ELSEVIER

DISSIM: A program for the analysis of chemical diversity

Darren R. Flower

Department of Physical and Metabolic Sciences, Astra Charnwood, Loughborough, Leicestershire, UK

As interest in database searching and compound selection has grown, there has been a concomitant growth in interest in the quantification of chemical similarity. Described here is a computer program called DISSIM, which addresses the problem of selecting diverse subsets from larger collections of chemical compounds. It is a pragmatic solution combining a maximum dissimilarity search algorithm and a general multidimensional measure of chemical similarity based on the combination of different molecular descriptors. The problem of correlation between descriptors is addressed and appropriate schemes for weighting and normalisation are described. The specific application of these techniques to the comparative analysis of topological indices and their use in the area of chemical diversity analysis and compound selection are also described. © *1999 by Elsevier Science Inc.*

Keywords: chemical diversity, compound selection, topological index, maximum dissimilarity search, variable selection, molecular descriptors

INTRODUCTION

Unmet medical need remains a constant spur to the discovery of new therapies and new therapeutic agents. Within the pharmaceutical industry, the identification of novel candidate drugs is the fountainhead of future success. It is the ultimate source of new products and sustainable profitability. There is an analogous situation in the agrochemical industry with regard to the development of new herbicides and pesticides. The discovery of candidate drugs begins with initial lead compounds and proceeds through an optimisation process familiar from decades of medicinal chemistry. Generally, lead compounds possess key properties, such as activity at a receptor or enzyme, but are deficient in others, such as selectivity, metabolic stability, or their pharmacokinetic profile. Traditionally, new lead compounds have arisen, largely as a result of serendipity, from analogy to the structures of known compounds. These may be natural ligands—enzyme substrates or receptor agonists—or they may be extant pharmacological agents—inhibitors or antagonists. Although there is no doubt that this approach has proved successful in the past, and will continue to do so, its limitations have led many to complement it with alternative strategies.

High-throughput screening (HTS) is able to assay large numbers of compounds in comparatively short times. At least in principle, HTS allows for the identification of novel lead compounds, in the absence of any structural information regarding ligand or receptor, for new areas of biological activity where, potentially, knowledge concerning the nature of either is lacking. It is generally agreed that the benefits of HTS technology in accelerating the drug discovery process cannot be overestimated. To capitalise on the potential power of HTS it is obviously necessary to access compounds for testing. Moreover, the number of compounds we test will have a profound influence on our results. At one extreme, if we test too few compounds we are unlikely to find active compounds; at the other extreme, we will have hits but finding them will prove an expensive business. Therefore our choice of compounds is a matter of considerable importance.

There are many sources, or types, of compounds for HTS. Natural products are one: extracts from plants, bacterial or fungal cultures, marine flora and fauna, etc., have all proven useful sources of novel lead compounds. By natural products we generally mean secondary metabolites, compounds that appear to have no explicit role in the internal economy of the organism that biosynthesised them. Of the competing arguments that seek to explain the existence of such redundant molecules, perhaps the most engaging is an evolutionary one: secondary metabolites enhance the survival of their producer organisms by binding specifically to macromolecular receptors in competing organisms with a concomitant physiological action. As a consequence of this intrinsic capacity for interaction with biological receptors, made manifest in their size and complexity, natural products will be generally predisposed to form macromolecular complexes. On this basis, one might expect that natural products would possess a high hit rate when screened and a good chance of high initial activity and selectivity. However, although potent, that same complexity makes natural products difficult to work with synthetically. When natural products are only weak hits, they do not represent

The Color Plate for this article is on page 264.

Address reprint requests to: Dr. Darren R. Flower, Edward Jenner Institute for Vaccine Research, Compton, Newbury, Berkshire, UK RG20 7NN. Phone: 01635 577954; Fax: 01635 577901; E-mail: darren.flower@jenner.ac.uk

Journal of Molecular Graphics and Modelling 16, 239–253, 1998

655 Avenue of the Americas, New York, NY 10010

1093-3263/99/$–see front matter
PII S1093-3263(99)00014-5

particularly attractive starting points for optimisation. Other natural products can prove to be potent and selective compounds that can, with little or no modification, progress directly to clinical trials. For example, cyclosporin, FK506, and taxol have all found clinical application.

Another source of compounds is represented by molecular libraries generated by combinatorial chemistry. Libraries have, in the past, been based on overly familiar templates or give rise to problems of novelty, variety, and, for peptides and peptidomimetics, metabolic stability and chemical tractability. Because this technology is new, it is only now having a sizable impact. The other principal sources of compounds for screening are synthetic organic molecules that have accumulated in public and corporate compound banks. These are generally chemically tractable starting points, but can suffer from lack of novelty and have unwanted additional activities. These three approaches are complementary, and all have some advantages and disadvantages.

It can be assumed, with certainty, that the more compounds we test the greater the likelihood of successfully identifying lead compounds, assuming that these compounds are, in some meaningful sense, different from one another. Although this subject has been studied since at least the mid-1980s, it is only in the era of combinatorial chemistry and high-throughput screening (HTS) that the analysis of the similarity, or dissimilarity, of chemical structures has assumed an important position in the fields of computational chemistry and chemical information science.[1]

MEASURES OF CHEMICAL SIMILARITY

Currently, there is no generally agreed quantitative definition of chemical similarity. Many proposed methods exist, each with different strengths and weaknesses. However, choice of an appropriate distance measure is important: while it may be that no single method is significantly better than the rest of the best, obviously inappropriate methods do exist. One of the most widely used is based on mapping fragments within a molecule to bits in a binary string. It has been shown that bit strings provide a nonintuitive encoding of molecular size, shape, and global similarity; and also that the observed behaviour of bit string-based searches have a large nonspecific component.[2] On this basis, one might wish to question whether bit string-based similarity methods have all the features desirable in a quantitative chemical distance measure.

Many approaches to molecular similarity have been suggested; of these, many have been based on the combination of several different properties of a chemical structure.[3] Such descriptors have included calculated properties based on representations of molecular structure at both the two-dimensional level (topological indices and constitutional descriptors[4,5]) and three-dimensional level (properties derived from MO calculations, surface area and volume, or COMFA[6–8]). Other descriptors include measured physical properties or biological activities.[9–11] To these primary descriptors, we can add complex, or combined, descriptors.[12] The number of potential quantities currently available is daunting.

Of these many alternatives, which are the most appropriate descriptors? We might wish to choose as descriptors those properties we feel we understand well. A near universally applicable quantity such as the octanol/water partition coefficient (so-called LogP) might be a better choice, say, than a particularly obscure and poorly characterised topological index. In the context of drug discovery, we might wish to concentrate on those descriptors that allow us mechanistic insights into the basis of some particular biological activity. Having selected a set of potential descriptors, we need to combine them appropriately. Highly correlated variables add little extra information. We may wish to remove such variables or, at least, compensate for their presence.

In this article, we describe a computer program called DISSIM, which implements a pragmatic approach to the problem of selecting diverse subsets of larger collections of chemical compounds. DISSIM makes use of a maximum dissimilarity search algorithm and a general multidimensional measure of chemical similarity based on combining different properties or descriptors. The correlation of descriptors and appropriate schemes for weighting and normalisation are discussed. In addition to discussion of methodology, the specific application of these techniques to topological indices is addressed. The use of DISSIM in the area of chemical diversity analysis and compound selection is also described.

DEFINING A DESCRIPTOR-BASED CHEMICAL DISTANCE

If decades of quantitative structure activity studies have shown anything, it is that single descriptors, even extraordinarily powerful ones such as LogP, account for only a part of observed biological activity within a series of similar or dissimilar chemical compounds. When attempting to formulate an approach to chemical diversity analysis one seeks to emulate the success of QSAR in capturing, in simple relationships, the dependence of biological activity on measured or calculated properties. The best such relationships are often no more than correlations without meaning in a mechanistic sense, but even these relations can provide useful ways of designing new compounds.

Having selected a set of descriptors, it is then necessary to combine them in order to derive a useful dissimilarity measure. In creating this multidimensional metric, we must account for inherent differences between the descriptors as numerical variables, essentially a matter of appropriate scaling, and also weight each contribution to reflect its relative importance and to compensate for any residual correlation between different variables.

