Similarity Metrics for Ligands Reflecting the Similarity of the Target Proteins

Ansgar Schuffenhauer,*,† Philipp Floersheim,‡ Pierre Acklin,† and Edgar Jacoby†

Novartis Pharma AG, Lead Discovery Center, Compound Management and Computation Unit, and Novartis Pharma AG, Nervous System Research, CH-4002 Basel, Switzerland

Received July 26, 2002

In this study we evaluate how far the scope of similarity searching can be extended to identify not only ligands binding to the same target as the reference ligand(s) but also ligands of other homologous targets without initially known ligands. This "homology-based similarity searching" requires molecular representations reflecting the ability of a molecule to interact with target proteins. The Similog keys, which are introduced here as a new molecular representation, were designed to fulfill such requirements. They are based only on the molecular constitution and are counts of atom triplets. Each triplet is characterized by the graph distances and the types of its atoms. The atom-typing scheme classifies each atom by its function as H-bond donor or acceptor and by its electronegativity and bulkiness. In this study the Similog keys are investigated in retrospective in silico screening experiments and compared with other conformation independent molecular representations. Studied were molecules of the MDDR database for which the activity data was augmented by standardized target classification information from public protein classification databases. The MDDR molecule set was split randomly into two halves. The first half formed the candidate set. Ligands of four targets (dopamine D2 receptor, opioid δ-receptor, factor Xa serine protease, and progesterone receptor) were taken from the second half to form the respective reference sets. Different similarity calculation methods are used to rank the molecules of the candidate set by their similarity to each of the four reference sets. The accumulated counts of molecules binding to the reference target and groups of targets with decreasing homology to it were examined as a function of the similarity rank for each reference set and similarity method. In summary, similarity searching based on Unity 2D-fingerprints or Similog keys are found to be equally effective in the identification of molecules binding to the same target as the reference set. However, the application of the Similog keys is more effective in comparison with the other investigated methods in the identification of ligands binding to any target belonging to the same family as the reference target. We attribute this superiority to the fact that the Similog keys provide a generalization of the chemical elements and that the keys are counted instead of merely noting their presence or absence in a binary form. The second most effective molecular representation are the occurrence counts of the public ISIS key fragments, which like the Similog method, incorporates key counting as well as a generalization of the chemical elements. The results obtained suggest that ligands for a new target can be identified by the following three-step procedure: 1. Select at least one target with known ligands which is homologous to the new target. 2. Combine the known ligands of the selected target(s) to a reference set. 3. Search candidate ligands for the new targets by their similarity to the reference set using the Similog method. This clearly enlarges the scope of similarity searching from the classical application for a single target to the identification of candidate ligands for whole target families and is expected to be of key utility for further systematic chemogenomics exploration of previously well explored target families.

INTRODUCTION

Similarity searching is a well-established method for the identification of new ligands of a biological target. In its typical application (Figure 1) single ligands with known biological activity are used to find ligands binding to the same target as the reference ligand. This method has the disadvantage that it is only applicable later in the discovery process when at least one ligand is known. However, in a recent publication[1] we introduced a target-based ontology for pharmaceutical ligands and discussed how the similarity by protein sequence and function between a new biological target and existing targets with well-known ligands can be exploited to discover compounds with biological activity related to new targets. The strategy uses a sequence similarity-based classification scheme to find among targets with several known ligands that target which is most similar to the new target for which ligands are to be found. Following the "structure−activity relationship homology" concept, the reference set formed by the ligands of the previously investigated target(s) is assumed to contain an implicit pharmacophore hypothesis not only valid for a single target but also for a whole family of targets.[2] Conservation of the sequence encoding the binding site within a target family leads putatively also to conservation of the shape and physicochemical properties of the ligand binding sites, resulting in a similarity of the structural requirements for

* Corresponding author phone: +41 61 32 45385; fax: +41 61 3242395; e-mail: ansgar.schuffenhauer@pharma.novartis.com.
† Novartis Pharma AG, Drug Discovery Center.
‡ Novartis Pharma AG, Nervous System Research.

10.1021/ci025569t CCC: $25.00
Published on Web 12/13/2002

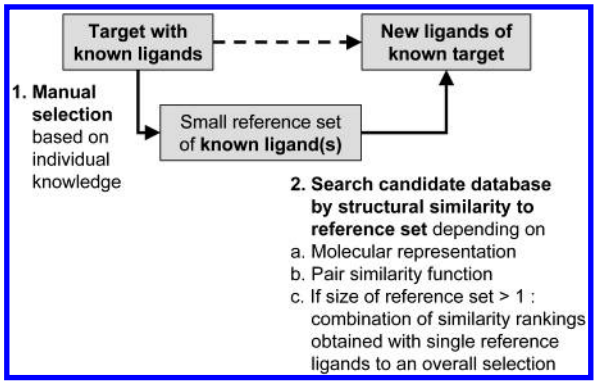

Figure 1. "Classical" similarity searching for protein ligands.

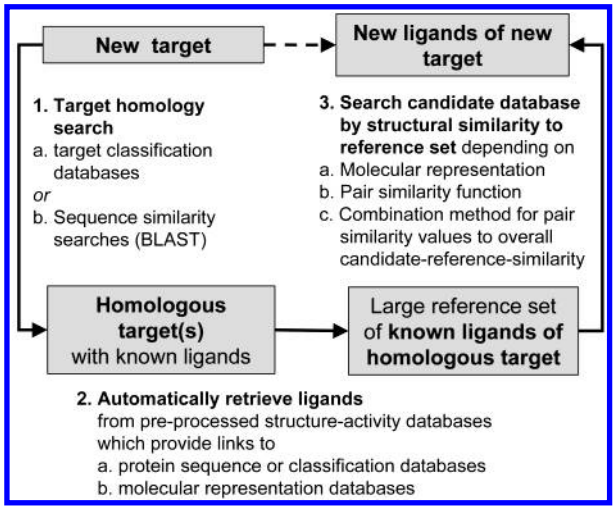

Figure 2. Homology-based similarity searching for protein ligands.

ligands.[3] Thus, molecules similar to molecules of a reference set formed by ligands of a previously investigated target can be expected to have some activity also on other targets within the same family. If L_1 is a ligand of a target T_1, whereas L_2 is a ligand of a target T_2, and T_1 is homologous to T_2 without necessarily being identical, then we will call the ligands L_1 and L_2 "activity homologues". With this definition one can rewrite the working hypothesis as the following: Activity homologues are expected to be structurally similar. The closer the activity homology of the ligands is, the higher the structural similarity of the ligands is expected to be. Activity as used here can mean ligand binding and any resulting effects such as agonism, antagonism, or inhibition (Figure 2).

One prerequisite for homology-based similarity searching is the integration of molecular structure, receptor binding, protein sequence, and annotation data in order to allow the automatic retrieval of reference sets. This data integration is still far from being completed. In addition, the requirements for molecular representations used to calculate the similarity between molecular structures in this method are high: They must represent the implicit pharmacophore hypothesis (or hypotheses) contained in the reference set. Our investigations with the Unity 2D-fingerprint[4] were encouraging enough to pursue this kind of approach further.[1] However, in another study[5] only a modest correlation between the target protein sequence similarity and the similarity of the corresponding ligands based on a fingerprint representation was found. As in that study only a very short (199 bits) and basic fingerprint, mainly based on the presence and absence of atom types, was used to calculate the molecular similarity, we aimed at different representations focusing on a ligand's ability to interact with proteins rather than on its molecular connectivity. Molecular fields are one possible representation to describe the scaffold independent 3D-receptor binding properties. However, similarity calculation based on molecular fields is computationally expensive,[6,7] because it requires the alignment of conformations. Representations based on pairs, triplets, or quadruplets of atoms characterized by distances in the 3D-space and types of atoms encoding molecular interaction properties have become increasingly popular.[8−13] To be used meaningfully, these conformation dependent representations require either the generation of multiple conformations representing approximately the total conformational space of the molecule, which also involves a lot of computational effort, or the determination of the relevant binding conformation which depends on the structure of the target protein. From a theoretical point of view, representations based on the stereochemical configuration of a molecule as the constant part of its structure would be desirable, but practical problems arise especially when using historical structure databases in which the stereochemical information is often incomplete and inconsistent. These problems can be avoided by using representations based on the connectivity only. The conformation dependent distance in the 3D-space can be replaced by the graph distance between two atoms depending only on the connectivity. This is done in case of the conformation independent Similog keys which were developed and applied successfully for over 10 years within Sandoz and Novartis.[14] These keys are described here. In contrast to the conceptually similar CATS descriptors,[15] the Similog keys are based on atom triplets instead of atom pairs. In addition, the atom-typing scheme of Similog is based on four atom properties and differs from the atom-typing scheme of CATS which is based on five atom properties. Atom triplets characterized by graph distances, but in contrast to the Similog keys without consideration of atom types, were also used in a representation of molecular shape only.[16]

In the retrospective study presented here, the suitability of the Similog keys for homology-based similarity searching was evaluated in comparison with other connectivity based representations. Four important target families, i.e., amine and peptide binding G-protein-coupled receptors, nuclear receptors, and endopeptidases, are covered by this study to ensure that the obtained results are not only specifically valid for one protein family.

METHODS

The Similog Keys. The Similog keys are built according to a compact "DABE" atom-typing scheme which is based on the following four atom properties: potential hydrogen bond **D**onor; potential hydrogen bond **A**cceptor; **B**ulkiness; and **E**lectropositivity.

