REVIEW

A REVIEW AND EXAMINATION OF THE MATHEMATICAL SPACES UNDERLYING MOLECULAR SIMILARITY ANALYSIS

M.A. JOHNSON

Computational Chemistry, The Upjohn Company, Kalamazoo, Michigan 49001, USA

Received 3 December 1988
(in final form 10 February 1989)

Abstract

As an intuitive concept, molecular similarity has played a fundamental role in chemistry. It is implicit in Hammond's postulate, in the principle of minimum structure change, and in the assumption that similar structures tend to have similar properties. With the advent of large computers, computable definitions of similarity are being used in the pharmaceutical industry for similarity searching, dissimilarity selection, molecular superpositioning, structure generation, and quantitative structure-activity analysis. The diversity of applications of computable definitions of molecular similarity has often obscured important mathematical commonalities underlying these definitions. The broadest commonalities are relationships based on equivalence, matching, partial ordering, and proximity. A mathematical space suitable for molecular similarity analysis consists of a set of mathematical structures and one or more of these similarity relationships defined on that set. This report surveys the mathematical spaces used in molecular similarity analysis. The survey covers the types of chemical information, similarity relationships, and applications associated with the use of each mathematical space in a molecular similarity context.

1. Introduction

As an intuitive concept, molecular similarity has played a fundamental role in chemistry. The concept occurs in the widely recognized statement that similar structures have similar properties [1]. Quantitative applications of this statement in the pharmaceutical industry include similarity searching, dissimilarity selection, molecular superpositioning, and quantitative structure-activity analysis. The intuitive concept occurs in the principle of minimum structure change [2] which states that chemical reactions proceed so as to minimize the redistribution of valence electrons.

Quantitative applications of this principle in organic chemistry include the classification of reactions and the generation of reaction intermediates and reaction pathways. Finally, the intuitive concept occurs in Hammond's postulate [3] that if a reaction is exothermic, the transition state resembles the reactants more closely than the products. Quantitative investigations of this postulate are beginning to appear in the physical-organic chemistry literature.

The advent of high-speed computers has opened up many possible applications of molecular similarity analysis (MSA) in computer-assisted chemistry. The diversity of the applications has resulted in the development of a wide variety of similarity measures. A need and opportunity have arisen for constructing a formalism for molecular similarity analysis. Such a formalism is necessarily based on features that characterize and distinguish molecular similarity concepts. However, research on computable concepts of similarity is widely dispersed in many unrelated journals and symposia proceedings. Not surprisingly, the underlying concept of molecular similarity goes unstated in the literature, if not unrecognized.

This study focuses on four distinct mathematical relationships, each of which constitutes a possible definition of similarity relevant to chemistry: equivalence, matching, partial ordering, and proximity. Most approaches to MSA share one or more of these relationships. These relationships arise in the context of a mathematical space. A mathematical space suitable for MSA will be defined to consist of a set X of mathematical representations of molecules and one or more similarity relationships defined on X. This study surveys the mathematical spaces that have been used in MSA and classifies molecular similarity measures according to their underlying mathematical spaces. In attempting this classification, most of the mathematical definitions of similarity that have been used in MSA are discussed.

Most concepts of molecular similarity have been proposed in applied contexts. Applications of similarity concepts in chemistry are so numerous that an exhaustive coverage would detract from our focus on the underlying mathematical spaces. The applications cited here are chosen to illustrate the diversity of the contexts for which *explicit* similarity concepts have been proposed. Although not specifically addressed, methods of computing the similarity between molecules are covered and often emphasized in the cited studies. Thus, this study can also be viewed as a fairly broad, but certainly not exhaustive, introduction to and review of the literature on mathematically explicit approaches to MSA.

Only approaches to similarity that involve reflexive relationships will be covered in this study. A similarity relationship will be called reflexive if an entity is at least as similar to itself in the relationship as it is to anything else. The similarity measures arising in the distance geometry approach of Crippen [4], the minimum topological difference method of Simon et al. [5], and the geometric approaches of Kuntz et al. [6] and DesJarlais [7] are not reflexive. Rather, these measures are based on the complementary relationship between a ligand and either an observed or hypothesized receptor site of an otherwise noncomparable macromolecule. We shall also exclude from this study those mathematical definitions of similarity that are not broadly transportable to other sciences. Later we will distinguigh between the chemical description and the mathematical representation of a molecule in MSA. The mathematical representations and associated similarity relationships described here are broadly relevant to a wide variety of nonchemical sciences. This is not the case for the similarity concepts occurring in DuChamp [8] and Hopfinger [9]. These approaches use the concept of a potential function whose form and parameters are, for the most part, unique to the chemical sciences. Finally, approaches defined only for special subclasses of compounds such as, for example, amino acids [10] will also be excluded due to space considerations. In doing so, we will ignore another broad area of research involving the comparison of protein and DNA sequences [11]. Although a significant amount of research involving concepts of molecular similarity is excluded from this study by the preceding restrictions, a broad, and we hope to show, integrated body of research remains. The next two sections define mathematical relationships that will be used to relate the various concepts of molecular similarity proposed in that research.

2. Basic mathematical concepts of similarity

Let C and C' denote two molecules. Chemically, we can envision many ways in which C and C' might be similar. For example, they may contain common functional groups, exhibit similar physicochemical properties, and/or have similar shapes. Different definitions of molecular similarity can have important mathematical commonalities. For example, one investigator may assess similarity between two van der Waals surfaces, while another investigator may assess the similarity between the contour surfaces of two molecular orbitals. Both of these surfaces are examples of what we shall term the *chemical descriptions* of their investigations. Although the two assessments of similarity differ markedly in their chemical descriptions, both descriptions represent examples of mathematical surfaces defined on a three-dimensional euclidean space. Consequently, the same mathematical formulas can be used to compute similarity once the chemical descriptions have been mapped to the appropriate mathematical surfaces. The surfaces defined on R^3 are examples of *mathematical structures* that can be associated with molecules. Other examples of mathematical structures are numbers, vectors, functions, scalar fields, graphs, and groups. The two methods of assessing similarity use the same type of mathematical structures to represent the information in their respective chemical descriptions. This is significant because the computation of similarity is always defined in terms of the mathematical representation of the chemical description.

One may ask what types of mathematical concepts of similarity have been used in MSA. We shall distinguish four types, which will be illutrated with labeled graphs as the mathematical structures. Although we shall use labeled graphs to illustrate these concepts of similarity, the concepts are associated with a variety of other mathematical structures being employed in MSA.

2.1. EQUIVALENCE

A labeled graph $L = (V, E, l)$ consists of a set V of vertices, a set E of edges, and a labeling function l mapping each vertex v to a vertex label $l(v)$ and each edge e to an edge label $l(e)$. The labeled graph L will be called a chemical graph of compound C if the labeled vertices denote the nonhydrogen atoms of C, and the labeled edges denote the types of bonds connecting two nonhydrogen atoms. The chemical graph of C is often represented as a connection table when stored or manipulated in a computer.

If A and B denote two molecules, we shall, for the moment, write $A \simeq B$ to mean A and B have identical chemical graphs. For example, D-glucose and L-glucose are stereoisomers and, consequently, have identical chemical graphs. Thus, we would write D-glucose $\simeq$ L-glucose. To write $A \simeq B$ is to indicate that A and B are similar in a fundamental way. Clearly, $A \simeq B$ implies A and B agree in all properties determined solely by the chemical graph of a molecule.

The relation $\simeq$ is an example of an equivalence relation. Equivalence relations in mathematics are defined by three properties: reflexivity, symmetry, and transitivity [12]. A relation is reflexive if $A \simeq A$ for all molecules A. It is symmetric if $A \simeq B$ implies $B \simeq A$. It is transitive if $A \simeq B$, and $B \simeq C$ implies $A \simeq C$. Let $[A]$ denote the family of those molecules equivalent to A with respect to $\simeq$. If B denotes some other molecule, it can be proven mathematically that either $[A]$ and $[B]$ denote the same set of molecules or the two sets have no members in common. The set $[A]$ is called an equivalence class.

Equivalence relations abound in chemistry. To illustrate, redefine $A \simeq B$ to mean A and B have identical chemical descriptions of type d. Here, a type of chemical description can range from a melting point to a set of chemical properties, from a chemical formula to a structural formula, from an infrared carbonyl stretching frequency to an infrared spectrum, and from a van der Waals surface to a molecular orbital contour surface. Regardless of the type of chemical description, $\simeq$ is an equivalence relation.

The types and uses of equivalence relations in chemistry are too extensive and too varied to be adequately covered within the scope of this study. Instead, attention will be focused on contexts involving an equivalence relation augmented by one or more additional similarity relationships based on matching, partial order, or proximity. Euclidean spaces involve all of these relationships plus the important operations of addition and multiplication. The elements of an n-dimensional euclidean space are vectors. However, the elements of many of the mathematical spaces of MSA are not vectors, but rather sequences, graphs, distance matrices, volumes, and so forth. We will exclude from this study the wide variety of work premised on euclidean spaces that does not readily carry over to these other interesting mathematical spaces. In particular, important work on predicting chemical properties using quantum-mechanical models and linear regression models will be excluded. These two approaches to predicting chemical properties are distinguished from similarity-based approaches in Johnson et al. [13].

2.2. MATCHING

Let us return to our original definition that $A \simeq B$ implies A and B have identical chemical graphs. Let G_a denote the chemical graph of A. One might ask what G_a and G_b have in common. A matching is a correspondence of features of G_a with features of G_b. Because G_a constitutes a specification of $[A]$, a matching can be viewed as a similarity relationship between equivalence classes. The methods of specifying a matching of two mathematical structures usually depends on the form of those structures. A method of expressing a matching of two chemical graphs will be illustrated.

