early 1960s in providing an outstanding CAS building facility has been a great aid in attracting and retaining good personnel. Individual offices are provided for most scientists, maintained beautifully with individual controls for light and temperature. Over 50 acres of landscaped ground surround the current two buildings. There is ample, easily accessed parking. Such facilities have been a factor in holding editorial staff turnover to only a few percentage points annually. This is a great asset because initial CAS training for a new Ph.D. scientist requires 12-18 months, a significant investment.

A final perspective is obvious. The described major organizational changes have provided an efficient editorial operation through which CAS can continue to supply timely, complete, high-quality chemical information into the future. The excellently trained and dedicated staff is CAS's greatest asset. It assures the integrity of the data base for the years to come.

REFERENCES AND NOTES

(1) Wigington, R. L. "Computer Architecture for Editorial Processing Within an Integrated Publishing Organization". *J. Res. Commun. Stud.* **1979/1980**, *2*, 25-38.
(2) Seybold, J. W. "Data Base and Journal Publishing at the American Chemical Society". *Seybold Rep. Publ. Syst.* **1982**, *11* (24), 3-16.
(3) "CAS Today", 1980 ed.; Chemical Abstracts Service: Columbus, OH, 1980; p 11.
(4) Platau, G. O.; Metanomski, W. V. "Productivity and Its Measurement at Chemical Abstracts Service". *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 8-11.
(5) "The First 75 Years of CAS". *CAS Rep.* **1982**, *No. 12*, 3-9.
(6) Baker, D. B.; Horiszny, J. W.; Metanomski, W. V. "History of Abstracting at Chemical Abstracts Service". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 193-201.
(7) Weil, B. H. "Standards for Writing Abstracts". *J. Am. Soc. Inf. Sci.* **1970**, *21*, 351-357.
(8) Platau, G. O. "Annual Report, Assignment and Abstracting Unit of Publications Division, 1970"; Chemical Abstracts Service: Columbus, OH, 1971.
(9) "Annual Report, CAS Editorial Operations Division, 1970"; Chemical Abstracts Service: Columbus, OH, 1971.
(10) Dittmar, P. G.; Stobaugh, R. E.; Watson, C. E. "The Chemical Abstracts Service Registry System. I. General Design". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 111-121.
(11) Weisgerber, D. W. "Finding Chemical Compounds by CAS Registry Numbers". *Ind. Res./Dev.* **1981**, *23* (5), 156-160.
(12) Rowlett, R. J., Jr.; Tate, F. A. "A Computer-Based System for Handling Chemical Nomenclature and Structural Representations". *J. Chem. Doc.* **1972**, *12*, 125-128.
(13) Donaldson, N.; Powell, W. H.; Rowlett, R. J., Jr.; White, R. W.; Yorka, K. V. "*Chemical Abstracts* Index Names for Chemical Substances in the Ninth Collective Period (1972-1976)". *J. Chem. Doc.* **1974**, *14*, 3-15.
(14) "Annual Report, CAS Editorial Processing Division, 1972"; Chemical Abstracts Service: Columbus, OH, 1973.
(15) "Annual Report, CAS Editorial Processing Division, 1976"; Chemical Abstracts Service: Columbus, OH, 1977.
(16) Weisgerber, D. W. "Applications of Technology to CAS Data Base Production". *Inf. Serv. Use* **1984**, *4*, 317-325.

ARTICLES

Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications

RAYMOND E. CARHART,*† DENNIS H. SMITH,† and R. VENKATARAGHAVAN

Lederle Laboratories, Pearl River, New York 10965

Received March 30, 1984

A simple type of substructure called an atom pair is defined in terms of the atomic environments of, and shortest path separations between, *all* pairs of atoms in the topological representation of a chemical structure. An algorithm is presented for computing atom pairs from such a representation. Two applications of atom pairs to structure-activity problems are described. In the first, a measure of similarity between compounds is defined, and the use of this measure in probing large databases of structures is discussed. In the second, a heuristic technique called trend vector analysis is described. The trend vector summarizes the correlation, within a set of structures, of the occurrence of atom pairs of different types with measured biological activity. These correlations can be used to estimate the biological activity of new compounds. A comparison of trend vector analysis with discriminant plane analysis is presented for one series of compounds.

INTRODUCTION

There currently are several computational methods for exploring relationships between chemical structures and measured properties, especially biological effects, of compounds. Blankley[1] has recently reviewed many of these methods. These techniques share a common problem; how does one express an irregular object like a chemical structure in a regular form that allows the quantitative comparing and contrasting of those structures? Most approaches rely upon *molecular descriptors*, which are numerical values representing selected features of the compounds. Each structure is represented as a list, or vector, of such numerical descriptors and thus may be thought of as a point in a high-dimensional space, with coordinates equal to (or related to) the corresponding descriptor values. With this change of representation, the problem of relating structure to biological activity becomes one of relating position (in the high-dimensional space) with activity. A variety of powerful mathematical techniques (e.g., pattern recognition,[2-4] multiple linear regression,[5] SIMCA[6]) is available for treating such problems.

Many different types of descriptors have been presented in the literature. Substituent parameters such as the Hammet σ[7] (electron-withdrawing power) and the Hansch π[8] (lipo-

†Present address: IntelliCorp, Menlo Park, CA 94025.

0095-2338/85/1625-0064$01.50/0 © 1985 American Chemical Society

philicity) constants are the most widely used. The typical use of these in Hansch analysis[9] requires a set of compounds that are closely related, sharing a common skeleton and differing only in the nature and positioning of a few substituents. Descriptors such as molecular connectivity,[10] counts of self-avoiding paths,[11,12] and Moreau's autocorrelation of topological molecular structure (ATS) values[13] have the advantage that they can be computed easily from the connection table of a structure and can be applied to much more diverse sets of compounds. The drawback is that they are complex, "holistic" measures of the topology of the structure and, hence, may be hard to interpret even if they do correlate with activity. Hopfinger[14,15] has defined parameters from molecular shape analysis that relate to the space-filling characteristics and three-dimensional electrostatic potential of molecules. While these can yield geometric models of biological activity, their computation requires a detailed conformational analysis of each structure in the series. A reasonable compromise between generality, ease of interpretation, and ease of automatic perception can be found in descriptors relating to topological substructures, either as counts (e.g., number of ester groups in a structure) or as indicator variables, which record simple presence or absence (e.g., 1 for the presence of one or more ester groups and 0 for none). The focus of this paper is an especially simple type of substructure dubbed an *atom pair* that we have found to be very useful in exploring structure–activity relationships, particularly among sets of structurally diverse compounds.