The approach to doing this is simple. Using reference distributions, each descriptor is first normalised, and then combined, with weighting, to form a Euclidean distance. A "phase space" of molecular descriptors is defined; a Euclidean distance within such a space forms a simple dissimilarity measure. Each descriptor is effectively transformed to a Z score and weighted before being used. The distance d_{ij} between two compounds i and j is defined as in Eq. (1):

$$d_{ij} = \sqrt{\sum_k (\hat{x}_k^i - \hat{x}_k^j)^2} \qquad (1)$$

where

$$\hat{x}_i^j = w_i\left(\frac{x_i^j - \bar{x}_i}{\sigma_i}\right) \quad \text{and} \quad w_i = \prod_j w_i^j$$

where $\hat{x}_i^j$ is the ith normalised descriptor of the jth compound, x_i^j is the raw nonnormalised descriptor, $\bar{x}_i$ is the mean value for

the ith descriptor, and σ_i is the standard deviation for the ith descriptor, w_i is the weight associated with descriptor i, and w_i^l is the lth component of weight w_i.

Because the raw value of each descriptor is on a different scale, it is necessary, and a standard statistical pretreatment, to transform the data in this way. This prevents particular descriptors from dominating the distance function. The values for the means and standard deviations used can be calculated from the data itself or taken from some reference set. This second option is attractive as it allows the distance values derived to be compared between data sets.

Weights can be used to reintroduce the relative importance of particular descriptors lost in the scaling process. For example, in a particular analysis we may know that the number of hydrogen bond donors or acceptors is more important—and so should contribute more to the distance function—than a topological index. Weights can also be used to help compensate for any residual undesirable correlation between variables. There are occasions when one may wish to base a distance function on descriptors, such as molecular weight and LogP, for example, which remain correlated, in a statistical sense, because they represent quantities that are well understood and are amenable to synthetic manipulation in a straightforward way.

The issue of reducing the influence of contributions from degenerate data has taxed the minds of workers in many disciplines, and arguably the most natural way to achieve this is through a tree structure relating the degree of correlation between variables. It is possible to represent the distances between objects characterised as vertices related by a tree structure in many ways, including hierarchical distances, centroid distances, and ultrametrics. More specific example applications include a number of methods, of differing sophistication, developed in the area of protein sequence analysis.[13–18] Methods based on a tree structure are necessarily complicated, and an adequate description of them is beyond the scope of the present work. Instead, the concept is illustrated by a simpler method described below.

SELECTING AN APPROPRIATE BASIS SET OF MOLECULAR DESCRIPTORS

In defining a multidimensional distance from a set of descriptors, we need to ask the question: how discriminating are our variables? Ideally we require each of the descriptors to have, in itself, a highly developed power to distinguish between compounds, and that the chosen descriptors not be highly correlated with each other. We might ask ourselves, how redundant is a particular descriptor? Can it distinguish between different chemical structures? Put another way: how correlated are the values attributed to different structures, as opposed to different descriptors of the same compound? We are interested, then, in the distinguishing power of each descriptor: the degree to which different structures give different measured or calculated values. The simplest expression of this is the percentage of structures inspected with unique values. Descriptors with a low distinguishing power may need to be rejected.

A more generally important question is as follows: how correlated are different descriptors? A variable that is highly correlated to another brings little extra discrimination. We may wish to eliminate highly correlated variables, compensate for them by modifying our analysis, or simply be aware of their presence. A particularly straightforward way of looking at the similarity of different descriptors is through the use of a so-called correlation matrix, where each element of a symmetric matrix is a correlation coefficient between variables corresponding to its indices.[19] By using the correlation coefficient as a simple similarity measure, it is possible to explore the relatedness of our different variables. We might wish to group highly related descriptors and then to eliminate those that are redundant.

The general expression for a correlation coefficient takes the form shown in Eq. (2)[19]:

$$R_{xy} = \sum xy \Big/ \sqrt{\sum x^2 \sum y^2} \tag{2}$$

where $x = X - \bar{X}$, $y = Y - \bar{Y}$, and $\bar{X} = \Sigma X/n$. There is a considerable number of more sophisticated ways of measuring the interaction of variables, but many are inappropriate when the number of variables and the size of the data set become large. The correlation matrix has been much used to examine the relationship within a set of variables, and has become a standard tool. However, when the number of variables itself becomes reasonably large, this approach, while simple, becomes clumsy and the implicit relationships difficult to visualise. However, taking a lead from macromolecular sequence analysis inter alia, we can chose to visualise the relationship between the descriptors as an unrooted tree.

A tree is a powerful means of both formally, and visually, representing the structure of a data set.[20] The type of tree used here for visualisation is a member of a class of trees, sometimes called X-trees, which includes phylogenetic, additive, and Steiner trees; they are characterised by vertices of two types: labelled or real vertices and latent or unlabelled vertices. Labelled vertices are typically, but not exclusively, terminal or leaf nodes. A combination of the branch lengths within the tree and its topological organisation allows us to represent pictorially how one variable relates to another, the inherent clustering of and organisation within our data, etc.

As a preliminary to our tree-based approach, our correlation matrix, which is a similarity matrix, must be converted to a distance matrix. There are many ways to achieve this transformation,[21] none of them ideal. A simple transformation has the form shown in Eq. (3):

$$D_{ij} = 1 - \exp(r_{ij}) \tag{3}$$

where r_{ij} is the correlation coefficient between variables i and j. This expression is sufficient for all practical purposes. Given a table of distances between descriptors, then existing software for induction of unrooted trees, such as PHYLIP,[22] can be used to calculate and display trees corresponding to the required data. DISSIM will write a correlation matrix summarising the interaction between descriptors for a given data set, as well as the corresponding distance matrix that can be read directly into PHYLIP. Of the different methods of tree generation available with the package, the Fitch algorithm has proved suitable for this purpose, particularly for small sets of descriptors. When the descriptors become large, the Kitsch algorithm provides a suitable and less computationally demanding alternative. A number of programs can be used to visualise tree structure apart from programs within the PHYLIP package; examples include Treetool,[23] Treeview,[24] and Treecon.[25] Figure 1 gives an example of using unrooted trees to visualise the relationships between a set of descriptors.