The electropositivity is used here as an estimate of the ability to undergo lipophilic interactions. Acidity and basicity, which are included in some of the other reported pharmacophore types,[9,15] are often correlated with donor- and acceptor properties and are therefore not taken into explicit representation in order to avoid redundance. For each non-H atom the presence or absence of these four properties is encoded in the form of an atom key which is a string of

Table 1: Hydrogen Bonding Properties, van der Waals Radii, and Electronegativity Assigned to the Sybyl Atom Types[a]

Sybyl atom type	D	A	R_{VDW}	e	Sybyl atom type	D	A	R_{VDW}	e
H	-	-	1.08	2.1	S.o	-	-	1.7	2.5
Li	-	-	0.6	1.0	S.o2	-	-	1.7	2.5
B.2	-	-	1.60	2.5	Cl	-	-	1.65	3.0
B.3	-	-	1.60	2.5	K	-	-	1.33	0.8
C.3	-	-	1.52	2.5	Ca	-	-	0.99	1.0
C.2	-	-	1.53	2.5	Fe.3	-	-	2.00	1.0
C.1	-	-	1.54	2.5	Fe.2	-	-	2.00	1.0
C.ar	-	-	1.53	2.5	Co.3	-	-	2.00	1.0
N.3	+	+	1.45	3.0	Co.2	-	-	2.00	1.0
N.2	+	+	1.48	3.0	Zn.2	-	-	2.00	1.0
N.1	-	+	1.5	3.0	Zn.1	-	-	2.00	1.0
N.ar	-	+	1.48	3.0	As.5	-	-	1.8	2.8
N.am	+	-	1.45	3.0	As.3	-	-	1.8	2.8
N.4	+	-	1.45	1.0	Se.3	-	-	1.9	2.5
N.2+	+	-	1.48	1.0	Se.o	-	-	1.9	2.5
N.o	-	-	1.5	3.0	Se.o2	-	-	1.9	2.5
N.o2	+	-	1.5	3.0	Br	-	-	1.8	2.8
N.pl3	+	-	1.5	3.0	Ag.2	-	-	2.00	1.0
N.lin	-	-	1.48	1.0	Ag.1	-	-	2.00	1.0
O.3	+	+	1.36	3.5	Cd.2	-	-	2.00	1.0
O.2	-	+	1.36	3.5	Cd.1	-	-	2.00	1.0
O.2+	-	-	1.36	1.0	Sn.2	-	-	3.00	1.0
F	-	-	1.3	4.0	Sn	-	-	3.00	1.0
Na	-	-	0.95	0.9	Te.2	-	-	2.05	2.5
Al	-	-	2.05	1.5	Te	-	-	2.05	2.5
Si	-	-	2.1	2.8	I	-	-	2.05	2.5
Si.2	-	-	2.1	2.8	Au.3	-	-	2.00	1.0
P.5	-	-	1.75	2.1	Hg.2	-	-	3.00	1.0
P.3	-	-	1.75	2.1	Hg.1	-	-	3.00	1.0
P.o2	-	-	1.75	2.1	Tl	-	-	3.00	1.0
S.3	-	-	1.7	2.5	Du	-	-	0.0	0.0
S.2	-	-	1.72	2.5	LP	-	-	0.85	0.0

[a] D: hydrogen bond donor (if there is at least one hydrogen attached), A: hydrogen bond acceptor, R_{VDW}: van der Waals radii, e: electronegativity. N.4 (ammonium) and N.2+ are deprotonated during preprocessing of the molecules and become eventually N.3 and N.2, if they have hydrogens connected to them. Thus, N.4 and N.2+ never occur as donor in the Similog keys of neutralized molecules.

four DABE bits. For instance, an alcohol oxygen atom having the property of donating or accepting a hydrogen bond is encoded by the atom key 1100 or the central carbon atom of a bulky *tert*-butoxy group by the atom key 0010. To determine the atom properties the Sybyl[17] atom-typing scheme is applied to the neutral, uncharged molecules. The hydrogen-bonding properties of an atom are read directly from an extended Sybyl atom definition data table (see Table 1) and the number of hydrogens (implicitly) attached to the atom. To determine the bulkiness and electropositivity of an atom *i*, its non-hydrogen neighbor atoms *j* have to be taken into account. The definition of bulkiness is based on the van der Waals radii r_{VDW} associated with the atom types. An atom *i* has the property of bulkiness if

$$r_i^3 + \sum_j r_j^3 > 10 \text{ Å}^3 \qquad (1)$$

As an example of using in eq 1 the threshold volume of 10 Å^3, a methine but not a methylene carbon atom is assigned the property of bulkiness. The definition of electropositivity is based on Pauling's electronegativity[18] *e* associated with the Sybyl atom types. An atom *i* has the property of electropositivity if

$$(e_i \le 2.5) \text{ and } (e_j \le 2.5 \text{ for all } j) \qquad (2)$$

Table 2: Graph Distance Intervals[a]

interval	graph distance
2	2,3
4	4,5
6	6,7
8	≥8

[a] No interval is defined for the graph distance 1. Atom triplets with two or three bonded atoms are intentionally omitted in Similog because the directly connected atoms are already taken into account when the atom keys are determined.

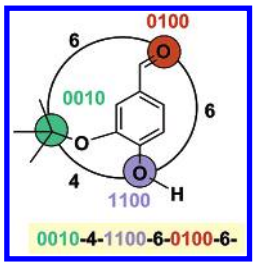

Figure 3. Example of a Similog key.

As a special case, carbon atoms of methyl groups are always taken as electropositive because of their contribution to lipophilicity. As an example of using in eq 2 the threshold electronegativity of 2.5, the substituted carbon atom of thiophenol but not of phenol is assigned the property of electropositivity.

Of the 16 combinatorially possible atom keys the following six keys 0101, 0111, 1001, 1011, 1101, and 1111 are not realized in neutral organic molecules because hydrogen bond donor or acceptor atoms have always an $E > 2.5$ and cannot become electropositive.

In the next step each triplet of atoms present in a molecule is represented by the keys of their three atoms and the three graph distances between them. Atom triplets containing one or more atoms without any of the DABE binding properties (atom key 0000) are omitted. The graph distance between two atoms, often called "topological distance", is defined as the count of bonds in the shortest path connecting the two atoms. The graph distances used for Similog are mapped to four intervals (Table 2).

The notation of the so obtained triple keys for the atoms *i*, *j*, *k* is

$$\text{DABE}(i) - \text{distance}(i,j) - \text{DABE}(j) - \text{distance}(j,k) - \text{DABE}(k) - \text{distance}(k,i)$$

In view of uniqueness, we take from the six equivalent notations of a triple key the lexicographically smallest one. An example of an atom triplet and the corresponding Similog key is shown in Figure 3.

Enumeration of all triple keys containing the nine remaining atom keys under consideration of symmetry and the triangle inequality gives 8031 possible triple keys. In total, only some 5989 triple keys have been found earlier in structures of the Novartis corporate database, and these triple keys are taken and alphabetically sorted to form the triple key basis of the molecular Similog representations as follows in the final step. The Similog representation of a molecule is formed by the vector of the occurrences of the sorted triple keys in the represented structure. This leads to a vector size equal to the size of the underlying triple key basis and to

vector elements, equal to zero or nonzero, representing the nonoccurrence or respective occurrence of those atom triplets present in the given molecule. In our case, the size of the underlying basis is 5989, and, as a tolerable consequence, some atom triplets absent from the basis may have been neglected in the molecular representations.

Other Representations and Similarity Measures Used. As a benchmark, we compared the results obtained from similarity searches using the Similog keys with those obtained using the Unity 2D-fingerprints[4] which denote the presence or absence of combinatorially enumerated connections paths in a bit string. As another key occurrence counting representation vector like the Similog keys we used the occurrence counts of the ISIS public keys[19] called further on "ISIS public key count". To determine the counts of the ISIS public keys the corresponding structural fragments were translated into SLN (Sybyl[17] Line Notation) strings in a Unity fingerprint definition file (provided as Supporting Information). For each fragment a range of 16 bits was allowed, so that the key occurrence count was capped at 15. On average over the keys, this limit was only reached in 0.2% of the MDDR molecules. For no key the percentage of MDDR molecules reaching the counting limit was higher than 1.5%, with the exception of the aromatic bond count for which the bit range was doubled as a consequence. The fingerprint definition file was used to create an intermediate fingerprint with dbmkfprint, the fingerprint generator of Unity.[4] From this intermediate fingerprint the counts of the structural keys in the fingerprint definition file were extracted and further on used as a vector of integers. Public ISIS key fragments defining "inorganic" elements or isotopes were omitted. Some other bit positions in the ISIS public key were already used for rudimentary counting, e.g., there is a bit to set for the presence of oxygen in general and additional bit for the presence of at least two oxygens. In our fingerprint definition file such bit positions were merged into one counter. Additional molecular representations included in our study were the E-state keys (electrotopological indices) by Kier and Hall[20] and a vector of topological descriptors, consisting of the Wiener,[21] Zagreb,[22] Balaban,[23] Phi,[24] and the three Kappa indices.[25] These topological and the electrotopological indices were calculated using the Cerius[II] software.[26]

As similarity function for a pair of molecules we used in this study always the Tanimoto coefficient, which in studies by Willett[27] et al. showed the best performance compared to Euclidean distance, cosine, and Dice coefficients for similarity searching in structural databases and is implemented in most of the publicly available structure similarity searching programs. However, the relevance of a similarity function depends on the molecular representation and its intended use.

Similarity Comparison Methods for Screening with a Multimolecule Reference Set. To use a set of molecules instead of a single molecule as a query for similarity searching, a method to calculate the similarity or distance of a candidate molecule to a reference (query) set of N molecules needs to be defined. This can be done in several ways. Let the representation vector of molecule i in the reference set R be r_i and the representation vector of the candidate molecule be x. The similarity between two representation vectors x and y is calculated with the function $S(x, y)$. Following the formalism introduced by Sheridan[28]

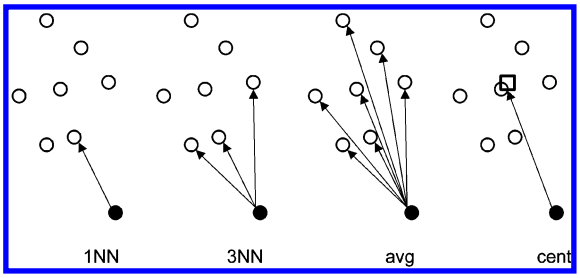

Figure 4. Similarity comparison methods.

for the representation of compound mixtures, we may then calculate the similarity between the candidate molecule and the reference set as the similarity of the candidate and the center of the reference set according to

$$S(x,R) = S(x,\bar{r}) \quad \bar{r} = \frac{1}{N}\sum_{N} r_i \tag{3}$$

This method will be further named *centroid* (in short *cent*) method, as the centroid vector $\bar{r}$ is used as a representative of the reference set. Another way to calculate the similarity between a candidate and a reference set is based on the average over single distances according to

$$S(x,R) = \frac{1}{N}\sum_{N} S(x,r_i) \tag{4}$$

This method will be further referred to as *average* (in short *avg*) method, as one takes the average similarity between all reference vectors and the candidate vector x. One may calculate the average similarity not over all similarity values $S(x, r_i)$ but only of the k highest values representing the similarities between x and its k nearest neighbors in the reference set R. Corresponding to the value of k, these methods will be referred to as k nearest neighbor, or in short *k-NN*, methods. *1NN* is a particular *k-NN* method in which case the similarity between R and x is defined as similarity between x and its nearest neighbor in R. When comparing the *cent* method to all others, it should be noted that the *cent* method requires only one computation of the similarity coefficient in order to express the similarity of a multimolecule reference set with a candidate structure, whereas for all the other methods the similarity coefficient has to be computed for each of the members of the reference set. The differences between the similarity methods are geometrically illustrated in Figure 4.