In establishing correlations between descriptor values and biological activity, some mathematical methods (notably linear regression and discriminant plane analysis) require adjustable parameters, one for each descriptor, i.e., one for each dimension in the space. To yield a mathematically well-behaved solution, the number of adjustable parameters must be kept rather small compared to the total number of structures.[16] But often the number of descriptors that *might* be used greatly exceeds this number, and some sort of selection process (feature reduction)[17] is needed. The selected descriptors are often those that, after much trial and error, are found to yield the best mathematical model. Topliss[18] has shown that this approach can introduce bias and can lead to artifically high "goodness of fit" parameters in linear regression analyses. To avoid this problem, we have utilized heuristic techniques that do not require feature selection but that deal uniformly with all descriptors. Our methods do not include adjustable parameters and thus avoid mathematical singularities, but this approach raises questions concerning the independence of descriptors. These questions are considered under Discussion.

DEFINITION OF ATOM PAIRS

In defining substructures that might serve as standard, general features in structure activity studies, we set forth a number of desirable characteristics that these substructures should have. First, they should be intrinsic properties of a structure that are defined algorithmically and that embody no ad hoc assumptions about the types of functional groups or ring systems that "should" be important, as sometimes is the case with predefined libraries[19-21] of substructures created primarily for structure retrieval and categorization. Second, these substructures should be capable of describing long-range relationships between atoms that are not captured by, for example, augmented atom-centered fragments.[22] Third, these features should be generalizable to three-dimensional structures or to molecular representations in which atom properties such as estimated charge or polarizability replace the more usual atom properties of chemical type, degree of substitution, etc. Fourth, the perception of these substructures should be facile enough, and the representation of them compact enough, to allow their use even for very large sets (perhaps many thousands) of structures. The atom pair substructure satisfies all of these criteria.

Table I. Atom Pairs Contained in Cyclohexane-1,4-dione (1)[a]

count	atom pair	example
2	CX2-(2)-CX2	X–CH_2–CH_2–X
4	CX2-(2)-C·X3	X–CH_2–C(=X)–X
2	C·X3-(2)-O·X1	X–C(=O)–X
2	CX2-(3)-CX2	X–CH_2–X–CH_2–X
4	CX2-(3)-C·X3	X–CH_2–X–C(=X)–X
4	CX2-(3)-O·X1	X–CH_2–X=O
2	CX2-(4)-CX2	X–CH_2–X–X–CH_2–X
1	C·X3-(4)-C·X3	X=C(–X)–X–C(–X)=X
4	CX2-(4)-O·X1	X–CH_2–X–X=O
2	C·X3-(5)-O·X1	X=C(–X)–X–X=O
1	O·X1-(6)-O·X1	O=X–X–X–X=O

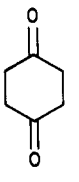
[a] A dot ("·") following an atom name indicates the presence of a bonding π electron. The suffix X*n* following an atom name indicates the presence of *n* non-hydrogen neighboring atoms.

We define an atom pair to be substructure composed of two non-hydrogen atoms and an interatomic separation:

⟨atom 1 description⟩–⟨separation⟩–⟨atom 2 description⟩

The two atoms in an atom pair need not be directly connected, and in fact, the ⟨separation⟩ tells how far apart they are, measured as the number of atoms in the shortest bond-by-bond path that contains both atoms 1 and 2. For instance, the methyl groups in ethane have a ⟨separation⟩ of 2, the methyl groups in propane have a ⟨separation⟩ of 3, in *n*-butane it is 4, and so on. The ⟨description⟩ of each atom tells its chemical type, the number of non-hydrogen atoms attached to it, and the number of bonding π electrons that it bears. Only the more common atom types C, O, N, S, F, Cl, Br, I, P, Si, B, Se, and As are represented explicitly in our current programs. The generic symbol "Y" is used for atoms of any other type. Defined in this way, atom pairs do not convey any stereochemical or conformational information. Thus, they can conveniently be computed from a constitutional representation of chemical structure (e.g., a connection table). The above definition applies to atom pairs as discussed in the remainder of this paper, though other definitions of the ⟨description⟩ and ⟨separation⟩ components of an atom pair are possible (see Discussion).

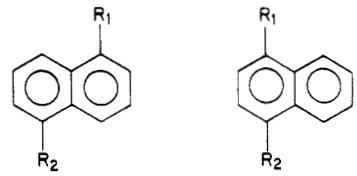
As an example of atom pairs, consider the compound cyclohexane-1,4-dione (**1**). It has eight non-hydrogen atoms



1

and thus 28 ($=n(n-1)/2$ for $n = 8$) atom pairs, 11 of which are distinct. Table I shows the types and counts of atom pairs for this structure. Clearly, no individual atom pair conveys very much structural information, but the *set* of atom pairs possessed by a molecule is a fairly characteristic structural property. These features are general enough that a significant number may be found in common among diverse structures yet specific enough that in the aggregate they can discriminate

even closely related topological isomers from one another. For example, among about 170 000 structures in a public database of structures (the SANSS master file—see below), fewer than 340 structures share identical sets of atom pairs with other, topologically distinct structures. Most of these correspondences are related to the pair of isomers **2** and **3**, which cannot be



2 **3**

distinguished by their sets of atom pairs.

