Analysis and Display of the Size Dependence of Chemical Similarity Coefficients

John D. Holliday, Naomie Salim, Martin Whittle,* and Peter Willett

Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield,
Western Bank, Sheffield S10 2TN, United Kingdom

Received January 2, 2003

We discuss the size-bias inherent in several chemical similarity coefficients when used for the similarity searching or diversity selection of compound collections. Limits to the upper bounds of 14 standard similarity coefficients are investigated, and the results are used to identify some exceptional characteristics of a few of the coefficients. An additional numerical contribution to the known size bias in the Tanimoto coefficient is identified. Graphical plots with respect to relative bit density are introduced to further assess the coefficients. Our methods reveal the asymmetries inherent in most similarity coefficients that lead to bias in selection, most notably with the Forbes and Russell-Rao coefficients. Conversely, when applied to the recently introduced Modified Tanimoto coefficient our methods provide support for the view that it is less biased toward molecular size than most. In this work we focus our discussion on fragment-based bit strings, but we demonstrate how our approach can be generalized to continuous representations.

INTRODUCTION

Virtual screening of databases is becoming a popular method of preselecting available compounds for physical screening. Of the methods available, similarity searching is a relatively simple, fast, and cheap technique that involves the conversion of chemical structures to a machine-readable representation and subsequent comparison of these using one of many available similarity coefficients.[1] Different representations reflect different aspects of compound structure, and, coupled with the use of different similarity coefficients, these lead to a rich field of subtly different similarity measures. Fragment-based bit-string representations, or fingerprints, are widely used because of their computational efficiency and robustness. In these, each element of a binary string is set to 1 or 0 to indicate the presence or absence of some substructural element in the corresponding molecule. Most of the work that will be described here has used standard Unity fingerprints from Tripos,[2] which encode 992 bits in terms of both a fragment dictionary and hashed atom-path substructures, but analogous results are obtained if alternative fingerprints are used. The similarity coefficient most frequently used with these fingerprint representations is the Tanimoto coefficient. This is despite the fact that several workers have suggested that it is biased toward smaller molecules when used for diversity selection[3,4] and that other coefficients may be more appropriate for searching some compound collections or for use with different representations.[5] Since the number of features encoded by nonzero bits in a fingerprint generally increases with molecular size and complexity, a rough correlation exists between the size of a molecule and the number of bits set in its associated string. It is this correlation that forms the primary link between a similarity coefficient and any bias toward the size of molecule involved. However, size here may refer to any of several properties such as number of atoms, molecular weight, or molecular volume.

In this paper we analyze the effect of molecular size on the magnitudes of a range of similarity coefficients, when used for quantifying the degree of resemblance between pairs of molecules represented by fragment bit-strings. In the final section the approach is extended to include continuous representations.

A SIMPLE MODEL OF BINARY SIMILARITY COEFFICIENTS

The Model. When fragment-based bit-strings are used as the representation, the similarity coefficient between two structures is a straightforward function of a simple bit count.[1] For bit-strings of length *n*, we will suppose that *a* bits are set in the string for a query compound *A*, *b* bits are set for a comparison molecule *B*, and *c* bits are set common to both strings. From these basic definitions all of the well-known similarity coefficients $S(A,B)$ can be calculated.[1] Note that an alternative Boolean way of counting the bits is more convenient for some purposes and is sometimes used.[6] It is also convenient at this point to define the number of bits set in neither string, *d*, which can be computed from the quantities already defined, i.e.

$$d = n - a - b + c \tag{1}$$

Expressions in terms of these quantities describing the basic coefficients that we will study are collected in Table 1, where each type of coefficient is also labeled with a suffix.

For a similarity coefficient to give meaningful rankings of dissimilar molecules the value obtained for each match must be compared with the value at perfect similarity. For the Tanimoto coefficient this is given by $S_T(A,A) = 1$. A comparison string that has fewer bits set than the query string is clearly as similar as it can be if all of its bits are completely matched by the query. If, on the other hand, it has more bits set than the query string, then maximum similarity is obtained if the query string is completely matched by some subset of the new string. Thus, upper bounds for values of the similarity coefficient are obtained under conditions of maximum overlap, which we now investigate for the Tanimoto coefficient.

[*] Corresponding author e-mail: m.whittle@sheffield.ac.uk.

10.1021/ci034001x CCC: $25.00 © 2003 American Chemical Society
Published on Web 04/09/2003

Table 1. Similarity Coefficients and Limits to the Ratio R, Eq 10, under Each of Three Conditions Described in the Text[a]

coefficient	expression	R(i)	R(ii)	R(iii)
Tanimoto	$S_T = \frac{c}{a+b-c}$	>1	<1	>1*
Cosine	$S_C = \frac{c}{\sqrt{ab}}$	>1	<1	>1*
Squared Euclidean	$S_E = \frac{a+b-2c}{n}$	<1	>1	∼*
Simple Match	$S_{SM} = \frac{c+d}{n}$	>1	<1	∼*
Russell-Rao	$S_R = \frac{c}{n}$	>1	1	>1
Forbes	$S_F = \frac{cn}{ab}$	1	<1	<1
Kulczynski(2)	$S_K = \frac{1}{2}\left[\frac{c}{a} + \frac{c}{b}\right]$	>1	<1	>1*
Baroni_Urbani	$S_{BU} = \frac{\sqrt{cd}+c}{\sqrt{cd}+a+b-c}$	>1	<1	>1*
Fossum	$S_{FS} = \frac{n(c-(1/2))^2}{ab}$	>1	<1	>1*
Simpson	$S_{SI} = \frac{c}{\min(a,b)}$	1	1	1
Yule	$S_Y = \frac{nc-ab}{cd+(a-c)(b-c)}$	1	1	1
Stiles	$S_{ST} = \log_{10}\frac{[\lvert nc-ab\rvert-(1/2)n]^2}{ab(n-b)(n-a)}$	>1*	<1*	∼*
Pearson	$S_P = \frac{nc-ab}{\sqrt{nab(n-b)(n-a)}}$	>1	<1	∼*
Dennis	$S_D = \frac{nc-ab}{\sqrt{nab}}$	>1	<1	∼*

[a] Starred entries (*) mean that this limit is true more often than not according to simulations that are described in the text. Limits quoted as "∼" were found to be indeterminate.