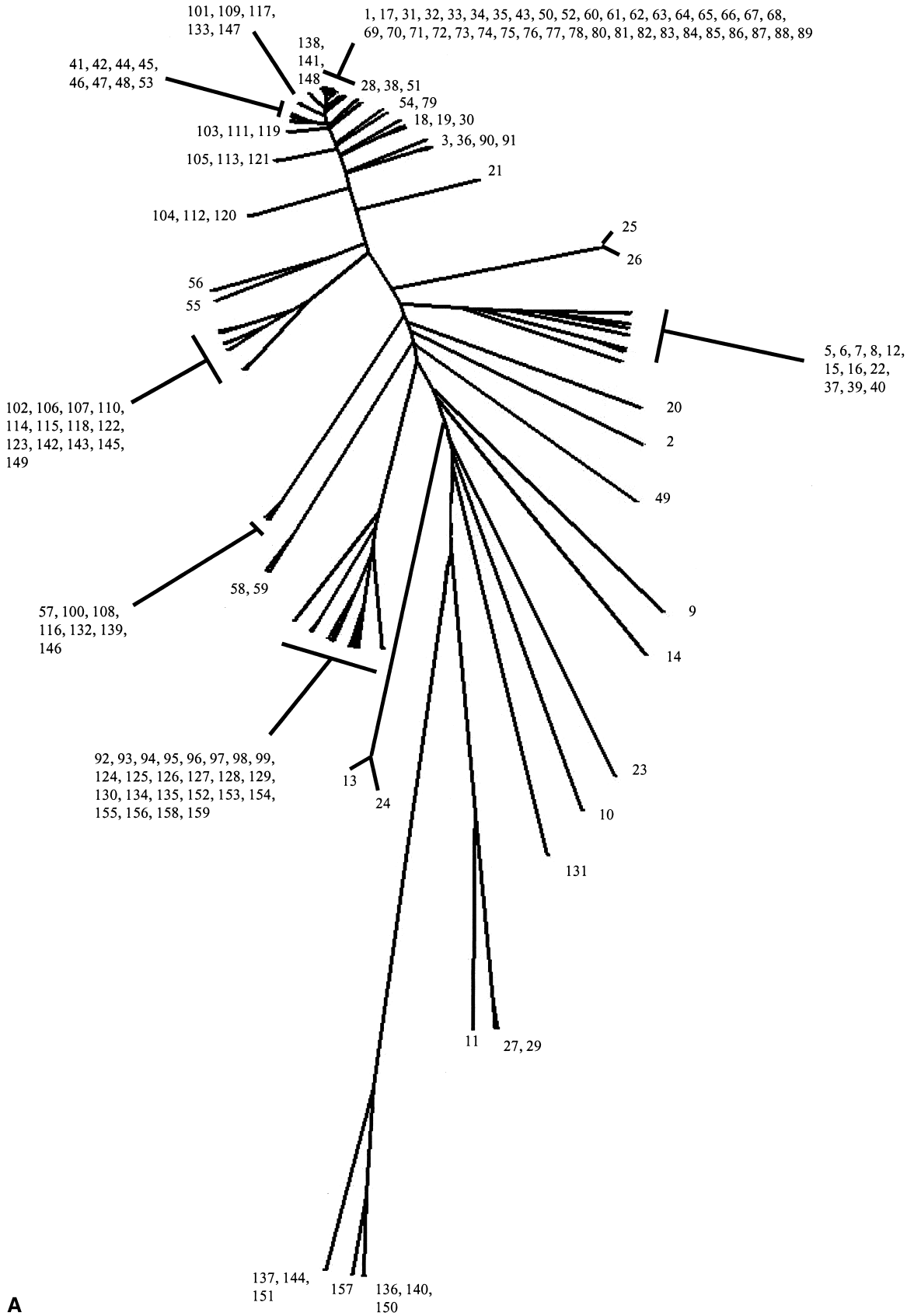


Figure 1. Using an unrooted tree to visualise the relationships between a set of descriptors: (A) 159 descriptors; (B) subset of 35 descriptors.The equivalent descriptors from Table 1 are as follows: 1, 3, 6, 9, 10, 11, 13, 14, 18, 20, 21, 23, 24, 25, 26, 27, 28, 36, 37, 41, 51, 54, 55, 56, 58, 92, 101, 103, 104, 105, 131, 137, 148, 150, 156.

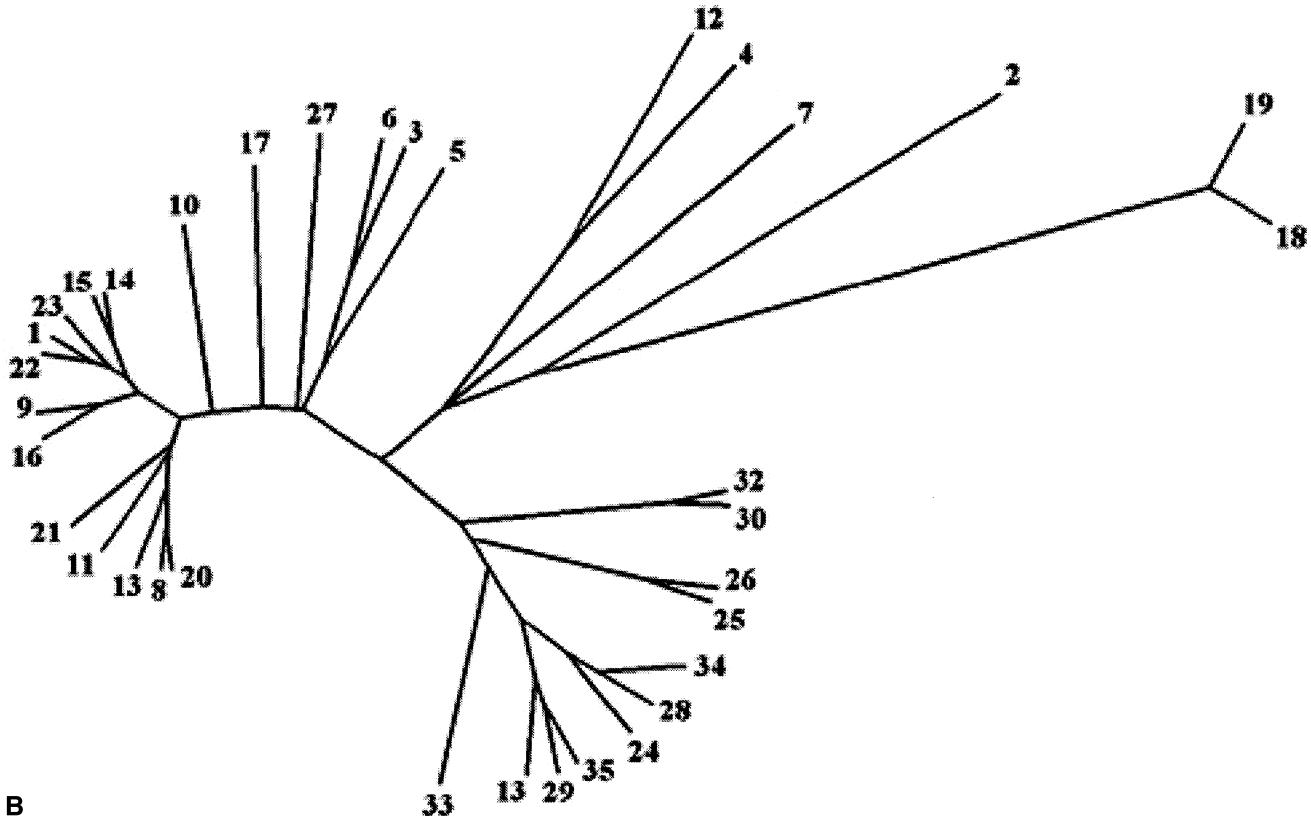


Figure 1. (continued)

The tree structure corresponding to the relative correlation between variables can be used to select a subset of useful descriptors, either manually by inspection or by an automatic method such as a clustering algorithm. As explained above, there will, to some greater or lesser extent, be residual correlation between the retained descriptors and one may choose to compensate for this by weighting the different variables appropriately. Any unidimensional scaling such as this is crude at best, lacking detailed structure, and so it is possible to reproduce weights of acceptable quality by using simple algorithms. For example, their respective distances from their common centroid can effectively weight a set of points in any Euclidean system. Moreover, these distances can be reconstructed directly from the set of distances relating each point to every other point, using a relationship owing to Lagrange[26] [Eq. (4)]:

$$d_{i0}^2 = \frac{1}{N}\sum_{j=1}^{N} d_{ij}^2 - \frac{1}{N^2}\sum_{k>j=1}^{N} d_{jk}^2 \qquad (4)$$

where N is the total number of points and d is the distance between points. This set of distances to the centroid can then be appropriately normalised to provide a set of weights for the descriptors.

APPLICATION TO TOPOLOGICAL DESCRIPTORS

The topological index has a long history.[27,28] In the present context it is synonymous with the terms topological, structural, and graph invariant. A topological index is a numerical index that seeks to characterise, or quantify, the topological, or structural, properties of a graph. Within the chemical literature, an enormous number of different topological indices have been proposed.[29–32] Within a chemical context, they have found uses in database indexing and other aspects of the management of chemical information, in the prediction of physical and biological properties, in formulating QSARs, and in drug design.[12,30,31]

Constitutional descriptors are a closely related quantity: atom counts, number of features (rings, functional groups, etc.). The distinction between topological indices and constitutional descriptors is to some extent semantic. Some constitutional descriptors require graph theoretical methods for their calculation, whereas others do not. One significant distinction is that constitutional descriptors do not, in general, characterise properties of the whole graph or molecule, whereas topological indices do, potentially, in so far as they generally contain information derived from all vertices and/or edges of a graph.

The potential number of topological indices and constitutional descriptors available for use is enormous. Even a cursory examination of the literature is sufficient to show their great number and variety. When choosing topological indices as a set of descriptors, ideally we want each of the descriptors to have a highly developed power to distinguish between compounds, and for none of the descriptors to be highly correlated with each other. There are a number of well-known relationships between different topological indices, for example, that be-

tween the Weiner number and the Altenburg polynomial. Many other, often analytical or even algebraic, relationships have been reported: recent examples include Chan et al.[33] Gutman and Mohar,[34] and Klavzar and Gutman.[35] The relationship between topological indices and other descriptors, such as MW, which are not graph-theoretical in nature, are also well known. As the degree of relatedness between descriptors is dependent on the structures of compounds used to generate the descriptors, in general, one needs to perform an analysis of the correlation between available descriptors.

A computer program has been described that, amongst many other functions, calculates a useful range of topological indices and constitutional descriptors. Although by no means complete, the range of indices generated by ALTER[36] compares favourably with that produced by other software. The availability of this software allows for the ready analysis of the correlation between a large number of topological descriptors over a large set of chemical structures. Table 1 summarises the available descriptors used in this analysis.