Conceptually, atom pairs may be considered to be generalizations of other types of structural features that have been described in the literature. First, they are related to the self-avoiding paths described by Randic and co-workers.[11,12] That method treats only the topological skeleton of a molecule without accounting for differences in atom types or hybridization. Self-avoiding paths thus have interesting applications to sets of structures such as monocyclic monoterpenes[23] in which these differences are not important, but are limited when applied to more varied populations of compounds. Atom pairs augment the path-length information with a description of the terminal atoms and thus convey more chemically meaningful information. Second, atom pairs are extensions of the two-atom substructures used by Varkony and co-workers[24] as building blocks in the constructive perception of maximal subgraphs common to two or more structures. In that work, the description of the terminal atoms is similar, but the two-atom substructures refer only to directly connected pairs of atoms. Third, the autocorrelation of a topological molecular structure (ATS) defined by Moreau & Broto[13] considers both the nature of the terminal atoms and the separation between them, but by summing products of the electronegativities (or other atomic features) of the terminal atoms for all paths of a given length, the ATS method yields a single number for each path length. Atom pairs contain all the information needed to compute ATS descriptors, but the representation is more detailed (and hence has a higher dimensionality) than that for the ATS technique. Finally, atom pairs appear to be related to the matrixes of descriptor sites discussed by Raevskii and co-workers[25] and to the functional group pairs described by Morita and Oka,[26] though detailed definitions of these descriptors are not available to us.

PERCEPTION OF ATOM PAIRS

Computing the set of atom pairs from the connection table of a structure requires an algorithm for finding the length of the shortest path between each pair of non-hydrogen atoms in the structure. Dreyfus[27] has reviewed shortest path algorithms in the general context of minimizing the total "cost" of a path when each link in the path has a specified cost associated with it. Below is presented a simplified algorithm that deals only with path length [equivalent to a constant cost associated with each link (bond)]. From a connection table describing a structure of N atoms, this algorithm finds the length of the shortest path from a selected atom to each atom in the structure, where length is measured as the number of atoms in a path counting both end atoms (i.e., directly connected atoms have a shortest path length of 2).

The algorithm is as follows. Let N represent the number of non-hydrogen atoms in the structure, which are assumed to be numbered from 1 to N. Let j represent the selected, starting-point atom and let D be the output vector, with $D(k)$ equal to the length of the shortest path from atom j to atom k. The length of the "path" from atom j to itself, that is, $D(j)$, is defined to be 1. Let nnbrs(k) be an array of, or function that returns, the number of non-hydrogen attachments to atom k, and let nbr(k,m) be an array of, or function that returns, the atom number of the mth non-hydrogen attachment to atom k. Finally, let todo be an internal scratch array of dimension N and let b, t, atm, and nbrnum be internal variables. All variables, arrays, and functions are integers. The steps are

(1) set the N elements of D to 0
(2) set $D(j)$ to 1, todo(1) to j, b to 1, and t to 1
(3) until $b > t$ do
  (3.1) set atm to todo(b)
  (3.2) for nbrnum from 1 to nnbrs(atm) do
    (3.2.1) if D(nbr(atm,nbrnum)) = 0 do
      (3.2.1.1) set D(nbr(atm,nbrnum)) to D(atm) + 1
      (3.2.1.2) increment t
      (3.2.1.3) set todo(t) to nbr(atm,nbrnum)
  (3.3) increment b

Initially, all path lengths are set to "no path" (value 0), and atom j is given a path length of 1. Progressively larger "shells" of atoms around atom j are explored. The portion of the vector todo between index b ("bottom") and t ("top") stores the numbers of the atoms that are currently on the periphery, and thus whose neighbors must still be considered. As each atom atm is removed from the bottom (smaller D) end of this list, any of its neighbors that do not already have path lengths are assigned a path length one greater than that for atm and are placed on the top (larger D) end of the list. When no more atom numbers remain in this list ($b > t$), no further growth of shells can take place and the process is complete. Note that some of the values of D may remain 0; in a multicomponent structure (i.e., one with two or more disconnected pieces), there will be no paths from atom j to atoms in other components. Some improvement in efficiency can be obtained for single-component structures by inserting an additional terminating condition ($t = N$) after step 3.2.1.2.

The properties of atom type and number of non-hydrogen neighbors are easily extracted from the topological description of the structure. The number of bonding π electrons on an atom can be inferred from the bond types in the structure (for aromatic atoms, 1 π electron; for multiply bonded atoms, the sum of the bond orders minus the number of neighbors). Tautomer bonds and dative bonds present special problems that can be approached in a variety of ways. Our approach to tautomeric representations is to "denormalize"[28] them, reducing the tautomer bonds to a specific pattern of double and single bonds. Since there may be more than one such pattern, this introduces a small degree of arbitrariness in our current implementation, though common cases such as carboxylic acids, amides, and ureas are handled in a self-consistent manner. Dative bonds, such as the N–O bond in an N-oxide, are treated as double bonds, so that, for example, the nitrogen atom in a nitro group is treated formally as a pentavalent atom with two N–O double bonds. Formal atomic charges and unpaired electrons are ignored in the current implementation.

All three of the atom properties can be expressed as small integers (with the atom-type value referring to a predefined table) and can conveniently be "packed" into a single atom property through a formula such as

atom property value =
64 × (atom-type value) + 16 × (number of bonding π electrons) + number of non-hydrogen neighbors

For up to three π electrons and up to 15 non-hydrogen neighbors, this formula provides an unambiguous numerical represention for all three properties. Unusual cases can arise in which these limits are exceeded, and in these cases, a possibly ambiguous value is generated. If, as is the case in this study, the number of possible atom types is 15 or fewer

(recall that we represent all "unusual" atom types by a single, generic type), the magnitude of this value will not exceed 1023. Thus, it can occupy a 10-bit field of a computer word that describes an entire atom pair (see below).

The algorithm for computing atom pairs is a straightforward application of the above:

(1) compute the atom property value for each of the N non-hydrogen atoms in the structure
(2) for non-hydrogen atom j from 2 to N do
  (2.1) use the above path-length algorithm to compute D, the vector of path lengths from atom j to the other non-hydrogen atoms
  (2.2) for non-hydrogen atom k from 1 to $j - 1$ do
    (2.2.1) if $D(k)$ is non-zero (i.e., if there is a path from j to k) do
      (2.2.1.1) process the atom pair composed of $D(k)$ and the atom property values for atoms j and k

What is meant by "process the atom pair" above depends upon the application, but in our implementation the first step is always to "pack" the information into a single integer (or *key*) via

$$\text{atom-pair key} = \min[\text{ap}(j),\text{ap}(k)] + 1024\{\max[\text{ap}(j),\text{ap}(k)] + 1024D(k)\}$$

where ap(j) and ap(k) are the atom property values for atoms j and k. The minimum (min) and maximum (max) functions are used to provide a standard ("canonical") value for atom pairs in which ap(j) and ap(k) are unequal. For values of ap(i) and ap(j) up to 1023 and values D up to 2047, the above formula provides an unambiguous representation for the atom pair that can be stored as a one-word, positive integer on computers whose word size is at least 32 bits. Subsequent processing of this value (perhaps simply writing it to a file) is completely dependent upon the application.

