Bioisosteric Similarity of Molecules Based on Structural Alignment and Observed Chemical Replacements in Drugs

Markus Krier and Michael C. Hutter*

Center for Bioinformatics, Saarland University, Campus Building C7.1, D-66123 Saarbruecken, Germany

Received September 19, 2008

The algorithmic concept used to assess the evolutionary relationship between protein sequences was adopted to the comparison of drug-like compounds. For this purpose, we have developed a method that uses the SMILES representation of the molecules to perform the corresponding pairwise alignment. The necessary exchange matrix was generated in an automated procedure that reflects the frequencies of chemical replacements in pharmaceutical substances. From the resulting alignment, the relationship between two molecules is computed as so-called bioisosteric similarity. This measure was used to perform virtual screening in several publicly available substance databases. We observed that databases containing drug-like compounds throughout showed higher bioisosteric similarities to the query compound than our reference set of confirmed nondrugs. Likewise, most actual drugs within a class show a higher bioisosteric similarity than the large background of other substances. The compounds obtained as highest ranking hits from the lead-like subset of the ZINC library showed distinct differences in comparison with corresponding results from a fingerprint-based similarity search, as well as the FTrees method. In particular the kind of chemical replacements as well as the conservation of substructures strongly reflect the underlying bioisosteric exchanges. Moreover, the bioisosteric similarity was used to assess the chemical diversity of the utilized drug classes and to compute the "average" molecule within the respective class.

INTRODUCTION

Screening of compounds is a deterministic and convergent process in drug development.[1] Corresponding experimental, as well as virtual, screening protocols have been established.[1,2] Choosing compounds is, however, a far less deterministic task. Although combinatorial chemistry allows us to cover certain regions of the vast chemical space in great detail, this increases in turn the haystack of substances to screen. The question is thus where to start. The empirical knowledge of medicinal chemists tells us that similar compounds are likely to have similar properties. Comparison of marketed drugs within a chosen therapeutic category will further corroborate this assumption. Therefore numerous approaches have been developed to express similarity between molecules.[3,4] The most obvious is search for the maximum common substructure that corresponds to a 2-dimensional similarity.[5] For this purpose, a series of approaches based on clique detection and subgraph similarities, such as the RASCAL approach have been used.[6,7] Extending the concept of finding maximum common substructures to 3-dimensional space yields pharmacophore models.[8] Comparisons based on molecular fields, such as the molecular electrostatic potential require likewise the 3D-structure of the molecules.[9,10] In turn, fragment-based approaches quantify similarity by comparison of present chemical groups or properties in the form of so-called fingerprints.[11−17] Eventually all of these representations have to be compared numerically. For this purpose, a number of distance metrics and similarity coefficients are commonly used.[3]

A drawback of most fingerprint descriptors is the loss of information regarding the spatial arrangement and neighborhood relationship of the underlying features, particularly in comparison to algorithms that search for the maximum common substructure. An intermediate approach is the so-called feature tree concept and its extensions.[18−20] There, molecules are decomposed into a small number of chemical features that are interconnected forming a tree-like topology. These features comprise physicochemical properties, such as hydrogen-bond donors, hydrogen-bond acceptors, positively and negatively ionizable groups, as well as hydrophobic and aromatic moieties. The comparison of molecules is thus transferred to the matching of corresponding subtrees. The concept of distinct chemical features is also exploited in the PhAST approach.[21] There, features are assigned to each non-hydrogen atom in the molecule, and a canonical sequence is derived that allows the application of alignment algorithms to obtain similarity scores. Alternative approaches using reduced molecular graphs have also been reported.[22−25] Furthermore, Weber et al. described an approach to calculate the similarity between functional groups in drug-like molecules for docking (Flexsim-R).[26]

All of the mentioned approaches make use of apparent chemical replacements that are more or less based on expert knowledge of medicinal chemistry. In the simplest case, only the maximum common substructure is regarded, but retaining charge and hydrogen-bonding properties is considered as foremost important. Elaborated concepts account for more specific biological properties that arise, for example, from the ADME optimization problem. For example, the according

* To whom correspondence should be addressed. Tel: +49 (0)681 30264178. Fax: + +49 (0)681 30264180. E-mail: michael.hutter@bioinformatik.uni-saarland.de.

10.1021/ci8003418 CCC: $40.75
Published on Web 04/29/2009

methyl-ester of a compound is better absorbed from the human intestine than the carbonic acid. Corresponding molecules are therefore called bioisosteres, in analogy to other isoster concepts. Comprehensive reviews regarding bioisosteric replacements can be found elsewhere.[27,28]

Comparisons and rankings of observed chemical replacements in drug-like molecules are commonly derived from the BIOSTER database that lists such pairwise exchanges.[12,29,30] Sheridan derived corresponding pairs from the MDL Drug Data Report.[31] Other approaches comprise the evaluation of privileged fragments from databases like the World Drug Index or corporate compound libraries.[32,33] There exist several commercial programs that carry out bioisosteric replacements on given molecules, or allow ranking of compounds. Among these are Drug Guru,[34] WABE,[35] BIOISOSTER,[36] BROOD,[37] as well as the "LASSO" tool of eHITS.[38] For most of these programs, sufficient information regarding the underlying scientific background and details on the algorithmic implementation are, however, limited for obvious reasons. In the results section we will therefore focus on the description of the functional differences between our approach and these programs.

Since most classical bioisosteric replacements are based on the exchange of single atoms and small groups, they are easily recognized. Nonclassical alterations, however, comprise less obvious exchanges, such as ring closure and ring-opening, or the replacement of small substituents by larger fragments.[27,28] Moreover, some types of bioisosteric exchanges might have been overlooked or still remain undiscovered by human experts. Therefore it is desirable to have an automated procedure to detect and assess all kinds of replacements across substance and drug classes. For this purpose, we have designed an approach that adopts the algorithmic concept used to assess the homology of amino acid sequences to chemical molecules.

METHODS

The methodological approach pursued in this study is summarized by the following overview:

(1) Explanation of the algorithmic concept and its realization.
(2) Alignment example and derivation of the bioisosteric similarity.
(3) Preselection criteria to handle larger sets of compounds.

Algorithm and Implementation. The pairwise similarity of proteins can be computed on the basis of the alignment of their amino acid sequences. Over time evolution has let to mutations in the DNA coding the proteins that give rise to different amino acid residues, as well as insertions and deletions. Since such substitutions do not occur with equal frequency, one is able to construct according substitution matrices for each possible combination of exchanges between residues.[39,40] For example, substitution of aspartate by glutamate is observed more frequently than by lysine which is attributed to the change in electric charge. The degree of relationship between two proteins can now be calculated by summation of the corresponding matrix entries for each matching residue pair, assuming an appropriate alignment of the sequences. Suitable algorithms for this purpose have been designed by Needleman and Wunsch, as well as by Smith and Waterman to obtain an optimal alignment by applying dynamic programming.[41,42]

Table 1. Atom Types Used in the Bioisosteric Substitution Matrix[a]

C	N	O	P	S
CX4&H3;R0	NX4&H3;R0	OX2&H1;R0	PX5&H0;R0	SX4&H1;R0
CX4&H2;R0	NX4&H2;R0	OX2&H0;R0	PX5&O4;R0	SX4&H0;R0
CX4&H1;R0	NX4&H1;R0	OX1&H0;R0	PX4&O3;R0	SX4&O3;R0
CX4&H0;R0	NX4&H0;R0	OX2&H0;R	PX4&O2;R0	SX4&O2;R0
CX3&H2;R0	NX3&H2;R0	OX2&H0;r	PX4&O1;R0	SX4&O1;R0
CX3&H1;R0	NX3&H1;R0		PX4&O0;R0	SX4&O0;R0
CX3&H0;R0	NX3&H0;R0		PX5	SX3&O2;R0
CX2&H1;R0	NX2&H1;R0		PX3	SX3&O1;R0
CX2&H0;R0	NX2&H0;R0			SX3&O0;R0
CX4&H2;R	NX1&H0;R0			SX2&H1;R0
CX4&H1;R	NX3&H1;R			SX2&H0;R0
CX4&H0;R	NX3&H0;R			SX4&O2;R
CX3&H1;R	NX2&H0;R			SX2&H0;R
CX3&H0;R	NX3&H1,r			SX2&H0;r
CX2&H0;R	NX3&H0;r			SX1&H0;R0
CX3&H1;r	N;O3			
CX3&H0;r	N;O2			
	N;O1			
#6	#7	#8	#15	#16

[a] The atom type is given in SMARTS notation. R0 denotes atoms that are not part of a ring system, R those are in an aliphatic ring, and r those that are members of an aromatic ring. The last row denotes any kind of the respective element. For the halogens only one atom type is used, respectively. Other elements are considered as "any atom".

The transfer of this concept to molecules is complicated by the fact that the constitution of the atoms is basically a 2-dimensional topology, namely the network of covalent bonds, whereas proteins contain a directed 1-dimensional amino acid sequence. Nevertheless, it is possible to construct a main chain through each molecule from which side chains branch off. Cyclic structures can be expressed likewise as side chains by cleaving rings accordingly. This scheme is utilized, for example, in the well-known SMILES and SMARTS notation of molecules.[43,44] Weininger et al. also introduced the concept of unique SMILES that allows determination of the main chain and the side chains unambiguously.[45]

As first step in our approach, we use the unique SMILES representation of a molecule to reconstruct the number and kind of covalent bonds, as well as the number and kind of substituents belonging to each non-hydrogen atom. According to this information the atom is assigned a certain atom type which corresponds to the residue type. The atom types are expressed in SMARTS notation, for example, CX4&H2;R denotes a sp^3-hybridized carbon with two attached hydrogens that is part of an aliphatic ring. We use 18 different atom types for carbon, 19 for nitrogen, 6 for oxygen, 9 for phosphorus, 16 for sulfur, and one for each of the halogens atoms, respectively (see Table 1). These give rise to a total of 74 different atom types plus the additional gap (see below). First, we constructed an appropriate substitution matrix on the basis of observed chemical replacements in drugs.[29,31] Likewise, substitution matrices used for protein sequence alignment were initially obtained by manual count of amino acids exchanges between functionally similar proteins.[39] We can, however, use our initial substitution matrix to perform the alignment of molecules within a distinct compound class and from that obtain the actual exchange frequencies in an automated procedure. For example, inhibitors of the angiotensin converting enzyme are quite similar to each other, and their alignment is therefore unequivocal. Collecting the

Table 2. Therapeutic Categories That Were Considered in Deriving the Bioisosteric Exchange Matrix[a]

class	number of compounds	class	number of compounds
ACE inhibitors	44	HIV protease inhibitors	18
anabolic steroids	51	leukotriene antagonists	60
androgens	39	local anesthetics (Caines)	65
angiotensin II-antagonists	25	nitrofuranes	29
antiarrhythmics (class III)	17	penicillines and derivatives	164
barbitals	23	phosophodiesterase IV inhibitors	11
benzodiazepams	97	pramines	22
beta-blockers	51	antiulcers (Prazoles)	18
calcium channel blockers	30	profenes	47
carbonic anhydrase inhibitors	8	progestogens	59
antifungals (Conazoles)	55	reverse transcriptase inhibitors	67
COX inhibitors	88	serotonin antagonists	25
dazoles	17	sulfonamides	55
diuretics	42	tetracyclines	19
floxacines	40	anticoagulants	31
histamine H1-antagonists	28	tyrosine kinase inhibitors	14
histamine H2-antagonists	26		

[a] Corresponding SMILES of these compounds are provided as Supporting Information.

corresponding exchange frequencies from a variety of different compound classes yields an updated substitution matrix that contains sufficient information to allow a general comparison on the basis of bioisosteric exchanges. For this purpose, a total of 1353 drugs from 33 different therapeutic classes were used (see Table 2).[46] Corresponding SMILES were obtained from the PubChem database and converted into unique SMILES using the CACTUS online SMILES translator at the National Cancer Institute.[47,48] All corresponding SMILES are provided in the Supporting Information.

The mutual alignments gave rise to about 10^4 exchanges between pairs of atom types. If no matching atom types can be found, that is, upon chain elongation or ring contraction, a gap is introduced that corresponds to insertions and deletions in one of the sequences. As gap penalty a value of -5 was chosen, whereas most other substitutions of atoms possess positive values. Thus the final exchange matrix closely reflects observed bioisosteric replacements in drugs.

