Chemical Similarity Searching

Peter Willett*

Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Sheffield S10 2TN, U.K.

John M. Barnard and Geoffrey M. Downs

Barnard Chemical Information, 46 Uppergate Road, Sheffield S6 6BX, U.K.

Received February 27, 1998

This paper reviews the use of similarity searching in chemical databases. It begins by introducing the concept of similarity searching, differentiating it from the more common substructure searching, and then discusses the current generation of fragment-based measures that are used for searching chemical structure databases. The next sections focus upon two of the principal characteristics of a similarity measure: the coefficient that is used to quantify the degree of structural resemblance between pairs of molecules and the structural representations that are used to characterize molecules that are being compared in a similarity calculation. New types of similarity measure are then compared with current approaches, and examples are given of several applications that are related to similarity searching.

1. INTRODUCTION

Databases of two-dimensional (2D) or three-dimensional (3D) molecular structures play an increasingly important role in modern chemical research.[1−3] The most common type of retrieval mechanism is *substructure searching,* which involves the retrieval of all those molecules in a database that contain a user-defined query substructure, irrespective of the environment in which the query substructure occurs.[4] Substructure searching has proved to be a valuable tool for accessing databases of chemical structures, especially now that 3D *pharmacophore searching* is available to complement the long-established facilities for 2D substructure searching.[5] It does, however, have several limitations that arise from the requirement that a database structure must contain the entire query substructure if it is to be retrieved,[6] which implies that the user who is posing a database query must already have formed a fairly clear view of the types of structure that will be retrieved.

The first limitation of substructure searching is that the specification of a pharmacophore query requires sufficient knowledge of the geometric requirements for activity to be able to specify distance and/or angular constraints to characterize those molecules, and just those molecules, that can fit into a biological receptor site. Such pharmacophores are generally identified by comparing several bioactive molecules to identify the pattern of features that they have in common,[7] which is clearly very difficult at the start of an investigation when perhaps just a single weak lead is available and when it is thus not possible to specify the particular feature(s) that are responsible for the observed activity. The user also has very little control over the size of the output that is produced by a particular query substructure. Thus, the specification of a broadly defined query and/or a common ring system can result in the retrieval of many thousands of compounds from a chemical database (unless it is also possible to apply additional filters, such as a user-defined range of values for some physicochemical property); alternatively, an initial query may prove to be too specific, retrieving very few, or even no, structures. In either case, it may be necessary to reformulate the query one or more times before an appropriate volume of output is available for subsequent analysis. Finally, a substructure search results in a simple partition of the database into two discrete subsets, i.e., those structures that do contain the query and those that do not. There is thus no direct mechanism by which the retrieved molecules can be ranked in order of decreasing similarity to the query, i.e., in order of decreasing probability of activity if the search is intended to identify possible bioactives in the database.

These characteristics of substructure searching have led to the development of the alternative, and complementary, access mechanism known as *similarity searching*.[6] A query here generally involves the specification of an entire molecule, the *target structure*, rather than the substructure that is required for substructure searching (although the target can be a substructure of another, larger molecule if desired). The target is characterized by one or more structural descriptors, and this set is compared with the corresponding sets of descriptors for each of the molecules in the database. These comparisons enable the calculation of a measure of similarity between the target structure and each of the database structures, and the latter are then sorted into order of decreasing similarity with the target. The output from the search is a ranked list in which the structures that are calculated to be most similar to the target structure, the *nearest neighbors*, are located at the top of the list. These neighbors form the initial output of the search and will be those that have the greatest probability of being of interest to the user, given an appropriate measure of intermolecular structural similarity.

* To whom all correspondence should be addressed. E-mail: p.willett@sheffield.ac.uk.

10.1021/ci9800211 CCC: $15.00 © 1998 American Chemical Society
Published on Web 07/21/1998

In this paper, we provide an overview of chemical similarity searching. The next section introduces the 2D (two-dimensional), fragment-based measures of structural similarity that are used in the current generation of similarity searching systems. This is followed by more detailed discussions of similarity coefficients and structural representations, and the paper concludes by discussing other applications of similarity searching and by highlighting areas where further work is required. A more extended review of these, and other, aspects of molecular similarity is provided by Downs and Willett.[6]

2. FRAGMENT-BASED SIMILARITY SEARCHING

The first two reports of similarity searching appeared in the mid-1980s, based on work carried out at Lederle Laboratories[8] and at Pfizer.[9] The starting points for these two, near-contemporaneous studies were very different, but both groups of workers realized that counts of the numbers of fragment substructures common to a pair of molecules provided a computationally efficient, and surprisingly effective, basis for quantifying the degree of structural resemblance between the two molecules under consideration.

The Lederle study was carried out as part of a project to develop simple, robust techniques for the prediction of biological activity that would not suffer from the sample-to-feature problems that affect many types of high-dimensionality descriptors.[10] Molecules were represented by their constituent *atom pairs*, where an atom pair is a substructural fragment comprising two non-hydrogen atoms together with the number of intervening bonds (see section 4 below). These characterizations were used for two applications: for similarity searching, with the set of atom pairs describing a user-defined target structure being matched against the corresponding sets for each of the database structures, and *substructural analysis*, where weights are calculated that relate the presence of a specific substructural moiety in a molecule to the probability that that molecule is active in some biological test system.[11] The similarity search allowed users to request either some number of the top-ranked molecules or all those that had a similarity with the target structure greater than a minimal value. The latter search option does require the user to have at least some feeling for the magnitudes of the values resulting from the chosen similarity measure, but it serves to restrict the output to those molecules that do have a significant level of resemblance to the target structure.

The work at Pfizer started out as a way of prioritizing the outputs of 2D substructure searches from their in-house chemical information system. A user of the Pfizer system would submit not only a conventional substructural query but also a target molecule typical of the sorts of structure that were required. The conventional screen search and atom-by-atom search[4] were used to identify the matches to the query substructure, and then a similarity measure based on the screens common to the target and each of these matches was used to rank the substructure-search output in order of decreasing similarity with the query; specifically, the similarities were calculated using the Tanimoto coefficient that is discussed in the next section of this paper. At least in part, the initial substructural query was used to minimize the elapsed time required for the calculation of the similarities, by restricting the similarity calculation to just that small fraction of the database not eliminated by the substructure search. The subsequent development of a much faster nearest neighbor search algorithm, based on an inverted file, allowed the ranking of an entire database against the target structure in real time, without the need for the specification of the initial substructural query.

Interactive, fragment-based similarity searching has proved to be extremely popular, both for property prediction purposes (as in the work at Lederle) and for allowing end-users to pose “give me 10 more like this” queries (as in the work at Pfizer), and it is now a standard retrieval mechanism in nearly all operational systems for chemical information management. However, it must be remembered that similarity searching provides a very crude way of accessing a structural database, since it is appropriate when just a single bioactive molecule is available. Progressively more sophisticated approaches are appropriate as more structural data become available: substructure or pharmacophore searching when sufficient bioactive molecules are available to generate a query specification and a *docking* search (see section 5 below) when the 3D structure of the biological target is known. Even so, it makes obvious sense to exploit whatever information is available, and much effort has thus gone into the development of similarity searching since the Lederle and Pfizer systems were reported just over a decade ago. This work has involved both enhancements of fragment-based searching and the use of different types of similarity measure.

An example of an enhanced fragment-based system is provide by Hagadone’s work on *substructure similarity* (or *subsimilarity*) *searching*.[12] Conventional similarity searching is appropriate when the need is to identify complete structures that are similar to the target structure. Such a *global* similarity search,[6] i.e., one in which the entire matching structures are involved in the similarity calculation, is far less effective when the need is to identify molecules containing a substructure that is similar to a target structure or target substructure. This is an example of a *local* similarity search,[6] i.e., one in which account must be taken of parts of the molecules that are being compared and in which a more detailed similarity calculation is required. In subsimilarity searching, a simple, fragment-based similarity search is used to calculate an upperbound to the size (in terms of the numbers of constituent atoms or bonds) of the maximal common substructure (or MCS) between the target (sub)-structure and each database structure; these upperbounds are then used to prioritize database structures for an MCS search that uses a rapid, but approximate, maximal common subgraph isomorphism algorithm.

Another example of a similarity search system that uses fragment occurrence information in combination with a second-level search is described by Fisanick et al. as part of a project to develop facilities for similarity searching in the Chemical Abstracts Service (CAS) Registry File, using 2D, 3D, and molecular property data.[13−15] The 2D studies involved subsets of the substructural fragments that comprise the CAS Online screen dictionary,[16] focusing on the different types of similarity relationships that can be identified between a target structure and a database structure when different classes of substructural fragment are employed. For example, the selection of *augmented atoms* (an atom and its pendant atoms and bonds) and *atom sequences* (unbranched chains of atoms) gives a very different view of the structural resemblances between a pair of molecules from that provided by the selection of ring composition fragments (the atoms within a ring and the bonds between them). This suggests that further analysis into mixed descriptor types could give users an even more flexible approach to similarity searching, perhaps using the data fusion techniques discussed in the final section of this paper. Another part of the CAS work includes a second-stage search based on *reduced graphs*,[17−19] which, unlike substructural fragments, retain some of the topological relationships between areas of a molecule and which are thus capable of providing a local measure of similarity.

