Core Trees and Consensus Fragment Sequences for Molecular Representation and Similarity Analysis

Eugen Lounkine and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology & Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstrasse 2, D-53113 Bonn, Germany

Received January 17, 2008

A new type of molecular representation is introduced that is based on activity class characteristic substructures extracted from random fragment populations. Mapping of characteristic substructures is used to determine atom match rates in active molecules. Comparison of match rates of bonded atoms defines a hierarchical molecular fragmentation scheme. Active compounds are encoded as fragmentation pathways isolated from core trees. These paths are amenable to biological sequence alignment methods in combination with substructure-based scoring functions. From multiple core path alignments, consensus fragment sequences are derived that represent compound activity classes. Consensus fragment sequences weighted by increasing structural specificity can also be used to map molecules and search databases for active compounds.

INTRODUCTION

The use of molecular substructures and fragment-type descriptors has a long tradition in chemoinformatics and pharmaceutical research, and fragment descriptors continue to be of significant interest for many applications.[1–3] Substructures are thought to be powerful molecular descriptors because they implicitly capture many chemical characteristics. Accordingly, substructure or fragment design is an active area of research, as it has been for many years. Different methodologies have been introduced to derive collections of molecular fragments that can be used as descriptors,[2–5] and these approaches often employ systematic or knowledge-based fragmentation protocols. Apart from these approaches, different concepts have also been applied to guide fragment generation. For example, synthetic reaction-oriented fragmentation schemes[6] are widely used to aid in compound or combinatorial library design. In addition, text-based molecular representations such as the pioneering SMILES language[7] lend themselves to fragment design and representation.[8,9] Moreover, departing from knowledge-based fragmentation, mining of randomly generated molecular fragment populations has also been introduced as an approach to identify substructure combinations that are characteristic of compound activity classes and delineate core regions in active molecules.[10–12]

Considering the spectrum of currently available substructure methods, questions that have thus far been little addressed include, for example, whether hierarchical and random fragmentation methods could be combined in a meaningful way or how generally applicable and compound class-directed fragmentation approaches might relate to each other. As a first step in this direction, we have attempted to utilize information encoded in random fragment populations as a basis for devising a novel hierarchical fragmentation method. This scheme combines fragments from core regions of compound activity classes with increasingly generic fragments representing peripheral regions. The methodology represents active molecules as sets of fragmentation pathways of increasing structural specificity. Individual fragmentation paths are combined using well-known algorithms for the alignment of biological sequences in conjunction with scoring functions designed to assess substructure similarity. From multiple path alignments, consensus fragment sequences are extracted to represent compound activity classes. In addition, these consensus sequences can also be transformed into queries to search databases for novel active compounds. Here we report the development and analysis of our "hybrid" fragmentation approach.

METHODS

Generation and Mining of Random Fragment Populations. Random molecular fragment populations of compounds represented as hydrogen-suppressed 2D molecular graphs were produced using MolBlaster calculations over 3000 iterations with randomized numbers of bond deletions per step, as described.[13] Activity class characteristic substructures (ACCS) are defined as random molecular fragments that occur in at least two molecules of an activity class but no background database compounds.[12] In our analysis, ACCS were generated for all activity classes (see below) against a background of 2000 randomly selected ZINC compounds.[14]

Core-Oriented Molecular Fragmentation. ACCS were mapped onto the active compounds from which they originated using a previously described protocol that assigns fragment match rates in the interval [0, 1] to each atom. Every time an atom is matched by a fragment, the atom-specific counter is incremented by one, and match rates are obtained by dividing the atom counter by the total number of mapped fragments. On the basis of these match rates, molecular cores can be defined at different levels, for example, by combining all atoms with 90%+ match rate, 80%+, 70%+, and so on.[12]

* Corresponding author phone: +49-228-2699-306; fax: +49-228-2699-341; e-mail: bajorath@bit.uni-bonn.de.

