

On the Properties of Bit String-Based Measures of Chemical Similarity

Darren R. Flower[†]

Department of Physical and Metabolic Sciences, ASTRA Charnwood, Bakewell Road, Loughborough, Leicestershire, UK, LE11 5RH

Received November 26, 1997

With the growth of interest in database searching and compound selection, the quantification of chemical similarity has become an area of intense practical and theoretical interest. One of the most widely used methods of measuring chemical similarity is based on mapping fragments within a molecule as bits within a binary string. We present empirical results which suggest that bit strings provide a nonintuitive encoding of molecular size, shape, and global similarity. Other results, this time statistical in nature, suggest that the observed behavior of bit string-based searches have a large nonspecific component. On this basis, we question whether bit string-based similarity methods possess all the features desirable in a quantitative chemical distance measure or metric and suggest that there are instances when they may not be the most appropriate tool for searching or segregating chemical structures.

INTRODUCTION

The definition and analysis of chemical similarity has long been an active area of study in theoretical and computational chemistry. With the growth of interest in both directed database searching and compound selection for high throughput screening (HTS), the quantification of chemical similarity has become a subject of great practical significance. Likewise, study of the subject has gained a hitherto unrecognized relevance.

Currently, there seems to be no unambiguous, or even generally agreed, quantitative, or even qualitative, definition of chemical diversity. We know it when we see it; or, more usually, we can recognize its absence. In formulating any description of quantitative chemical distance, therefore, one is obliged to make approximations and to use heuristically derived solutions.

In the context of chemical structure searching, many ways have been suggested of measuring the degree of resemblance between molecules. How might one make sense of them? Fundamental to any attempt to address the issue of chemical diversity is the following question: what is the most appropriate representation of a molecule for use in the assessment of chemical similarity? Clearly, this will depend on the context, or objective, we have in mind. Many different ways have been used to represent chemical structures leading to many different approaches to assessing their similarity. These include methods based on three-dimensional representations, such as those utilizing projected properties—ASP¹ and related methods,² for example—or those based on so-called pharmacophore diversity.³ Two-dimensional approaches are, perhaps, even more numerous; the term 2D is a convention, as it is in general the properties of the molecular graph which are of interest, and not its pictorial representation in the plane. There have also been

recent attempts to reduce the dimensionality of the problem still further by considering the measured biological properties of compounds as the basis for diversity analysis.^{4,5}

What data there are suggests that a 2D description can be adequate for most purposes,^{6,7} and it is 2D based similarity measures that are by far the most commonly used in both similarity searching and in clustering or selection procedures. It is the properties of this kind of representation which are addressed here; other approaches will not be discussed further.

Most approaches which use a 2D structural description to address the problem of quantifying chemical similarity fall into either of two broad classes. First, there are parametric approaches. These may, for example, be based on some statistical analysis of molecular descriptors.^{8,9} In this regard, such methods show considerable conceptual similarity to QSAR methods. The descriptors may take the form of measured or computed physical properties—such as LogP—or structural invariants—such as topological or constitutional indices.

Second, there are those approaches which are based on some count of shared features. Such features include atom or element types, bonds, topological torsions, etc.,^{10,11} or bits set in a bit string. The raw counts become measures of chemical distance when combined in some form of similarity coefficient. A common class of such similarity measures are the association coefficients. There are many such coefficients, but all involve some normalized form of the dot product of vector representations of the molecules being compared, with coefficients differing only in the form of normalization used. The Tanimoto coefficient is perhaps the most widely known example of such an association coefficient. It has been widely used as an effective measure of intermolecular similarity in both the clustering and searching of databases. Similar measures of association include the cosine coefficient. Holliday et al.¹² have shown that such coefficients are typically highly correlated.

[†] Tel: 44 (0)1509 644882. Fax: 44 (0)1509 645576. darren.flower@charnwood.gb.astra.com.

The Tanimoto coefficient τ , takes the form

$$\tau = N_{AB}/N_A + N_B - N_{AB}$$

where N_A is the number of features in A, N_B is the number of features in B, and N_{AB} is the number of features common to A and B.

As one might expect, hybrid methods have also been proposed which make use of a combination of these two approaches.^{13,14}

Of all the similarity measures outlined above, perhaps the most used—at least in day to day work at the bench, so to speak—are those based on comparison of common bits in compared bit strings. It is also these approaches which have been implemented in commercially available software packages.^{15–17} In the following, we explore some of the properties of bit string-based methods for quantifying chemical similarity.

ENCODING STRUCTURAL INFORMATION IN BIT STRINGS

A common strategy to increase the efficiency of database searching has been to encode the structure of a molecule as a pattern of bits set within a bit string or fingerprint. The most common encoding of bit strings has been based on a 2D description. The process involves splitting a molecule up into fragments. If a particular fragment is present, then a corresponding bit is set in the bit string. However, it is the identity of a fragment which determines if a bit is set and not its quantity. A fragment can be present once or 100 times, and it would still only set one bit. It is the number of different types of fragment that determines the number of bits set in a fingerprint. Generally, these fragments tend to be either specific functional groups or substructures (carboxylic acids, say, or a certain type of ring system) or different linear atom paths through the corresponding molecular graph (see Figure 1).

The first of these ways of setting bits is closest to the group screens used in the MDL MACCS system.¹⁵ The path based approach is implemented by Daylight Chemical Information.¹⁶ By default, all paths through the molecular graph of length 1 to 8 atoms are found. Bits corresponding to each possible type of path are set if present. The resulting bit string is then *folded* to reduce storage requirements and speed searching. Note that these are default options and can be changed by the user.