Both of the studies above thus involve the combination of a global, fragment-based, similarity algorithm with a more sophisticated, local, graph algorithm that allows some degree of substructural matching: ways of combining substructural constraints in a global similarity measure are discussed by Willett[20] and by Grethe and Hounshell.[21]

Further developments have focused on the similarity measure that is used to quantify the degree of structural resemblance between the target structure and each of the structures in the database that is to be searched. Many different types of similarity measure have been discussed in the literature,[6,22−24] but they generally involve three principal components: the *representation* that is used to characterize the molecules that are being compared, the *weighting scheme* that is used to assign differing degrees of importance to the various components of these representations, and the *similarity coefficient* that is used to provide a quantitative measure of the degree of structural relatedness between a pair of structural representations. While there has been some interest in the extent to which the weighting scheme affects the utility of a similarity measure,[13,25,26] there is a much more extended literature relating to the other two components, and these are hence reviewed in the next two sections of this paper.

3. SIMILARITY AND DISTANCE COEFFICIENTS

The idea of determining a numerical measure of the similarity (or conversely, the distance) between two objects, each characterized by a common set of attributes, is common to a wide range of disciplines, including biology, psychology, and bibliographic information retrieval. Because of the diversity of these application areas, and the lack of communication between them, there has been a great deal of duplication of effort, and commonly used similarity coefficients have been reinvented a number of times; this partly accounts for the variety of different names applied to some of these coefficients. This section reviews those coefficients that have found widespread use in chemical information systems; more comprehensive surveys of the very many coefficients available are provided by Hubálek,[27] Gower,[28] and Ellis et al.,[29] inter alia.

An object A can be described by means of a vector $\mathbf{X}_{\mathrm{A}}$ of n attributes such that

$$X_{\mathrm{A}} = \{x_{1\mathrm{A}}, x_{2\mathrm{A}}, x_{3\mathrm{A}}, ..., x_{j\mathrm{A}}, ..., x_{n\mathrm{A}}\}$$

where $x_{j\mathrm{A}}$ is the value of the jth attribute of object A, as detailed in Table 1 (which provides a complete list of the symbols used in this section of the paper). The values of the attributes may be real numbers over any range (and may involve some weighting factor applied to the basic property value involved), or they may be confined to dichotomous (i.e., binary) values, indicating the absence (0) or presence (1) of some particular feature of the object. In the case of a molecular object, the attributes might be a set of n topological indexes or calculated physicochemical properties, or the on/off state of each of the n bits in the fingerprint representing the molecule.

Table 1. Symbols Used

i,j	attributes
A, B	objects (or molecules)
n	total number of attributes of an object (e.g., bits in a fingerprint)
$\mathbf{X}_{\mathrm{A}}$	attribute vector describing object A
$x_{j\mathrm{A}}$	value of jth attribute in object A
χ_{A}	set of "on" bits in binary vector $\mathbf{X}_{\mathrm{A}}$
a	number of bits "on" in molecule A
b	number of bits "on" in molecule B
c	number of bits "on" in both molecules A and B
d	number of bits "off" in both molecules A and B
$S_{\mathrm{A,B}}$	similarity between objects A and B
$D_{\mathrm{A,B}}$	distance between objects A and B

Some coefficients are measures of the distance, or dissimilarity between objects (and have a value of 0 for identical objects), while others measure similarity directly (and have their maximum value for identical objects). In most cases the values that can be taken by a coefficient lie in the range from 0 to 1 or can be normalized to that range: this is typically effected by means of a function based on the values of the attributes for the two objects that are being compared, with the resulting coefficients being referred to as *association coefficients*. The zero-to-unity range provides a simple means for converting between a similarity coefficient and a complementary distance coefficient, namely, subtraction from unity. In some cases a similarity coefficient and its complement have been developed independently and are known by different names; e.g., the Soergel distance coefficient is the complement of the Tanimoto (or Jaccard) association coefficient.

Distance coefficients are analogous to distances in multidimensional geometric space, though they are not necessarily precisely equivalent to such distances. For a distance coefficient to be described as a *metric* it must have the following properties:

(1) Distance values must be zero or positive, and the distance from an object to itself must be zero:

$$D_{\mathrm{A,B}} \geq 0, \quad D_{\mathrm{A,A}} = D_{\mathrm{B,B}} = 0$$

(2) Distance values must be symmetric:

$$D_{\mathrm{A,B}} = D_{\mathrm{B,A}}$$

(3) Distance values must obey the *triangular inequality*:

$$D_{\mathrm{A,B}} \leq D_{\mathrm{A,C}} + D_{\mathrm{C,B}}$$

(4) The distance between nonidentical objects must be greater than zero:

$$A \neq B \Leftrightarrow D_{\mathrm{A,B}} > 0$$

A distance coefficient which has only the first three of these properties is called *pseudometric*, and one which does not have the third property is *nonmetric*. Though a particular distance coefficient may have all four properties, this is not sufficient to imply that the distances involved can be embedded in a Euclidean space of any given dimensionality (e.g., n, the number of properties). Certain other properties are necessary, and even then the dimensionality of the space required may be much larger than n. The requirements for Euclidean embedding are discussed by Gower.[28]

Though a large number of similarity and distance coefficients have been defined (and often redefined, by different authors), many of them are closely related to each other. In some cases, the same coefficient can be obtained by different routes; in other cases coefficients which are different when calculated for continuous attributes become equivalent when applied to binary attributes. Certain coefficients are described as being *monotonic* with each other, which means that it can be shown analytically that they will always produce identical similarity rankings of objects against a specified target, even though the actual coefficient values are different. Even though two coefficients may not be completely monotonic, the values resulting from their use may well exhibit a high degree of correlation, as demonstrated by Holliday et al. in a comparison of the Cosine and Tanimoto coefficients.[30] Some pairs of coefficients, conversely, exhibit very low correlations, suggesting that they are reflecting very different characteristics of the objects that are being compared;[29] an extended empirical study of the monotonicity relationships existing between no less than 43 different coefficients is reported by Hubálek.[27]

Where the attribute values are restricted to 0 and 1, the expressions used for the various similarity and distance measures can often be substantially simplified. In this context a number of useful symbols can be defined. For objects A and B characterized by vectors $\mathbf{X}_A$ and $\mathbf{X}_B$ containing n binary values (such as fingerprints) we can write

$$a = \sum_{j=1}^{j=n} x_{jA} \qquad \text{number of bits “on” in A}$$

$$b = \sum_{j=1}^{j=n} x_{jB} \qquad \text{number of bits “on” in B}$$

$$c = \sum_{j=1}^{j=n} x_{jA}x_{jB} \qquad \text{number of bits “on” in both A and B}$$

$$d = \sum_{j=1}^{j=n} (1 - x_{jA} - x_{jB} + x_{jA}x_{jB}) \qquad \text{number of bits “off” in both A and B}$$

and hence

$$n = a + b - c + d$$

Note that the definitions of a and b shown here differ from those given by Gower[28] and by Ellis et al.;[29] they are, however, the definitions that have been more commonly used in the chemical information literature. The various quantities above can also be expressed in set-theoretic notation, if we define χ_A as the set of all elements x_{jA} in vector $\mathbf{X}_A$ whose value is 1 (the “on” bits) and χ_B as the set of all elements x_{jB} in vector $\mathbf{X}_B$ whose value is 1. Then

$$a = |\chi_A|$$

$$b = |\chi_B|$$

$$c = |\chi_A \cap \chi_B|$$

$$d = n - |\chi_A \cup \chi_B|$$

and, as a corollary to the above, the number of bits “on” in at least one of the molecules is given by

$$a + b - c = |\chi_A \cup \chi_B|$$

Given the above definitions, Table 2 describes a number of similarity and distance coefficients commonly used in chemical information, with expressions for calculating them for continuous-variable or dichotomous attributes or using set notation.

Both the Hamming distance and the Euclidean distance are examples of a more general class of distance metrics called Minkowski distances which are given by the general formula

$$D_{A,B} = [\sum_{j=1}^{j=n} (|x_{jA} - x_{jB}|)^t]^{1/t}$$

where $t = 1$ for the Hamming distance and $t = 2$ for the Euclidean distance.

A fundamental difference between the Hamming and Euclidean distances, on one hand, and the Tanimoto, Dice, and Cosine coefficients, on the other, is that the former effectively consider a common absence of attributes (or common low values in the case of continuous variables) as evidence of similarity, whereas the latter do not. This is a basic philosophical argument, which has been much discussed in the literature. In the context of numerical taxonomy, Sokal and Sneath[31] have commented:

“The absence of wings ... among a group of distantly related organisms (such as a camel, a horse, and a nematode) would surely be an absurd indication of affinity. Yet a positive character such as the presence of wings...could mislead equally...for a similar heterogeneous assemblage (for example, bat, heron, and dragonfly).”

In the chemical context, James et al.[32] have suggested that Hamming and Euclidean distances are useful only for “relative” distance comparisons (i.e., the distance of two molecules to the same target) but not for “absolute” comparisons (between two independent pairs of molecules), for which they prefer the Tanimoto coefficient. Nevertheless, Euclidean distance comparisons form the basis of Ward’s hierarchical agglomerative clustering method,[33] which has been shown to be particularly effective on the basis of empirical studies and which is discussed later in this paper. It is also worth noting that a number of familiar chemical concepts are essentially negatively defined; for example, the common feature of carbocycles is the lack of heteroatoms and the common feature of aliphatic systems is the lack of aromaticity.

