Four Association Coefficients for Relating Molecular Similarity Measures

Cheng Cheng,† Gerald Maggiora, Michael Lajiness, and Mark Johnson*

Pharmacia & Upjohn, Inc., Kalamazoo, Michigan 49007-4940

Received February 12, 1996⊗

Four association coefficients are defined for assessing the relatedness between an arbitrary pair of molecular similarity measures. The utility of these coefficients is illustrated by assessing the relatedness of two molecular similarity measures in use at Pharmacia & Upjohn, one based on topological indices and the other on the occurence of structural fragments. The high coaggregation and discrimination coefficients imply that the two similarity measures preserve neighborhood relationships. Consequently, the two similarity measures should perform comparably with respect to property prediction if one used a locally-defined predictive method such as nearest neighbor prediction. The small density correlation coefficient indicates that the two measures could lead to quite different choices with regard to compound acquisition programs designed to fill in the "sparse" regions of structure space. The small distance correlation coefficient indicates that the two methods would perform differently when used for dissimilarity selection of subsets of diverse compounds. These coefficients replace our cruder assessments of the relatedness of two molecular similarity measures based on a series of similarity searches with more precise and statistically-defined statements regarding the manner in which they are related.

INTRODUCTION

Measures of molecular similarity are finding diverse applications in computational chemistry.[1−3] These measures are extremely varied,[4] and the number of similarity measures one might conceive in a related class of similarity measures can be very large.[5] Consequently, the referenced similarity measures are illustrative rather representative of all that are appearing in the literature.

The availability of such a wide variety of similarity measures raises issues as to which similarity measures are preferable and in which contexts. In many cases, whole classes of similarity measures can be excluded based on the nature of the problem under consideration. The computational cost of selecting 1000 most dissimilar structures from a database of 100 000 structures[6] or clustering of the compounds in such a database[7] usually restricts our attention to molecular representations based on vectors of chemical descriptors even though the problem can be formally defined for all similarity measures. Similarity measures used in comparing the presence and absence of bonds in a molecular structure traditionally employ the chemical graph[8,9] as the molecular representation of choice, while similarity measures used to describe the "taxonomy" of molecular orbitals naturally employ quantum mechanical representations.[10] In these two cases, one need only examine the chemical representation on which the similarity measure is defined in selecting the best class of similarity measures for the job. We shall say such selection methods are *representation based*.

Once attention has been restricted to a particular class of measures because of the nature of the problem, we must still select a similarity measure from that class. In cases where very diverse similarity measures have been used for the same type of problems this class can be very large. For example, almost the entire range of similarity measures have been used in the study of the biological effects of compounds at one time or another[4] although the presence of an obvious three-dimensional component in drug-receptor interactions has tended to focus effort in the direction of similarity measures based on 3D representations.[3] In these situations, similarity measures should be ranked based on performance independently of the different forms of molecular representations under consideration. We shall say such selection methods are *performance based*. In the study of the biological effects of compounds, performance is based on the similar property principle that similar compounds tend to have similar properties.[5,11−14] Good similarity measures are expected to segregate active compounds from inactive compounds in structure space better than poor ones. Surprisingly, 2D representations have performed comparably if not better than 3D methods in these comparisons.[12,14] This may be only one of many surprises to be encountered as we begin the study of molecular similarity measures utilizing performance-based methods rather than representationally-based methods.

In this report, some new methods of relating similarity measures are proposed in which performance is defined by the "structure" space induced by the similarity measures. Such methods are said to be *space based*. To illustrate, suppose d_1 and d_2 are two molecular similarity measures possibly based upon two quite different representations of chemical structure. If, however, for every pair of structures, **s** and **t**, we have $d_1(\mathbf{s},\mathbf{t}) = d_2(\mathbf{s},\mathbf{t})$, then d_1 and d_2 would be indistinguishable via any space-based comparison. Clearly, they would also be indistinguishable via any performance-based method founded on the similar property principle.

We present four space-based methods for examining pairwise relations between similarity measures. The coaggregation method measures the extent to which the joint behavior of two measures differs from what would be expected if the two measures were to structure their respective spaces in a statistically independent manner. The pairwise distance and density methods correlate pairwise distances and the relative densities of structures induced by

† Johns Hopkins University.
* To whom correspondence should be directed: Email majohns1@upj.com.
⊗ Abstract published in *Advance ACS Abstracts,* June 1, 1996.

S0095-2338(96)00460-X CCC: $12.00

the two measures. The discrimination method assesses the extent to which two measures induce similar neighboring relationships.

Each method constructs a index of relatedness for any pair of similarity measures, and consequently, serves as a basis for arraying the molecular similarity measures themselves in "similarity-measure space." This is of little consequence to the present study as only two molecular similarity measures have been related. However, the actual values of the four association indices give insights into how the two measures differently array compounds in structure space. We will see that perceptions of chemical diversity depend very markedly on our choice of a similarity measure even when these two measures intuitively perform quite comparably.

THE SIMILARITY MEASURES

The same two molecular similarity measures were used in the construction of each of the four association indices. For the topological index or TI distance, each compound is represented by a 10-dimensional vector consisting of its values for the first 10 principal components determined over the nonpeptide structures in the Cousin[15] database for all but the 21 IC_r, SIC_r, and CIC_r, $r = 0,...,6$ indices of the 90 topological indices as defined in Basak et al.[16] The similarity measure between two compounds is the Euclidean distance between their respective TI-vector representations. (Here, we follow the usual convention of referring to the measure as a "similarity" measure even though it employs a distance formula in its computation. However, when referring to the formula/index/coefficient of the measure, we will use whichever term, "similarity" or "distance," is most specific.)