Alignment Example and Derivation of the Bioisosteric Similarity. As an example of the alignment procedure lisinopril **1** and zabiciprilat **2** are used (see Figure 1, top). Their unique SMILES contain rather similar main chains (marked bold). Because of the nature of SMILES in general, this sequence of atoms is not necessarily the longest possible pathway in the molecule. Here, the main chain is used only as continuous path through the molecule from which side chains branch off. The use of unique SMILES, however, ensures that the directionality of the paths is equal, for example, from left to right. In the first step only these main chains are aligned to each other on the basis of the exchange matrix. This corresponds to the alignment procedure of amino acid sequences. Thus the resulting alignment A contains a number of gaps at position where no matching atom types are present. In cases where the alignment of the main chains produces only a very short sequence of matching atoms, a matching side chain is searched instead.

In the second step, for each of the remaining side chains (denoted B−G), a matching counterpart of the other molecule is assigned. To obtain the highest scoring combination, all mutual pairwise alignments of side chains are computed. This results in an $n \times m$ matrix of alignments for n and m side chains. That is, lisinopril possess five side chains (B−F), whereas zabiciprilat contains an additional sixth side chain G. The assignment of side chain counterparts is now performed iteratively in a greedy fashion: The highest scoring combination of two side chains is chosen in each step which is then removed from the matrix of available combinations. Here, the identical side chains C of both molecules would be selected first, as they result in the longest alignment. In cases where two or more side chain alignments yield equal scores, the pair that appears first in the matrix is chosen. This strategy therefore enables the matching of disconnected fragments that is also referred to as scaffold hopping. Depending on the chosen value for the gap penalty, the hopping is more or less pronounced (see Results). Here, most of the side chains (B−F) can be matched one-to-one between **1** and **2** (bottom of Figure 1a). To side chain G, no counterpart can be assigned, and therefore, gaps are introduced.

In the final step, the alignments of the main chain and the side chains are assembled (Figure 1b). Using this alignment of two molecules, a numerical score is obtained that can be used to express the distance or in turn the similarity between the molecules. Here, the bioisosteric similarity S_{AB} between two molecules A and B is calculated as

$$S_{AB} = \frac{2}{\text{length(A)} + \text{length(B)}} \times \sum_{A_i,B_j}^{\text{alignment}} \frac{M_{ij}}{\max(M_{ii}, M_{jj})} \quad (1)$$

The alignment assigns to each atom A_i on molecule A a matching atom B_j on molecule B or a gap, respectively. The according entry of the substitution matrix M_{ij} is divided by the maximum of the two diagonal elements. This ensures that matches of identical atom types always yields unity, whereas substitutions are considered according to their exchange probability. The sum of these quotients is finally normalized by the sum of the lengths of the molecules A and B that corresponds to their respective sequence lengths, that is equal to the count of non-hydrogen atoms in both molecules. Thus the bioisosteric similarity ranges from zero for molecules that have nothing in common, to unity for identical molecules. For the example lisinopril against zabiciprilat a similarity of 84.7% is obtained. The according (evolutionary) distance between molecules can be computed

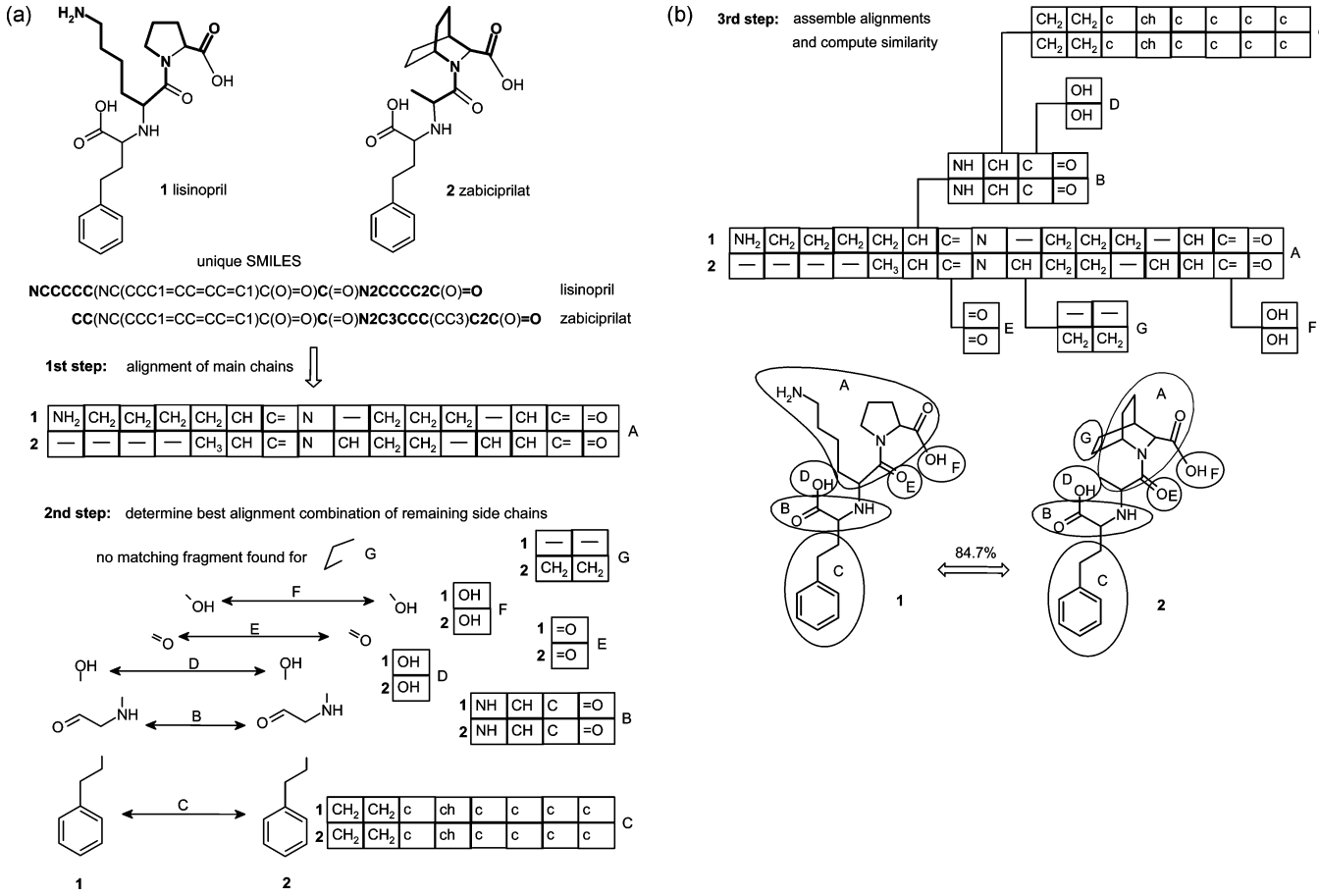

Figure 1. Structural alignment of lisinopril and zabiciprilat based on the exchange frequencies of bioisosteric replacements in pharmaceutical compounds. (a) First, the main chains (marked bold) of both compounds as obtained from unique SMILES are aligned. For the remaining side chains, the highest scoring combination of pairwise alignments is determined in an iterative way. Here, corresponding matches are found for all side chains B−F, except fragment G. (b) Finally, the alignments of main and side chains are assembled and the similarity is computed according to eq 1.

as the reciprocal of their mutual alignment score. These distances can now be used to perform clustering within the compound class or similar to protein sequences to construct a phylogenetic tree.[49]

Preselection Criteria to Handle Larger Sets of Compounds. The generation of alignments between molecules is of course only feasible for a limited number of compounds. To allow the virtual screening of substance databases against a query, a preselection of potentially similar molecules is thus necessary. For this purpose, we compare the presence and count of the above-mentioned atom types (see Table 1) between the query structure and each molecule of the database. The corresponding similarity index was originally introduced by Carhart et al. using descriptors.[50]

$$F_{AB} = \frac{\sum_k \min(f_{Ak}, f_{Bk})}{0.5\left(\sum_k f_{Ak} + \sum_k f_{Bk}\right)} \quad (2)$$

Here, we define f_{Ak} as the count of atom types of type k in molecule A, and likewise is f_{Bk} the count of atom types of type k in molecule B. The according similarity index F_{AB} ranges from zero (no atom types in common) to unity for a pair of molecules sharing identical count of the respective atom types. Since the count for each atom type in a molecule can be stored in a separate database, corresponding assignment of atom types has to be carried out only once for each molecule of a substance database. The comparison to a query structure applying F_{AB} is thus computationally inexpensive. As a threshold value we use 0.5 to account for the large number of atom types (74) since eq 2 was observed to yield lower similarities on average when an increased number of descriptors k is used.[51] This threshold was chosen on the basis of a comparison of atom type-based similarity F_{AB} and bioisosteric similarity S_{AB} of diazepam against all molecules in the lead-like subset of the ZINC library. None of the compounds with an atom type similarity below 0.5 achieved a corresponding bioisosteric similarity above 87%. A corresponding plot is provided as Supporting Information. Compounds exhibiting higher values than the threshold are taken into consideration for the alignment to the SMILES string of the query. Within this study, this preselection criterion was only applied upon screening the lead-like subset of the ZINC library but not for all other compound sets.

Compound Set. SMILES strings of the investigated molecules were either obtained from the PubChem database,[47] the ZINC library,[52] or taken from a previous study.[53] The substances comprised the lead-like subset (972 608 compounds, version 7) of the ZINC library, 33 291 organic compounds from the Sigma-Aldrich catalog,[54] 12 959 ligands of the PDB database,[55] 4603 pharmaceutical agents from Prous Science Drugs of the Future, which are predominantly new drugs in clinical trials,[56] and 2238 nondrugs without any pharmaceutical function.[53] Entries containing metals

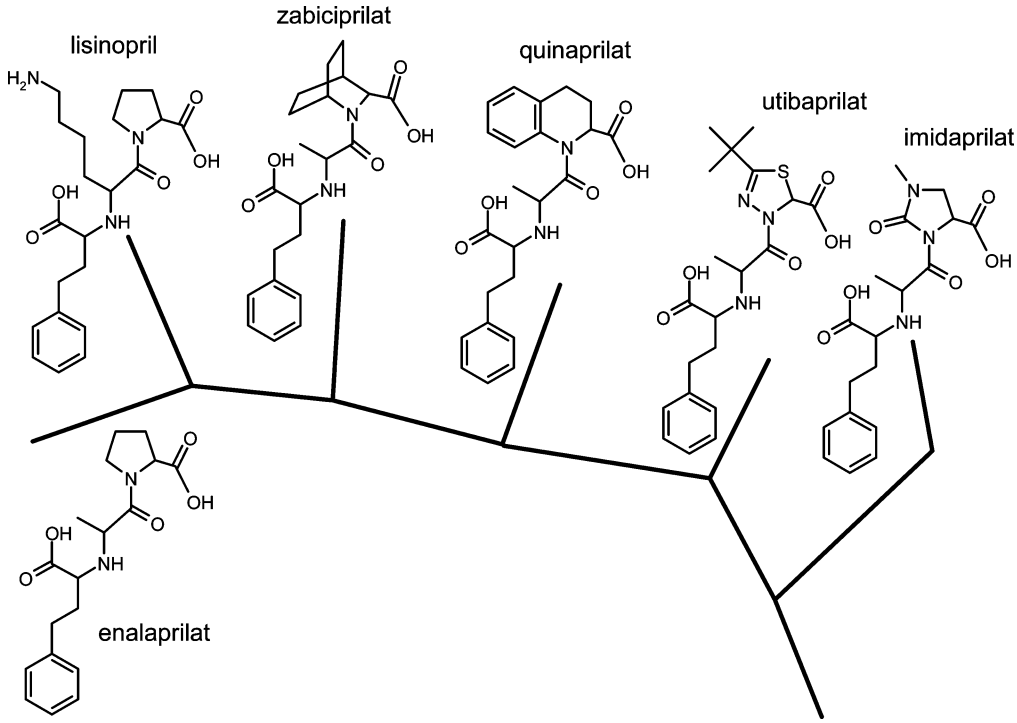

Figure 2. Bioisosteric similarity between compounds can be used to construct a phylogenetic tree that reflects their chemical relationship resulting from from chemical replacements. Shown here is a single branch of the complete tree of ACE inhibitors.

were removed. Salts were converted into their neutral counterparts, thereby replacing lithium, sodium, potassium, and calcium. Entries consisting of multiple substances or mixtures were converted into the respective active agent. All SMILES strings were converted into unique SMILES using the CACTUS online SMILES translator at the National Cancer Institute.[48] Query compounds are either "average" molecules of the respective drug class or external substances with known binding affinity, that is, SC-558 and UC-781, where corresponding X-ray structures are available showing the inhibitors bound to the target enzyme. All SMILES used are either available from the cited sources or provided in the Supporting Information.