A further fundamental difference is that the association coefficients involve a normalization factor that helps to lessen molecular size effects in some cases. Thus, in a similarity search using fragment bit-strings or fingerprints, a large molecule in the database is a priori much more likely to have

Table 2. Descriptions of Some Distance Metrics and Similarity Coefficients Commonly Used in Chemical Information[a]

	Hamming Distance
other names	Manhattan distance City-Block distance normalized complement for dichotomous data called simple matching coefficient
formula for continuous variables	$D_{A,B} = \sum_{j=1}^{j=n} \lvert x_{jA} - x_{jB} \rvert$
formula for dichotomous variables	$D_{A,B} = a + b - 2c$
set-theoretic definition	$D_{A,B} = \lvert \chi_A \cup \chi_B \rvert - \lvert \chi_A \cap \chi_B \rvert$
range	∞ to 0 (continuous), n to 0 (dichotomous)
metric properties	obeys all four metric properties
notes	equivalent to the squared Euclidean distance for dichotomous variables can be normalized to the range 1 to 0 if the values of all attributes are normalized to this range and the result divided by n
	Euclidean Distance
other names	none
formula for continuous variables	$D_{A,B} = [\sum_{j=1}^{j=n} (x_{jA} - x_{jB})^2]^{1/2}$
formula for dichotomous variables	$D_{A,B} = [a + b - 2c]^{1/2}$
set-theoretic definition	$D_{A,B} = [\lvert \chi_A \cup \chi_B \rvert - \lvert \chi_A \cap \chi_B \rvert]^{1/2}$
range	∞ to 0 (continuous), n to 0 (dichotomous)
metric properties	obeys all four metric properties
notes	frequently used as its square (with which it is, of course, monotonic), which avoids the need to take the square root in the calculation monotonic with the Hamming distance in all cases (and its square is equivalent to the Hamming distance for dichotomous variables) can be normalized to the range 1 to 0 if the values of all attributes are normalized to this range and the result divided by n
	Soergel Distance
other names	none
formula for continuous variables	$D_{A,B} = [\sum_{j=1}^{j=n} \lvert x_{jA} - x_{jB} \rvert]/[\sum_{j=1}^{j=n} \max(x_{jA}, x_{jB})]$
formula for dichotomous variables	$D_{A,B} = 1 - c/[a + b - c] = [a + b - 2c]/[a + b - c]$
set-theoretic definition	$D_{A,B} = [\lvert \chi_A \cup \chi_B \rvert - \lvert \chi_A \cap \chi_B \rvert]/\lvert \chi_A \cup \chi_B \rvert$
range	1 to 0
metric properties	obeys all four metric properties provided all attributes have nonnegative values
notes	for dichotomous variables only, the Soergel distance is identical to the complement of the Tanimoto coefficient
	Tanimoto Coefficient
other names	Jaccard coefficient
formula for continuous variables	$S_{A,B} = [\sum_{j=1}^{j=n} x_{jA} x_{jB}]/[\sum_{j=1}^{j=n} (x_{jA})^2 + \sum_{j=1}^{j=n} (x_{jB})^2 - \sum_{j=1}^{j=n} x_{jA} x_{jB}]$
formula for dichotomous variables	$S_{A,B} = c/[a + b - c]$
set-theoretic definition	$S_{A,B} = \lvert \chi_A \cap \chi_B \rvert / \lvert \chi_A \cup \chi_B \rvert$
range	-0.333 to $+1$ (continuous), 0 to $+1$ (dichotomous)
metric properties	complement does not obey the triangular inequality in general, though does obey it if dichotomous variables are used
notes	monotonic with the Dice coefficient complement of the dichotomous version is identical to the Soergel distance
	Dice Coefficient
other names	Czekanowski coefficient Sørenson coefficient essentially equivalent to the Hodgkin index for overlap of electron density functions
formula for continuous variables	$S_{A,B} = [2\sum_{j=1}^{j=n} x_{jA} x_{jB}]/[\sum_{j=1}^{j=n} (x_{jA})^2 + \sum_{j=1}^{j=n} (x_{jB})^2]$
formula for dichotomous variables	$S_{A,B} = 2c/[a + b]$
set-theoretic definition	$S_{A,B} = 2\lvert \chi_A \cap \chi_B \rvert / [\lvert \chi_A \rvert + \lvert \chi_B \rvert]$
range	-1 to $+1$ (continuous), 0 to $+1$ (dichotomous)
metric properties	complement does not obey the triangular inequality
notes	monotonic with the Tanimoto coefficient

Table 2. (Continued)

Cosine Coefficient	
other names	Ochiai coefficient essentially equivalent to the Carbo index for overlap of electron density functions
formula for continuous variables	$S_{\mathrm{A,B}} = [\sum_{j=1}^{j=n} x_{j\mathrm{A}}x_{j\mathrm{B}}]/[\sum_{j=1}^{j=n}(x_{j\mathrm{A}})^2\sum_{j=1}^{j=n}(x_{j\mathrm{B}})^2]^{1/2}$
formula for dichotomous variables	$S_{\mathrm{A,B}} = c/[ab]^{1/2}$
set-theoretic definition	$S_{\mathrm{A,B}} = \lvert\chi_\mathrm{A} \cap \chi_\mathrm{B}\rvert/[\lvert\chi_\mathrm{A}\rvert\lvert\chi_\mathrm{B}\rvert]^{1/2}$
range	-1 to $+1$ (continuous), 0 to $+1$ (dichotomous)
metric properties	complement does not obey the triangular inequality
notes	highly correlated with the Tanimoto coefficient, though not strictly monotonic with it

[a] Definitions of the symbols used are shown in Table 1. Note that the negative lower-bound values for the three association coefficients apply only if negative attribute values are possible.

bits in common with the target structure than is a small molecule, and it is thus appropriate to include some degree of size normalization in the coefficient to avoid a bias in the nearest neighbors toward the largest database molecules. The converse of this problem can arise in diversity applications, for example in dissimilarity-based compound selection procedures where one seeks to identify database subsets for which the constituent molecules are as dissimilar as possible.[34] Small molecules are likely to have few bits set in a fingerprint: since the Tanimoto coefficient, for example, does not take account of a common absence of features and since $c \leq \min(a,b)$ (in the coefficient's expression in Table 2), low-similarity (and thus high-dissimilarity) values will be obtained with small molecules, thus possibly biasing the size distribution in the final subset that is selected. A solution adopted at Pharmacia and Upjohn[35] is to use a composite coefficient essentially involving both the Tanimoto coefficient and the simple matching coefficient (the complement of the normalized Hamming distance).

Following earlier work by Adamson and Bush,[36] Willett and Winterman[37] compared the performance of a range of similarity and distance coefficients by the extent to which they obeyed the *similar property principle* of Johnson and Maggiora;[23] specifically, they assessed the effectiveness of a coefficient by the extent to which it was able to predict correctly a compound's measured property or activity value as the value of the most similar compound in the same dataset. In this study, the Tanimoto and Cosine coefficients performed rather better than the Hamming and Euclidean distance measures, and in an operational system implemented subsequently,[9] the Tanimoto coefficient was preferred, partly on the basis of a subjective evaluation of the similarity search rankings it produced and partly because its calculation does not involve a square root, making it faster. Since then, the Tanimoto coefficient has been the measure of choice for fragment-based chemical similarity work, though the Hamming distance (equivalent to the squared Euclidean distance for binary data) retains its adherents and the Euclidean distance remains the most popular measure for continuous data.

Other criteria can be used to evaluate similarity coefficients. For example, Cheng et al.[38] have described four association coefficients for assessing the degree of relatedness between pairs of different similarity coefficients. Their study, which again draws upon the similar property principle, was used to compare different coefficients based on different descriptor sets (Euclidean distances with topological indexes and Tanimoto coefficient with 2D bit-strings), but the same principles could also be applied to coefficients based on the same descriptors. Computational efficiency can also merit consideration as a basis for comparison. For example, the Cosine coefficient allows the calculation of the average similarity between all pairs of compounds in two disjoint datasets extremely rapidly, something that is not possible with the Tanimoto coefficient[30] and that may be necessary for some similarity applications. Finally, the behavior of a coefficient over its range of possible values may give guidance as to its suitability for use in a particular application domain, as evidenced by the continuing discussion as to which similarity coefficient is most appropriate for the calculation of field-based similarities.[26,39−42]

Bradshaw[43] has recently drawn attention to the use of asymmetric similarity coefficients (in which $S_{\mathrm{A,B}} \neq S_{\mathrm{B,A}}$) based on the ideas of Tversky.[44] The general form for Tversky similarity is defined for binary data as follows:

$$S_{\mathrm{A,B}} = \frac{c}{\alpha(a-c) + \beta(b-c) + c}$$

where α and β are user-defined constants. If α and β are equal, the resulting similarity coefficient is symmetric, and in the case of certain values, the expression reduces to one of the commonly known coefficients: the Tanimoto coefficient when $\alpha = \beta = 1$ and the Dice coefficient when $\alpha = \beta = {}^1/_2$. If α and β are different, the resulting coefficient is asymmetric, and when $\alpha = 1$ and $\beta = 0$ yields $S_{\mathrm{A,B}} = c/a$, which can be interpreted as the "fraction of A" which it has in common with B; the coefficient will become equal to 1 when all the features of A are also in B—i.e., when A is (within the constraints of a fragment-based representation) a substructure of B, and features of B which do not occur in A are irrelevant to the similarity value. This type of subsimilarity expression has also been derived by Maggiora et al.[45] using an approach based in fuzzy-set theory and provides an alternative subsimilarity measure to the MCS described by Hagadone[12] and discussed in section 2.

