Ties in Proximity and Clustering Compounds

John MacCuish,* Christos Nicolaou,[†] and Norah E. MacCuish[‡]

Bioreason, Inc., 150 Washington Avenue, Santa Fe, New Mexico 87501, and
Daylight Chemical Information Systems, 441 Greg Avenue, Santa Fe, New Mexico 87501

Received June 30, 2000

Hierarchical clustering algorithms such as Wards or complete-link are commonly used in compound selection and diversity analysis. Many such applications utilize binary representations of chemical structures, such as MACCS keys or Daylight fingerprints, and dissimilarity measures, such as the Euclidean or the Soergel measure. However, hierarchical clustering algorithms can generate ambiguous results owing to what is known in the cluster analysis literature as the ties in proximity problem, i.e., compounds or clusters of compounds that are equidistant from a compound or cluster in a given collection. Ambiguous ties can occur when clustering only a few hundred compounds, and the larger the number of compounds to be clustered, the greater the chance for significant ambiguity. Namely, as the number of "ties in proximity" increases relative to the total number of proximities, the possibility of ambiguity also increases. To ensure that there are no ambiguous ties, we show by a probabilistic argument that the number of compounds needs to be less than $2(n^{1/4})$, where n is the total number of proximities, and the measure used to generate the proximities creates a uniform distribution without statistically preferred values. The common measures do not produce uniformly distributed proximities, but rather statistically preferred values that tend to increase the number of ties in proximity. Hence, the number of possible proximities and the distribution of statistically preferred values of a similarity measure, given a bit vector representation of a specific length, are directly related to the number of ties in proximities for a given data set. We explore the ties in proximity problem, using a number of chemical collections with varying degrees of diversity, given several common similarity measures and clustering algorithms. Our results are consistent with our probabilistic argument and show that this problem is significant for relatively small compound sets.

1. INTRODUCTION

The clustering of compounds based on chemical descriptors or chemical representations has had many applications in the pharmaceutical industry. In the mid-1980s, screening large corporate databases was a formidible task. Selecting subsets from vast amounts of compounds led to early work by Willett et al.,[1] in which clustering applications were utilized to automate screening subset selection from Pfizer's compound collection.

Pharmaceutical companies of today are finding themselves with larger and larger compound collections. Chemical warehouses are expanding owing to mergers, acquisitions, and the synthetic explosion brought about by combinatorial chemistry. The focus of the pharmaceutical industry in the 1990s has been on "diversity".[2] Companies have been diversifying their corporate databases through either compound acquisition from compound vendors or through proprietary synthesis of combinatorial libraries. In either case, large numbers of compounds need to be analyzed for either their internal diversity or the diversity in which they "add" to the current corporate compounds. Such decisions are commonly made through the use of clustering applications.[3−12] In the 1990s, high-throughput screening (HTS) has helped to move the bottleneck for lead discovery away from the screening arena and toward the analysis arena. The sheer volume of leads and the emphasis on automation has brought about the creation of lead discovery algorithms that often utilize cluster analysis techniques.[11,13−15]

Clustering applications of chemical structures require chemical representations that most often take the form of binary strings, e.g., MACCS keys,[16] Daylight fingerprints,[17] or BCI keys.[18] Such representations are suited for measure or metric comparisons, e.g., the Tanimoto coefficient or the Euclidean distance. With the ability to represent and compare chemical structures using such measures, clustering algorithms have flourished. With the recent work of Brown and Martin[19] and Wild and Blankley,[20] hierarchical applications such as Wards, complete-link, and group average have been shown to be, in varying degrees, at least reasonable at grouping chemically active molecules together. However, these same results are not entirely conclusive: it is still uncertain as to what combination of clustering algorithm, measure, representation, and rule to determine the number of clusters (e.g., level selection techniques in the case of hierarchical algorithms) is the best or most robust with respect to data dependencies. Some of this confusion and difficulty in determining a clustering methodology may in part be a result of the ties in proximity problem.

We approach this problem in two ways. First, we try to understand the ties in proximity problem from a probabilistic

* To whom all correspondence should be addressed. E-mail: mesaac@earthlink.net.Tel: (505) 983-3449.
[†] E-mail: nicolaou@bioreason.com.
[‡] E-mail: mesaac@earthlink.net.

10.1021/ci000069q CCC: $20.00 © 2001 American Chemical Society
Published on Web 12/02/2000

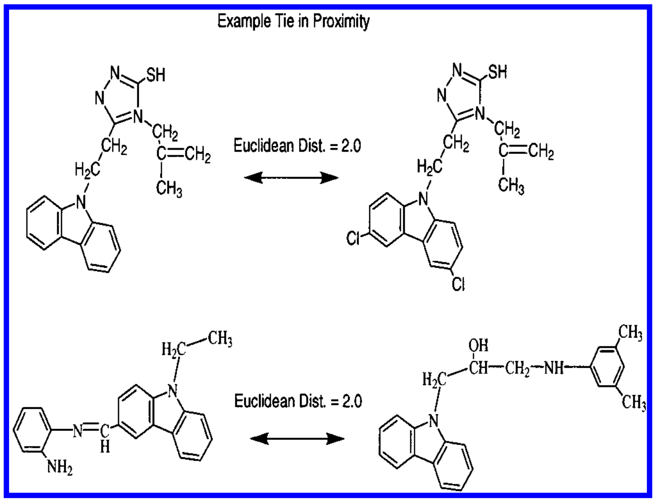

Figure 1. Example of a generic tie in proximity from the AsInEx data set, using the BR-MACCS key representation.

standpoint assuming random bit vectors in place of bit vector representations. And second, since the distribution of the bit vectors that represent any specific set of compounds is likely to be quite different from a set of random bit vectors, we need to empirically explore the impact of "ties in proximity" to understand the true nature of this problem with respect to clustering compounds. In section 2, we discuss the probabilistic argument and show an upper bound on the number of compounds that can be clustered without ties in proximity becoming a problem. This gives us a benchmark to compare our theoretical results with real sets of compounds. We show the total number of possible proximities of the Soergel measure (1 − the Tanimoto similarity), one of the most common measures of comparison used in clustering and database searching. In addition, we discuss how the Euclidean−Soergel product (a measure that helps remove the respective large and small compound biases of the Euclidean and Soergel measures[21]) increases the total number of possible proximities but not so much as to significantly increase the number of compounds that can be clustered without having a ties in proximity problem. In the last part of section 2, we show how the fractallike nature of the distribution of the Soergel proximities,[22] given all bit vectors of a given size or random subset of those same bit vectors, plays a role in determining the number of ties in proximities that occur. In section 3, we explain our experimental design and describe the data and methods used. In section 4, we present and discuss our results. Our conclusions and notes on future work follows in sections 5 and 6, respectively.

2. THE PROBLEM DESCRIPTION

In Figure 1, we show an example of a tie in proximity from the AsInEx data set (see section 3.3), using the BR-MACCS keys represention. The BR-MACCS is a slightly modified version of the public MACCS keys. A tie in proximity is simply a number of compounds or clusters of compounds that are equidistant from a compound or cluster in a given collection. This form of tie is not to be confused with a representation collision, as shown in Figure 2, which can occur where two or more compounds have the same representation. This is a encoding problem and has ramifications in clustering and database searching that are beyond the scope of this paper. In our empirical studies below, we remove the few compounds involved in representation ties from consideration so that they do not impact our results (see ref 23 for a lucid discussion of the problems of bit representations). In Figure 3, we show an example of an ambiguous tie: three compounds, where two are equidistant, or equisimilar to the third. This may or may not be a problem when clustering these same compounds in a larger data set, and it would depend on whether they are candidates for merging at a specific level of a hierarchy in a hierarchical clustering algorithm.

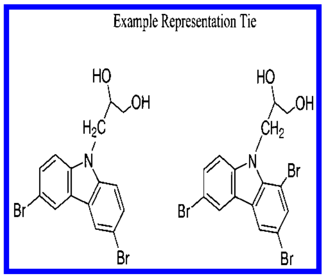

Figure 2. Representation tie from the AsInEx data set, using the BR-MACCS key representation.

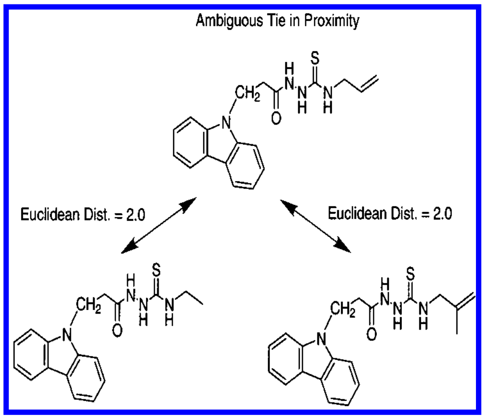

Figure 3. Example of an ambiguous tie in proximity from the AsInEx data set, using the BR-MACCS key representation.

The problem of ties in proximity with respect to hierarchical clustering algorithms is well-known,[24] but rarely considered as problematic if only a few ambiguities exist in the resulting hierarchy, and those ambiguities do not substantially change the structure of the hierarchy. However, if there are numerous ambiguities, then there are many possible resulting hierarchies and very likely some with considerably different structure from one another. Thus, determining where to cut a hierarchy, often done with level selection techniques to find the most effective or natural clusters, becomes difficult when the hierarchy was formed by making arbitrary decisions at each tie ambiguity found while generating the hierarchy. This can lead to inconsistent results where algorithms that are normally stable with respect to input order are no longer so, creating different hierarchies, which in turn provide different clusters using the same level selection technique.