SIMILARITY PROBE APPLICATION

The simplest problem in structure–activity studies is one of similarity among chemical structures. When a compound has been found that exhibits an interesting biological effect, often the first question is whether other closely related structures are readily available. This question can be addressed via substructure-search computer programs,[29] which allow the rapid retrieval of chemical structures containing selected partial structures. The problem, of course, is that initially one may have no indication of which substructures are important and must thus rely upon heuristics concerning the probable importance of unusual ring systems or functional groups. When there is no reason to emphasize any particular aspect of a structure, one may wish to seek other structures that most resemble it in a more holistic sense.

Randic and Wilkins[23,30] have presented a technique for measuring structural similarity based upon counts of self-avoiding paths in molecules. We have extended their method to account for differences in atom type and hybridization by substituting counts of atom pairs of different types for their counts of self-avoiding paths of different lengths. Also, we have chosen a distance measure (see below), which is somewhat simpler to compute yet which appears to yield comparable results.

Given a set of numerical descriptors for each of two structures, one can define a variety of measures of the degree to which the descriptors of one structure parallel those of the other. In a number of statistical methods, these measures take the mathematical form of a generalized distance, or *metric*, between the two structures, viewed as points in a high-dimensional space. The so-called *city block distance*[31] is especially simple to compute; it is the sum of the absolute differences between descriptor values for the two structures.

Acetone atom pairs	Shared pairs	Isobutylene atom pairs
CX — (2) — C.X3	< — >	CX — (2) — C.X3
CX — (2) — C.X3	< — >	CX — (2) — C.X3
C.X3 — (2) — O.X1	x	C.X3 — (2) — C.X1
CX — (3) — CX	< — >	CX — (3) — CX
CX — (3) — O.X1	x	CX — (3) — C.X1
CX — (3) — O.X1	x	CX — (3) — C.X1
6 atom pairs	3 shared pairs	6 atom pairs

Figure 1. Illustration of the calculation of the similarity measure between acetone (**2**) and isobutylene (**3**). $S = (2 \times 3)/(6 + 6) = 6/12 = 0.5$.

Taking the descriptors to be counts of atom pairs of distinct types, the following "atom-pair distance", $D(s,t)$, between structures s and t can be defined:

$$D(s,t) = \sum_{\substack{\text{distinct types } i \text{ of} \\ \text{atom pairs}}} \text{ABS}[n(i,s) - n(i,t)] \qquad (1)$$

where $n(i,j)$ is the number of atom pairs of type i possessed by structure j and where ABS is the absolute value function. If the two structures have precisely the same set of atom pairs, the value of this distance is 0. If they share no atom pairs in common, the distance takes on its maximum value of $d(s) + d(t)$, where $d(i)$ is the "length" of structure i, that is, the total number of atom pairs that are contained in structure i. [For a single-fragment structure, $d(i)$ has the value $N(N - 1)/2$, with N representing the number of non-hydrogen atoms in the structure.]

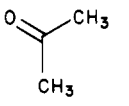
A unitless similarity measure, $S(s,t)$ varying from 0 (dissimilar) to 1 (completely similar), can then be defined as

$$S(s,t) = 1 - D(s,t)/[d(s) + d(t)] \qquad (2)$$

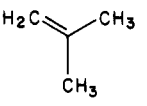
An alternative formula, equivalent to the above and sometimes easier to compute, is

$$S(s,t) = \frac{2}{d(s) + d(t)} \sum_{\substack{\text{distinct types } i \text{ of} \\ \text{atom pairs}}} \text{MIN}[n(i,s),n(i,t)] \qquad (3)$$

where MIN represents the standard minimum function. So defined, $S(s,t)$ may be interpreted as the fraction of atom pairs in s and t together that are shared between s and t. For example, consider the similarity measure S (eq 2 or 3) between acetone (**4**) and isobutylene (**5**). As shown in Figure 1, each



4



5

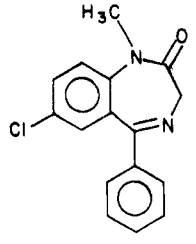
structure has six atom pairs. Three of these are shared, so the numerator of the similarity measure is twice 3 or 6. The denominator is the sum of the numbers of atom pairs in the structures, that is, 6 + 6 or 12. The similarity score between acetone and isobutylene is thus 6/12 or 0.5.

Using the algorithm described under Perception of Atom Pairs, we have written a set of BCPL programs for the DEC TOPS-10 system that can compute, record, and collate the atom pairs occurring among a large set of structures. The input consists of a list of connection tables, and the output consists of (1) a list of all distinct types of atom pairs appearing in any of the structures, sorted by increasing value of the packed integer describing the atom pair, (2) a corresponding list of index values that are assigned sequentially to the distinct types of atom pairs as they are encountered during the pro-

cessing of the connection tables, and (3) for each structure, a list of doubles (index, count) enumerating by type the atom pairs occurring in that structure. At a rate of about five average-sized structures (say, 22 non-hydrogen atoms) per s of CPU time (KL-10 processor), it is practical to apply the programs to databases of up to a few hundred thousand connection tables, and we have done so both for our local database ("CL file"—about 200 000 structures) and for the publicly available SANSS master file,[32] which contains connection tables for about 170 000 compounds referenced in the Chemical Information System (CIS)[33,34] developed by the National Institutes of Health and the Environmental Protection Agency. Processing a file of this size requires about 10 CPU h and requires a substantial amount of disk storage [on the order of 400 8-bit bytes per average structure when the (index, count) doubles are packed into single words] for the output files.