Consider the chemical graphs L and L' depicted in fig. 1. A matching of L and L' is a one-to-one correspondence between some vertices of L and some of L'. In fig. 1, matched vertices are subscripted with identical indices and unmatched vertices are not subscripted. If vertex v of L is matched with vertex v' of L', then $l(v)$ must equal $l(v')$, i.e. v and v' must agree in their vertex type. If, in addition, vertex u of L is matched with vertex u' of L' and if uv is an edge of L, then $u'v'$ must be an edge of L' of the same edge type. In the same manner, if $u'v'$ is an edge of L', then uv must be an edge of L of the same edge type.

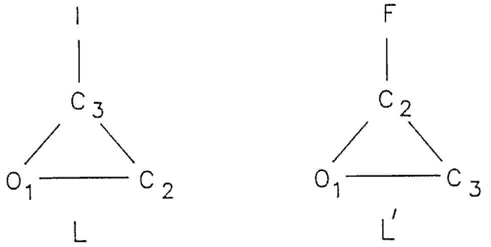

Fig. 1. A matching of chemical graphs L and L'.

A matching resembles an analogy. In an analogy, features of one entity are paired with features of another entity. Which features are paired is often a subjective choice. Similarly, there are many possible matchings between the features of a mathematical representation of one molecule and of a mathematical representation of another molecule. In fig. 1, other matchings are obtained by permuting indices 2 and 3 in L or by deleting index 2 in both L and L'. The suitability of a particular matching depends on the context of the problem. In the next section, we will discuss some of the methods by which a particular matching is selected.

2.3. PARTIAL ORDERING

Let A and B denote two labeled graphs where some of the vertices of each graph are subscripted so as to define a matching of A and B. If all of the vertices of A are subscripted, A is called a subgraph of B. For example, by deleting the vertex labeled I in chemical graph L of fig. 1, we see that the resulting graph G (representing ethylene oxide) is a subgraph of L'. For the moment, we shall write $A \leq B$ to mean A is a subgraph of B.

The relation $\leq$ is called the subgraph partial order. A relation $\leq$ is called a partial order if it is reflexive, antisymmetric, and transitive. The reflexive and transitive properties of a relation were defined earlier. A relation is antisymmetric if $A \leq B$ and $B \leq A$ implies A and B are identical. When the labeled graphs refer to chemical graphs, the subgraph partial order will also be called the substructure partial order.

A partial order $\leq$ on a finite set X can be represented by a graph as follows. Assume u and w are members of X, and that $u \leq w$. If $u \leq v \leq w$ implies either $v = u$ or $v = w$, i.e. there are no members of X between u and w, then w is said to cover u. Define the graph $G = (X, E)$ such that uw is an edge of G if and only if u covers w or w covers u. Call G the graph of the partial order $(X, \leq)$ and denote G by $G(X, \leq)$. Figure 2 depicts the graph of the substructure partial ordering of six alkanes.

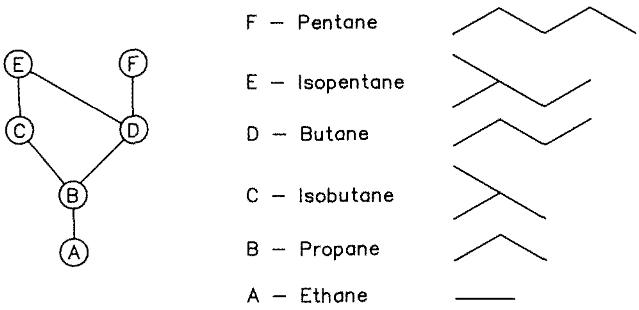

Fig. 2. The graph of the substructure partial ordering of six alkanes.

Several concepts of similarity relevant to chemistry are based on the concept of a partial order. For example, when we speak of a ketone, we refer to a set of molecules having the chemical graph C=O as a substructure. Let G denote the ketone structure C=O. The set of ketones is defined mathematically as those chemical graphs G' for which $G \leq G'$. This illustrates a membership relationship based on a partial order. A betweenness relationship based on the graph of a partial order can be illustrated using fig. 2. In this figure, isopentane (E) is between isobutane (C) and pentane (F) because isopentane lies on a shortest path connecting isobutane and pentane. Finally, a distance relationship between the vertices of the graph of a partial order is given by the length of the shortest path connecting two vertices. For example, in fig. 2 the distance between isobutane and pentane is 3.

Although the concept of matching arises quite naturally in most developments of similarity based on partial orderings, a partial ordering of chemical descriptions which is developed without the use of the concept of matching is described in section 13 on groups.

2.4. PROXIMITY

The preceding distance relationship is an example of using a number to express the similarity between two molecules. Although terminology and notation varies [14,15], we shall denote a measure of similarity between mathematical structures D and D' by $\nu(D, D')$. Following the terminology of Borg and Lingoes [14], we shall call the function ν a proximity measure.

All the proximity measures we shall encounter use nonnegative numbers to express similarity. When large numbers represent similar compounds, the proximity measure will be called a similarity measure or similarity coefficient. Correlation coefficients are similarity measures whose values vary from 0 to 1, with 1 denoting an exact correlation between the features of two chemical descriptions. We will write $s(D, D')$ for $\nu(D, D')$ to emphasize that ν is a similarity measure.

When large numbers represent dissimilar molecules, the proximity measure will be called a dissimilarity measure or distance function. We will write $d(D, D')$ for $\nu(D, D')$ to emphasize that ν is a dissimilarity measure. Most of the dissimilarity measures we discuss are metrics. A dissimilarity measure d is a metric if d is a nonnegative function satisfying (1) $d(x, y) = 0$ if and only if $x = y$, (2) $d(x, y) = d(y, x)$, and (3) $d(x, z) \leqslant d(x, y) + d(y, z)$. Property (3) is called the triangle inequality. We shall write μ for d when we wish to stress that d is a metric.

2.5. SIMILARITY CONCEPTS ON DERIVED REPRESENTATIONS

Frequently, one encounters cases in which one mathematical representation of a molecule is derived from another representation. For example, given a configuration of n atoms in R^3, one can derive an $n \times n$ distance matrix giving the pairwise distances between the atoms.

Derived representations provide a method of transferring a similarity concept defined on one mathematical space to another space. To illustrate the transference of a proximity measure, let U and V denote any two sets of mathematical structures being used to represent molecular information. Let $f(u)$ denote the structure in V derived from the structure u in U. For example, u may be a 3D configuration, and $f(u)$ may be the pairwise interatomic distance matrix of u. Let ω be a proximity measure defined on V. Define the proximity measure ν on U by

$$\nu(u, u') = \omega(f(u), f(u')), \tag{2.1}$$

where u and u' are any two members of U.

The properties of ω that are transferred to ν depend on the nature of f. For example, suppose ω is a metric. It is easy to show that ν is symmetric and satisfies the triangular inequality. Yet ν need not be a metric. To see this, let f be our earlier mapping of steroechemical structures to labeled graphs in which u = D-glucose and u' = L-glucose were mapped to the same chemical graph G. We have

$$\begin{aligned}\nu(u,u') &= \omega(f(u), f(u')) \\ &= \omega(G, G') \\ &= 0\,. \end{aligned} \tag{2.2}$$

Since we know u and u' denote different stereochemical structures, ν cannot be a metric, since property (1) of metrics is not guaranteed. However, it can be shown that ν is a semimetric.

Still, mappings deriving one representation of molecules in terms of another are widely used in MSA. Figure 3 is a "derivational" flowchart in which the arcs indicate derivations used in MSA. We will parenthatically indicate each derivational arc when the associated derivation is covered in the body of the text.

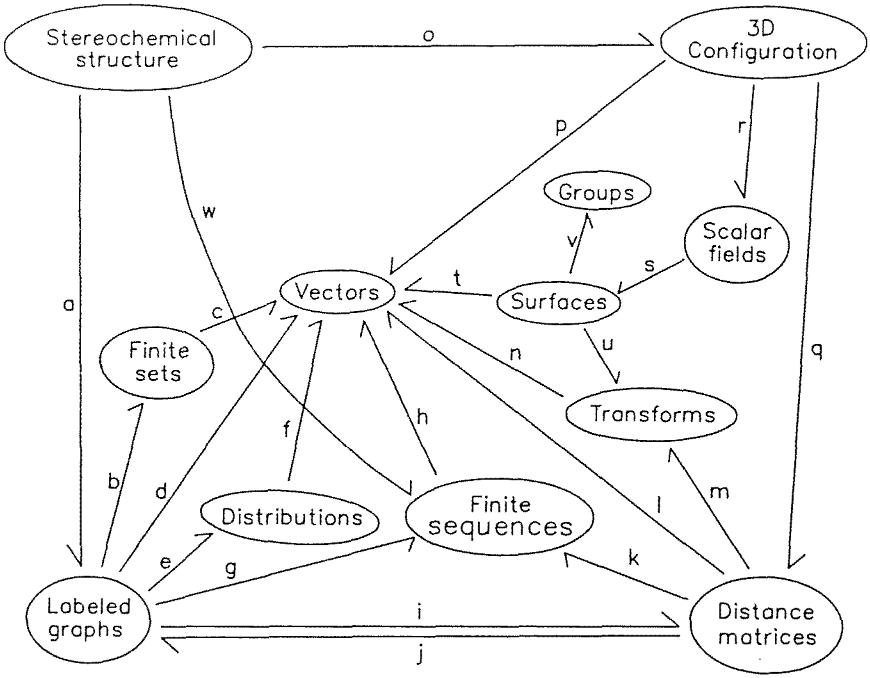

Fig. 3. A flowchart of derivations connecting mathematical representations of molecules.

3. Types of matchings

As we have seen, matching is a basic similarity concept and is frequently used in the development of partial orders and proximity measures. The number of possible matchings tends to grow explosively with the complexity of the mathematical structure. A variety of methods of selecting one matching from amongst the many possibilities have been developed. We group these methods into three types.

