Performance of Similarity Measures in 2D Fragment-Based Similarity Searching: Comparison of Structural Descriptors and Similarity Coefficients

Xin Chen*,† and Charles H. Reynolds‡

Computer Assisted Drug Discovery, Johnson & Johnson Pharmaceutical Research and Development, L.L.C., Raritan, New Jersey 08869, and Spring House, Pennsylvania 19477

Received April 24, 2002

2D fragment-based similarity searching is one of the most popular techniques for searching a large database of chemical structures and has been widely applied in drug discovery. However, its performance, especially its effectiveness in retrieving active structural analogues, has not been adequately studied. We report a series of computational experiments, where we systematically studied the influence of structural descriptors and similarity coefficients on the effectiveness of similarity searching. The study was conducted using two public large data sets, NCI anti-AIDS and MDDR. Four sets of 2D linear fragment descriptors, based on the original definitions of atom pairs and atom sequences, were compared. The effect of using the Tanimoto coefficient and the Euclidean distance was studied as a function of descriptor set. The results clearly indicate that the Tanimoto coefficient is superior to the Euclidean distance in 2D-fragment based similarity searching, in terms of hit rate, while atom sequences demonstrate the best overall performance among the structural descriptors we studied.

INTRODUCTION

Quantitative measurement of chemical similarity plays a central role in a variety of computer assisted drug discovery (CADD) techniques, such as database searching,[1] diversity analysis,[2] and QSAR analysis,[3] etc. The basic assumption underlying these applications is *the similar property principle*,[4] which states that similar chemical structures should lead to similar physicochemical properties and biological activities. Many similarity measures have been developed to try quantifying the degree of resemblance between chemical structures. Although most of these measures have demonstrated their usefulness in many successful CADD applications,[5,6] they are rather poorly characterized. Particularly, the quantitative evaluation and comparison of the effectiveness of different similarity measures remains a challenge. More studies are still needed for us to fully understand what computational schemes are most effective for measuring chemical similarity under various conditions.

Generally speaking, similarity measures involve three principal components:[1] *structural descriptors* that represent chemical structures in a way so that they can be easily compared, *weighting schemes* that assign the relative importance to different structural descriptors, and *similarity coefficients* that provide the mathematical function for calculating a similarity value based on the weighted values of structural descriptors. For each component, many choices can be made, their combination leading to a large number of possible similarity measures. Consequently, evaluating the performance of all these measures is a daunting task. Furthermore, the performance of an individual similarity measure may be influenced by many factors, like application types, biological targets, etc., which makes any comprehensive evaluation even more complicated.

However, a number of studies have been reported that try to address this challenge of comparing the performance of different similarity measures in different applications. Adamson and Bush[7] first compared the performance of eight different similarity/dissimilarity measures, using augmented atoms[8] as structural descriptors. The performance was evaluated by two independent "simulated property prediction" experiments. The first approach was to estimate the biological property of a compound based on the observed property value of its nearest neighbor. The second one was to estimate the property as the average property value of a cluster after clustering classification. A set of 39 compounds with known local anesthetic activities was used as the test data set. Later, Willett and Winterman[9] examined the performance of the combination of two structural descriptors, six weighting schemes and six similarity coefficients in a similar fashion, in which the prediction was based on the nearest neighbor. A collection of sixteen small QSAR/QSPR data sets (about 795 compounds in total) was used as the test data set in their work. Recently, Brown and Martin[10] reported their investigation of the performance of several clustering methods, which are all based on the calculation of similarity/dissimilarity between chemical structures, for compound selection. The performance was evaluated by the average hit rate in the active clusters generated by clustering classification. A couple of public and in-house data sets, containing hundreds to thousands of compounds in each, were tested in their work. Some other comparison studies have also been reported by Matter et al.[11,12] and Patterson et al.[13]

Our primary interest is evaluating and comparing the performance of various similarity measures in 2D fragment-based similarity searching. Similarity searching has demon-

* Corresponding author e-mail: xchen3@prdus.jnj.com.
† Raritan.
‡ Spring House.

10.1021/ci025531g CCC: $22.00
Published on Web 10/15/2002

strated itself as a productive tool of lead identification in the early stage of drug discovery.[5,6] Among the similarity searching methods that have been developed, those based on 2D fragment descriptors have the obvious advantage of simplicity and speed, making them ideal for handling large chemical databases. They have been chosen as the default set in almost all the commercial chemical database searching systems.[14,15] Also, they have, surprisingly, shown better performance than those based on more sophisticated structural descriptors, like 3D pharmacophore descriptors, in some comparison studies.[10]

We report a series of "simulated similarity searches", conducted on the NCI anti-AIDS[16] and MDDR[17] data sets. Four sets of 2D linear fragment descriptors, based on the original definitions of atom pairs[18] and atom sequences,[19] were used in this study as were three forms of the Tanimoto coefficient and the Euclidean distance. The influence of these structural descriptors and similarity coefficients on the effectiveness of retrieving active structural analogues was studied systematically. It should be noted that the effect of weighting schemes was not considered as part of this work, or state in another way, all the descriptors were treated equally.

METHODS

Test Data Sets. The choice of test data set is critical for fairly evaluating the performance of different similarity measures. An ideal data set should satisfy the following two requirements. First, it should cover the whole chemical space as broadly as possible so that any evaluation based on it would not be biased to a particular region in chemical space. Second, it should sample the chemical regions it covers as thoroughly as possible so that it can really test the capability of a similarity measure to tell between active and inactive structural analogues.