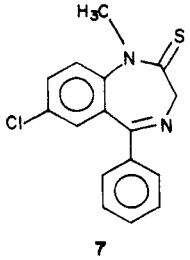
Another BCPL program called SIMPRB (for "similarity probe") has been written that accepts the connection table of a new structure, computes the corresponding set of atom pairs, and then sequentially scans the above output files evaluating the similarity measure between the new structure and each of the database structures. The user may specify NSTRUCS, the maximum number of structures that the program will select, and SIMMIN, the minimum value of similarity that a database compound must show to be selected. Within these limits, the program will select from the database the structures that are most similar (highest similarity value) to the new structure and will create an output file of compound numbers and similarity values, sorted by decreasing similarity, for the selected compounds. The program has been highly tuned and critical sections have been written in assembler language. At an average search rate of about 1000 database structures/s, atom pair files from a few hundred thousand structures can be processed in several minutes of CPU time.

As an illustration of this technique, we have carried out a similarity probe of the SANSS master file for diazepam (**6**),

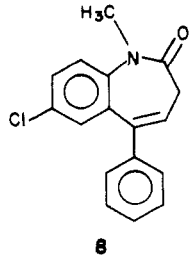


6

a widely prescribed antianxiety drug with some sedative and anticonvulsant properties. The 100 most similar compounds give similarity values with **6** ranging from $S = 1.000$ (diazepam itself) to $S = 0.660$. Of these, 78 are easily recognizable as analogues of the probe compound. They range from structures such as **7** and **8** ($S = 0.900$ for both), which differ

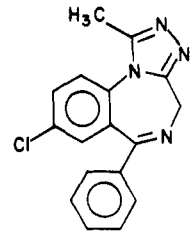


7

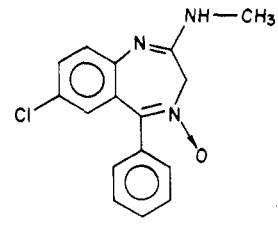


8

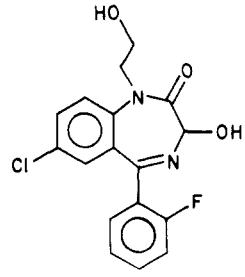
from **6** by only one atom, through structures such as **9** (alprazolam, $S = 0.790$) and **10** (chlordiazepoxide, $S = 0.740$), which possess a few different atoms and bonds, to more heavily altered analogues such as **11** ($S = 0.661$) and **12** ($S = 0.660$).



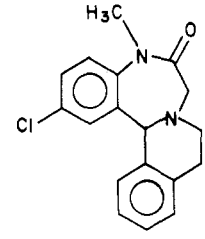
9



10

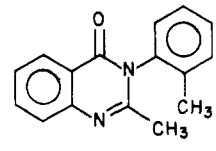


11

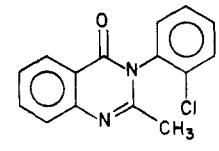


12

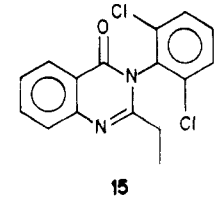
Interspersed with these analogues are 22 other structures that bear resemblance to **6** but that are not usually thought of as benzodiazepine analogues. An unusual number of these show psychotropic activity or analogous activities in animal models. Ten of them are analogues of methaqualone (**13**) (an anti-



13

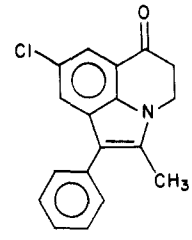


14

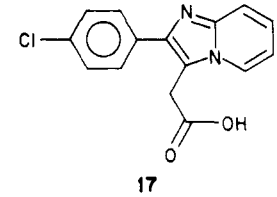


15

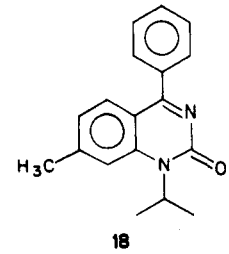
convulsant and hypnotic) such as **14** (mecloqualone, $S = 0.687$) and **15** (chloroqualone, $S = 0.665$). Methaqualone itself is the 106th structure in the list, with $S = 0.654$. The remaining 12 structures include **16** ($S = 0.705$), for which



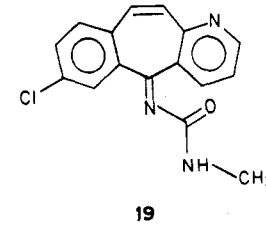
16



17



18



19

sedative properties have been observed,[35] **17** ($S = 0.679$), for which anticonvulsant activity has been noted,[36] the antiinflammatory and analgesic **18** (proquazone, $S = 0.665$), which shows CNS-depressant effects at high doses,[37] and **19** ($S = 0.660$), which is closely related to compounds with central nervous system depressant activity.[38]

It is perhaps not surprising to find this degree of clustering of psychotropic activity, even among non-benzodiazepines, around diazepam; that is consistent with the expectation that similar structures will frequently show similar properties. But

by providing a quantitative and holistic similarity measure that is not biased by concepts of functional groups and ring systems, the similarity probe can complement substructure-search techniques and can reveal relationships between classes of compounds that might otherwise be missed. The SIMPRB program has become an important in-house tool for selecting compounds for testing from the CL file, especially useful in the early stages of a project when only one, or perhaps a few, active compound(s) has (have) been identified. Extensions of this concept have been and are being applied to the computer-assisted recognition of classes (i.e., clusters) of compounds in our database of structures and to the evaluation of the uniqueness of proposed or newly synthesized compounds.

TREND VECTOR ANALYSIS APPLICATION

The similarity probe described above is especially useful when only a few structures are known that show a particular biological activity. When larger numbers of structures are available for which the biological response has been measured, one can begin to ask about molecular features that distinguish the more active compounds from the less active, or inactive, ones. In the best circumstances, answers to such questions can shed light on the fundamental biological action of the compounds being analyzed. A model need not embody such deep understanding, though, to be useful. A statistical model, one that estimates probabilities of activity rather than activity directly, may perform well on populations of structures. When such a model is sufficiently general, it can aid in the selection of compounds for synthesis and testing and can "on the average" steer one in fruitful new directions.