For generality, we define bit densities for the two strings as

$$\rho_A = a/n; \quad \rho_B = b/n \tag{2}$$

We also introduce the ratio, $\alpha = b/a$, to signify the number of bits set in string B relative to the query A. Then, if the comparison string is less occupied than the query, we have $\alpha < 1$ and, for the condition of maximum overlap, $c = b$. Then for the Tanimoto coefficient

$$S_T(A,B) = \frac{b}{a} = \alpha \tag{3}$$

Conversely, if the second string has more bits set than the query, $a > 1$ and $c = a$, giving

$$S_T(A,B) = \frac{a}{b} = \frac{1}{\alpha} \tag{4}$$

These upper bounds define the first-order dependence of the Tanimoto coefficient on relative bit densities and have important consequences for the size distribution of molecules chosen in a similarity search. A molecule with lower relative bit density than a given query can never have a Tanimoto coefficient larger than that defined by eq 3, and a molecule with a higher relative bit density can never have a Tanimoto coefficient larger than that defined by eq 4. A corollary is that by choosing comparison molecules in a similarity search with a Tanimoto value greater than a given value, 0.7 for example, these limits ensure that their bit density ratio is within a range defined by $0.7 < \alpha < 1/0.7$.

The lower bound for a coefficient is obtained for minimum overlap and provided the occupancies are sufficiently low to satisfy $\rho_A + \rho_B < 1$, the value of this is zero: $c = 0$. The lower bound of the Tanimoto coefficient is consequently zero, but for some similarity coefficients the lower bound always depends on the values of a and b.

The expectation value of the similarity coefficient for random strings is of some theoretical interest.[7,8] This value delineates the watershed between random strings that are positively similar and those that are positively dissimilar, in the sense that their set bits are negatively correlated relative to random matching. The expectation value is readily calculated for bit strings of relatively low occupancy. For binomial trials with probability of success p and $q = 1 - p$, the mean and standard deviation over N trials is $\mu = Np$ and $\sigma = \sqrt{Npq}$. Setting $p = \rho_A$ these values give the mean and standard deviation over N trials for the probability that a single bit in B coincides with a set bit in A. There are b bits set in string B and for low occupancy we can treat them as independent trials so that $N = b$. Then the mean number of bit matches, c_R, on the basis of this model, is

$$c_R = b\rho_A = n\rho_A\rho_B \tag{5}$$

Similarly the standard deviation is obtained as

$$\sigma_R = \sqrt{n\rho_B\rho_A(1-\rho_A)} \tag{6}$$

Substituting the string densities ($\rho_A \sim 0.4$) used by Flower[7] into these expressions gives values that are in excellent agreement with plots for the probability of bit string matching presented in that paper and obtained using a potentially much more precise and comprehensive model. This agreement hence suggests that the assumption of independence is reasonable for the densities normally encountered in bit string representations.

Replacing ρ_B with $\alpha\rho_A$ gives expressions in terms of the relative occupancy and bit string density that are successful for low values of α if ρ_A is taken as the average occupancy. However, there is no reason that ρ_A should be chosen as the average density instead of ρ_B, and indeed this choice leads to large errors for high values of α. A better model is obtained by setting the mean bit density for the two compared strings

$$\frac{\rho_A + \rho_B}{2} = \rho_m \tag{7}$$

equal to the average occupancy ρ_0 over all N structures in the sample $\rho_0 = (1/N)\sum_{i=1}^{N}\rho_i$, thereby introducing the additional constraint $\rho_m = \rho_0$. With this substitution we have

$$c_R = n\alpha\left(\frac{2}{1+\alpha}\right)^2{\rho_0}^2 \tag{8}$$

Equation 6 can be similarly expressed in these terms. By using c_R in place of c in a similarity coefficient an estimate of the random expectation value is readily obtained. Putting $\rho_0 = 0.5$ to represent all possible fingerprints, with $n = 54$ the number of bits, and with $\alpha = 1$ to represent an average value (strictly, an integration is needed) we obtain a mean expectation value for the Tanimoto coefficient of 0.333 and a standard deviation of 0.172. These values correspond very well with a frequency distribution given by Godden et al.[8] for the complete distribution of the Tanimoto coefficient. However, this correspondence could well be fortuitous, as we would not expect our assumption of independence to hold for bit densities that span the whole range.

Comparison with Data. The interesting point about these bounds and expectation values for the Tanimoto coefficient is that they depend simply on the ratio of bit string occupancies or, indirectly, on relative molecular size. In Figure 1(a) they are compared with some real results taken from the Unity version 9.1 of the Dictionary of Natural Products (DNP)[9] encoded using Unity bit strings.[2] This was first treated to remove blank and duplicate records.[5] Two sets of roughly 100 compounds each were chosen at random from the DNP, and the first set was used as queries and the second as a matching set giving a total of $\sim 10^4$ values. Results are plotted in Figure 1(a) against the relative occupancy α and compared with limits calculated in the previous section. It will be noticed how few compounds are actually similar in the commonly accepted sense that S_T(A,B) > 0.7. Quite a few points fall on or near the lines delimiting the upper bounds, these points representing comparisons for which part of one structure is almost wholly contained by the other. This sample of compounds represents rather a small fraction (~0.1%) of the database, nevertheless, different random samples give consistent results and can be computed in a few minutes on an R10000 processor. The results that we present are typical of repeated trials.

Because values of $b < a$ are represented by points in the region $0 < \alpha < 1$ and values of $b > a$ represented by points $1 < \alpha < \infty$ Figure 1(a) is rather misleading, particularly since the Tanimoto coefficient is clearly symmetric with respect to interchange of a and b. The results are more evenly represented by plotting on a scale that measures the difference in bit-string occupancy relative to the total occupancy

$$\Delta = \frac{b - a}{a + b} = \frac{\alpha - 1}{\alpha + 1} \tag{9}$$

Viewed on this scale, Figure 1(b), the density of results for the Tanimoto coefficient become bilaterally symmetrical.

As a further check on our calculations of the random expectation values we generated two sets of 100 random bit strings with bit densities randomly chosen from a model symmetrical triangular distribution with mean 0.15. This distribution is a close match to that found for the DNP. As before, these were used to generate 10^4 values of the Tanimoto coefficient, and these are plotted against the corresponding values of Δ in Figure 1(c) along with the prediction of eq 8 with $\rho_0 = 0.15$. They demonstrate excellent agreement between our simple theory and simulation for bit densities in the range encountered in a real database. Comparison of Figure 1(b),(c) shows that compounds chosen to maximize diversity using the Tanimoto coefficient (i.e. comparison molecules with low values of S_T, say $S_T < 0.2$) occupy some of the same space on a Δ-plot as those that are randomly generated. This is particularly true near the center of the plot that relates to the comparison of bit-strings that are of similar density. This observation seems to support to the view that the Tanimoto coefficient may be a poor tool for discriminating between structures when low values are involved.[7] However, the bit-strings generated from a database of chemical compounds are selected from a relatively small and special population compared to the full set of combinatorial possibilities, and this renders any direct comparison