To satisfy these two requirements as far as possible, two large, public data sets, NCI anti-AIDS[16] and MDDR,[17] were selected as the test data sets. The NCI anti-AIDS data set[16] contains structure and activity data for tens of thousands of compounds screened by the NCI AIDS antiviral screening program,[20] which uses a soluble formazan assay to measure the protection of human CEM cells from HIV-1 infection. It represents a large data set composed of both active and inactive compounds against a specific therapeutic target and hopefully provides a thorough sampling of a particular region in chemical space. The MDDR data set[17] is a commercial database containing over hundreds of thousands of pharmaceutically relevant compounds, with data on their activity categories. It represents a large data set comprised of active compounds against multiple therapeutic targets and should give a broad coverage of the chemical space that is pharmaceutically relevant. We recognize that neither data set alone is perfect for the evaluation purpose. However, they probably represent the best test data sets that are available, and they are complementary to each other for the evaluation purpose. We believe that together they should provide a good basis for evaluating the performance of different similarity measures.

Before any analysis, both data sets were first cleaned, removing all the registrations with no structural information and all compounds with a molecular weight over 800. After cleaning, the NCI anti-AIDS data set contains 41,588 compounds, including 339 confirmed actives (CAs), 987 confirmed moderately actives (CMs) and 40,262 confirmed inactives (CIs), while the MDDR data set contains 104,790 compounds related to around 600 different activity classes. In the NCI anti-AIDS data set, both the CAs and the CMs were treated equally as actives and the CIs as inactives for this work. In the MDDR data set, 9 particular activity classes, as listed in Table 1, were selected for study, based on the large number of the compounds that belong to these classes and the clear definition of their biological targets. A total of 6677 compounds in these 9 classes were then treated as actives, while all other compounds were treated as inactives.

Table 1. Activity Classes Selected as Actives for Study in the MDDR Data Set

class	no. of compds
ACAT inhibitor	1398
farnesyl protein transferase inhibitor	857
HIV-1 protease inhibitor	750
phospholipase A2 inhibitor	712
H+/K+-ATPase inhibitor	696
acetylcholinesterase inhibitor	678
cyclooxygenase-2 inhibitor	557
collagenase inhibitor	526
ACE inhibitor	503

Structural Descriptors. There are three popular ways of generating 2D fragment descriptors. The first one, represented by MACCS keys,[21] uses a fragment library that usually contains a couple of hundred or even less fragments that are frequently observed in chemical structures. This approach has the limitation that it is biased to a predefined fragment library and the structural features not defined in the library will not be reflected in the subsequent similarity calculations. The second one, represented by Daylight fingerprints,[22] uses a set of predefined rules to enumerate all the fragments that satisfy these rules from a chemical structure. This approach avoids the "library bias" problem, which exists in the first approach. Since a large number of fragments can be generated in this way, they are typically hashed/folded into a bit string with the fixed length to save memory space. This may lead to loss of certain structural information and introduces additional noise into subsequent calculations. The third approach, exemplified by the work at Lederle,[18] is basically the same as the second, except that it does not use the hashing/folding technique. Instead, it usually uses a linked list to save all the unique fragments together with their number of occurrences in a chemical structure. Compared to the first two approaches, this linked-list approach keeps more structural information of a chemical compound, with the tradeoff that it requires more memory space and is more computationally expensive. Nevertheless, with the current hardware and careful programming, this tradeoff is becoming more and more favorable. In our experience, the third approach (rule-based enumeration without hashing) can be used to search a typical corporate database, containing hundreds of thousands of compounds, in a very comfortable time range. Therefore, we elected to use this approach to represent chemical structures for our work.

The descriptors studied in this work are based on the original definitions of atom pairs[18] and atom sequences,[19] with some modifications. An *atom pair (AP)* is defined as two atoms and the length of the shortest bond path between

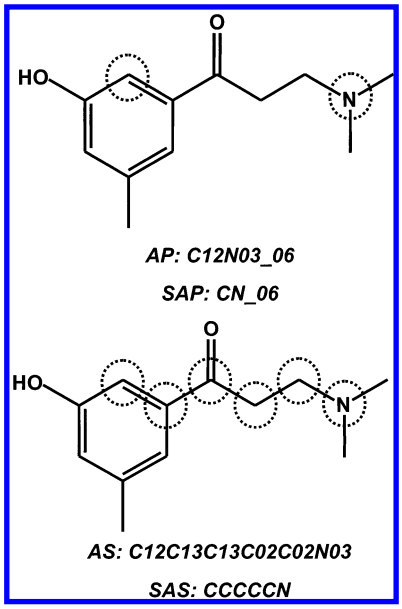

Figure 1. Illustration of the definitions of atom pair (AP), atom sequence (AS), simplified atom pair (SAP) and simplified atom sequence (SAS). Each atom is described either by its element type, the number of π electrons that it bears and the number of non-hydrogen atoms bonded to it (e.g. C12 means a carbon bearing 1 π electron and having 2 non-hydrogen bonded neighbors) or by its element type only.

them, which is counted as the number of atoms along the path including two terminal atoms. An *atom sequence (AS)* is defined as all the atoms in the shortest bond path between two atoms, including two terminal atoms. In APs and ASs, each atom can be described by its element type, the number of π electrons that it bears and the number of non-hydrogen atoms bonded to it, as in the original work of Carhart et al.[18] We have also considered describing an atom simply by its element type only and called the atom pair and the atom sequence defined in this way *simplified atom pair (SAP)* and *simplified atom sequence (SAS)*, respectively, to distinguish them from the definitions given above. Examples of these definitions are given in Figure 1. All four descriptors are linear fragment descriptors. They were chosen because they are easy to implement and compare. In contrast, it is rather difficult to manipulate and match nonlinear fragment descriptors, like multi-level near neighbor descriptors,[23] especially when their sizes go beyond one layer of nearest neighbors.