Cramer and co-workers[39] have described such a technique, which they call substructural analysis, and Hodes and co-workers[40-42] have refined and extended the technique, which they refer to as a statistical-heuristic method, to such an extent that it is now used routinely in prescreening compounds that are candidates for acquisition and testing in the National Cancer Institute's screening program.[43] Tinker[44] has also described the application of a very similar method to the estimation of mutagenicity based upon chemical structure.

Each of these methods makes use of a library of substructural features, either predefined or extracted from the structures being analyzed. This library defines the "space" in which the analysis is carried out. A substructure is assigned a probabilistic score on the basis of the number of active compounds containing the substructure compared to the total number (either tested or in an entire structure database) of compounds containing the substructure. These scores may be viewed collectively as a vector in the n-dimensional space. Structures are then assigned scores proportional to the sum of the scores of the substructures that they contain. This corresponds mathematically to taking the dot product (projection) of the score vector with the vector representing the structure. In both methods, these dot products have been found to be of modest but significant predictive value in estimating the activities of new compounds.

We have followed the same general principles in this work but with three major differences. First, we have chosen a form for the substructure scores that is especially easy to compute and that is based on a heuristic rather than a probabilistic model. Second, the features used in our method are all the atom pairs that occur among the structures being analyzed. Typically, the number of atom pairs greatly exceeds the number of structures, and questions may arise concerning the independence of the atom pairs. The Discussion addresses these questions. Third, our approach includes an explicit evaluation of the significance of a model based upon a comparison with other models that would result from randomly scrambled activities.

In considering fast, heuristic methods for describing a "cloud" of compounds and their measured activities in n-dimensional space, we considered the analogy of describing a cloud of charged particles in three-dimensional space. The dipole moment is a simple, first-order measure of the degree to which such a charge distribution is "lopsided" with positive charges preferentially located at one side of the cloud and negative charges at the other. Thus, we have experimented with the analogous vector, which might be called an "activity moment", computed in n-dimensional space with the measured activity values treated mathematically as "charges". We use the term *trend vector* rather than activity moment for this vector to emphasize the relationship of this vector to trends in the n-dimensional distribution of activities.

If $\mathbf{S}_i$ is the n-dimensional vector (i.e., list of descriptor values, see below) representing structure i, a_i is a numerical measure of the activity of structure i, N is the number of structures comprising the n-dimensional cloud, and A is the average activity for the N structures, then the formula for the trend vector $\mathbf{T}$ is the vector sum

$$\mathbf{T} = (1/N) \sum_{\text{structures } i}^{N} (a_i - A)\mathbf{S}_i \qquad (4)$$

We note that the trend vector is insensitive to the addition of a constant scalar to every activity value and to the addition of a constant vector to every S_i.

Sometimes only qualitative data (i.e., active/inactive) are available for a particular test. In this case, activity "values" of 0 (inactive) and 1 (active) are used in eq 4. The trend vector is then collinear with the vector interconnecting the center of gravity (in the n-dimensional space) of points for inactive structures with the center of gravity for the actives. In this form, the trend vector is related to the binary classifier described by Varmuza,[45] which uses the symmetry plane between the two centers of gravity to predict whether a given new structure is more or less likely to be active. We emphasize, however, that the current method applies to quantitative measures of activity as well as qualitative ones and is thus more general.

As described so far, the trend vector could be computed for any set of descriptors. In our standard applications, though, the dimensions of the space represent unique atom pairs. Thus, if $\mathbf{S}$ is the vector for a structure, then $S(j)$ describes the presence of atom pairs of type j. It is possible to take $S(j)$ as the count of atom pairs of type j, but we have found that $\mathbf{T}$ (eq 4) is less dominated by very large structures if $S(j)$ is taken as an indicator variable. If the structure contains at least one atom pair of type j, then $S(j) = 1$; otherwise, $S(j) = 0$. In this case, a component $T(j)$ of the trend vector can be easily interpreted as

$$T(j) = f(j)[A(j) - A] \qquad (5)$$

where $f(j)$ is the fraction of structures in the set containing atom pairs of type j, where $A(j)$ is the average activity for that fraction, and where A is the average activity for all structures in the set. If atom pairs of type j occur in only a few structures, $T(j)$ will be low because $f(j)$ will be low, while, if they occur in almost all structures, $T(j)$ will be low because $A(j)$ cannot differ much from A. Only if atom pairs of type j occur in an intermediate number of structures showing substantially greater (less) than average activity can $T(j)$ be very positive (negative). Thus, $T(j)$ can be viewed as a "score" for atom pairs of type j indicating the degree to which their presence or absence correlates with activity.

The length of this vector (that is, the square root of the sum of the squares of the scores) is important in deciding whether the trend vector is significant or simply the result of random differences between structures and between activities. A

Monte Carlo technique is used in which the given activity values are randomly reassigned to the "wrong" structures, and a spurious trend vector is computed by the above scheme. If the length of the real trend vector is significantly larger than the average of the lengths of, say, 10 such spurious vectors (where significance is measured by the standard deviation in the lengths of the several spurious vectors), the real vector is considered to express meaningful information.

If a significant trend vector is obtained for a given series, it can be used in two ways. First, examination of the atom pairs with the highest scores (or most highly negative scores) may help one to pinpoint portions of the structures that show the greatest relationship with high (or low) activity values. Second, the trend vector may be used to rank structures according to how well they capture the "good" atom pairs while avoiding the "bad" ones. Among the structures used to construct the trend vector, such ranks may be compared to the original activity values, and the (presumably high) correlation between rank and activity can be evaluated. For new structures, the rank may be taken as a rough prediction of activity.

The rank R of a structure is computed as the dot product of the trend vector $\mathbf{T}$ with the vector $\mathbf{S}$ representing the structure:

$$R = \mathbf{S}\cdot\mathbf{T} = \sum_{\substack{\text{distinct}\\ \text{types } i \text{ of}\\ \text{atom pairs}}} S(i)T(i) \qquad (6)$$

As in the definition of the trend vector, we usually restrict $S(i)$ to values of 0 and 1: If a particular atom pair occurs more than once in the structure, it is only counted once. When a new structure is ranked, atom pairs may be encountered that never appeared in the original series. Such pairs are given a score of 0, and a record is kept of the number of such "undefined" pairs. The rank of a new structure is not considered to be meaningful if this count becomes too high (say, greater than 5% of all atom pairs in the structure).