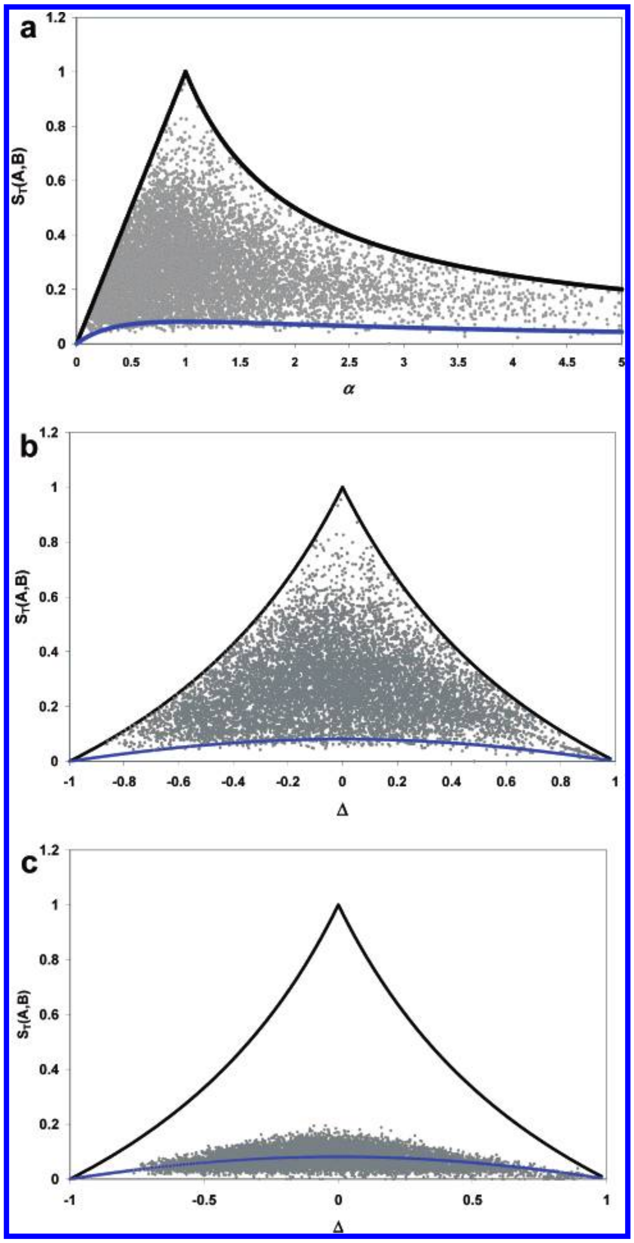

Figure 1. Comparison of results for the Tanimoto coefficient (Table 1) upper bounds from the simple model eqs 3 and 4 are shown in black; random expectation value from the simple model eq 8 are shown in blue. (a) Tanimoto similarities between two sets of ~100 randomly selected compounds from the DNP (grey circle); plotted against α. (b) Tanimoto similarities between two sets of ~100 randomly selected compounds from the DNP (grey circle); plotted against Δ (eq 9) with random expectation value for $\rho_0 = 0.176$. (c) Tanimoto similarities between two sets of 100 randomly generated bit strings with a density distribution centered on $\rho_0 = 0.15$ (grey circle); plotted against Δ (eq 9) with random expectation value for $\rho_0 = 0.176$.

with randomly generated strings unsound.

The plots shown in Figure 1(b),(c) essentially relate bit-matching similarity (S_T) to a comparison of the number of features present (Δ). Consequently, different regions of the plot are related to different aspects of the similarity between a set of compounds and plots for different sets of compounds may reveal differences in pattern that could be useful to know. This becomes apparent when the technique is applied to other databases, and in Figure 2(a),(b) we show the Δ-plots obtained with ~100 molecule subsets randomly chosen from the MDDR[10] and AIDS99[11] collections. Comparison of these plots with Figure 1(b) for the DNP shows that these sets have even fewer data points in the region of close similarity. For these databases, the range of similarities and the range of comparative sizes (inferred from the range of Δ) is narrower, and there are few close sub- or superstructures occupying the space near the boundaries. This would suggest that these databases contain a more diverse set of compounds than the DNP. This may be by design or simply because the DNP contains a large number of derivatives with high similarity to the parent compounds.[5] The plots also show that the DNP is actually a good database to use for the exploration of similarity space since points derived from it (Figure 1(b)) fill the gap between the calculated upper and lower bounds more uniformly than the drug collections. We will therefore continue to use it as the primary source of data for our comparative studies of similarity coefficients.

The figures presented so far show the results for similarities between two randomly chosen sets, but these plots can also be obtained for the comparison of a set with itself. In Figure 2(c), such results are plotted for a set of 101 Aminoglycoside Antibiotics taken from the DNP. These have previously been found to have a high degree of self-similarity by recall experiments.[5] In this case the individual points are symmetrically positioned because each compound in the group was used both as a query and comparison. More significantly, the self-similarity of the group is immediately apparent with perhaps only one or two compounds responsible for the unusually low values. Used in this way, the plots can thus give a rapid assessment of the overall similarity for a group of compounds using a particular coefficient. Comparative studies using the same group of compounds could help to suggest an optimal choice of similarity tools for that structure class.

Limits to the Upper Bounds. The simple model that we have introduced for the upper bounds of the Tanimoto coefficient can be applied to some other coefficients, such as the Cosine; but it is not so easily applied for many of the alternative forms of coefficient in Table 1. Similarity searching invariably involves ranking of the coefficient values relative to the value returned for the comparison of a molecule with itself. But for some coefficients, notably the Russell-Rao and the Forbes, this self-similarity value is itself dependent on the chosen query molecule.[5] For this reason, and in order to develop a robust and general method to probe the boundaries in a systematic way, it is helpful to consider comparisons between three molecules. The comparison of a target with at least two other structures is also needed to study any bias of a coefficient toward large or small molecules. We thus consider a target A associated with a string for which a bits are set, and two comparison compounds, B_1 and B_2 for which the bit counts b_1 and b_2 are subject to the condition that $b_2 < b_1$ (i.e., in molecular terms we would normally find that molecule B_2 is smaller than B_1). We suppose that B_1 and B_2 have c_1 and c_2 bits in common with A, respectively. There are then three possible distinct conditions depending on the relative bit density of the target structure and the comparison structures