In conclusion, we reiterate the fact that the discussion here has concentrated on those coefficients, and their close relations, that have been most extensively used, thus far, for chemical applications. There are many others that have been discussed in the literature of, e.g., multivariate statistics,[28] information retrieval,[29] and numerical taxonomy,[31] and it must not be assumed that there is any single "best" coefficient even if we restrict attention to the domain of chemical structure handling. Indeed, as noted by Jones and Curtice[46] in a discussion of the association between indexing terms in information retrieval systems:

> "What is annoying is that no clear-cut criterion for choice among the alternatives has emerged. As a result, few candidate measures have been permanently dismissed from consideration, and a rather large set of formulas remains available."

There is hence a continuing need for both empirical and analytical comparisons of the available coefficients to ensure that the most appropriate one(s) are employed in any specific similarity-based system.

4. STRUCTURAL REPRESENTATIONS FOR SIMILARITY SEARCHING

Similarity searching in large chemical databases needs representations of the molecules that are both *effective*, i.e., can differentiate between molecules that are different, and *efficient*, i.e., quick to calculate, in operation. There is a general conflict between these two requirements in that the most effective methods of representation tend to be the least efficient to calculate, and *vice versa*, so a suitable compromise needs to be made. For instance, quantum-mechanical descriptions, such as the electron probability density function described by Carbo and Calabuig,[47] take too long to calculate whereas, at the other extreme, simple atom and bond counts are generally too trivial to discriminate between many molecules. In the middle lie descriptors based on 2D and 3D substructural fragments or properties. These are the ones that are currently most commonly used for similarity searching and that form the principal focus of the subsequent discussion. This overview is necessarily brief; further details are given in the recent review by Brown.[48]

The representation of molecules by descriptors involves the generation of suitable descriptors and, if desired, the selection of a subset of them and then the encoding of the chosen descriptors in a form that will enable similarity calculation between pairs of representations. Many of the descriptors are described in the literature along with a particular encoding method; however, the two are largely independent, and it is usually possible to encode a given descriptor in a variety of ways. We have thus deliberately separated descriptor selection and descriptor encoding in this section to highlight the two stages.

Descriptor Selection. There is an infinite variety of potential descriptors, so descriptor selection is necessary as an exercise in data reduction to select those most appropriate to a given application. The following subsections will examine examples of counts, 2D-fragment, 3D-fragment, and physicochemical property descriptors, topological indices, whole molecule comparisons, and the issue of descriptor choice.

The simplest descriptors are counts of individual atoms, bonds, degrees of connectivity, etc. These can be extended to counts of rings, pharmacophore points, and any other feature that can be represented as a single node or arc in the graph or reduced graph representation of the molecule.

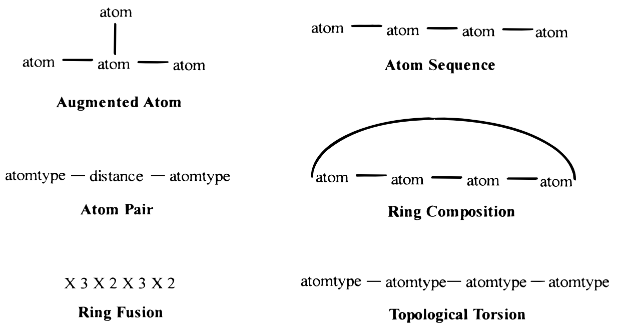
atom
atom — atom — atom
Augmented Atom
atom — atom — atom — atom
Atom Sequence
atomtype — distance — atomtype
Atom Pair
atom — atom — atom — atom
Ring Composition
X 3 X 2 X 3 X 2
Ring Fusion
atomtype — atomtype— atomtype — atomtype
Topological Torsion

Figure 1. Example 2D-fragment descriptors.

Two-dimensional fragment descriptors were first studied in detail by Lynch and co-workers (see, e.g., ref 49), who investigated the use of various types of atom-centered, bond-centered, and ring-centered fragments for substructure searching. This work led to the widespread adoption of augmented atom, atom sequence and ring fragments in substructure search systems, e.g., the fragments in CAS Online.[16] Some typical 2D fragment definitions are shown in Figure 1. Augmented atoms comprise a central atom with the neighboring attached atoms and intervening bonds. Atom sequences are linear sequences of a given number of connected atoms, with their intervening bonds. Ring fragments can be of several different types, for instance the *ring sequence* (atom sequence round a ring) and *ring fusion sequence* (ring-connectivity counts round a ring) fragments. Other fragment definitions, originally developed at Lederle Laboratories, that have become popular for similarity searching are the *atom pair*[8] and *topological torsion*[50] fragments. Atom pairs comprise a pair of atom types and the intervening distance between them, in terms of the shortest bond-by-bond path between them. The atom type describes the elemental type, the number of non-hydrogen attachments, and the number of π-bonds. The topological torsion fragment comprises a linear sequence of four connected atoms, with each atom type described in the same way as for atom pairs.

Workers at CAS found that the use of specific elemental and bond types for atoms and bonds can be too specific for substructure searching, and generalized forms of these fragments are thus often used (and similarly so for similarity searching). For instance, atoms can be generalized to groups such as their group in the periodic table, and bonds to ring or chain, allowing the specification of many combinations of generalized atoms and bonds, and similar definitions can be used for similarity searching. The generalization of atom pairs and topological torsions by use of physicochemical atom types is mentioned later.

An alternative to the algorithmically generated descriptors described above is to define and to search for particular functional groups, which may be expressed in specific or general terms. Once defined, the functional groups can be detected by scanning the connection tables for instances of them. A more efficient way is to use string-searching of a linear representation of the molecule, for instance by defining the functional groups in terms of SMARTS strings and using them to search SMILES representations.[51]

As noted in section 2, 2D fragment descriptors rapidly established themselves as the basis for operational similarity searching, and it was some years before attempts were made to develop fragments for 3D similarity searching: some examples of 3D fragment descriptors are shown in Figure 2. Many of the 2D fragments that can be generated from a 2D connection table have equivalents in the 3D fragments that can be generated from a 3D connection table. However,

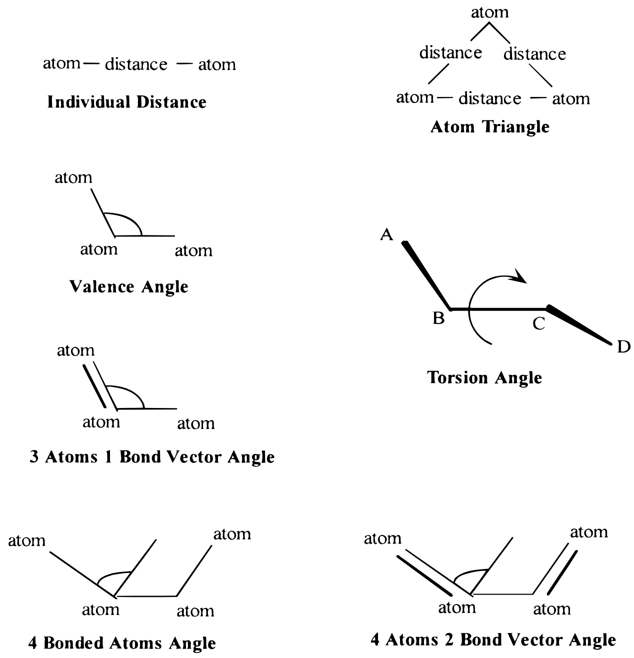

Figure 2. Example 3D-fragment descriptors.

there is much less consensus in the 3D area as to which are the best descriptors to use; moreover, the variety of available descriptors is greater, and new fragment types continue to be developed. Due to the fully connected nature of a 3D connection table, and the flexibility of 3D structures, the number of 3D fragments generated for a given class can be much larger than for the 2D equivalent, and the generation process can be more time-consuming.

Willett and co-workers have described both distance-based[52] and angle-based[53] descriptors for the calculation of 3D similarity. The simplest of the distance-based descriptors is the *distance distribution* in which each distance in a molecule increments a count in an associated distance-range bin. The resultant frequency distribution of distances is used to represent the molecule. To include elemental types, *individual-distance* descriptors comprise a pair of atoms and the interatomic distance between them. The angle-based descriptors are based on generalized valence angles and torsion angles, in which the atoms comprising the angle do not need to be directly bonded to each other. The distance-based descriptor described by Bemis and Kuntz[54] is an extension of the distance-distribution descriptor and uses the distances between triplets of atoms. For each triplet in a molecule, the three interatomic distances are squared and summed to give a single value, and the distribution of these values is used to describe the molecule. Closely related to this is the *atom triplet* descriptor of Nilakantan et al.[55] Here, the three distances between the three atoms comprising the triplet are sorted into increasing length; the first is left alone, the second is multiplied by 10^3 and the third by 10^6, and then they are summed to produce a single integer value.

The group at Merck have described 3D variants of the atom pair, referred to as *geometric atom pairs* and *geometric binding property pairs*.[56] In geometric atom pairs, the atom types are defined as for standard atom pairs, but the distance between them is the through-space distance rather than the through-bond distance. In geometric binding property pairs, the distance is through-space distance and the atom type is generalized to one of seven binding classes (cation, anion, H-bond donor, H-bond acceptor, polar, hydrophobic, and other). Similarly, the group at Abbott Laboratories has compared a wide variety of descriptors[51] including two in-house descriptors based on *potential pharmacophore points* (PPPs). Five points are defined: H-bond donor, H-bond acceptor, positively charged, negatively charged, and hydrophobic. All atoms of the molecule are analyzed to see whether they can be classed potentially as one of the point types. The descriptors are *PPP-pairs* and *PPP-triangles*. PPP-pairs are similar to geometric atom pairs, with the atom types represented by PPP types. PPP-triangles are triplets of PPPs and their associated distances (categorized into bin ranges).