In discrete settings, such as clustering compounds with binary representations, this problem is far more likely to occur. A mitigating factor in hierarchical algorithms is the rule by which clusters are merged: the chances of ambiguity may decrease or increase on the basis of whether the merging rule increases or decreases the number of possible measures compared to those in the original proximity matrix.

The set of feature descriptors and a dissimilarity measure define the feature space associated with a set of compounds. This space is not necessarily uniformly covered by an arbitrary set of compounds, however. The nature of the measure also defines the coarseness of the space. For instance, this coarseness can be seen in the substantial difference between the number of possible dissimilarities given the Euclidean distance and those given the Soergel measure. Thus, given a measure and a finite set of N binary features, the number of proximities can be described by a function of N. The number of possible proximities given the Euclidean distance is simply $N + 1$, which is equivalent to the total number of unique sums of bits up to N including 0. In the symmetric measure case, the number of pairwise proximities given C compounds is $C(C - 1)/2$. Using the Euclidean distance, there will be many ties in proximity if $C(C - 1)/2 \gg N$. Thus, even with 1024 features, only a small number of compounds can be clustered without there being a strong likelihood of ambiguous ties in the hierarchy.

We can frame the above discussion in terms of a famous problem in probability, known as the *Birthday Problem* (see ref 25). Namely, what is the probability that no two persons in a room with m people have the same birthday, given that there are 365 possible birthdays? It turns out that when $m = 23$ the probability is less than $^1/_2$ that there will be no ties. In our case, the compound proximities represent each person's birthday and the total possible proximities represent the total number of days in a year. For example, using a purely probabilistic argument, if the total number of proximities equaled 365 (e.g., a binary representation of length 364, using the Euclidean distance) and there were 23 compound proximities, we would have a probability less than $^1/_2$ that there would be no ties. This is very few compounds (between 7 and 8).

As we add more compound proximities, the probability of ties grows very quickly. Using the example above, let us say that we would like to increase the number of compound proximities, and therefore the number of compounds, by allowing a higher probability of ties. With the probability of less than $^1/_e$ (slightly less than $^2/_3$) that there will be no ties (an almost $^2/_3$ probability that there will be ties), we can add roughly just 6 more proximities ($m = 29$ total), or 1-2 more compounds. So, in this example as the number of compounds increases significantly beyond 9 or so, the number of ties in proximity also increases dramatically. Again, from a purely probabilistic point of view, if the number of proximities to be considered equals the total possible number of proximities, it can be shown that the expected number of ties in proximity is equal to the total number of proximities over e. This means that somewhere between $^1/_3$ and $^1/_2$ of the proximities will be ties.

2.1. Probabilistic Argument. We can take advantage of some simple probability results concerning the Birthday Problem to answer the following questions:

a. Assuming that the compound representations are randomly chosen bit vectors of a fixed length, how many compounds can one effectively cluster given a specific similarity measure, such that the probability of having many ties in proximity is low, so that the chance of ambiguous ties is largely nonexistent?

b. Using the assumptions above, how does one know when there are far too many compounds, such that there will be a great many ties in proximity and hence a strong likelihood of a significant number of ambiguous ties?

If we let $C =$ the number of compounds, $N =$ the number of bits in the representation, $n =$ the total number of possible proximities, given the similarity measure, and $m =$ the total number of proximities, given C compounds (or $C(C - 1)/2 = m$), then

1. With probability $^1/_e$ there will be no ties in proximity if

$$m = \lceil \sqrt{2n} + 1 \rceil$$

2. If $m = n$, and as n grows large, the expected number of ties in the proximity approaches

$$\frac{n}{e}$$

Equations 1 and 2 help us approach answers to questions a and b, respectively. To help get a feeling for question a, note that to obtain a greater than $^1/_e$ probability that there will be no ties in the proximity matrix

$$C \cong 2(n^{1/4})$$

Namely, for $m = \lceil \sqrt{2n} + 1 \rceil$ and $C(C - 1)/2 = m$, it follows that $C \cong 2(n^{1/4})$, under the assumption that any of the n possible similarities are equiprobable. Thus, even though n may be very large, say in the millions, its one-fourth root—and hence the number of compounds C—may not be more than a few hundred. From eq 2 we will almost surely have many ambiguous ties if

$$C \cong \sqrt{2n}$$

As will be noted below, this is not particularly large number of compounds given the various proximity measures or the typical size of the bit vectors used in this connection.

2.2. Calculating the Total Possible Number of Proximities. 2.2.1. Soergel or Tanimoto Proximities. The Soergel measure defined for bit vectors is simply one minus the Tanimoto similarity. For simplicity we can count the number of possible Tanimoto proximities, given bit vectors of a fixed length. The Tanimoto similarity for bit vectors is defined as $a/(a + b + c)$, where $a + b + c \leq N$, a is the number of bits turned on in both vectors, b is the number of bits turned on in one vector and not the other, and c is the same as b but for the other vector.[26] Thus, we can count the number of proximities for the Tanimoto similarity measure by counting the number of irreducible fractions whose denominators do not exceed N, and whose numerators are always less than N. We can calculate this number from a well-known theorem in number theory:[27,28] the sum of the Farey series N is the number of reduced fractions in the range from 0 to 1 whose denominators do not exceed N. This turns out to be the sum of the Euler φ function, from $k = 1, 2, ..., N$, or

$$\Phi(N) = \sum_{1 \leq k \leq N} \varphi(k) = \frac{3}{\pi^2} N^2 + O(N \log N)$$

Note that the $N \log N$ term is dominated by the $(3/\pi^2)N^2$ term as N grows large. To get a sense as to how well this function approximates the actual number of proximities,[27]

Table 1. Bounding the Number of Compounds *C* by the Number of Possible Proximities *n* Given the Length of the Bit Vector *N* and the Measure Used

measure	N	total proximities (n)	$2(n^{1/4}) > C$
Euclidean	166	167	7
Euclidean	1024	1025	11
Euclidean	2048	2049	13
Euclidean	4096	4097	16
Soergel	166	9600	20
Soergel	1024	328969	48
Soergel	2048	1297444	67
Soergel	4096	5148814	95

if $N = 100$, the actual sum equals 3043, whereas $(3/\pi^2)N^2$ equals 3039.64. In the case of $N = 500$, the actual sum equals 76 115 and $(3/\pi^2)N^2$ equals 75 990.89. A bit vector of length 1024 would have approximately 329 000 possible proximities given the Soergel measure.

2.2.2. Euclidean−Soergel Product. In ref 21, the authors show that, for reasons that have to do with biases associated with compound size in either the Euclidean or Soergel measure, a possibly better proximity measure used in diversity and clustering than either the Euclidean distance or Soergel measure is the Euclidean−Soergel product. This increases the number of possible proximities of the Soergel by only a small factor. One cannot simply multiply the total possible number of Euclidean distances, $N + 1$, by the total possible number of Soergel measures, $\Phi(N)$. The factor is much smaller than $N + 1$. There are several reasons why a factor of $N + 1$ would be overcounting. First, the Euclidean distances are integral. Thus, the product contains some of the same irreducible fractions that would be found in the Soergel measure and a few additional ones. Second, only certain Euclidean distances, not necessarily all, can be associated with a specific Soergel measure. For example, all possible combinations of bit vectors that generate the Soergel measure of $^1/_3$ do not necessarily produce all possble Euclidean distances, $0-N$. Thus, all combinations of bit vectors only produce a limited subset of $N + 1$ times the number of Soergel proximities. Though experiments show that this is an increasing function of N, very few additional fractions are added, given the likely sizes of bit vectors used as compound representations. A proof of a bound for this function akin to the bound on the sum of the Farey series is unknown.

2.2.3. Examples. In Table 1, we show how many compounds (the last column) could be clustered with very little chance that there would be ambiguous ties in clustering those compounds, since in general there would either be no or just a few ties in proximity. This is under the assumption that the bit vectors representing the compounds are equivalent to random vectors and the distribution of the proximities is drawn from a uniform random distribution. Both of these assumptions are unrealistic, and the latter can be seen in Figures 4−6. In these figures we show the distributions of all possible proximities for three measures given bit vectors of length 10. They are not uniform random distributions, but rather have subsets of statistically preferred values. In the case of the Euclidean distance in Figure 4, the distribution is Gaussian-like and the preferred values are centered at the mean of the midpoint of the distances. The Soergel and Euclidean−Soergel in Figures 5 and 6 have statistically preferred values at numerous peaks in fractallike distributions. It is important to note that the character of the distributions of the three measures and real data remains the largely the same. Thus, the results of Table 1 can only give us a very rough benchmark by which we can assess the likelihood of obtaining ambiguous ties.

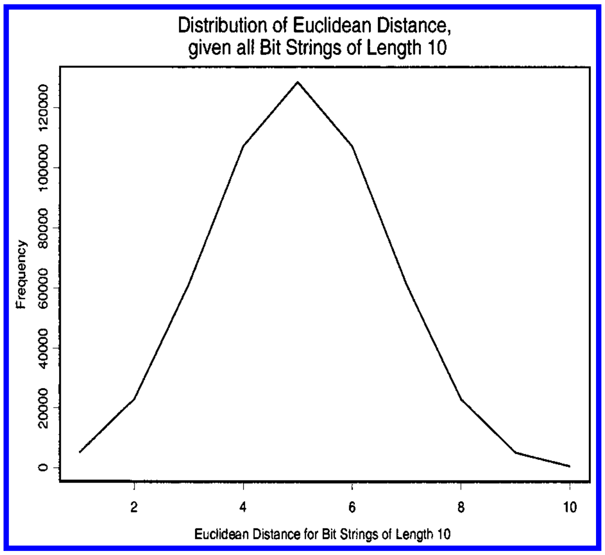

Figure 4. Distribution of Euclidean distances given all pairwise combinations of bit vectors of length 10.