All fragments with lengths of 2 to 20 atoms were recorded, with the counts of their occurrences in a chemical structure. All the fragments longer than 20 atoms were discarded. As in the work of Carhart et al.,[18] thirteen explicit element types, C, O, N, S, F, Cl, Br, I, P, Si, B, Se and As, were considered, and one new type, Y, was used to represent elements of any other types. In cases where there were more than one shortest bond paths between two atoms, all the paths were enumerated as AS/SAS, while only one was counted as AP/SAP. All the descriptors for a compound were saved as a linked list in memory.

Similarity Coefficients. In general, there are two major classes of similarity coefficients: association and distance coefficients.[1] The essential difference between them is that the latter considers the common absence of certain structural features as the evidence of similarity between two chemical structures whereas the former does not.[1] The Tanimoto coefficient and the Euclidean distance were chosen to represent these two classes, respectively, since they are probably the most widely used coefficients in similarity searching and they also show high correlations with many of the other members in their classes.

One interesting issue related to fragment-based similarity searching is whether we should count the occurrence of fragment in a chemical structure. It has been speculated that including the occurrence information may improve the "distinguish-ability" of a descriptor set and consequently lead to an increased hit rate in similarity searching.[24] However, some studies have shown that counting fragment occurrences did not significantly improve performance.[10] As part of this work, we investigated the effect of including or omitting the count.

When the fragment occurrences in a chemical structure are considered, the counts of occurrences can be treated as algebraic variables, or each fragment can be viewed as an individual member of a set. This leads to two different definitions of similarity coefficients: *algebraic form* and *set-theoretic form*, respectively. If we only consider the absence or presence of unique fragments, these two forms converge into one: *binary form*. These three different definitions of the Tanimoto coefficient and the Euclidean distance are summarized in Table 2, where a, b and c are the numbers of unique fragments in compound A, in compound B and shared by compounds A and B, respectively, while $n_{A,i}$ and $n_{B,i}$ are the numbers of fragment i in compounds A and B, respectively.

Computational Experiment Design. Using the NCI anti-AIDS and MDDR data sets as test data sets, a series of "simulated similarity searching" studies was conducted to compare the effectiveness of different similarity measures. In each data set, each active compound was used as a probe to search the remainder of the data set and order compounds according to the calculated similarity values, from most to least similar. The averaged accumulative percentages of actives, i.e., hit rates, versus the number of nearest neighbors were recorded. In the MDDR data set, since 9 different activity classes are represented (cf. Table 1), all the compounds in the same activity class as the probe were treated as the actives in that round of similarity searching, while all the others were treated as inactives.

The choice of the number of nearest neighbors, instead of the absolute similarity value, as the abscissa is based on the observation that similarity values calculated in different ways may have different significance and cannot be directly compared.[24] Also, the real interest in similarity searching is actually the capability to separate active and inactive structural analogues, so the relative closeness rankings, instead of the absolute similarity values, are more important in practical use.

RESULTS

Comparison of the Tanimoto Coefficient and the Euclidean Distance. The Tanimoto coefficient and the Euclidean distance were first compared using the four sets of descriptors, ASs, APs, SASs and SAPs. A similar trend of relative performance was observed for all of them, and the results obtained with ASs are shown in Figure 2, where

Table 2. Definitions of the Tanimoto Coefficient and the Euclidean Distance[a]

	Tanimoto coefficient	Euclidean distance
binary form	$S_{A,B} = \frac{c}{a+b-c}$	$D_{A,B} = [a+b-2c]^{1/2}$
algebraic form	$S_{A,B} = \frac{\sum_{i=1}^{m} n_{A,i} n_{B,i}}{\sum_{i=1}^{m} n_{A,i}^{\,2} + \sum_{i=1}^{m} n_{B,i}^{\,2} - \sum_{i=1}^{m} n_{A,i} n_{B,i}}$	$D_{A,B} = [\sum_{i=1}^{m} (n_{A,i} - n_{B,i})^2]^{1/2}$
set-theoretic form	$S_{A,B} = \frac{\sum_{i=1}^{m} \min(n_{A,i}, n_{B,i})}{\sum_{i=1}^{m} n_{A,i} + \sum_{i=1}^{m} n_{B,i} - \sum_{i=1}^{m} \min(n_{A,i}, n_{B,i})}$	$D_{A,B} = \sum_{i=1}^{m} \lvert n_{A,i} - n_{B,i} \rvert$

[a] a: the number of unique fragments in compound A; b: the number of unique fragments in compound B; c: the number of unique fragments shared by compounds A and B; $n_{A,i}$: the number of fragment i in compound A; $n_{B,i}$: the number of fragment i in compound B.

the Tanimoto results (filled circles) are compared to the Euclidean results (unfilled circles), defined in the three different forms (cf. Table 2) and tested in the two data sets. The *x*-axis plots the incremental number of nearest neighbors, *n*, around an active compound, up to ca. 10% of the total number of compounds in the data set, i.e., 4,000, for the NCI anti-AIDS data set and 10,000 for the MDDR data set. It is plotted on a logarithmic scale, to show the difference at the near end. The *y*-axis plots the averaged percentage of the actives, i.e., hit rate, among the *n* nearest neighbors of an active compound. In all the plots, the horizontal dotted line represents the average percentage of active compounds in the whole data set, which is equal to 3.19 for the NCI anti-AIDS data set or 0.71 for the MDDR data set and can be interpreted as the hit rate for a random selection.