The candidate structure can be interpreted as a cluster with only one member which center is represented by the vector x. From this point of view, the similarity as defined by the *cent* method is analogue to the joining condition for the clusters x and R in Ward's[29] clustering algorithm, since it also depends on the centroid distance. Equally the similarity as defined by the *1NN* method is applied in the cluster joining condition in the Jarvis-Patrick[30] algorithm in its dependence on nearest neighbors.

Centroid Fingerprints for a Molecule Reference Set. In the case of representations in the form of numeric vectors the definition of the corresponding centroid vector of a vector set is according to eq 3. If one interprets a binary fingerprint as a vector of the numeric values 0.0 and 1.0, the same definition may be also used for the centroid of a set of binary

fingerprints. The *j*th component of the centroid fingerprint $\bar{r}$ having any value between 0.0 and 1.0 represents then the fraction of fingerprints in the reference set *R* in which the bit in position *j* is set. The Tanimoto coefficient between the centroid fingerprint of the reference and the candidate fingerprint is then calculated using the general definition of the Tanimoto coefficient for numerical instead of binary vectors.[27] For computational and storage reasons, the numerical accuracy for each vector element was reduced to 1/255, so that a vector element of the centroid vector requires the storage of only one byte memory.

Design of the Retrospective In Silico Screening Experiments. An ideal data set for a retrospective evaluation study would cover a large set of diverse molecules together with a comprehensive set of target proteins and contain activity data for each molecule-target combination. Since such a data set is currently not publicly available, we chose to examine the MDDR[31] database which we enhanced by an annotation of the contained molecules based on target classification information as described previously.[1] Structure−activity databases such as the MDDR, which are compiled from patent literature, cover a wide range of diverse ligands and targets. However, the absence of reported activity on a specific target does not necessary imply that a molecule is actually inactive but can mean also that the activity was never determined. The risk to miss yet undiscovered activity on the targets studied was reduced by choosing such targets for the study which are well-known for a longer time. This is making it likelier that a large fraction of the reported compounds was actually examined for activity on them. The molecular representations to be studied were calculated for all purely organic molecules with fully defined constitution (105 051) in the MDDR database in their neutral form. This set was randomly split into two halves. For each target to be studied a reference set of molecules was extracted from the first half of the database using the target classification information with which the ligands were annotated. Each of the reference sets contains compounds binding to the target regardless of their agonistic, antagonistic, or inhibitory activity. These reference sets were used as queries to search the other half of the MDDR database forming the candidate set. The representation vectors except the purely binary Unity 2D-fingerprints were normalized to avoid the dominance of keys related to frequently occurring functional groups in the similarity measure. As several normalization methods for structural key occurrence count representations lead to comparable retrieval results we used the well-established Z-score.[32] Consequently, for each representation normalization was done by subtraction of the corresponding average vector and subsequent division of the elements by their standard deviations obtained from the representation vectors of the candidate set remaining constant in this study.

With each reference set and for each combination of molecular representation and similarity comparison methods a similarity search of the candidate set was carried out. The resulting similarity rankings of the candidate sets were analyzed in several steps as depicted for the dopamine D2 reference set in Figure 5. The accumulation of structures binding to the reference target was examined as a function of the similarity rank in the form of cumulative recall curves.[33] Then the structures binding to the reference target were removed from the ranked candidate set and the

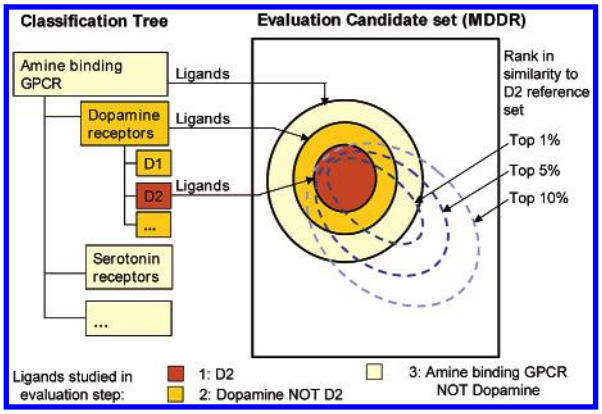

Figure 5. Classification tree of amine binding GPCRs and diagrams of the corresponding candidate ligand sets (straight lines) and their relation to the sets obtained by similarity rankings (dashed line). The candidate set is examined consecutively for ligands with a decreasing activity homology to the reference set.

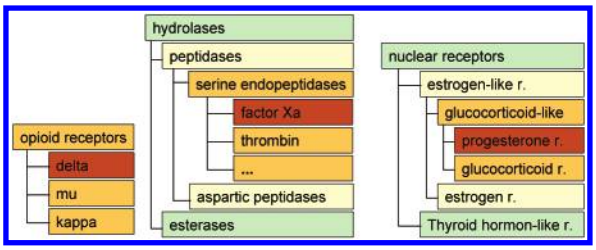

Figure 6. Classification trees for reference targets opioid δ-receptor, coagulation factor Xa, and progesterone receptor. Only selected targets proteins are shown for illustration.

accumulation of the next closest activity homologues of the reference target, which were still present in the remaining ranking, was examined. In the dopamine D2 example, the accumulation of all ligands of any dopamine receptor except D2 ligands was examined. These hits were then removed, and the previous step was repeated. In the D2 example this time the ranked candidate set was scanned for all amine binding GPCR (G-protein coupled receptor) ligands except dopamine receptor ligands, and their accumulation was evaluated. Molecules binding to more than one target were only evaluated once as ligand of that target with the highest homology to the reference target and then removed from the ranking. Therefore, the accumulation of candidate ligands with lower activity homology does not include unselective ligands binding also to targets with a higher homology to the reference. For example, reported unselective ligands of the D2 and the D1 receptor were evaluated together with the selective D2 receptor ligands in the first step; they were then removed from the ranking and not included in the accumulation of ligands binding to any dopamine receptor except the D2 receptor. The analysis described above emulated searches for ligands with decreasing activity homology to the reference set. It is expected that the degree of accumulations decreases at each evaluation step with the decreasing homology between the reference target and the targets examined. The results obtained with the other reference targets (opioid δ-receptor, coagulation factor Xa, and progesterone receptor) were analyzed in the same way using the classification information shown in Figure 6. The selected targets cover together the classes of amine and

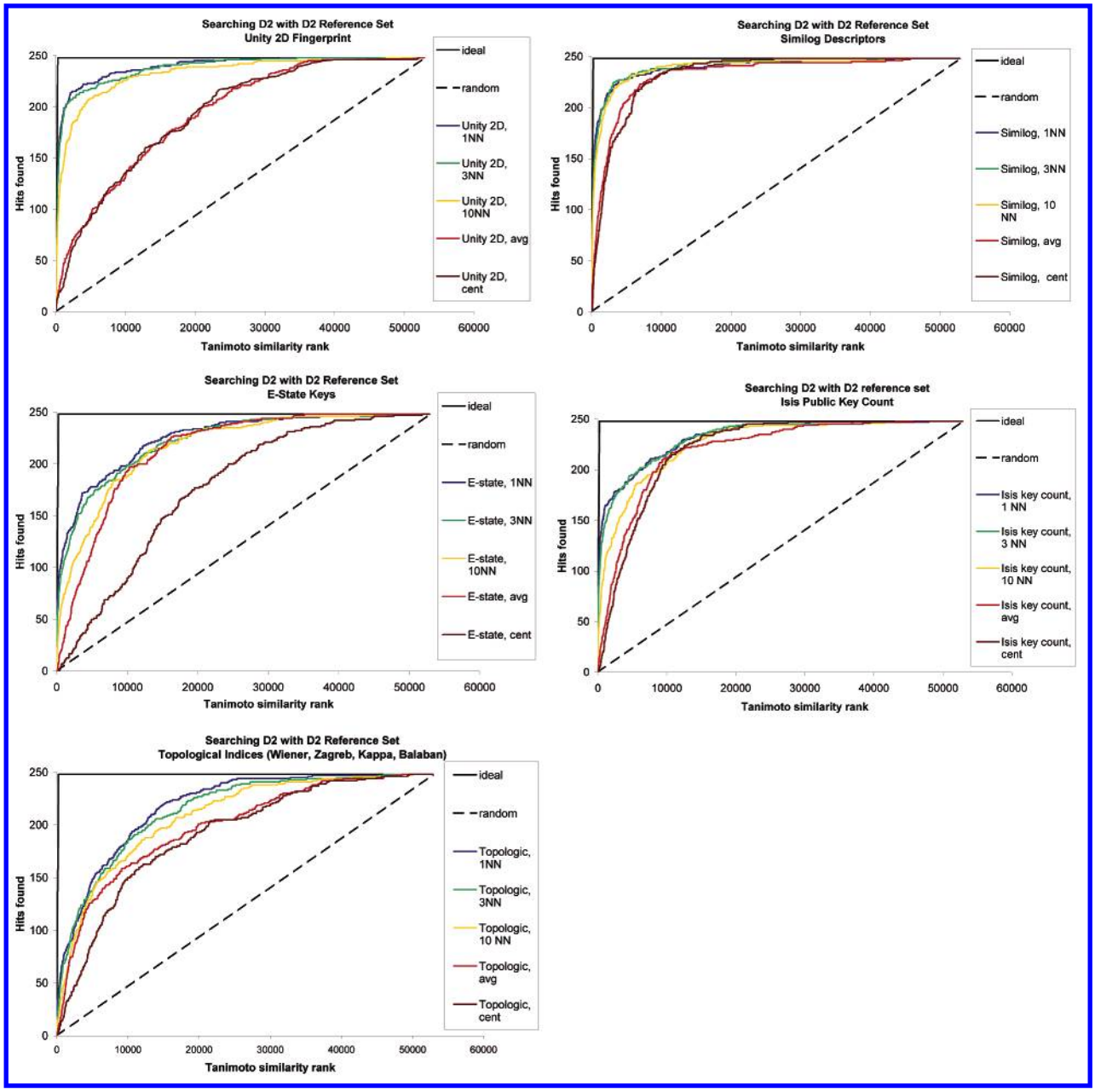

Figure 7. Searching D2 receptor ligands with a D2 reference set. Reference compounds and evaluated compounds are binding to the same target.

peptide binding GPCRs, endopeptidases, and nuclear receptors (NR).