10.1021/ci800020s CCC: $40.75 © 2008 American Chemical Society
Published on Web 05/21/2008

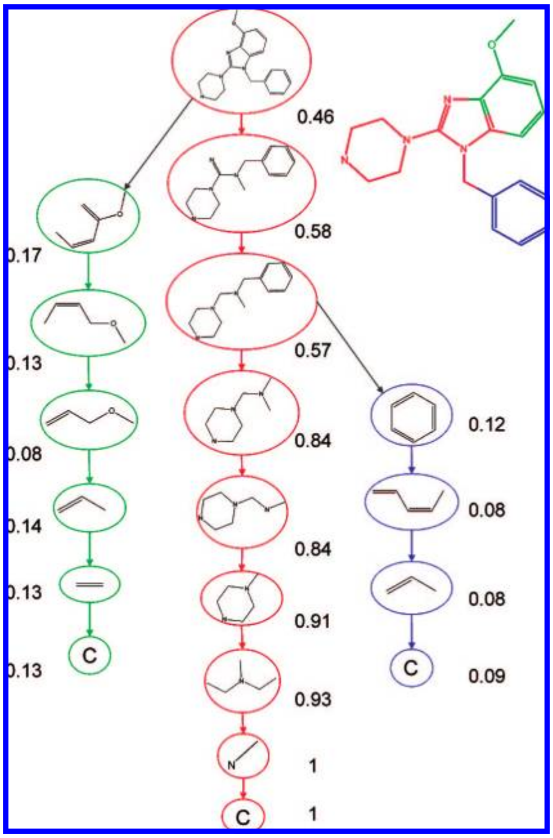

Figure 1. Exemplary core tree. The core tree for a serotonin receptor antagonist is shown. For clarity, the core tree is simplified by omitting nodes that do not constitute a path. Three paths can be distinguished, colored red (core path), green, and blue, respectively. The numbers report average atom match rates of ACCS. The molecule is color-coded according to pathway origin.

Here we introduce a molecular fragmentation scheme that depends on differences in match rates. After an ACCS set has been used to determine match rates, the difference in match rates between bonded atoms is assigned to each bond in a molecule. Then bonds are sorted in decreasing order of match rate differences. This ranking defines the order in which bonds are iteratively deleted in a subsequent fragmentation operation. Bonds that have the same match rate difference are deleted during the same iteration. When a fragment is separated into subfragments, these fragments are considered children of the parent fragment. A tree structure is constructed that represents all parent-child relationships, organizes the resulting fragments accordingly, and assigns to each fragment the average match rate of its atoms. The tree structure is termed the core tree. It is a rooted tree with the original molecule from which the fragments are derived at the root (see Figure 1).

Identification of Fragmentation Pathways. A $\mathrm{score}_{\mathrm{edge}}$ is assigned to each edge in the core tree that is calculated from the fragment size (N, number of atoms) and average match rate (MR) of the corresponding nodes. If the parent's average match rate is 0, then the edge is also assigned a score of 0. Otherwise, it is calculated as

$$\mathrm{score}_{\mathrm{edge}} = \sqrt{\frac{\mathrm{MR}_{\mathrm{child}}}{\mathrm{MR}_{\mathrm{parent}}}\frac{N_{\mathrm{child}}}{N_{\mathrm{parent}}}}$$

For each fragment (node) in the core tree, a fragmentation pathway is identified by iteratively following maximal edge scores among the node's children (that then become nodes). The fragmentation pathway of the root following the maximal edge score is called core path. Starting with the core path, peripheral paths are identified as follows: F_{Path} is defined as the set of fragments forming a path and C_{Path} as the set of nodes that are not elements of F_{Path} but children of an F_{Path} fragment. Then, for each C_{Path} node that is larger than five atoms the fragmentation path is identified, and its fragments are added to F_{Path}. The procedure is repeated for each newly delineated pathway until no more children with more than five atoms can be found. Five atoms were chosen to ensure that fragmentation pathways could originate from five-membered rings. Figure 1 shows distinct pathways in a core tree.