As can be seen, the Tanimoto coefficient gives considerably better results than the Euclidean distance in both the data sets, regardless of the specific formalism used. The two coefficients converge to give a similar hit rate at both the near and the far ends of the neighborhood, indicating that the major difference between them exists in their relative ability to distinguish moderately similar structures.

From Figure 2, we can also see that the highest hit rate that can be reached is only about 40% for the NCI anti-AIDS data set while it is around 90% for the MDDR data set. This reflects the difference between these two data sets: NCI anti-AIDS data set is a more focused data set containing many inactive structural analogues while MDDR data set is a more broad data set composed of multiple activity classes. The performance curve shown for the NCI anti-AIDS data set therefore gives us a good estimate of the hit rates that can be probably achieved in practical use, although it is biased toward one particular, i.e., anti-AIDS, therapeutic area.

Comparison of the Different Definitions of the Tanimoto Coefficient. Since the Tanimoto coefficient consistently showed better performance than the Euclidean distances, it was examined further by comparing the performance of its three different definition formalisms (cf. Table 2). All the four sets of descriptors were used, respectively. The results obtained using ASs are shown in Figure 3, and the similar trends were also observed for the other three descriptors. In Figure 3, the set-theoretic results (filled circles), the algebraic results (unfilled diamonds) and the binary results (unfilled circles) are compared in the two data sets, respectively.

As can be seen from Figure 3, the binary and the set-theoretic forms give highly similar results for both the test data sets, with the latter being slightly better. Interestingly, the algebraic form gives the performance inferior to the binary form for both data sets, even though the occurrences of fragments have been counted. The closeness of the performance curves shown in this figure may be due to the fact that many of the descriptors used for our data sets are unique. For example, the percentages of the unique AS descriptors are 54.07% in the NCI anti-AIDS data set and 59.29% in the MDDR data set. However, the relative performances of the three forms are consistent for both data sets, and we expect to see more obvious separation with less unique descriptors.

Comparison of the Different 2D Fragment Descriptors. Based on the results described in the previous sections, the Tanimoto coefficient defined in the set-theoretic form emerged as the best choice. Even though the Tanimoto coefficient defined in the binary form showed very comparable performance, we prefer the set-theoretic form because using the binary form would not save appreciable memory space when the data structure of linked list is employed. Furthermore, counting occurrences keeps more structural information and may be important in some cases.

The performance of the four descriptors (AS, AP, SAS, SAP) was compared using the set-theoretic Tanimoto coefficient. The results are shown in Figure 4. The AS results (filled circles), the AP results (unfilled circles), the SAS results (filled diamonds) and the SAP results (unfilled diamonds) are compared for each of the two test data sets.

Generally, performance decreases in the order of ASs, APs, SASs and SAPs for both data sets, as shown in Figure 4. The performance difference among them is more obvious in the MDDR data set. In the NCI anti-AIDS data set, ASs and APs have essentially the same performance, but both outperform SASs and SAPs. Interestingly, this trend is consistent with increasing degrees of fuzziness as one goes