RESULTS

The dopamine D2 receptor was examined as the first of four targets. The following five molecular representations were used: Unity 2D-fingerprints, Similog keys, E-state keys, ISIS public key count, and topological indices. For each representation all of the following similarity comparison methods were used for the comparison of a candidate molecules with the reference set: *1NN*, *3NN*, *10NN*, *avg*, and *cent*. The resulting similarity rankings were analyzed with respect to the accumulation of candidate molecules known as ligands of 1. the dopamine D2 receptor; 2. any dopamine receptor except dopamine D2; and 3. any amine binding class A GPCR except all dopamine receptors.

With each target, the accumulation was compared to the theoretical average accumulation by random selection, in the figures denoted as "random" and the maximum possible accumulation obtained by an ideal searching method, in the figures denoted as "ideal". The results are shown in Figures 7−9.

The accumulations obtained with selected combinations of representations and similarity comparison methods (Unity 2D/*1NN*, Similog/*1NN*, Similog/*cent*, E-state/*1NN*, ISIS key count/*1NN*, ISIS key count/*cent*, topological/*1NN*) are shown in Table 3. These combinations contain the most effective similarity comparison method(s) in the case of the D2 receptor for each representation and were used to examine the other reference targets. The similarity rankings obtained with factor Xa inhibitors were examined for factor Xa inhibitors (Figure 10), inhibitors of serine endopeptidases except factor Xa (Figure 11), and inhibitors of non-serine peptidases (Figure 12). The rankings obtained with the progesterone receptor ligands as reference set were examined

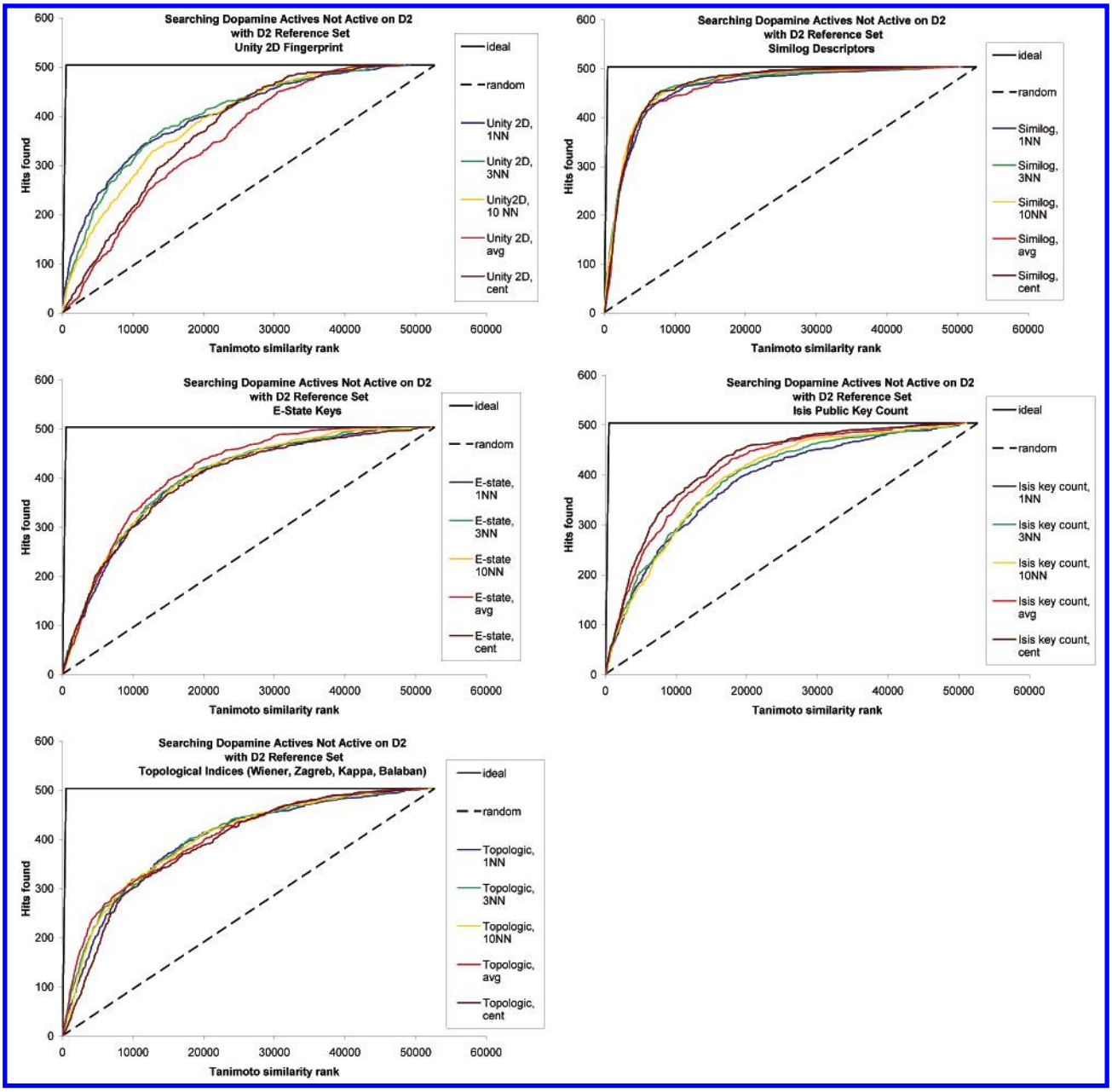

Figure 8. Searching dopamine except D2 receptor ligands with a D2 reference set. Reference compounds and evaluated compounds are close activity homologues.

for ligands of the progesterone receptor itself (Figure 13), glucocorticoid-like NRs except progesterone (Figure 14), estrogen-like NRs except glucocorticoid-like NRs (Figure 15), and finally non-estrogen-like NRs (Figure 16). The rankings obtained with the opioid δ-receptor were examined for opioid δ-receptor ligands (Figure 17) and for molecules binding to other non-δ opioid receptors (Figure 18).

It needed to be verified that the similarity searches did not generally accumulate drug-like molecules, but indeed with preference those molecules displaying the structural properties specific for ligands of the protein family represented by the reference target. In a control experiment, the D2 reference set was used to search the candidate set, and the accumulation of such molecules was evaluated which are drug-like according to the Lipinski rules[34] based on properties calculated with Cerius^{II} [26] but without reported affinity to any GPCR. In contrast to the accumulation of amine binding GPCR ligands (Figure 9) the accumulation of these drug-like molecules was, as shown in Figure 19, not higher than expected by random selection with the exception of the accumulations obtained with the topological key representation. In a second control experiment (data not shown), we searched with the factor Xa reference set and looked at the accumulation of amine binding GPCR ligands having no reported activity on any hydrolase enzymes. Again, the observed accumulation was equal to the expectation of random.

DISCUSSION

General Trends of Retrieval Performance. As expected, it was observed in all examples studied that ligands with lower activity homology are accumulated less effectively in the top similarity ranks than those of targets having a closer homology to the reference target. In some cases the decrease of accumulation with the decrease of the activity homology of the ligands is very high, as it is observed in the peptide

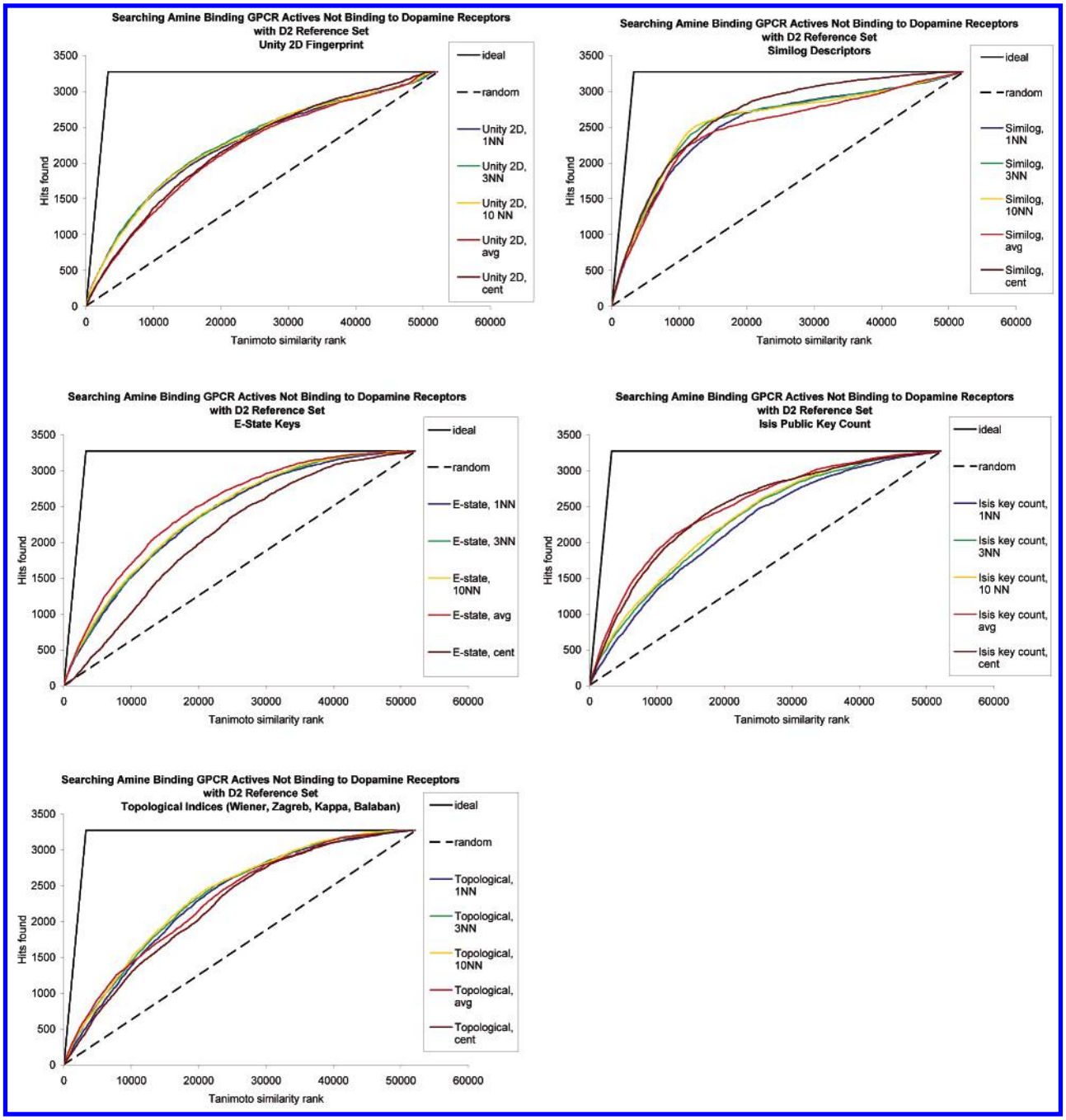

Figure 9. Searching amine binding GPCR except dopamine receptor ligands with a D2 reference. Reference compounds and evaluated compounds are distant activity homologues.