Path Alignment. For path alignment, fragments are encoded as unique SMILES[7] strings. Two fragmentation pathways are aligned using the Needleman-Wunsch algorithm[15] utilizing a SMILES-based scoring function instead of a standard amino acid substitution matrix. Relative insertions and deletions are accommodated using affine gaps with a gap opening penalty of −5 and a gap extension penalty of −2. The newly designed SMILES-based scoring function combines a size similarity and a string similarity term. The size similarity term is calculated as

$$\mathrm{sizeterm} = 10 * \frac{\min(N_a, N_b)}{\max(N_a, N_b)}$$

N_a and N_b are the numbers of atoms in two fragments a and b, respectively. Fragments that have equal size are assigned a score of 20 instead.

The string similarity is evaluated and scored in three steps: (i) if SMILES strings of two fragments are identical, a string similarity score of 15 is returned; (ii) if they are not identical, branches are eliminated and if the resulting "cropped" strings are identical, a score of 5 is returned; (iii) if the cropped stings are not identical, but a largest common substring exists, a score of 2 is returned; otherwise, 0 is returned. The latter case applies to small fragments in pathways terminating at heteroatoms, where a largest common substring is not always found. Score levels with an approximately three-time change in magnitude (i.e., 2, 5, and 15) have been empirically determined to produce reasonable fragmentation pathway alignments.

Values for the size and the string similarity terms are added to produce the fragment alignment score. This scoring scheme is found to produce meaningful and robust pairwise path alignments, as shown in the Results section. In order to make pathway alignments comparable, the global alignment score is normalized to the interval [0, 1] by dividing the alignment score by the average of the alignment scores of each path aligned with itself.

Multiple Core Path Alignment. From the normalized alignment scores, a guidance tree is calculated using the Unweighted Pair Group Method with Arithmetic Mean (UPGMA).[16] The scores for each position in a multiple alignment are derived as unweighted averages of pairwise comparisons.[17] The scoring function is adjusted in order to avoid gaps at the termini: the first and the last fragment are each substituted with a placeholder ("*") that yields a very high score of 100 when matched with another "*". After completion of the alignment, the placeholders are replaced by the original fragments. For a multiple core path alignment,

Table 1. Activity Classes for Similarity Searching

class designation	biological activity	no. of compds	source
AA2	adrenergic (alpha 2) agonists	34	MDDR[21]
ACE	ACE inhibitors	15	CMC-3D[22]
ARO	aromatase inhibitors	10	CMC-3D
HIV	HIV protease inhibitors	48	literature[23]
XAN	xanthine oxidase inhibitors	35	MDDR

a "Consensus Fragment Sequence" (CFS) can be derived by combining nonredundant fragments at each alignment position.

Position-Dependent Fragment Weights. When a multiple core path alignment is generated for molecules of an activity class, the consensus fragment sequence can be used as a signature of an activity class and also as a search query. Each CFS fragment is weighted according to its relative position in the alignment. The first position corresponding to the largest fragments is assigned a weight of 1 and the last position a weight of (1/length of CFS). The other positions are assigned position-dependent multiples of this weight unit. For example, for an alignment with five positions, the corresponding positional weights are 1, 0.8, 0.6, 0.4, and 0.2. Thus, weights become lower in the order of decreasing fragment size.

Structure−Activity Relationships. CFS provides a molecular representation that is amenable to the analysis of structure−activity relationships. Database molecules can be mapped to the weighted CFS of an activity class by evaluation of substructure matches. From all matching substructures, the maximally weighted fragment is selected and used to calculate a compound score based on the weight and relative size of this fragment:

$$\text{score}_{\text{compound}} = \text{weight}_{\text{fragment}} * \frac{\text{size}_{\text{fragment}}}{\text{size}_{\text{compound}}}$$

The size terms report the number of heavy atoms in the fragment and test compound, respectively. Thus, a compound obtains a high score if it matches a large CSF fragment. If a molecule is identical to an active reference compound, a score of 1 is obtained. Thus, for database searching, this scoring scheme corresponds to a flexible and size-oriented substructure search over a set of active reference compounds.