The UNITY system from TRIPOS uses a similar algorithm to that of Daylight and also uses the Tanimoto coefficient as its similarity index. The examples presented below are all results taken from UNITY¹⁷ (version 2.7).

The structural key approach uses representations of structures which reflect the presence or absence of predefined functional groups and is, therefore, dependent on a predefined list of structural fragments. The main alternative approach involves the *hashing* of unique structural paths. The advantage is said to be that the fingerprints generated are characterized by the nature of the chemical structures in the database rather than as a function of the fragments in some predefined list which often leads to superior database screening and more efficient 2D and similarity searching.

Within UNITY, fingerprints are generated using a combination of the keyed and hashed fingerprinting approaches.

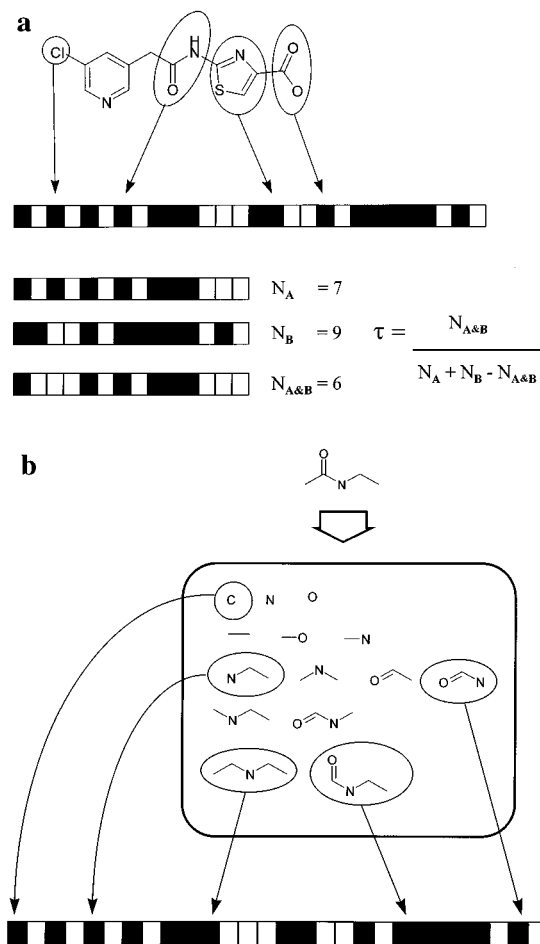


Figure 1. Encoding chemical structure as a bit string. (a, top) The concept of encoding is illustrated in a general way. Fragments within a molecular structure map to particular bits set within a bit string. The lower part of the diagram shows how comparison of two different bit strings leads to counts of set and common bits, which are, in turn, combined to give the Tanimoto coefficient. (b, bottom) A more explicit, but still conceptual, representation of how bits are set in UNITY and other chemical information systems. A molecule is decomposed into a set of atom paths of all possible lengths. Each of these paths is then mapped to a bit set in a corresponding binary string. A simple structure is used to allow all paths to be enumerated and shown easily.

The hashed fragments encode all unique linear, branched, and cyclic fragments, including overlapping fragments. Each fragment is then mapped to a pseudorandom integer in the range 0 to $(2^{31}-1)$ using a cyclic redundancy check algorithm. The integer generated by this algorithm is unique and reproducible for each unique structure. The hashing then occurs by folding the pseudorandom integer for a particular compound into the bin range defined. Since the length of the bit string is considerably smaller than the integer to which the molecule is mapped, it is possible for different structures to hash to the same bin.

ENCODING SIZE AND SHAPE

Experience of performing similarity searches using UNITY and other systems as well as consideration of the nature of the method itself suggests that a bit string based representation may not capture all aspects of molecular structure.

Consider the question of molecular size. Figure 2 gives an artificial but illuminating example of size effects in bit

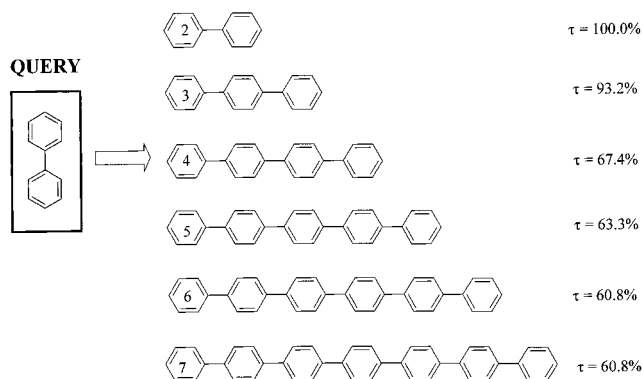


Figure 2. Using bit strings to encode molecular size. A biphenyl query is compared to a series of analogues of increasing size. The Tanimoto coefficient, which is shown next to the corresponding structure, decreases with increasing size, until a limiting value is reached. Increasing this series *ad infinitum* will not decrease the Tanimoto value further. This is because the maximum path length is fixed, and the larger molecules already generate the largest possible path. The setting of a bit indicates the presence or absence of a path and not the number of a such paths in a molecule.