Eight classes of 3D descriptors have been investigated at CAS.[13] The *atom pair distance* and *three-bonded atoms angle* are generalized distance-distribution and valence angle descriptors, respectively (with atom types carbon, hetero, and any). The *three atoms and one bond vector* angle descriptor is a hybrid atom triangle, and the *four-bonded atoms* and *four atoms and two bond vectors* angle are hybrid topological torsion descriptors, each with selected angle information added to the generalized atom type information. The "*atom triangles*" are atom triangles with generalized atom types, and the *atom triangle three-slot* and *atom triangle five-slot* are reduced and generalized atom triangle descriptors using two atoms and one or three of the interatomic distances, respectively. The detailed search results provided by Fisanick et al.[13] demonstrate that these triangle-based features provide a simple and effective mechanism for similarity searching based on size and shape.

The CAS workers have also investigated the use of calculated molecular properties.[13] Twenty whole-molecule properties were tested, such as ClogP, molar refractivity, ionization potential, HOMO, and LUMO. The resultant values can be used directly as descriptors. In addition, several localized properties were included, such as atomic electron densities and eigenvalues for molecular orbitals, with the resultant values being binned into ranges to provide sets of descriptors for the whole molecule. A subset of the global properties subsequently formed the basis for the similarity measures used in a comparison of various clustering methods.[57] Similar work has been reported by Kearsley et al.,[58] who have generalized the atom pair and topological torsion descriptors by replacing the atom types by physicochemical properties.[58] *Binding property pairs* and *binding property torsions* have the atom types replaced by one of seven binding property groups (cations, anions, H-donor, H-acceptor, polar, hydrophobic, and other). Hydrophobic pairs and torsions, and charge pairs and torsions have the continuous values split into seven overlapping bins. Unlike most descriptors, the charge pairs and torsions consider hydrogen atoms as atoms.

Topological indices are similar to physicochemical properties in that they characterize some aspect of molecular data by a single value. Very many different topological indices have been, and continue to be, described in the quantitative structure−activity relationship (QSAR) literature, but most are highly correlated. One of the few published uses of topological indices for similarity calculation in large databases is that by Basak et al.,[59] who generated 90 topological indices, encoding shape, size, bonding pattern, and branching pattern. Principal components analysis of these indices identified 10 principal components that were then used as the descriptors for the similarity analysis. Topological indices are often used in conjunction with other descriptors, as exemplified by work at Rhone-Poulenc Rorer.[60] This study of 49 molecular properties (of which just over half were topological indices and a quarter were counts) identified 6 descriptors that were relatively uncorrelated and that covered steric, electronic, and hydrophobic aspects: the selected descriptors comprised three topological indices (flexibility, normalized electrotopological, and aromatic density), two fragment-based properties (H-donor and H-acceptor), and one physicochemical property (ClogP).

With the increase in computing power and development of more efficient algorithms, there are now several relatively efficient ways of mapping whole, or large parts, of molecules onto each other for comparison.

The superpositioning by permutations (SPERM) method has been developed at Organon[61] following work by Dean et al.[62] A molecule is placed at the center of a tessellated icosahedron, the vertices of which denote points at which some chosen physical property is calculated. In the case of shape similarity, the property at each vertex is either the minimum distance or the radial distance from the vertex to the molecule's surface. Several optimizations have been developed to enable alignment of pairs of molecules in reasonable time and then calculation of the similarity between the values at each point pair. Dean's group has described a range of methods for calculating 3D similarities;[63] thus far, only the tessellated icosahedron approach has been applied to database searching, but it is likely that improvements in computing speeds will enable others of his methods to be used in this context in the future.

The *atom mapping* method[52] compares the 3D environment of each atom in one molecule with the 3D environment of each atom in a second molecule. The resultant list of interatomic similarities is used to identify pairs of geometrically similar atoms, which are used as descriptors in the calculation of the overall intermolecular similarity. The atom mapping method was compared with an MCS procedure, where the 3D MCS was defined to be the largest set of atoms, in common between two molecules, that have matching interatomic distances (within a given tolerance). Comparative experiments using the similar property principle suggested that the atom mapping and MCS procedures were of comparable effectiveness; however, the former is far more efficient to calculate and formed the subsequent basis for a 3D similarity searching system developed at Zeneca Agrochemicals.[64]

Given the wide variety of descriptor types available, it is necessary to select the most appropriate structural representation for a given application. A recent, detailed comparative study is reported by Brown and Martin,[65] who have analyzed various descriptors (and encoding methods) to find those most relevant to ligand–receptor binding. Two-dimensional structural descriptors contain a lot of information about the physical properties and reactivity of a molecule and are quick to calculate. Augmented atoms are very localized, atom sequences are less so, and atom pairs span the whole of a 2D structure. Ring descriptors are essential for cyclic structures, and topological torsions correlate well with 3D torsions except in highly folded molecules. Physical properties and topological indices can be useful representations of hydrophobic and electrostatic interactions. Three-dimensional shape descriptors can give useful information about dispersion and steric interactions. The larger 3D descriptors tend to be the most effective, but the flexibility of many molecules can increase the time required to generate 3D descriptors and/or can also decrease their effectiveness. Generalization of atom types from specific elements to groups or properties can help to reveal broader similarities. Overall, the trend is to use combined descriptors or descriptor sets which contain many different descriptor types, at many different levels of generalization.

Descriptor Encoding. Having discussed some of the types of descriptor that are available, we now describe how they can be encoded to enable similarity calculations to be carried out.

The representation that is overwhelmingly used as a basis for similarity calculations in large databases is the fixed-length bit-string. This contains a fixed number of bits in which each bit can represent the absence (0) or presence (1) of some feature, either on its own or in conjunction with other bits in the bit-string. The binary bit-string is usually used for 2D and 3D fragment descriptors. Discrete variables, with more than two values, can be represented in the binary bit-string by using a bit for each possible value or for given ranges of values. Continuous variables can be represented by defining ranges of values and then assigning a bit to each range, a process known as *binning*. The ranges covered by each bin can be separate or be overlapping, as is done, e.g., in the geometric atom pair descriptors of Sheridan et al.[56] The ranges can also be equidistant, equifrequent, or user-/application-defined. Equidistant ranges, as the name implies, have the same interval. Equifrequent ranges have different intervals, each interval being derived from examination of the frequency distribution of the descriptor being represented[56] or an equation of the distribution.[51] For specialized applications, where distinct peaks in the distribution are known, the user may define the bin ranges manually.

Bit-strings can be directly, dictionary, or hash assigned, as described below. Examples of alternatives (dataprints and distribution-comparisons) applicable to certain descriptor types are also mentioned.

Descriptors with fixed limits (e.g. number, range of sizes, elemental composition) can be directly assigned to positions in a bit-string, with offsets being calculated to assign different groups or different descriptors to separate areas of the bit-string. For example, the ring descriptors devised by Downs et al.[66] were directly assigned to the end of a bit-string, offset to avoid the beginning (which was reserved for dictionary-assigned augmented atom and atom sequence descriptors). The diverse-property derived (DPD) code developed at Rhone-Poulenc Rorer[60] could also be directly assigned. The DPD contains the six descriptors described earlier, each split into a number of classes (from 2 to 4) giving a total of 17 bit positions (432 combinations).

Systems based on dictionary-assigned bit-strings employ a dictionary that specifies correspondences between particular functional groups or fragments and bit positions in a bit-string, with each entry (structural key) in the dictionary being assigned a bit position (screen number). Dictionaries of functional groups tend to be fairly small, so all groups can be listed in the dictionary and assigned to a short bit-string.

However, analysis of a database to generate several different fragment types typically produces many tens or hundreds of thousands of distinct fragments, with a highly skewed distribution (i.e., a few fragments occur very frequently and many occur very infrequently). Methods to reduce this number to fit into a fixed-length bit-string of a few thousand bits, while retaining those fragments that act as the best descriptors, include the following: statistical analysis of the fragment frequencies to remove very frequent/infrequent fragments (for substructure searching equifrequent occurrence of fragments gives better screenout; for similarity calculations frequency is less important); generalization of specific fragments to less specific forms which cover many different, but related, specific fragments; assignment of the same screen number to several different, but related (e.g., by co-occurrence or composition) fragments (co-occurrence is particularly relevant for similarity calculations since it biases the measure toward that feature).

Development of the CAS ONLINE Screen Dictionary used such methods to create a dictionary suitable for substructure searching.[16] Careful selection of very generalized fragments can give good representations for similarity calculations using relatively few bits; for example, Brown and Martin found that effective searches could be achieved using a small subset of MDL Information Systems' MACCS keys.[51] However, selection of appropriate fragments to include in a dictionary is tedious, at best; features not represented in the dictionary can never be reflected in the similarity measure, and the resulting structural resemblances may be strongly database-dependent.