3.1. A PRIORI MATCHING

As noted earlier, a matching is a one-to-one correspondence between a subset of the elements of one mathematical structure with a subset of the elements of another mathematical structure. In fig. 1, the matching was indicated by indexing the elements with subscripts.

The elements in fig. 1 could have been indexed in many different ways. As noted earlier, subscripts 2 and 3 of L could be permuted to obtain another matching. This choice might reflect an hypothesis that the carbons attached to the halogens should be matched. The grouping of the letters F and I under the term "halogen" introduces a chemical concept in fig. 1 not explicitly present for either the nonchemist or the computer. The basis of this matching hypothesis, although valid, lies outside our confine of computable similarity concepts. Consequently, we shall call the selected matching an a priori matching.

The rationale of a matching hypothesis is sometimes questioned and sometimes not. One might reasonably question the preceding hypothesis that the carbons attached to the halogens should correspond to each other. On the other hand, consider the matching implicit in the comparison of two vectors $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$ of chemical descriptors [16]. Here, we match u_1 with v_1, u_2 with v_2, etc., without giving this a priori matching a second thought.

3.2. CANONICAL MATCHING

Other approaches to selecting matchings begin by putting each mathematical structure in a standard form. The matching of two structures is then defined in terms of the standard form. For example, in section 9 we show how a chemical graph G of molecule C can be represented as a sequence of letters. It turns out that a number of different sequences can represent G. The first member in the lexicographical ordering of these sequences can be viewed as a standard form of C. Assume atoms u and v of the chemical graph of molecule C are of the same type and u precedes v in the standard form of C. Assume also that u and v are matched with atoms u' and v' of the chemical graph of molecule C'. A canonical matching might require that u' precede v' in the standard form C'.

A canonical matching can also be viewed as the outcome of an hypothesis on what elements of two chemical descriptions should be paired. However, in this case the hypothesis is embedded in a computable algorithm.

3.3. MAXIMAL MATCHING

A priori and canonical matchings are essentially determined before any elements of two mathematical structures are *explicitly* paired. To develop a maximal matching, one first associates a numerical value with each matching. A matching is maximal if it has the smallest/largest associated numerical value.

To illustrate using chemical graphs, let the numerical value associated with a matching be the number of atoms and bonds without counterparts in the other graph. For example, the value for the matching in fig. 1 would be 4 since two atoms (1 fluorine and 1 iodine) have no counterparts in the other graph and neither do their associated bonds. One can show that the numerical value for any other matching of the chemical graphs in fig. 1 is at least 4. Thus, the matching in fig. 1 is a maximal matching with respect to this method of associating a numerical value with a matching.

As noted earlier, the concepts of equivalence, matching, partial ordering and proximity are defined on a variety of other mathematical spaces being employed in MSA. We now consider each of these mathematical spaces indicated in fig. 3, beginning with the real numbers, an important special case of a vector space.

4. Real numbers

Possibly the simplest chemical description D of a molecule C is a single number. Because of its importance, a single number used as a chemical description will be called a chemical *descriptor*. Any scalar chemical property can be viewed as a chemical descriptor. The use of chemical descriptors in regression analysis has been excluded from this study because such analyses fail to extend to the majority of the mathematical spaces under consideration. Under this restriction, the only chemical descriptors used in MSA appear to be topological indices. A topological index is a number calculated from the chemical graph that is independent of the method of representing the graph (see arc d, fig. 3). Balaban et al. [17] review a number of topological invariants or indices. When the number of chemical graphs having the same value for the topological index is small, Randić [18,19] refers to the index as a molecular identification number. Unique molecular identification numbers have been conjectured to exist, but have yet to be found [10,21]. Razinger et al. [20] and Szymanski et al. [21] review progress toward the development of unique molecular identification numbers.

As numbers, the values of a topological index are linearly ordered. A partial order $\leq$ on a set X is a linear order if for every x and y in X, either $x \leq y$ or $y \leq x$. The natural metric on numbers is the absolute value of the difference between two numbers. We say two compounds C and C' with associated numbers r and r' are similar if $|r - r'|$ is small.

Recently, molecular identification numbers have been proposed for the indexing of structures in a chemical database [18,19]. Randić [22] has proposed that the absolute value of the difference between two such numbers be used as a dissimilarity measure for clustering compounds.

5. Product spaces

In the preceding section, the chemical description was based on a single chemical descriptor. Here, the description is based on a vector of chemical descriptors. Vector descriptions proposed in an MSA context include mass spectrum relative intensities [23], counts of various lengths of paths in a chemical graph [24], topological distances between types of atom pairs [25], counts of atom-centered fragments [26], counts of bond-centered fragments [27], the first ten principle components of a selected set of 90 topological indices [28], and counts of atoms and bonds [29] (arc d of fig. 3). Hansch [30] uses substituent constants as descriptors of substituents. Sections 9 and 11 give additional applications of vector descriptions.

Since we always compare the ith component of one vector with the ith component of the other vector, the matching problem does not arise with vectors. If we let $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$, then an important partial ordering on vectors is defined by $\boldsymbol{x} \leq \boldsymbol{y}$ if $x_1 \leqslant y_1, \ldots, x_n \leqslant y_n$. Many proximity measures have been used for comparing vectors [15,31]. Some common ones include the L_1 or city-block metric $\mu_1(D, D')$ given by

$$\mu_1(D, D') = \sum |x_k - y_k|, \tag{5.1}$$

the L_2 metric $\mu_2(D, D')$ or euclidean metric given by

$$\mu_2(D, D') = \sqrt{\sum (x_k - y_k)^2}, \tag{5.2}$$

the correlation coefficient $s_1(D, D')$ given by

$$s_1(D, D') = \frac{\sum (x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum (x_k - \bar{x})^2 \times \sum (y_k - \bar{y})^2}}, \tag{5.3}$$

where $\bar{x}$ denotes the mean of $x_i, \ldots, x_n$, and the Tanimoto coefficient $s_2(D, D')$ is given by

$$s_2(D, D') = \frac{\sum x_k y_k}{\sum x_k^2 + \sum y_k^2 - \sum x_k y_k}. \tag{5.4}$$

When structural fragments are used for descriptors, one has the choice of letting x_i denote the simple occurrence or nonoccurrence of the ith fragment or letting x_i denote the number of times the fragment occurs. These are special cases of the more general problem of assigning weights to the fragments [15,27].

The descriptive flexibility and computational accessibility of vectors has resulted in their wide use as chemical descriptions. The partial ordering of vectors underlies their popular use in the screening step of substructure searches of large chemical databases [32]. The preceding proximity measures are routinely used in similarity analysis involving multidimensional scaling [33], nearest neighbor prediction [34], and cluster analysis [35–37], These measures are being used in querying databases containing spectral information [23,38], molecular structures [24], and reactions [39]. They are being used to select compounds for screening [29] based on testing similar compounds [26] and on testing dissimilar compounds [40], and they are being used to select substituents in lead optimization [30,41,42]. Finally, they have been used to identify active substructures in structure-activity studies [43] and to suggest that members of a series of biologically active compounds are acting by different mechanisms [44].

6. Distributions

A vector $(p_1, \ldots, p_n)$ of nonnegative numbers that sum to one constitutes a parameter vector of a multinomial distribution (arc f in fig. 3). Such a vector can be associated with a molecule in many ways [45]. One used in MSA [46,47] is defined as follows (arc e in fig. 3): Let δ_i denote the number of self-avoiding (no vertex occurs more than once) paths of length i, $i = 1, \ldots, k$, in the chemical graph of the compound [24]. For example, in chemical graph L of fig. 1, $\delta_2 = 10$ since there are 3 paths of length 2 starting from O_1, 3 from C_2, 2 from C_3 and 2 from I_4. Define p_i by

$$p_i = \frac{\delta_i}{\sum_i^k \delta_i} \qquad (6.1)$$

so that the p_i's sum to one.

The fact that the vector $(p_1, \ldots, p_n)$ characterizes a distribution gives rise to special similarity measures based on the concepts of information theory. One such measure that has been used in MSA is the informationally-based measure of Jeffreys [48]. He defines the similarity between two multinomial parameter vectors $P = M(p_1, \ldots, p_k)$ and $Q = M(q_1, \ldots, q_k)$ by

$$d_1(P, Q) = \sum (p_i - q_i) \log_2 (p_i/q_i) . \qquad (6.2)$$

Since log(0) and log(∞) are not defined, we must restict the value of k in (6.1) so that $\delta_k > 0$ for all compounds being compared. Rao [49] presents a wide variety of proximity measures that have been developed for other distributions.

To my knowledge, similarity measures based on distributions have always been used to construct other chemical descriptions. The construction procedure generates descriptions which are finite sequences. Since the construction procedure is defined for any finite set of compounds for which a proximity measure is available, the procedure will be described in section 9 on finite sequences. This construction procedure has been used as an intermediate step in quantitative structure-activity analyses [46,47].

7. Finite sets

A finite set is the natural chemical description when comparing compounds with respect to molecular fragments. The molecular formula is probably the foremost example. For example, the molecular formula for chemical graph L in fig. 1 can be represented by the set $\{O_1, C_1, C_2, I_1\}$ (arc b in fig. 3). Here we require that the subscripts start with 1 and end with the number of occurrences of the associated atom type. This requirement ensures that the intersection of the set representations of the molecular formulas for butane and pentane is four carbons, as we would expect. Section 5 on product spaces reviewed lists of structural fragments that have been used in assessing molecular similarity.

As we have seen, vectors have been used to represent the set of structural fragments occurring in a molecule. Here, the ith value of the vector represents the number of times the ith fragment occurs in the molecule (arc c in fig. 3). This vector representation is possible only if all fragments come from some fixed list of fragments. The list of atom types occurring in molecular formulas would be an example of such a fixed list. Any fixed list must have a finite number of members. Consider the list L of all fragments [50] of chemical graphs. If L were finite, then there would exist a fragment with at least as many atoms as any other fragment. Clearly, this is not the case. Consequently, L cannot be a fixed list. However, the set of fragments associated with any one molecule is finite. Thus, the study of molecular similarity based on structural fragments might be more naturally formulated in terms of finite sets.