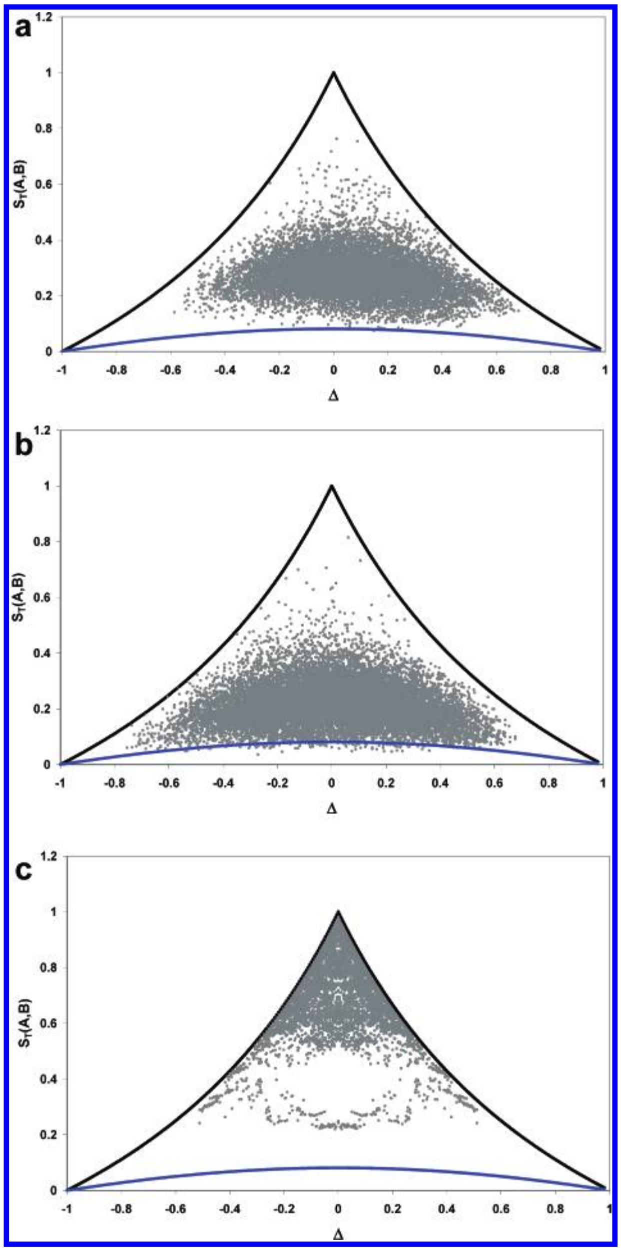

Figure 2. Tanimoto similarities between two sets of ~100 randomly chosen compounds (a) from the MDDR database plotted against Δ (eq 9) and (b) from the AIDS99 database plotted against Δ (eq 11). (c) Tanimoto similarities for a set of 101 Aminoglycoside Antibiotics taken from the DNP plotted against Δ (eq 9).

$$b_2 < b_1 < a;\ \max(c_1) = b_1;\ \max(c_2) = b_2 \tag{i}$$

$$a < b_2 < b_1;\ \max(c_1) = a;\ \max(c_2) = a \tag{ii}$$

$$b_2 < a < b_1;\ \max(c_1) = a;\ \max(c_2) = b_2 \tag{iii}$$

where we have also indicated the maximum values of the common count obtained when A is a perfect sub- or superstructure of one of the comparison compounds. Ranking of B_1 and B_2 in a similarity search depends on the relative sizes of the evaluated coefficients, and we express this as the ratio R

$$R = \frac{S(A, B_1)}{S(A, B_2)} \tag{10}$$

For the Tanimoto coefficient we have

$$S_T(A, B_1) = \frac{c_1}{a + b_1 - c_1};\ S_T(A, B_2) = \frac{c_2}{a + b_2 - c_2} \tag{11}$$

Substituting the maximum common counts for c_1 and c_2 using condition (i) and obtaining the ratio we get

$$R = b_1/b_2 \tag{12}$$

and since $b_2 < b_1$ we can assert that $R > 1$. For this condition, therefore, the compound with fewer set bits will have the lower upper bound. Note that since c_2 is a maximum, all of the bits for B_2 are contained in the string for A, and, thus, given a well-designed dictionary, the molecule B_2 must be substructure of A and will therefore be smaller. Similarly, for condition (ii) we obtain

$$R = b_2/b_1 \tag{13}$$

for which $R < 1$ and in this case the compound with the higher number of bits set will have the lower upper bound. In this case since c_1 is a maximum, all of the bits for A are contained in the string for B_1 and thus the molecule B_1 must be superstructure of A and therefore be larger. Taken together, these results do not imply any bias for the coefficient, rather that either smaller or larger compounds can return a lower coefficient depending on their size relative to the target. Comparison structures fulfilling condition (i) are those with negative Δ (eq 9) in our Δ-plots (e.g. Figure 1(b)) while those fulfilling (ii) have positive Δ. In fact for these conditions the ratio R is just related to the gradient of the leading edges, which are positive and negative, respectively.

Comparison structures that fulfill condition (iii) lie on either side of the origin in a Δ-plot. In this case B_2 is a substructure and B_1 is a superstructure of A. For these the ratio becomes

$$R = a^2/b_1 b_2 \tag{14}$$

It is not possible to put bounds on this expression algebraically since $a/b_1 < 1$ and $a/b_2 > 1$. However, any bias in this ratio is significant since it will reveal any inherent preference in the selection of large or small structures. It is therefore worth probing further.

To focus on this problem and for simplicity of notation we identify a, b_1, and b_2 with new variables y, z, and x, respectively. One approach would be to evaluate the expectation value $E\{y^2/xz\}$, which is an analytically accessible quantity that we would expect to be 1 if the Tanimoto coefficient is unaffected by size. Although this appears straightforward, it is the conditions that $x < y$ and $z > y$ that make the problem nontrivial. Suppose that the normalized bit distribution for a set of compounds is $\phi(s)$ for s bits, then by this definition, the probability density of choosing a bit string with y bits is

$$P(y) = \phi(y) \tag{15}$$

For continuous variables the probability density of choosing a second value x: $x < y$ from the same distribution can be shown to be

$$P(x|y) = \phi(x)/\int_0^y \phi(s)ds \tag{16}$$

Similarly the probability density of choosing a value z: $z > y$ is

$$P(z|y) = \phi(z)/\int_y^m \phi(s)ds \tag{17}$$

where m is the maximum value of s that has a nonzero value of $\phi(s)$. With these definitions we can write down an expression for the expectation value

$$E\left\{\frac{y^2}{xz}\right\} = \int_0^m dy \int_0^y dx \int_y^m \left(\frac{y^2}{xz}\right) P(y)P(x|y)P(z|y)dz \tag{18}$$

The reciprocal term in x leads to infinity when attempting to evaluate this integral for a simple flat distribution, and a sum over a discrete distribution is needed rather than integration over continuous variables. Although the integral may prove tractable for other distributions that vanish at 0 and m the problem rapidly becomes algebraically complex. It is simpler, more direct, and informative in this case to perform a simulation using appropriately chosen random variables to represent the values of x, y, and z. Random variables can be chosen from any given distribution using a rejection method[12] which we have adapted as follows. A random number, s ($0 < s < m$), is first chosen from a flat distribution. A second random number, r ($0 < r < 1$), is then chosen and compared with the desired distribution $f(s)$. If $r < f(s)$ the value of s is accepted, but if $r > f(s)$ it is discarded and the process is repeated. For our simulations, $f(s)$ is just a scaled version of the probability density $\phi(s)$ with a peak value of 1 to minimize the number of rejections needed. The conditional probabilities in eq 18 represent the fact that the target compound is chosen first in a similarity search, and as a result the order of choosing the values in this simulation is important. Consequently, it is the variable y, representing the target, that is chosen first using the rejection procedure. Values of x and z are subsequently chosen from the same distribution subject to the additional conditions $x < y$ and $z > y$. For each simulation 10^5 sets of random numbers representing x, y, and z were used. From each independent set the value of $R = y^2/xz$ was computed, and from these values, rather than evaluate an expectation value, the problem could be tackled directly by keeping a count of the number of times that the inequality $R > 1$ was true.