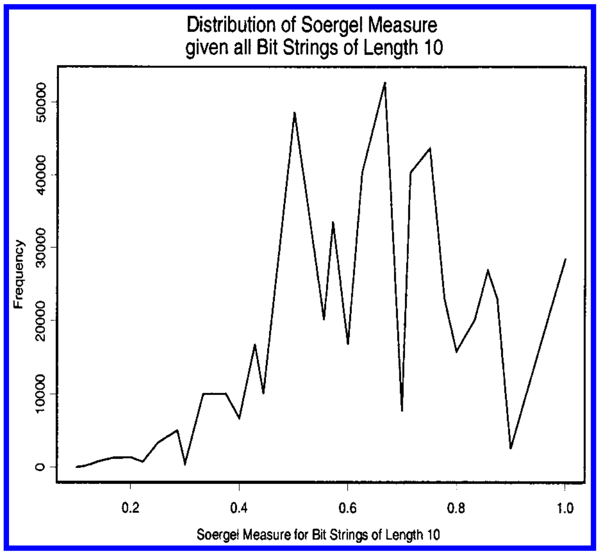

Figure 5. Distribution of Soergel proximities given all pairwise combinations of bit vectors of length 10.

As mentioned above the distributions of possible values is not uniform. In fact, the distribution of the Soergel proximities has very definite statistically preferred values. The low-order relative primes (i.e., 1/3, 1/2, 1/4, 1/5, 2/5, etc.) are preferred over relative primes closer to N.[22] This property in effect lowers the number of compounds that can be clustered with a certain measure and bit vector length, since there is a higher probability that the proximities will be drawn from the statistically preferred values.

2.2.4. Statistically Preferred Values for the Tanimoto. Godden et al.[22] showed how to obtain the statistically preferred values of the Tanimoto coefficient on bit vectors of a given length using a different but equivalent to the $a/(a + b + c)$ formulation of the Tanimoto measure. Here for simplicity and in the context of irreducible fractions, we use the $a/(a + b + c)$ definition of the Tanimoto measure to

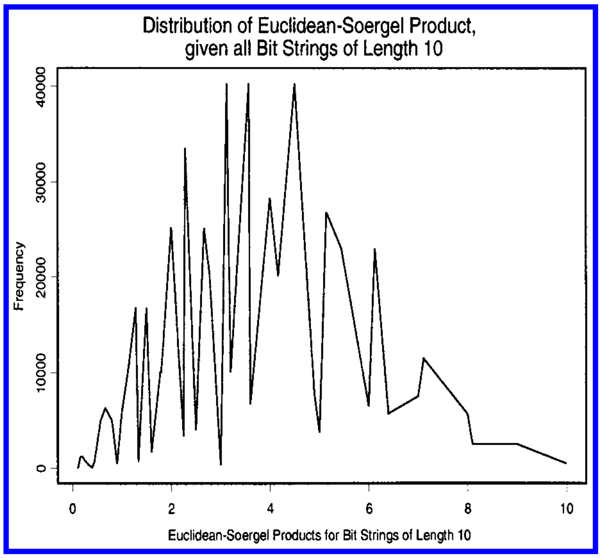

Figure 6. Distribution of Euclidean Soergel product proximities given all pairwise combinations of bit vectors of length 10.

show how statistically preferred values are obtained and how they can substantially differ in frequency from those that are not preferred.

To get a sense of how the statistically preferred values can occur, we start by showing the numeration of a specific Tanimoto fraction, given N, a, and $b + c$. This amounts to the sum of N choose $a + k$, from $k = 0$ to $b + c$:

$$\sum_{k=0}^{b+c} \binom{N}{a+k}$$

Thus, the probability of Tanimoto reduced fraction R is

$$P(R) = \frac{\sum_{i=1}^{\lfloor N/(a+b+c) \rfloor} \sum_{j=0}^{i(b+c)} \binom{N}{ia+j}}{2^{N-1}(2^N + 1)}$$

where a is the numerator of the reduced fraction. The double sum generates all of the bit vectors that can be used in creating a fraction that is reduced or can be reduced to R. It can be easily seen that the reduced Tanimoto fraction, where $a = N - 1$ and $a + b + c = N$, has a probability far less than where the reduced fraction is composed of $a = 1$ and $a + b + c = 2$. With statistically preferred values, $C \ll O(n^{1/4})$. In fact, empirical evidence suggests that most of the fractions have a very low probability, and as N increases, we conjecture that $C = O(n^{1/4}/\log N)$. This means that a large increase in the length of the bit vector has only a marginal reduction in the number of ties in proximity.

3. EXPERIMENTAL DESIGN

There are many degrees of freedom in which the ties in proximity problem can be explored. We chose our degrees of freedom from the most prevalent methods in use today, such as common key sets, bit vector length, measures, representative sizes of data sets used in clustering, and range of diversity of the data sets. Exploring all of the combinations of the above tools in a more formal design of experiments setting would be far beyond the scope of this paper. Thus, we selected certain combinations to help reveal to the reader at least some of the properties of the ties in proximity problem.

3.1. Chemical Representations. We picked two common bit vector representation systems for our experiments: the Daylight fingerprints (a graph and unique path hashed fingerprint) and the BR-MACCS keys (a structural key based representation). BR-MACCS keys are a slightly modified and a bit smaller subset (157 versus 166) than the original public MACCS keys. The Daylight fingerprints we used are unfolded with a length of 1024 bits. This gives us a relatively large bit vector in the Daylight fingerprint and a much smaller bit vector in the BR-MACCS keys, bordering on the mini-fingerprint size[29] to explore the bit vector length in our experiments.

3.2. Measures. We chose the Euclidean and Soergel measures largely because of their ubiquity within the chemoinformatics clustering literature, and the Euclidean–Soergel product because this is a newer and largely unfamiliar measure with hopefully improved properties over the Euclidean and Soergel measures. All of the measures have different distributions of statistically preferred values, and this gives us a chance to explore these properties as well.

Precision can be an important factor in studies of this type if the proximity values are stored in a file. For instance, it takes at least six digits of precision to ensure that we can discriminate between all possible proximity values when using a Daylight fingerprint of length 1024 and the Soergel measure.

3.3. Data Sets. We chose the following publicly available data sets on the basis of structural diversity and size. We wanted a range of focus, from very focused to very diverse. Our size was determined by a small range of what would be considered a common number of compounds to cluster, to show that within this range this problem exists and can grow as the number of compounds grows.

3.3.1. Data Set Size. The smallest and most focused data set is the AsInEx dataset. It is a subset of 289 carbazole-containing compounds from the AsInEx Organic Chemical Catalog.[30] These compounds were obtained through a substructure search on the full AsInEx Organic Chemical Catalog of more than 100 000 compounds. Our next set of compounds is the NCI data set, comprised of 675 compounds that have been shown to have activity against the AIDS virus.[31] The data set is of intermediate size and can be characterized as being semifocused compared to the other two data sets. Our largest and most diverse set of compounds is the Interbioscreen data set, a random subset of 1049 compounds taken from the Natural Products Catalog of Screening Compounds, which contains more than 10 000 compounds.[32] From each of these data sets, we removed a small fraction of compounds (anywhere from 8% to less than 1% of the compounds per data set) that were involved in representation ties, so that each compound in the sample set has a unique representation. We deemed that the slight difference in size and composition between data set samples was insignificant with respect to the number of ties in proximity and ambiguous ties that we would find. In Table 3 we show the number and type of representation ties found in each data set, given either the Daylight fingerprint representation or BR-MACCS keys representation.

Table 2. Number of and Type of Representation Ties

representation type	AsInEx	NCI-AIDS	Interbioscreen
BR-MACCS	10, 2-way ties	35, 2-way; 5, 3-way; 3, 4-way ties	9, 2-way ties
Daylight fingerprints	3, 2-way ties	23, 2-way; 4, 3-way; 1, 4-way ties	6, 2-way ties

Table 3. Ambiguous Ties in Wards Clustering of Data Sets Using BR-MACCS Keys

measure	AsInEx	NCI-AIDS	Interbioscreen
Euclidean	46, 2-way α-ties	28, 2-way α-ties	46, 2-way α-ties; 3, 3-way α-ties
Soergel	20, 2-way α-ties	5, 2-way α-ties	25, 2-way α-ties

3.3.2. The Number and Type of Representation Ties. In Table 2 we show the number of representation ties and their composition. From these we selected a subset of compounds to omit from each data set per representation. In the case of the AsInEx and Interbioscreen data sets we simply removed all of the compounds involved in representation ties, whereas in the NCI-AIDS data set we removed only so many compounds as to avoid representation ties, leaving in one compound per representation. Both key sets have representation tie problems for all three data sets; the Daylight fingerprints have fewer primarily as a result of having a longer bit vector to represent the compounds. As a result of its diverse nature, the Interbioscreen has relatively very few representation ties despite having a larger number of compounds than the other two data sets.