Data Sets and Search Calculations. For the generation and characterization of core trees, a previously generated data set was used consisting of fragment populations of 1025 molecules belonging to 45 different activity classes.[12] The ability of the CFS scoring scheme to capture structure−activity relationships was evaluated on five compound classes (Table 1) in comparison with MACCS structural keys, the most widely used molecular fragment descriptors. For each activity class, ten reference sets were randomly chosen consisting of half-or nearly half (e.g., seven of 15) of the compounds. For each reference set, the CFS was derived starting with the generation of ACCS. In each case, the remaining active molecules were added to 150,000 randomly chosen ZINC compounds,[14] and the resulting database was mapped to the weighted CFS of the corresponding reference set. MACCS similarity search calculations were also carried out with each reference set using the centroid approach[18] (i.e., fingerprint averaging) as a search strategy for multiple reference compounds and the general form of the Tanimoto coefficient (Tc)[19] for comparing "active" centroid fingerprints to those of individual database compounds. For comparison, recovery rates were calculated for the top scoring 100 database compounds, a selection set size that is often used to evaluate virtual screening calculations.

RESULTS AND DISCUSSION

Core Tree Design. Activity class characteristic substructures have been selected from random fragment populations and mapped onto the molecules from which they originated.[12] Mapping of a series of ACCS produces match rates of atoms in a source molecule that are used to define molecular cores as unions of atoms that are most frequently matched by characteristic substructures. Moreover, match rates can also be used to classify bonds in molecules according to match rate differences, which we have done here. Large differences between match rates of bonded atoms distinguish molecular regions that are most characteristic of a given activity class from more generic regions. Therefore, this match rate-based classification scheme can also be used to guide molecular fragmentation procedures. We introduce a tree representation for individual active molecules termed core tree that organizes substructures according to an iterative fragmentation scheme where bonds connecting atoms with high differences in match rates are deleted first (Figure 1). The core tree describes the whole molecule and encodes distinct molecular regions as fragment pathways of decreasing specificity. This means that most characteristic fragments are found at the top and most generic ones at the bottom. Importantly, although the fragmentation scheme is based on match rates derived from mapping of ACCS, it becomes independent of ACCS sets, because fragment paths do not necessarily contain ACCS.

Fragmentation Pathways. Different pathways can be identified in core trees on the basis of atom match rates and fragment sizes. The fragmentation path that starts at the original molecule and represents the molecular core most frequently matched by ACCS is termed the core path; other pathways describe peripheral molecular regions. The exemplary core tree in Figure 1 illustrates that average atom match rates are much larger for fragments of the core path than peripheral paths. We statistically analyzed the core tree representations of our 1025 test molecules belonging to 45 different activity classes and found that, on average, a molecule was described by one core path and 1.7 peripheral paths. Figure 2 reports the corresponding path length distribution. It shows that core paths most frequently consist of seven to 10 fragments, whereas most peripheral paths consist of only three to seven fragments.

Multiple Core Path Alignment. When represented as SMILES strings, core paths can be regarded as fragment sequences and subjected to alignment methods for biological sequences, provided an appropriate scoring scheme is applied. We have designed a scoring function based on unique SMILES strings that combines size and string similarity terms, as described in the Methods section. This function can be generally applied to compare SMILES strings and score substructure similarity because graph resemblance is implicitly encoded in a fragmentation pathway. Small fragments toward the end of the core pathway alignments focus on substructures that are characteristic of an activity class and are well-represented by SMILES strings. We have