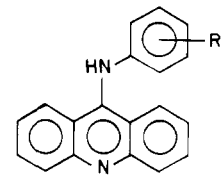
Sometimes it is convenient to express the rank values on a relative scale via the transformation

$$r = (\text{DPMAX} - R)/(\text{DPMAX} - \text{DPMIN}) \qquad (7)$$

where DPMAX is the sum of all positive components in $\mathbf{T}$ and where DPMIN is the sum of all negative components. So defined, r can vary from 1.0 for the "best possible" structure (an imaginary entity containing all of the positively scored atom pairs and none of the negatively scored ones) to 0.0 for the "worst possible" structure.

A suite of BCPL and FORTRAN programs has been written for the DEC-10 computer to compute the trend vector from given activities, to apply the Monte Carlo test for significance, and, finally, to rank structures either in the training set, in a test set, or (using the atom-pair file described in the prior section) in an entire database of structures. In the latter two cases, the percentage of undefined atom pairs in each structure is also recorded. Once connection tables have been defined and entered into the computer along with activity values, a typical analysis of a training set containing several hundred structures requires perhaps 20 min of computer time and an afternoon or so of elapsed time.

As an example of the use of this method, we explore here a literature study in which pattern recognition, in particular discriminant plane analysis, yielded a model of antitumor activity among 9-anilinoacridines.[46] To develop the model, Henry and co-workers used a set of 213 compounds with the generic structural formula **20**, selected at random from a more extensive list compiled by Denny and co-workers.[47] Activity in this study was measured as the ability of a compound to prolong the life of mice inoculated intraperitoneally with L1210 leukemia cells, specifically as the maximal percent increase



20

in life span at a given level of toxicity. To be classed as active in this two-category analysis, a compound must have caused at least a 35% increase in lifespan. Of special interest in this study was the evaluation of the model on a prediction set of 50 compounds that had not been included in the training set. The 82% rate of correct predictions is impressive; we find that a χ^2 test[48] of the relationship between observed and predicted activity indicates an extremely slim chance (on the order of 1 in 10000) that the correlation could have happened "by accident". One reason for choosing this series was to determine whether trend vector analysis could give a comparable predictive ability.

Given the connection tables for the 213 training set structures, the perception of atom pairs yielded a total of 95 364 pairs, 1498 of which were distinct. Thus, the dimensionality of the space in this case was 1498. A trend vector was computed from the corresponding qualitative activity values (1 for active and 0 for inactive, relative to the threshold of a 35% increase in lifespan). As an example of this computation, consider the highest scoring atom pair

$$\langle \text{O}\cdot\text{X1}\rangle\text{–}(12)\text{–}\langle \text{N}\cdot\text{X2}\rangle$$

which occurred in 158 of the 213 structures. The ratio of these, 0.742, was the fraction of structures containing the atom pair and corresponded to $f(j)$ in eq 5. Of the 158 structures containing the atom pair, 133 were active. Because the numerical value of 1 represented activity and 0 inactivity, the average activity for structures containing the atom pair was just the active/total ratio 133/158 = 0.842, which corresponded to $A(j)$ in eq 5. Similarly, the average activity A in eq 5 was just the active/total ratio for the entire set of structures, that is, 153/213 or 0.718. The final score obtained via eq 5 was thus 0.742(0.842–0.718), or 0.092. Scores for the atom pairs in this series ranged from this value down to –0.052.

The length of the 1498-dimensional trend vector was 0.66 (arbitrary activity units). Ten spurious vectors were constructed with randomly scrambled activities, and these yielded lengths from 0.21 to 0.37, with a mean of 0.27 and a standard deviation of 0.05. Thus, the real trend vector was over twice as long as one would expect randomly, and if it is assumed that the distribution of lengths arising from the randomizations was roughly normal, there was a very small chance indeed that the observed trend vector could have arisen "by accident". By our usual criteria, we judged this to be a very strong and significant trend vector.

Ranks for the structures in the training set ranged from R = –0.16 to R = 6.49 (arbitrary activity units) or, on the relative scale, from r = 0.14 to r = 0.66. Unlike discriminant plane analysis, our method does not define an inherent "decision value" of rank, which allows the categorization of structures into predicted-active and predicted-inactive classes. For purposes of comparison to the previous work,[46] we determined the value of relative rank (r = 0.31) that gave the highest percentage of correct prediction. Of the 153 active compounds, 136 (or 89%) had relative ranks at or above this threshold while, of the 60 inactive compounds, 36 (or 60%) had relative ranks below, for an overall correct percentage of 81%. This was substantially lower than the 94% correct classification reported from the discriminant plane analysis. It should be emphasized, though, that in the trend vector analysis no parameter except the decision value itself was adjusted to op-

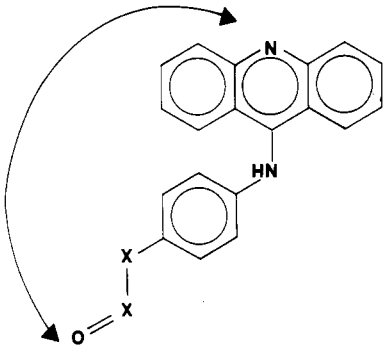


Figure 2. Highest scoring atom pair [N·X2–(12)–O·X1] from analysis of 9-anilinoacridines illustrated in its most common structural context.

timize this separation; the discriminant plane analysis allowed 18 additional parameters, which were mathematically varied to increase the separation.

Of greater interest was the predictive ability of the model when it was applied to compounds outside of the training set. Ranks were computed for the 50 compounds in the test set[46] by using the trend vector derived from the training set. Using the above threshold of 0.31 (relative rank) with higher values as predicted active and with lower values as predicted inactive, we found that 29 (83%) of 35 actives and 13 (87%) of 15 inactives were correctly predicted, for an overall correct rate of 84%. These numbers are quite comparable to those obtained from discriminant plane analysis (30 of 35 actives correct, 11 of 15 inactives correct, and 82% correct overall). In our opinion, the best measure of the information content of a model is its ability to predict beyond what was given at the outset; on this basis, it appears that, at least for this example, trend vector analysis is as informative as discriminant plane analysis.