Fingerprint Similarity. For comparison with our approach, we computed the similarity between the query compounds and the remaining molecules in the respective drug class based on the Tanimoto index of the default 1024 bit fingerprint method implemented in Open Babel.[57,58] This so-called FP2 fingerprint is of the Daylight type, indexing a molecule based on the occurrence of linear fragments up to 7 atoms in length. Chemically identical fragments (reverse order of atoms or rings starting at different atom positions) are discarded. Occurring fragments set a bit in the 1024 bit vector forming the fingerprint.

RESULTS

Alignment. To demonstrate the functionality of the bioisosteric alignment algorithm, Angiotensin converting enzyme (ACE) inhibitors were used.[59] From the pharmacological view, they belong to the antihypertensive agents and are derived from the tripeptide Phe-Ala-Pro that exhibits micro molar affinity toward the human ACE that cleaves the decapeptide angiotensin I to angiotensin II, which works as vasoconstrictor. Shown in Figure 1 is the alignment procedure between lisinopril **1** and zabiciprilat **2**. From the unique SMILES, the main chain (denoted A) is obtained (top of Figure 1a). This is a peptide-like fragment that contains the proline-related heterocycle found in all ACE inhibitors. To the amino-butyl substituent of lisinopril no corresponding counterpart in zabiciprilat can be found and thus gaps are introduced in the alignment of the main chain A. Most of the remaining side chains (B−F) can be matched one-to-one (bottom of Figure 1a). More instructive is the processing of the bicyclic ring system of zabiciprilat. To align the 6-membered ring onto the 5-membered ring of lisinopril, a gap is introduced that accounts for the additional $-CH_2-$ group. The remaining $-CH_2-CH_2-$ bridge is treated as side chain G, which is unique to zabiciprilat. These differences eventually give rise to a bioisosteric similarity of 84.7% between lisinopril and zabiciprilat after assembly of the alignments of main and side chains (Figure 1b).

Using the computed mutual similarities for all ACE inhibitors, we constructed a corresponding phylogenetic tree by applying the Neighbor-Joining algorithm that allows a visual inspection of the relationships.[49] Shown in Figure 2 is a single branch of the obtained unrooted tree for ACE inhibitors. The complete tree is given in the Supporting Information. The close vicinity of lisinopril and zabiciprilat, which is the result of their strong similarity, is apparent, whereas compounds that contain even larger or substituted ring systems are placed increasingly distant. As a consequence of the Neighbor-Joining algorithm, closest neighbors are stepwise connected first, whereas more distant members are joined to the tree at last. Therefore obviously loosely related molecules might be found surprisingly close but at the very end of a branch. We observed this case for captopril, which is most distant to all other inhibitors. This can be attributed to its small molecular size and limited number of functional groups. Moreover, such trees must be seen as a

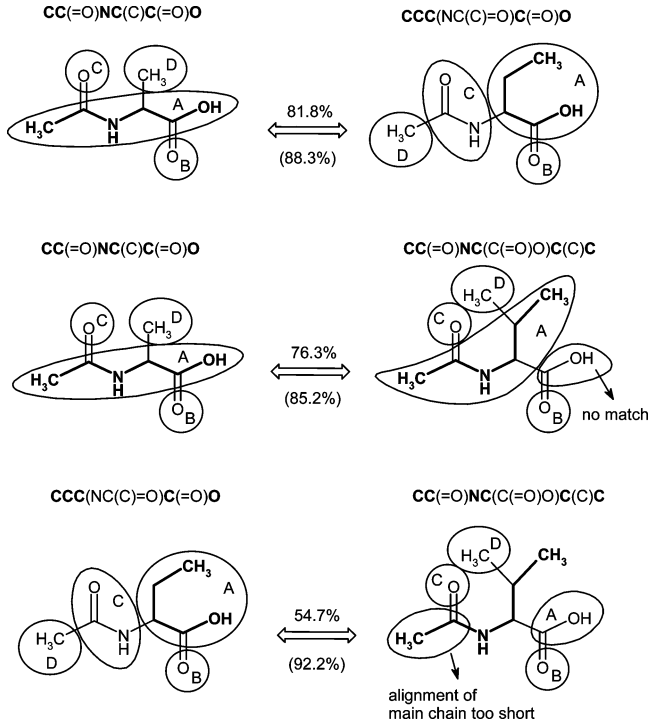

Figure 3. Mutual pairwise alignment of three similar molecules where the main chains (marked bold) of their unique SMILES strongly diverge. In cases where the resulting alignment of the main chains would be too short in comparison to the number of mismatches, the remaining C−OH side chain is matched instead (bottom, right). For comparison, the similarities obtained with manually created SMILES on the basis of the first molecule are given in parentheses.

2D-projection of a multidimensional space similar to clustering and therefore cannot account for all issues simultaneously.

Alignment of Diverging Main Chains. Because of the nature of unique SMILES cases occur in which obviously similar compounds exhibit strongly diverging main chains. This is caused by the reordering of the atoms by the applied canonicalization algorithm that does not necessarily produce the main chain as longest path in the molecule, but rather a unique description of the molecular graph. Figure 3 shows a corresponding example of three small molecules and the obtained pairwise alignments between them. Apparent are the different main chains (marked bold). It is obvious that manually generated SMILES would resemble that of the molecule shown on top left and corresponding bioisosteric similarities are given in parentheses. In all three alignments based on the unique SMILES matching counterparts of the side chains B−D are assigned unambiguously. The alignment of the main chains, however, reflects the mutual exchange probabilities between different atom types. In the second case (Figure 3, middle), the −OH group of the carboxylic acid (left) is matched to the terminating -CH_3 group (right). In the last case (Figure 3, bottom) the corresponding alignment of the main chains would yield only a very short uninterrupted segment (two carbon atoms) being outnumbered by mismatches between the remaining atoms. After matching of the side chains B−D, the remaining C−OH group (right) is instead matched to the main chain A (left) causing the lower similarity. In turn, manually generated SMILES yield the similarities shown in parentheses, that are in agreement with the expect trend. For these rather small molecules, unique SMILES do only allow an insufficient matching of the main chains by the present implementation of the alignment algorithm. In contrast to these three molecules, drug-like substances contain substantially longer main chains that allow a practically usable alignment.

Diversity among Drug Classes. The mutual distances can also be used to assess the diversity of a given group of compounds. In a class that contains only very similar drugs, throughout high mutual bioisosteric similarities are found. Thus the according chemical diversity of this class is low. For a highly diverse class the corresponding similarities are comparatively low. Here, we computed the diversity using the median of the according distances. The median was used instead of the average to account for nonsymmetrical distributions. For example, if more drugs exhibit higher similarities than the actual average of compounds in that class, the distribution is skew. The resulting diversities of the 33 investigated drug classes are shown in Figure 4. The lowest diversity is found for tetracycline antibiotics (at the bottom in Figure 4). These contain throughout the same ring system with only minor modifications of the substituents. Conversely, antiulcer drugs of the Prazol group (e.g., omeprazole) show an emphasized variability that renders them as highly diverse (at the top in Figure 4). At first sight it may seem surprising that ACE inhibitors are ranked less diverse than angiotensin II antagonists, despite the distinct and reappearing framework of the latter compounds. The comparison reveals, however, that the modifications to the skeleton of ACE inhibitors are limited and reoccurring. For example, all compounds in Figure 2 contain the same substructure derived from the Phe-Ala-Pro tripeptide. Other approaches to assess diversity, for example, those based on the comparison of fingerprints assign a low diversity to angiotensin II antagonists.[60] This can be attributed to the nature of the respective fingerprints that usually emphasize the presence of substituents and functional groups. That is, these contain more descriptors for such properties than for the skeleton.[11,14] Here, any kind of bias regarding framework or substituents is avoided, while simultaneously the frequency of chemical modifications is taken into account. This also allows an estimate whether the considered substance class covers the available chemical space adequately.

The mutual distances between substances in a class can furthermore be used to identify the "average" molecule. This compound is assumed to be representative for the corresponding class. For this purpose, we computed the average of the mutual distances for each compound. The molecule being most central should thus possess the smallest standard deviation of the averaged distances. In contrast to other clustering algorithms no further weighting is used. Among the ACE inhibitors we obtained imidaprilat (see Figure 2) as average molecule. It is situated closer to the inner part of the tree than, for example enalaprilat that is found at the end of this branch.

Restrospective virtual screening. To investigate the suitability of the bioisosteric similarity for virtual screening, we compared the recovery of ACE inhibitors against a background of other substances using imidaprilat as query. Figure 5a shows that the majority of ACE inhibitors possess a higher bioisosteric similarity to imidaprilat than substances from any other database. The distribution of the bioisosteric similarities shows an apparent trend: compounds from databases that entirely contain drug-like matter, that is, those

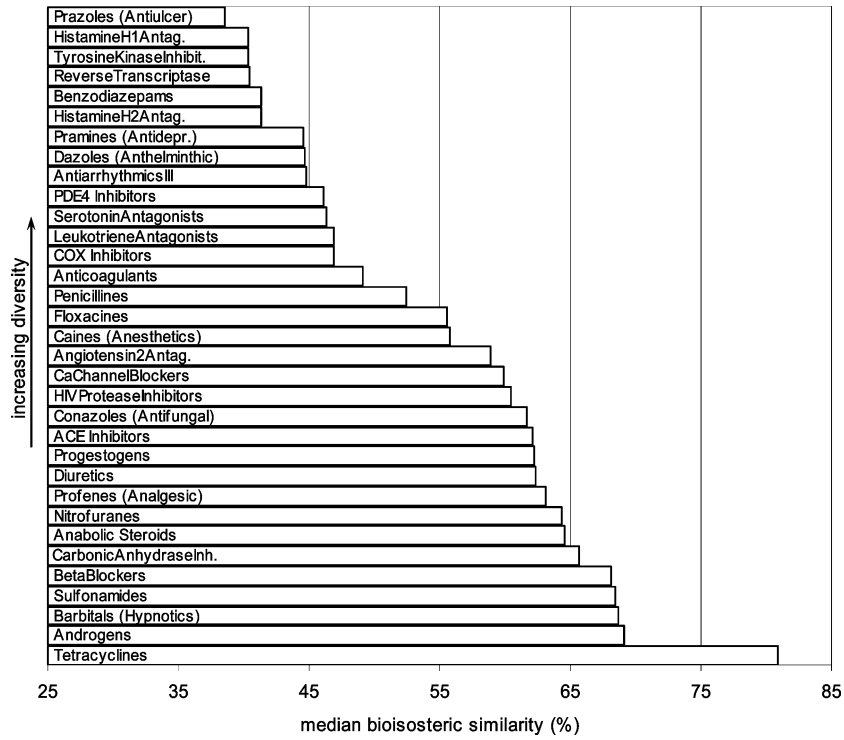

Figure 4. Chemical diversity within a class of compounds is related to the median bioisosteric similarity of the corresponding compounds. Classes containing drugs that exhibit high mutual similarities possess a low diversity (bottom). Low similarities conversely indicate highly diverse classes (top).

from Prous Science Drugs of the Future exhibit throughout higher bioisosteric similarities than the set of actual nondrugs. Both the PDB database as well as the Sigma-Aldrich catalog comprise any kind of substances and therefore lie in between. The curve of the PDB database shows a steep increase around 30% similarity that can be attributed to the nature of its compounds. Typical ligands of enzymes are usually larger than the smallest molecules that are present in other substance catalogs.