binding class A GPCRs represented by the opioid *δ*-receptor (Figures 17 and 18). In contrast to this, with the reference set of progesterone receptor ligands even the ligands of non-estrogen-like NRs three classification levels apart from the reference target are effectively accumulated (Figures 13−16). There are two reasons for this, one resulting from the molecular structure of the ligands, which will be discussed later, and the other from the nature of classification tree in which the discrete classification levels are only qualitatively related to the sequence similarity of the target proteins. The degree to which the accumulation of ligands decreases with decreasing activity homology depends also on the representations used. The Unity 2D-fingerprint similarity (Figures 7−9 and 10−18) lead in all examples studied to high accumulations of ligands with a close activity homology, whereas ISIS public key count (Figures 7−9 and 10−18) and Similog keys (Figures 7−9 and 10−18) result also in significant accumulations of ligands with a lesser activity homology to the reference set.

Comparison of the representations regarding the accumulation of the ligands of the reference targets themselves (Figures 7, 10, 13, and 17) reveals that Similog keys, Unity 2D-fingerprints, and the ISIS public keys are generally more effective than the E-state keys and the topological indices.

Examination of the similarity comparison methods shows that *1NN* is the most effective method if the accumulation of the ligands of the reference target itself should be maximized. This changes in the case of some representations when the accumulation of activity-homologous ligands not binding to the reference target is examined. Here the *cent*

Table 3: Fraction of Hits Obtained by Similarity Searches as a Function of Reference Target and Target Aimed for Different Representation and Similarity Comparison Methods[b]

reference target (size of reference set)	target aimed at (total number of possible hits)	classification distance[a]	representation, similarity comparison method	% of possible hits found in *x*% most similar structures		
				$x = 1$	$x = 5$	$x = 10$
D2 receptor (270)	D2 receptor (248)	0	Unity 2D, 1NN	**69**	**87**	90
			Similog, 1NN	68	**87**	**92**
			Similog, cent.	20	62	77
			E-state, 1NN	2	9	21
			ISIS key count, 1NN	56	72	79
			ISIS key count, cent.	5	35	56
			topological, 1NN	23	42	61
	dopamine receptors except D2 receptor (504)	1	Unity 2D, 1NN	12	33	48
			Similog, 1NN	**18**	55	78
			Similog, cent.	12	**56**	**80**
			E-state, 1NN	7	2	38
			ISIS key count, 1NN	8	23	39
			ISIS key count, cent.	7	30	51
			topological, 1NN	8	24	44
	amine binding class A GPCRs except dopamine receptor (3274)	2	Unity 2D, 1NN	6	18	31
			Similog, 1NN	6	24	40
			Similog, cent.	**7**	**26**	**44**
			E-state, 1NN	5	16	28
			ISIS key count, 1NN	3	13	23
			ISIS key count, cent.	5	20	35
			topological, 1NN	3	13	24
factor Xa (212)	factor Xa (194)	0	Unity 2D, 1NN	77	**94**	**96**
			Similog, 1NN	**86**	**94**	**96**
			Similog, cent.	43	78	87
			E-state, 1NN	56	72	77
			ISIS key count, 1NN	69	85	93
			ISIS key count, cent.	30	61	75
			topological, 1NN	28	40	49
	serine endopeptidases except factor Xa (687)	1	Unity 2D, 1NN	**14**	**35**	**48**
			Similog, 1NN	10	31	**48**
			Similog, cent.	9	27	42
			E-state, 1NN	3	14	22
			ISIS key count, 1NN	11	24	37
			ISIS key count, cent.	10	29	43
			topological, 1NN	*1*	6	13
	peptidases except serine endopeptidases (3174)	2	Unity 2D, 1NN	*2*	9	17
			Similog, 1NN	*1*	**10**	**20**
			Similog, cent.	*0*	*2*	*5*
			E-state, 1NN	*1*	7	14
			ISIS key count, 1NN	*1*	*5*	*8*
			ISIS key count, cent.	*1*	*4*	*8*
			topological, 1NN	*1*	*5*	12
progesterone receptor (58)	progesterone receptor (67)	0	Unity 2D, 1NN	96	96	96
			Similog, 1NN	**99**	**100**	**100**
			Similog, cent.	73	82	94
			E-state, 1NN	88	94	97
			ISIS key count, 1NN	88	96	96
			ISIS key count, cent.	72	91	94
			topological, 1NN	54	72	78
	glucocorticoid-like NRs except progesterone receptor (290)	1	Unity 2D, 1NN	5	19	34
			Similog, 1NN	3	30	51
			Similog, cent.	**16**	48	63
			E-state, 1NN	8	38	54
			ISIS key count, 1NN	6	18	36
			ISIS key count, cent.	12	**64**	**70**
			topological, 1NN	2	*4*	*16*
	estrogen-like NRs except glucocorticoid-like NRs (228)	2	Unity 2D, 1NN	6	12	24
			Similog, 1NN	13	30	48
			Similog, cent.	**15**	**45**	**75**
			E-state, 1NN	3	18	33
			ISIS key count, 1NN	5	12	18
			ISIS key count, cent.	6	27	41
			topological, 1NN	2	6	*10*

Table 3: (Continued)

reference target (size of reference set)	target aimed at (total number of possible hits)	classification distance[a]	representation, similarity comparison method	% of possible hits found in *x*% most similar structures		
				$x = 1$	$x = 5$	$x = 10$
	NRs except estrogen-like NRs (315)	3	Unity 2D, 1NN	36	44	50
			Similog, 1NN	**43**	65	78
			Similog, cent.	15	**69**	**87**
			E-state, 1NN	23	41	50
			ISIS key count, 1NN	5	34	53
			ISIS key count, cent.	*0*	63	80
			topological, 1NN	*0*	*1*	*4*
opioid delta receptor (81)	opioid δ-receptor (89)	0	Unity 2D, 1NN	90	**97**	**98**
			Similog, 1NN	**97**	**97**	97
			Similog, cent.	80	89	93
			E-state, 1NN	64	78	85
			ISIS key count, 1NN	81	94	94
			ISIS key count, cent.	30	71	87
			topological, 1NN	33	53	63
	opioid receptors except opioid δ-receptor (159)	1	Unity 2D, 1NN	13	25	39
			Similog, 1NN	16	26	34
			Similog, cent.	**21**	**43**	**48**
			E-state, 1NN	8	14	18
			ISIS key count, 1NN	11	19	25
			ISIS key count, cent.	9	19	26
			topological, 1NN	*1*	*3*	*8*

[a] In correspondence to the classification trees the classification distance is the number of levels which separate the parent node common to the nodes of the reference target and the targets aimed at from the node of the reference target. [b] Accumulations less or equal to random expectation are printed *italic*. The best accumulations obtained for each combination of reference and target aimed at are printed **bold**.

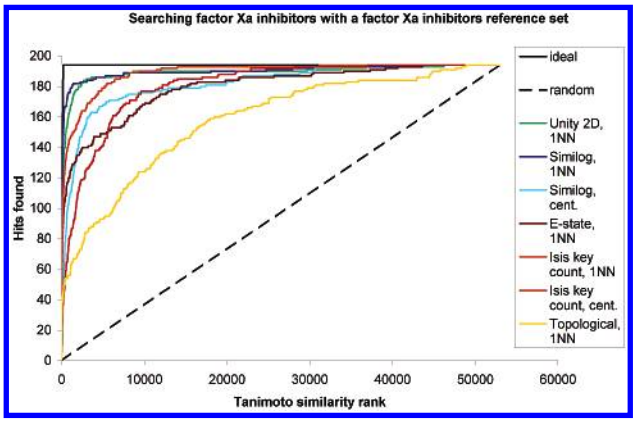

Figure 10. Searching factor Xa ligands with a factor Xa inhibitor set.

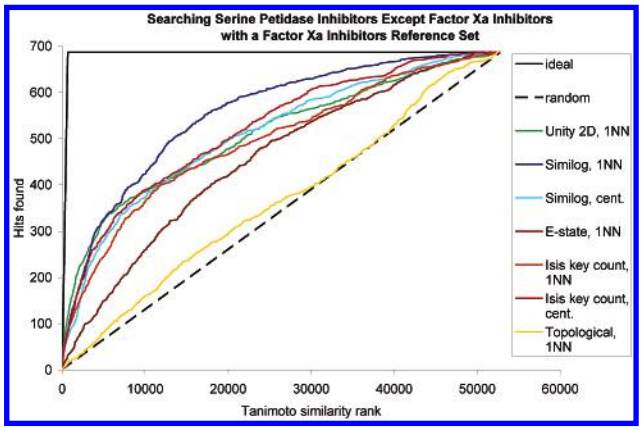

Figure 11. Searching serine peptidase except factor Xa ligands with a factor Xa inhibitor set.

method is superior to the *1NN* method in combination with the Similog keys or the ISIS Public keys, whereas in combination with the Unity 2D-fingerprints the *1NN* method is always superior to the *cent* method, at least in the case of the D2 reference target.