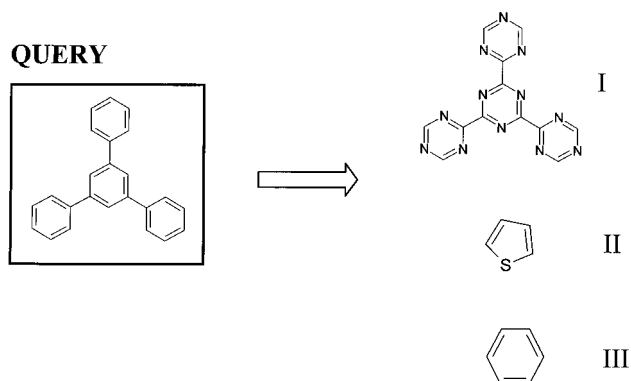


Figure 3. Using bit strings to encode molecular shape. A molecule is compared to three others. The Tanimoto value versus the query of the three target compounds is **III**: 22.0%; **II**: 7.3%; **I**: 5.7%. The lack of paths common to the query and **I** results in a low value of the Tanimoto for molecules with the same overall shape. Bits are set on the basis of local structure and atom identity, leading to a poor encoding of the overall shape of a molecule.

string-based similarity searching. Using a biphenyl group to search against a set of multiphenyl structures of increasing size, we see that for [Ph]₁ to [Ph]₆ there is some discrimination of size. Beyond that, [Ph]₇ and larger, there is no discrimination. Indeed, in extending the series *ad infinitum* we observe the same set of bits are set irrespective of size. Why? Bits are set only once regardless of how many times the corresponding path appears in the molecule. Because the maximum length of path has already been exceeded, no new types of path are found and hence no new bits are set.

Consider now molecular shape. Figure 3 shows four molecules: what is the similarity, in terms of the Tanimoto coefficient, of the three targets to the query? How do the three rank and what are their absolute similarity values? Try to estimate this for yourself before reading further.

The possibly surprising result is (**III**: 22.0%; **II**: 7.3%; **I**: 5.7%). Because of its pattern of heteroatom substitution, molecule **I** has very few paths in common with the Query. This results in a lack of measurable similarity, despite these molecules having the same overall shape.

Bits are set on the basis of local structure and atom identity, whether we look at functional groups or paths.

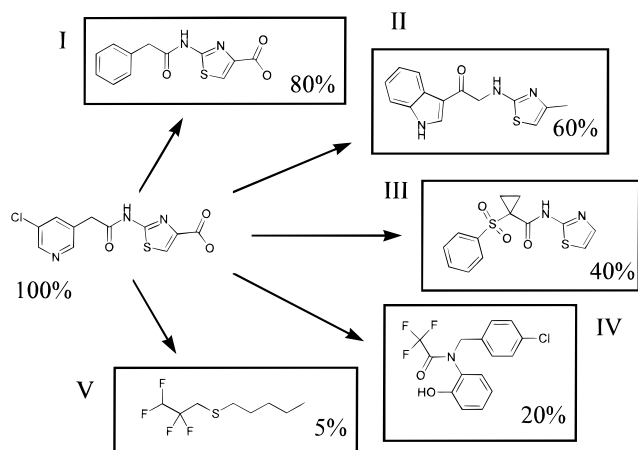


Figure 4. Molecular similarity at a range of Tanimoto coefficient values. Using a database of 50 000 randomly selected compounds, a query molecule is compared first to itself—giving a Tanimoto coefficient of 1.0 or 100%—and then to other molecules sampling a range of Tanimoto values. No molecule near zero was found, and so an arbitrary value of 5% is chosen instead.

Because these methods use only local information there is only a poor encoding of global structural properties, such as the overall size and shape of a molecule. This is also true to a large extent, though we show no example here, of global molecular symmetry.

DO BIT STRINGS ENCODE A QUANTITATIVE MEASURE OF CHEMICAL SIMILARITY?

The discussion above highlights potential problems with using bit string-based measures of structural similarity. How do they perform in more realistic examples?

Consider the example given in Figure 4. A database of 50 000 randomly chosen compounds was constructed and searched using the query molecule. Figure 4 shows hits at different similarity values as expressed by the Tanimoto coefficient. The query itself represents a value of 100%, while a compound with 5% similarity was chosen rather than 0%. This is because only compounds with exceptionally exotic chemistry are capable of this low level of similarity; indeed, few if any scores at this level could be found. Equally, the 80% target is an artificial hit created from the query; this was necessary since no database entry was found to exceed 60% similarity to the query.

Armed with this information, what is the similarity, in terms of the Tanimoto coefficient, of the three targets to the query molecule in Figure 5a? How do the three rank and what are the absolute values of the similarity? Try this for yourself before reading further.

The surprising result is (**III**: 20.252%; **II**: 20.252%; **I**: 20.252%). Or put another way, they are all equally similar and at a level which suggests they are very dissimilar. This is perhaps most surprising for compound **I**, but again this results from the pattern of heteroatom substitution and its effect on the number of paths common to the two molecules. Now consider Figure 5b and repeat the ranking process. This time the molecules are all more obviously similar. The reader may, or may not, be surprised to learn that again all of the three molecules have the same Tanimoto value relative to the query, but this time at the 50% level.

Whatever your own answers were to these puzzles, try them on other people in order to assess the validity of the