The intersection of two sets is a unique and natural matching of the members of the sets. The natural partial ordering of sets is based on the subset relation. (The preceding use of vectors in the screening step of substructure searching in large chemical databases is possibly more naturally viewed as being based on this partial ordering of sets of fragments.) Although many proximity measures have been defined on finite sets [15], we shall only present two common measures that have analogies in numerous other mathematical spaces. Assume D and D' are finite sets. The cardinality $|D|$ of D is defined to be the number of points in D. The Hamming metric $\mu_3(D, D')$ for finite sets is defined by

$$\begin{aligned}\mu_3(D, D') &= |D| + |D'| - 2|D \cap D'| \\ &= |D \backslash D'| + |D' \backslash D|, \qquad (7.1)\end{aligned}$$

where $D \backslash D'$ denotes the members of D that are not members of D'. The square of the Ochiai similarity coefficient [15] is defined by

$$s_3(D, D') = \frac{|D \cap D'|^2}{|D| \times |D'|} . \tag{7.2}$$

(The square of the Ochiai similarity coefficient relates more directly to equations (8.3), (9.2) and (11.7) than does the Ochiai coefficient itself.) Using the atoms sets $D = \{C_1, C_2, O_1, I_1\}$ and $D' = \{C_1, C_2, O_1, F_1\}$ from chemical graphs L and L' of fig. 1, we see that $\mu_3(D, D') = 2$ and $s_3(D, D') = 9/16$. When the list L of possible fragments is finite, one can easily set up vector descriptions under the L_1 metric and set descriptions under the Hamming metric that give identical measures of molecular similarity.

As already noted, most applications of similarity based on vectors are perhaps more appropriately viewed as applications of similarity based on finite sets. The Hamming distance is used as a prescreen for finding nearest neighbors based on the maximum common subgraph metric on labeled graphs [51]. The preceding proximity measures have been used to assess the clustering of active compounds in chemical description spaces [29].

8. Labeled graphs

Chemical graphs are by far the most popular chemical descriptions based on labeled graphs. A labeled graph with multiple loops and edges is sometimes called a labeled pseudograph [52]. Information on the free valence electrons and bond orders [53] and on the group and period of the atoms [54] can be added to the chemical graph using multiple loops and edges (arc a in fig. 3). By considering graphs with more than one component, Ugi et al. [53] represents the reactants (products) by a single labeled pseudograph. Crandell and Smith [55] let the edge labels of a complete graph (all vertices are connected) denote interatomic distances (arc j of fig. 3).

Sometimes, two labeled graphs are essentially defined on the same vertex set. For example, radioactive labeling often enables us to know which atoms in the reactants correspond with which atoms in the products. Such a correspondence can be viewed as an a priori matching. We will call it a complete matching since all vertices in both labeled graphs are paired. Assume $L = (V, E, l)$ and $L' = (V, E', l')$ are two labeled graphs with a common vertex set V. Then

$$\mu_4(L, L') = |E \backslash E'| + |E' \backslash E| \tag{8.1}$$

is a city-block metric on labeled graphs with vertex set V, where $E \backslash E'$ denotes the set of labeled edges in E that are not in E'.

Let L and L' be pseudographs in which the number of loops at a vertex corresponds to the number of free valence electrons of the corresponding atom and the number of edges connecting two vertices corresponds to the bond order of the corresponding bonded atoms. Then μ_4 becomes the chemical distance defined by Ugi et al. [53] under a fixed correspondence between all of the atoms of the products and all of the atoms of the reactants. Wochner et al. [56] call the smallest value of (8.1), under all complete matchings, the exact minimum of the chemical distance.

Proximity measures, not restricted to labeled graphs with a common vertex set, can be defined as follows [57,58]. Let $L = (V, E, l)$ be any labeled graph. Define the cardinality $|L|$ of L by $|V| + |E|$. Let $L' = (V', E', l')$ be any other labeled graph. Let $|\mathrm{Ma}CS(L, L')|$ denote the cardinality of the largest labeled graph which is a subgraph of both L and L'. The maximum common subgraph metric is then defined by

$$\mu_5(L, L') = |L| + |L'| - 2|\mathrm{Ma}CS(L, L')| . \tag{8.2}$$

A closely associated similarity measure [59] is defined by

$$s_4(L, L') = \frac{|\mathrm{Ma}CS(L, L')|^2}{|L| \times |L'|} . \tag{8.3}$$

Let M be a subgraph of both L and L'. If $|M| = |\mathrm{Ma}CS(L, L')|$, then M is called a maximum common subgraph of L and L'. The maximum common subgraph of the labeled subgraphs L and L' in fig. 1 is the epoxide ring. In this case there is only one maximum common subgraph, but generally there may be more than one [56,57]. Since the epoxide ring contains 6 elements (3 vertices and 3 edges) and since L and L' both contain 8 elements, respectively, $\mu_5(L, L') = 4$.

It should be noted that the subgraph partial ordering is directly involved in defining the maximum common substructures. If we consider only *connected* subgraphs of connected labeled graphs [60] or only *complete* subgraphs of complete labeled graphs [55], we obtain other useful definitions of maximum common substructures. It might also be noted that if two labeled graphs have a common vertex set and if each vertex is assigned a unique label, then eq. (8.2) reduces to eq. (8.1). In that sense, μ_5 can be viewed as a generalization of μ_4.

Hendrickson and Braun-Keller [61] define a dissimilarity measure given by

$$d_2(C, C') = \frac{1}{2} \sum_i (|\Delta h_i| + |\Delta z_i|) , \tag{8.4}$$

where C and C' are chemical graphs and where Δh_i and Δz_i refer, respectively, to the overall change in the number of hydrogens and heteroatoms on the ith carbon atom. As in eq. (8.1), this dissimilarity measure assumes that C and C' have common vertex sets. (Hendrickson and Braun-Keller [61] suggest that d_2 is a metric on chemical graphs. This is unlikely because $d_2(L, L') = 0$ on all 3-regular graphs in which every vertex is connected to 3 other vertices. Thus, one can show d_2 is not a metric if one can find two 3-regular chemical graphs of the same order which are not isomorphic.)

Substructure searching of chemical databases is an important application of MSA which uses the substructure partial ordering of chemical graphs [62]. The subgraph, connected subgraph, and complete subgraph partial orders have all been used to define maximal commonalities between compounds [53,56,60,63]. The subgraph partial order has also been used to computationally generate generic structures in Markush representations of compounds [64] and identify bond changes in reactions [65,66]. The counterpart of a maximum common subgraph is the minimum common subgraph [57], which has been used in representing and classifying chemical reactions [67,68]. Maximum common subgraphs have also been used to computationally generate structures lying on "lines" connecting two reference structures [54].

Ugi et al. [53] has developed an extensive theory of reaction transforms based on the vertex matching underlying eq. (8.1). By means of these transforms, one can classify reactions [56], suggest new reaction categories [69], generate reaction sequences [70] and reaction intermediates [56]. Equation (8.1) and related ideas have been incorporated into a principal of minimum chemical distance [2] as a "computable" version to the principal of minimal structural change noted in the introduction. Hendrickson and Braun-Keller [61] use their reaction distance to prune the generation of irrelevant reaction pathways connecting a set of reactants with a set of products.

The maximum common subgraph metric has been used for nearest neighbor property prediction [71] and for quantitative structure-activity analysis [59]. Johnson et al. [13] argue that concepts of proximity defined on *chemical graphs* are implicit in *all* general, nonrandom methods of searching for compounds with desired chemical properties.

9. Finite sequences

Chemical names, and consequently finite sequences, were probably the first chemical descriptions of molecules. The IUPAC and common chemical names are currently important finite sequences associated with molecules (arc w in fig. 3). Perhaps more useful from a computational viewpoint are the various forms of canonical codes which can be used to generate the connection table of the associated chemical graph [72,73–76] (arc g in fig. 3). For example, IC1CO1 and FC1CO1 are canonical SMILES codes [73] for L and L' in fig. 1. Two successive letters denote a single bond between the two corresponding atoms. Two letters followed by the same integer indicate a single bond between the two corresponding atoms that completes a ring. In the two codes, carbon C1 and oxygen O1 are bonded.

Finite sequences have been used in other ways to represent chemical information. Randić [77] assigns to each bond of a molecule a number reflecting the local environment of that bond. The bonds of the molecule are inversely ordered by the magnitude of their assigned number. The ith element of the finite sequence is the value assigned to the ith bond in this ordering (arc g in fig. 3). Along a different line, Randić [78] shows how a finite sequence description of a molecule can be generated from any proximity measure ν defined for a set of n molecules (arc k in fig. 3). To obtain the finite sequence for molecule C, the molecules are ranked by their similarity to C. Let $C_i(C)$ denote the ith molecule based on this ranking where, of course, $C_1(C)$ is C itself. Then, $C_1(C), \ldots, C_n(C)$ is the finite sequence for C. We shall call this a *permutation* sequence because each molecule appears once and only in the sequence.

An important partial ordering of sequences found in chemistry is the lexicographical ordering used in indexing molecules. A second partial ordering is based on the subsequence relation. A sequence $S = x_1, \ldots, x_k$ is a subsequence of $S' = y_1, \ldots, y_n$, $k \leqslant n$, if S can be obtained by delting one or more terms from S'. For example, the sequence AFAAB is the subsequence obtained from BABFAACACB by deleting the 1st, 3rd, 7th, 8th and 9th terms. A third partial ordering of sequences has been defined [77] when the elements of the sequence are numbers of decreasing magnitude. In this case, the sequences are partially ordered by the majorization relation, where S' is said to majorize S if $\Sigma x_i = \Sigma y_i$ and

$$x_1 + \ldots + x_j \leqslant y_1 + \ldots + y_j$$

for all j.