Retrieval Performance of the Different Molecular Representations and Similarity Measures. The Unity 2D-fingerprints belong to a frequently used class of molecular representations based on combinatorially enumerated connection paths. Although introduced as so-called screens to accelerate substructure searching, they are now also widely used for chemical similarity detection. Similarity searching based on comparison of these fingerprints is fast and known to identify close structural analogues quite effectively. This is observed here as well. If accumulations of ligands of the reference target(s) themselves are examined, the combination of the Unity 2D-fingerprint and the *1NN* similarity averaging method is among the most effective virtual screening procedures (Figures 7, 10, 13, and 17). Since for most structure classes more than one molecule is present in the MDDR database, it is very likely that after a split into a reference and a candidate set of equal size, many molecules in the reference sets have close analogues in the candidate set. Due to this distribution of close structural analogues, the usage of Unity 2D-fingerprints gives good results. It was found that the *cent* and *avg* similarity comparison methods give only poor results compared to the *1NN* and *3NN* method when combined with the Unity 2D-fingerprint representation. This shows that in the Unity 2D-fingerprint space substructural analogues are located closely together. However, ligands binding to the same target are not aggregated in one region of the representation space and, as qualitatively depicted in Figure 20a, are likely to form multiple clusters instead. When ligands of lesser activity homology not active on a reference

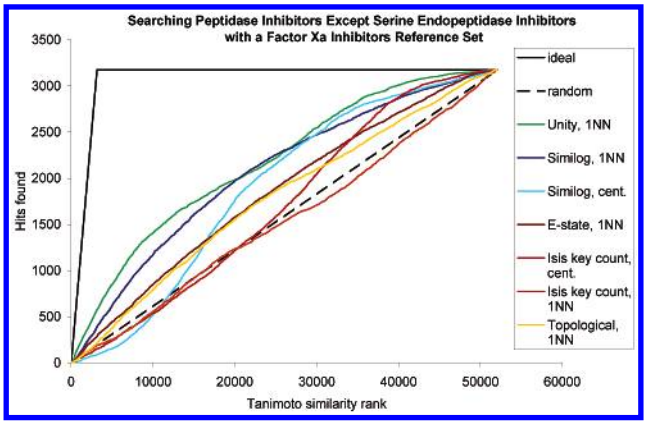

Figure 12. Searching peptidase except serine peptidase ligands with a factor Xa inhibitor set.

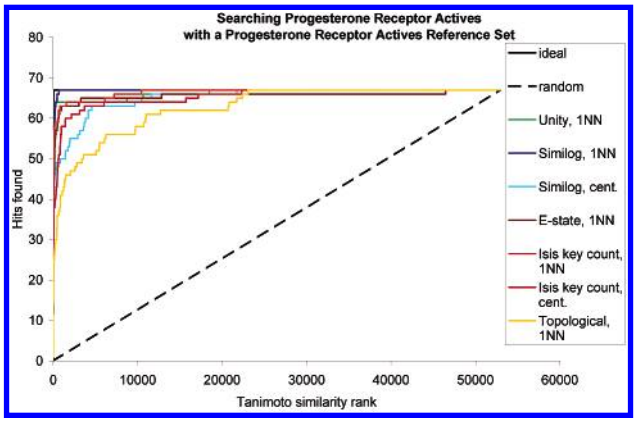

Figure 13. Searching progesterone receptor ligands with a progesterone receptor ligand reference set.

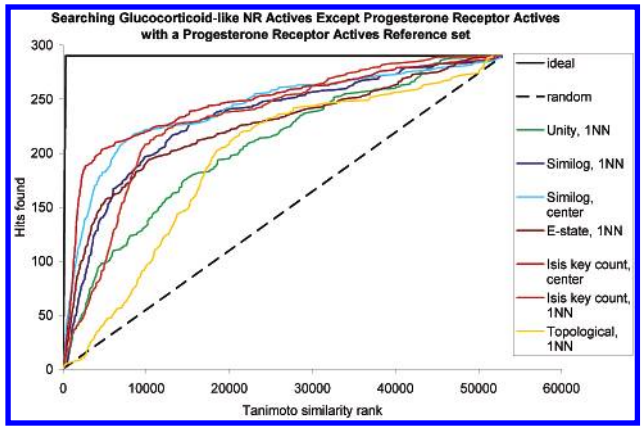

Figure 14. Searching glucocorticoid-like except progesterone nuclear receptor ligands with a progesterone receptor ligand reference set.

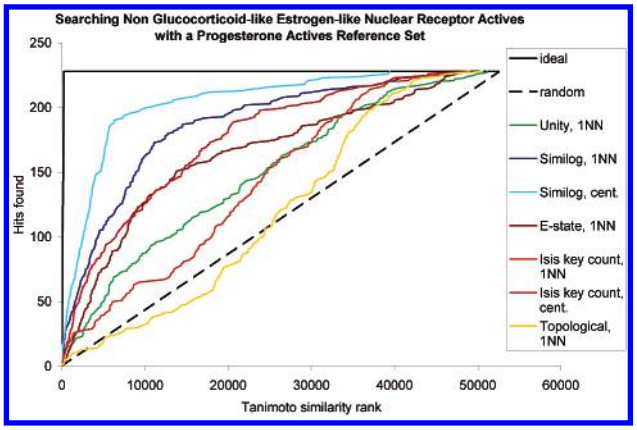

Figure 15. Searching estrogen-like except glucocorticoid-like nuclear receptor ligands with a progesterone receptor ligand reference set.

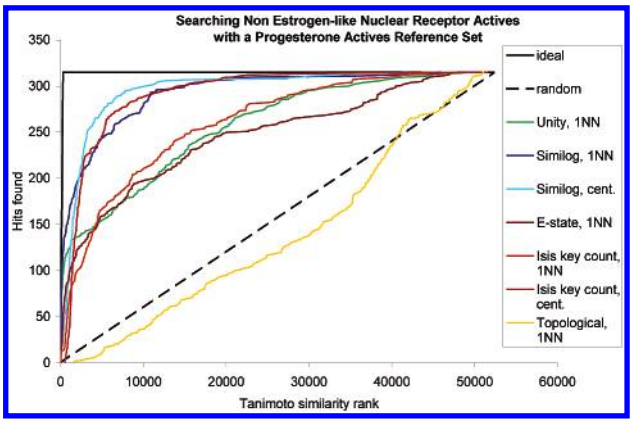

Figure 16. Searching nuclear except estrogen-like receptor ligands with a progesterone receptor ligand reference set.

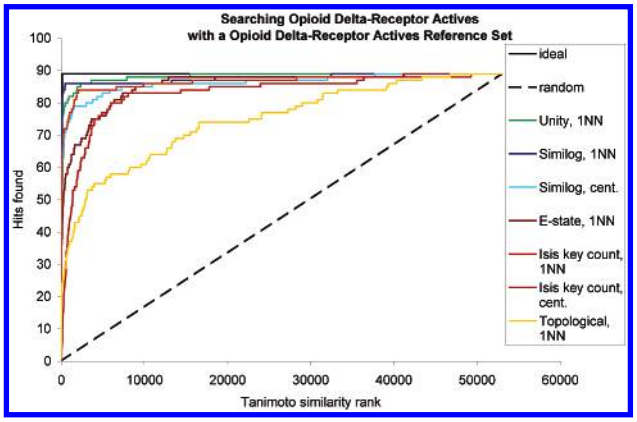

Figure 17. Searching opioid δ-receptor ligands with an opioid δ-receptor ligand reference set.

target are to be identified, the ligands are not any longer close structural analogues of the reference ligands and often share only some more general structural features. As a consequence, the accumulations obtained with Unity 2D-fingerprints are rapidly decreasing in these cases.

The Similog keys give the highest accumulations of all molecular representations evaluated. They are effective in combination with the *1NN* comparison method as well as with the *cent* method. This indicates that the averages of the Similog key occurrence counts in the reference molecules represent implicitly a kind of pharmacophore model which is common to all ligands of the reference target. Since the Similog atom keys are covering interaction specific properties of an atom, which are abstracted from the chemical elements, they can be applied to detect bioanalogous[35] or bioisosteric equivalence between functional groups. Due to the atom properties included, the Similog keys represent structural aspects responsible for intermolecular interactions. If the binding sites of two related targets are similar to each other, it can be consequently expected that the difference in Similog keys between two corresponding ligand sets is small.

In the Similog space the reference ligands are likely to form a single cluster. In this case the centroid appears as a representation of the entire reference set and the *cent* method

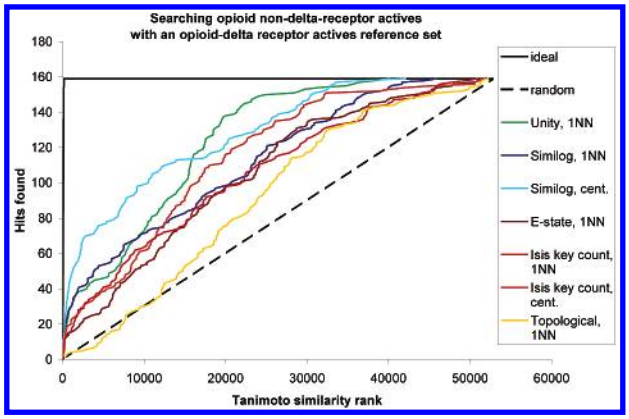

Figure 18. Searching opioid except δ-receptor ligands with an opioid δ-receptor ligand reference set.

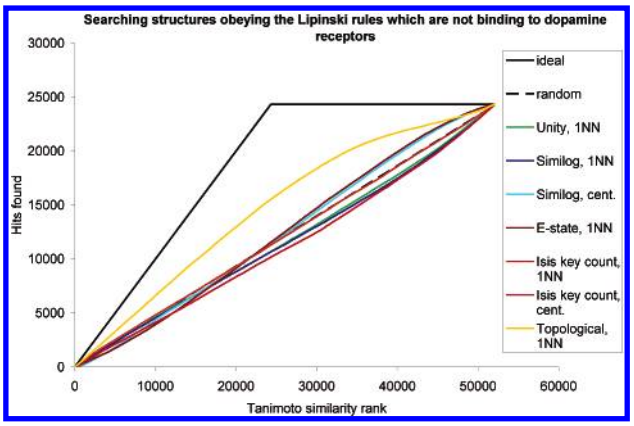

Figure 19. Searching molecules obeying the Lipinski rules and not binding to any GPCR with the D2 reference set.