3.4. Algorithms. Wards is considered one of the best and most commonly used hierarchical clustering algorithms in the chemoinformatics literature (e.g., refs 12, 19, 20, and 33). However, other hierarchical algorithms have been applied such as complete-link[34] and group average[19] with varying success. We selected the latter two as well on the basis of the different merging criterion that each has in contrast to Wards. This was meant to give us a sense of how the ties problem occurs with different merging criteria. The merging criterion for complete-link clustering only uses those values expressed by the measure used and the compound set. The merging criteria of Wards (a sum of squares calculation) and group average (an averaging of dissimilarities calculation) clustering, however, can create ancillary numbers that are a slightly larger superset of numbers than those generated by the measure and compound set. Since these numbers are formed from the finite set generated by the measure, with a distribution of preferred values dependent on the compound set, they in turn are finite and have a related distribution of preferred values.

3.5. The Design. Each experiment has its own independent focus so that we can compare results within each experiment. At the same time we can compare across the first three experiments, since they have two representations, two measures, and one of the data sets in common. The last experiment is meant to explore the Euclidean−Soergel product with all three algorithms used, while keeping a common data set used in all four experiments, so that we can compare these results across the other experiments.

1. *One clustering algorithm, two representations, two common measures, and three data sets:* In this experiment, we use just one clustering algorithm, Wards. The two common bit vector representations are used since they have considerably different lengths. The two common measures are used, where the number of possible distances and the distribution of possible statistically preferred values differ. Last, we use all three data sets (AsInEx, NCI-AIDS, and Interbioscreen) to identify the scope of the ties problem in terms of data set size and the data set diversity.

2. *One clustering algorithm, two representations, two common measures, and one data set:* This experiment's focus is on the ties problem and complete-link clustering with the NCI data set.

3. *One clustering algorithm, two representations, two common measures, and one data set:* Here we try to understand the ties problem with the group average clustering algorithm and the NCI data set.

4. *Three algorithms, two representations, one (Euclidean−Soergel) measure, and one data set:* All three of the above algorithms are explored with the Euclidean−Soergel measure and the NCI data set.

4. EMPIRICAL RESULTS AND DISCUSSION

For each measure and each data set we generate a proximity matrix that contain the pairwise proximities that will be used by each of the algorithms. We can plot the distribution of the *n* choose 2 proximities and observe how the statistically preferred values dominate the distribution. Since there is a large number of equivalent values (ties in proximity), we can see that there is a strong chance of obtaining ambiguous ties.

4.1. Distributions of Ties in Each Proximity Matrix. 4.1.1. Distributions with BR-MACCS Keys. The distributions of proximities with BR-MACCS keys are more granular but otherwise have much the same character as those distributions generated via the Daylight fingerprints. We show and describe the distributions of the proximities with BR-MACCS keys in this subsection and the distributions with the Daylight fingerprints in the next subsection. We have squared the square root values of the Euclidean distance to show the distributions in terms of the number of bits in the BR-MACCS keys bit vector. We have not done this for the Euclidean or Euclidean−Soergel distances generated via the Daylight fingerprints. These values are normalized between 0 and 1.

The distribution of the distances in the proximity matrix is highly skewed to the left for the AsInEx data set—comprising only about a third of the possible proximity values—owing to the focused nature of the data set as can be seen in Figure 7. Some values occur greater than 1500 times, and most of the values fall between 10 and 50, where the greatest possible distance is 157. Thus, given the 36 046 pairwise proximities from 269 compounds in this data set, almost all of the proximities are drawn from the 40 or so values between 10 and 50, and only 60 proximities in all of the 158 are used. We would expect there to be many chances for ambiguous ties with this proximity matrix.

Given the AsInEx data with the BR-MACCS keys and the Soergel measure, the focused data set has a highly skewed distribution with no values above 0.5. Note how in Figure 8 this distribution has a probabilistic fractallike nature in that it has the properties of scale invariance and self-similarity. Again, only a small fraction of the roughly 7700 possible

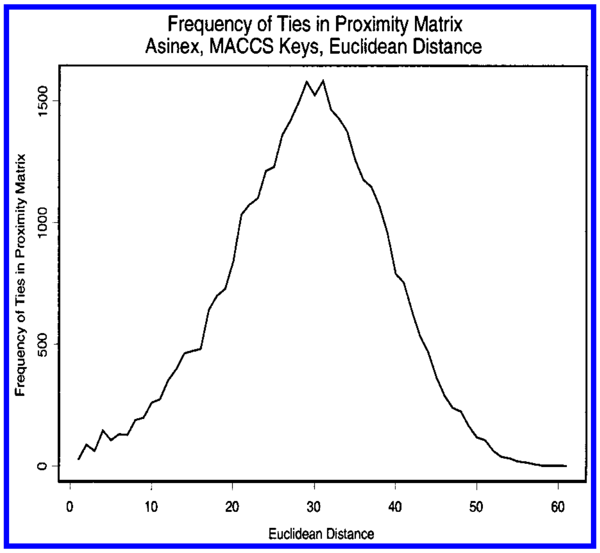

Figure 7. Distribution of proximities in the proximity matrix of AsInEx data, given the BR-MACCS keys and Euclidean distance.

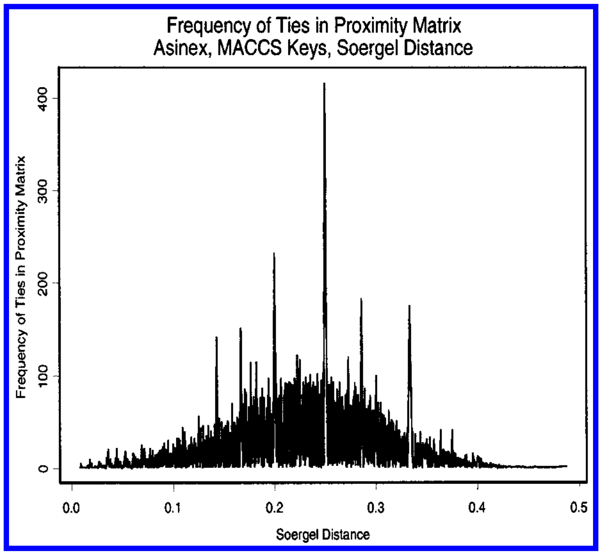

Figure 8. Distribution of proximities in the proximity matrix of AsInEx data, given the BR-MACCS keys and Soergel distance.

proximities are used, so a greal deal of the 36 046 pairwise proximities will have ties. Many of the statistically preferred values have more than 50 occurrences making it likely that there will be some chance for ambiguous ties.

In Figure 9 we show the AsInEx data set with the BR-MACCS keys and the Euclidean−Soergel distance. The distribution displays many statistically preferred values with a large frequency that we believe plays a role in our empirical results. Note how almost all values have many occurrences and the values are only in the interval from 1.0 to 35.0. This is roughly one-fifth of the possible range. It is hard to know exactly how many possible proximities there are given the BR-MACCS key length of 157. Our guess is that it is some small multiple of the number of possible 7700 Soergel proximities, perhaps between 20 000 and 100 000 proximities. Nevertheless, these seem to be dominated by statistically preferred values, even more so than the Soergel proximities.

In Figure 10, we can see the same Gaussian-like distribution as in the AsInEx distribution in Figure 7. The majority of the proximities are larger in general as would be expected from a less focused data set. There are some proximities with nearly 7000 ties. We would expect that this data set with this measure will have a strong likelihood of ambiguous ties. Only 94 proximities are used of the 158, although there are nearly 200 000 pairwise proximities.

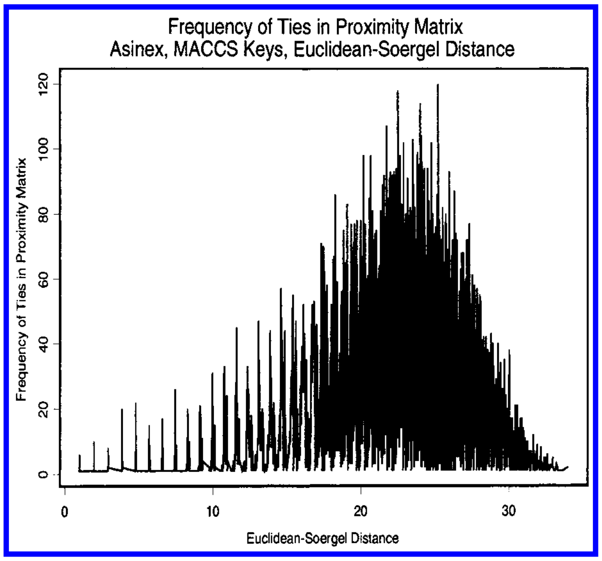

Figure 9. Distribution of proximities in the proximity matrix of AsInEx data, given the BR-MACCS keys and Euclidean−Soergel product.

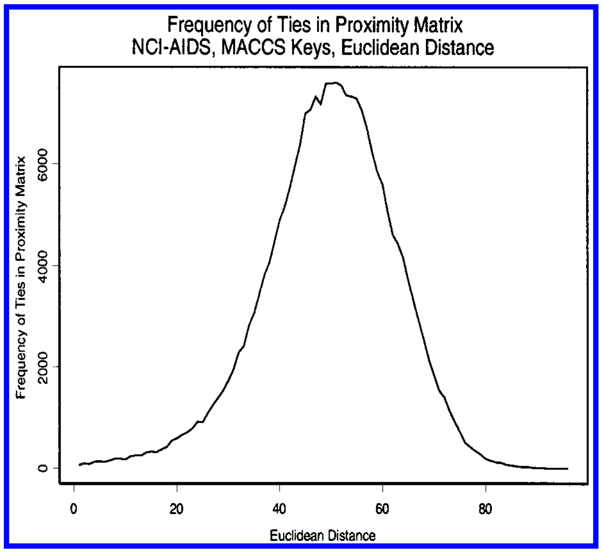

Figure 10. Distribution of proximities in the proximity matrix of NCI-AIDS data, given the BR-MACCS keys and Euclidean distance.