Barysz et al. [46] have proposed a proximity measure for permutation sequences. Assume $C_1, \ldots, C_n$ and $C_{i_1}, \ldots, C_{i_n}$ are two permutation sequences. The ordered subscripts form two vectors $(1, \ldots, n)$ and $(i_1, \ldots, i_n)$ (arc h in fig. 3). Substituting into eq. (5.3), we obtain a similarity measure called Spearman's rank correlation coefficient. A number of correlation coefficients are available [79] and have been tried [46,47].

The subsequence partial ordering underlies a number of proximity measures available for sequences [11,47]. Let $|S|$ denote the length of the sequence S, and let $|\mathrm{LCS}(S, S')|$ denote the length of the longest common subsequence of S and S'. Then, Coggins [80] notes that μ_6 defined by

$$\mu_6(S, S') = |S| + |S'| - 2|\mathrm{LCS}(S, S')| \qquad (9.1)$$

is a metric. For example, the longest common subsequence of the SMILES codes IC1OC1 and FC1OC1 is C1OC1. Thus, μ_6(IC1OC1,FC1OC1) = 2.

Alternatively, think of FClOCl as being obtained from IClOCl by deleting I and inserting F. Deletions (insertions) of single characters are examples of elementary operations one can perform on a sequence to obtain another sequence. The smallest number of such elementary operations that transforms one sequence into another defines a metric on finite sequences [76,81]. Different metrics arise depending on what elementary operations are admitted and what weights are associated with them [81]. The metric in (9.1) is obtained in this manner by restricting the elementary operations to insertions and deletions, and assigning each of these operations a weight of one.

A similarity measure on sequences is given by

$$s_5(S, S') = \frac{|\mathrm{LCS}(S, S')|^2}{|S| \times |S'|} \quad . \tag{9.2}$$

Although this similarity measure has yet to appear in the chemical literature, it is presented here because of its obvious analogy with the similarity measures given in eqs. (7.2) and (8.3). A related similarity measure

$$s_6(S, S') = 1 - \mu_6(S, S')/(|S| + |S'|) \tag{9.3}$$

has been proposed by Herndon and Bertz [76].

The lexicographical ordering of sequences is routinely used to locate particular structures in chemical databases. Subsequence matching is implicit in the generation of structural fragments from Wiswesser codes [82] (arc *h* in fig. 3). Randić and Wilkins [78] present a method of selecting molecules for screening that involves a matching of permutation sequences. Randić [77] shows how the partial ordering based on majorization can be used to examine if a particular chemical property can be expressed an an additive function of bond contributions. Jerman-Blažič et al. [47] use a clustering of finite sequence descriptions of molecules to enhance some quantitative structure-activity relationships. Herndon and Bertz [76] use the similarity measure in eq. (9.3) to compare two different canonical coding schemes.

10. Transforms

Let Z_i denote the atomic number of the ith atom in chemical graph C, and let δ_{ij} denote the topological distance between atoms i and j. The topological distance between atoms i and j is the number of edges in a minimal path connecting atom i to atom j (arc i in fig. 3). Following the approach of Soltzberg and Wilkins [83], Gabanyi et al. [84] define the topological transform t(s) for chemical graph C by

$$\mathrm{t}(s) = k(C) \sum \sum Z_i Z_j \frac{\sin(s\,\delta_{ij})}{s\,\delta_{ij}} \ , \tag{10.1}$$

where $k(C)$ is a normalizing constant such that

$$\int_0^\infty |t(s)|^2 \, ds = 1 \quad \text{(arc } m \text{ in fig. 3).}$$

Let t_1 and t_2 denote two such transforms for chemical graphs C_1 and C_2. Then, the distance between t_1 and t_2 based on the L_2 metric is given by

$$\mu_7(t_1, t_2) = \left(\int_0^\infty |t_1(s) - t_2(s)|^2 \, ds \right)^{1/2} . \tag{10.2}$$

Gabanyi et al. [84] use this metric as a dissimilarity measure for establishing quantitative structure-activity relationships.

11. Three-dimensional structures

Compounds exist in three-dimensional space. Included in the wide variety of 3D chemical descriptions that have been addressed from a similarity perspective are finite sets such as atomic configurations [85], surfaces such as molecular surfaces and potential contours [86], and various integrable functions such as molecular volumes [87] and molecular density functions [88]. Standard molecular modeling programs provide procedures for deriving 3D configurations from stereochemical structures (arc *o* in fig. 3), deriving scalar fields from 3D configurations (arc *r* in fig. 3), and deriving surfaces from scalar fields (arc *s* in fig. 3).

Proximity measures are computed quite differently, depending on whether or not the 3D chemical description is a finite set, a surface, or an integrable function. However, there are significant commonalities in the preliminary problem of appropriately superimposing two 3D descriptions.

A priori matchings use rotations and translations to optimally superimpose features of one 3D-structure that have been paired with features of another 3D-structure. The particular paired features may be selected based on hypotheses relating the behavior of particular atoms or functional groups [87,89], or they may be selected based on maximal matchings of distance matrices (see section 12) associated with the two molecules [55,90].

Canonical matchings arise when internal canonical coordinates are defined on the 3D-structures. For example, the first, second and third moments of inertia of the mass distribution may serve as canonical coordinates with the origin at the center of mass. A matching between two structures is then defined by simply aligning their respective internal canonical coordinates.

Maximal matchings arise if the translations and rotations are iteratively selected by gradient methods defined in terms of the proximity measure of interest [86]. The gradient methods often generate matchings which are locally maximal, but not globally maximal. Frequently, a matching is a composite of a priori, canonical and maximal processes [88].

11.1. LABELED FINITE SETS IN R^3

The most common use of labeled finite sets in R^3 is an atomic configuration. It can be specified by giving the position vector $\boldsymbol{u} = (u_x, u_y, u_z)$ for each atom. If u_i denotes the position vector of atom i, then a configuration on n atoms with respect to a reference frame is simply a vector $\boldsymbol{U}_n = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ in R^{3n} (arc p in fig. 3). The root mean square distance between two configurations $\boldsymbol{U}_n = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ and $\boldsymbol{V}_n = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$ is given by

$$\mathrm{rms}(\boldsymbol{U}_n, \boldsymbol{V}_n) = \sqrt{\left(\sum \delta(\boldsymbol{u}_i, \boldsymbol{v}_i)^2/n\right)} \ , \tag{11.1}$$

where $\delta(\boldsymbol{u}_i, \boldsymbol{v}_i)$ denoted the euclidean distance between the triples $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$.

Since the chemical features of a configuration of n points in R^3 are invariant under translations and rotations, eq. (11.1) is usually minimized with respect to those transformations. Let $\mathcal{T}$ denote the set of all transformations expressible as combinations of translations and rotations. Assume T is a transformation in $\mathcal{T}$, and write $\mathrm{T}(\boldsymbol{U}_n)$ for the vector that results from letting T act on $\boldsymbol{U}_n = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$. For example, assume T translates a configuration along the x-axis a distance of 1. Then, $\mathrm{T}(\boldsymbol{U}_n)$ is the vector $(v_1, \ldots, v_n)$ where

$$v_i = (u_{ix}, u_{iy}, u_{iz}) + (1,0,0), \qquad i = 1, \ldots, n.$$

Write $\mathcal{U}_n = \{\mathrm{T}(\boldsymbol{U}_n) | \mathrm{T} \in \mathcal{T}\}$ for the set of configurations in R^{3n}, where T varies over all transformations in $\mathcal{T}$. Then, $\mathcal{U}_n$ contains essentially the same information as $\boldsymbol{U}_n$ would if $\boldsymbol{U}_n$ were to be expressed in terms of internal coordinates. Assume $\boldsymbol{V}_n$ is a configuration in R^{3n}, and write $\mathcal{V}_n = \{\mathrm{T}(\boldsymbol{V}_n) | \mathrm{T} \in \mathcal{T}\}$. The root mean square distance between $\mathcal{U}_n$ and $\mathcal{V}_n$ can be defined as

$$d_3(\mathcal{U}_n, \mathcal{V}_n) = \min_{\mathrm{T} \in \mathcal{T}} \mathrm{rms}(\boldsymbol{U}_n, \mathrm{T}(\boldsymbol{V}_n)) \, . \tag{11.2}$$

By replacing the expression $\delta(\boldsymbol{u}_i, \boldsymbol{v}_i)^2$ in eq. (11.1) with $m_i \delta(\boldsymbol{u}_i, \boldsymbol{v}_i)^2$ where m_i is the mass of atom i and then using ëq. (11.2), we obtain the metric dèveloped by Mezey [91,92] using mass-weighted cartesian coordinates [93].

The distance in eq. (11.1) can be viewed as the length of a straight line connecting $\boldsymbol{U}_n$ and $\boldsymbol{V}_n$. One can also think of distance along a continuous curve originating at $\boldsymbol{U}_n$ and terminating at $\boldsymbol{V}_n$. These generalizations are made in Mezey [94] using mass-weighted coordinates.