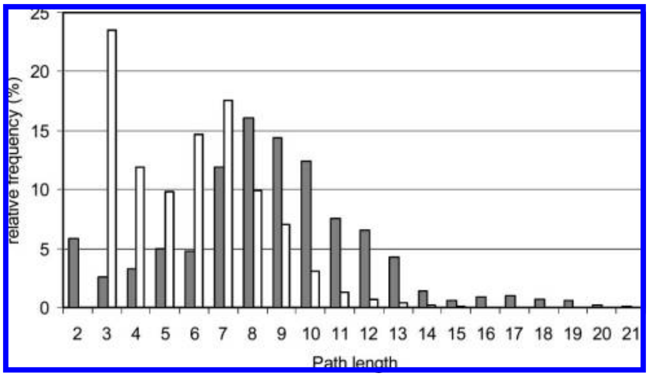

Figure 2. Path length distribution. For 1025 active molecules, the path length distributions are reported for core (dark) and peripheral (white) paths.

used the scoring function in combination with the Needleman-Wunsch algorithm to globally align core paths of compound activity classes. Using normalized scores for pairwise path alignments, a phylogenetic tree was generated using UPGMA to guide multiple core path alignment. Figure 3 shows a multiple alignment for an activity class. Core paths in the alignment begin with active molecules and end with terminal atoms. Given our match rate-based fragmentation scheme, fragments proximal to the molecule are more class-specific than distant fragments. Thus, from the left to the right in Figure 3, fragment specificity is decreasing. Core paths of individual molecules within an activity class typically have overlapping yet distinct composition. Therefore, substructure similarity-based sequence alignment approaches are particularly suitable for comparison of core paths.

Consensus Fragment Sequence. From a multiple path alignment, a consensus fragment sequence is derived for an activity class by combining unique fragments at each alignment position. The CFS is also illustrated in Figure 3. As a molecular representation tool, it organizes nonredundant substructures found in an activity class as groups of similar fragments. In our basis set of 1025 active molecules belonging to diverse classes, the number of unique CFS fragments ranged from 11 to 215, with an average of 63 fragments per CFS.

Molecular Similarity Analysis. We also explored whether the CFS activity class representation could be utilized for similarity analysis and the study of structure−activity relationships. Therefore, we have complemented the CFS with position-dependent fragment weights, as shown in Figure 3, which prioritizes fragment groups in the order of increasing specificity. Then we devised an approach to map database compounds to a weighted CFS, as summarized in Figure 4. Multiple core path alignments of an activity class are reduced to a weighted CFS to which arbitrary molecules can be compared by determining substructures that match the CFS and selecting the maximally weighted one. For a database compound, a score is calculated that combines the