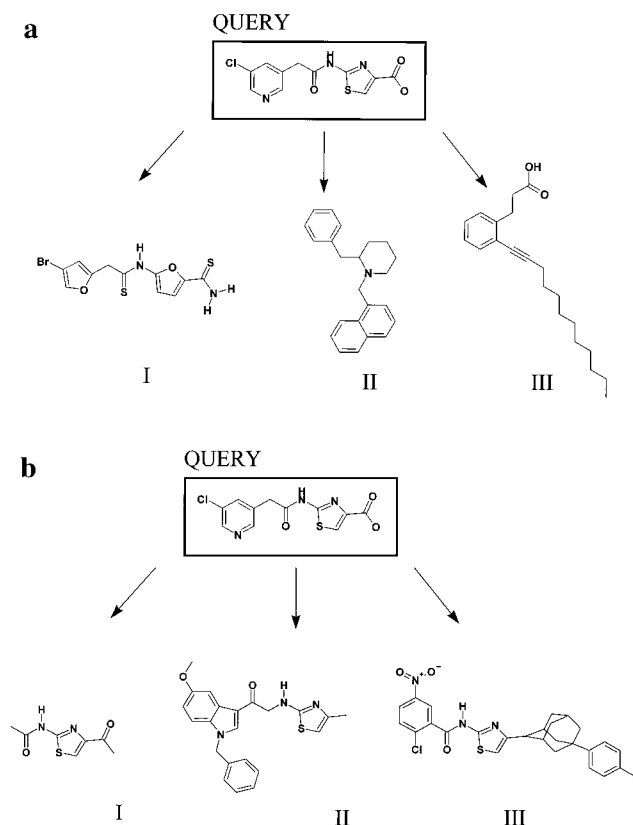


Figure 5. Judging similarity at low Tanimoto values. (a, top) A query molecule is compared to three other, obviously distinct compounds. Contrary to naïve expectation all compounds have a value of 20.252% compared to the query. (b, bottom) A similar search to that shown above but at a Tanimoto value of 50.0%.

point we make here. Bit string-based measures of global similarity can give nonintuitive results, particularly at low similarity levels.

THE BEHAVIOR OF SEARCHES

Turning from individual pairwise comparisons of compounds to the behavior of searches as a whole, consider the set of molecules shown in Figure 6. These compounds represent molecules of increasing structural complexity. The first three are perhaps of the complexity seen in starting materials or chemical reagents, perhaps to be used in a combinatorial chemistry experiment. Compound **V** is of the complexity of an interesting hit from HTS, while **VI** has the extreme complexity typically associated with natural products. Using each compound in turn to search the same 50 000 compound database as above, what are the distributions of Tanimoto coefficients of all molecules in the database? These distributions are plotted out in Figure 6.

At first glance, what appears to be happening in these searches is that the larger and more complicated the queries become, the more the distributions of similarities are shifted to the right. That is to say that the average similarity appears to increase, implying, for example, that there are more compounds such as **VI** than compounds **V** or **IV**. Visual examination of the contents of the database suggested this is not the case. There were no compounds with anything like the overall size and chemical complexity of **VI**, while there were significant numbers of molecules similar in size and complexity to **VI** or **V**.

Moreover, as the query grows the corresponding distribution flattens out. This suggests that the larger queries are more discriminating since the Tanimoto values are more evenly spread over a greater range. A narrow range of values, on the other hand, would suggest that a query is not able to distinguish between molecules because most compounds have the same, or very similar, values. A wider, more even spread places compounds into a greater number of different classes.

Both of these implications also seem counterintuitive, again for obvious reasons. One explanation for this behavior is that the effects we see here result from general properties of bit string matching unrelated to the chemical information they encode. Specifically, as structures get larger and more complicated they tend to set more bits: can this saturation of the bit string give rise to these effects? To answer this, we need a probabilistic statistical model for bit string matching.

A PROBABILITY DENSITY FUNCTION FOR BIT STRING MATCHING

Within a collection of chemical structures, each different molecule will, typically, set a different pattern of bits within a bit string. This is the essence of the ability of bit strings to encode structure and the basis by which they provide a similarity measure. The set of strings corresponding to a set of compounds will, to a large extent, be different not only in which bits are set, but also how many. The number of bits set per string will form a distribution. Figure 7 shows such distributions for three different databases.

The number of ways that bits in a bit string can be set to yield a given total number of bits set can be huge. Assuming a 1000 bits in a bit string, then the total number of different ways of setting half the bits is approximately 2.7×10^{299} . This is a large number—the number of protons in the visible universe is believed to be only about 10^{80} .

Leading on from this discussion and aiming to address the issues raised at the end of the last section, we can use this approach to help us understand the statistics of matching bit strings. This question is analogous to that of comparing protein and nucleic acid sequences, and there have been many attempts at such an analysis. However, none were suitable for application to the present problem, and so we turned to the mathematics of basic combinatorics for a solution.

As a first step, a probability density function is required for the matching of bit strings purely by chance; this random model assumes each different pattern of set bits is equally probable. For generality, we need to know what the probability of a given score (*i.e.*, the number of equivalent bits set in both strings) given the total number of bits (*i.e.*, the length of the fingerprint), and the number of bits set in the two strings to be compared—the query string used to search and the target string.

Applying standard results for calculating combinations¹⁸ yields the following

$$P(s) = \frac{T_s! Q_s! (N - T_s)! (N - Q_s)!}{s! N! (T_s - s)! (Q_s - s)! (N - Q_s - T_s + s)!}$$