Equation (11.2) is based on an a priori matching specifying that atom i of $\boldsymbol{U}_n$ is to be paired with atom i of $\boldsymbol{V}_n$. A maximal dissimilarity measure can be specified by means of permutations defined on the indices. Let P denote a permutation of the indices $(1, \ldots, n)$, and let $\mathrm{P}(U_n)$ denote U_n after the indices have been permuted. For example, assume P interchanges indices 1 and 2. If $U_3 = (\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3)$, then $\mathrm{P}(U_n) = (\boldsymbol{u}_2, \boldsymbol{u}_1, \boldsymbol{u}_3)$. When a permutation maps index i to index j, we require that any associated labels, such as the types of the atoms, be identical. Let $\mathcal{P}$ denote the set of admissible permutations. A minimal root mean square dissimilarity measure between $\mathcal{U}_n$ and $\mathcal{V}_n$ is given by

$$d_4(\mathcal{U},\mathcal{V}) = \min_{\substack{\mathrm{T} \in \mathcal{T} \\ \mathrm{P} \in \mathcal{P}}} \mathrm{rms}(U_n, \mathrm{T} \circ \mathrm{P}(V_n)), \tag{11.3}$$

where $\circ$ denotes the composition of two functions. To our knowledge, equation (11.3) has never been computed except for configurations involving only a small number of atoms. The preceding methods of superimposing 3D chemical descriptions provide computationally accessible approximations to (11.3).

The dissimilarity measures in eqs. (11.1) to (11.3) do not seem to have been extended to configurations involving differing numbers of points. They could be, at least in principle, using the group similarity methods of cluster analysis [95] or a slight modification of the scoring methods proposed in Kuntz et al. [6] and Des-Jarlais et al. [7]. The minimum steric difference proposed by Balaban et al. [96] can compare configuration sets with differing numbers of points. Here, if a pair of points u and v satisfy an arbitrarily defined proximity constraint, they are called superimposable. The minimum steric difference is a weighted count of the number of nonsuperimposable atoms, where the weight is based on the group of the atom in the periodic table. The authors do not clearly specify the rigid transformations used to set up their matching.

Root mean square measures are commonly used in checking proposed functional correspondences between molecules. They are increasingly being used in the retrieval of 3D substructure querying [97,98]. They are also used to study how the shape of the charge density distribution changes with a change in conformation [99]. The minimum steric difference has been used in quantitative structure-activity studies of pharmacophores [96]. Finally, the computation of the distance along a reaction path in R^{3n} [94] is used in the study of Hammond's postulate using explicit concepts of similarity [100].

11.2. SURFACES

Chemical descriptions of surfaces in R^3 are just now being introduced into MSA. Van der Waals surfaces and electrostatic potential contours [88] are examples. (Section 1 references related work on surface complementarity, and section 13 covers groups abstracted from surfaces.)

In comparing two surfaces of two molecules, Chau and Dean [101] begin with a canonical matching of the two molecules which superimposes the molecular centroids. A set of rays emanating from the superimposed centroids are defined. Let y_k and z_k denote the length along the ray where surfaces S_y and S_z first intersect the kth ray. Chau and Dean compute the correlation coefficient between $(y_1, \ldots, y_n)$ and $(z_1, \ldots, z_n)$ defined by eq. (4) as a measure of similarity. They also calculate the correlation coefficient based on the ranks [79]. To date, their work has been primarily directed toward finding optimal matches between two surfaces.

Along a different line, Leicester et al. [102] represents a surface by a convergent series of spherical harmonics as given by

$$r(\theta, \emptyset) = \sum_{l=0}^{\infty} \sum_{m=-1}^{+1} a_{lm} Y_l^m(\theta, \emptyset), \tag{11.4}$$

where $(r(\theta, \emptyset), \theta, \emptyset)$ denotes a point on the surface expressed in spherical coordinates (arc u in fig. 3). A vector consisting of the coefficients in (11.4) is formed (arc n in fig. 3). For computational reasons, an upper limit is set for the index l. A dissimilarity measure is obtained by multiplying the vector of coefficients of one surface by the complex conjugate of the vector of coefficients of the other surface. The complex conjugate is used because the coefficients may be complex numbers. This dissimilarity measure is a generalization to complex numbers of the square of the dissimilarity measure of eq. (5.2). Leicester et al. [102] note the possibility of using interactive graphics to obtain various a priori matchings so as to minimize their dissimilarity measure.

11.3. INTEGRABLE SCALAR FIELDS

Chemical descriptions used in molecular similarity analysis that involve integrable functions defined over R^3 are quite varied. Perhaps the most common are 3D volumes [89,103–106]. However, electron densities [107] and molecular orbitals [88] are also being studied.

Some proximity measures for integrable functions f and g include the L_1 metric

$$\mu_7(f, g) = \int |f - g| \mathrm{d}\boldsymbol{u}, \tag{11.5}$$

the L_2 or euclidean metric

$$\mu_8(f, g) = \left(\int |f(\boldsymbol{u}) - g(\boldsymbol{u})|^2 \, \mathrm{d}\boldsymbol{u} \right)^{1/2}, \tag{11.6}$$

and the correlation coefficient

$$s_7(f, g) = \frac{\int f(\boldsymbol{u}) g(\boldsymbol{u}) \mathrm{d}\boldsymbol{u}}{\left[\int f^2(u) \mathrm{d}\boldsymbol{u} \int g^2(\boldsymbol{u}) \mathrm{d}\boldsymbol{u}\right]^{1/2}}. \tag{11.7}$$

If V is a volume in R^3, then by defining $f(\boldsymbol{u}) = 1$ if $\boldsymbol{u}$ is inside V and $f(\boldsymbol{u}) = 0$ if $\boldsymbol{u}$ is not in V, we obtain a number of common proximity measures based on volumes. For example, the numerator in (11.7) would correspond to the volume intersection of Hopfinger [89] and the metric in eq. (11.5) would correspond to a measure of the excluded volume of Motoc et al. [105]. Implicit in the definition of the proximity measures associated with eqs. (11.5) to (11.7) is a reasonable matching of the domains of the two integrable functions.

Similarity based on integrable functions in R^3 have been used primarily in pharmacophore analysis [105] and in developing quantitative structure-activity relationships [89,96,107]. Carbó and Domingo [88] prese' a similarity analysis of molecular orbitals.

12. Distance matrices

Any finite set $U = \{u_1, \ldots, u_k\}$ with a distance $\delta(u_i, u_j) = m_{ij}$ defined between every pair of points u_i and u_j defines a distance matrix $M = \{m_{ij}\}$, where m_{ij} denotes the matrix element in the ith row and jth column. Call M the distance matrix of U. For example, m_{ij} could denote the through space distance between atoms i and j of a configuration in R^3 (arc q in fig. 3). Note that enantiomeric distinctions are lost in this mapping of a configuration to a distance matrix.

Let $M = \{m_{ij}\}$ and $N = \{n_{ij}\}$ be two distance matrices of order k. By representing M by the vector $(m_{11}, \ldots, m_{1k}, \ldots, m_{k1}, \ldots, m_{kk})$, one can easily see that

$$\mu_9(M, N) = \left[\sum_{ij}^{n} (m_{ij} - n_{ij})^2\right]^{1/2} \tag{12.1}$$

is a metric on the space of distance matrices of order k (arc l of fig. 3). Denote $\mu_9(M, N)$ divided by $\sqrt{[k(k-1)]}$ by rmsd(M, N). Danziger and Dean [90] call rmsd(M, N) the root mean square of the difference distance matrix.

Equation (12.1) is based on an a priori matching of the ith row of M with the ith row of N. Such an a priori matching is natural when assessing the similarity between the distance matrices of two conformations of a single compound. At other times, a maximal matching is desired [90]. Proceeding as we did in the development of eq. (11.3), let $M = \{m_{ij}\}$ denote the distance matrix associated with a set $U = \{u_1, \ldots, u_k\}$. Let P denote a permutation of the indices of the elements of U, and let $\mathrm{P}(M) = \{m_{p(i)p(j)}\}$. Again, let $\mathcal{P}$ denote the set of admissible permutations which preserves the values of any labeling function defined on U. Then

$$d_5(M, N) = \min_{\mathrm{P} \in \mathcal{P}} \mathrm{rmsd}(M, \mathrm{P}(N)) \tag{12.2}$$

is a dissimilarity measure defined on distance matrices.

Difference distance matrices have been used for visualizing similarities and dissimilarities in compounds based on their interatomic distances [108,109]. Although computational solutions of (12.2) are still beyond current computational speeds, useful approximations are available [6,90,98]. These approximations are used in screening matchings used in computational approximations to eq. (11.3) for 3D finite sets. Root mean square dissimilarity measures are also being used to screen compounds in 3D substructure querying [98].

13. Groups

Mezey [111,112] shows how a number of different groups can be used to describe relationships between the concave, convex and saddle point regions of, for example, van der Waals surfaces and potential energy hypersurfaces (arc v in fig. 3). The details of abstracting these groups from a three-dimensional surface involve concepts in algebraic topology, and the reader is referred to the original articles for details.

For the most part, the similarity concepts employed in the articles are based on the assumption that molecules with shapes mapped to isomorphic groups will exhibit similar chemical behaviors. Those molecules mapped to the same group form an equivalence class. As noted earlier, the use of the equivalence concept of similarity is not in itself sufficient to include a study in this investigation. However, Mezey [112] also makes use of the fact that the groups are partially ordered by the subgroup relation. A group is associated with a truncated reaction surface in which the surface has been removed at points in which the potential energy exceeds a truncation value. This results in a surface with holes. Different hole structures associated with different truncation values may or may not be mapped to the same group. The partial ordering of groups was used to relate the changes in the hole structures. Interestingly, the development of this partial ordering makes no use of matchings.

In the work just described, a group is associated with a single entity, a reaction surface with a particular hole structure. In what follows, a group [113] and a set of Betti numbers [114] is associated with a *pair* of molecules. This association will be called a group association and denoted by $\gamma(D, D')$ in the first case, and will be called a Betti association and denoted by $\beta(D, D')$ in the second case. A proximity measure $\nu(D, D')$ also associates something (a number) with a pair D and D' of mathematical structures. A proximity measure takes values in a set whose members are numbers. In constrast, group and Betti associations take values in sets whose members are groups and sets of Betti numbers, respectively. Because notions of distance, addition and multiplication are defined on the set R of real numbers, we can do many things with proximity measures. It remains to be seen what can be done with group and Betti associations. Mezey and coworkers [114] provide one illustration of how Betti associations might be used to explain the regioselectivity of chemical reactions.