For the fragment or FR-similarity measure, each compound is represented by a 320-component binary or bit vector. Each component corresponds to one or more molecular fragments as implemented in the initial screening step of the substructure search algorithm in Cousin.[15] The magnitude of the FR-similarity measure between two compounds, i and j, is the Jaccard coefficient, defined on the respective bit vectors by $\sum x_{1j}x_{2j}/(\sum x_{1j} + \sum x_{2j} - \sum x_{1j}x_{2j})$ where x_{ij} is 1 if the jth fragment is present on the ith compound, and 0 otherwise. It is also referred to as the Tanimoto coefficient. In the case of bit vectors, the Soergel coefficient is one minus the Jaccard coefficient.[17] The Soergel coefficient is a metric[17] and thus satisfies the triangular inequality. In this study, it is convenient to use the Soergel coefficient rather than the Jaccard coefficient, and we will do so.

In practice, a similarity measure is defined on a set or database of compounds. The density study was carried out on the complete Cousin database of nonpeptide compounds in use at Pharmacia & Upjohn. The coaggregation, distance, and discrimination studies were carried out on a subset of 15 500 compounds randomly selected from the database used in the density study.

In the density and coaggregation studies, the number of compounds falling within prescribed query searches were counted. Since neither representation distinguishes stereoisomers, we counted the number of distinguishable chemical graphs in the hit sets rather than the number of distinguishable stereoisomers. In the case of ionic compounds, the TI representation is based solely on the component of the salt with the most atoms. On the other hand, FR-search bits will be turned on if the corresponding fragment is present in the smaller component. Thus, the TI representation will not distinguish salts, but the FR representation might. We counted the number of salt "parent" classes (those with a common largest component) rather than the number of distinguishable salts. Finally, only one representation of a tautomer is registered in our database, and only that representation is used in computing the similarity searches and counting the hits.

COAGGREGATION

Suppose we have a molecular similarity measure. We might demonstrate its reasonability by showing that a similarity search with a rigid compound such as DDT as a query structure will return other rigid analogs of DDT when executed over any database containing such analogs. Such a demonstration is given in Johnson et al.[18] with the Available Chemicals Directory (marketed by Molecular Design Ltd.) as the database. In that study, the hits from an FR-similarity search are compared to those for one of our in-house 3D-similarity measures. All of the six closest hits to DDT for each similarity measure were analogs of DDT, and two of the hits were common to both hit sets.

Such outcomes give strong empirical support to the notion that reasonable molecular similarity measures are quite highly related from the perspective of similarity searching. But these outcomes are quite subjective from two standpoints. First, how does one agree upon which rigid analogs to use for query structures? Second, how does one decide if the hits are in the same structural classes as the query structure? Obviously, one would like more objective and quantitative measures of the functional relatedness of two similarity measures. Here we present a new measure of association between two molecular representations called the coaggregation coefficient. We begin by relating the coaggregation coefficient, which derives from the notion of random quadrat sampling in spatial statistics,[19−21] to the correlation coefficient.

The correlation coefficient measures the association between two scalar properties, say the height and weight of a random sample of individuals. If X and Y represent these properties and if X and Y are normalized to have zero mean and unit variances, then the correlation coefficient is formally defined by

$$\rho = \int\int xy\, \mathrm{f}(x,y)\mathrm{d}x\mathrm{d}y$$

where $\mathrm{f}(x,y)$ denotes the density describing how the two variables are jointly distributed. The bivariate normal density with zero mean and unit variances is a commonly studied form for a joint density between two scalar variables. It is given by

$$\mathrm{f}(x,y) = \frac{1}{2\pi(1-\rho^2)}\exp\{(x^2 - 2\rho xy + y^2)/(1-\rho^2)\}$$

The correlation coefficient, ρ, is the sole parameter of the joint density in this special case.

Taken by itself, the height property X has its own density $\mathrm{f}(x)$ called the marginal density of X. It is obtained by integrating the joint density $\mathrm{f}(x,y)$ over all values of y and is given by

$$\mathrm{f}(x) = \int_{-\infty}^{\infty}\mathrm{f}(x,y)\mathrm{d}y = \frac{1}{\sqrt{2\pi}}e^{x^2/2}$$

Table 1. The Median, and the 5 and 95% Quantiles Given in Parentheses, of 100 Simulated Estimates of the κ Coaggregation Coefficient[a]

correlation coefficient	no. of points in the simulated database		
	1000	2500	5000
0	0.48 (0.32−0.60)	0.44 (0.25−0.62)	0.47 (0.28−0.63)
0.7	0.57 (0.42−0.70)	0.55 (0.40−0.72)	0.58 (0.42−0.70)
0.9	0.68 (0.55−0.85)	0.70 (0.58−0.83)	0.70 (0.57−0.80)
0.95	0.78 (0.65−0.87)	0.78 (0.65−0.87)	0.78 (0.70−0.88)
0.99	0.90 (0.82−0.95)	0.90 (0.82−0.95)	0.90 (0.82−0.92)
0.999	0.97 (0.92−1.00)	0.97 (0.92−1.00)	0.97 (0.92−1.00)

[a] The joint density is defined by the bivariate normal density. Sixty query structures were randomly selected from both the joint and product densities in each simulation. The two search radii were individually readjusted in each simulation to give approximately the same number of hits and jointly readjusted so that the proportion of zero hits in the intersection never exceeded 0.5 when sampling from the joint density. The medians for the latter proportions were between 0.46 and 0.47 for all 18 tabulated cases.