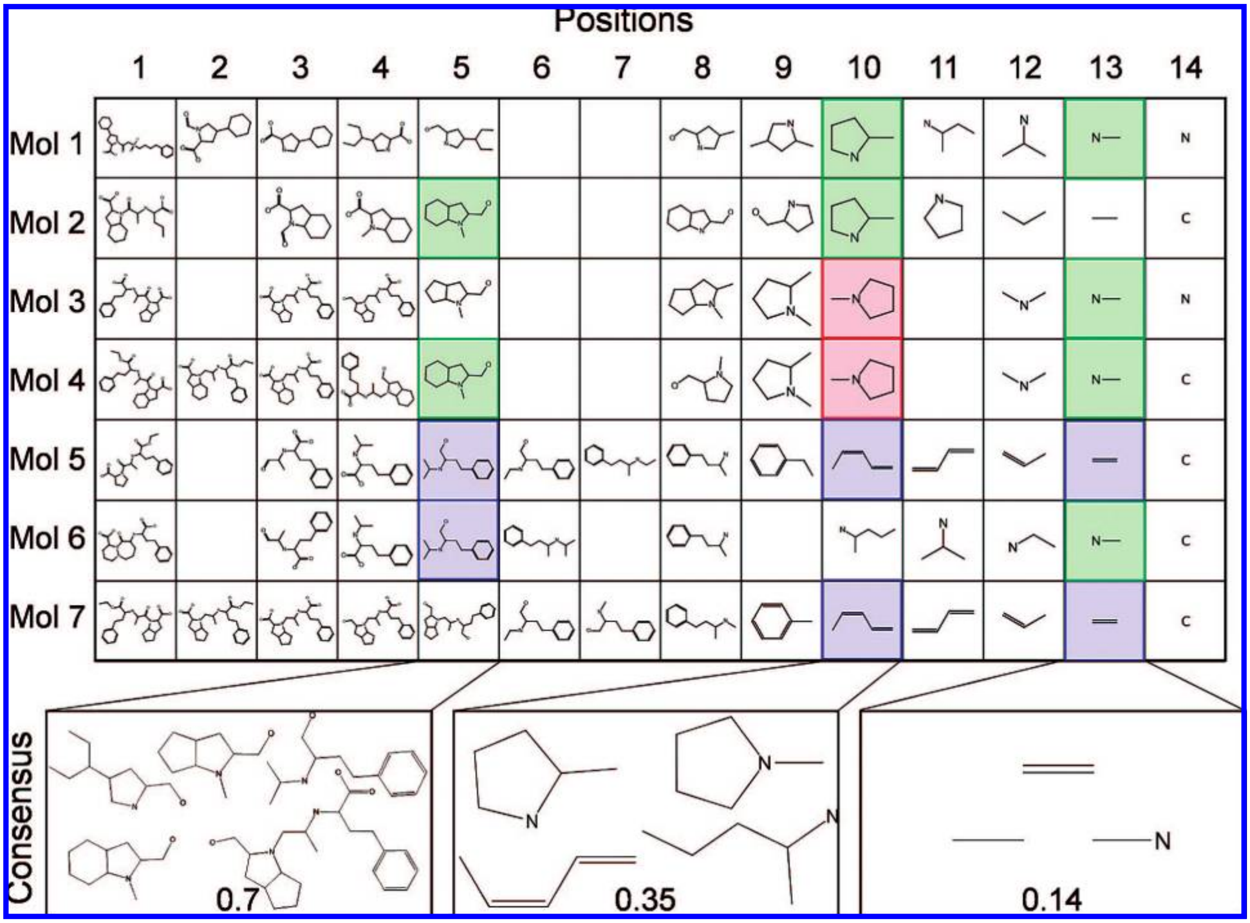

Figure 3. Multiple core path alignment. An exemplary multiple core path alignment is shown for a set of acetylcholine esterase inhibitors. Columns correspond to positions in the alignment. On the left, core paths start with the original molecules and, on the right, end with terminal atoms. For alignment positions 5, 10, and 13, identical fragments are color-coded and consensus fragments are shown. Fragment weights are calculated from the relative position in the consensus fragment sequence and are reported for each of the three positions.

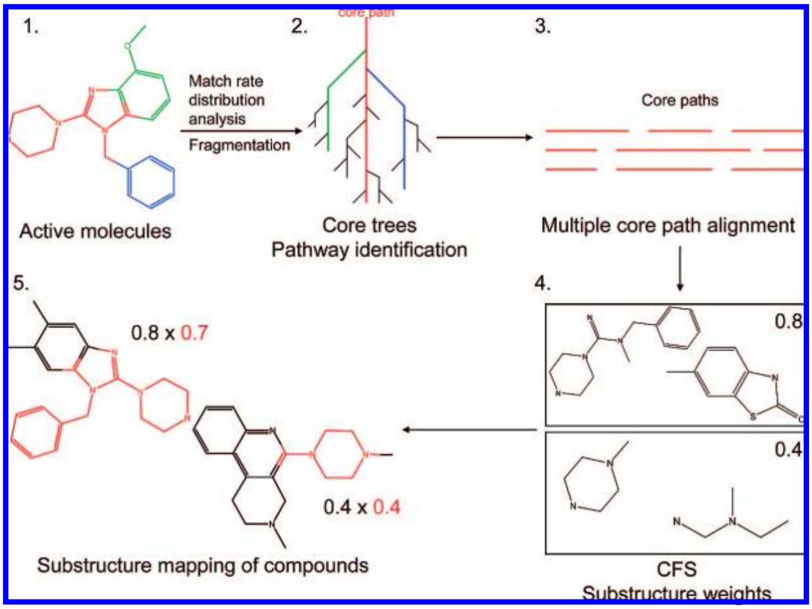

Figure 4. From core trees to database mining. The figure illustrates how molecular information contained in core trees of an activity class can ultimately be utilized to search for compounds having similar activity. Beginning from individual active molecules, five stages are distinguished, as discussed in the text.