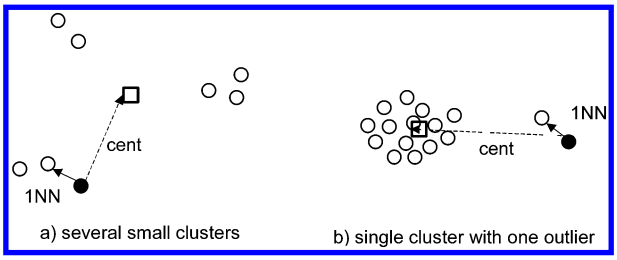

Figure 20. Multicluster reference set (a) and a single cluster reference set with an outlier (b). The reference ligands are represented by the white dots and the candidate ligand is drawn as a black dot. The centroid of the reference set is represented by the white square.

leads to good accumulations. In addition it is likely that the accumulations obtained with the cent method are less affected by single outliers in the reference set than those obtained with the 1NN method in which an outlier in the reference can cause the incorrect identification of many inactive molecules in the candidate set similar to it (Figure 20b).

Good accumulations were also obtained with the ISIS public key counts. In contrast to the Unity 2D-fingerprint they are based on selected chemical fragments rather than on combinatorially generated connection paths. The ISIS keys are partly derived directly from chemical elements or functional groups, which makes them sensitive to bioisosteric exchange of atoms in the same way as connection paths of the Unity 2D-fingerprints. Two-thirds of the keys however are defined by structural fragments consisting at least in part of augmented atoms such as "any non-hydrogen atom" or "any heteroatom". As a result, the corresponding keys are

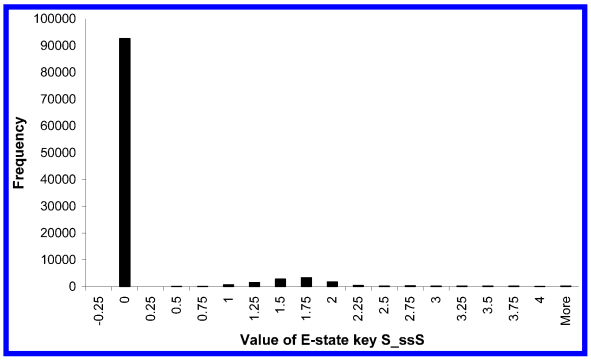

Figure 21. Histogram of the values of E-state key S_ssS calculated for the MDDR structures. The key S_ssS represents a sulfur atom with exactly two non-hydrogen atoms connected to it by single bonds.

less sensitive to the bioisosteric exchange of atoms or groups. This is in line with results of a study showing that α_1-adrenoceptor ligands can be recognized more efficiently with the ISIS keys or minifingerprints constructed partially from fragments defined by generic atoms than with the Daylight fingerprints,[36] which are like Unity 2D-fingerprints generated from combinatorially enumerated connection paths.[37,38] The accumulations obtained with the ISIS public key count representations are the second highest after those obtained by Similog when the identification of activity-homologues is considered, as it can be seen in Figures 9, 11, and 16. The Similog and the ISIS public key count share the commonality that both are key counting representations in contrast to the Unity 2D-fingerprints which only indicate the presence or absence of a structural key. The key occurrence count provides implicit information on the molecular size which has an important influence on protein binding. Some compound classes are characterized by the multiple occurrences of structural features, such as oligopeptides containing multiple amide bonds or sugars containing multiple OH groups. Key counting can for example discriminate between structures containing a single amide group and an oligopeptide.

The E-state keys[20] are a representation combining electrostatic and topological information of molecular structures. The E-state keys are based on estimations for atom charges which are derived from the electronegativity of the atoms and their connectivity. The atoms are divided in atom types which are defined by connectivity information. For each of the atom types the estimated charges are summed up, and these sums form the representation vector. If an atom type is not present in the molecule, the value of the associated key element is zero. Frequency analysis of the components of the E-state keys of the MDDR structures show often a sharp peak at zero for structures not containing the atom-type associated with the key and a bell-shaped distribution at a completely different value for the other structures. This is shown in Figure 21 for the S-ssS atom type, representing a sulfur atom with two single bonds to two non-hydrogen atoms. Similar to a fingerprint, the E-state key is approximately a binary indicator for the presence or absence of atom types. Like the Unity 2D-fingerprint, the E-state keys can be used to identify close structural analogues, as in the query for ligands of the reference targets itself. But if the queried target is only similar to the reference target, the accumulations of its ligands drop to an extent equal in

magnitude to those observed with the Unity 2D-fingerprint, as it can be seen for example in the series of Figures 13−16.

Rather ineffective and often leading to no accumulation above random at all (e.g., Figures 11, 15, 16, and 18) was the combination of topological indices. In addition, they were the only representation which leads in the control experiment shown in Figure 19 to the accumulation of generally drug-like structures. Thus, it cannot be excluded that the observed accumulation of activity homologous ligands obtained with this representation is partly caused by a target family unspecific drug-likeness. It may be argued that the essential functional groups to achieve receptor binding can be attached to different core scaffolds as long as the groups can occupy the required position in space. This is an assumption widely used in combinatorial chemistry which often tries to cover with a single scaffold more than one type of biological activity. Variations of the scaffold, which are not affecting the orientation of their substituents and the protein binding ability, may be overestimated by the topological keys. Even more important is that the set of seven topological keys used here is almost certainly insufficient information to represent a ligand with all its properties relevant for protein binding. It is by far that molecular representation with the lowest dimensionality, compared to the E-state keys containing sums for 30 atom types, the ISIS public key count with 135 integers, and the 5989 used Similog keys. The decrease of dimensions in the representations from the Similog keys to the topological keys reflects the decrease in ligand accumulation obtained. The Unity 2D-fingerprints with the second highest dimensionality of 988 are not directly comparable to the other representations due to their binary nature. In addition, it had been observed that more complex fingerprints do not necessarily lead to improved recognition of biological activity, as small minifingerprints were shown to be superior to the longer Daylight fingerprints.[36]

Differences between the Target Families. There are differences between the target families regarding the maximum number of classification levels a target may be apart from the reference target to get an above random accumulation of its ligands. In the case of the D2 dopamine receptor as reference it is possible to obtain accumulations of ligands of those amine binding class A GPCRs which are not dopamine receptors. As the name of this receptor class expresses, all its members are known to bind amines as endogenous ligands which are structurally closely related to each other. Therefore, it is not surprising that also most synthetic analogues have several structural features in common. Since the known endogenous ligands are synthetically accessible and have molecular weights low enough for suitable pharmacokinetics, it is not surprising that the most drugs interacting with amine binding class A GPCR are developed on the basis of synthetic ligands directly derived from the endogenous ones.

The opioid δ-receptor as the other GPCR investigated is different from the D2 receptor with respect to its endogenous ligands. It belongs together with the other opioid receptors (μ and κ) and other receptor families to the class of the peptide binding GPCRs which have oligopeptides as their endogenous ligands. In the case of the opioid receptors these are the enkephalins. Enkephalin derivatives or close analogues have been investigated as drug candidates and form

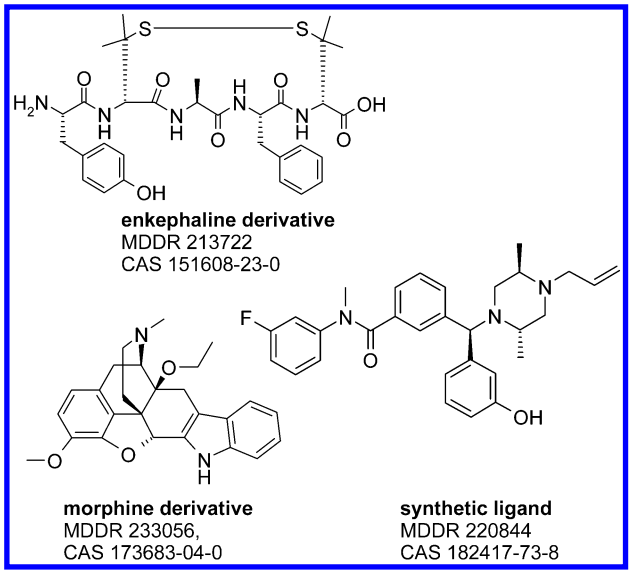

Figure 22. Examples of opioid δ-receptor ligands.

a large part of the structures with opioid-receptor activity in the MDDR database. The best known class of opioid receptor ligands, which gave the receptor its name, is formed by the natural product morphine (the effective component of opium) and its derivatives: complex, rigid, and polycyclic structures. As a third class there are completely synthetic ligands reported with a far less complex scaffold (Figure 22). It is a challenging problem to detect the structural features in common to these structural classes. Nevertheless, ligands for the δ-receptor itself can be effectively accumulated with the δ-receptor reference set (Figure 17) because this set contains representatives from each structural class. The *cent* method combined with the Similog representation is able to produce good accumulations in this case. Since this method can only work if all structures are aggregated in one area of the representation space, it can be assumed that the Similog method perceives some of the structural features in common to the three ligand classes. However, considerable smaller accumulations are obtained for the ligands of the non-δ opioid receptors (Figure 18).

Like the opioid receptors, the serine endopeptidases (serine proteases) have peptides as endogenous substrates. There exist two classes of serine endopeptidase inhibitors: close peptide analogues and non-peptidic organic molecules. This makes the recognition of the common structural features difficult. When searching with the factor Xa reference set, good accumulations are obtained for ligands of factor Xa itself (Figure 10) and moderate accumulations for other serine endopeptidase inhibitors (Figure 11). Only very low accumulations, with the cent similarity comparison method not better than random, are obtained for other endopeptidase ligands. This was not unexpected, as the structural key feature of the substrate changes when moving on from serine endopeptidases to other endopeptidases. Accumulations obtained in this case especially are likely to result from the fact that many peptidase ligands have a peptidic structure in general which does always bear some similarity to other peptides; even in a molecular representation like Similog there are some atom triplets related to the backbone which all peptides have in common. However, we did expect better accumulations for the ligands of the nonfactor Xa serine endopeptidases. The high similarity of the backbones of the

peptidic ligands may hinder the recognition of the binding relevant properties residing in the side chains. This indicates that the studied similarity measures are less suitable for targets which have peptides as natural ligands. In the case of targets which have peptides as natural ligands or have peptides among the reference ligands for other reasons, it might be helpful to remove these from the reference set to avoid the accumulation of structures similar to the reference only due to the peptide backbone.