We have performed this simulation with $m = 1000$ values for: a flat distribution $f(s) = 1$; triangular distributions:

$$f(s) = s/a;\ s < a$$

$$f(s) = \frac{m - s}{m - a};\ s > a \tag{19}$$

specifically to investigate any effect of the skew, which can be controlled by the parameter *a*; normal distributions with a range of means μ and standard deviations σ; and normal distributions with means and standard deviations fitted to the average frequencies of set bits derived from real data sets. Results are reported in Tables 2−5, and in all cases, regardless of the type of distribution chosen, we find that the condition $R > 1$ is true more often than it is not. Given the rough correspondence between bit density and size, the conclusion is hence that there is an overall tendency to choose larger structures for the selection of compounds that are sub- or superstructures of the target. Conversely, if small values of the Tanimoto coefficient are sought, as in dissimilarity selection, there is an overall tendency to select smaller structures. The latter bias has been observed and discussed in earlier studies,[3,4,7] though whether this latter bias is, in fact, a real problem is open to debate. Indeed, recent studies highlighting the general increase in molecular complexity that occur during lead optimization[13,14] would suggest that a bias toward less complex, and hence generally smaller, molecules might well be a desirable trait of a coefficient. Our analysis places no restriction on the size of the Tanimoto coefficient, and the conclusions for similarity and dissimilarity are therefore complementary.

Other Similarity Coefficients. Algebraic quotients for the three conditions (i)−(iii) have been examined for a number of similarity coefficients, and the results are reported in Table 1. The results for *R*(i) and *R*(ii) are effectively the same for many of the coefficients given that the negative of the Squared Euclidean coefficient, $- S_E$, is frequently used in practice (so that results can be ranked descending), and for this the results for conditions (i) and (ii) would be reversed. More notable exceptions are the Russell-Rao and the Forbes coefficients, which are opposite to one another, and the Simpson and the Yule coefficients, which give ratios of 1 for all three conditions. For the Russell-Rao we find that $R = 1$ for condition (ii) implying that all superstructures to the target *A*, including itself, will return the same value of coefficient. However, we find that $R > 1$ for the substructures accessed by condition (i) indicating that smaller compounds return smaller values of the coefficient. Significantly, the Russell-Rao is one of the few coefficients for which condition (iii) gives an unambiguous algebraic result. In this case the result that $R > 1$ is a clear statement that large superstructures return a higher value of the coefficient than small structures and indicate in more general terms that this coefficient is biased toward large structures in a similarity search. The Forbes is entirely complementary to the Russell-Rao and by inverting the arguments we find that the Forbes is biased toward small structures in a similarity search. By this analysis, the Yule and the Simpson are the only coefficients completely unbiased toward the size of sub- or superstructures.

The results for condition (iii) have in all other cases been found to be algebraically indeterminate, and for these coefficients we have performed simulations as discussed above, after replacing eq 14 with the ratio appropriate to the coefficient concerned. These simulations were limited to the skewed triangular distributions of Table 3 and Normal distributions with extreme values from Table 5 (ID Alert/Daylight and AIDS/BCI). For many of these tests, values of *R*(iii) gave consistent results for all of the distributions tested, and these are recorded in Table 1. A few results remained indeterminate in the sense that some values were obtained that conflicted with others. For the Squared Euclidean, Simple Match, Stiles, and Pearson coefficients the result depended on the skew parameter, but the Normal distributions nevertheless gave *R*(iii) >1 most of the time for the Stiles and Pearson coefficients. For the Squared Euclidean and Simple match coefficients we found that *R*(iii) was greater than or less than 1 the same number of times within a 2% error when using the Normal distributions. The Dennis coefficient gave *R*(iii) < 1 more often than not for all the skewed triangular distributions tested, but the Normal distribution for AIDS/BCI (Table 5) opposed the trend. These dependencies on bit distributions indicate a more subtle relationship between molecular size and results obtained using these coefficients. In the next section we discuss size bias in more general terms that are not limited to sub- and superstructures.

Table 2. Relative Number of Times that $R > 1$ under Condition (iii) for the Tanimoto Coefficient (See Eq 14) for Samples Taken from a Flat Distribution

run	true	false
1	0.645	0.355
2	0.647	0.353
3	0.646	0.354

Table 3. Relative Number of Times that $R > 1$ under Condition (iii) for the Tanimoto Coefficient (See Eq 14) for Samples Taken from Skewed Triangular Distributions Eq 19 with Parameter *a*

a	true	false
0	0.610	0.390
100	0.571	0.429
200	0.573	0.427
300	0.578	0.422
400	0.585	0.415
500	0.598	0.402
600	0.613	0.387
700	0.625	0.375
800	0.626	0.364
900	0.642	0.358
1000	0.646	0.354

Table 4. Relative Number of Times that $R > 1$ under Condition (iii) for the Tanimoto Coefficient (See Eq 14) for Samples Taken from Normal Distributions with Mean μ and Standard Deviation σ

m	*s*	true	false
300	50	0.535	0.465
700	50	0.516	0.482
300	100	0.589	0.421
700	100	0.533	0.467
200	25	0.525	0.475
800	25	0.508	0.492
500	100	0.543	0.457

Table 5. Relative Number of Times that $R > 1$ under Condition (iii) for the Tanimoto Coefficient (See Eq 14) for Samples Taken from Normal Distributions with Means μ and Standard Deviations σ Fitted to the Average Frequencies of Set Bits Derived From Real Data Sets

data set	bit string	μ	σ	true	false
AIDS	BCI	81.60	28.66	0.586	0.414
AIDS	Daylight	235.70	106.83	0.600	0.400
AIDS	UNITY 2D	184.70	66.56	0.587	0.413
ID Alert	BCI	97.40	31.80	0.577	0.423
ID Alert	Daylight	270.30	111.59	0.595	0.405
ID Alert	UNITY 2D	208.30	71.11	0.582	0.418