A database of 100 000 unique structures was prepared. An initial set containing a large number of commercially available structures, obtained in a variety of different formats, was converted to a single common format—Weininger's SMILES notation—using the program ALTER.[36] By using the program CCT, part of the Daylight software package,[37] these arbitrary SMILES strings were transformed into canonicalised or "uniquified" SMILES, which provides an unambiguous description of a particular structure. An ordered leader-type clustering algorithm—similar to that employed successfully in the construction of nonredundant composite sequence databases[38]—was used to remove redundant data from our final set of SMILES. A priority was assigned to the set of initial databases. By working down this ordered list of SMILES, any duplicated string identical to a higher-priority string—i.e., above it in the combined list—could be identified and eliminated. A subset of 100 000 compounds was then selected randomly from the final nonredundant set.

The results of our correlation analysis are summarised in Figure 1. Correlation coefficients were generated between all pairs of variables and converted to distances. Tree induction was used to generate an unrooted tree summarising the interaction between the 159 variables. As can be seen, such a tree is a useful way of visualising the relationship between these descriptors (Figure 1A), and it is possible to obtain a representative set of variables, 35 in this case, which correlate weakly with each other (Figure 1B). One may approach this selection in many ways. One may assess it by eye, one may apply minimum branch length selections, or one may select a defined number of variables where the total terminal branch length is a maximum.

One might wish to analyse these results in terms of the nature of the relationship between particular descriptors. Such a lengthy analysis is not our present focus. One specific example should suffice to illustrate the potential usefulness of such an analysis. The molecular identification number of Burden[39] is not highly correlated with any other descriptor. However, those descriptors to which it is most similar, albeit weakly, include the number of rings, element counts, and topological indices such as the Balaban index and various Kier and Hall path indices. These support the usefulness of Burden's formalism, as it is able both to distinguish and discriminate between compounds, while succinctly capturing information about the structural topology and elemental composition of a compound. We may also note in passing that descriptors looking at properties of the uncoloured graph underlying the molecular graph, and therefore taking no account of heteroatom substitution, are generally poorly discriminating.

Some descriptors are highly correlated despite seeming to be conceptually distinct. This may be due in part to chance effects,

Table 1. One hundred and fifty-nine descriptors generated by ALTER and used in the correlation analysis

Descriptor number	Description of descriptor
1	MW
2	HB donor and acceptors
4	# heavy atoms
5–16	% element type
17, 18	# rings, # bonds
19, 20	#/% rotatable bonds
21–24	% carbon carbon bonds, % carbon hetero bonds, % hetero-hetero bonds same, % hetero-hetero bonds other
25–27	% atoms in rings, linkers, and side chains
28	Weiner number
29	Balaban index
30	Centric index
31, 32	Zagreb index 1 and 2
33, 34	Randic index 0 and 1
35–38	Electrotopological values sum of all atoms, heteroatoms, halogens, and carbons
39–42	Kappa indices 0–3
43–45	Hetero kappa 1–3
46	KH Phi
47–49	PetiJohn radius, diameter, and eccentricity/ shape I2
50	Harary number
51	Schultz index
52, 53	Balaban mean distance deviation and RMSD indices
54	General distance index
55, 56	Symmetry indices
57	Information Weiner index
58, 59	Burden molecular identification numbers 1 and 2
60–69	# Path length 1–10
70–79	Kier Hall path indices length 1–10
80–89	KH path cube root length 1–10
90–99	KH path electropological length 1–10
100–107	# Cluster size 3–10
108–115	Cluster index KH 3–10
116–123	KH cluster 1/3 KH 3–10
124–131	KH cluster electropological 3–10
132–138	Cluster Path length 4–10
139–145	KH cluster path index 4–10
146–152	KH cluster path index 1/3 4–10
153–159	KH cluster path electrotopological 4–10

but probably also reflects the fact that many structural or topological properties are themselves coupled within series of molecules. As new topological indices are suggested they need to be compared with other, extant descriptors using an analysis such as this. If they are highly correlated with existing quantities it is unlikely that they describe new information or new properties of molecules. One might also wish to take this into account before descriptions of new topological indices are published.

SUBSET SELECTION USING A MAXIMUM DISSIMILARITY SEARCH ALGORITHM

The maximum dissimilarity search algorithm implemented in DISSIM is shown diagrammatically in Figure 2.[40] It can operate in two modes. In one, an initial set of compounds is chosen manually, either from within the data set itself or from another set of compounds, such as a corporate compound bank. In the other mode, the algorithm chooses its own starting point. In the first of these modes it is possible to bias selection away from compounds already present in particular compound collections. This is potentially useful when databases overlap in chemical space.

In many dissimilarity programs, the initial choice of compounds corresponds to the two most chemically distant compounds. This requires calculation of an all-against-all distance matrix, which can be computationally slow. Because the dissimilarity metric used here is Euclidean, it is possible to calculate the centroid of the data set in our multidimensional space, and then find the compound most distant from the centroid. This would be more difficult were one using a dissimilarity metric based on, say, bit strings. Identifying the compound most distant from this initial choice results in a pair that is a good approximation to the most distant pair, yet requires only a square root of the number of calculations. The algorithm proceeds in standard fashion, iterating toward the goal number of selected compounds. Because the minimum distance for each molecule is stored, only the distance of unselected compounds to the newly selected molecule need be calculated at each step. This results in a computationally efficient algorithm. Other forms of this algorithm, in various other incarnations, have been described elsewhere.[41–43]

When used to find a small number of compounds, the maximum dissimilarity search method employed in DISSIM will identify a set that is broadly spread over the whole chemical space explored. Members of this set should bear relatively little resemblance to each other. At this stage, the results are similar to those produced by D-optimal design. However, as the number to be found increases, the dissimilarity, to previously selected molecules, of newly selected compounds falls rapidly. It quite quickly begins to select molecules from the centre of the compound population. D-optimal design, by contrast, will continue to pick compounds at the extremity of the distribution. Clearly, as the number found approaches the number in the initial set, the distance between successive selections closes in on the minimum distance in the overall population of molecules. Thus as one progresses through a list of compounds generated in this way, one will increasingly see compounds reminiscent of those already encountered. As the number selected approaches some reasonable percentage of the total then these will appear uniformly distributed through the total population: in effect, the algorithm has performed an even sampling of the search space.

We have compared D-optimal selection and the maximum dissimilarity search algorithm implemented in DISSIM in a test study. Sixteen reagents were selected from 90, using the two selection methods and four different sets of descriptors: 39 descriptors from ALTER; 20 diverse molecular orbital properties including dipole moment, HOMO and LUMO energies, heat of formation, etc., obtained from SPARTAN; substructure fingerprints[44]; and pharmacophore triplets.[1] For different descriptor sets, the overlap between selections (in terms of number of equivalent compounds) made using D-optimal design were, typically, two or three times the overlap between sets picked using maximum diversity. D-optimal selection identified compounds from the edge of the distributions, while DISSIM chose an even spread through the data space. Inspection of the resulting sets confirmed the more extreme overall nature of the D-optimal sets. The effect of differences in parameter set was relatively weak compared with the effect of the different selection methods.

Dissimilarity selection can be compared with clustering. Dissimilarity searching finds compounds that are well separated by a distance measure. The initial selection corresponds to a representative sampling of the total population. Clustering tends to be highly parameter dependent, often time consuming, and occasionally unreliable; in our hands at least, dissimilarity search is both simpler, more effective, and more efficient. The key to compound selection is even sampling, as performed by DISSIM, and the avoidance of near analogues. It also has the additional benefit of producing a good approximation of a representative subset, potentially summarising succinctly the chemical space covered by the initial compound population.

APPLICATION TO COMPUTER-AIDED COMPOUND SELECTION

There are many situations in which one may wish to select a representative collection of compounds from a larger set of molecules. One such situation is compound acquisition, often undertaken to supplement an existing compound bank by procurement from diverse suppliers in order to increase the number and range of compounds available for screening and thus the likelihood of successfully identifying lead compounds. One of the main ways in which this has been achieved has been through the procurement of compounds identified from electronic catalogues using a computational approach to compound selection. The descriptor-based approach to defining a chemical distance, as described above, has been incorporated into our protocol for computer-aided compound selection, or CACS.