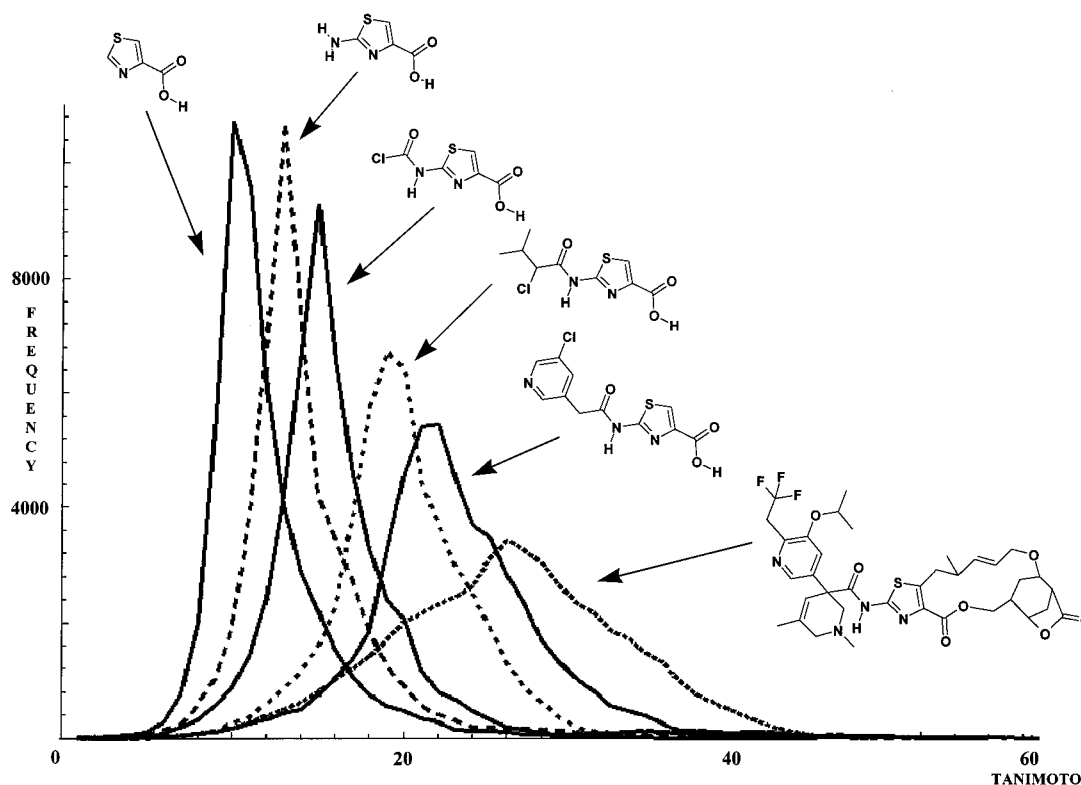


Figure 6. Performance of bit string-based similarity searches. The distribution of Tanimoto coefficient values found in database searches with a range of query molecules of increasing size and complexity.

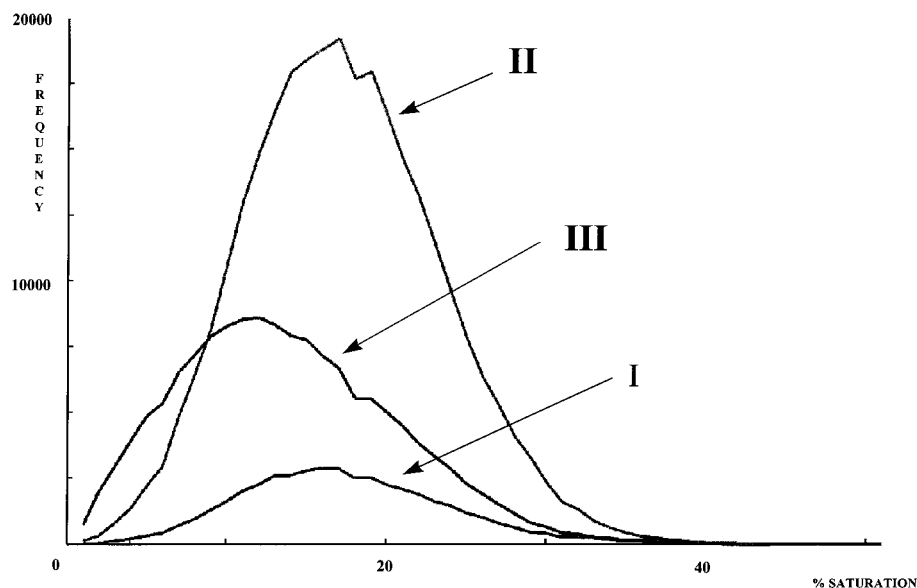


Figure 7. Distributions of bit string saturation. Distributions of the relative saturation of bit strings in three databases. By saturation, we mean the proportion, or percentage, of bits that are set in a bit string. The distribution of bit string saturation for the 50 000 compound test database shown labeled **I**. This distribution peaks below 20% indicating that most compounds in the database are relatively simple. For comparative purposes, distributions for two large composite databases also shown. Database **II** contains 300 000 compounds sold for screening. Database **III** contains 120 000 structures of mixed type: both reagents/starting materials and compounds for screening. Consequently, this distribution peaks at a lower value.

where s is valid in the range

$$\text{MAX}(0, T_s + Q_s - N) \leq s \leq \text{MIN}(T_s, Q_s)$$

s is the score of common bits sets, N is the length of the string in bits, Q_s is number of bits set in the query, and T_s is the number set in the target string.

Figure 8 illustrates some properties of this probability distribution, showing how it varies with Q_s for fixed values

of T_s and N . In this particular example $T_s = 200$ and $N = 512$. As Q_s increases the distribution shifts to the right, the most probable score increasing in proportion to Q_s . As expected, the spread of probabilities, and therefore the standard deviation, is greatest when values of T_s and Q_s are similar. It is only here that the intersection of T_s and Q_s can become equal to the union of T_s and Q_s in the Tanimoto expression and thus give the widest range of values for the