GRAPHICAL ANALYSIS OF SIZE DEPENDENCE

Analysis of the Tanimoto Coefficient. The graphs shown in Figures 1(b),(c) and 2 plot separate points along the Δ-axis according to relative bit-string density, but the target structures, represented by the apex, themselves span a range of sizes and bit densities representative of the population sampled. So they do not make clear any dependencies on "absolute" size in the sense of size relative to the database population as a whole. To examine this it is possible to filter out just those structures of the second set with the highest and lowest bit-densities and color code the results accordingly. However, the statistics are potentially poor with this approach, and instead we preferred to choose new sets of 100 random compounds from subsets that represent those 1% of structures with the highest or lowest bit densities of the whole database. These were then matched with 100 randomly chosen query compounds chosen from the whole database. The results are superimposed on a Δ-plot for the full database obtained as before with two randomly chosen sets. A typical result for the Tanimoto coefficient is shown in Figure 3(a) with the high and low bit-density compounds color-coded red and blue, respectively. The extreme molecules reveal an interesting and significant asymmetry. Since dissimilar molecules are those for which $S_T(A,B)$ is smallest, it is clear from the figure that more low bit-density structures contribute to these regions of the plot (e.g. $S_T < 0.1$) than high. It is less clear that there is any bias for similarity selection since the few red and blue points appear in roughly equal numbers for $S_T > 0.7$. Given the rough correspondence between bit-density and molecular size the plot thus confirms that use of the Tanimoto coefficient will result in a bias toward small molecules in a diversity selection. Note that the results of the previous section, which refer to sub- and superstructures of the target compound, are equivalent to the points that occupy the outer edges of this plot. With this observation, Figure 3(a) supports our previous finding that $R > 1$ for this coefficient, in that red points ($\Delta > 0$) along the edge of the diagram have a higher average Tanimoto coefficient than the blue points ($\Delta < 0$) that appear along the opposite edge.

Recently, Fligner et al.[15] have introduced a Modified Tanimoto coefficient S_{MT} that includes a contribution for the unset bits in compared strings

$$S_{MT} = \left(\frac{2 - \rho_0}{3}\right) S_T + \left(\frac{1 + \rho_0}{3}\right) S_{T0} \tag{20}$$

where S_{T0} is the Tanimoto coefficient for absent features

$$S_{T0} = \frac{d}{n - c} \tag{21}$$

A Δ-plot for the Modified Tanimoto coefficient prepared in exactly the same way is shown in Figure 3(b). It can be seen immediately that there is a much smaller range of values for this new coefficient but also that values for small and large compounds appear more symmetrically placed. The plot indicates that roughly equivalent numbers of high and low bit-density structures are found at the lowest reported values (e.g. $S_{MT} < 0.3$) suggesting, as the authors proposed, that this coefficient does indeed overcome the bias of the normal Tanimoto coefficient. A surprising development is that the high bit-density structures, depicted in red, do not hug the upper boundary of the plot as they did for the ordinary Tanimoto coefficient. This requires some explanation since we have argued previously that this group touches the upper bound on the Tanimoto Δ-plot because it contains superstructures of the targets. Applying the same simple model that we used for the ordinary Tanimoto coefficient we obtain for the upper bound of the Tanimoto coefficient for absent features

$$S_{T0} = \frac{1 + \alpha - 2\rho_m}{1 + \alpha - 2\alpha\rho_m};\ \alpha < 1 \text{ and}$$

$$S_{T0} = \frac{1 + \alpha - 2\alpha\rho_m}{1 + \alpha - 2\alpha_m};\ \alpha > 1 \tag{22}$$

where ρ_m is the mean bit density for the two compared strings as defined in eq 7. Unlike the results for the ordinary Tanimoto coefficient, eqs 3 and 4, these limits cannot be expressed solely in terms of relative bit densities α but depend on the absolute bit densities of the two compounds, here expressed as the mean. Therefore, for the Modified Tanimoto coefficient, each individual comparison has a different upper bound. The upper boundary of the Δ-plot for the Modified Tanimoto is therefore a maximal line for all the target and comparison combinations that have been accessed. Comparisons with large structures involve a large value of ρ_m, and inspection of eq 22 shows that this leads to lower values of S_{T0} for $\alpha > 1$. Thus the upper bound of the Modified Tanimoto, eq 20, is reduced for comparison of high bit-density structures compared with the upper bound for comparison of lower bit-density structures for which $\alpha > 1$. In summary, the upper boundary that we observe on the positive side of the Δ-plot for this coefficient corresponds to comparisons of relatively low bit-density compounds; comparisons of higher bit-density compounds have a reduced upper bound and appear displaced from the edge.

Analysis of Other Coefficients. Size bias in the Tanimoto coefficient has been explained on a combinatorial basis to be a consequence of the fact that the fewer set bits there are in a fingerprint, the more possible fingerprints exist that will give near-zero values when a comparison is made.[15] Our study of upper bounds suggests that there is also another, numerical contribution that arises from the structure of the coefficient itself. Bias in other coefficients may not be so easily assessed or predicted. However, Δ-plots can be applied to any other similarity coefficient although it is sometimes useful to apply some linear scaling to adjust the range: such scaling cannot affect the ranking of returned values and thus does not affect the most important characteristic of a coefficient. Among other coefficients the Squared Euclidean is one of the most frequently used (Table 1). For a particular pair of molecules it has an upper bound of $(a+b)/n$ and for a group of molecules the maximum value is therefore max-$(a+b)/n$. The minimum value is zero obtained at perfect similarity. We have chosen to transform values to fit within a scale of 0−1 with unity the value at perfect similarity by setting

$$S_E^0 = 1 - nS_E/M \tag{23}$$

To achieve the aim of this exercise we need to set $M >$ max(a,b) for all comparisons and for our data set $M = 550$