Whenever $f(x,y) = f(x)f(y)$, X and Y are said to be statistically independent. In the special case of the bivariate normal density, this will occur if and only if $\rho = 0$. However, there are numerous examples of joint densities for which $\rho = 0$, and yet the joint density does *not* factor into its two marginal densities. This will be illustrated in a later example.

Although this discussion assumed that X and Y were one-dimensional variables, the notion of the joint densities $f(x,y)$, the marginal densities $f(x)$ and $f(y)$ defined by integrating out one of the variables in the joint density, and the product density $f(x)f(y)$ are all defined even if X and Y are multidimensional. In fact, the definitions of these terms make no assumption that X and Y even take values in a Euclidean space. For example, the values of X and Y could be labeled graphs, configurations of points in R^3, or three-dimensional scalar fields. In this section, we measure how far the joint density $f(x,y)$ departs from the product density $f(x)f(y)$ using the coaggregation coefficient in a manner applicable to these more general contexts.

Estimation of the coaggregation coefficient basically involves two steps. First, similarity searches are run for a number of query structures randomly selected according to the joint density $f(x,y)$. The details of how to do this will be given shortly. The number of hits in each search neighborhood is counted. Specifically, for any particular similarity measure M and query structure q, it is the number of compounds C, other than q, for which $d_M(q,C) < r$, where r is the search radius. Second, the first step is repeated, but this time the query structures are randomly selected according to the product density $f(x)f(y)$. All similarity searches must be based upon the same radius. This radius is adjusted so that roughly half the counts in the first step are zero. The coaggregation coefficient, denoted by κ, is the proportion of zeros in the second step. If X and Y are independent of each other, the joint and product densities are identical, and κ will vary about 1/2 by construction. For a wide variety of different types of extreme dependencies κ is close to 1.

The coaggregation coefficient is calibrated against the correlation coefficient in Table 1 when the joint density is the bivariate normal. The details of this calibration are given in the appendix. The first row is reassuring. When there is no association, the κ estimates vary about the desired value of 1/2 when the correlation coefficient is 0. The spread between the 5 and 95% quantiles are also relevant as these

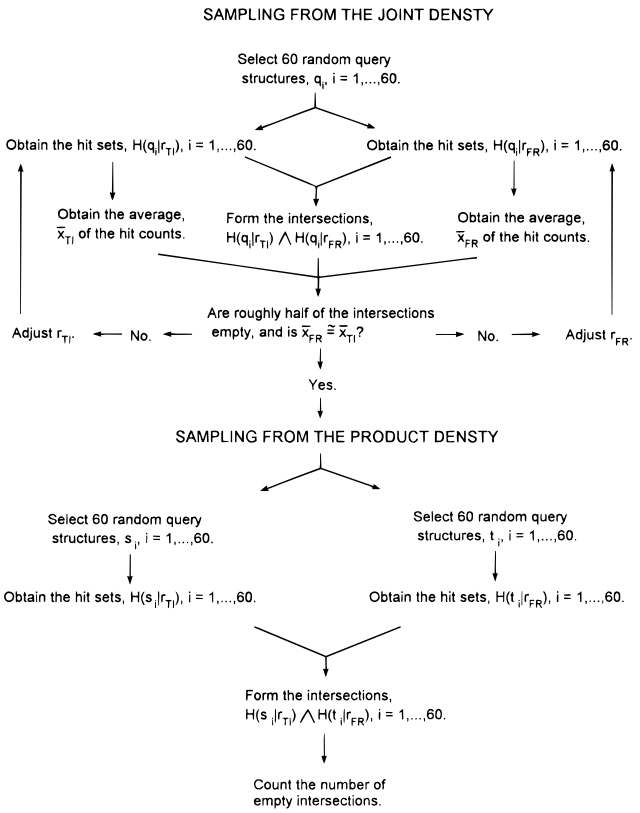

Figure 1. A flow chart of the procedure for obtaining the search radii and the hit counts when computing the coaggregation coefficient.

simulations are based on the same number (60) of randomly selected query structures as was used in estimating κ for the coaggregation of the TI and FR similarity measures. It is also reassuring that the κ estimates do not depend on the size of the database once that number exceeds 1000, although the search radii necessarily do as noted in the Appendix. Note that a κ of 1 based on 60 query points randomly selected from the two design densities is associated with an extremely high degree of relatedness (>0.99).