Our approach to this problem is based on the notion of chemical diversity: our goal is the identification of compounds that are palpably dissimilar from each other. Thus selection of trivial analogues is to be minimised. Our selection protocol is conceptually simple and straightforward. Initially, the atom and connectivity data present in some source database from which we seek to make a selection is converted into a common format: SMILES notation.[37] The second stage of the process involves removal of duplicates within the data set using canonicalised SMILES notation, as described above, as well as the removal of certain specific compounds, such as those

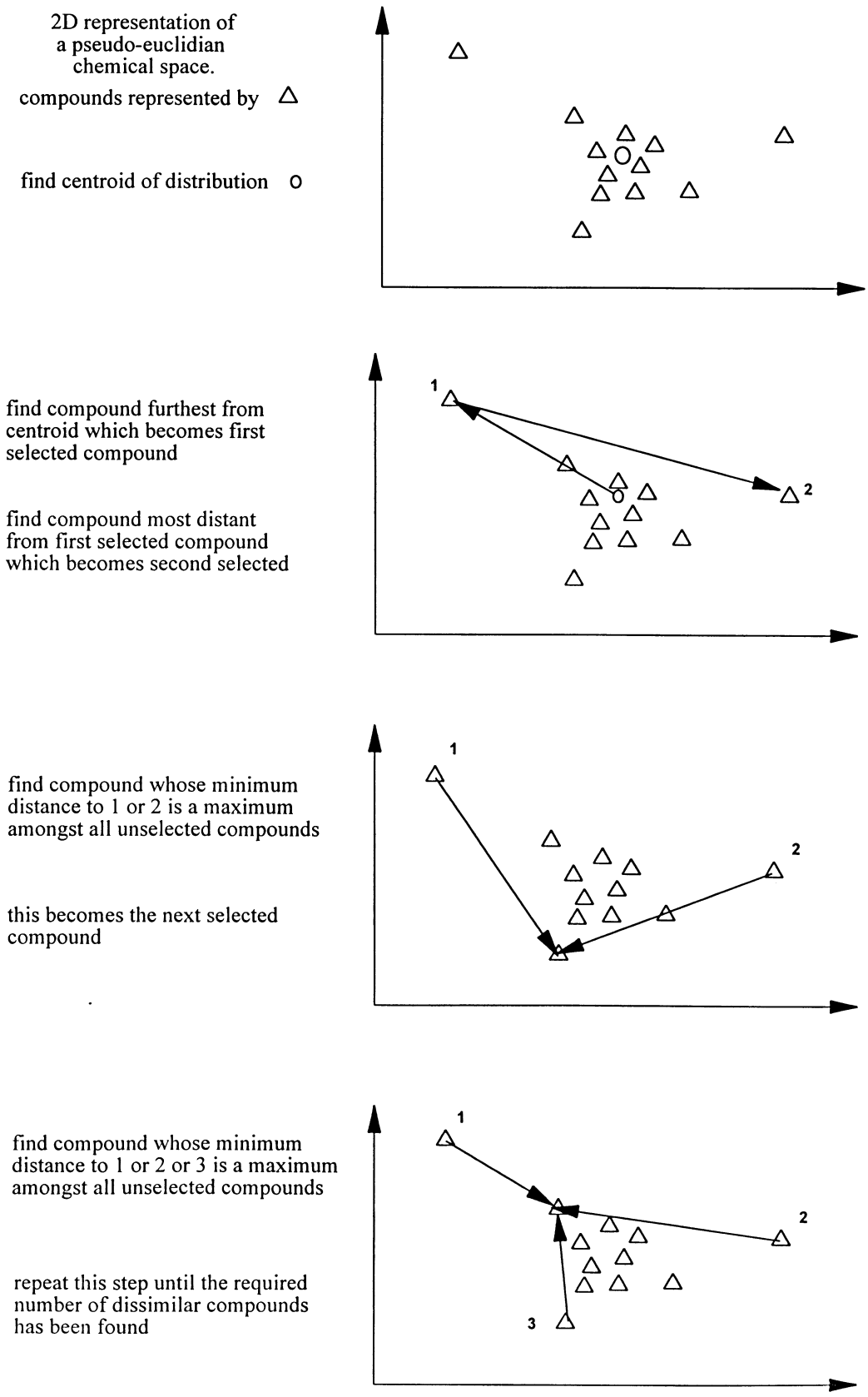


Figure 2. Maximum dissimilarity search algorithm used by DISSIM.

present in an existing compound bank. The third stage involves "cleaning" the database of "undesirable" compounds: those with unacceptably high, or low, molecular weights, reactive or poorly functionalised molecules, etc. In the fourth stage, the cleaned data are then subjected to a diversity analysis using DISSIM, which combines our descriptor-based metric with a maximum dissimilarity algorithm, to sample the data set. The resulting selection can then be screened manually to arrive at a final selection. This final interactive filtering step, where individual molecules are included or excluded, ensures that all compounds purchased have been manually vetted for their suitability. This is a good final safeguard and check on the performance of the protocol.

One of the key steps involved in performing a compound selection involves "cleaning" the data,[45] that is to say, removing compounds whose structure or properties would preclude their selection. We summarise the exclusion criteria as the Good, the Bad, and the Ugly. The Good refers to retaining compounds with some desirable feature. The presence of certain interactive atoms or groups, for example (such as oxygen or nitrogen), or the balance between cyclic and acyclic structures (less than 5% of oral drugs are totally acyclic, while essentially none contain only cyclic bonds). By the Bad we mean potentially reactive functional groups (see, for example, Figure 3A), and by the Ugly we mean the presence of certain features, such as certain functional groups or combinations of functional groups, which might render a compound an unattractive starting point for optimisation (Figure 3B). Rejection criteria that fall into this latter category also include those often referred to as Lipinski analysis—upper and/or lower bounds on quantities such as molecular weight or LogP.[46] We might wish to screen out abnormally large or small molecules, using broad cutoffs such $160 < MW < 900$, or to tailor our choice of molecular weight to favour oral bioavailability.

Few, if any, databases are totally clean and lose no structures at this stage. The proportion varies, as does the type of structure screened out. For example, databases biased toward chemical reagents should, by their very nature, suffer considerably higher knockout rates than databases whose compounds have supposedly been preselected for their suitability for screening.

For this particular application, a chemical similarity metric was developed based on the quasi-unsupervised selection of topological descriptors. A simple genetic algorithm was used to refine the weights associated with our initial selection of descriptors in order to produce an acceptable level of chemical diversity in selections made. The objective function used combined a bit string-based comparison with various substructure counts. A list of descriptors and their weights is given in Table 2. Means and standard deviations, used in normalisation, were taken from a small collection of orally bioavailable drugs. Such normalisation inherently biases our definition of chemical similarity to a region of chemical space that is most interesting to us.

In practice, compound selection by this method is quite straightforward. ALTER is again used to generate descriptors, which are read into the program DISSIM. These are optionally scaled, the number of molecules required is defined, and the maximum dissimilarity algorithm is run. The resulting dissimilar set of compounds is written out ready for final manual filtering.

We have also applied this general approach to other compound selection problems. A diverse set of compounds required for the calibration of in vitro bioavailability assays based on the permeability of cultured cell monolayers was selected by covering the range of descriptors—such as LogP and Lipinski variables[46]—deemed to be important determinants of absorption. The approach has also been used successfully in the selection of reagents for combinatorial chemistry experiments.

As noted above, the choice of descriptors may be less important than the choice of selection method. By using PLS modelling,[40] we found high correlations between the 39 descriptors of ALTER and our test set of 20 SPARTAN-derived molecular orbital descriptors. Again, the innate structural variation in compounds is sufficiently limited that different, seemingly unrelated, descriptors would adequately capture the same information.

DISCUSSION

With the growth of interest in compound selection and library design, the measurement of chemical similarity has become a subject of considerable practical significance. Yet there is no unambiguous quantitative definition of chemical diversity. It will be instructive to consider a few examples illustrating issues in this area.

Fifteen medicinal chemists were asked to choose the most diverse 20 compounds from 100 randomly chosen molecules. An all-against-all comparison was performed for these selections, 105 comparisons in total, and the overlaps determined between them. The higher the overlap between the selections made by two chemists the more similar is their notional internal model of chemical similarity. Using standard combinatorics we can work out the probability density function for the overlaps when choosing M objects from N—in this case 20 from 100—assuming picking at random. Most of the overlaps are small and well explained by a random model. Some overlaps are significantly nonrandom—an overlap of 10 has a probability of 0.05%, for example—but does not constitute a large fraction of the total number of possible pairwise overlaps. We can use our distribution to generate a random distribution corresponding to our 105 comparisons and compare it with our observed distribution (see Color Plate 1). It is clear that the two distributions are different. Indeed, the chi-squared value suggests that the difference between the distributions is statistically significant. However, if one compares the observed distribution with the distribution we would expect if all the chemists had the same internal model for chemical similarity—that is, if they all picked the same compounds, then it is clear that the observed distribution is still a long way from this ideal. There is not a single definition of chemical similarity common to all the chemists in our set, or to chemists in general. Although each chemist probably uses some combination of molecular features (size, shape, lipophilicity, interactivity, reactivity, etc.) as the basis on which to make his or her selections, other individuals will weight each of these contributions differently.