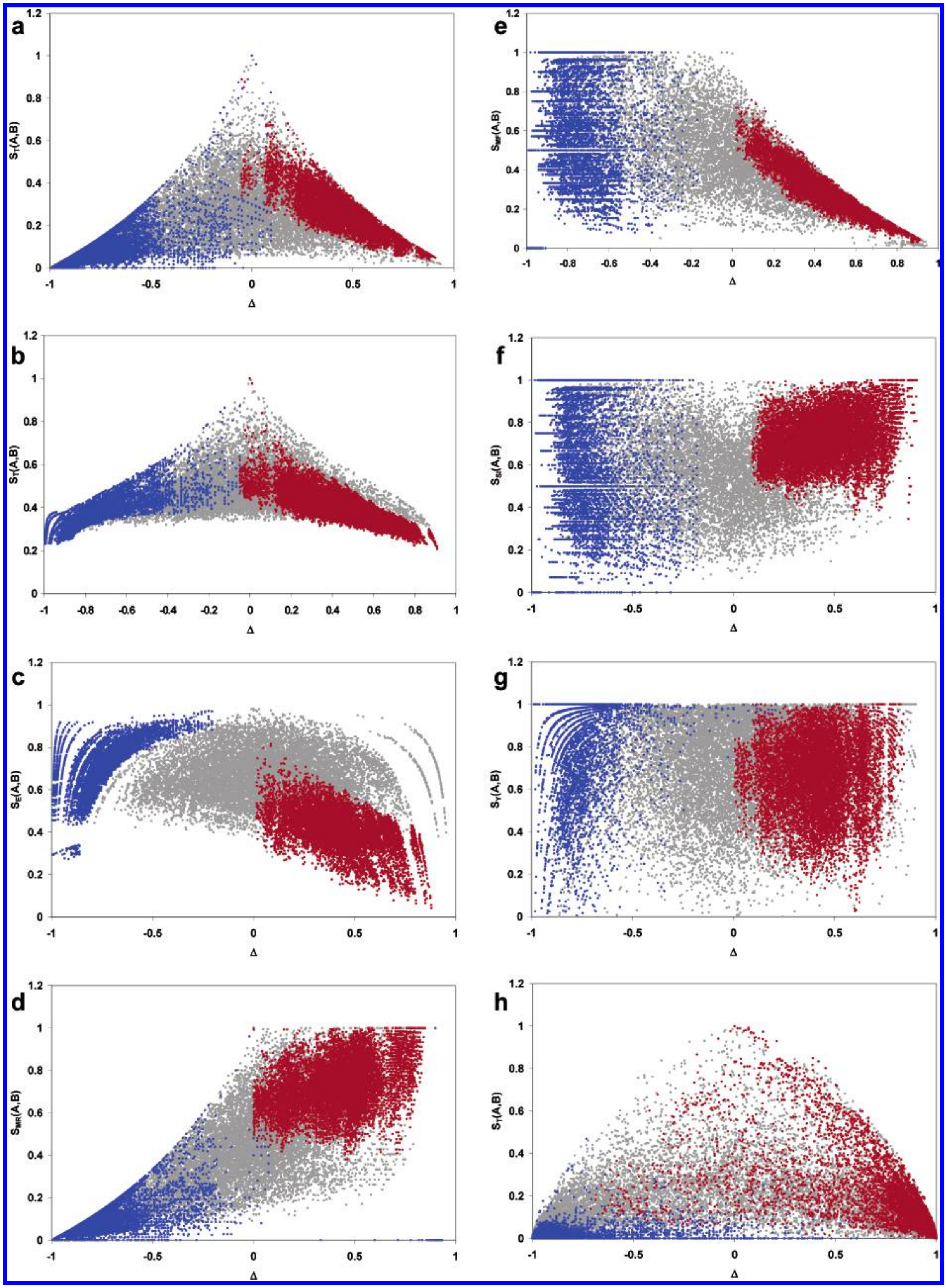

Figure 3. Plots of similarity coefficient values against Δ (eq 9) between two sets of ~100 randomly selected compounds from the DNP (grey circle), compared with similarities between a randomly chosen set and a set of molecules chosen from those with a bit-density in the lowest 1% (blue circle), and similarities between a randomly chosen set and a set of molecules chosen from those with a bit-density in the highest 1% (red circle), for (a) the Tanimoto coefficient; (b) the Modified Tanimoto coefficient eq 20; (c) the scaled Squared Euclidean coefficient with $M = 550$ eq 23; (d) the modified Russell-Rao coefficient eq 25; (e) the modified Forbes coefficient eq 26; (f) the Simpson coefficient (Table 1); (g) the Yule coefficient (Table 1); (h) for the Tanimoto coefficient computed using molecular holograms.

proved to be sufficient. A Δ-plot including overlays of structures chosen with bit-densities in the lowest and highest 1% of the range is shown in Figure 3(c). Given the correspondence with size, this plot suggests that the Squared

Euclidean is biased toward the selection of small compounds in a similarity search but biased toward large compounds in diversity selection. The latter effect has been documented by Dixon and Koehler[3] and can be linked directly to the dependence of the upper bound on bit density for this coefficient (or equivalently the lower bound for S_E^0).[16] Notably, the largest compounds are displaced from the upper boundary, and we find that, like the modified Tanimoto, the upper bound for this coefficient varies with the absolute bit density. Similarly, the observed upper boundary in the plot is due to structures with relatively low bit-density. By using the identity eq 1, the Simple Match coefficient, is easily shown to be a special case of eq 23 with $M = n$

$$S_{SM}(A, B) = 1 - S_E(A, B) \tag{24}$$

The Russell-Rao and Forbes coefficients (Table 1) were shown to have unusual and complementary behavior in our study of upper bounds. For the standard form of the Russell-Rao all comparisons are made with respect to the value at perfect similarity, which is $S_R(A, A) = a/n$. This varies for different target structures making comparison on a Δ-plot impossible. However, there is no loss in generality by replacing the Russell-Rao with a modified form, which has a value of 1 at perfect similarity

$$S_{MR}(A, B) = c/a \tag{25}$$

A Δ-plot for this coefficient is shown in Figure 3(d) and has a completely different shape to those presented previously, in being asymmetric about the center. From this figure it can be seen immediately that selection of similar compounds is strongly biased toward those with high bit-densities as suggested by the results in Table 1. Note that the result that $R = 1$ for condition (ii) for this coefficient is directly related to the zero gradient of the upper bound for positive Δ in Figure 3(d). By parallel arguments we can define a modified Forbes coefficient as

$$S_{MF}(A, B) = c/b \tag{26}$$

and we find that Δ-plots using this coefficient, Figure 3(e), are almost a reflection of those for the Russell-Rao and biased toward the selection of structures associated with strings of low bit density. The complementary nature of these coefficients is effectively combined in the Simpson coefficient (Table 1), which is now seen to be a mixture of the modified Russell-Rao and the modified Forbes that depends on the sign of Δ. A Δ-plot, Figure 3(f), suggests that it is relatively unbiased toward the selection of similar compounds and, as reported in Table 1, has the unusual property that the upper bound is independent of size. However, like the Tanimoto, the plot suggests that it remains biased toward the selection of small dissimilar compounds. The Yule coefficient, which is seen in Figure 3(g) to have a Δ-plot comparable to the Simpson in appearance, also has upper bounds independent of bit-density, but additionally it appears to be one of the least biased for dissimilarity selection.

Of the other common coefficients tested, the Cosine, Kulczynski(2) (which is seen to be another combination of the modified Russell-Rao and modified Forbes), Fossum, Pearson, and Stiles (scaled in much the same way that we have treated Squared Euclidean eq 23) display different shapes but are comparable to the basic pattern of the Tanimoto coefficient in crucial aspects. Each shows different degrees of bias toward low bit-density structures for low values of the coefficients and large superstructures appear close to the upper limit. In contrast, the Baroni/Urbani and Dennis have more in common with the modified Tanimoto in the sense that large superstructures are separated from the upper limit. A visual inspection suggests that, of all the standard coefficients tested, the Pearson is probably the least biased for dissimilarity selection followed closely by the Dennis. These conclusions are in broad agreement with extensive studies using target sets with different average bit densities.[16] None of the coefficients listed in Table 1, however, appears quite as even-handed toward large and small compounds as the Modified Tanimoto.