In estimating the coaggregation coefficient for the Cousin database, a TI-distance search with radius $r_{TI} = 0.797$ and a FR-distance search with radius $r_{FR} = 0.285$ (based on the Soergel distance coefficient) was run for each of 60 structures randomly selected from the subset of 15 500 Cousin compounds. For each pair of TI and FR neighborhoods having common query structure, the number of compounds in the corresponding joint neighborhood in "TI × FR space", the xy-space in the preceding equations, was determined by counting the number of compounds in the intersection of the two TI and FR neighborhoods. The values of r_{TI} and r_{FR} were adjusted so that (1) the average number of hits per query was the same for the TI and FR searches, and (2) roughly half (34 of 60) of the corresponding joint neighborhoods had 0 hits. The query structures were not included in the hit counts. See the upper half of Figure 1 for details. It might be noted that these two conditions essentially fix the choices for the two radii. As a consequence, the coaggregation coefficient estimates a parameter of the relationship between the two similarity measures that is relatively independent of the size of the search radii used in its estimation in any particular case. This invariance to the size of the search radii was seen in Table 1 where the values of the estimated aggregation coefficients did not depend on the size of the simulated database to any significant extent.

Table 2. Summary of Hits from Random Query Sampling with a TI-Distance Radius of 0.797 and a FR-Similarity Radius of 0.715

design density	no. with no hits	no. with hits
joint density	34	26
product density	60	0

Since the counts in the joint neighborhoods are based on the intersections of the TI and FR-hit sets having common query structures, they represent neighborhoods randomly selected from the *joint* density. To randomly sample neighborhoods from the product density, one simply removes the restriction that the intersected neighborhoods be associated with the same query structure. Thus, a TI-similarity search with radius r_{TI} was run on each of 60 additional randomly selected structures. Denote these structures by $s_1,...,s_{60}$. In the same manner, a FR-similarity search with radius r_{FR} was run on each of 60 more randomly selected structures. Denote these by $t_1,...,t_{60}$. Although it is extremely unlikely, it does not matter if some structures are common to both sets. For each value of i, $i = 1,...,60$, we count the number of compounds in intersection of the TI-neighborhood having query structure s_i and the FR neighborhood having query structure t_i. See the low half of Figure 1 for details. In this case, the counts reflect joint neighborhoods in xy-space randomly selected from the *product* density. All 60 counts were zero giving a coaggregation coefficient of 1.

The counts are summarized in Table 2. If the joint and product densities are identical, we expect the numbers in the first column to differ only by amounts attributable to sampling variation. This assumption is tested by Pearson's χ-square test for 2×2 tables.[22] In our case, $\chi^2 = 33.2$, which is statistically significant at 0.0001. Thus, there is strong evidence that the joint density is not identical to the product of its TI and FR densities. This does not imply that the differences between these two densities need be large. However, the value of 1 for κ suggests that the joint and product densities differ considerably from each other. We base this conclusion on the results of the calibration study in Table 1. There we see that when the κ values were centered about 0.90 (the row with a correlation coefficient of 0.99), fewer than 5% of the time the κ values exceeded 0.95. Thus it is unlikely that our value of 1 for κ can be attributed to sampling variation. Moreover, we also see from Table 1 that a κ of 0.9 corresponds to a correlation coefficient of 0.99.

We conclude that the TI and FR representations induce neighborhood relationships between compounds in a highly related manner. This means that close neighbors of a structure in FR space tend to occur in related subregions of TI space as opposed to being randomly strewn about that space. Similarly, close neighbors of a structure in TI space tend to occur in related subregions of FR space. This is consistent with our intuitive experience of the high performance relatedness of the two representations based upon similarity searches involving common query structures. However, we are given no hint as to the nature of these related subregions other than that they exist.

METRIC DEPENDENCIES

The high value for the coaggregation coefficient rigorously confirms that the TI and FR representations position molecules in their respective similarity spaces in a related manner. But to simply say that two variables are highly related does not say anything about the way in which they are related. To illustrate, suppose X is a scalar variable that always takes on values in the set $\{x_1,...,x_m\}$. Let p be any permutation of the integers $\{1,...,m\}$, and let Y_p be another scalar variable such that $Y_p = x_{p(i)}$ if $X = x_i$. For every permutation p, X and Y_p are statistically highly related. In fact, a knowledge of the value of X gives complete knowledge about the value of Y_p given we know p. If p is the identity function, then the sample correlation coefficient between X and Y_p would be 1. However, there are m factorial other possible relationships in this class of permutation relationships all of which exhibit an equally high statistical dependence. As correlation coefficients measure linear dependencies, many of these relationships will have sample correlation coefficients near 0.

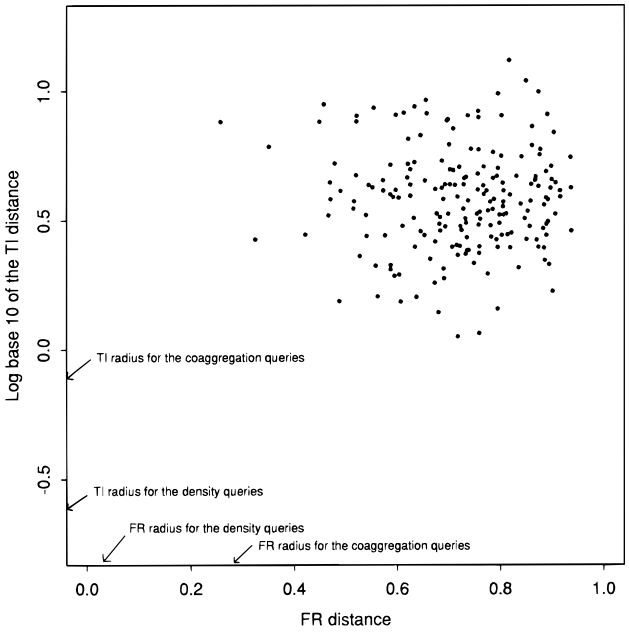

Figure 2. FR distances versus TI distances on 200 structure pairs randomly selected from the Cousin database of nonpeptide compounds.