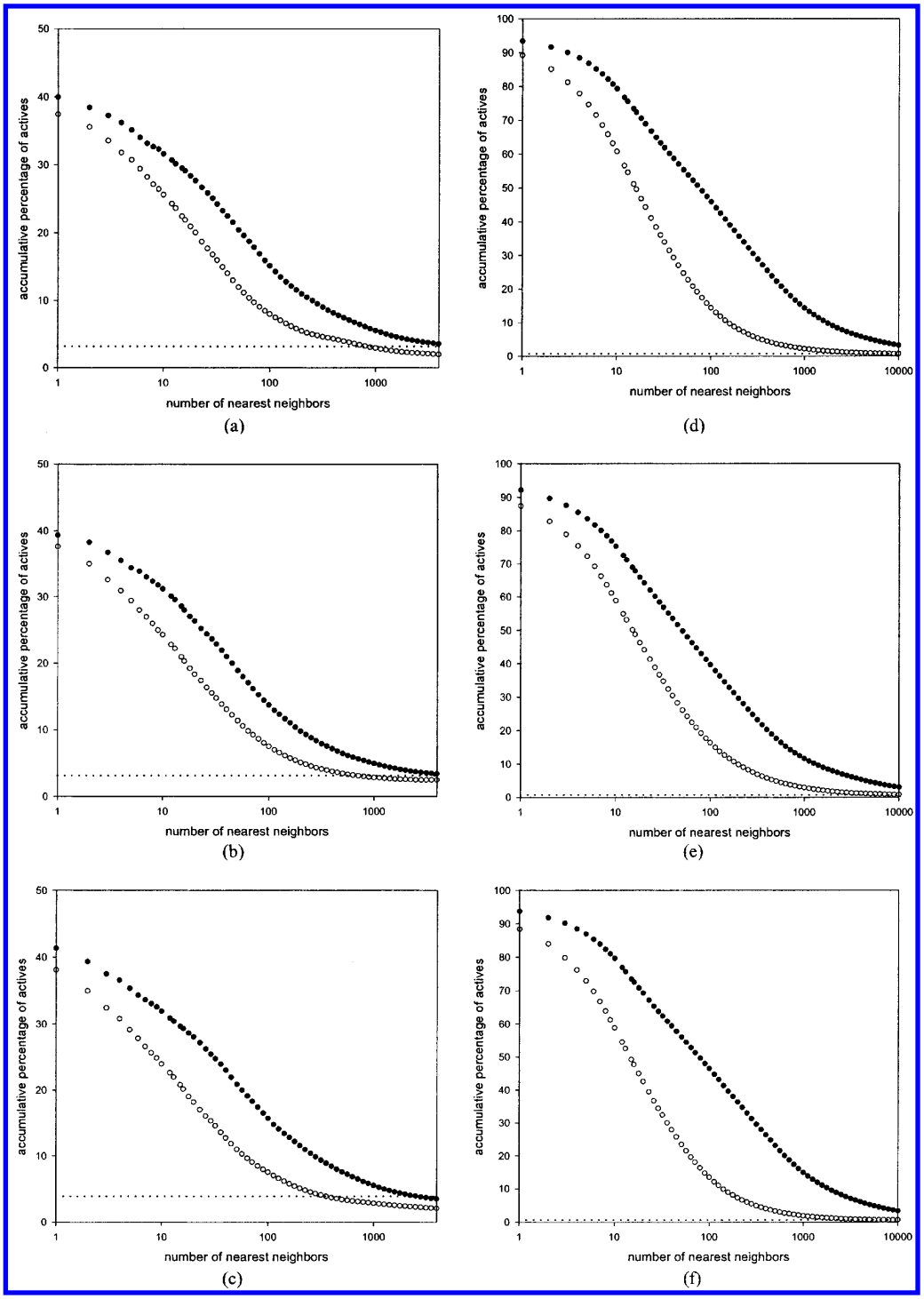

Figure 2. Comparison of the performance of the Tanimoto coefficient (filled circles) and the Euclidean distance (unfilled circles), using atom sequences (ASs) as structural descriptors. Three different definitions (cf. Table 2) were tested, separately, in two data sets: (a) binary form, in the NCI anti-AIDS data set; (b) algebraic form, in the NCI anti-AIDS data set; (c) set-theoretic form, in the NCI anti-AIDS data set; (d) binary form, in the MDDR data set; (e) algebraic form, in the MDDR data set; (f) set-theoretic form, in the MDDR data set.

from AS to SAP. Figure 5 shows the average numbers of unique fragments per compound in the two test data sets, which can be viewed as the indication of the degree of "fuzziness" of these four descriptors. Using this criterion the degree of fuzziness clearly increases in the order AS, AP, SAS and SAP.

Comparison with ISIS Similarity Search Method. Since the combination of ASs and the Tanimoto coefficient (defined in set-theoretic form) emerges as the best combination in this study, we further compare this combination with the similarity search method implemented in the popular ISIS chemical database management system.[14]

ISIS system uses the Tanimoto coefficient in the binary form as the default similarity coefficient and the MACCS keys as the default structural descriptors. MACCS keys are composed of 166 pre-selected structural keys and were originally designed for preprocessing sub-structural searching. In a previous study,[10] they gave the best performance of separating active and inactive compounds into clusters. Therefore, the ISIS similarity searching method provides a

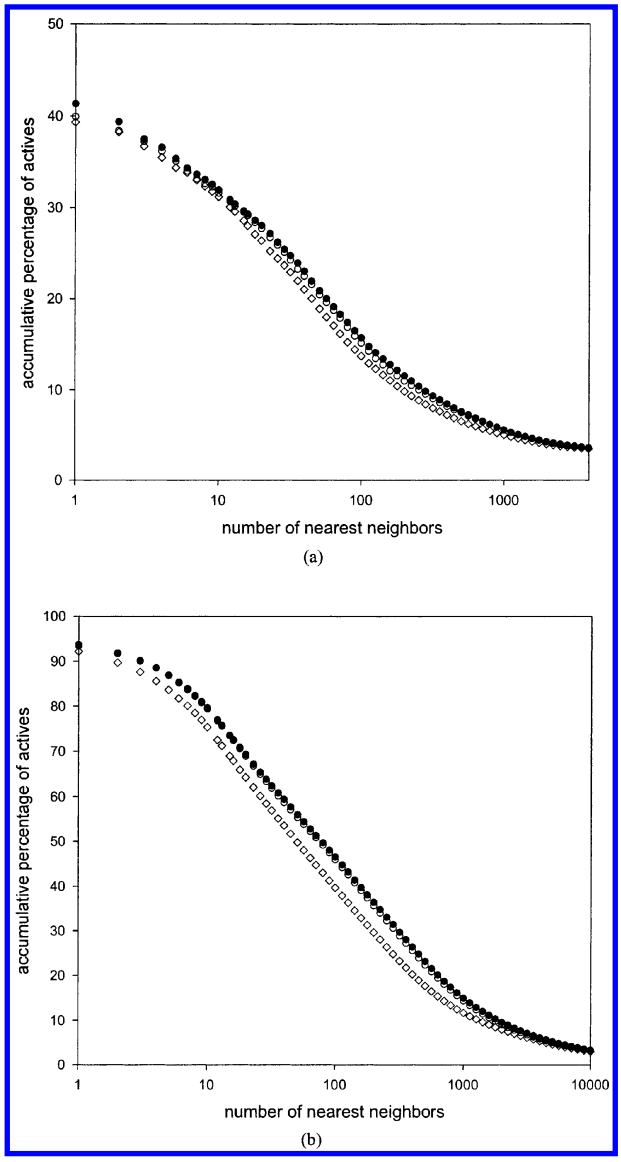

Figure 3. Comparison of the performance of the three different definitions of the Tanimoto coefficient (cf. Table 2): binary form (unfilled circles), algebraic form (unfilled diamonds), and set-theoretic form (filled circles), in the two data sets: (a) NCI anti-AIDS data set and (b) MDDR data set. Atom sequences (ASs) were used as structural descriptors.