The bioisosteric similarity of imidaprilat to all other ACE inhibitors ranges from 88.6% for zabiciprilat to 21.6% for rentiapril. For comparison, we computed the mutual similarities using other ACE inhibitors. The corresponding curves looked rather similar for inhibitors that exhibit high bioisosteric similarities to imidaprilat, for example, utibaprilat (86.3%) and enalaprilat (86.7%) (data not shown). Using captopril (42.8%) as query instead resulted in a curve that stayed around 25% below that obtained with imidaprilat. This is because of the small size and low content of hetero atoms in comparison to all other ACE inhibitors. Thus it is seemingly better to use query molecules for screening that are close to the "average" molecule of that class. Comparing the bioisosteric similarities of the average molecule imidaprilat to the other ACE inhibitors shows that the highest value is obtained for zabiciprilat (88.6%), which is in the same branch of the tree, but not for the nearest neighbor utibaprilat (86.3%) (see Figure 2). This seeming dichotomy is resolved recalling that the relationships within the tree reflect the mutual distances between all compounds rather than pairwise similarities of neighbors. Eventually, enalaprilat shows an even higher similarity (86.7%) despite being located at the very end of this branch. The remaining inhibitors in the same branch comprise lisinopril (78.3%) and quinaprilat (79.4%). In this context it is interesting to note the common substructure in this branch which is derived from the tripeptide Phe-Ala-Pro.

As result of screening Prous Science Drugs of the Future using imidaprilat we obtained as highest scoring hits enalaprilat (86.7%), imidapril (82.9%), lisinopril (78.3%), the clinical ACE inhibitors CGS-13928C (77.9%), and CL-242817 (77.8%), as well as the aminopeptidase inhibitor bestatin (73.2%). From the ligands in the PDB database, the highest scoring hits were enalaprilat (1UZE), the inhibitors WR-112 (1EWM), and WR-99 (1EWL) of the cysteine protease cruzain, lisinopril (1O86), sialoadhesin complexed with 2-phenyl-prop5ac (2BVE), and the aspartic acid−based inhibitor of the Matrix metalloproteinase-8 (1A86). Further ligands with bioisosteric similarities higher than 70% compared to imidaprilat comprised peptide-like inhibitors of thermolysin (5TMN and 1TMN). Thus, on one hand other ACE inhibitors are recovered and on the other hand similar peptide-like ligands of other enzymes are detected. The latter functionality, namely, to find similar substances in other compound classes, is also referred to as scaffold-hopping. Likewise, two penicillin derivatives, carbenicillin (80.2%) and Penicillin-G (74.9%) were found among the highest scoring hits from the Sigma-Aldrich catalog. They also contain peptide-like substructures. The remaining compounds exhibiting bioisosteric similarities above 70% are mostly substituted di- and tripeptides without pharmaceutical function.

Comparison with Other Similarity Methods. In virtual screening, it is also of interest to be able to estimate a lower margin for the similarity, for example, down to which threshold most of the active compounds are recovered. This value is depending on the used similarity method and also on the class of compounds being screened. The effect of using different exchange matrices is shown in Figure 5b. Here again imidaprilat was used as query but against the

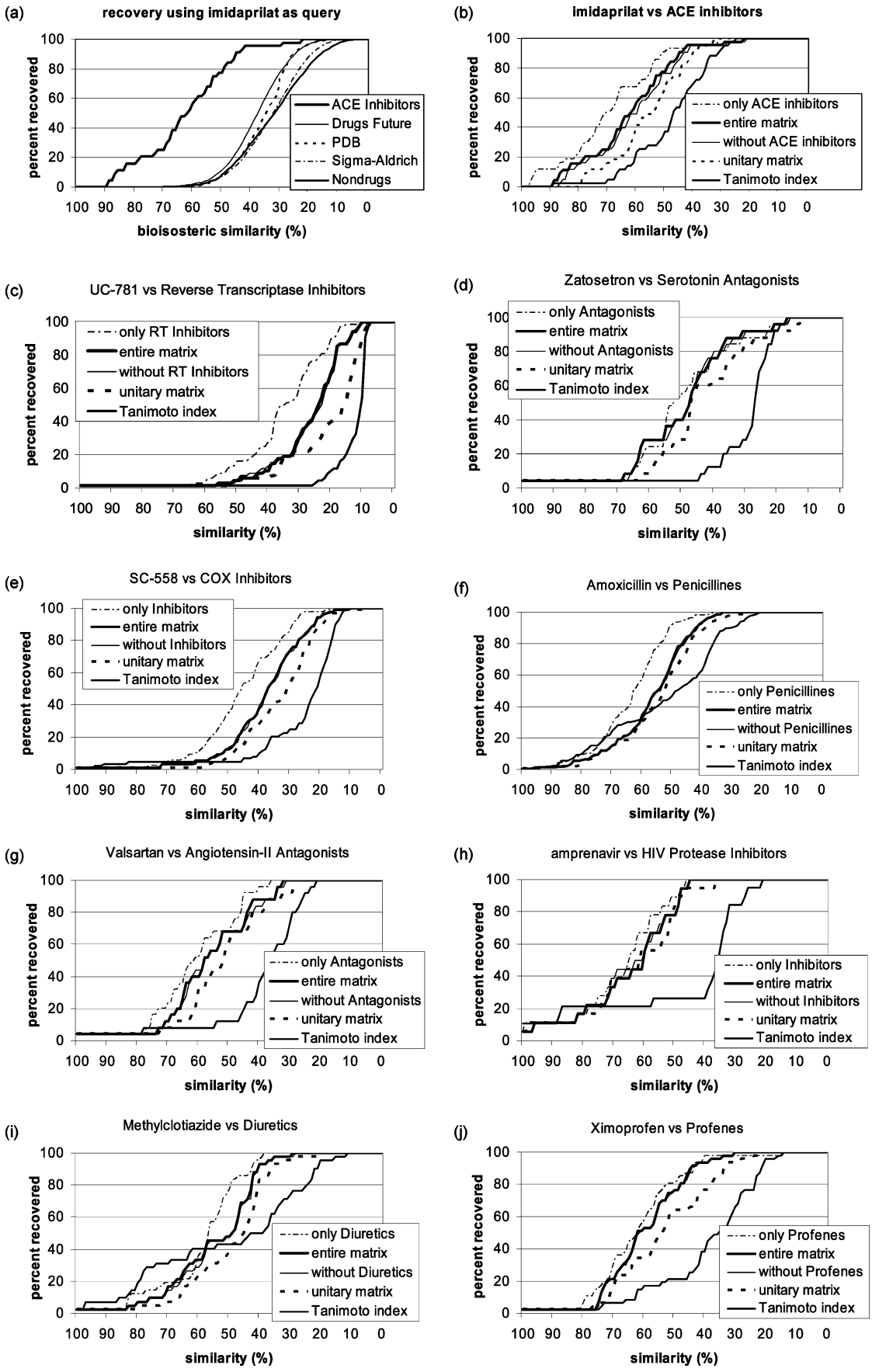

Figure 5. Comparison of different substitutions matrices, as well as fingerprint similarity for the recovery of drugs in virtual screening. (a) of ACE inhibitors using imidaprilat as query among various substance databases (b−j) within the respective drug class using the given query molecule. Dashed line: substitution matrix derived only from the respective drug class. Bold line: matrix including all drug classes. Thin line: matrix excluding the respective drug class. Bold dotted line: unitary matrix that corresponds to maximum common substructure search. Emphasized line: Tanimoto index of fingerprint similarity.

remaining ACE inhibitors in this drug class. The first matrix (dashed line) was derived solely from bioisosteric exchanges within the same drug class, namely, ACE inhibitors. It shows a somewhat higher recovery rate than that of the complete matrix (bold line) that was constructed using all of the investigated drug classes in Table 2. Conversely, the third matrix (thin line) was derived from all drug classes, but excluding ACE inhibitors. The corresponding rate of recovery is, however, almost identical compared to the complete matrix. This indicates that bioisosteric exchanges are rather similar across the investigated drug classes. Assigning equal probabilities to exchanges between the various atom types gave rise to a unitary matrix (bold dotted line). Here, nonhydrogen atoms are matched to each other on a one-to-one basis that corresponds to a maximum common substructure approach. The rate of recovery for this matrix is below that of the other matrices that account for bioisosteric exchanges down to about 45% similarity, where all of the matrices recovered 95% of the compounds. Finally we computed the similarity of the remaining ACE inhibitors toward the query molecule as the Tanimoto index of a 1024 bit fingerprint (emphasized line). Details of the fingerprint method are given in the corresponding section of the methods. The performance of this similarity method is well below that of the exchange matrix-based alignment approaches.

The same comparison of different exchange matrices and fingerprint similarity was carried out on further drug classes of varying diversity (see Figure 5c−j). The respective molecules used as queries are either "average" molecules of the respective drug class according to the bioisosteric similarity or, where available, external substances with known binding affinity that have been used as queries in other studies in the context of retrospective screening. The chemical structures of all query molecules are shown in Figure 7. The drug classes shown in Figure 5c−j are arranged in order of decreasing diversity (c.f., Figure 4). In general the rate of recovery is almost identical for the complete exchange matrix and the matrix without the respective drug class. The curve shape of the unitary matrix (maximum common substructure) is usually well below that of the other matrices, indicating the benefit of incorporating bioisosteric exchange probabilities. Particularly for drug classes exhibiting high diversity e.g. Reverse Transcriptase inhibitors and Serotonin antagonists the performance is much worse using fingerprint similarity. Only for less diverse classes where the corresponding drugs additionally share a common framework, for example, diuretics, the recovery of highly similar compounds is facilitated by fingerprints. However, for any of the investigated classes, the majority of compounds is recovered with higher similarity values using the matrix-based alignment approaches than using fingerprint similarity.

Since the comparisons shown in Figure 5b−j were carried out within the same drug classes of the respective query molecules, the performance of the bioisosteric similarity and the fingerprint similarity were furthermore tested against the Prous Science Drugs of the Future set of compounds. To this collection of substances comprising predominantly new drugs, the compounds of the respective drug class were added. The resulting databases were ordered in decreasing similarity toward the query molecule. Shown in Figure 6a−i are the recoveries of drugs in the respective class obtained by the bioisosteric approach using the entire exchange matrix (bold line), the unitary matrix (dashed line) that corresponds to a maximum common substructure approach, and the Tanimoto index of the fingerprint similarity method (dotted line). Here, the compounds were sorted in order of decreasing similarity from left to right, according to the respective similarity method. Thus, each fraction of the database comprises those compounds that possess equal similarity toward the query molecule. The curves indicate that the bioisosteric similarity obtained with the entire exchange matrix recognizes the active drugs at higher values of similarity for most drug classes compared to using the unitary matrix, as well as the fingerprint approach.