This permutation method for defining relationships between two variables illustrates how statistical dependencies can be constructed without taking metric properties into account. However, when statistical dependence exists, it is informative to see what metric relationships, if any, exist between the two variables. This section will examine metric relationships between the two similarity measures that may explain their high statistical dependence.

Pairwise Distances. Pairwise distances among the structures in the two similarity spaces were examined to check for global space-based relatedness. Two hundred (200) pairs of structures were drawn at random from our randomly selected subset of 15 500 structures. The results are given in Figure 2 where FR similarity is based on the Soergel distance.

Figure 2, with an associated correlation coefficient of 0.04, clearly demonstrates that the two similarity measures are distancewise unrelated. To know two structures are relatively far apart in TI space says nothing about their relative separation in FR space and vice versa.

In addition, by projecting the points back to the axes, we can sense the distributions for the distances between two randomly selected structures. FR distances generally range from 0.4 to just under 1, while TI distances generally range

from 1 to 10. (These latter numbers are obtained from the ordinate of Figure 2 after taking the antilogarithm.) From these ranges, we see that the FR distance search radius, $r_{FR} = 0.285$, and the TI distance search radius, $r_{TI} = 0.797$, (log-(0.797) = −0.099) used in the coaggregation study is quite small. Rarely would two randomly selected structures be separated by such small distances. On the other hand, these radii were selected in the coaggregation study so that roughly half of the structures have at least one nearest neighbor within those distances.

Density. Before we look at the relatedness of the TI and FR similarity measures from the perspective of the densities they induce, it is necessary to see how the size of the database can influence our perception of extremely dense regions in structure space which may arise, for example, from lead optimization programs.

Suppose, for ease of argument, that compounds are randomly strewn over structure space except for one very small, but very dense region. By very dense, we mean that a query centered in that region would be a 1000 times more likely to have a hit than a query of the same radius located elsewhere. By a very small region, we mean that only 1 in 1000 structures lies in that region. Suppose our database has 10 000 structures. Then we would expect there to be about 10 compounds in the dense region. If we ran a number of random query searches of radius r_1 where r_1 is large enough so that the average hit count is 1, then we would expect hit counts of 0, 1, and 2 with an occasional 3 or 4. If by chance (1 in 1000) we selected a query from the dense region, all of the structures in that region would be hits as might a neighboring structure or two. Thus, we would expect a hit count of around 10, which would not be too impressive relative to the other hit counts. On the other hand, suppose our database has 1 000 000 structures so that there would be about a 1000 compounds in our dense region. Suppose we again ran a number of random query searches of radius r_2 where r_2 is large enough so that our average hit count is 1. We would expect r_2 to be smaller then r_1, and we would again expect to see hit counts of 0, 1, and 2 with an occasional 3 and 4. However, this time we would obtain a hit count of around 1000 if the query is located in the dense region. Thus, the range in hit counts associated with the larger database more correctly conveys the difference in the densities that we have defined.

This points out a problem with studies dealing with the classification of sparse and dense regions. If there exist very small, but very dense regions, it takes a large database of structures to convey the magnitude of the differences in the possible densities. For this reason, our density study uses the complete Cousin file of nonpeptide structures rather than the subset of 15 500 randomly selected structures used in the coaggregation, distance, and discrimination studies.

In this study, TI and FR-similarity searches were run for each of 600 structures randomly selected from the complete Cousin database of nonpeptide compounds. All TI-similarity searches were run at a TI-distance radius of $r_{TI} = 0.25$, and all FR-similarity searches were run at a FR-distance radius of $r_{FR} = 0.025$ (which corresponds to a Tanimoto coefficient of 0.975.) These particular cutoff values were chosen so that the average number of hits in both cases was approximately the same, and so that the maximum number of hits in any one search was less than 50. The average number of counts was 1.05 for the TI searches and 0.98 for the FR searches. For each pair of searches, the number of hits was recorded. The results are given in Figure 3.

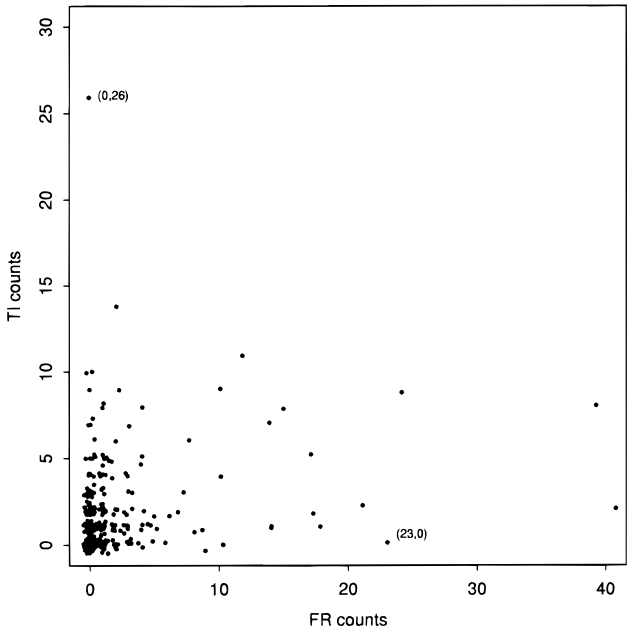

Figure 3. Hit counts for FR and TI-similarity searches for 600 query structures randomly selected from the Cousin database of nonpeptide compounds. Positions of plotted points are jittered by adding to each a randomly generated normal deviate to its x and y coordinates. The (0,0) position corresponds to 313 structures.