The ligands for nuclear receptors were very effectively accumulated with the progesterone reference set. In contrast to the expectation that the accumulation should decrease with the increasing dissimilarity of the targets to the reference target, the accumulation of ligands increases when moving from the non-progesterone glucocorticoid-like receptor ligands (Figure 14) to the non-glucocorticoid-like estrogen-like receptor ligands (Figure 15) and again when increasing the classification distance by one to the non-estrogen-like nuclear receptor ligands (Figure 16, for classification see Figure 6). The group of the non-progesterone glucocorticoid-like NRs is almost exclusively formed by compounds with the MDDR activity key "antiandrogenic" which were interpreted as antagonist of the testosterone receptor. The activity keyword "antiandrogenic" is not strictly target related and it used as well for testosterone-5α-reductase inhibitors, which inhibit the reduction of testosterone to the dihydrotestosterone, a compound with increased activity on the testosterone-receptor. Compounds binding to the estrogen receptor were found as the almost exclusive members of the group of the non-glucocorticoid-like estrogen-like NR ligands. In this class there are many synthetic compounds without a steroid-like scaffold, whereas in one-third of the non-estrogen-like NR ligands the steroidal C and D rings are found, which are also present in the endogenous ligands of many other NRs (e.g., vitamin D3). This improves the identification of the ligands of the non-estrogen-like NRs. The good accumulations obtained for all NR ligands with one reference set implies that all NRs for which ligands are included in the MDDR database have a conserved binding site with a common shape. This is supported by analyses of the sequences of the NR ligand binding domains.[39]

Selectivity and Scope. Accumulation was observed for sets of activity homologous ligands from which the ligands of the reference target itself were explicitly excluded. Therefore, homology-based similarity searching requires not necessarily additional unselective activity on the reference target to identify ligands of targets homologous to the reference target. Nevertheless, it can be expected that ligands binding to the reference target are accumulated preferentially. This has the consequence that ligands of targets homologous to the reference target are likely to be unselective. Therefore, the method described here is less suitable for the lead optimization process where target selectivity is one of the main objectives, but it is most suitably applied early in the lead finding process when not much information about the target is available. In this situation the possibility to identify ligands of targets without known protein structure and without initially known ligands is a clear advantage of homology-based similarity searching. In a first approximation, the only knowledge required is the sequence of the target protein to identify homologous targets with known ligands; more rigorous searches should be based on more detailed analysis and comparison of the binding site residues, requiring the knowledge of a 3D structure model.[3] Homology-based similarity searching is also suitable for the acquisition of compounds and composition of screening libraries used for the exploration of a whole target family.

CONCLUSION

We have introduced the Similog keys as a new molecular representation which can be used to recognize ligands binding to biological targets in sets of chemically diverse molecules. Using a reference set of ligands of a known target, it is not only possible to screen effectively for further ligands of this target but also to identify ligands of targets which are similar but not identical to the reference target. This dramatically enhances the scope of similarity searching to the further systematic chemogenomics[40] exploration of new receptors belonging to previously well explored target families and enables machine-intelligent virtual screening approaches ("smart screening"[41]). Our comparative study illustrates that for such applications the Similog keys are superior to other molecular connectivity representations at least for the target families studied. Homology-based similarity searching with adequate similarity measures, as shown here, makes it possible to exploit previous screening data effectively in an ongoing ligand identification process once a close integration of molecular structure, screening, and protein sequence and annotation databases is achieved.

ACKNOWLEDGMENT

Drs. P. Ertl, G. Paris, B. Rohde, P. Selzer, and J. Zimmerman (Novartis) are acknowledged for insightful discussion. Parts of this work were done within the frame of the research project "Information-based Approaches in Drug Design" sponsored by Swiss KTI, which is therefore gratefully acknowledged.

Supporting Information Available: The SLN-based fingerprint definition file used for calculation of the ISIS Public key counts. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Schuffenhauer, A.; Zimmermann, J.; Stoop, R.; van der Vyver, J.-J.; Lechini, S.; Jacoby, E. An Ontology for Pharmaceutical Ligands and its Application for In Silico Screening and Library Design. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 947−955.

(2) Frye, S. V. Structure−Activity Relationship Homology (SARAH): A Conceptual Framework for Drug Discovery in the Genomic Era. *Chem. Biol.* **1999**, *6*, R3−R7.

(3) Jacoby, E. A Novel Chemogenomics Knowledge-Based Ligand Design Strategy − Application to G- Protein-Coupled Receptors. *Quant. Struct.-Act. Relat.* **2001**, *20*, 115−123.

(4) UNITY 4.2.1 Tripos Inc., 1699 South Hanley Rd., St. Louis, Missouri, 63144, U.S.A., http://www.tripos.com.

(5) Mitchell, J. B. O. The Relationship Between the Sequence Identities of Alpha Helical Proteins in the PDB and the Molecular Similarities of their Ligands. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1617−1622.

(6) Wild, D. J.; Willett, P. Similarity Searching in Files of Three-Dimensional Chemical Structures: Alignment of Molecular Electrostatic Potential Fields with a Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 159−167.

(7) 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.

(8) Good, A. C.; Kuntz, I. D. Investigating the Extension of Pairwise Distance Pharmacophore Measures to Triplet-based Descriptors. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 373−379.
(9) Pickett, D. S.; Mason, J. S.; Mcay, I. M. Diversity Profiling and Design using 3D-Pharmacophores: Pharmacophore-Derived Queries (PDQ). *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1214−1223.
(10) Matter, H. Selecting Optimally Diverse Compounds from Structure Databases: A Validation Study of Two-dimensional and Three-dimensional Molecular Descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.
(11) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-point Pharmacophore Method for Molecular Similarity and Diversity Applications: Overview of the Method and Applications, Including a Novel Approach to the Design of Combinatorial Libraries containing Privileged Substructures. *J. Med. Chem.* **1999**, *42*, 3251−3264.
(12) Martin, E. J.; Hoeffel, T. J. Oriented Substituent Pharmacophore PRopErtY Space (OSPPREYS): A Substituent-based Calculation that Describes Combinatorial Library Products Better than the Corresponding Product-based Calculation. *J. Mol. Graph. Model.* **2000**, *18*, 383−403.
(13) Makara, G. M. Measuring Molecular Similarity and Diversity: Total Pharmacophore Diversity. *J. Med. Chem.* **2001**, *44*, 3563−3571.
(14) Floersheim, P. Sandoz, 1991, unpublished results.
(15) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. Scaffold-hopping by Topological Pharmacophore Search: A Contribution to Virtual Screening**.** *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 2894−2896.
(16) Bemis, B. W.; Kuntz, I. D. A Fast and Efficient Method for 2D and 3D Molecular Shape Description. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 607−628.
(17) SYBYL 6.7.1 Tripos Inc., 1699 South Hanley Rd., St. Louis, Missouri, 63144, U.S.A., http://www.tripos.com.
(18) Pauling, L. *Grundlagen der Chemie*; Verlag Chemie: Weinheim, 1973.
(19) The ISIS public keys are already used in MACCS-II and are also known as MACCS keys. ISIS/Base and MACCS are both products of MDL Information Systems, Inc., San Leandro, CA, http://www.mdli.com.
(20) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039−1045.
(21) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.
(22) Gutman, I.; Rušic, B.; Trinajstic, N.; Wilcox, C. W. Graph Theory and Molecular Orbitals. Part 12. Acyclic Polyenes. *J. Chem. Phys.* **1975**, *62*, 3399−3405.
(23) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.
(24) Kier, L. B. An Index of Molecular Flexibility from Kappa Shape Attributes. *Quant. Struct.-Act. Relat.* **1989**, *8*, 221−224.
(25) Kier, L. B. Shape Indexes of Orders One and Three from Molecular Graphs. *Quant. Struct.-Act. Relat.* **1986**, *5*, 1−7.
(26) Cerius^{II} Package by Accelrys Inc. San Diego, CA, http://www.accelrys.com.
(27) Willett, P.; Winterman, V. A. Comparison of Some Measures for the Determination of Intermolecular Structural Similarity. *Quant. Struct.-Act. Relat.* **1986**, *5*, 18−25.
(28) Sheridan, R. P. The Centroid Approximation for Mixtures: Calculating Similarity and Deriving Structure−Activity Relationships. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1456−1469.
(29) Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **1963**, *58*, 236−244.
(30) Jarvis, R. A.; Patrick, E. A. Clustering Using a Similarity Measure Based on Shared Nearest Neighbors. *IEEE Trans. Comput.* **1973**, *C-22*, 1025−1034.
(31) MDL Drug Data Report Version 2001.1, MDL ISIS/HOST software, MDL Information Systems, Inc. San Leandro, CA, http://www.mdli.com.
(32) Bath, P. A.; Morris, C. A.; Willett, P. Effect of Standardization on Fragment-based Measures of Structural Similarity. *J. Chemometrics* **1993**, *7*, 543−550.
(33) Edgar, S. J.; Holliday, J. D.; Willet, P. Effectiveness of Retrieval in Similarity Searches of Chemical Databases: A Review of Performance Measures. *J. Mol. Graph. Model.* **2000**, *18*, 343−357.
(34) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **1997**, *23*, 3−25.
(35) Floersheim, P.; Pombo-Villar, E.; Shapiro G. Isosterism and Bioisosterism Case Studies with Muscarinic Agonists. *Chimia* **1992**, *46*, 323−334.
(36) James, C. A.; Weininger, D. Daylight theory manual. Daylight Chemical Information Systems, Inc., Irvine, CA, 1995, http://www.daylight.com.
(37) Xue, L.; Stahura, F. L.; Godden, J. W.; Bajorath, J. Mini-Fingerprints Detect Similar Activity of Receptor Ligands Previously Recognized Only by Three-Dimensional Pharmacophore-Based Methods. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 394−401.
(38) Xue, L.; Godden, J. W.; Bajorath, J. Database Searching for Compounds with Similar Biological Activity Using Short Binary Bit String Representations of Molecules, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 881−886.
(39) Francoijs, C. J.; Klomp, J. P.; Knegtel, R. M. Sequence Annotation of Nuclear Receptor Ligand-Binding Domains by Automated Homology Modeling. *Protein Eng.* **2000**, *13*, 391−394.
(40) Caron, P. R.; Mullican, M. D.; Mashal, R. D.; Wilson, K. P.; Su, M. S.; Murcko, M. A. Chemogenomic Approaches to Drug Discovery. *Curr. Opin. Chem. Biol.* **2001**, *5*, 464−470.
(41) Engels, M. F.; Venkatarangan, P. Smart Screening: Approaches to Efficient HTS. *Curr. Opin. Drug Discov. Devel.* **2001**, *4*, 275−283.

CI025569T