For the purpose of finding similar molecules in larger databases, ACE inhibitors are less suited because of their peptide-like structures. Such compounds are found abundantly in the lead-like subset of the ZINC library. Therefore, we used diazepam from the class of benzodiazepams to compare the functionality of our bioisosteric similarity method to the corresponding results obtained with the online version of FTrees.[20] Diazepam contains the prototypic tricyclic structure (see Figure 8) that is shared by most other benzodiazepines. Due to the large number of different bioisosteric modifications in this class, the according diversity is, however, rather high (see Figure 4). This makes diazepam an interesting query compound for the comparison of different similarity methods. Shown in Figure 8 are the highest scoring hits obtained using the SwiFT algorithm of FTrees (left) and our bioisosteric similarities applying the default gap penalty of −5 (right) using the preselection criterion according to eq 2 (see Method section). The exchange of the chlorine substituent by bromine is in both methods among the first hits. Our approach, however, ranks the bioisosteric exchange by methyl somewhat higher. Likewise is the arrangement and connectivity of the tricyclic structure strongly conserved. Compounds containing an 8-membered ring instead of the 7-membered ring, for example, ZINC04579500 are also found by our approach and are among the 20 highest scoring compounds (data not shown). Interesting are the chemical alterations to the tricyclic framework. Along the introduction of substituents such as hydroxyl, amino, and acetyl, reduction of the C═N double bond is observed. Moreover, the exocyclic carbonyl oxygen is replaced by sulfur. Many compounds obtained with FTrees in contrast show pronounced rearrangements of the ring systems that go along with contraction to 6-membered and 5-membered rings. Thereby a methylene group is lost that leads to accordingly lower similarities computed by our approach. The corresponding similarity values are given in parentheses in Figure 8. Applying the default gap penalty of −5 apparently leads to a pronounced conservation of substructures, since disconnection of fragments becomes unfavorable. We therefore repeated the screening using a smaller penalty of −2 that should allow "hopping" between fragments and therefore produce rearrangements. Now the substances containing the 8-membered ring are ranked considerably higher (see Figure 9). Likewise compounds with larger aliphatic substituents instead of the N-methyl group (e.g., ZINC03138544), rearranged fragments and ring contraction (e.g., ZINC03200237), as well as additional aromatic rings (ZINC00221193) are found from rank 12 on (data not shown) but with bioisosteric similarities below 93%. Thus bioisosteric similarities above 90% indicate not exclusively

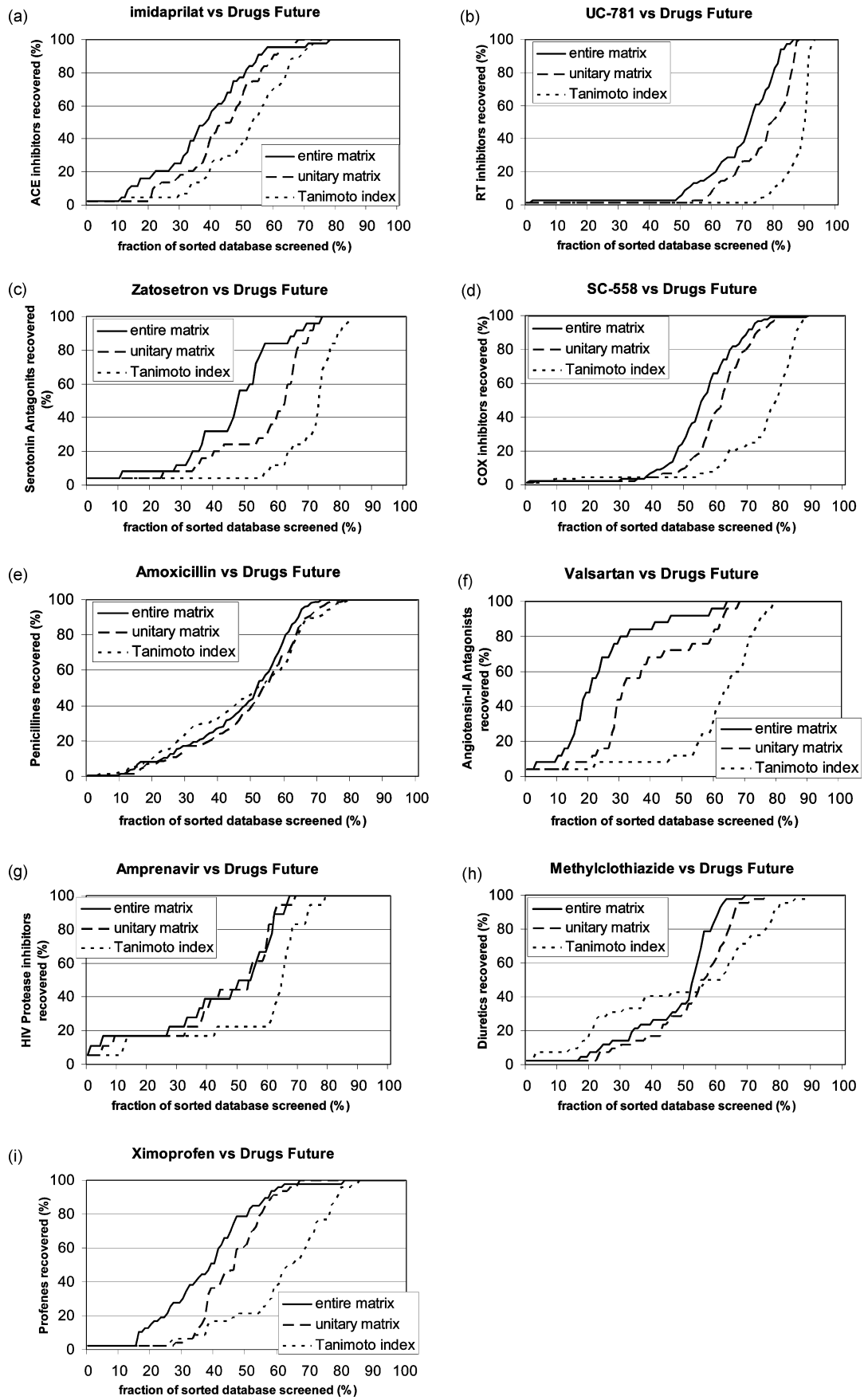

Figure 6. Comparison of different substitutions matrices as well as fingerprint similarity for the recovery of drugs in virtual screening. The given query molecule was used to screen against a database containing drug-like substances plus all drugs of the respective class. The screened compounds were sorted in order of decreasing similarity according to the respective similarity method. Thus, each fraction of the database comprises those compounds that possess equal similarity toward the query molecule. Bold line: matrix including all drug classes. Dashed line: unitary matrix that corresponds to maximum common substructure search. Dotted line: Tanimoto index of fingerprint similarity.

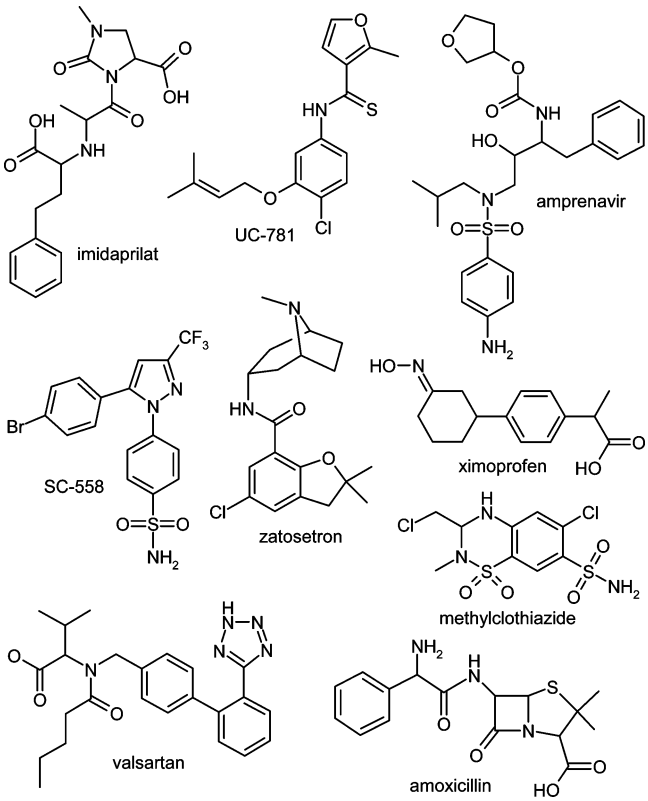

Figure 7. Chemical structures of the molecules used as queries for virtual screening.

compounds containing the same substructure but also substances where corresponding exchanges have produced "highly" similar compounds.

For a further comparison, the fingerprint-based similarity function in PubChem[47] was also used to screen diazepam against the lead-like subset of ZINC. The highest scoring hits are shown on the left-hand side of Figure 9. Apparent among the structures obtained by the fingerprint similarity are those where the *N*-methyl group is replaced by hydrogen. Since this exchange introduces an additional hydrogen-bond donor, the corresponding compounds, for example, nordiazepam and delorazepam, are not found among the high scoring compounds in both the FTrees, as well as our approach. Likewise extending the *N*-methyl group to longer hydrocarbon chains also causes lower bioisosteric similarities.

Most Common Replacements. The exchange matrix used to generate the alignments is based on observed one-to-one replacements between the various atom types among the investigated drug classes. Since the corresponding matrix entries are normalized with respect to the occurrence of the individual atom types in these compounds, it is possible to extract the relative frequency of replacements. Shown in Table 3 are the rankings for the corresponding one-to-one exchanges of the most relevant atoms. Since the approach used here entirely consists in such one-to-one exchanges, replacements of larger fragments or chemical groups cannot be considered directly.

DISCUSSION

So far, the results demonstrate the usability of our approach expressing bioisosteric similarity by the means of exchange probabilities for both retrospective screening and scaffold-hopping. In contrast to tree-based methods for comparing molecules, here an alignment is generated that uses distinct atom types. The sequence of atoms obtained from unique SMILES is used only as an empirical guidance to the pairwise alignment. Since similar molecules are likely to share similar main chains, this will speed up the generation of the alignment. With increasing differences compounds become more and more partitioned, for example, the common main chain becomes small in comparison to the number of side chains. Likewise, those parts of ring systems that cannot be matched to the main chain, are treated as side chains (see for example side chain G in Figure 1). Regarding such extensions, deletions, and more complex substitutions, our approach reflects the algorithms being used to align protein sequences in contrast to maximum common substructure algorithms that mainly maximize the size of connected fragments. Here, most computational effort arises from mutual alignment of possible side chain combinations, and therefore, any empirical approach to reduce their amount is reasonable. In this respect, it is thus more obvious to use unique SMILES than arbitrary SMILES.

In case of the three small molecules (see Figure 3), graph-based similarity methods can be expected to yield higher pairwise similarities as a result of matching the respective maximum common substructure. Here, the low similarity obtained for the last pair is the consequence of the inappropriate unique SMILES, whereas manually generated SMILES yield the expected results (values given parentheses). Nevertheless, the obtained alignment matches corresponding side chains and, if necessary, also considers parts of the main chain. Thus, even in the case of poor main chain alignment because of awkward SMILES a practically useable assignment between molecules is obtained. Moreover, the exchange probabilities allow matching of bioisosterically related atoms in a different way than graph-based or tree-based method, where a limited number of fragments is considered. Since matching of side chains is pronounced in this approach, scaffold hopping results as a side effect. In so far, this approach does not attempt to produce an optimal alignment such as a maximum common substructure, but rather to find even loosely related compounds in terms of bioisosterical similarity, with an emphasis on side chains and fragments. Thus we see this approach an addition to existing similarity methods for screening.

The actual position of molecular fragments also influences the behavior of the algorithm. For example, the alignment and scoring upon the exchange of cyclohexane against hexane depends on the subsitution pattern and placement of those fragments within the considered molecule: If cyclohexane or hexane are terminating substituents, then no gap penalty is assigned because, for each atom, a corresponding counterpart is found. The scoring differs only for the ring-opening atom and the final atom resulting from different numbers of attached hydrogens. In cases where a linear hexane chain replaces cyclohexane as linker, the additional $-CH_2-$ fragments that cannot be matched directly give rise to gaps in the alignment. Regarding the consideration of such disconnections, this alignment strategy is related to the maximum common edge subgraph procedure realized in the RASCAL method.[6] In contrast to other maximum common substructure approaches where only connected graphs are matched, corresponding gaps between fragments are allowed.