Rather than selecting a subset of fragments for inclusion in a dictionary, so that the number of screens is reduced to the same as the length of the bit-string, hash-assigned bit-strings are created by fitting all of the fragments into the bit-string. This can be achieved by hashing the fragment to generate one or more integers that fall within the length, or a given subrange of the length, of the bit-string (fingerprint). The more integers generated by the hash function, the more unique patterns can be superimposed on the bit-string, so the more fragments can be included. This fingerprint approach is used, for example, by Daylight Chemical Information Systems Inc. and Tripos Inc. for both substructure searching and similarity searching. However, overlaps between patterns can lead to many patterns being overlaid by other patterns, with a consequent loss of information. For similarity calculations this can give rise to false similarities since common bits in two bit-strings may have been set by completely unrelated fragments. Adding all fragments can also give problems by including many co-occurring fragments, by including large numbers of fragments that are unrelated to the similarity relationships that the measure is seeking to quantify and by swamping the effects of those fewer fragments that are so related. These problems are exacerbated by the technique of folding the bit-strings to condense the information further.

Physicochemical property and topological index descriptors are usually represented using a fixed-length string of real numbers (sometimes referred to as a *dataprint*). Dataprints typically have far fewer elements than bit-strings, and each element has a value. Dataprints thus describe molecular space by a full matrix rather than the sparse matrix description given by bit-strings. To avoid biases caused by differences in magnitude of the descriptors (particularly physicochemical properties), it is usual to normalize each element of a dataprint by the range or standard deviation of that element throughout the dataset.[26] The frequency distribution of many descriptors (particularly 3D) can also be used directly as an encoding of the descriptors for similarity calculation (see, e.g., refs 52 and 55). If the distributions have the same number of elements, then a similarity coefficient or distance can be calculated in much the same way as that for dataprints.

Finally, descriptors such as those developed for whole molecule comparisons are often already encoded in a form suitable for similarity calculations, so no further encoding is necessary, e.g., the similarity measures based on surface comparisons that are discussed by Dean and Perkins.[63]

5. APPLICATIONS OF SIMILARITY SEARCHING

There are many applications of the similarity measures we have described above, including database clustering, docking searches, reaction similarity searching, and the analysis of molecular diversity. Here, we give just a few leading references to work in the first three of these domains, with more detailed discussions being provided in the literature cited; molecular diversity is discussed elsewhere in this special issue of *J. Chem. Inf. Comput. Sci.*[67]

Cluster analysis, or clustering, is the process of subdividing a group of objects (chemical molecules in the present context) into groups, or clusters, of objects that exhibit a high degree of both intracluster similarity and intercluster dissimilarity.[31,68] It is thus possible to obtain an overview of the range of structural types present within a dataset by selecting one (or some small number) of the molecules from each of the clusters resulting from the application of an appropriate clustering method to that dataset. The representative molecule for each cluster is either selected at random or selected as being the closest to the center of that cluster. These representative compounds can be used to maximize the efficiency of random screening in lead-discovery programs: if a representative compound proves active when tested in the bioassay of interest, then it is appropriate to assay the other compounds in its cluster since these may also exhibit the activity of interest; alternatively, if it proves inactive, then attention should be transferred to another cluster.[69,70]

Very many different clustering methods have been described in the literature. An early study of over 30 hierarchic and nonhierarchic methods[22] showed that the best results were obtained from Ward's hierarchical-agglomerative method,[33] with the nonhierarchical nearest neighbor method of Jarvis and Patrick[71] performing almost as well. In the mid-1980s, when these comparative experiments were carried out, computer limitations (in terms of both raw CPU speeds and the clustering algorithms available) meant that Ward's method could not be applied to databases of substantial size. The Jarvis–Patrick method was thus rapidly adopted as the clustering method of choice in commercial chemical database software, not only to select compounds for random screening but also to cluster the outputs of substructure searches that retrieve very large numbers of molecules, thus providing the searcher with an overview of the structural classes that contain the substructure of interest.[72] However, the method does have limitations (see, e.g., ref 73), and subsequent comparisons[51,57,65] have reaffirmed the general superiority of Ward's method. The availability of improved computer hardware and of the efficient *reciprocal nearest neighbors* algorithm[74] means that sequential implementations of this method can now be used on databases containing up to perhaps a quarter of a million structures in an acceptable amount of time; larger datasets, however, still require use of the Jarvis−Patrick method or of an appropriate parallel machine. Thus far, the great majority of clustering studies have used the simple fragment-based similarity measures described in the second section of this paper; however, any measure could be used if it could be calculated sufficiently rapidly to encompass the inherent quadratic time complexity of the Jarvis−Patrick and Ward methods.

A similarity search finds database structures that are similar to the target structure; a docking search finds database structures that are *complementary* to the binding site of a 3D protein structure and that might thus be putative ligands for it.[75,76] The first program for this purpose, called DOCK,[77] described the geometries of ligands and binding sites by sets of spheres, and the shape similarity of the ligand to the site was then estimated by the extent to which the corresponding sets of spheres could be overlapped by means of an approximate clique-detection procedure. More recent versions of DOCK augment these steric matching-scores with electrostatic and molecular-mechanics interaction energies for the ligand−receptor complex and consider the use of atomic hydrophobicity descriptors in scoring docked orientations. Database searching is effected by calculating the degree of fit for each database structure in turn and then ranking the molecules in decreasing order of the calculated scores. There have been many reports in the literature that describe this use of DOCK to support the design of novel inhibitors,[78] and its success has spurred the development of many other docking programs, such as CLIX[79] and FLOG.[80] The extension of docking programs so that they can handle ligand flexibility (as well as, ideally, protein flexibility) is one of the two main problems facing workers on docking[81−83] (as well as 3D similarity searching, as discussed in section 6 below). The other problem is the identification of scoring functions that can be calculated sufficiently rapidly to permit database searching but that are, at the same time, sufficiently accurate to provide a reliable basis for prioritizing the database structures for biological testing.[76,84]

Database systems for the handling of chemical reactions did not become widely available until the mid-1980s,[85] almost two decades after the implementation of the first in-house database systems for chemical molecules. Early systems permitted structure and substructure searches to be carried out on the reacting molecules and/or the reaction centers, i.e., those parts of the reacting molecules where the substructural transformation had taken place, but it was not long before facilities were introduced for reaction similarity searching,[86] in which similarities are calculated between pairs of reactions rather than between pairs of molecules as in a conventional similarity searching system. An early example of a reaction similarity measure was included in the REACCS software package.[21,87] This used the atom pair fragments originally developed at Lederle Laboratories[8] to characterize both reacting molecules and reaction sites and allowed not only conventional similarity searching but also subsimilarity searching and supersimilarity searching (as discussed in section 2), which provide "fuzzy" versions of substructure and superstructure searching, respectively. Grethe and Hounshell provide several examples of the use of the various types of search options and emphasize the complementary natures of the resulting outputs.[21] A related task is that of similarity searching to support computer-aided synthesis design programs.[88] Here, the similarity of a molecule (either the intended final product of the synthesis or an intermediate in the synthetic tree) to the molecules in a database of readily available starting materials can help to identify low-cost, synthetically feasible reaction pathways, with the similarity measure typically being based on an MCS definition of some sort.[89−91]

6. CONCLUSIONS

The previous sections of this paper have reviewed the origins and current status of similarity searching in databases of 2D and 3D chemical structures. Although we have discussed a large number of ways in which the similarity between a pair of molecules can be quantified, it must be emphasized that we have restricted our attention to those that can be computed sufficiently rapidly to enable them to be used for searching databases of nontrivial size. Many other types of similarity measure have been reported in the literature but are only applicable, given current hardware and software technology, when small numbers of similarities need to be calculated;[23,24] the development of fast implementations for such measures is one of the main challenges facing workers in the field.[6] The challenge is particularly pressing in the context of developing similarity measures that can be used to search databases of flexible 3D structures. The past few years have seen much work on the implementation of 3D substructure searching,[2,5] and techniques that have been developed there are now starting to find application in the similarity domain. Specifically, attempts are being made to encompass conformational flexibility by the use of multiple low-energy conformations to represent each of the flexible molecules that are to be compared or by the exploration of their conformational spaces during the comparison operations. Examples of these two approaches are exemplified by Perkins et al.[92] and by Thorner et al.,[93] who use simulated annealing and genetic algorithm approaches, respectively, to calculate measures of steric and electrostatic similarity.

The need for improved algorithmic techniques is also apparent, even if attention is restricted to 2D structural representations. This is because of the vastly increased file sizes that need to be processed as a result of the use that is being made of *virtual databases*, i.e., notional sets of molecules such as those represented by a combinatorial library specification. Two radically different approaches suggest themselves. The simpler approach is to enhance the efficiency of current nearest neighbor algorithms so that they permit rapid searching of even the largest libraries when they are fully enumerated: a useful review of such algorithms is provided by Murtagh.[74,94] Alternatively, it may be possible to develop representational methods that can describe all of the molecules in a virtual database in such a manner that they can be searched en masse, without the need for explicit enumeration: some initial work in this area has been reported by Barnard et al.[95]

A further challenge is to develop robust methodologies for the quantitative evaluation and comparison of the effectiveness of different similarity measures (as against their efficiency, which can be determined both by theoretical algorithmic analysis and by timed implementations). This is normally done by means of simulated property prediction experiments based on the similar property principle of Johnson and Maggiora,[21] as exemplified by the extended investigations of different representations reported by Brown and Martin.[51,65] The growing number of similarity measures available means that such comparative investigations will become increasingly important, and their execution will be facilitated by the development of standard datasets containing structural and biological data for nontrivial numbers of compounds. The steroids dataset of Cramer et al.[96] has become a de facto standard for validating new QSAR methods, and recent papers in the literature suggest that the *World Drug Index*[97] and NCI AIDS[98] databases are starting to play a similar role for validating new similarity and diversity methods; this trend is to be applauded since it serves to facilitate the comparison of the results obtained by different research groups and, consequently, to progress the entire field.