In Figure 11, the distribution, given the Soergel measure has longer tails that the distribution of the AsInEx data with the same measure. Again, this is due to the less focused nature of the data set. Nevertheless there are still the characteristic spikes that denote the prevalence of statistically preferred values. There were 2924 proximities used out of the roughly 7700 possible proximities.

The distribution of proximities for the Interbioscreen data with BR-MACCS keys seen in Figure 12, given the Euclidean distance, is very similar to the proximity distribution for the NCI-AIDS data set. With some of the proximities containing over 8000 ties, it is clear that we would expect some number of ambiguous ties generated by this data set. There were 88 proximities used of the 158 possible.

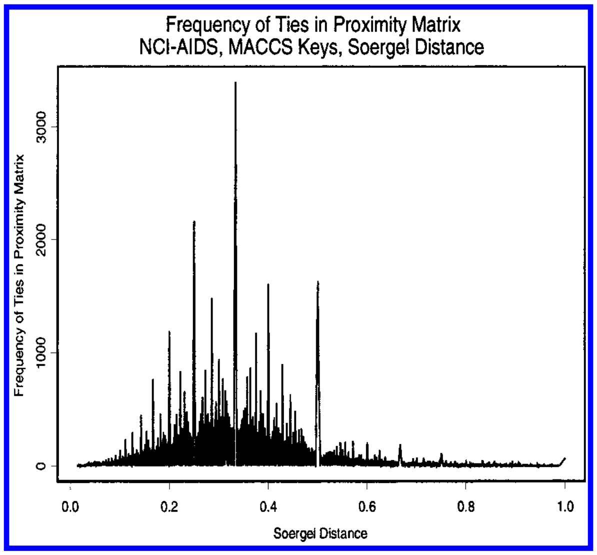

Figure 11. Distribution of proximities in the proximity matrix of NCI-AIDS data, given the BR-MACCS keys and Soergel distance.

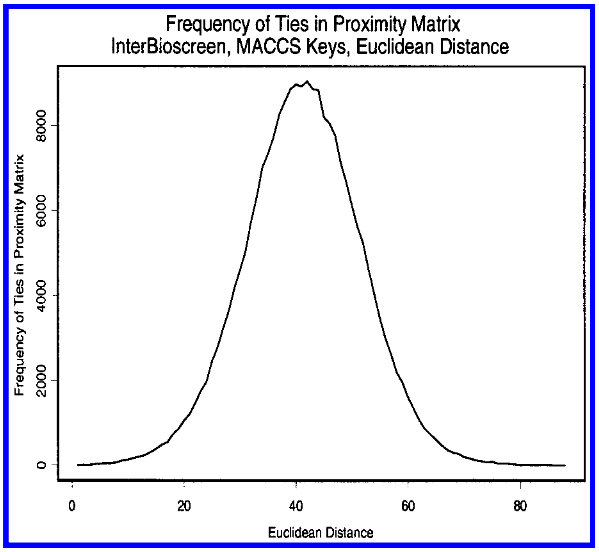

Figure 12. Distribution of proximities in the proximity matrix of Interbioscreen data, given the BR-MACCS keys and Euclidean distance.

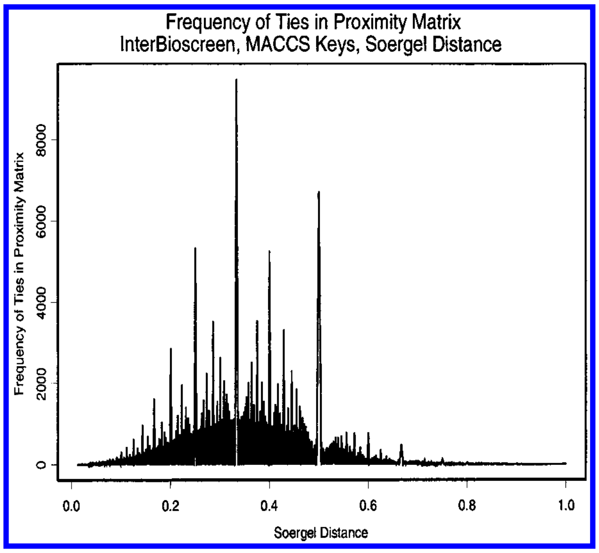

Figure 13. Distribution of proximities in the proximity matrix of Interbioscreen data, given the BR-MACCS keys and Soergel distance.

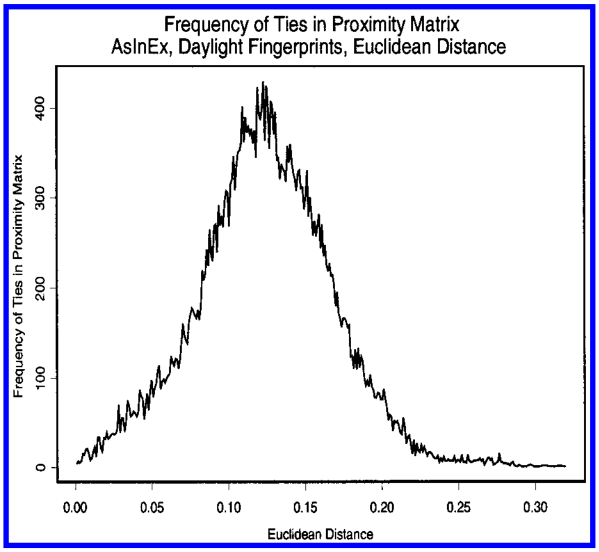

Figure 14. Distribution of proximities in the proximity matrix of AsInEx data, given the Daylight fingerprints and Euclidean distance.

In Figure 13, the distribution of proximities with the Soergel measure and the Interbioscreen data has even more values above 0.5 than the NCI-AIDS distribution. There are many statistically preferred values with greater than 2000 ties, even below the common 0.3 cutoff. There are even fewer proximities than the NCI-AIDS data with the same measure, 2480, although the total number of proximities in the matrix is roughly 500 000. We can guess that this data set with the measure and BR-MACCS keys will produce a significant number of ambiguous ties.

4.1.2. Distributions with Daylight Fingerprints. In general the distributions, given the Daylight fingerprints, are more fine-grained than those found with the BR-MACCS keys, but they have much the same character per measure.

The proximity distribution in Figure 14 of the AsInEx data with Daylight fingerprints and the Euclidean distance has very much the same type of distribution of proximities as with the BR-MACCS keys shown in Figure 7. However, only a few proximities have just over 400 ties at the peak of the distribution. This is far smaller than the frequency of ties that can be found with the BR-MACCS. The Gaussian-like distribution is again highly skewed to the left as would be expected by a focused data set. There are 307 proximities used out of a total of 1025. This is a slightly smaller ratio of proximities used than for the BR-MACCS keys of the same data set and measure.

In Figure 15, note how the distribution of proximities given Daylight fingerprints and the Soergel measure is not nearly as skewed to the left as it is when this data set and measure are used with BR-MACCS keys, as shown in Figure 8. The largest statistically preferred value in the former latter figure is 0.25, whereas in Figure 15 it is 0.5. This may be a byproduct of the way Daylight fingerprints are generated. Of the approximately 329 000 possible proximities, 10 784 are used.

The distribution of the proximities in Figure 16 is skewed further left by use of the Euclidean−Soergel measure. Even though there are more proximities in this distribution, 12 927, there is a dramatic increase in the statistically preferred values. The ratio of proximities used over the possible

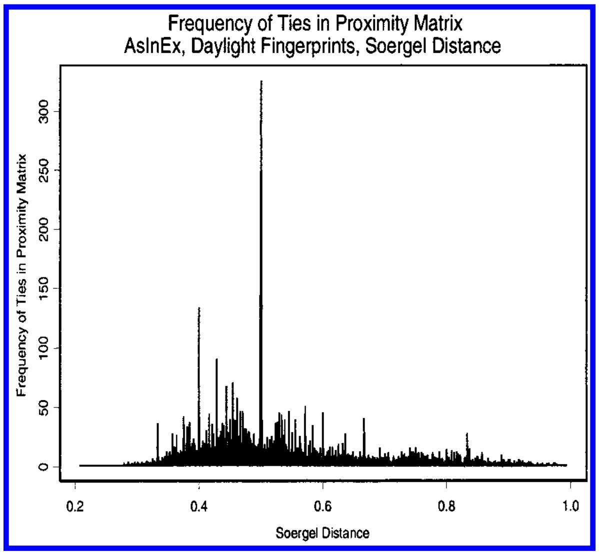

Figure 15. Distribution of proximities in proximity matrix of AsInEx data, given Daylight fingerprints and Soergel distance.