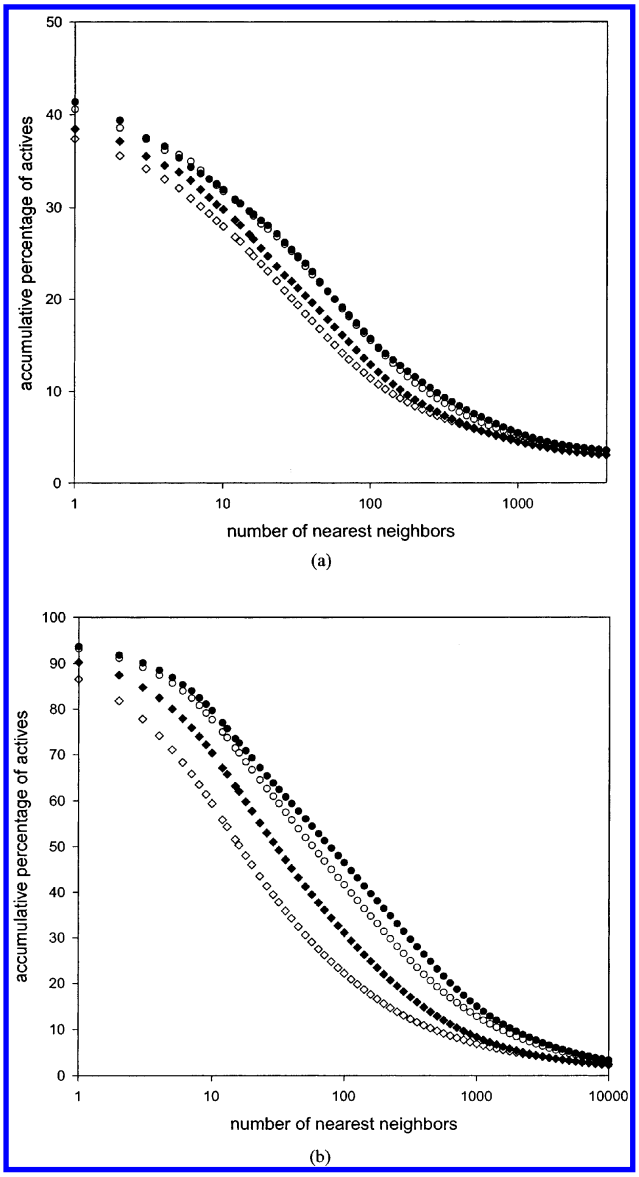

Figure 4. Comparison of the performance of the four different 2D fragment descriptors: atom sequences (filled circles), atom pairs (unfilled circles), simplified atoms sequences (filled diamonds), and simplified atom pairs (unfilled diamonds), in the two data sets: (a) NCI anti-AIDS data set and (b) MDDR data set. The set-theoretic form of the Tanimoto coefficient was used as similarity coefficient.

good benchmark for testing the performance of other methods.

A comparison of the results using ASs and the Tanimoto coefficient in set-theoretic form (filled circles), and the results of using MACCS keys and the Tanimoto coefficient in binary form (unfilled circles), for both data sets is given in Figure 6. As can be seen, the combination of ASs and the set-theoretic Tanimoto gives significantly better results than ISIS for both data sets.

DISCUSSIONS

The performance of similarity measures is expected to depend, at least to some degree, on the specific applications where they are applied. Therefore, understanding their behavior in a variety of circumstances is useful for practical drug discovery. We have presented a systematic study of a series of structural descriptors and similarity coefficients in the context of 2D fragment-based similarity searching. Some practical guidelines that emerged from this study are summarized as follows.

First, the Tanimoto coefficient performs considerably better than the Euclidean distance in all the tests. This result is consistent with the former finding of Willett et al.'s[9] but is more convincing since it was obtained with much larger, and more diverse, test data sets. Since the Tanimoto coefficient only treats the common presence of a structural feature as the evidence of similarity while the Euclidean distance also considers the common absence of a structural feature, it is tempting for us to attach a general conclusion to this result. It is plausible that the common presence of certain structural features is the primary factor in determining whether two chemical structures should have a similar biological activity, while the common absence of some structural features is at best secondary. This conclusion is compatible with the popular model of drug-receptor interactions, which states that two compounds need first to share,