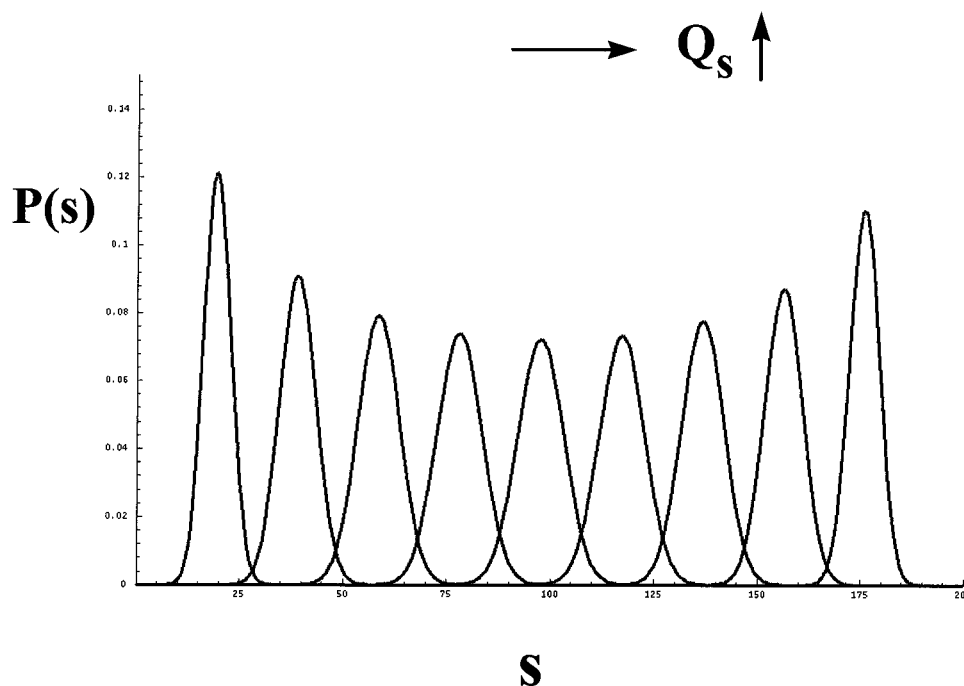


Figure 8. Properties of the probability distribution for bit string matching. Some properties of the random probability distribution are illustrated, showing how it varies with Q_s for fixed values of T_s and N . In this example $T_s = 200$ and $N = 512$. Each of the Gaussian-shaped curves represents a probability distribution for a different value of Q_s . For the first curve Q_s is set to 50, and this value is incremented by 50 for each subsequent curve. As Q_s increases the distribution shifts to the right: the most probable score increasing in proportion to Q_s . As expected, the spread of probabilities, and therefore the standard deviation, is greatest when values of T_s and Q_s are similar. It is only here that the intersection of T_s and Q_s could possibly become equal to the union of T_s and Q_s in the Tanimoto expression and thus give the widest range of values for the coefficient.

coefficient. It is also worth noting at this point that for highly saturated or dense bit strings there will be a minimum number of bits that can be in common, and for sparse strings there will be a maximum.

This model for the probability of bit-string matching can be used to simulate the characteristics of database searching. Given the distribution of bits set for entries in a database and the bits set in a query string, it should be possible to calculate, using the above model, the distribution of scores and Tanimoto coefficients if matching occurs at random. Figure 9 shows the results from such a simulation. Figure 9a shows the distribution of scores for increasing Q_s . Figure 9b shows the equivalent distributions expressed as Tanimoto coefficients.

These simulations suggest that many properties of bit string matching can be explained by a random model. A possible interpretation of this is that matching contains a significant nonspecific component related to the saturation of bit strings. Small molecules set few bits. Typically, as molecules get bigger, more bits are set. As our simulation shows, this leads to an increased intrinsic probability of matching with a high score. Another way of expressing this is to say that the absolute value of a score, or Tanimoto coefficient, is dependent on the saturation of bit strings and therefore on the precise structures under consideration. Again this might lead the cynically minded to question the validity of bit string based representations as a basis for chemical similarity measures. One way to increase the sensitivity and meaningfulness of searches is to use the probability density function above to provide p-values, say, expressing the probability that two strings have matched at random.

DISCUSSION

We have presented a number of empirical observations, supported by the results of statistical simulations, which highlight some interesting features of bit string-based measures of chemical similarity.

In hindsight, it is seemingly obvious how and why bit string based similarity measures behave as they do. Initially however, we found our observations surprising, and they might lead the sceptic to question whether bit strings are able to provide an intuitive encoding of size, shape, or overall molecular similarity. In turn, this sceptical individual might also wish to question, on this basis, the potential usefulness of the technique. It is perhaps more reasonable to say that the method is not completely general, and there may be times when it is not the most appropriate search or clustering tool.

At high values of Tanimoto coefficients using UNITY fingerprints, molecules are undoubtedly very similar; indeed, above a value of about 0.8 they are often virtually identical. This underlies the observation that molecules identified by searches at, or above, 0.85 will have a biological activity similar to that of the query structure.¹⁷ Rather than identify real similarities which we might feel were unexpected but “interesting”, the near identity of compounds at these Tanimoto values equate, essentially, to spotting trivial analogues: methyl, ethyl, futile, to coin a phrase.

At very low values of the Tanimoto coefficient, compounds are indeed typically, yet not exclusively, dissimilar, at least in the sense that they have few features in common; but here bit string based measures are only poorly quantitative.