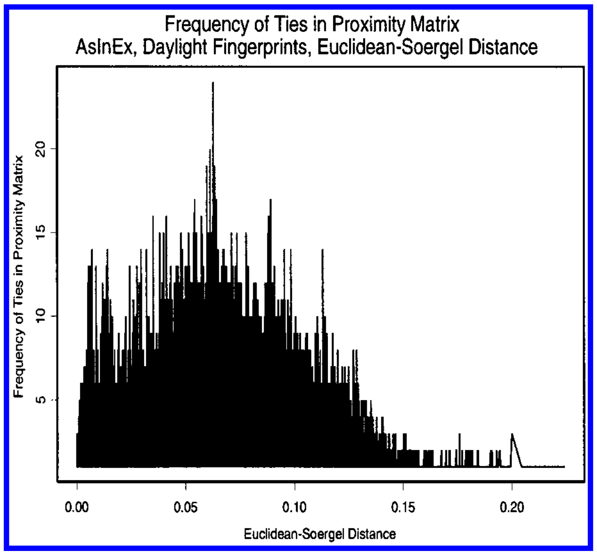

Figure 16. Distribution of proximities in the proximity matrix of AsInEx data, given the Daylight fingerprints and Euclidean−Soergel product.

number of proximities is smaller, since the total number possible is a great deal larger than for that of the Soergel measure alone.

In Figure 17, we can see the preferred values of the Euclidean distance given the NCI data and Daylight fingerprints. The distances have been normalized to between 0 and 1. Note how the distribution has been skewed to below 0.5. Only a little more than half of the 1025 possible proximity values are used, and the vast majority of those are between 0.2 and 0.4, representing roughly 200 values, some of which occur as many as 1200 times. Since there are 641 compounds, there are 205 120 pairwise proximities—almost all of which are involved in ties.

In Figure 18, we show the frequency of ties in proximity in the proximity matrix of the NCI-AIDS data, using Daylight fingerprints and the Soergel measure. The sharp spikes of the statistically preferred values of the Soergel distances can be seen quite clearly, some of which occur over 600 times. The values are spread across the entire 0 to 1 proximity range, but they are only a small fraction of the approximately 329 000 possible Soergel measures that can be obtained from a bit vector with a length of 1024. This is why there are so many ties in proximity even though the number of pairwise proximities given this data set is 205 120.

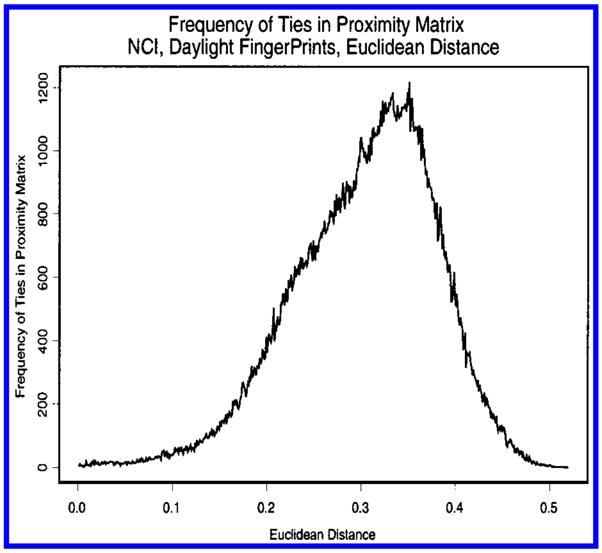

Figure 17. Distribution of proximities in the proximity matrix of NCI-AIDS data, given the Daylight fingerprints and Euclidean distance.

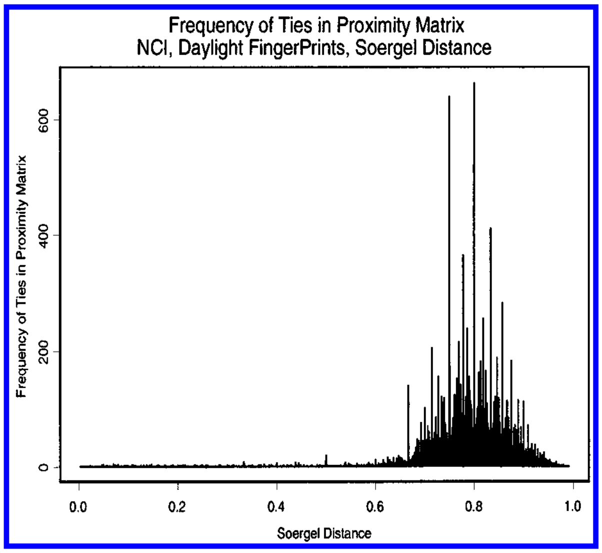

Figure 18. Distribution of proximities in the proximity matrix of NCI-AIDS data, given the Daylight fingerprints and Soergel distance.

In the Interbioscreen data proximity distribution shown in Figure 19, the Euclidean distances have been normalized between 0 and 1 and the distribution of the proximity matrix distances looks very much like the NCI Euclidean proximity matrix distribution in Figure 17. There is a 200 proximity value range where the values have more than 1000 occurrences, so we might expect a few ambiguous ties. However, the number of proximities used is 503, nearly half of all the proximities. This is a higher ratio than for the other data sets.

In Figure 20, note how similar the distribution is of the Interbioscreen data with Daylight fingerprints and the Soergel measure to the NCI Soergel measure with Daylight fingerprints shown in Figure 18. Many of the statistically preferred values are identical. Both of these sets are far more diverse

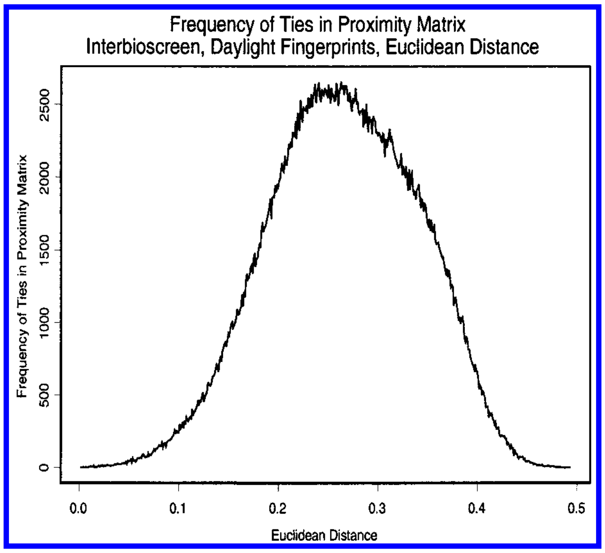

Figure 19. Distribution of proximities in the proximity matrix of Interbioscreen data, given the Daylight fingerprints and Euclidean distance.

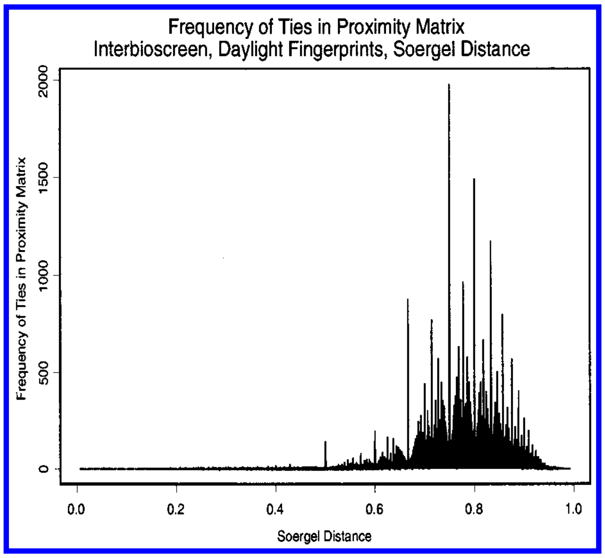

Figure 20. Distribution of proximities in the proximity matrix of Interbioscreen data, given the Daylight fingerprints and Soergel distance.

than the AsInEx data set, but the distributions show that there still is some chance for ambiguous ties given that there are so many ties in proximity. The number of proximities used is quite high in this case, with 41 559 used out of the approximately 329 000 possible.

Often a threshold is chosen to select only those compounds that are the most similar. For instance, the preferred value for the Tanimoto similarity is 0.7, 0.3 being the equivalent Soergel value. In Figure 21, we show that portion of the distribution including and below the 0.3 Soergel threshold value. There are still many ties in proximity relative to all of the proximities generated below the threshold. Equivalent thresholds used with either the Euclidean or Euclidean−Soergel measures would also show many ties in proximity.

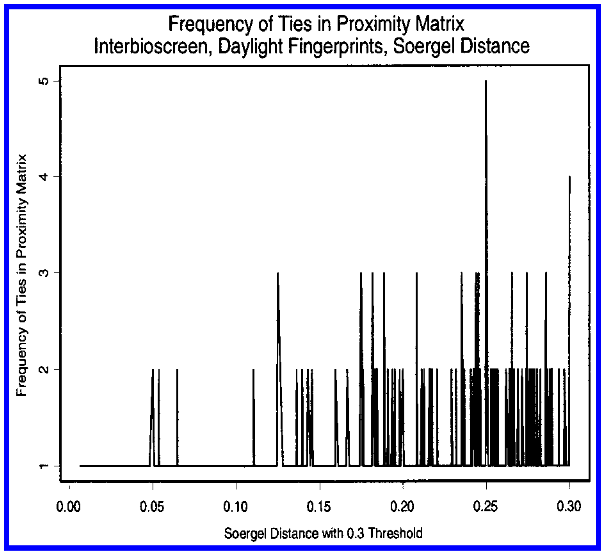

Figure 21. Distribution of proximities below 0.3 threshold in the proximity matrix of Interbioscreen data, given the Daylight fingerprints and Soergel distance.