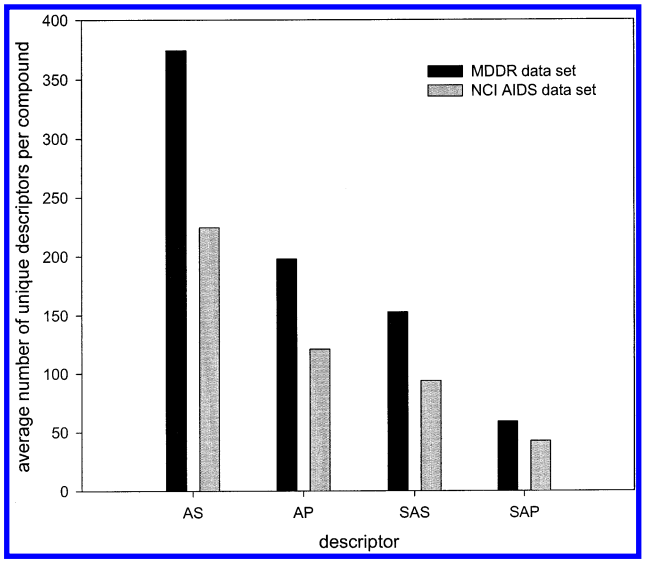

Figure 5. Distribution of the average number of unique descriptors: atom sequence (AS), atom pair (AP), simplified atom sequence (SAS) and simplified atom pair (SAP), per compound in MDDR data set (in black) and NCI anti-AIDS data set (in gray).

not miss, certain structural features (i.e. pharmacophore) in order to bind to the same receptor and exert the similar pharmacological effect. Missing certain structural features may also be important in some cases, but only after the pharmacophore has been satisfied. Therefore, the performance difference observed here for the Tanimoto coefficient vs the Euclidean distance may be safely generalized to the association coefficients vs the distance coefficients, if fragment descriptors, like the atom pairs and the atom sequences used in this work, are used to represent chemical structures. It is also intriguing to speculate that a reversed preference may hold for toxicity studies, where certain structural features should be avoided.

Second, when considering the occurrences of different fragments in a chemical structure, the set-theoretic definition of similarity coefficients appears to be superior to the algebraic definition. This result indicates that individual fragments should be treated as set members when occurrences are considered. Treating the count of occurrences as a continuous variable is not helpful for improving performance, compared to not counting the occurrences. Interestingly, the binary definition performs almost equally well in this study, as compared to the set-theoretic definition. This may be due to the characteristics of the descriptors we studied in this work. They are generally quite specific and lead to a large number of unique descriptors, which may dominate any statistical comparison. So, we speculate that using the set-theoretic definition should show obvious superiority to using binary definition in some other cases.

Third, with the four sets of structural descriptors studied in this work, an increased degree of fuzziness leads to a decreased hit rate and the atom sequences show the best performance overall. More fuzzy descriptors are expected to bring more remotely similar structures that are also active closer to the probe in similarity searching. However, the result indicates that they may introduce even more inactive structures and lead to a decreased hit rate. Generally speaking, a "good" descriptor set for similarity searching should have a balance between fuzziness and specificity, in

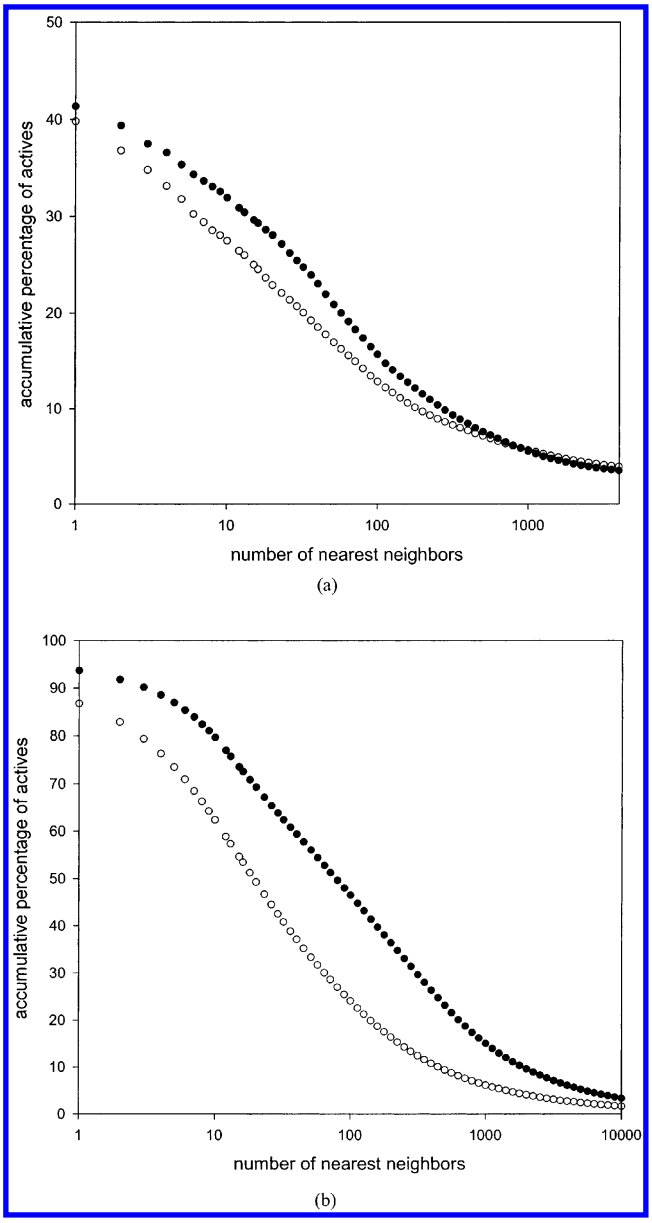

Figure 6. Comparison of the performance of using atom sequences (ASs) and Tanimoto coefficient (in set-theoretic form) (filled circles) and using the ISIS method (unfilled circles): MACCS keys and Tanimoto coefficient (in binary form), in the two data sets: (a) NCI anti-AIDS data set and (b) MDDR data set.

terms of the ability to tell active and inactive structural analogues. Descriptors that are too fuzzy may bring in more false positives, while descriptors that are too specific lead to more false negatives. Both will give a decreased hit rate in similarity searching. We expect that further specifying the atom sequences as we defined in this work may decrease the performance.