CFS weight and the fragment to compound size ratio. This approach ranks database compounds and favors the detection of molecules that share large substructures with one or more molecules belonging to an activity class. In contrast to substructure searching on the basis of predefined fragment dictionaries, CFS-based searching uses multiple path alignment information and consensus fragments derived from active molecules. It thus represents an activity class-centric substructure search strategy.

Similarity Search Trials. We tested CFS-based database searching on five different activity classes of increasing intraclass structural diversity, which influences the composition of fragment sets.[13] However, for structurally diverse reference sets ACCS are consistently identified. The more diverse active compounds are, the smaller ACCS sets often become. Nevertheless, even small ACCS sets are found to produce stable cores in active compounds that permit the generation of core trees.[12] As reference calculations, centroid fingerprint searches were carried out using the publicly available set of 166 MACCS structural keys and evaluated on the basis of Tanimoto similarity. It should be noted that we have shown in a recent study that the performance of conventional fingerprint search calculations can be further improved by support vector machine classification using fingerprint descriptors.[20] However, we have used MACCS fingerprint searching here as a conventional and widely used reference calculation. The results are summarized in Figure 5. As expected, similarity search performance increases with increasing structural homogeneity of activity classes. Overall recovery rates between approximately 10% and close to 90% were observed and relative performance of CFS- and MACCS-base similarity searching varied dependent on the activity class and the reference sets used, which is reflected in the standard deviations reported in Figure 5. Standard deviations were comparable for CFS and MACCS. In our test calculations, CFS performed considerably better than MACCS for classes AA2 and XAN, whereas MACCS was superior on ARO and ACE. Thus, the results of these initial trials suggest that database searching using the weighted CFS of a compound class is comparable in performance to conventional similarity searching and might provide an advantage depending on the activity class. Given the scope of our analysis, a key aspect of these observations is that fragments isolated from core trees and multiple path alignments also have predictive value as descriptors for the evaluation of structure−activity relationships.

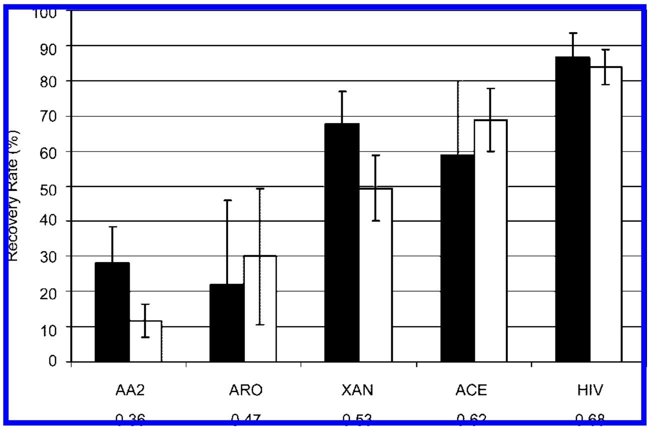

Figure 5. Database search trials. Recovery rates for 100 top-ranked database compounds are reported for CFS-based (black) and MACCS (white) similarity searching as averages over 10 independent trials with randomly selected compound reference sets. Activity classes are abbreviated according to Table 1, and average values of the Tanimoto coefficient for pairwise compound comparisons using MACCS are reported at the bottom as a measure of intraclass structural diversity. Error bars indicate the standard deviations from 10 independent trials.