The aim of a comparative study is generally to identify the most effective method(s) from among those that are being compared. An alternative approach recognizes that different similarity measures reflect different types of molecular characteristic, and the multifaceted nature of biological activity would thus suggest that no single measure will be optimal for all sorts of similarity search that one might wish to carry out. Instead, one can use several different similarity measures for searching, as advocated by Fisanick et al. in their work with different subsets of the CAS Online screen dictionary,[13] or even combine them into a new integrated measure. Examples of this latter approach are discussed by Sheridan et al.,[56] by Kearsley et al.,[58] and by Ginn et al.,[99] who have all combined the rankings produced by different measures to give a single resultant ranking, an example of the more general procedure known as *data fusion* [100]. We expect that such approaches will become increasingly attractive as further new similarity measures continue to be developed that are sufficiently rapid for use in a database searching context. For example, Thorner et al.[93] have recently described a system for field-based similarity searching that enables the identification of database molecules with electrostatic fields similar to those of the target structure; drawing on previous work by Moreau and Broto[101] on the use of autocorrelation vectors for similarity calculations, Bauknecht et al.[102] discuss a similarity searching system in which the vectors encode various electronic characteristics of the atoms in a molecule; and Robinson et al.[103] describe a representation of the 3D structure of a molecule that is derived from its 2D structure and that is processed extremely rapidly using techniques from digital image processing.

In conclusion, similarity searching has rapidly grown in importance since its introduction just over a decade ago and now plays an important role in lead-discovery programs in the pharmaceutical and agrochemical industries. Its importance can only increase further as 3D-based similarity measures become established, complementing the 2D similarity measures in current chemical information systems, and as new applications are investigated, e.g., the use of similarity measures in diversity analysis, data mining, and pattern recognition applications. It will be interesting to review the field after a further decade of development.

ACKNOWLEDGMENT

We thank the reviewers for their comments on an earlier version of this paper. The Krebs Institute for Biomolecular Research is a designated Biomolecular Sciences Centre of the Biotechnology and Biological Sciences Research Council.

REFERENCES AND NOTES

(1) Ash, J. E., Warr, W. A., Willett, P., Eds. *Chemical Structure Systems*; Ellis Horwood: Chichester, U.K., 1991.

(2) Good, A. C.; Mason, J. S. Three-Dimensional Structure Database Searches. *Rev. Comput. Chem.* **1995**, *7*, 67−117.

(3) Martin, Y. C., Willett, P., Eds. *Designing Bioactive Molecules: Three-Dimensional Techniques and Applications*; American Chemical Society: Washington, D.C., 1998.

(4) Barnard, J. M. Substructure Searching Methods: Old and New. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 532−538.

(5) Willett, P. Searching for Pharmacophoric Patterns in Databases of Three-Dimensional Chemical Structures. *J. Mol. Recognit.* **1995**, *8*, 290−303.

(6) Downs, G. M.; Willett, P. Similarity Searching in Databases of Chemical Structures. *Rev. Comput. Chem.* **1995**, *7*, 1−66.

(7) Martin, Y. C. Pharmacophore Mapping. In *Designing Bioactive Molecules: Three-Dimensional Techniques and Applications*; Martin, Y. C., Willett, P., Eds.; American Chemical Society: Washington, D.C., 1998; pp 121−148.

(8) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure−Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.

(9) Willett, P.; Winterman, V.; Bawden, D. Implementation of Nearest Neighbor Searching in an Online Chemical Structure Search System. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 36−41.

(10) Topliss, J. G.; Edwards, R. P. Chance Factors in QSAR Studies. *ACS Symp. Ser.* **1979**, *112*, 131−145.

(11) Cramer, R. D.; Redl, G.; Berkoff, C. E. Substructural Analysis. A Novel Approach to the Problem of Drug Design. *J. Med. Chem.* **1973**, *17*, 533−535.

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

(13) Fisanick, W.; Cross, K. P.; Rusinko, A. Similarity Searching on CAS Registry Substances. 1. Global Molecular Property and Generic Atom Triangle Geometric Searching. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 664−674.

(14) Fisanick, W.; Cross, K. P.; Forman, J. C.; Rusinko, A. An Experimental System for Similarity and 3D Substructure Searching of CAS Registry Substances. 1. 3D Substructure Searching. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 548−559.

(15) Fisanick, W.; Lipkus, A. H.; Rusinko, A. Similarity Searching on CAS Registry Substances. 2. 2D Structural Similarity. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 130−140.

(16) Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. The CAS Online Search System. 1. General System Design and Selection, Generation, and Use of Search Screens. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 93−102.

(17) Gillet, V. J.; Downs, G. M.; Ling, A.; Lynch, M. F.; Venkataram, P.; Wood, J. V.; Dethlefsen, W. Computer Storage and Retrieval of Generic Chemical Structures in Patents. Part 8. Reduced Chemical Graphs and Their Applications in Generic Chemical Structure Retrieval. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 126−137.

(18) Takahashi, Y.; Sukekawa, M.; Sasaki, S. Automatic Identification of Molecular Similarity Using Reduced-Graph Representation of Chemical Structure. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 639−644.

(19) Fisanick, W. The Chemical Abstracts Service Generic Chemical (Markush) Structure Storage and Retrieval Capability. 1. Basic Concepts. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 145−154.

(20) Willett, P. An Algorithm for Chemical Superstructure Searching. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 114−116.

(21) Grethe, G.; Hounshell, W. D. Similarity Searching in the Development of New Bioactive Compounds: An Application. In *Chemical Structures 2*; Warr, W. A., Ed.; Springer-Verlag: Berlin, 1993; pp 399−407.

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

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

(24) Dean, P. M., Ed. *Molecular Similarity in Drug Design*; Chapman and Hall: Glasgow, 1994.

(25) Bath, P. A.; Morris, C. A.; Willett, P. Effect of Standardisation on Fragment-Based Measures of Structural Similarity. *J. Chemomet.* **1993**, *7*, 543−550.

(26) Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. W. Similarity Searching in Files of Three-Dimensional Chemical Structures: Evaluation of Similarity Coefficients and Standardisation Methods for Field-Based Similarity Searching. *SAR QSAR Environ. Res.* **1995**, *3*, 101−130.

(27) Hubálek, Z. Coefficients of Association and Similarity, Based on Binary (Presence-Absence) Data: an Evaluation. *Biol. Rev. Cambridge Philos. Soc*. **1982**, *57*, 669−689.

(28) Gower, J. C. Measures of Similarity, Dissimilarity and Distance. In *Encyclopaedia of Statistical Sciences*; Kotz, S., Johnson, N. L., Read, C. B., Eds.; Wiley: Chichester, U.K., 1982; pp 397−405.

(29) Ellis, D.; Furner-Hines, J.; Willett, P. Measuring the Degree of Similarity between Objects in Text-Retrieval Systems. *Perspect. Inf. Manage.* **1994**, *3*, 128−149.

(30) 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.

(31) Sokal, R. R.; Sneath, P. H. *Principles of Numerical Taxonomy*; Freeman: San Francisco, 1963.

(32) James, C. A.; Weininger, D.; Delaney, J. Fingerprints−Screening and Similarity. *Daylight Theory Manual*; Daylight Chemical Information Systems Inc.: 1997; URL http://www.daylight.com/dayhtml/doc/theory/theory.toc.html.

(33) Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc*. **1963**, *58*, 236−244.

(34) Lajiness, M. S. Dissimilarity-Based Compound Selection Techniques. *Perspect Drug Discovery Des.* **1997**, *7/8*, 65−84.

(35) Lajiness, M. S. Personal communication, February 1998.

(36) Adamson, G. W.; Bush, J. A. A Comparison of the Performance of Some Similarity and Dissimilarity Measures in the Automatic Classification of Chemical Structures. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 55−58.

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

(38) Cheng, C.; Maggiora, G.; Lajiness, M.; Johnson, M. A. Four Association Coefficients for Relating Molecular Similarity Measures. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 909−915.

(39) Carbo, R.; Arnau, M.; Leyda, L. How Similar Is a Molecule to Another? An Electron Density Measure of Similarity Between Two Molecular Structures. *Int. J. Quantum Chem.* **1980**, *17*, 1185−1189.

(40) Reynolds, C. A.; Burt, C.; Richards, W. G. A Linear Molecular Similarity Index. *Quant. Struct.-Activ. Relat*. **1992**, *11*, 34−35.

(41) Good, A. C. The Calculation of Molecular Similarity: Alternative Formulas, Data Manipulation and Graphical Display. *J. Mol. Graph*ics **1992**, *10*, 144−151.

(42) Petke, J. D. Cumulative and Discrete Similarity Analysis of Electrostatic Potentials and Fields. *J. Comput. Chem*. **1993**, *14*, 928−933.

(43) Bradshaw, J. Introduction to the Tversky Similarity Measure. Presented at Daylight MUG Meeting, Laguna Beach, CA, February 1997. URL http://www.daylight.com/meetings/mug97/agenda97/Bradshaw/MUG97/tv_tversky.html.

(44) Tversky, A. Features of Similarity. *Psychol. Rev.* **1977**, *84*, 327−352.