In Figure 3, the distribution of FR counts is obtained by projecting all plotted points down on the "FR counts" axis. Clearly, most queries were 0-hit queries (448) or 1-hit queries (77) with a fair number of larger hit queries and a 39 and a 41-hit query. This supports a picture not too unlike what we have been supposing. Most of FR space, at least (448 + 77)/600 or 88% is sparse, but there are some small, but very dense regions. The density of these regions is also apparent from Figure 2 where the FR radius of 0.025 is put in perspective with the average pairwise distance in FR space between two randomly selected structures. Similar comments can be made about TI space. In this case there were 357 0-hit queries and 106 1-hit queries with fair number of larger hit queries, and a 26-hit query. One comes away with the feeling, supported by Hodes and Feldman,[23] that structure space somewhat resembles the occupied space of a molecule where an essentially vacant space is sparsely peppered with dense nuclei.

Extremely dense regions of compounds exist in both spaces, and it is natural to ask if structures associated with dense regions in TI space are also associated with dense regions in FR space, i.e., are the two similarity measures densitywise related? Figure 3, with an associated correlation coefficient of 0.28, strongly suggests not. For example, the query structure for the 26-hit query in TI space gave rise to a 0-hit query in FR space. This corresponds to the (0,26) point in Figure 3. Conversely, the (23,0) point in Figure 3 corresponds to a query structure for a 23-hit query in FR space that gave rise to a 0-hit query in TI space. Although these represent extreme cases, Figure 3 suggests that knowing the density of compounds in the local neighborhood of a structure in FR space tells us little about the density of compounds in the local neighborhood of that structure in TI space, and vice versa.

Discrimination. So far it has been shown that the TI and FR representations are neither distancewise nor densitywise

Table 3. Number of Times a Nearest Neighbor of **u** in TI/FR-Space Is Nearer **u** in the Other (FR/TI) Space Than to a Randomly Selected Structure **v**

FR discriminates **u** from **v** with respect to TI	TI discriminates **u** from **v** with respect to FR		
	yes	no	total
yes	167	17	184
no	13	3	16
total	180	20	200

related. Now consider a discriminantwise relationship. Let **u** and **v** denote two structures in a database. Let **u′** denote the nearest neighbor of structure **u** in TI space, and let d_{FR}(**u**,**u′**) denote their FR distance. Let **v** be any other structure. We shall say FR discriminates **u** from **v** with respect to TI if $d_{FR}(\mathbf{u},\mathbf{u}') < d_{FR}(\mathbf{u},\mathbf{v})$. Now suppose **u** and **v** are selected at random. Sometimes FR will discriminate **u** from **v** with respect to TI and sometimes not. Denote the proportion of time it does by δ(FR|TI). Define δ(TI|FR) analogously, and define the discrimination coefficient δ by the average of δ(FR|TI) and δ(TI|FR). The discrimination coefficient measures the extent to which the two representations induce similar neighborhood relationships. If δ is close to 1, we know that a nearest neighbor of randomly selected structure **u** in one space is likely to be closer to **u** in the other space than to another randomly selected structure.

To estimate the discrimination coefficient, we randomly sampled 200 compound pairs (**u**,**v**) from the subset of 15 500 compounds selected from Cousin. For each structure **u**, we found its nearest neighbor **u′** in TI space and then tallied a 1 if **u′** was closer to **u** than to **v** in FR space. Similarly, for each structure **u**, we found its nearest neighbor **u″** in FR space and then tallied a 1 if **u″** was closer to **u** than to **v** in TI space. The results are given in Table 3.

From Table 3, obtain estimates of 184/200 or 0.92 for δ(FR|TI) and 180/200 or 0.9 for δ(TI|FR). Their average gives an estimate of 0.91 for the discrimination coefficient δ. Thus, we see that the two similarity measures basically preserve neighborhood relationships.

SUMMARY AND CONCLUSIONS

Although this study has presented some approaches for relating any two similarity measures with respect to their positioning of molecular structures in their respective spaces, our conclusion will be based on the results for the TI and FR-similarity measures. Most obviously, the two measures position molecular structures in a highly related manner as judged by the high coaggregation coefficient. The representations of these measures must be encoding similar types of information which is obscured by their disparate forms. However, the coaggregation coefficient gives no hint as to what this common information might be.

The analysis of the metric dependencies further examines the nature of the relationship between the TI and FR similarity measures. Global relationships, as expressed by pairwise distances, were not preserved, but local relationships, as expressed by the discrimination coefficient, were largely preserved. On the other hand, the lack of any relationship between which structures are associated with high density regions and which structures are associated with sparse regions suggest that the two similarity measures, relatively speaking, shrink and stretch their respective spaces in an unrelated manner.

This picture of the relationship between the two similarity measures has important implications regarding the uses to which similarity measures are being put. The preservation of local neighborhoods suggests that the two similarity measures should behave rather comparably with respect to property prediction performed by locally-defined predictive methods such as nearest neighbor prediction.