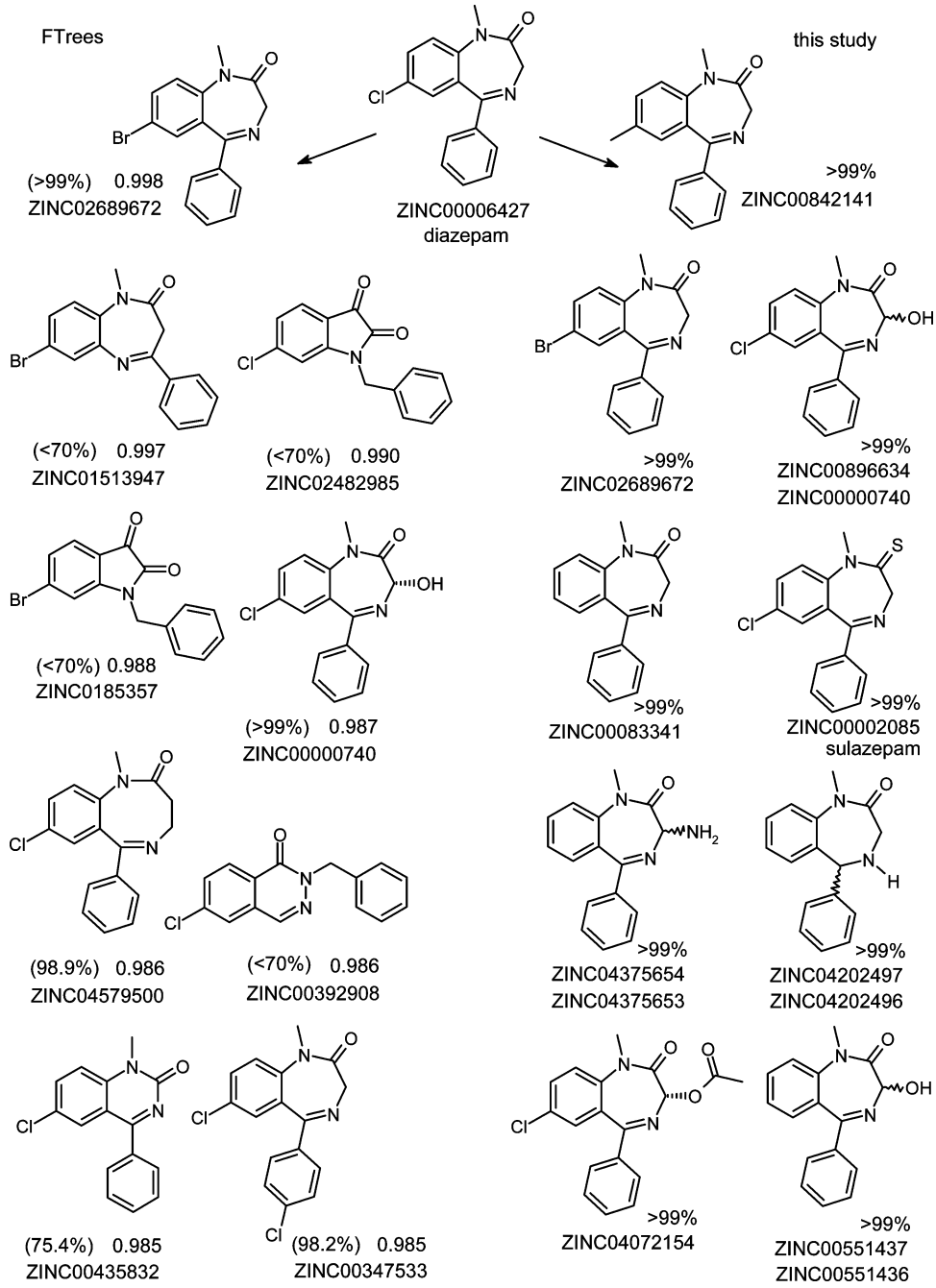

Figure 8. Comparison of the most similar compounds to diazepam from the lead-like subset of the ZINC library. Left: obtained with the online version of FTrees. Right: obtained with the bioisosteric method described in this study. The corresponding bioisosteric similarities for the compounds obtained with FTrees are given in parentheses. Two ZINC entry numbers indicate stereoisomers sharing identical bioisosteric similarity.

Likewise bonds between atoms are considered instead of the vertices of the graph. Whether the implemented clique detection method is actually carried out is subject to a preselection system that uses cost matrices to determine a minimum acceptable similarity between the molecules.

Despite apparent similarities of the results obtained with our alignment-based method to those of FTrees (see Figure 8) that predominantly should conserve features and the fingerprint-based method used in PubChem on the other hand (see Figure 9), there are distinct differences beyond the usual differences between closely related similarity methods.[51] These concern the introduction of additional substituents and the rearrangement of fragments, ring contraction and ring extension, respectively. Thereby the following tendencies are observed: (1) Substructures are conserved in favor of rearrangements of fragments. (2) Exchange of groups is carried out according to their observed frequency in bioisosteric compounds. This is most apparent for one-to-one replacements of terminating substituents (see also Table 3). (3) Insertions into aliphatic side chains compete with the corresponding extension of rings. The magnitude and preference for such insertions depends on the chosen gap penalty. The same behavior is observed for the contraction of chains and rings. Thus all of these tendencies reflect the nature of bioisosteric replacements.

Some other studies have previously applied SMILES or SMILES-derived notation of compounds to assess similarity between molecules. Vidal et al. have devised a method called

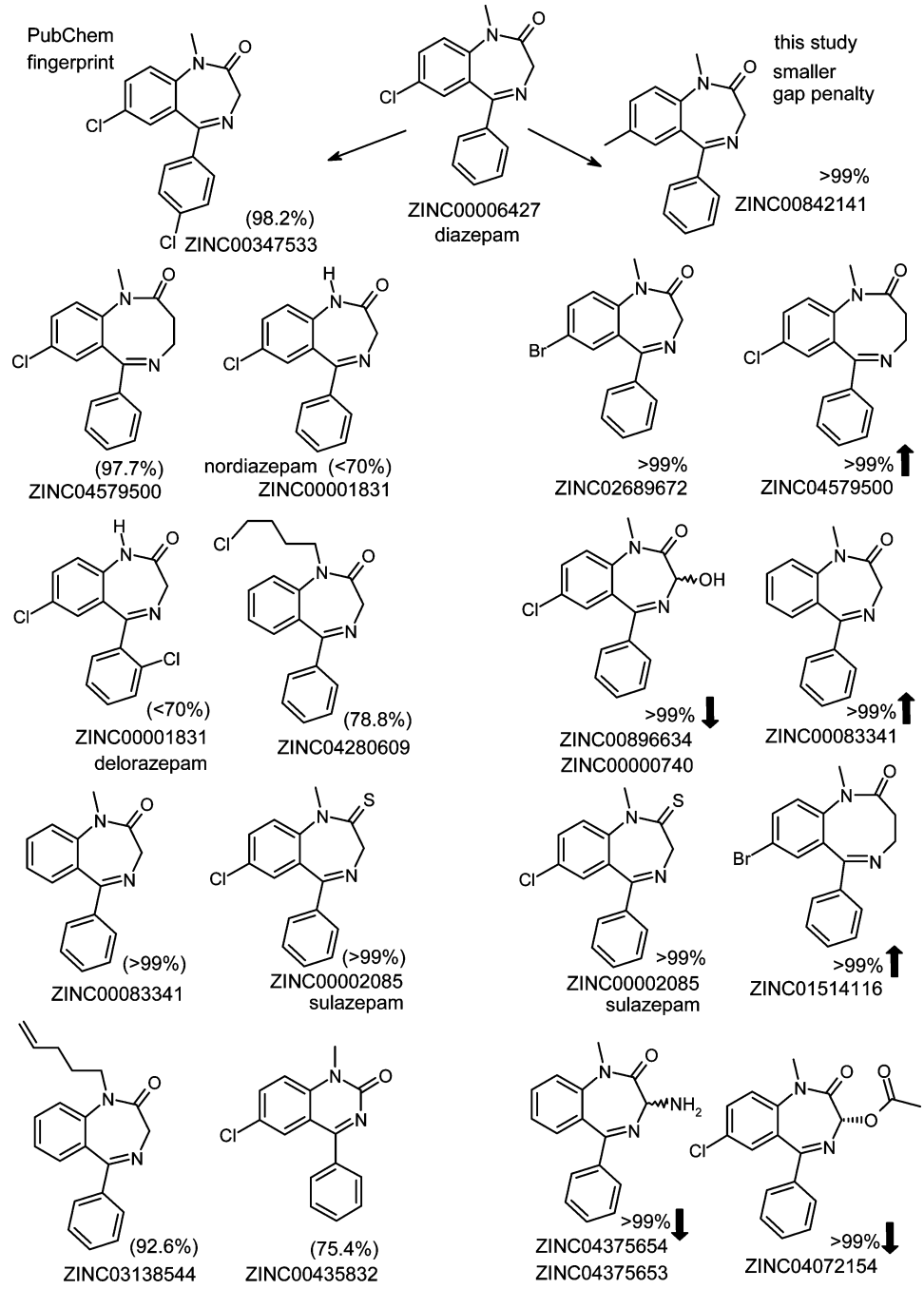

Figure 9. Comparison of the most similar compounds to diazepam from the lead-like subset of the ZINC library. Left: obtained with the fingerprint-based method of PubChem; Right: obtained with the bioisosteric method described in this study, but applying a smaller gap penalty. The arrows indicate the changes in ranking compared to the results with the default gap penalty. The corresponding bioisosteric similarities for the compounds from PubChem are given in parentheses.

LINGO where canonical SMILES are fragmented into substrings that can be used as binary descriptors in quantitative structure−property prediction such as log *P*.[61] The resulting bitstrings can also be used to express the similarity between molecules in a fingerprint-like fashion. They observed that bioisosterically related compounds tend to higher similarity values than randomly chosen pairs of molecules. Our results (see Figure 5a) corroborate these findings: The similarity between related substances, that is, ACE inhibitors is significantly higher than compared to unrelated molecules from other databases. Within the framework of the LINGO method, the number of compared substrings is, however, dependent on the size of the molecules. Since pairs of bioisosteric compounds are usually also similar in size, in contrast to random pairs of unrelated molecules, the observed higher similarity is possibly a result of that, at least to certain degree. On the other hand, the considered compounds can all be assumed to be within the typical size range of drug-like compounds.

Lounkine and Bajorath presented a method for similarity analysis between molecules that also applies the SMILES notation.[17] Similar to our method alignments are constructed using dynamic programming and gap penalties. In their approach, however, similarities are derived from the overlap of largest common SMILES substrings that are subject to iterative fragmentation, thus producing a multitude of possible pathways. Furthermore, the alignment score is computed from only three possible values, comprising identical strings,

Table 3. Most Common Bioisosteric Substitutions Regarding One-to-One Exchanges of Atoms[a]

$-CH_3$	$-CH_2-$	$=CHR$	C_{ar}	$-OR$	$-OH$	$=O$	$-NH_2$	$-NHR$	$-NR_2$	F	Cl
Cl	C_{ar}	C_{ar}	$-CH_2-$	$-CH_2-$	$=O$	C_{ar}	$=O$	$-CH_2-$	$-CH_2-$	CH_3	CH_3
CH_2R	$=CR_2$	$-CH_2-$	$=O$	CH_3	C_{ar}	OH	OH	OR	$=CR_2$	C_{ar}	C_{ar}
OR	CH_3	$=CR_2$	$=CR_2$	NHR	CH_3	CH_3	CH_3	$=CR_2$	CH_3	Cl	$-CH_2-$
OH	OR	CH_3	CH_3	C_{ar}	NH_2	$-CH_2-$	C_{ar}	C_{ar}	CHR	$=O$	F
F	$=O$	N_{ar}	N_{ar}	$=CR_2$	$-CH_2-$	N_{ar}	$-CH_2-$	N_{ar}	OR	$-CH_2-$	OH
NH_2	NR_2	$=O$	$=CHR$	NR_2	Cl	NH_2	Cl	$=NR$	N_{ar}	OH	NH_2
Br	NHR	NR_2	CHR	N_{ar}	F	F	F	$=NH$	OH	NH_2	N_{ar}
I	N_{ar}	NHR	NR_2	OH	N_{ar}	$=CHR$	N_{ar}			N_{ar}	$=CHR$
	$=CHR$	SR	OH	SR	OR	S_{ar}	NHR			OR	$=CH_2$
		Cl	Cl	F		OR				$\equiv N$	NR_2
			S_{ar}			NHR					NHR
			OR			$\equiv N$					

[a] Shown is the sequence for the respective atoms. The relative exchange frequencies decrease from top to bottom, for example, for the methyl group more exchanges to chlorine were observed than to iodine. R denotes non-hydrogen substituents. Atoms in aromatic rings are indicated by "ar".

matching of shortened strings, and matching of largest common substrings, in decreasing order. Thus no values for replacements of atoms are included, in contrast to our methods that consequently uses exchange probabilities. Nevertheless, they obtained comparable or better recovery rates in comparison to fingerprint searches, in particular for compound classes with low diversity.

