple Fast Algorithms for the Editing Distance between Trees and Related Problems. *SIAM J. Comput.* **1989**, *18*, 1245−1262.
(22) Shi, L.; Javitch, J. A. The binding site of aminergic G protein-coupled receptors: The transmembrane segments and second extracellular loop. *Annu. Rev. Pharmacol. Toxicol.* **2002**, *42*, 437−467.
(23) Hamaguchi, N.; True, T. A.; Goetz, A. S.; Stouffer, M. J.; Lybrand, T. P. et al. Alpha 1-adrenergic receptor subtype determinants for 4-piperidyl oxazole antagonists. *Biochemistry* **1998**, *37*, 5730−5737.
(24) Jacoby, E.; Schuffenhauer, A.; Floersheim, P. Chemogenomics knowledge-based strategies in drug discovery. *Drug News Perspect.* **2003**, *16*, 93−102.
(25) Klabunde, T.; Evers, A. GPCR antitarget modeling: Pharmacophore models for biogenic amine binding GPCRs to avoid GPCR-mediated side effects. *Chembiochem.* **2005**, *6*, 876−889.
(26) Bender, A.; Glen, R. C. Molecular similarity: A key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204−3218.

(27) Maggiora, G. M.; Shanmugasundaram, V. Molecular Similarity Measures. *Methods Mol. Biol.* **2004**, *275*, 1−50.
(28) Tai, K.-C. The Tree-to-Tree Correction Problem. *J. ACM* **1979**, *26*, 422−433.
(29) Jiang, T.; Wang, L.; Zhang, K. *Alignment of Trees*—An Alternative to Tree Edit; McMaster University, Department of Computer Science and Systems: Hamilton, Canada, 1993.
(30) Gupta, A.; Nishimura, N. Finding Largest Subtrees and Smallest Supertrees. *Algorithmica* **1998**, *21*, 183−2000.
(31) Bellmann, R. *Dynamic Programming*; Princeton University Press: Princeton, NJ, 1957.
(32) Ahuja, R. K.; Magnati, T. L.; Orlin, J. B. *Network flows*; Prentice Hall: Englewood Cliffs, NJ, 1993.
(33) World Drug Index (WDI), Derwent Information, London WC2B 5DF, UK. http://www.derwent.co.uk.

JM050078W