EXTENSIONS TO CONTINUOUS VARIABLES

So far we have limited our discussion to bit-strings, which are widely used in chemical similarity searching. However, many representations use a more general number system, and in these cases the similarity coefficients must be defined in a different way. For representations of length n using continuous variables we define the following quantities

$$a = \sum_{j=1}^{n} (x_{jA})^2; \quad b = \sum_{j=1}^{n} (x_{jB})^2; \quad c = \sum_{j=1}^{n} x_{jA} x_{jB} \tag{27}$$

where x_{jA} and x_{jB} are the jth elements of the representation vector for compounds A and B. The definitions already given for bit-string representations are actually just a special case of these. In many cases, the definitions of similarity coefficients given in Table 1 generalize to the continuous case by straight substitution, but this is not always true, in part because there is no counterpart to eq 1. It is also true that some representations may not give values of a and b that scale with molecular size and complexity and the interpretation of the associated Δ-plots cannot be extrapolated to these features. In these cases, if one wishes to examine size dependence in this way, it may be worthwhile redefining Δ (eq 9) directly in terms of some property such as molecular weight or number of atoms.

As an example of results using a continuous representation we have used molecular holograms, which are an extension of 2D fingerprints developed by Tripos for HQSAR studies.[2] Rather than using a binary string the fingerprint is represented as a string of integers corresponding to the number of times fragments are hashed into each bin. A Δ-plot for the Tanimoto coefficient, prepared in exactly the same way as Figure 3(a) using molecular holograms, is shown in Figure 3(h). The plot shares many characteristics of Figure 3(a) but is slightly different in shape. There remains a preponderance of compounds recording low values of b for small values ($S_T < 0.1$) of the coefficient, and so, as we would expect, this change of representation has not affected the size-bias of the results. For some of the coefficients of Table 1 that do have a counterpart for use with continuous representations the range and shape assumed by the Δ-plot can be very different to that found for binary representations. In some cases, what appears to be an obvious way of extending a similarity coefficient to continuous representations may in fact produce similarity values with a range that extends above and below the value obtained at self-similarity and which therefore cannot be ranked. Using molecular holograms, Δ-plots have been able to confirm this finding, which was originally deduced on mathematical grounds for the Russell-Rao, Forbes, and Kulczynski(2) coefficients.[5] For more mathematically complex coefficients the development of these plots should yield this information while circumventing the need to do some awkward algebra.

CONCLUSIONS

We have introduced a novel means of plotting the results obtained from similarity comparisons and, through the approximate relationship between bit-density and molecular size, used them to investigate the size bias of similarity coefficients using data from the DNP. The plots help to explain and understand some of the differences between coefficients that have been previously discussed in the literature. We have also investigated these differences by examining the limits to the upper bounds of 14 standard similarity coefficients. This study identified some unusual characteristics for a few of the coefficients which we have further explored using the graphical technique. We have particularly focused on the tendency of the Tanimoto coefficient to choose small compounds in dissimilarity selection, and an additional numerical contribution has been investigated that arises from the structure of the coefficient rather than any combinatorial preference. Our plots for the Tanimoto coefficient reveal this tendency as an asymmetry in the treatment of the highest and lowest bit-density compounds collected from a database. They also confirm that the Modified Tanimoto coefficient[15] is less biased than most toward diversity selection. We have also found a preference of the Squared Euclidean coefficient for the selection of small molecules seen in Figure 3c that accounts for the observed correlation of the performance of this index with the Forbes in our studies on the DNP.[5] The Russell-Rao and the Forbes have been shown to be complementary to each other and strongly biased toward selection of large and small compounds, respectively. The Δ-plots for these coefficients show a “first order” asymmetry in the overall shape that is distinct from all of the other coefficients. Extending the work to continuous variables is straightforward and allows the comparison of similarity space for different representations. The plots have an added use of mapping the self-similarity of small selections of compounds, a feature that may aid the search for effective similarity methods for compounds of a particular type or activity.[17] We have also demonstrated that by sampling large databases our methods can give a rapid assessment of the range of associated intrinsic similarities and thus some indication of the diversity of compounds available when probed by a particular technique.

ACKNOWLEDGMENT

We thank the following: Unilever Research for funding (M.W.), the Malaysian government for funding (N.S.); Barnard Chemical Information Ltd., Current Drugs Ltd., Daylight Chemical Information Systems Inc. and Tripos Inc. for software and data; and the Royal Society and the Wolfson Foundation for hardware and laboratory support. The Krebs Institute for Biomolecular Research is a Biomolecular Sciences Centre of the Biotechnology and Biological Sciences Research Council. The referees are thanked for some useful comments.

REFERENCES AND NOTES

(1) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(2) The Unity software packages are available from Tripos Inc. at URL http://www.tripos.com.

(3) 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.

(4) Lajiness, M. S. Dissimilarity-based Compound Selection Techniques. *Persp. Drug Discuss. Design* **1997**, *7/8*, 65−84.

(5) Whittle, M.; Willett, P.; Klaffke, W.; van-Noort, P. Evaluation of Similarity Measures for Searching the Dictionary of Natural Products Database. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 449−457.

(6) Holliday, J. D.; Hu, C.-Y.; Willett, P. Grouping of Coefficients for the Calculation of Intermolecular Similarity and Dissimilarity Using 2D Fragment Bit-Strings. *Combin. Chem. High-Through. Screening* **2002**, *5*, 155−166.

(7) Flower, D. R. On the Properties of String-Based Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **1997**, *38*, 379−386.

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

(9) The *Dictionary of Natural Products* database is available from Chapman & Hall/CRC at URL http://www.crcpress.com/

(10) The MDDR database is available from MDL Information Systems at URL http://www.mdli.com/dats/pharmdb.html.

(11) The AIDS database is available from the NCI/NIH Developmental Therapeutics Programme at URL http://dtp.nci.nih.gov/

(12) Press: W. H., Teukolsky, S. A, Vetterling, W. T., Flannery, B. P. *Numerical Recipies in C*, 2nd ed.; Cambridge University Press: 1994.

(13) Hann M. M., Leach, A. R., Harper, G., Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 856−864.

(14) Oprea, T. J., Current trends in lead discovery: Are we looking for the appropriate properties? *J. Comput.-Aided Mol. Des.* **2002**, *16*, 325−334.

(15) Fligner, M. A.; Verducci, J. S.; Blower, P. E. A Modification of the Jaccard-Tanimoto Similarity Index for Diverse Selection of Chemical Compounds Using Binary Strings. *Technometrics* **2002**, *44*, 110−119.

(16) Salim, N. Analysis and Comparison of Molecular Similarity Measures. Thesis submitted to the University of Sheffield, 2002.

(17) Sheriden, R. P.; Kearsley, S. K. Why do we Need so Many Chemical Similarity Search Methods? *Drug Discov. Today* **2002**, *17*, 903−911.

CI034001X