Tanimoto coefficients of intermediate value—say, for argument’s sake, between 0.2 and 0.8—show some gradua-

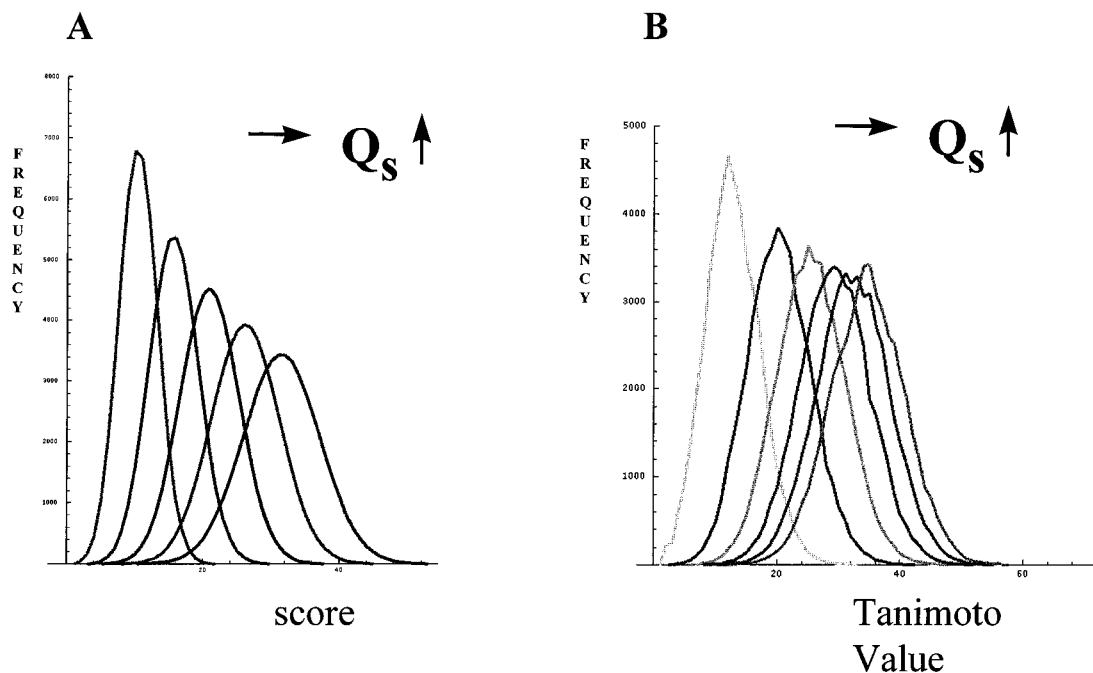


Figure 9. Statistical simulation of bit string matching for a whole database. Results from a statistical simulation of bit string matching within a database of 50 000 compounds assuming a random model. (a) The distribution of raw score values for increasing Q_s . (b) Equivalent distributions expressed as Tanimoto coefficients. It should be noted that the curves correspond to ascending values of Q_s and are in the same order in both graphs.

tion in the similarity of the matches, but looking at the compounds **II** and **III** in Figure 4, for example, it is open to question whether the difference between them, in terms of their respective relative similarity to the query, is really as great as 20%.

However, overall the fingerprint approach gives only a semiquantitative measure of similarity with a significant random or nonspecific component, a nonintuitive encoding of size and shape, and Tanimoto values which are not absolute but reflect the level of saturation in the strings being compared. We can say they are different but not by how much; occasionally they are actually surprisingly similar (Figure 5a). We might argue from this that bit string based methods are appropriate for clustering very similar molecules together but less appropriate for performing dissimilarity searches.

There is a corresponding phenomenon in the world of protein sequence analysis—the so-called “Twilight Zone”—where below a value of percentage amino acid identity relationships based on aligned protein sequences lose statistical significance.¹⁹ The most extreme manifestation of this behavior are so-called structural superfamilies, where a group of proteins related by their obviously similar three-dimensional structures exhibit little or no significant sequence similarity.^{20,21} This is a situation not unlike that seen in Figure 5a.

We have seen above how poor the encoding of molecular size and shape is by conventional bit string representations of chemical structure. Similarity of global shape is of particular concern. Without some explicit definition of overall geometry it is difficult to see how a method based on small fragments can deal with this problem, unless we allow such fragments to become arbitrarily large and complex. Encoding size is addressed more easily. The most obvious modification we can make is, as suggested by Brown

and Martin,^{6,7} to set bits for the number of matches to a given fragment or path, rather than simply indicating its presence or absence.

Molecular Holograms are a recently introduced variant form of bit string. Rather than using a binary bit string, a molecular hologram is represented as a string of integers corresponding to the number of times fragments are hashed into each bin. In its hashed form, this integer string is hashed into occupied bins: this representation is purely binary and will, potentially, suffer many of the drawbacks we have described—it will be interesting to explore the behavior of the unhashed representation in the context of the present results.

Although the statistical model presented in the previous section reproduces much of the behavior seen for database searches, in reality the model is too simple to be deemed realistic. Clearly, the likelihood that a particular bit will be set by a given molecule is far from random. The path C–C will be set by virtually every organic molecule. Likewise, the path C=O will be set far more often than say the path Cl–C=C–O–N. Moreover, there is also an obvious Markov dependence of the setting of one bit on the setting of another: *i.e.*, if a bit corresponding to a 5 carbon path is set, then bits for 1, 2, 3, and 4 carbon paths must be set also. A corollary of this is that bit strings based on 2D structural fingerprints exhibit redundancy of information. If nothing else, this helps vindicate the hashing, compression, or folding of such bit strings. A statistical model for the setting of bits able to encompass these complex and mutually dependent probabilities would be difficult to implement. It would have to rely on empirically derived probabilities for the setting of each bit, which may not be general, and this situation is further complicated by the asymmetry of some of these mutual probabilities. The effects have a practical impact as well. The presence of some fragments with

unusually high frequencies can limit the range of possible similarity values, while the co-occurrence of fragments introduces unwanted bias into similarity measures by placing unwarranted emphasis on certain features.