14. Summary notes

Table summarizes where in the text that notes, references or discussion can be found relevant to the derivational arcs in fig. 3. Proximity measures have been proposed in MSA for all of the mathematical spaces in fig. 3 with the exception of the space of stereochemical structures. Mathematical representations of a stereochemical structure are currently being proposed. Possible candidates include numbers [115], canonically defined character strings [116], and *n*-ary relations [117]. Further work needs to be done in this area because of the critical role stereochemical structures play in fig. 3.

Table 1

Sections containing material relevant to the various conversion arcs in fig. 3

Arc	Section	Arc	Section	Arc	Section	Arc	Section	Arc	Section
a	2,8	*b*	7	*c*	7	*d*	4,5	*e*	6
f	6	*g*	9	*h*	9	*i*	10	*j*	8
k	9	*l*	12	*m*	10	*n*	11.2	*o*	11
p	11.1	*q*	12	*r*	11	*s*	11	*t*	11.2
u	11.2	*v*	13	*w*	9				

Matchings, partial orders and proximity measures are distinct mathematical concepts of similarity. Matching is emphasized when searching for interesting commonalities between compounds, partial orderings are emphasized when querying databases for compounds with specified structural attributes, and proximity measures are emphasized when clustering compounds and predicting chemical properties. However, as we have seen, these three similarity concepts are often closely intertwined. Maximal matchings are usually defined in terms of partial orderings and many similarity measures are defined in terms of maximal matchings. In particular, the city-block metrics (eqs. (5.1), (7.1), (8.1), (8.2), (9.1), (11.5)) fall under the general theory of value functions defined on partially ordered sets [118–121]. In fact, the concepts of equivalence, matching, partial order, and proximity all fall under a more general theory of fuzzy relations [122]. However, fuzzy relations have not yet been used in MSA.

From another angle, we see that the computational accessibility of numbers, codes, finite sets, and product spaces result in their wide use in similarity applications involving chemical databases. The fact that one can visually superimpose 3D structures on the computer has resulted in the wide use of 3D superpositioning in finding interesting commonalities between molecules. The discrete nature of labeled graphs possibly underlies their wide use in computationally generating structures and modeling reaction pathways. Almost all of these mathematical spaces are being used in predicting chemical properties.

This study has pointed out the mathematical structures and mathematical concepts of similarity that help to relate the diverse chemical descriptions and concepts of molecular similarity in MSA. A better knowledge of the relationships between the chemical descriptions and similarity concepts of MSA should lead to a better understanding of the development and application of these concepts.

Acknowledgements

The author is indebted to Gerald Maggiora and Dennis Rouvray for encouraging this work and for critical discussions on the concept of molecular similarity, and to The Upjohn Company for providing a stimulating research environment. The referees are to be thanked for a number of helpful comments.

References

[1] C.L. Wilkins and M. Randić, Theor. Chim. Acta 58(1979)45.

[2] C. Jochum, J. Gasteiger and I. Ugi, Angew. Chemie Int. 19(1980)495.

[3] G.W. Klump, *Reactivity in Organic Chemistry* (Wiley, New York, 1982).

[4] G.M. Crippen, J. Med. Chem. 22(1979)988.

[5] Z. Simon, A. Chiriac, S. Holban, D. Ciubotaru and G.I. Mihalas, *Minimum Steric Difference: The MTD Method for QSAR Studies* (Research Studies Press Ltd., Letchworth, 1984).

[6] I.D. Kuntz, J.M. Blaney, S.J. Oatley, R. Langridge and T.E. Ferrin, J. Mol. Biol. 161(1982) 269.

[7] R.L. DesJarlais, R.P. Sheridan, G.L. Seibel, J.S. Dixon, I.D. Kuntz and R. Venkataraghavan, J. Med. Chem. 31(1988)722.

[8] D.J. DuChamp, in: *Computer-Assisted Drug Design*, ed. E.C. Olson and R.E. Christoffersen (ACS Symp. Ser. 112, Amer. Chem. Soc., Washington D.C., 1979) p. 79.

[9] A.J. Hopfinger, in: *Quantitative Structure-Activity Relationships (QSAR) in Drug Design*, ed. J.L. Fauchère (Alan R. Liss, Inc., New York), to appear.

[10] P.H.A. Sneath, J. Theor. Biol. 12(1966)157.

[11] D. Sankoff and J.B. Kruskal, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison* (Addison-Wesley, London, 1983).

[12] J.P. Tremblay and R. Manohar, *Discrete Mathematical Structures with Applications to Computer Science* (McGraw-Hill, New York, 1975).

[13] M.A. Johnson, G.M. Maggiora and S. Basak, in: *Proc. Sixth Int. Conf. on Mathematical Modeling*, ed. X. Avula and E.Y. Rodin (Pergamon Press, 1987) 630.

[14] I. Borg and J. Lingoes, *Multidimensional Similarity Structure Analysis* (Springer-Verlag, New York, 1987).

[15] J.C. Gower, in: *Encylopedia of Statistical Sciences*, ed. S. Kotz and N.L. Johnson (Wiley, New York) 5(1985)397.

[16] A.J. Stupor, W.E. Brugger and P.C. Jurs, *Computer-Assisted Studies of Chemical Structure and Biological Function* (Wiley, New York, 1979).

[17] A.T. Balaban, I. Motoc, D. Bonchev and O. Mekenyan, in: *Steric Effects in Drug Design*, ed. M. Charton and I. Motoc (Springer-Verlag, Berlin, 1983) 23.

[18] M. Randić, J. Chem. Inf. Sci. 24(1984)164.

[19] M. Randić, J. Chem. Inf. Sci. 26(1986)134.

[20] M. Razinger, J.R. Chrétien and J.E. Dubois, J. Chem. Inf. Sci. 25(1985)23.

[21] K. Szymanski, W.R. Muller, J.V. Knop and N. Trinajstić, Int. J. Quant. Chem.: Quant. Chem. Symp. 20(1986)173.

[22] M. Randić, Int. J. Quant. Chem.: Quant. Biol. Symp. 11(1984)137.

[23] G.T. Rasmussen and T.L. Isenhour, J. Chem. Inf. Comput. Sci. 19(1979)179.

[24] M. Randić and C.L. Wilkins, J. Chem. Inf. Comput. Sci. 19(1979)31.

[25] R.E. Carhart, D.H. Smith and R. Vankataraghavan, J. Chem. Inf. Comput. Sci. 25(1985)64.

[26] P. Willett, V. Wintermann and D. Bawden, J. Chem. Inf. Comput. Sci. 26(1986)109.

[27] P. Willett and V. Wintermann, Quant. Struct.-Act. Relat. 5(1986)18.

[28] S.C. Basak, V.R. Magnuson, G.J. Niemi and R.R. Regal, Discrete Appl. Math. 19(1987)17.

[29] M.A. Johnson, M.S. Lajiness and G.M. Maggiora, in: *Quantitative Structure-Activity Relationships (QSAR) in Drug Design*, ed. J.L. Fauchère (Alan R. Liss, Inc., New York, 1989) p. 167.

[30] C. Hansch, S.H. Unger and A.B. Forsythe, J. Med. Chem. 16(1973)1217.

[31] P. Willett, *Similarity and Clustering in Chemical Information Systems* (Research Studies Press, Letchworth, 1987).

[32] M.F. Lynch, in: *Chemical Information Systems*, ed. J.E. Ash and E. Hyde (Ellis Horwood, Chichester, 1975) Ch. 12.

[33] R.N. Shepard, A.K. Romney and S.B. Nerlove, *Multidimensional Scaling*, Vol. I, (Seminar Press, New York, 1972).

[34] T.M. Cover, IEEE Trans. Inf. Theory 14(1968)50.

[35] J. van Ryzin, *Classification and Clustering* (Academic Press, New York, 1977).

[36] B.R. Kowalski and C.F. Bender, J. Amer. Chem. Soc. 94(1972)5632.

[37] J.W. McFarland and D.J. Gans, J. Med. Chem. 29(1986)505.

[38] M.F. Delaney, J.R. Hallowell, Jr. and F.V. Warren, Jr., J. Chem. Inf. Comput. Sci. 25 (1985)27.

[39] D. Grier, W.D. Hounshell, T. Moock and G. Grethe, Poster at Amer. Chem. Soc. Mtg., Los Angeles, CA (1988).

[40] M.S. Lajiness, M.A. Johnson and G.M. Maggiora, in: *Quantitative Structure-Activity Relationships (QSAR) in Drug Design*, ed. J.L. Fauchère (Alan R. Liss, Inc., New York, 1989) p. 173.

[41] W.J. Streich, S. Dove and R. Franke, J. Med. Chem. 23(1980)1452.

[42] R. Wootton, J. Med. Chem. 26(1983)275.

[43] M. Randić, B. Jerman-Blažič, D.H. Rouvray, P.G. Seybold and S.C. Grossman, Int. J. Quant. Chem.: Quant. Chem. Biol. Symp. (1987), forthcoming.

[44] C.L. Wilkins, M. Randić, S.M. Schuster, R.S. Markin, S. Steiner and L. Dorgan, Anal. Chem. Acta 133(1981)637.

[45] D. Bonchev and N. Trinajstić, Int. J. Quant. Chem.: Quant. Chem. Symp. 16(1982)463.

[46] M. Barysz, N. Trinajstić and J.V. Knop, Int. J. Quant. Chem.: Quant. Chem. Symp. 17 (1983)441.

[47] B. Jerman-Blažič, I. Fabič and M. Randić, J. Comp. Chem. 7(1986)176.

[48] H. Jeffreys, *Theory of Probability* (Clarendon Press, Oxford, 1961).