Table 4. Ambiguous Ties in Wards Clustering of Data Sets Using Daylight Fingerprints

measure	AsInEx	NCI-AIDS	Interbioscreen
Euclidean	4, 2-way α-ties	5, 2-way α-ties	9, 2-way α-ties
Soergel	2, 2-way α-ties	1, 2-way α-tie	1, 2-way α-tie

4.2. Wards Clustering with Both Fingerprints, Euclidean and Soergel Measures, and All Three Data Sets. In Table 3, we show the number of ambiguous ties using Wards clustering algorithm on the three data sets, given the two measures. With Wards clustering, there are a great many ambiguous ties using either the Euclidean or Soergel measures for the AsInEx data and BR-MACCS keys. This is most likley due to the very focused nature of the AsInEx data and the small number of bits of the BR-MACCS keys relative to the number of compounds. Even with the use of the Soergel measure there would be in the neighborhood of 1 000 000 possible hierarchies, and these would most certainly contain substantially different hierarchies. Curiously, the NCI-AIDS data set had fewer ambiguous ties than the AsInEx data set despite having more than twice the number of compounds. This is likely due to the more diverse nature of the NCI-AIDS data set. Note that in the NCI-AIDS data set with the Soergel measure, the number of possible hierarchies is only 32, many of which happen to be inconsequential, namely, simple transpositions in the hierarchy. On the other hand, using the NCI-AIDS data we did find substantial number of ambiguous ties using the Euclidean distance, and these could be found throughout the hierarchy, independent of the level. The Interbioscreen data set is also diverse, but there are nearly double the number of compounds as are in the NCI-AIDS data set. With the increase in the number of compounds, the number of ambiguous ties also increases substantially—nearly 2 to 1 in the Euclidean case, and 5 to 1 when using the Soergel measure.

The use of Wards with Daylight fingerprints changes the results substantially. Only the Interbioscreen data with the Euclidean distance has enough ambiguous ties to make the number of hierarchies worrisome. In Table 4 one finds that the length of the Daylight fingerprint helps to prevent ambiguous ties from becoming a significant problem when using up to roughly 1000 compounds as long as the Soergel measure is used with Wards clustering. Note how with the Daylight fingerprints the level of diversity of the data set

Table 5. Ambiguous Ties in Complete-Link Clustering of the NCI Data Set

measure	BR-MACCS	Daylight
Euclidean	98, 2-way α-ties; 9, 3-way α-ties; 1, 4-way α-tie	19, 2-way α-ties
Soergel	14, 2-way α-ties	3, 2-way α-ties

Table 6. Ambiguous Ties in Group Average Clustering of the NCI Data Set

measure	BR-MACCS	Daylight
Euclidean	31, 2-way α-ties; 1, 3-way α-ties	7, 2-way α-ties
Soergel	6, 2-way α-ties	1, 2-way α-ties

Table 7. Ambiguous Ties in Wards, Complete-Link, and Group Average Clustering, Using the NCI Data Set and the Euclidean−Soergel Product Measure

representation	Wards	complete-link	group average
BR-MACCS	6, 2-way α-ties	15, 2-way α-ties	7, 2-way α-ties; 3, 3-way α-ties
Daylight	1, 2-way α-ties	3, 2-way α-ties	1, 2-way α-ties

seems to play no real role in the number of ties. It is hard to say whether this is an artifact of Daylight fingerprints in general.

4.3. Complete-Link Clustering with Both Fingerprints, Euclidean and Soergel Measures, and the NCI Data Set. Complete-link clustering clearly has significant problems with ambiguous ties, especially with Euclidean distance shown in Table 5. Even the use of the Soergel and Daylight keys contains a few ambiguous ties. Larger data sets would likely have even more ambiguous ties. As noted above, the complete-link algorithm uses no ancillary values to determine the merging of clusters within the hierarchy and thus is confined to using just those values from the distribution of pairwise proximities. These results are a good indication of this fact.

4.4. Group Average Clustering with Both Fingerprints, Euclidean and Soergel Measures, and the NCI Data Set. Group average clustering performs more like Wards clustering and has only slightly more ties in general, regardless of the measure or the representation used shown in Table 6. Both of these clustering algorithms compute new measures that decrease the likelihood of ambiguous ties in general. Nevertheless, since the computation of those measures is based on the discrete nature of the original proximities, they in turn are a finite set of numbers that only augment the original set of proximities. The BR-MACCS keys perform substantially worse in terms of the number of ambiguous ties than the Daylight keys with this algorithm and data set as they do with the other algorithms and data sets. The number of ambiguous ties is close to one-third that of complete-link, whether in terms of the measure or the keys.

4.5. All Three Algorithms with Both Fingerprints, the Euclidean−Soergel Product, and the NCI Data Set. Remarkably, the Euclidean−Soergel product is slightly worse than the Soergel for all of the algorithms and representations—although there are more possible values than the Soergel shown in Table 7. This is due to the complex combinatorial nature of the distribution of statistically preferred values of the product. The distribution of the Euclidean distances has preferred values owing to its Gaussian-like shape. These preferred values are then multiplied in a combinatorial fashion such that they coincide with many of the Soergel statistically preferred values. This creates a distribution where the statistically preferred values come to dominate the distribution even more than they do in the Soergel distribution alone. Thus, there are more chances for there to be ties and hence more ambiguous ties. Note, however, this property may be independent of the product's efficacy at removing the compound size biases of the Euclidean and Soergel measures.

4.6. Discussion. If we only consider the number of ties, Wards clustering—with the Soergel measure and the Daylight fingerprints—has the fewest ties regardless of the data set. With the exception of focused data sets, the diversity of the data set may only play a minor role in how many ties are found. None of the data sets had a significant number of ties with the above combination. Even the Interbioscreen data set was not large enough to produce more than one ambiguous tie. Nevertheless, in general, the larger the data set, the more ambiguous ties we noticed, especially if the length of the bit vector is relatively small. That the focused nature of the AsInEx data set would produce a large number of ambiguous ties came as somewhat of a surprise to us, since the size of the data set was relatively small. The narrow distribution of the proximities, however, helps to explain why this is so.

In Brown and Martin's paper[19] they show that the Wards clustering algorithm, using the Euclidean distance and MACCS keys, to be the best for grouping compounds with respect to structure−activity data. Yet the results of our experiments suggest that this combination creates a considerable number of ambiguous ties of a serious nature, especially given the data set sizes that they and others (e.g., ref 20) commonly use. How much ambiguous ties may have confounded past experimentors' results, and hence their conclusions, thus becomes an open question.

5. CONCLUSIONS

Our results show that ambiguous ties can impact the hierarchies generated by various algorithms and measures, across a common range of data set sizes and bit vector representation sizes. Of the algorithms that we tested, complete-link was the most likely to produce large numbers of ambiguous ties. The production of ancillary numbers due to the merging criteria of Wards and group average ameliorates the ambiguous tie problem somewhat. However, even these algorithms can produce a significant number of ambiguous ties throughout the hierarchy while using the Euclidean distance and small number of compounds. The best combination of clustering algorithm, measure, and representation for reducing the number of ties, given our data sets, was Wards algorithm, the Soergel measure, and Daylight fingerprints.

We found that even though there are more possible proximities given the Euclidean−Soergel product than when using the Soergel, in general this measure produces more ambiguous ties than are typically found when using the Soergel measure. Though we have only explored a few data sets it appears that a very focused data set can produce a significant number of ambiguous ties, more so than more diverse sets of compounds.

6. FUTURE WORK

The fact that there are ambiguous ties in all of our examples shows that one must realize that there may be an inherent ambiguity in clustering compound data with bit vector representations and the various common measures and clustering algorithms. We have expanded on this idea and created an algorithm that uses ties to create a multidomain hierarchy.[35] A fuller discussion of the relationship between chemical hierarchies and ambiguous ties can also be found in this work.

Given our results presented above, we believe that more thorough study of ambiguous ties, incorporating other structural representations, such as BCI keys or minifingerprints, other data sets—both larger and with a wider range of diversity—and other measures (e.g., ref 36) is warranted. In this connection, we would additionally like to perform a more thorough review of statistically preferred values and their distributions and how they effect the generation of ambiguous ties. We also envision exploring a quantitative measure of the difference in results due to ambiguous ties, namely, a measure of the utility of a set of ambiguous clusterings, possibly using agreement subtrees (e.g., ref 37) or consensus partitions (e.g., ref 38), as this would be of considerable value in determining the efficacy of the clusterings. Work is ongoing[39] on the effects of ambiguous ties with asymmetrical clustering algorithms (e.g., refs 40 and 41).

Another important avenue of research relates to nonhierarchical clustering algorithms with the ambiguous ties problem. It is not well-known that other forms of clustering algorithms aside from hierarchical algorithms can also have significant problems with ties in proximity. A shared nearest-neighbor clustering algorithm such as Jarvis−Patrick[42] suffers from the choice of a fixed number of neighbors to be considered. Namely, if we let k be the number of nearest-neighbors to be considered, and we sort those neighbors so that the kth has the greatest dissimilarity, it is possible that there can be more than one nearest-neighbor with that same dissimilarity. The shared nearest-neighbors ($j < k$) among any two points should be taken from the pool of all k nearest-neighbors of each point, but those points tied to the kth nearest-neighbor should also be in that pool. Such problems are likely to become more acute as the number of compounds to be clustered increases but the size of the binary representation remains fixed. The choice of k and j also plays a role in how ties will effect the clustering. As the ratio of j to k gets larger, the more likely it is that such ties will impact the clustering. Graph clustering algorithms such as the minimum spanning tree (MST) clustering algorithm and various divisive algorithms[3,19] also can fall prey to ties in proximity. MST forms clusters by cutting the longest edges, but if many of those edges have the same distance, making arbitrary decisions as to which edge to cut introduces ambiguity in the results. A simple divisive algorithm based on the nearest-neighbors of furthest-neighbors as a splitting criterion can easily be shown to be subject to ties in proximity, as the nearest-neighbor groups can both have the same candidate members that are equidistant from the furthest-neighbors.