In the context of substructure search and compound retrieval the biased nature of fragment distributions and their complex interdependence has been recognized for some time,²² and work in this area has attempted to design appropriate, and more unbiased, fragment sets for use by bit string-based structure searching systems. It is generally true that the results of similarity searching will depend on the choice of features, fragments, or atom-paths upon which the encoding of bits is based. A different set of features could give rise to very different similarity values. It may be that problems associated with bit string-based measures of similarity arise from an inappropriate choice of features encoded in the bit string, rather than from any problem inherent in the bit string concept itself. It may, in principle, be possible to encode different features in the fingerprint and obtain results which are both more intuitive and more useful.

The Soergel distance function (1 - Tanimoto) is often used to transform the Tanimoto association coefficient to a distance measure. For the reasons discussed above, such measures of molecular similarity derived from bit strings based on hashed 2D structural fingerprints do not seem to possess all the features one might hope for in an accurate chemical distance. Because it lacks the properties of a metric, its suitability for certain types of application, such as assessing molecular diversity, can be questioned. At present, one can safely use these measures to categorize molecules in a qualitative way as similar or distinct but not to form a quantitative chemical distance. However, it is fair to say that using bit strings based on hashed 2D structural fingerprints does increase the efficiency, if not the efficacy, of database searching. It would be wrong to abandon the approach, but it is clear that there is considerable scope to improve it.

ACKNOWLEDGMENT

I should like to thank the referees for their helpful and constructive comments.

REFERENCES AND NOTES

- (1) Hodgkin, E. E.; Richards, W. G. Molecular Similarity Based On Electrostatic Potential And Electric-Field. *Int. J. Quant. Chem.* **1987**, *14*, 105–110.
- (2) Blaney, F. E.; Edge, C.; Phippen, R. W. Molecular surface comparison. 2. Similarity of electrostatic vector fields in drug design. *J. Molecular Graph.* **1995**, *13*, 165–174.
- (3) Ashton, M. J.; Jaye, M. C.; Mason, J. S. New Perspectives in Lead generation II: evaluating molecular diversity. *Drug. Discovery Today* **1996**, *1*, 71–78.
- (4) Kauver, L. M.; Higgins, D. L.; Villar, H. O.; Sportsman, J. R.; Engqvist-Goldstein, A.; Bukar, R.; Bauer, K. E.; Dilley, H.; Rocke, D. M. Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.* **1995**, *2*, 112–118.
- (5) Weinstein, J. N.; Myers, T. G.; O'Connor, P. M.; Friend, S. H.; Fornace, A. J., Jr.; Kohn, K. W.; Fojo, T.; Bates, S. E.; Rubinstein, L. V.; Anderson, N. L.; Buolamwini, J. K.; van Osdol, W. W.; Monks, A. P.; Scudiero, D. A.; Sausville, E. A.; Zaharevitz, D. W.; Bunow, B.; Viswanadhan, V. N.; Johnson, G. S.; Wittes, R. E.; Paull, K. D. An information-intensive approach to the molecular pharmacology of cancer. *Science* **1997**, *275*, 343–9.
- (6) Brown, R. D.; Martin, Y. C. Use Of Structure - Activity Data To Compare Structure-Based Clustering Methods And Descriptors For Use In Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572–584.
- (7) Brown, R. D.; Martin, Y. C. The Information Content Of 2D And 3D Structural Descriptors Relevant To Ligand-Receptor Binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1–9.
- (8) Lewis, R. A.; Mason, J. S.; McLay, I. M. Similarity measures for rational set selection and analysis of combinatorial libraries: the Diverse Property-Derived (DPD) approach. *J. Chem. Inf. Comp. Sci.* **1997**, *37*, 599–614.
- (9) Basak, S. C.; Grunwald, G. D. Molecular Similarity And Estimation Of Molecular-Properties. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 366–372.
- (10) Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical Similarity Using Geometric Atom Pair Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128–136.
- (11) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical Similarity Using Physicochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118–127.
- (12) Holliday, J. D.; Ranade, S. S.; Willett, P. A Fast Algorithm For Selecting Sets Of Dissimilar Molecules From Large Chemical Databases. *Quant. Struct.-Act. Relat.* **1995**, *14*, 501–506.
- (13) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring Diversity - Experimental-Design Of Combinatorial Libraries For Drug Discovery. *J. Med. Chem.* **1995**, *38*, 1431–1436.
- (14) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* **1996**, *39*, 3049–3059.
- (15) MACCS-II; MDL Ltd.: San Leandro, CA, 1992.
- (16) James, C. A.; Weininger, D. *Daylight Theory Manual*; Daylight Chemical Information Systems, Inc.: 1995.
- (17) *UNITY Reference Manual*; Tripos Inc.: St. Louis, MO, 1995.
- (18) Hall, M. *Combinatorial Theory*; Wiley: New York, 1986.
- (19) Vogt, G.; Etzold T.; Argos, P. An assessment of amino acid exchange matrixes in aligning protein sequences: the twilight zone revisited. *J. Mol. Biol.* **1995**, *249*, 816–31.
- (20) Flower, D. R.; North, A. C. T.; Attwood, T. K. Structural and sequence relationships in the lipocalins and related proteins. *Prot. Science* **1993**, *2*, 753–761.
- (21) Flower, D. R. Structural relationship of Streptavidin to the Calycin Protein Superfamily. *FEBS Lett.* **1993**, *333*, 99–102.
- (22) Lynch, M. F. The microstructure of chemical databases and the choice of representation for retrieval. In *Computer Representation and Manipulation of Chemical Information*; Wipke, W. T., et al. Eds.; Wiley: New York, 1974.

CI970437Z