[49] C.R. Rao, in: *Classification and Clustering*, ed. J. van Ryzin (Academic Press, New York, 1977) p. 175.

[50] S.H. Bertz and W.C. Herndon, in: *Artificial Intelligence Applications in Chemistry*, ed. T.H. Pierce and B.A. Hohne (ACS Symp. Ser. 306, Amer. Chem. Soc., Washington D.C., 1986) 169.

[51] V. Nicholson, C.-C. Tsai, M. Johnson and M. Naim, in: *Graph Theory and Toplogy in Chemistry*, ed. R.B. King and D.H. Rouvray (Elsevier, Amsterdam, 1987) p. 226.

[52] G. Chartrand and L. Lesniak, *Graphs and Digraphs* (Wadsworth and Brooks, Monterey, 1986).

[53] I. Ugi, J. Bauer, J. Brandt, J. Friedrich, J. Gasteiger, C. Jochum and W. Schubert, Angew. Chem. Int. Ed. Engl. 18(1979)111.

[54] M. Johnson, M. Naim, V. Nicholson and C.-C. Tsai, in: *Graph Theory and Topology in Chemistry*, ed. R.B. King and D.H. Rouvray (Elsevier, Amsterdam, 1987) p. 219.

[55] C.W. Crandell and D.H. Smith, J. Chem. Inf. Comput. Sci. 23(1983)186.

[56] M. Wochner, J. Brandt, A. van Scholley and I. Ugi, Chmia 42(1988)217.

[57] M.A. Johnson, in: *Graph Theory and Its Applications to Algorithms and Computer Science*, ed. Y. Alavi, G. Chartrand, L. Lesniak, D.R. Lick and C.E. Wall (Wiley, New York, 1985) p. 457.

[58] V. Baláž, J. Koča, V. Kvasnička and M. Sekanina, Casopis Pro Pest. Mat. 111(1986)431.

[59] C.-C. Tsai, V. Nicholson, M.A. Johnson and M. Naim, in: *Graph Theory and Topology in Chemistry*, ed. R.B. King and D.H. Rouvray (Elsevier, Amsterdam, 1987) p. 231.

[60] M.M. Cone, R. Venkataraghavan and F.W. McLafferty, J. Amer. Chem. Soc. 99(1977) 7668.

[61] J.B. Hendrickson and E. Braun-Keller, J. Comput. Chem. 1(1980)323.

[62] J. Ash, P. Chubb, S. Ward, S. Welford and P. Willett, *Communication, Storage and Retrieval of Chemical Information* (Horwood, Chichester, 1985).

[63] A.T. Brint and P. Willett, J. Mol. Graphics 5(1987)200.

[64] Fraser Williams (Scientific Systems) Ltd. brochure, Cheshire, U.K. (1988).

[65] M.F. Lynch and P. Willett, J. Chem. Inf. Comput. Sci. 18(1978)154.

[66] M.A. Johnson, in: *Proc. Sixth Int. Conf. on the Theory and Applications of Graphs*, ed. Y. Alavi, G. Chartrand, O. Oellermann and A.J. Schwenk (Wiley, New York, 1988), to appear.

[67] S. Fujita, J. Chem. Inf. Comput. Sci. 26(1986)205.

[68] S. Fujita, J. Chem. Inf. Comput. Sci. 27(1987)120.

[69] J. Bauer, R. Herges, E. Fontain and I. Ugi, Chimia 39(1985)43.

[70] E. Fontain, J. Bauer and I. Ugi, Chem. Lett. (1987)37.

[71] M.A. Johnson, M. Naim, V. Nickolson and C.-C. Tsai, in: *QSAR in Drug Design and Toxicology*, ed. D. Hadži and B. Jerman-Blažič (Elsevier, Amsterdam, 1987) p. 67.

[72] E.G. Smith and P.A. Baker, *The Wiswesser Line-Formula Chemical Notation* (Chemical Information Management, Inc., Cherry Hill, 1975).

[73] *Med. Chem. Software Manual* (Medicinal Chemistry Project, Pomona College, Claremont, CA, 1984).

[74] W.T. Wipke and T.M. Dyott, J. Amer. Chem. Soc. 96(1974)4834.

[75] R.C. Read, J. Chem. Inf. Comput. Sci. 23(1983)135.

[76] W.C. Herndon and S.H. Bertz, J. Comput. Chem. 8(1987)367.

[77] M. Randić, Int. J. Quant. Chem.: Quant. Chem. Biol. Symp. 5(1978)245.

[78] M. Randić and C.L. Wilkins, Int. J. Quant. Chem.: Quant. Chem. Biol. Symp. 6(1979)55.

[79] W.J. Conover, *Practical Nonparametric Statistics* (Wiley, New York, 1971).

[80] J.M. Coggins, in: *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, ed. D. Sankoff and J.B. Kruskal (Addison-Wesley, London, 1983) Ch. 11.

[81] J.B. Kruskal, in: *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, ed. D. Sankoff and J.B. Kruskal (Addison-Wesley, London, 1983) Ch. 1.

[82] G.W. Adamson and D. Bawden, J. Chem. Inf. Comput. Sci. 15(1975)215.

[83] L.J. Soltzberg and C.L. Wilkins, J. Amer. Chem. Soc. 99(1977)439.

[84] Z. Gabanyi, P. Surjan and G. Naray-Szabó, Eur. J. Med. Chem. 17(1982)307.

[85] A.D. McLachlan, Acta Cryst. A38(1982)871.

[86] P.M. Dean and P.-L. Chau, J. Mol. Graphics 5(1987)152.

[87] M. Marsili, P. Floersheim and A.S. Dreiding, Comput. & Chem. 7(1983)175.
[88] R. Carbó and L. Domingo, Int. J. Quant. Chem. 32(1987)517.
[89] A.J. Hopfinger, J. Amer. Chem. Soc. 102(1980)7196.
[90] D.J. Danziger and P.M. Dean, J. Theor. Biol. 116(1985)215.
[91] P.G. Mezey, Int. J. Quant. Chem. 26(1984)983.
[92] P.G. Mezey, Int. J. Quant. Chem.: Quant. Chem. Biol. Symp. 17(1983)137.
[93] E.B. Wilson, Jr., J.C. Decius and P.C. Cross, *Molecular Vibrations* (McGraw-Hill, New York, 1955) p. 14.
[94] P.G. Mezey, *Potential Energy Hypersurfaces* (Elsevier, Amsterdam, 1987).
[95] B. Everitt, *Cluster Analysis* (Halstead Press, New York, 1980) Ch. 2.
[96] A.T. Balaban, A. Chiriac, I. Motoc and Z. Simon, *Steric Fit in Quantitative Structure-Activity Relations* (Springer-Verlag, Berlin, 1980) Ch. 4.
[97] P. Gund, W.T. Wipke and R. Langridge, in: *Proc. Int. Conf. on Computers in Chemical Research and Education,* ed. D. Hadzi (Elsevier, Amsterdam, 1973) p. 5.
[98] A.T. Brint and P. Willett, J. Mol. Graphics 5(1987)49.
[99] G.A. Arteca and P. Mezey, Int. J. Quant. Chem.: Quant. Biol. Symp. 14(1987)133.
[100] G.A. Arteca and P. Mezey, J. Comput. Chem. 9(1988)728.
[101] P.L. Chau and P.M. Dean, J. Mol. Graphics 5(1987)97.
[102] S.E. Leicester, J.L. Finney and R.P. Bywater, J. Mol. Graphics 6(1988)104.
[103] J.E. Moore, G. Palmieri and E. Wanke, Nature 216(1967)1084.
[104] G.R. Marshall, C.D. Barry, H.E. Bosshard, R.A. Dammkoehler and D.A. Dunn, in: *Computer-Assisted Drug Design,* ed. E.C. Olson and R.E. Christoffersen (American Chemical Society, Washington D.C., 1979) p. 205.
[105] I. Motoc, G.R. Marshall, R.A. Dammkoehler and J. Labanowski, Z. Naturforsch. 40a(1985) 1108.
[106] T.R. Stouch and P.C. Jurs, J. Chem. Inf. Comput. Sci. 26(1986)4.
[107] R. Carbó, L. Leyda and M. Arnau, Int. J. Quant. Chem. 17(1980)1183.
[108] K. Nishikawa and T. Ooi, J. Theor. Biol. 43(1974)351.
[109] M.N. Liebman, *Molecular Structure and Biological Activity,* ed. J. Griffin and W.L. Duax (Elsevier, New York, 1982) p. 193.
[110] P. Mezey, Int. J. Quant. Chem.: Quant. Biol. Symp. 12(1986)113.
[111] P. Mezey, J. Comput. Chem. 8(1987)462.
[112] P. Mezey, Theor. Chim. Acta (Berl.) 67(1985)91.
[113] P. Mezey, Int. J. Quant. Chem.' Quant. Biol. Symp. 14(1987)127.
[114] G.A. Arteca, V.B. Jammal and P.G. Mezey, J. Comp. Chem. 9(1988)608.
[115] H. Beierbeck, J. Chem. Inf. Comput. Sci. 22(1982)215.
[116] W.T. Wipke and T.M. Dyott, J. Amer. Chem. Soc. 96(1974)4834.
[117] K. Wirth, J. Chem. Inf. Comput. Sci. 26(1986)242.
[118] B. Monjardet, Discrete Math. (1981)173.
[119] S.A. Boorman and P. Arabie, in: *Multidimensional Scaling,* Vol. I, ed. R.N. Shepard, A.K. Romney and S.B. Nerlove (Seminar Press, New York, 1972) p. 225.
[120] S.A. Boorman and D.C. Oliver, J. Math. Psychol. 10(1973)26.
[121] J.-P. Barthélemy, Math. Sci. Hum. 16(1978)39.
[122] D. Dubois and H. Prade, *Fuzzy Sets and Systems: Theory and Applications* (Academic Press, Orlando, 1980).