Occasionally, feature selection methods such as principle component analysis are used to to pick the most appropriate features and also to reduce the number of features considered. In general, with fewer features one would expect this to lead to a greater number of ambiguous ties. This might not always be the case if simple natural groups (partitions) are easily found with the new features set. This would be an interesting area of empirical research.

In some representation systems, counts rather than binary values are used. The value of counts over binary representations could be explored in the context of ties in proximity as well. For instance, if we assume that counts are bounded by a relatively small number, say k, then we can simply encode the feature counts as a binary vector, where in a simple encoding the number of bits equals k times the number of features. Each feature may have a different bound, but if we assume that the bound is the mean of the maximum counts for all the features, $\bar{k}$, we can get a rough estimate on the size of n on the basis of the size of the bit encoding of the count vector, $N\bar{k}$. Since $\bar{k}$ is typically small with respect to N, and may be in large measure inversely proportional to N, it will not in general appreciably increase n with any of the common measures mentioned above. $\bar{k}$ is also very data set dependent and thus empirically testing how much it increases n would be of some value.

REFERENCES AND NOTES

(1) Willett, P.; Winterman, V.; Bawden, D. Implementation of Non-Hierarchic Cluster Analysis Methods in Chemical Information Systems: Selection of Compounds for Biological Testing and Clustering of Substructure Search Output. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 109−118.

(2) Shemetulskis, N. E.; Dunbar, J. B.; Dunbar, B. W.; Moreland, D. W.; Humblet, C. H. Enhancing the Diversity of a Corporate Database Using Chemical Database Clustering and Analysis. *J. Comput-Aided Mol. Des.* **1995**, *9*, 407−416.

(3) Barnard, J. M.; Downs, G. M. Clustering of Chemical Structures on the Basis of Two-Dimensional Similarity Measures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 644−649.

(4) Martin, E. J.; Blaney, J. M.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring Diversity: Experimental Design of Combinatorial Libraries for Drug Discovery. *J. Med. Chem.* **1995**, *38*, 1431−1436.

(5) Sadowski, J.; Wagener, M.; Gasteiger, J. Assessing Similarity and Diversity of Combinatorial Libraries By Spatial Autocorrelation Functions and Neural Networks. *Agnew. Chem., Int. Ed. Engl.* **1995**, *34* (23), 2674−2677.

(6) Martin, E. J.; Critchlow, R. E. Beyond Mere Diversity: Tailoring Combinatorial Libraries for Drug Discovery. *J. Comb. Chem.* **1999**, *1*, 32−45.

(7) Warr, W. A. Combinatorial Chemistry and Molecular Diversity. An Overview. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 134−140.

(8) Menard, P. R.; Mason, J. S.; Morize, I.; Bauerschmidt, S. Chemistry Space Metrics in Diversity Analysis, Library Design, and Compound Selection. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1204−1213.

(9) Turner, D. B.; Tyrell, S. M.; Willett, P. Rapid Quantification of Molecular Diversity for Selective Database Acquisition. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 18−22.

(10) Ashton, M. J.; Jaye, M. C.; Mason, J. S. New Perspectives in Lead Generation ii: Evaluating Molecular Diversity. *Drug Discovery Today* **1996**, *1*, 71−78.

(11) Engels, M. F. M.; Thielemans, T.; Verbinnen, D.; Tollenaere, J. P.; Verbeeck, R. Cerberus: A System Supporting the Sequential Screening Process. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 241−245.

(12) Rhodes, N.; Willett, P.; Dunbar, J. B.; Humblet, C. H. Bitstring Methods for Selective Compound Acquisition. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 210−214.

(13) Stanton, D. T.; Morris, T. W.; Roychoudhury, S.; Parker, C. N. Application of Nearest-Neighbor and Cluster Analyses in Pharmaceutical Lead Discovery. *J. Chem. Inf. Comput. Sci.* **1998**, *39*, 21−27.

(14) Lewis, R. A.; Menard, P. R.; Mason, J. S. Rational Screening Set Design and Compound Selection: Cascaded Clustering. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 497−505.

(15) Weber, L.; Wallbaum, S.; Broger, C.; Gubernator, K. Optimization Of The Biological Activity Of Combinatorial Compound Libraries By A Genetic Algorithm. *Angew. Chem., Int. Ed. Engl.* **1995**, 34.

(16) MDL Information Systems, Inc., San Leandro, CA. Home page: http://www.mdli.com/.

(17) Daylight Chemical Information Systems, Inc., Mission Viejo, CA. Home page: http://www.daylight.com.

(18) Barnard Chemical Information Ltd. Sheffield, U.K. Home page: http://www.bci1.demon.co.uk/.

(19) 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.

(20) Wild, D. J.; Blankley, C. J. Comparison of 2D Fingerprint Types and Hierarchy Level Selection Methods for Structural Grouping Using Wards Clustering. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 155−162.

(21) Dixon, S. L.; Koehler, R. T. The Hidden Component of Size in Two-Dimensional Fragment Descriptors: Side Effects on Sampling in Bioactive Libraries. *J. Med. Chem.* **1999**, *42*, 2887−2900.

(22) Godden, J. W.; Xue L.; Bajorath, J. Combinatorial Preferences Affect Molecular Similarity/Diversity Calculations Using Binary Fingerprints and Tanimoto Coefficients. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 163−166.

(23) Flower, D. R. On the Properties of Bit String-Based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 379−386.

(24) Jain, A. K.; Dubes, R. C. *Algorithms for Clustering Data*; Prentice Hall Advanced Reference Series: Englewood Cliffs, NJ, 1998.

(25) Motwani, R.; Raghavan, P. *Randomized Algorithms*; Cambridge University Press: Cambridge, U.K., 1995.

(26) Hubalek, Z. Coefficients of Association and Similarity, Based on Binary (Presence−Absence) Data: An Evaluation. *Biol. Rev.* **1982**, *57*, 669−689.

(27) Beiler, A. H. *Recreations in the Theory of Numbers: The Queen of Mathematics Entertains*; Dover Publications: New York, 1966.

(28) Graham, R. L.; Knuth, D. E.; Pastashnik, O. *Concrete Mathematics*; Addison-Wesley: Reading, MA, 1994.

(29) 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.

(30) AsInEx. Organic Chemical Catalog, 6 Schukinskaya ulitsa, Moscow 123182, Russia. AsInEx home page: http://www.asinex.com.

(31) National Cancer Institute, Bethesda, MD. Home page: http://www.nci.nih.gov/.

(32) InterBioScreen Ltd., 121019 Moscow, Russia. Home page: http://www.ibscreen.com.

(33) Downs, G. M.; Willett, P.; Fisanick, W. Similarity Searching and Clustering of Chemical Structure Databases Using Molecular Property Data. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1094−1102.

(34) Matter, H.; Potter, T. Comparing 3D Pharmacophore Triplets and 2D Fingerprints for Selecting Diverse Compound Subsets. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 1211−1225.

(35) Nicolaou, C. A.; MacCuish, J. D.; Tamura, S. Y. A new multidomain clustering algorithm for lead discovery that exploits ties in proximities. In *Rational Approaches to Drug Design*; Proceedings of the 13th European Symposium on Quantitative Structure−Activity Relationships, Dusseldorf, Aug 27−Sept 1, 2000; Prous Scientific, in press.

(36) Kelley, B. P.; MacCuish, J. D.; Tamura, S. Y. Multi-Domain Clustering for Lead Discovery With a MCS and Recursive MCS Based Similarity Measure. Presented at the Society for Biomolecular Screening Conference 6th Annual Conference and Exhibition, Seattle, WA, Sept 6−9, 2000; Poster 416.

(37) Goddard, W.; Kubicka, W.; Kubicki, G.; McMorris, F. R. Agreement Subtrees, Metric and Consensus for Labeled Binary Trees. In *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*; American Mathematical Society: Providence, RI, 1995; Chapter 19, pp 97−104.

(38) Barthelemy, J.; Leclerc, B. The Median Procedure for Partitions. In *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*; American Mathematical Society: Providence, RI, 1995; Chapter 19, pp 3−34.

(39) Johnson, E.; MacCuish, J. D. Efficacy of Asymmetrical Clustering Algorithms for Lead Discovery. 2000, unpublished manuscript.

(40) Eppstein, D. Fast Hierarchical Clustering and Other Applications of Dynamic Closest Pairs. In *9th ACM-SIAM Symosium on Discrete Algorithms*; ACM and SIAM: 1998.

(41) Tarjan, R. An Improved Algorithm for Hierarchical Clustering Using Strong Components. *Inf. Process. Lett. (IPL)* **1983**, *17*, 37−41.

(42) Jarvis, R. A.; Patrick, E. A. Clustering Using a Similarity Measure Based on Shared Near Neighbors. *IEEE Trans. Comput.* **1973**, *C-22* (11), 1025−1034.

CI000069Q