Finally, it is worth noting that all the conclusions drawn from this study are in the context of 2D fragment-based similarity searching. Whether they can be extended to 3D or other descriptors is open to question.

CONCLUSIONS

The performance of some similarity measures was systematically investigated by a series of "simulated similarity searching" studies. The results provide some new insights on how to conduct 2D fragment-based similarity searching

for retrieving active structural analogues in drug discovery. Association coefficients may be more suitable for 2D fragment-based similarity searching than distance coefficients; each fragment descriptor of a chemical structure should best be treated as an independent member of a set; and atom sequences seem to perform best among the descriptors we studied. Consequently, we strongly recommend using the combination of atom sequences and the set-theoretic Tanimoto coefficient in similarity searching. This work also demonstrates that the performance of similarity searching methods can be easily improved, compared to those implemented in the commercial searching systems, by carefully choosing similarity coefficients and structural descriptors.

REFERENCES AND NOTES

(1) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−997.
(2) Lajiness, M. S. Dissimilarity-based Compound Selection Techniques. *Perspective Drug Discovery Design* **1997**, *7/8*, 65−84.
(3) Good, A. C.; So, S.-S. Richards W. G. Structure−Activity Relationships from Molecular Similarity Matrices. *J. Med. Chem.* **1993**, *36*, 433−438.
(4) Maggiora, G. M.; Johnson, M. A. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York, 1990; pp 99−117.
(5) Kulagowski, J. J.; Broughton, H. B.; Curtis, N. R.; Mawer, I. M.; Ridgill, M. P.; Baker, R.; Emms, F.; Freedman, S. B.; Marwood, R.; et al. 3-[[4-(4-Chlorophenyl)piperazin-1-yl]methyl]-1H-pyrrolo[2,3-b]pyridine: An Antagonist with High Affinity and Selectivity for the Human Dopamine D4 Receptor. *J. Med. Chem.* **1996**, *39*, 1941−1942.
(6) Hipskind, P. A.; Lobb, K. L.; Nixon, J. A.; Britton, T. C.; Bruns, R. F.; Catlow, J.; Dieckman-McGinty, D. K.; Gackenheimer, S. L.; Gitter, B. D.; Iyengar, S.; Schober, D. A.; Simmons, R. M. A.; Swanson, S.; Zarrinmayeh, H.; Zimmerman, D. M.; Gehlert, D. R. Potent and Selective 1,2,3-Trisubstituted Indole NPY Y-1 Antagonists. *J. Med. Chem.* **1997**, *40*, 3712−3714.
(7) Adamson, G. W.; Bush, J. A. Comparison of the Performance of Some Similarity and Dissimilarity Measures in the Automatic Classification of Chemical Structures. *J. Chem. Inf. Comput. Sci*. **1975**, *15*, 55−58.
(8) Adamson, G. W.; Lynch, M. F.; Town, W. G. Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-based File. Part II. Atom-center Fragments. *J. Chem. Soc.* **1971**, *C*, 3702−3706.
(9) Willett, P.; Winterman, V. A. Comparison of Some Measures for the Determination of Intermolecular Structural Similarity. Measures of Intermolecular Structural Similarity. *Quant. Struct.-Act. Relat.* **1986**, *5*, 18−25.
(10) Brown, R. D.; Martin, Y. C. Use of Structure−Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 572−584.
(11) 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.
(12) Matter, H.; Potter T. Comparing 3D Pharmacophore Triplets and 2D Fingerprints for Selecting Diverse Compound Subsets. *J. Chem. Inf. Comput. Sci*. **1999**, *39*.
(13) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.
(14) ISIS/Base, MDL Information Systems, Inc. 14600 Catalina Street, San Leandro, CA 94577, USA.
(15) Daylight Chemical Information Software. Daylight Information Systems, Inc. 441 Greg Avenue, Santa Fe, NM 87501, USA.
(16) The NCI AIDS data set was downloaded from the NCI Developmental Therapeutics Program website (http://dtp.nci.nih.gov/docs/aids/aids_data.html).
(17) MDDR (MDL Drug Data Report) data set is available from MDL Information Systems Inc., San Leandro, CA, 94577.
(18) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure−Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.
(19) Randic, M.; Wilkins, C. L. Graph Theoretic Approach to Recognition of Structural Similarity in Molecules. *J. Chem. Inf. Comput. Sci*. **1979**, *19*, 31−37.
(20) Weislow, O. S.; Kiser, R.; Fine, D. L.; Bader, J.; Shoemaker, R. H.; Boyd, M. R. New Soluble-formazan Assay for HIV-1 Cytopathic Effects: Application to High-flux Screening of Synthetic and Natural Products for AIDS-antiviral Activity. *J. Natl. Cancer Inst.* **1989**, *81*, 577−586.
(21) MACCS II Manual. MDL Information Systems, Inc. 14600 Catalina Street, San Leandro, CA 94577, USA.
(22) Daylight Theory Manual. Daylight Information Systems, Inc. 441 Greg Avenue, Santa Fe, NM 87501, USA.
(23) Filimonov, D.; Poroikov, V.; Borodina, Y.; Gloriozova, T. Chemical Similarity Assessment through Multilevel Neighborhoods of Atoms: Definition and Comparison with the Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 666−670.
(24) Flower, D. R. On the Properties of Bit String-based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379−386.

CI025531G