Schneider and co-workers also used dynamic programming to obtain alignments for comparing compounds.[21] Instead of applying SMILES-based substrings, each non-hydrogen atom in the molecule is assigned one of 10 so-called potential pharmacophoric points that are similar to the features used in the FTrees approach. Based on the topology a canonical PhAST-sequence is derived that is unique for each molecule and thus allows straightforward use of the Needleman–Wunsch alignment algorithm. The authors applied two different canonicalization algorithms that were modified. The required exchange matrix was constructed with respect to the normalized frequencies of these features as well as retrospective screening in their COBRA compound library. In contrast our substitution matrix is based on the actual exchange frequencies of atom types within the considered therapeutic classes, and thus should allow unbiased alignment and scoring.

Other approaches consider such bioisosteric replacements for the assessment of similarity likewise. Drug Guru comprises a (large) set of distinct reaction rules using the SMIRKS notation to perform transformations to functional groups as well as framework modifications.[34] It is interesting to note the principal distinction of bioisosteric replacements into substitutions of terminating side chains, namely, the functional groups on the one hand and alterations that concern the molecular skeleton itself on the other hand. Whereas the replacement of a carboxylate group by a tetrazole group is easily been recognized, the latter modifications comprise increasingly complex structural changes. For example, quinaprilat contains a bicyclic ring system instead of the substituted heterocyclic 5-membered ring in imidaprilat (see Figure 2). To cover all of such known modifications that occur in pairwise combinations of related drugs would thus require an enormous set of distinct modification rules. Considering the actual number of potential drug candidates on top of the actual drugs, it is moreover highly likely to overlook certain bioisosteric exchanges.

Instead of pursuing similar top-down approaches that cover molecular fragments of various size, we used a bottom-up strategy. Our smallest molecular units consist of distinct atom types that are combined to sequences. With respect to the set of 74 different atom types one may thus consider this approach as a more general strategy in comparison to the use of a more limited set of features. The feature trees approach makes use of physicochemical properties, such as hydrogen-bonding capabilities or hydrophobicity, that are assigned to either atoms or larger molecular fragments.[18] Likewise, Gillet and co-workers transform the molecule to a reduced graph on the basis of fragments that are assigned a strongly limited number of features.[24] Whereas hydrogen-bonding is seemingly unambiguously assignable to nitrogen and oxygen atoms, it is difficult to quantify aromaticity accordingly. A naphthyl group is clearly aromatic and hydrophobic as well. These properties change, however, dramatically upon introduction of hetero atoms. For example, the equally 10-membered pteridin ring system perfectly obeys Hückel's rule for aromaticity, although the 4 nitrogen atoms introduce hydrogen-bonding ability, give rise to increased polarity and decreased hydrophobicity. This reasoning furthermore motivated us to avoid any higher-level property and instead use atom types as "features" or smallest units. Moreover, atom types have been used successfully to derive molecular properties, for example, log *P*.[62] Likewise, force fields apply a limited set of atom types to describe molecules. The use of atom types instead of predefined substructures furthermore circumvents another problem. Newer chemical modifications and fragments might remain undiscovered by using a limited set of substructures, whereas atom type-based approaches are able to take these alterations into account.

A second group of programs that applies predominantly replacements of terminating groups makes use of molecular fields and shape information. They express similarity in terms of molecular shape and electrostatic potential. If interactions with the binding pocket are considered such approaches are moreover limited to cases where appropriate crystallographic data about the target is available. Likewise, the problem of finding an according scoring function becomes similar to that in docking approaches. For example, WABE produces isosteric molecules to a given query structure that contain the same heavy atom connectivity.[35] In contrast, the BROOD program considers the electrostatic similarity of the given

query fragment to corresponding analogs.[37] Furthermore, matching of molecular shape and atom types, as well as replication of the attachment geometry is used to generate the overlay with the given fragment. Unfortunately, no detailed information regarding the representation of atom types and molecular shape nor about the compound set that was used to validate the corresponding similarity function is given.

A general drawback of using shape and electrostatic information is that these properties are usually derived from the respective conformation. Thus ensembles of different molecular conformations for each molecule or fragment must be considered. The LASSO tool circumvents this shortcoming by using a so-called interacting surface point type descriptor for each molecule.[38] This descriptor is basically a vector that contains the count of 23 different properties that are derived from the hybridization and connectivities of the atoms. These comprise a more detailed representation of hydrogen-bonding capabilities, charge, hydrophobicity, π-electron, and lone electron pair characteristics than, for example, the 15 atom pair types of the CATS descriptor[25] or the 7 atom types of TOPOSIM.[51] Since these quantities are obtained from the 2D-topology of the molecules, the according descriptor is conformationally independent. The similarity between two molecules can now be calculated by two completely different approaches. The obvious way would be to compute the Tanimoto coefficient of the corresponding descriptors.[3] The LASSO tool, however, exploits further knowledge about active and assumed inactive compounds with respect to specific targets. According compound sets for each drug class were used to train according neural networks. Similarity is now obtained as the output score of the given molecule with respect to a chosen target. It is, however, debatable if the available number of active substances for certain targets was large enough to allow the use of neural networks, also regarding an adequate sampling of possible descriptor ranges. Moreover, this is the only approach where similarity between two molecules is dependent on the target. Likewise, around 300 target-specific modifications were used in the commercial package BIOISOSTER to derive a set of predefined transformation.[36] Thus new lead-like compounds can be generated from an input query, as well as searches for bioisosteric compounds in databases.

It is, however, not evident why bioisosteric modifications should be target-specific at all. The vast majority of binding pockets is hydrophobic and offers only limited possibilities for hydrogen-bonding. A strategy of lead optimization is thus to increase the molecular size to fill these cavities, thereby exploiting the entropic contribution of the displaced water molecules to the binding energy. Drugs typically possess a higher molecular weight and log *P* in comparison to nondrugs as well as to lead-like compounds.[53,63,64] Conversely, size and hydrophobicity are limiting factors to oral bioavailability. Thus, bioisosteric modifications are carried out to preserve or improve certain properties. For example, the replacement of a methyl group by chlorine does not alter log *P* notably, but usually increases metabolic stability.[65] It is obvious that the introduction of additional substituents and exchanges by larger fragments have to be carried out with respect to the shape of the binding pocket.

Sheridan and Hauberin and Bruneau have analyzed the frequencies of general chemical one-to-one exchanges in large sets of drugs and drug-like substances.[29,31] Their results show striking agreement regarding the most common replacements. Likewise, we observed that the corresponding exchange matrices for the respective targets given in Table 2 did not show pronounced mutual differences. Furthermore almost identical results were obtained for the recovery of substances within the same drug class using the complete exchange matrix compared to the matrix where the respective class was excluded (see Figure 5b−j). Pronounced differences are only visible in comparison with the exchange matrices that were exclusively derived from the respective drug classes. Drawing common conclusions is, however, complicated by the small number of drugs in certain categories. It is difficult to state a minimum number of substances that would allow the unambiguous assignment of specific differences between categories. We therefore united the exchange frequencies for all classes in one single matrix. In contrast to exchange matrices that might be class specific, the complete matrix reflects a wider range of common bioisosteric substitutions that in turn enables scaffold-hopping between different activity classes. The use of specific matrices for screening of certain activity classes corresponds to the use of profiles or position specific weight matrices in protein sequence alignment. The results using such specific matrices are, however, 2-fold. Although they allow a somewhat better enrichment of compounds in the regime of high to medium bioisosteric similarity, such adapted matrices mainly reflect exchanges with high probability in that particular activity class. Other exchanges that are successful in general might thus be disregarded. For the purpose of identifying new and likewise diverse compounds containing reasonable modifications it is therefore obvious to use the complete matrix based on all considered drug classes.

Screening results carried out against a background of drugs from Prous Science Drugs of the Future (Figure 6a−i) show that active compounds of the respective drug class are recovered at higher values of similarity compared to the fingerprint-based method, except for classes where a common substructure is highly conserved, that is, penicllins and diuretics. In contrast to other fingerprint methods that mainly encode chemical groups and substructures, the FP2 fingerprint used here encodes linear atom sequences up to 7 atoms in length (see Method section). It seems that this unbiased and more systematic encoding approach is also a cause for the worse recovery rate. This indicates, however, that our bioisosteric approach considers a larger variety of chemically meaningful modifications or rearrangements caused by scaffold hopping. A more thorough comparison between the different similarity methods, however, requires the analysis of the respective similarity distributions for each of the compound sets and from that the derivation of corresponding curves of random distributions as reference.

Since our exchange frequencies are normalized to the total number of exchanges involving a given atom type, a one-to-one comparison to the most common replacements appearing in databases is difficult, because these are not normalized. For example, we observed a higher probability for the exchange of $-OR$ against $-CH_2-$ than for $-NHR$ and $-SR$, whereas Sheridan found more exchanges involving

sulfur (see Table 3).[31] In general the most likely exchanges occur between various carbon atoms, especially the replacement of aromatic rings by aliphatic rings. Likewise introduction of nitrogen into aromatic and heterocyclic rings containing double bonds is widespread. This exchange (=CH− by =N−) was also found most often by Sheridan. After normalization to the total number of exchanges we found, however, that aromatic carbons (C_{ar}) are more likely to be converted to nonaromatic counterparts, for example, $-CH_2-$ than to aromatic nitrogens (N_{ar}). Haubertin and Bruneau showed that the most common replacements of side chains comprise the exchange of methyl by either ethyl, chlorine, methoxy, or fluorine, in decreasing order.[29] Again the frequencies of these replacements were not normalized to the actual number of side chains of the respective kind. We observed the sequence Cl > CH_2R > OR > OH > F for the corresponding exchange frequencies of methyl (see Table 3). Since CH_2R comprises, in addition to ethyl, propyl and further hydrocarbon substituents, both sequences highlight the same trend.

Normalization of the exchange frequencies also allows the analysis of the conservation of hydrogen-bonding properties upon substitution. The hydrogen-bond accepting groups −OR and $-NR_2$ are most frequently replaced by $-CH_2-$, thereby removing all possible functionality. Disruption is likewise found for −NHR fragments and carbonyl oxygens because of the replacement by methylene or aromatic carbons, respectively. Glen and Adams, however, pointed out that corresponding $>CH_2\cdots O{=}C$ interactions in protein−ligand complexes are among the most frequent in the protein databank.[4] Thus, the loss of hydrogen-bonding functionality is seemingly not always crucial. Moreover these exchange probabilities must be seen as a further indication that certain hydrogen-bonds are more favorable than others. For example, those involving the carbonyl oxygen of an ester group are energetically preferred to sp^3-hybridized oxygen.[66] Pronounced conservation of hydrogen-bond acceptor properties is only observed for −OH and $-NH_2$ that are both most frequently converted into carbonyl oxygens, thereby losing their possible function as hydrogen-bond donors. On first sight these findings seem to contradict the observed strong conservation of hydrogen-bonds that are responsible for selectivity in ligand recognition. On the other hand such specific hydrogen-bonds comprise only the minority of all possible hydrogen-bond functionalities in a typical drug. Since our approach includes normalization of all observed exchange frequencies, these highly target specific hydrogen-bond interactions contribute only little to the actual bioisosteric modifications in drugs. In turn this should cause the most reasonable drug candidates to appear as highest scoring hits in virtual screening.

CONCLUSIONS

We showed that the consistent adoption of the algorithmic concept used to assess the evolutionary relationship between protein sequences to molecules is a useful strategy to compute bioisosteric similarities. Likewise, the chemical diversity of a substance class can be estimated by calculating the median of the mutual similarities between all molecules in that class. We found that chemical replacements do not differ substantially among the investigated drug categories. Virtual screening for ACE inhibitors was performed using the "average" molecule imidaprilat in various publicly available substance databases. In all cases, we obtained known actives as well as similar inhibitors of related enzymes among the highest scoring hits. Recovery of drugs within the same activity class is dependent from the nature of the applied exchange matrix. Those derived solely from the respective drug class showed a somewhat better performance than the complete matrix that was based on bioisosteric substitutions including all classes. The results obtained with this matrix were throughout better than using a unitary matrix that corresponds to a maximum common substructure approach. For all investigated classes, the alignment-based approach recovered the majority of compounds with higher similarity values compared to fingerprint similarity. To furthermore demonstrate the feasibility for virtual screening, diazepam was used as query in the lead-like subset of the ZINC library. Comparison with the results obtained by FTrees and the fingerprint-based similarity used in PubChem showed distinct differences regarding the conservation of properties and arrangement of fragments. As a general tendency we found that chemical modifications to the query compounds closely reflect the frequency of corresponding bioisosteric exchanges in drugs. By using smaller gap penalties, extensions, contractions, and rearrangement of fragments can be intensified. These results suggest that our concept of bioisosteric similarity is helpful in rational drug design, not only for identifying substances with potential biological activity from virtual compound libraries but furthermore to assess the suitability of chemical modifications during drug optimization. We plan to make this program generally available as online application.