(45) Maggiora, G. M.; Mestres, J.; Hagadone, T. R.; Lajiness, M. S. Asymmetric Similarity and Molecular Diversity. Presented at the 213th National Meeting of the American Chemical Society, San Francisco, CA, April 13−17, 1997.

(46) Jones, P. E.; Curtice, R. M. A Framework for Comparing Document Term Association Measures. *Am. Doc*. **1967**, *18*, 153−161.

(47) Carbo, R.; Calabuig, B. Quantum Similarity Measures, Molecular Cloud Description and Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 600−606.

(48) Brown, R. D. Descriptors for Diversity Analysis. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 31−49.

(49) Adamson, G. W.; Cowell, J.; Lynch, M. F.; McLure, A. H. W.; Town, W. G.; Yapp, A. M. Strategic Considerations in the Design of a Screening System for Substructure Searches of Chemical Structure Files. *J. Chem. Doc.* **1973**, *13*, 153−157.

(50) Nilakantan, R.; Bauman, N.; Dixon. J. S.; Venkataraghavan, R. Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82−85.

(51) 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.

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

(53) Bath, P. A.; Poirrette, A. R.; Willett, P.; Allen, F. H. Similarity Searching in Files of Three-Dimensional Chemical Structures: Comparison of Fragment-Based Measures of Shape Similarity. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 141−147.

(54) Bemis, G. W.; Kuntz, I. D. A Fast and Efficient Method for 2D and 3D Molecular Shape Description. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 607−628.

(55) Nilakantan, R.; Bauman, N.; Venkataraghavan, R. A New Method for Rapid Characterization of Molecular Shape: Applications in Drug Design. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 79−85.

(56) Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical Similarity Using Geometric Pair Descriptors. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 128−136.

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

(58) 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.

(59) Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R. Determining structural similarity of chemicals using graph-theoretic indices. *Discrete Appl. Math.* **1988**, *19*, 17−44.

(60) 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. Comput. Sci.* **1997**, *37*, 559−614.

(61) Perry, N. C.; van Geerestein, V. J. Database Searching on the Basis of Three-Dimensional Molecular Similarity Using the SPERM Program. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 607−616.

(62) Dean, P. M.; Callow, P.; Chau, P. L. Molecular Recognition: Blind Searching for Regions of Strong Structural Match on the Surfaces of Two Dissimilar Molecules. *J. Mol. Graphics* **1988**, *6*, 28−34.

(63) Dean, P. M.; Perkins, T. D. J. Calculation of Three-Dimensional Similarity. In *Designing Bioactive Molecules: Three-Dimensional Techniques and Applications*; Martin, Y. C., Willett, P., Eds.; American Chemical Society: Washington, D.C., 1998; pp 199−218.

(64) Pepperrell, C. A.; Taylor, R.; Willett, P. Implementation and Use of an Atom-Mapping Procedure for Similarity Searching in Databases of 3-D Chemical Structures. *Tetrahedron Comput. Methodol.* **1990**, *3*, 575−593.

(65) 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.

(66) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch M. F. Computer Storage and Retrieval of Generic Chemical Structures. 10. Assignment and Logical Bubble-Up of Ring Screens for Structurally Explicit Generics. *J. Chem. Inf. Comput. Sci*. **1989**, *29*, 215−224.

(67) Cramer, R. D.; Patterson, D. E.; Clark, R. D.; Soltanshahi, F.; Lawlett, M. S. Virtual Compound Libraries: A New Approach to Decision Making in Molecular Discovery Research. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, xxx−xxx.

(68) Everitt, B. S. *Cluster Analysis*; 3rd ed.; Arnold: London, 1993.

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

(70) Downs, G. M.; Willett, P. Clustering of Chemical-Structure Databases for Compound Selection. In *Advanced Computer-Assisted Techniques in Drug Discovery*; van de Waterbeemd, H., Ed.; VCH: New York, 1994; pp 111−130.

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

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

(73) Doman, T. N.; Cibulskis, J. M.; Cibulskis, M. J.; McCray, P. D.; Spangler, D. P. Algorithm. 5. A Technique for Fuzzy Similarity Clustering of Chemical Inventories. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 1195−1204.

(74) Murtagh, F. *Multidimensional Clustering Algorithms*; Physica Verlag: Vienna, 1985.

(75) Blaney, J. M.; Dixon, J. S. A Good Ligand is Hard to Find: Automated Docking Methods. *Perspect. Drug Discovery Des.* **1993**, *1*, 301−319.

(76) Dixon, J. S.; Blaney, J. M. Docking: Predicting the Structure and Binding Affinity of Ligand−Receptor Complexes. In *Designing Bioactive Molecules: Three-Dimensional Techniques and Applications*; Martin, Y. C., Willett, P., Eds.; American Chemical Society: Washington, D.C., 1998; pp 175−197.

(77) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecule-Ligand Interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.

(78) Kuntz, I. D. Structure-Based Strategies for Drug Design and Discovery. *Science* **1992**, *257*, 1078−1082.

(79) Lawrence, M. C.; Davis, P. C. CLIX−A Search Algorithm for Finding Novel Ligands Capable of Binding Proteins of Known 3-Dimensional Structure. *Proteins: Struct., Funct., Genet.* **1992**, *12*, 31−41.

(80) Miller, M. D.; Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P. FLOG: A System to Select “Quasi-Flexible” Ligands Complementary to a Receptor of Known Three-Dimensional Structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 153−174.

(81) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* **1997**, *267*, 727−748.

(82) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. *J. Mol. Biol*. **1996**, *261*, 470−489.

(83) Westhead, D. R.; Clark, D. E.; Murray, C. W. A Comparison of Heuristic Search Algorithms for Molecular Docking. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 209−228.

(84) Ajay, Murcko, M. A. Computational Methods To Predict Binding Free Energy in Ligand−Receptor Complexes. *J. Med. Chem*. **1995**, *38*, 4953−4967.

(85) Willett, P., Ed. *Modern Approaches to Chemical Reaction Searching*; Aldershot: Gower, U.K., 1986.

(86) Barth, A. Status and Future Development of Reaction Databases and Online Retrieval Systems. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 384−393.

(87) Grethe, G.; Moock, T. E. Similarity Searching in REACCS. A New Tool for the Synthetic Chemist. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 511−520.

(88) Ihlenfeldt, W. D.; Gasteiger, J. Computer-Assisted Planning of Organic Syntheses: The Second Generation of Programs. *Angew. Chem., Int. Ed. Engl.* **1995**, *34*, 2613−2633.

(89) Wipke, W. T.; Rogers, D. Artificial Intelligence in Organic Synthesis. SST: Starting Material Selection Strategies. An Application of Superstructure Search. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 71−81.

(90) Gasteiger, J.; Ihlenfeldt, W. D.; Rose, P. A Collection of Computer Methods for Synthesis Design and Reaction Prediction. *Recl. Trav. Chim. Pays-Bas* **1992**, *111*, 270−290.

(91) Johnson, A. P.; Marshall, C. Starting Material Oriented Retrosynthetic Analysis in the LHASA Program. 2. Mapping the SM and Target Structures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 418−425.

(92) Perkins, T. D. J.; Mills, J. E. J.; Dean, P. M. Molecular Surface-Volume and Property Matching to Superpose Flexible Dissimilar Molecules. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 479−490.

(93) Thorner, D. A.; Wild, D. J.; Willett, P.; Wright, P. M. Similarity Searching in Files of Three-Dimensional Chemical Structures: Flexible Field-Based Searching of Molecular Electrostatic Potentials. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 900−908.

(94) Murtagh, F. Search Algorithms for Numeric and Quantitative Data. In *Intelligent Information Retrieval: The Case for Astronomy and Related Space Sciences*; Heck, A., Murtagh, F., Eds.; Kluwer: Dordrecht, The Netherlands, 1993; pp 29−48.

(95) Barnard, J. M.; Downs, G. M.; Turner, D. B.; Tyrrell, S. M.; Willett, P. Rapid Diversity Analysis in Combinatorial Libraries Using Markush Structure Techniques. Paper presented at the American Chemical Society National Meeting, San Francisco, April 1997.

(96) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc*. **1988**, *110*, 5959−5967.

(97) The *World Drug Index* database is available from Derwent Information, URL http://www.derwent.co.uk/.

(98) The AIDS database is available from the Developmental Therapeutics Program, National Cancer Institute, URL http://epnws1.ncifcrf.gov:2345/dis3d/aids_screen/aidstitle.html.

(99) Ginn, C. M. R.; Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. W. Similarity Searching in Files of Three-Dimensional Chemical Structures: Evaluation of the EVA Descriptor and Combination of Rankings Using Data Fusion. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 23−37.

(100) Hall, D. L. *Mathematical Techniques in Multisensor Data Fusion*; Artech House: Boston, 1992.

(101) Moreau, G.; Broto, P. Autocorrelation of Molecular Structures: Application to SAR Studies. *Nouv. J. Chim*. **1980**, *4*, 757−764.

(102) Bauknecht, H.; Zell, A.; Bayer, H.; Levi, P.; Wagener, M.; Sadowski, J.; Gasteiger, J. Locating Biologically Active Compounds in Medium-Sized Heterogeneous Datasets by Topological Autocorrelation Vectors: Dopamine and Benzodiazepine Agonists. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 1205−1231.

(103) Robinson, D. D.; Barlow, T. W.; Richards, W. G. The Utilization of Reduced Dimensional Representations of Molecular Structures for Rapid Molecular Similarity Calculations. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 943−950.

CI9800211