CONCLUDING REMARKS

In this study, we have introduced a molecular representation and fragmentation scheme that is based on differences in atom match rates of mapped random fragment populations. Molecules are represented as fragmentation pathways of decreasing structural specificity. Most specific and activity class characteristic molecular information is encoded in core paths that usually contain the largest number of fragments. Core paths of individual molecules having similar activities are organized in alignments analogous to multiple biological sequence alignments. From core path alignments, consensus fragment sequences are generated for activity classes that organize unique fragments occurring within an activity class as alignment position-dependent sets of similar ones. Moreover, consensus fragment sequences can be weighted according to structural specificity and used to assess molecular similarity or search database for compounds having similar activity. Substructures in consensus fragment sequences have predictive power comparable to structural keys. Thus, consensus fragment sequences might also serve as a source for identifying novel compound class-directed structural descriptors. Multiple core path alignments and consensus fragment sequences add to the current repertoire of systematic or knowledge-based fragment design approaches. Further studies will investigate whether fragment sequence alignments can be extended to directly compare structural features of different compound classes.

REFERENCES AND NOTES

(1) Merlot, C.; Domine, D.; Cleva, C.; Church, D. J. Chemical substructures in drug discovery. *Drug Discovery Today* **2003**, *8*, 594–602.

(2) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(3) Barnard, J. M.; Downs, G. M. Chemical fragment generation and clustering software. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 141–142.

(4) McGregor, M. J.; Pallai, P. V. Clustering of large databases of compounds using MDL "keys" as structural descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443–448.

(5) 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.

(6) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP−retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.

(7) 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.

(8) Vidal, D.; Thormann, M.; Pons, M. LINGO, an efficient holographic text based method to calculate properties and intermolecular similarities. *J. Chem. Inf. Model.* **2005**, *45*, 386–393.

(9) Karwath, A.; De Raedt, L. SMIREP: predicting chemical activity from SMILES. *J. Chem. Inf. Model.* **2006**, *46*, 2432–2444.

(10) Batista, J.; Godden, J. W.; Bajorath, J. Assessment of molecular similarity from the analysis of randomly generated structural fragment populations. *J. Chem. Inf. Model.* **2006**, *46*, 1937–1944.

(11) Batista, J.; Bajorath, J. Mining of randomly generated molecular fragment populations uncovers activity-specific fragment hierarchies. *J. Chem. Inf. Model.* **2007**, *47*, 1405–1413.

(12) Lounkine, E.; Batista, J.; Bajorath, J. Mapping of activity-specific fragment pathways isolated from random fragment populations reveals the formation of coherent molecular cores. *J. Chem. Inf. Model.* **2007**, *47*, 2113–2119.

(13) Batista, J.; Bajorath, J. Chemical database mining through entropy-based molecular similarity assessment of randomly generated structural fragment populations. *J. Chem. Inf. Model.* **2007**, *47*, 59–68.

(14) 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.

(15) 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.

(16) Eidhammer, I.; Jonassen, I.; Taylor, W. R. Multiple Global Alignment and Phylogenetic Trees. In *Protein Bioinformatics: An Algorithmic Approach to Sequence and Structure Analysis*, 1st ed.; John Wiley & Sons Ltd.: West Sussex, England, 2004; p 83.

(17) Thompson, J. D.; Higgins, D. G.; Gibson, T. B. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **1994**, *29*, 4673–4680.

(18) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity Metrics for Ligands Reflecting the Similarity of the Target Proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391–405.

(19) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

(20) Geppert, H.; Horváth, T.; Gärtner, T.; Wrobel, S.; Bajorath, J. Support-Vector-Machine-Based Ranking Significantly Improves the Effectiveness of Similarity Searching Using 2D Fingerprints and Multiple Reference Compounds. J. Chem. Inf. Model. 2008, in press.

(21) Molecular Drug Data Report (MDDR). Elsevier MDL: San Leandro, CA. http://www.mdl.com (accessed Sep 1, 2006).

(22) Comprehensive Medicinal Chemistry Database (CMC-3D), Version 99.1; Elsevier MDL: San Leandro, CA. http://www.mdl.com (accessed Sep 1, 2006).

(23) Xue, L.; Godden, J. W.; Gao, H.; Bajorath, J. Identification of a preferred set of molecular descriptors for compound classification based on principal component analysis. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 699–704.

CI800020S