The lack of correlation between the densities associated with the local neighborhoods has some important implications regarding the objectivity of compound acquisition programs directed toward filling in the sparser regions of structure space. This study suggests that a relatively sparse region in one similarity space could be an extremely dense region in another similarity space even when the associated similarity measures are highly related as judged by either the coaggregation coefficient or the discrimination coefficient. However, this study also suggests that there are many regions of structure space that are sparse by both similarity measures. It would make sense to augment these mutually sparse regions before augmenting regions that are sparse by one measure, but not by the other.

The lack of correlation between the pairwise distances also has some important implications regarding the objectivity of programs associated with selecting subsets of diverse compounds. Again, use of two highly related measures, as judged by either the coaggregation coefficient or discrimination coefficient, could result in the selection of compounds from markedly different regions of structural space. One similarity measure could lead to the selection of two dissimilar compounds which are deemed to be quite similar by the other measure.

The four coefficients of association have added precision to our perceptions of how two similarity measures might be related. Prior to these studies, we would have said the TI and FR similarity measures are highly related from a performance viewpoint based upon running a series of similarity searches with respect to a number of common query structures. Now we would say these two measures have an extremely high coaggregation coefficient of 1, a fairly strong discrimination coefficient of 0.91, and negligible distance and density correlation coefficients of 0.04 and 0.27, respectively.

These coefficients have also forced a reassessment of our intuition into "structure space". Molecular similarity space based upon vector representations of structure has a dimensionality that far exceeds common experience. Typical relationships between point clouds in very high dimensions may differ significantly from typical relationships between point clouds in three-dimensional Euclidean space. This all suggests that there may be many more surprises that lie ahead in our endeavors to relate the varied molecular similarity measures presently available.

APPENDIX

Let Z_1, Z_2, and Z_3 represent numbers from an algorithm for generating random normal deviates with mean 0 and variance 1. Let $X = aZ_1 + Z_2$ and let $Y = aZ_1 + Z_3$. Then (X,Y) represents a random outcome from a bivariate normal density with correlation coefficient $\rho = a^2/(a^2 + 1)$.

Using this representation, the calculation of a single coaggregation coefficient can be described as follows:

1. Generate n values for each of Z_1, Z_2, and Z_3 where n corresponds to the size of the database. Let D_x and D_y denote

the sets of X and Y values, each of size n, computed according to the preceding equations. Here each X value and Y value is analogous to a single structure in TI and FR space, respectively, only in this case X and Y values are one-dimensional as in the discussion of the bivariate normal.

2. Select 60 integers at random from the set $\{1,...,n\}$ such that there are not repeats and denote this set of "query" integers by Q.

3. Pick or adjust the radii r_x and r_y. Again this is analogous to selecting the radii for the TI and FR query neighborhoods.

4. For each q in Q, compute the set $Q_x(q)$ of integers i, $i \neq q$, for which $d(x_i,x_q) < r_x$ for x_i in D_x. Compute $Q_y(q)$ analogously. Let $N_x(q)$ and $N_y(q)$ be the number of "hits" in $Q_x(q)$ and $Q_y(q)$.

5. Compute the number $N_{xy}(q)$ of "joint hits" by counting the number of integers in the intersection of $Q_x(q)$ and $Q_y(q)$.

6. Let P_{xy} be the proportion of times (out of 60) that $N_{xy}(q) = 0$, and let A_x and A_y be the average of the 60 $N_x(q)$ and $N_y(q)$ values, respectively. If either P_{xy} is not sufficiently close to 0.5 or if A_x and A_y are not sufficiently close to each other, go to 3; otherwise continue.

7. Select 60 more integers at random from the set $\{1,...,n\}$ and redefine this set of "query" integers to be Q_x and select 60 more integers at random from the set $\{1,...,n\}$ and redefine this set of "query" integers to be Q_y.

8. For each q in Q_x, recompute the set $Q_x(q)$ of integers i, $i \neq q$, for which $d(x_i,x_q) < r_x$ for x_i in D_x. Recompute $Q_y(q)$ analogously. Let $N_x(q)$ and $N_y(q)$ be the number of "hits" in $Q_x(q)$ and $Q_y(q)$.

9. Recompute the number $N_{xy}(q)$ of "product hits" in the intersection of $Q_x(q)$ and $Q_y(q)$.

10. Calculate the proportion of time out of 60 that $N_{xy}(q) = 0$. This is the coaggregation coefficient.

Steps 1 through 10 were each repeated 100 times for each of the 18 combinations of ρ and n given in Table 1.

The decision criteria at step 6 are intentionally ambiguous. Exact mathematical definitions were too time consuming to compute. The algorithm that was used is not particularly interesting to the purposes of this study. To see the extent to which these criteria were achieved in the simulation study, note how the minimum of 0.25 for the 5% quantiles in Table 1 gives a fair lower bound for the simulated κ values. Similar statistics evaluated on the adjustment parameters, P_{xy}, A_x, and A_y, computed in step 6 give a fair impression on how well the algorithm achieved the desired decision criteria. At step 6, we attempt to set P_{xy} close to, but not exceeding 0.5. The minimums of its 5% and 50% quantiles across the 18 combinations of ρ and n were 0.32 and 0.46, respectively, and the maximums of its 50% and 95% quantiles were 0.47 and 0.5, respectively. We can look at the agreement between A_x and A_y by examining their ratio. The minimums of the 5% and 50% quantiles of these ratios were 0.82 and 0.98, and the maximums of the 50% and 95% quantiles were 1.06 and 1.23. Thus the desired values for P_{xy}, A_x, and A_y were achieved quite well.