Now consider the compound pairs in Figure 4. The first pair comprises enkephalin and morphine. To the chemist these compounds are quite different. Yet to the receptor they are similar in as much as they are both active ligands. If we contrast this with the second (hypothetical) example that mirrors the SAR familiar to most practicing medicinal chemists, then to the eyes of the chemist these are similar, yet one compound is a potent inhibitor while the other is inactive. So to

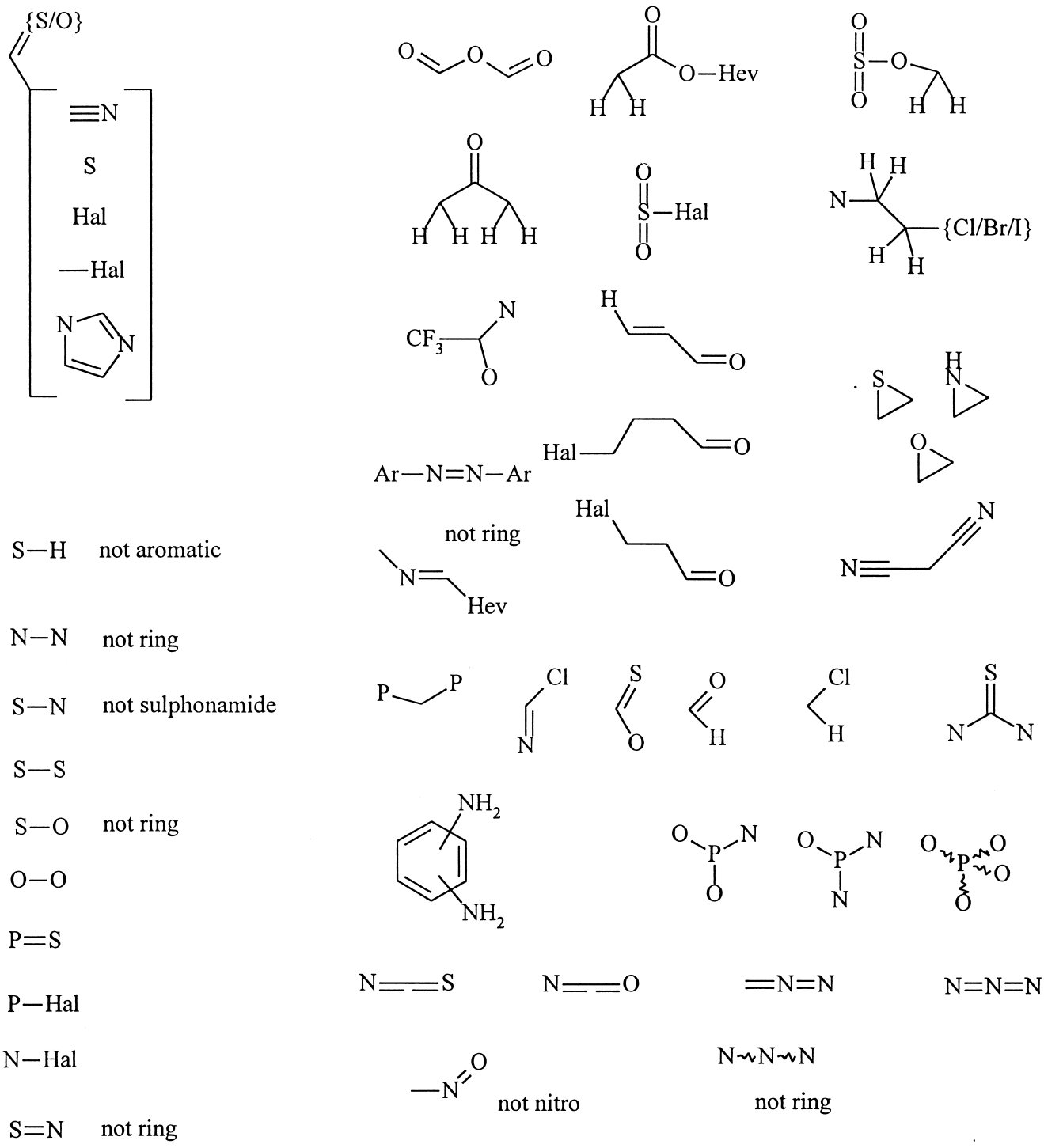


Figure 3. Substructures used to exclude compounds during computer-aided compound selection. (A) The Bad: compounds excluded owing to problems of reactivity, etc. (B) The Ugly: compounds excluded for reasons of chemical complexity and synthetic intractability. Compounds are also excluded because they contain "exotic" elements, particularly heavy metals, which may produce spurious results when screened. Only elements of the "organic subset" (C, N, O, S, P, Br, F, Cl, I, etc.) or benign salts (K, Na, etc.) are accepted; all compounds containing other elements are rejected. Again, this is conveniently achieved using ALTER. Compounds containing three or more iodines, three or more bromines, or four or more chlorines are excluded. Compounds with a $(CF_2)_4$ group or four or more CF_3 groups are excluded. Compounds containing three or more NO_2 groups are also excluded.

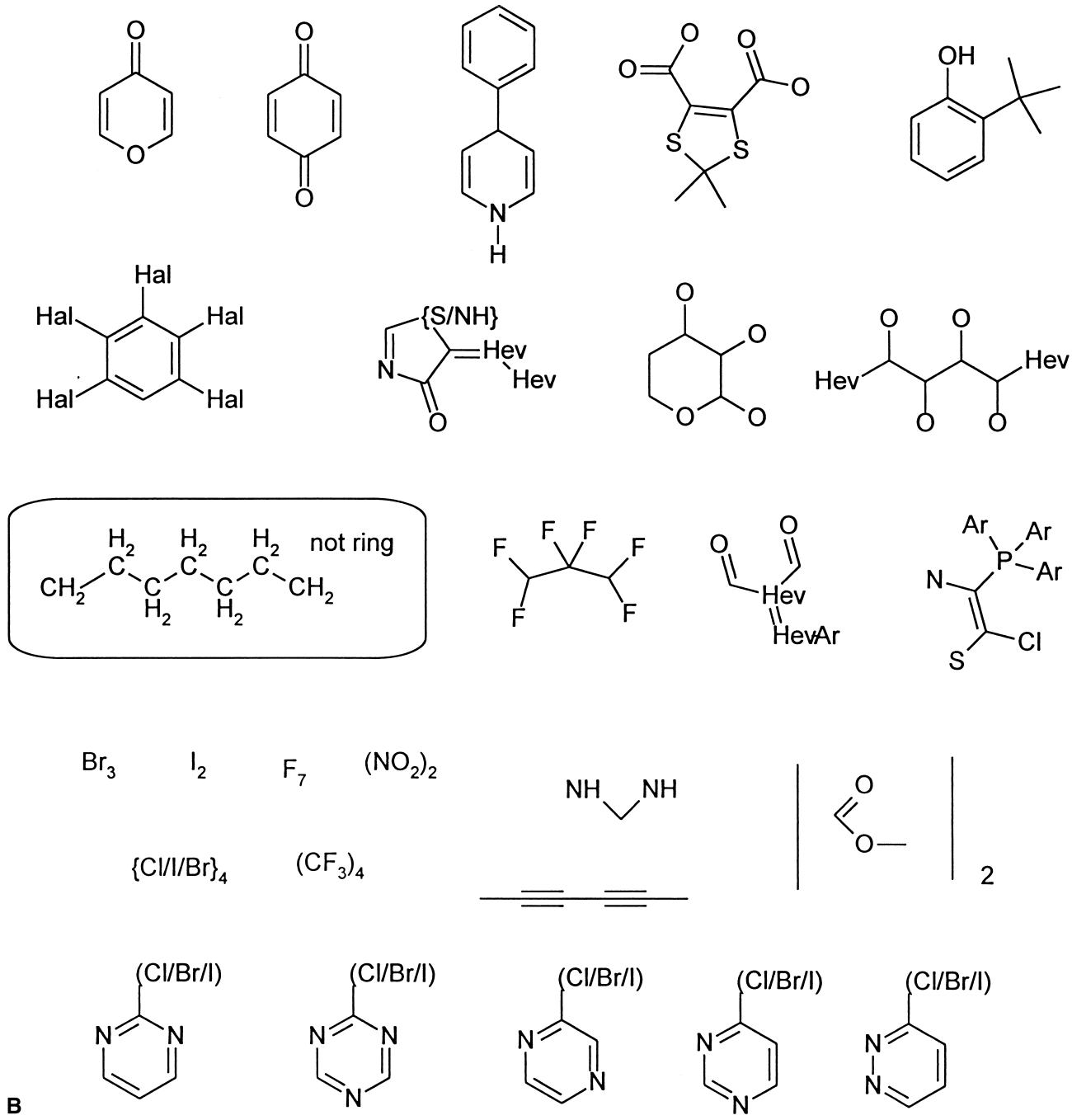


Figure 3. (continued)

the receptor these compounds, for all their apparent similarity, are quite different. This is an example of microdiversity within a congeneric series relative to the macrodiversity apparent between distinct chemical classes.