ACKNOWLEDGMENT

This work was supported in part by funds of the Saarland University.

Supporting Information Available: The obtained exchange matrix, as well as the SMILES strings of the compounds used to derive the matrix (including all drug classes used for screening), and those of the molecules used as additional query molecules, the complete phylogenetic tree that was obtained for the ACE inhibitors, as well as the plot showing the relationship between atom type-based similarity used as preselection criterion according to eq 2 and the bioisosteric similarity. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) van de Waterbeemd, H.; Gifford, E. ADMET *In Silico* Modelling: Towards Prediction Paradise. *Nat. Rev. Drug Discovery* **2003**, *2*, 192–204.
(2) Bajorath, J. Selected Concepts and Investigations in Compound Classification, Molecular Descriptor Analysis, and Virtual Screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882–894.
(3) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
(4) Glen, R. C.; Adams, S. E. Similarity Metrices and Descriptor Spaces−Which Combinations to Choose? *QSAR Comb. Sci.* **2006**, *25*, 1133–1142.
(5) Barnard, J. M. Substructure Searching Methods: Old and New. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 532–538.
(6) Raymond, J. W.; Gardiner, E. J.; Willett, P. RASCAL: Calculation of Graph Similarity Using Maximum Common Edge Subgraphs. *Computer J.* **2002**, *45*, 631–644.

(7) Raymond, J. W.; Blankley, C. J.; Willett, P. Comparison of Chemical Clustering Methods Using Graph- and Fingerprint-Based Similarity Methods. *J, Mol. Graph. Model.* **2003**, *21*, 421–433.
(8) Martin, Y. C.; Bures, M. G.; Danaher, E. A.; DeLazzer, J.; Lico, I.; Pavlik, P. A. A Fast New Apporach to Pharmacophore Mapping and its Application to Domapinergic and Benzodiazepine Agonists. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 83–102.
(9) Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857.
(10) Cramer, R. D. I.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
(11) Flower, D. R. On the Properties of Bit String-Based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379–386.
(12) Schuffenhauer, A.; Gillet, V. J.; Willett, P. Similarity Searching in Files of Three-Dimensional Chemical Structures: Analysis of the BIOSTER Database Using Two-Dimensional Fingerprints and Molecular Field Descriptors. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 295–307.
(13) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.
(14) Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorath, J. Design and Evaluation of a Molecular Fingerprint Involving the Transformation of Property Descriptor values into a Binary Classification Scheme. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1151–1157.
(15) Senese, C. L.; Duca, J.; Pan, D.; Hopfinger, A. J.; Tseng, Y. J. 4D-Fingerprints, Universal QSAR and QSPR Descriptors. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1526–1539.
(16) Wagener, M.; Lommerse, J. P. M. The Quest for Bioisosteric Replacements. *J. Chem. Inf. Model.* **2006**, *46*, 677–685.
(17) Lounkine, E.; Bajorath, J. Core Trees and Consensus Fragment Sequences for Molecular Representation and Similarity Analysis. *J. Chem. Inf. Model.* **2008**, *48*, 1161–1166.
(18) 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.
(19) Rarey, M.; Stahl, M. Similarity Searching in large Combinatorial Chemistry Spaces. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 497–520.
(20) Fischer, J. R.; Rarey, M. SwiFT: An Index Structure for Reduced Graph Descriptors in Virtual Screening and Clustering. *J. Chem. Inf. Model.* **2007**, *47*, 1341–1353; http://public.zbh.uni-hamburg.de/ftrees/query.py (accessed May 12, 2008).
(21) Hänke, V.; Hofmann, B.; Grgat, T.; Proschak, E.; Steinhilber, D.; Schneider, G. PhAST: Pharmacophore Alignment Search Tool. *J. Comput. Chem.* **2009**, *30*, 761–771.
(22) Fröhlich, H.; Wegner, J. K.; Sieker, F.; Zell, A. Kernel Functions for Attributed Molecular Graphs—A New Similarity-Based Approach to ADME Prediction in Classification and Regression. *QSAR Comb. Sci.* **2006**, *25*, 317–326.
(23) Wegner, J. K.; Fröhlich, H.; Mielenz, H. M.; Zell, A. Data and Graph Mining in Chemical Space for ADME and Activity Data Sets. *QSAR Comb. Sci.* **2006**, *25*, 205–220.
(24) Gillet, V. J.; Willett, P.; Bradshaw, J. Similarity Searching Using Reduced Graphs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 338–345.
(25) Schneider, G.; Werner, N.; Giller, T.; Schmid, G. "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed.* **1999**, *38*, 2894–2896.
(26) Weber, A.; Teckentrup, A.; Briem, H. Flexsim-R: A Virtual Affinity Fingerprint Descriptor to Calculate Similarities of Functional Groups. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 903–916.
(27) Patani, G. A.; LaVoie, E. J. Bioisosterism: A Rational Approach in Drug Design. *Chem. Rev.* **1996**, *96*, 3147–3176.
(28) Lima, L. M.; Barreiro, E. J. Bioisosterism: A Useful Strategy for Molecular Modification and Drug Design. *Curr. Med. Chem.* **2005**, *12*, 23–49.
(29) Haubertin, D. Y.; Bruneau, P. A Database of Historically-Observed Chemical Replacements. *J. Chem. Inf. Model.* **2007**, *47*, 1294–1302.
(30) Ujvary, I. BIOSTER—A Database of Structurally Analogous Compounds. *Pestic. Sci.* **1997**, *51*, 92–95.
(31) Sheridan, R. P. The Most Common Chemical Replacements in Drug-like Compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 103–108.
(32) Ertl, P. Cheminformatics Analysis of Organic Substituents: Identification of the Most Common Substituents, Calculation of Substituent Properties, and Automatic Identification of Drug-like Bioisosteric Groups. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 374–380.
(33) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP—Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Priviledged Molecular Fragements with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.
(34) Stewart, K. D.; Shiroda, M.; James, C. A. Drug Guru: A Computer Software Program for Drug Design Using Medicinal Chemistry Rules. *Bioorg. Med. Chem.* **2006**, *14*, 7011–7022.
(35) Sayle, R.; Nicholls, A. Electrostatic Evaluation of Isosteric Analogues. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 191–208.
(36) Balakin, K. V.; Tkachenko, S. E.; Okun, I.; Skorenko, A. V.; Savchuk, N. P.; Ivanshchenko, A. A.; Nikolsky, Y. Bioisosteric Morphing in Primary Hit Optimization. *Chim. Oggi—Chem. Today* **2004**, *22*, 15–18.
(37) *Brood: Fragment replacement for Medicinal Chemistry*, version 1.1.1; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2007; http://www.eyesopen.com.
(38) Reid, D.; Sadjad, B. S.; Zsoldos, Z.; Simon, A. LASSO-Ligand Activity by Surface Similarity Order: A new Tool for Ligand Based Virtual Screening. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 479–487.
(39) Dayhoff, M. O.; Schwartz, R. M.; Orcutt, B. C. A Model of Evolutionary Change in Proteins *Atlas of Protein Sequence and Structure*; National Biomedical Research Foundation: Washington, DC, 1978; pp345−352.
(40) Henikoff, S.; Henikoff, J. G. Amino Acid Substitution Matrices from Protein Blocks. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 10195–10919.
(41) Needleman, S. B.; Wunsch, C. D. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of two Proteins. *J. Mol. Biol.* **1970**, *48*, 443–453.
(42) Smith, T. F.; Waterman, M. S. Identification of Common Molecular Sequences. *J. Mol. Biol.* **1981**, *147*, 195–197.
(43) Weininger, D. SMILES, A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
(44) Daylight Chemical Information Systems Inc., Suite 550, Aliso Viejo, CA 92656. For full details of SMILES and SMARTS, see http://www.daylight.com (accessed May 12, 2008).
(45) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.
(46) World Health Organization. The Use of Stems in the Selection of International Nonproprietary Names (INN) for Pharmaceutical Substances. http://www.who.int/medicines/services/inn/en/ (accessed Oct 18, 2007).
(47) The PubChem Database at the National Center for Biotechnology Information. http://pubchem.ncbi.nlm.nih.gov/ (accessed Oct 18, 2007).
(48) Oellien, F.; Nicklaus, M. C. Online SMILES Translator.http://cactus.nci.nih.gov/translate (accessed Jan 7, 2008).
(49) Saitou, N.; Nei, M. The Neighbor-Joining Methods: A new Method for Reconstructing Phylogenetic Trees. *Mol. Biol. Evol.* **1987**, *4*, 406–425.
(50) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure−Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.
(51) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical Similarity Using Physiochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118–127.
(52) Irwin, J. J.; Shoichet, B. K. ZINC—A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
(53) Schneider, N.; Jäckels, C.; Andres, C.; Hutter, M. C. Gradual in Silico Filtering for Druglike Substances. *J. Chem. Inf. Model.* **2008**, *48*, 613–628.
(54) Sigma-Aldrich Co. http://www.sigmaaldrich.com (accessed Feb 27, 2008).
(55) Berman, M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
(56) *Prous Science Drugs of the Future*; Thompson Reuters: Philadelphia, PA, 2008.
(57) Guha, R.; Howard, M. T.; Hutchison, G. R.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Wegner, J. K.; Willighagen, E. The Blue Obelisk—Interoperability in Chemical Informatics. *J. Chem. Inf. Model.* **2006**, *46*, 991–998.
(58) *The Open Babel Package*; version 2.2.0b4. http://openbabel.sourceforge.net/ (accessed May 6, 2008).
(59) Acharya, K. R.; Sturrock, E. D.; Riordan, J. F.; Ehlers, M. R. W. ACE Revisited: A New Target for Structure-Based Drug Design. *Nat. Rev. Drug Discovery* **2003**, *2*, 891–902.
(60) Tovar, A.; Eckert, H.; Bajorath, J. Comparison of 2D Fingerprint Methods for Multiple-Template Similarity Searching on Compound Activity Classes of Increasing Structural Diversity. *ChemMedChem* **2007**, *2*, 208–217.
(61) Vidal, D.; Thormann, M.; Pons, M. LINGO, an Efficient Holographic Text Based Method To Calculate Biophysical Properties and Intermolecular Similarity. *J. Chem. Inf. Model.* **2005**, *45*, 386–393.
(62) Viswanadhan, V. N.; Ghose, A. K.; Rebankar, G. R.; Robins, R. K. Atomic Physicochemical Parameters for Three Dimensional Structure

Directed Quantitative Structure−Activity Relationships. 4. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally Occuring Nucleoside Antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163–172.

(63) Hutter, M. C. Separating Drugs from Nondrugs: A Statistical Approach Using Atom Pair Distributions. *J. Chem. Inf. Model.* **2007**, *47*, 186–194.

(64) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is There a Difference between Leads and Drugs? A Historical Perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308–1315.

(65) Leach, A. G.; Jones, H. D.; Cosgrove, D. A.; Kenny, P. W.; Ruston, L.; MacFaul, P.; Wood, J. M.; Colclough, N.; Law, B. Matched Molecular Pairs as a Guide in the Optimization of Pharmaceutical Properties: A Study of Aqueous Solubility, Plasma Protein Binding and Oral Exposure. *J. Med. Chem.* **2006**, *49*, 6672–6682.

(66) Böhm, H.-J.; Klebe, G.; Brode, S.; Hesse, U. Oxygen and Nitrogen in Competitive Situations: Which Is The Hydrogen-Bond Acceptor? *Chem.—Eur. J.* **1996**, *2*, 1509–1513.

CI8003418