Finally, it should be noted that the average search radii, r_x and r_y, for the cases with $n = 5000$ were roughly half of the average search radii for the cases in which $n = 1000$. For the cases with $\rho = 0.999$, the average search radii were roughly a fifth of the average search radii for the cases with $\rho = 0$.

REFERENCES AND NOTES

(1) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press Ltd.: Letchworth, 1987.

(2) *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; Wiley Interscience: New York, NY, 1990.

(3) *Molecular Similarity in Drug Design*; Dean, P. M., Ed.; Blackie Academic & Professional: London, 1995.

(4) Johnson, M. A. A Review and Examination of the Mathematical Spaces Underlying Molecular Similarity Analysis. *J. Math. Chem.* **1989**, *3*, 117−145.

(5) Willett, P.; Winterman, V. A. Comparison of Some Measures for the Determination of Inter-Molecular Structural Similarity - Measures of Inter-Molecular Structural Similarity. *Quant. Struct.-Act. Relat.* **1986**, *5*, 18−25.

(6) Lajiness, M. S. Molecular Similarity-Based Methods for Selecting Compounds for Screening, *Computational Chemical Graph Theory*; Rouvray, D. H., Ed.; Nova Science Pub.: New York, NY; 1990; pp 299−316.

(7) Whaley, R.; Hodes, L. Clustering a Large Number of Compounds. 2. Using the Connection Machine. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 345−347.

(8) Dugundji, J.; Gillespie, P.; Marquarding, D.; Ugi, I; Ramirez, F. Metric Spaces and Graphs Representing the Logical Structure of Chemistry*; Chemical Applications of Graph Theory*; Balaban, A. T., Ed.; Academic Press: New York, NY, 1976.

(9) Johnson, M. Structure−Activity Maps for Graphically Analyzing Data in Nondimensional Metric Spaces Arising in Drug Design. *J. Biopharm. Statist.* **1993**, *3*, 203−236.

(10) Carbo, R.; Domingo, L. LCAO-M.O. Similarity Measures and Taxonomy. *Int. J. Quantum Chem.* **1987**, *32*, 517−545.

(11) Johnson, M. A.; Naim, M.; Nicholson, V.; Tsai, C.-c. Comparing the Substructure Metric to Some Fragment-Based Measures of Intermolecular Similarity. In *QSAR in Drug Design and Toxicology*, Hadzi, D., Jerman-Blazic, B., Eds.; Elsevier Science Pub.: Amsterdam, 1987; pp 67−69.

(12) Pepperrell, C. A.; Willett, P. Techniques for the Calculation of Three-Dimensional Structural Similarity Using Inter-Atomic Distances. *J. Comput.-Aided Mol. Design.* **1991**, *5*, 455−474.

(13) Basak, S. C.; Grunwald, G. D. Predicting Mutagenicity of Chemicals Using Topological and Quantum Chemical Parameters: A Similarity Based Study. *Chemosphere*. **1995**, *31*, 2529−2546.

(14) 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.* In press.

(15) Hagadone, T. R. Molecular Substructure Similarity Searching: Efficient Retrieval in Two-Dimensional Structure Databases. *J. Chem. Inf. Comput. Sci*. **1992**, *32*, 515−521.

(16) Basak, S. C.; Magnuson, V. R.; Niemi, G. J; Regal, R. R. Determining Structural Similarity of Chemicals Using Graph-Theoretic Indices. *Discrete Appl. Math*. **1988**, *19*, 17−44.

(17) Gower, J. C. Measures of Similarity, Dissimilarity, and Distance. In *Encyclopedia of Statistics*; Kotz, S., Johnson, N. L., Eds.; Wiley-Interscience: New York, NY, 1985; Vol. 5, pp 397−405.

(18) Johnson, M. A.; Maggiora, G. M.; Lajiness, M. S.; Moon, J. B.; Petke, J. D.; Rohrer, D. C. In *Advanced Computer-Assisted Techniques in Drug Discovery*; Van de Waterbeemd, H., Ed.; VCH: Weinheim, 1994; Vol. 2, pp 89−110.

(19) Douglas, J. B. *Analysis with Standard Contagious Distributions*; Int. Coop. Pub. House: Fairland, MD, 1980.

(20) Cheng, C.; Johnson, M. Relative Aggregation and Random Quadrat Sampling. In *Computing Science and Statistics: Proceedings of the 26th Symposium on the Interface*; Sall, J., Ed.; Interface Foundation of North America: Fairfax, 1994; pp 415−418.

(21) Johnson, M. A.; Cheng, C.; Maggiora, G. M.; Lajiness, M. A Generalized Measure of Dependence with an Application to Molecular Similarity Analysis. In *Computing Science and Statistics: Proceedings of the 26th Symposium on the Interface*; Sall, J., Ed.; Interface Foundation of North America: Fairfax, 1994; pp 81−85.

(22) Conover, W. J. *Practical Nonparametric Statistics*; John Wiley & Sons: New York, NY, 1980.

(23) Hodes, L.; Feldman, A. Clustering a Large Number of Compounds. 3. The Limits of Classification. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 347−350.

CI9604605