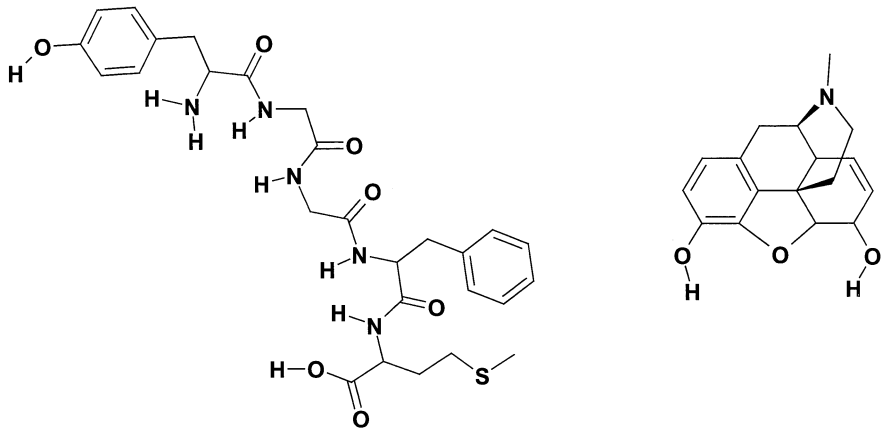
The two final molecules in Figure 4C (ranitidine and cimetidine) appear quite similar to the chemist, and, indeed, they are both active at the receptor. Yet to the patent lawyer, these compounds are different; or at least different enough that two competing companies have been able to make a great deal of money from their respective compounds. Similarity and dissimilarity remain difficult concepts. On what basis are comparisons made? There seems to be no "right" answer to this. The only correct set of rules would be those that a receptor chooses to select molecules: but these will be different for different receptors.

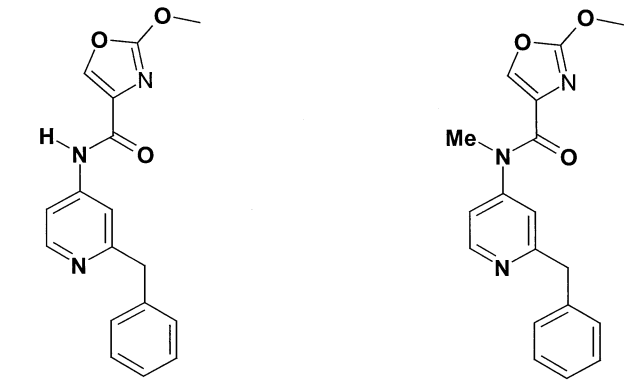
The approach to CACS outlined above represents a pragmatic solution to the various issues inherent in the selection process. Generally, when formulating a description of quantitative chemical distance one must make recourse to heuristic solutions. In discussing the merits of this method relative to any other, one encounters a fundamental problem: the difficulty of assessing the results of such an exercise. There is no

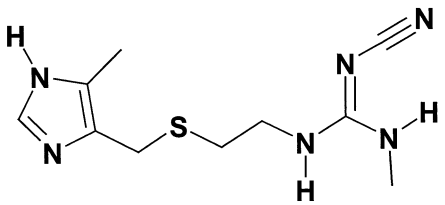
Table 2. Thirty-nine descriptors generated by ALTER and used in CACS protocol

Descriptor	Mean	Standard deviation	Weighting	Description
MW	317.456	117.343	5.0	Molecular weight
Idon	1.352	1.298	3.0	Number of H-bond donors
Iacc	3.043	2.368	3.0	Number of H-bond acceptors
thyd	0.923	0.281	0.2	Proportion of hydrogens
thet	0.263	0.131	0.2	Proportion of heteroatoms
thal	0.018	0.039	0.2	Proportion of halogens
tf	0.005	0.026	0.2	Proportion of fluorine
tcl	0.012	0.028	0.2	Proportion of chlorine
tbr	0.001	0.001	0.2	Proportion of bromine
ti	0.001	0.013	0.2	Proportion of iodine
tcarbon	0.737	0.131	0.2	Proportion of carbon
tphos	0.001	0.010	0.2	Proportion of phosphate
tsulph	0.013	0.031	0.2	Proportion of sulpher
toxy	0.132	0.093	0.2	Proportion of oxygen
tnitro	0.099	0.084	0.2	Proportion of nitrogen
Nring	2.631	1.442	2.0	Number of rings
tiribo	16.031	8.393	0.2	Number of bonds
tirobo	0.514	0.971	0.2	Number of rotatable bonds
tiprbo	0.675	0.273	1.2	Proportion of rotatable bonds
tibab	1.838	0.487	0.4	Balaban index
ticent	0.176	0.031	0.4	Centric index
tizag1	47.529	18.358	0.4	Zagreb M1 index
tizag2	137.155	60.756	0.4	Zagreb M2 index
tiran0	16.045	5.739	0.4	Randic 0th index
tiran1	10.544	3.856	0.4	Randic 1th index
tiesum	66.513	24.047	0.4	Sum of electrotopological values over all atoms
tiehet	31.784	17.949	1.0	Sum of electrotopological values over heteroatoms
tiehal	2.136	4.758	1.0	Sum of electrotopological values over halogens
tiecar	34.729	15.404	1.0	Sum of electrotopological values over all carbons
tikap1	7.596	6.474	0.5	Kier and Hall Kappa 1 index
tikap2	7.598	3.038	0.5	Kier and Hall Kappa 2 index
tikap3	4.335	2.151	0.5	Kier and Hall Kappa 3 index
tirad2	5.671	1.836	0.2	PetitJohn R2 index
tidia2	10.710	3.635	0.2	PetitJohn D2 index
tii2	0.881	0.125	0.2	PetitJohn I2 shape index
tihar	41.389	18.529	0.2	Harary number
tischul	6120.165	7991.119	0.2	Schultz index
tisyml	0.812	0.143	1.0	Total symmetry index
tisym2	0.108	0.161	1.0	Dyad symmetry index

obvious criterion by which one can prove that one selection is better than another. Given a particular measure of similarity then one can gauge the relative success of different search or selection algorithms, but there is no "gold standard" by which to judge the performance of different similarity measures. There is no consensus between chemists, or computer algorithms, and there isn't one between receptors either. There is no universally applicable definition of chemical diversity, only local, context-specific ones. We are in the realm of relative values; the success and failure of different measures is largely dependent on the context in which they are used, without any particular one consistently outperforming the others. Ultimately, the success of compound selection must be judged by its influence, through the success of HTS, on drug discovery. In passing, it is worth noting that the failure to identify good hits from HTS can reflect stringent structural requirements of a receptor, problems with the accuracy and reliability of particular screens, as well as insufficient diversity in compounds tested. Only the accumulation of HTS results will confirm the benefits of enhanced diversity.







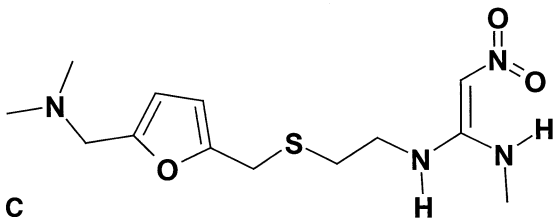


Figure 4. Pairs of compounds illustrating different aspects of chemical diversity. (A) Enkephalin and morphine; (B) two hypothetical compounds mirroring a common SAR trend; (C) ranitidine (top) and cemitidine (bottom).

SOFTWARE

DISSIM is written in standard Fortran 77 and was developed to run under UNIX on a series of Silicon Graphics workstations. DISSIM is controlled via a simple command line interface through a set of keywords. The parameters used by the program are fully configurable. Textual output from DISSIM is flexible and includes simple results files (statistical analyses or lists of selected codes, etc.) and data files suitable for input to other software (for example, the distance matrix read by the PHYLIP package).

DISSIM, and the descriptor generation program ALTER,

can be downloaded from the home page of the Royal Society of Chemistry Molecular Modelling Group: http://www.rsc.org/lap/rsccom/dab/ind006.htm.

ACKNOWLEDGMENTS

I thank the following for their help and encouragement: N.P. Tomkinson, D.P. Marriott, N.P. Gensmantel, A.M. Davis, D.H. Robinson, S. Teague, N. Kindon, A. Baxter, T. Birkinshaw, F. Lindgren, and S. Hellberg. I also thank the following for their participation in the manual selection exercise: A. Baxter, T. Birkinshaw, M. Coombs, S. Guile, S. Hirst, N. Kindon, D. Marriott, A. Mete, M. Mortimore, G. Pairaudeau, M. Stocks, S. Teague, S. Thom, A. Tinker, and I. Walters.

REFERENCES

1 Ashton, M.J., Jaye, M.C., and Mason, J.S. New perspectives in lead generation. II. Evaluating molecular diversity. *Drug Discovery Today* 1996, **1,** 71–78
2 Flower, D.R. On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.* 1998, **38,** 379–386
3 Martin, E.J., Blaney, J.M., Siani, M.A., Spellmeyer, D.C., Wong, A.K., and Moos, W.H. Measuring diversity—experimental design of combinatorial libraries for drug discovery. *J. Med. Chem.* 1995, **38,** 1431–1436
4 Brown, R.D., and 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
5 Lewis, R.A. Mason, J.S., and McLay, I.M. Similarity measures for rational set selection and analysis of combinatorial libraries: The diverse property-derived (DPD) approach. *J. Chem. Inf. Comput. Sci.* 1997, **37,** 599–614
6 Hodgkin, E.E., and Richards, W.G. Molecular similarity based on electrostatic potential and electric-field. *Int. J. Quantum Chem.* 1987, **14,** 105–110
7 Blaney, F.E. Edge, C., and Phippen, R.W. Molecular surface comparison. 2. Similarity of electrostatic vector fields in drug design. *J. Mol. Graphics* 1995, **13,** 165–174
8 Patterson, D.E., Cramer, R.D., Ferguson, A.M., Clark, R.D., and Weinberger, L.E. Neighborhood behavior: A useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* 1996, **39,** 3049–3059
9 Kearsley, S.K., Sallamack, S., Fluder, E.M., Andose, J.D., Mosley, R.T., and Sheridan, R.P. Chemical similarity using physiochemical property descriptors. *J. Chem. Inf. Comput. Sci.* 1996, **36,** 118–127
10 Kauver, L.M., Higgins, D.L., Villar, H.O., Sportsman, J.R., Engqvist-Goldstein, A., Bukar, R., Bauer, K.E., Dilley, H., and Rocke, D.M Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.* 1995, **2,** 112–118
11 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., and Paull, K.D. An information-intensive approach to the molecular pharmacology of cancer. *Science* 1997, **275,** 343–349
12 Katritzky, A.R., and Gordeeva, E.V. Traditional topological indices vs electronic, geometrical and combined molecular descriptors in QSAR/QSPR research. *J. Chem. Inf. Comput. Sci.* 1993, **33,** 835–857
13 Altschul, S.F., Carroll, R.J., and Lipman, D.J. Weights for data related by a tree. *J. Mol. Biol.* 1989, **207**(4), 647–653
14 Vingron, M., and Argos, P. A fast and sensitive multiple sequence alignment algorithm. *Comput. Appl. Biosci.* 1989, **5,** 115–121
15 Sibbald, P.R., and Argos, P. Weighting aligned protein or nucleic acid sequences to correct for unequal representation. *J. Mol. Biol.* 1990, **216,** 813–818
16 Vingron, M., and Sibbald, P.R. Weighting in sequence space: A comparison of methods in terms of generalized sequences. *Proc. Natl. Acad. Sci. U.S.A.* 1993, **90,** 8777–8781
17 Henikoff, S., and Henikoff, J.G. Position-based sequence weights. *J. Mol. Biol.* 1994, **243**(4), 574–578
18 Krogh, A., and Mitchison, G. Maximum entropy weighting of aligned sequences of proteins or DNA. *ISMB* 1995, **3,** 215–221
19 Kendall, M.G. *Correlation Methods.* Griffin, London, 1948
20 Barthélemy, J.-P., and Guénoche's, G. *Trees and Proximity Representations.* John Wiley & Sons, New York, 1991
21 Taylor, W.R., and Jones, D.T. Deriving an amino acid distance matrix. *J. Theor. Biol.* 1993, **164,** 65–83
22 Felsenstein, J. PHYLIP—Phylogeny inference package (version 3.2). *Cladistics* 1989, **5,** 164–166
23 Maciukenas, M. *TreeTool. Ribosomal RNA Database Project.* University of Illinois (ftp://rdp.life.uiuc.edu/rdp/programs/TreeTool)
24 Page, R.D.M. TREEVIEW: An application to display phylogenetic trees on personal computers. *Comput. Applic. Biosci.* 1996, **12,** 357–358
25 Van de Peer, Y., and De Wachter, R. TREECON for Windows: A software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. *Comput. Applic. Biosci.* 1994, **10,** 569–570
26 Langrange, J.L. *Oeuvres,* Vol. 5. Paris, 1870. [See also, Flory, P.J. *The Statistical Mechanics of Chain Molecules.* Wiley Interscience, New York, 1969]
27 Whitney, H.A. A set of topological invariants for graphs. *Am. J. Math.* 1933, **55,** 321–325
28 Weiner, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* 1947, **69,** 17–20
29 Kier, L.B., and Hall, L.H. *Molecular Connectivity in Chemistry and Drug Research.* Academic Press, New York, 1976
30 Kier, L.B., and Hall, L.H. *Molecular Connectivity in Structure Activity Analysis.* Research Studies Press, Letchworth., UK, 1986
31 Basak, S.C., Niemi, G.J., and Veith, G.D. Predicting properties of molecules using graph invariants. *J. Math. Chem.* 1991, **7,** 243–272
32 Bonchev, D. *Information Theoretic Indices for Characterisation of Chemical Structures.* Research Studies Press, Letchworth, UK, 1983
33 Chan, O., Gutman, I., Lam, T.-K., and Merris, R. Algebraic connections between topological indices. *J. Chem. Inf. Comput. Sci.* 1998, **38,** 62–65
34 Gutman, I., and Mohar, B. The quasi-Weiner and Kir-

choff indices coincide. *J. Chem. Inf. Comput. Sci.* 1996, **36,** 982–985
35 Klavzar, S., and Gutman, I. A comparison of the Schultz molecular topological index with the Weiner index. *J. Chem. Inf. Comput. Sci.* 1996, **36,** 1001–1003
36 Flower, D.R. ALTER: Eclectic management of molecular structure data. *J. Mol. Graphics Modelling* 1997, **15,** 161–169
37 James, C.A., and Weininger, D. *Daylight Theory Manual*. Daylight Chemical Information Systems, Inc., Santa Fe 1995
38 Bleasby, A.J., and Wootton, J,C. Construction of validated, non-redundant composite protein sequence databases. *Protein Eng.* 1990, **3,** 153–159
39 Burden, F.R.A. Chemically intuitive molecular index based on the eigenvalues of a modified adjacency matrix. *Quant. Struct.-Activity Relationships* 1997, **16,** 309–314
40 Lindgren, F., Hellberg, S., and Flower, D.R. Unpublished observations, 1998
41 Holliday, J.D., Ranade, S.S., and Willett, P. A fast algorithm for selecting sets of dissimilar molecules from large chemical databases. *Quant. Struct.-Activity Relationships* 1995, **14,** 501–506
42 Snarey, M., Terrett, N.K., Willett, P., and Wilton, D.J. Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graphics Modelling* 1997, **15,** 372–385
43 Clark, R.D. Optisim: An extended dissimilarity selection method for finding diverse representative subsets. *J. Chem. Inf. Comput. Sci.* 1997, **37,** 1181–1188
44 Tripos, Inc. *UNITY Reference Manual*. Tripos, Inc., St. Louis, MO, 1995
45 Rishton, G.M. Reactive compounds and in vitro false positives in HTS. *Drug Disc. Today* 1997, **2,** 382–384
46 Lipinski, C.A., Lombard, F., Dominy, B.W., and Feeney, P.J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 1997, **23,** 3–25