One advantage of the trend vector model is that the major contributing factors can usually be easily interpreted. For example, as mentioned above the highest scoring atom pair in the example analysis is

$$\langle \text{O·X1} \rangle\text{–(12)–}\langle \text{N·X2} \rangle$$

Examination of the structures in the training set shows that essentially every occurrence of this atom pair is in the context illustrated by Figure 2. The interpretation of this high score is that the probability of a structure showing activity is enhanced significantly by the presence of a doubly bonded oxygen three bonds removed from the 1′-position of the 9-anilinoacridine skeleton (**20**). In most cases, this oxygen is part of a sulfonamide moiety attached to the 1′-position, but examples of several types of carbonyl groups (urea, amide, ketone, acid) are also represented among the active compounds. Among the 50 compounds of the prediction set, 28 of 35 active structures possess this atom pair as compared to only 1 of the 15 inactive structures; almost all of the predictive ability of the trend vector model would be captured in a very simple model that predicts activity solely on the basis of this atom pair. In contrast, the most important descriptor[49] from the previously reported discriminant plane analysis of these compounds is the average number of paths per atom, a quantity that is generally related to the degree of branching and cyclicity of a structure but that is difficult to estimate or manipulate as a mental concept.

There are several other atom pairs that are nearly as important as the one examined above, including ones that correlate with lack of activity (bad atom pairs). Their presence in this example model does little to improve the overall predictive ability compared to the simple "one-pair model" above, but they do serve to make the model more general; the actual model does not depend upon the presence or absence of a specific atom pair but incorporates contributions from all atom pairs in a structure. It is thus a more holistic model than the "one-pair" interpretation would suggest. In our experience with this technique it is unusual for the major predictive power of a trend vector model to be concentrated in a single atom pair. More commonly, the trend vector expresses a summation of influences from a number of modestly correlated atom pairs.

Due to its simplicity, a trend vector model can be applied to large populations of compounds to select the highest ranked members. If a model has been developed over a diverse training set (not the case with this example), it may be reasonable to apply the model to an entire structure database such as the CL file or the SANSS master file discussed above in the context of the similarity probe. A BCPL program has been written for the DEC-10 that can accomplish this in a matter of several CPU minutes, making use of the same atom pair file described under Similarity Probe Application. The user specifies thresholds for the relative rank and for the percentage of atom pairs covered by the trend vector and obtains a file containing compound numbers, relative ranks, and percentages for all structures in the database that exceed the thresholds. We have developed similar programs for ranking computer-generated families of isomers and for ranking single structures input by the user.

We have found these tools to be quite valuable, especially in the selection of compounds for testing from our CL file; samples of compound are available for a substantial percentage of the structures stored in the database. For some biological tests, compounds have historically been selected at random from the file, and a record of the "hit rate" for random screening is available. Our experience is that compounds selected by a trend vector model usually show a severalfold increase in this hit rate (say, 2% increasing to 10%). By itself, this increase is not necessarily impressive; chemists and biologists often achieve higher hit rates when selecting analogues of known actives by using substructure-search techniques. But typically, the active structures found from trend vector ranking are much more diverse than those resulting from analogue searches, and thus, the probability is increased for the discovery of unexpected new classes of active compounds.

DISCUSSION

Implicit in the techniques described here is a particular philosophy concerning the representation of chemical structures in a high-dimensional space. The variety and diversity of chemical structures is great, and we feel that any generally useful representation must have a very high inherent dimensionality. For specific families of structures, lower dimensional representations may be sufficient, but these are seldom unique, and the selection between alternate sets of descriptors can consume a major fraction of the effort in a structure–activity study. If automatic methods are to be developed for correlating large numbers of diverse structures and their respective activities, it seems most appropriate that these methods deal directly with a standard, high-dimensional representation, such as the space of atom pairs.

When the dimensionality (i.e., number of features) exceeds the number of structures in an analysis, some redundancy among features is implied. A set of n points can at best span an $(n - 1)$-dimensional subspace; at most, $n - 1$ features, or linear combinations of features, can be linearly independent in describing structures corresponding to those points. Any additional features are necessarily expressible as linear combinations of the linearly independent ones. The methods described in this paper do not use adjustable parameters, so some of the mathematical difficulties associated with an excess of descriptors are avoided. But our methods do treat each type of atom pair as an independent dimension, contributing additively to the distance between structures (similarity probe)

or to the rank of a structure (trend vector analysis). Our experience indicates that, despite this apparent inconsistency, the results from both methods are routinely useful. Perhaps any redundancies that do occur are more or less evenly distributed over the dimensions of the space, due to the fairly constant amount of structural information conveyed by each atom pair. This would lead to systematic errors, primarily a consistent increase in the lengths of the trend vectors (including spurious vectors in the randomization process), which should not greatly alter similarity values, relative trend vector ranks, or the Monte Carlo evaluation of the significance of the trend vector. Other rationalizations can be constructed; but in essence, we view our methods as empirical ones that should be evaluated on their practical usefulness.

Beyond the question of redundancy, atom pairs are limited in their treatment of atoms that are qualitatively similar but that have different descriptions. A methyl carbon is viewed as totally different from a methylene carbon, just as a carbonyl oxygen is different from a boron atom; chemically, the former pair ought to be treated as "more similar" than the latter, but the atom pair descriptions do not allow this. It is possible to generalize the atom pairs by eliminating some portion of the description of each atom (say, the count of non-hydrogen neighbors, thus making all sp^3 carbons the same) or by grouping atoms of differing chemical type (say, equating fluorine, chlorine, bromine, and iodine atoms). Alternatively, the ⟨description⟩ of each atom in a pair may be generalized to refer not to atom type and hybridization but to estimated atomic properties such as polarizability[50] or local atomic charge.[51] Sometimes, these generalizations improve the significance of a trend vector model, but they do not adequately express the concept of "partial identity" between different types of atom pairs.

A further limitation of atom pairs is their topological nature. To the extent that a particular biological activity depends upon molecular geometry, even the best correlations between atom pairs and activity are limited by the crude relationship between through-space distance and shortest path separation. Some aspects of the three-dimensional shape of a structure can be captured by atom pairs when the ⟨separation⟩ value is defined as a through-space distance or as an integer representing a range of through-space distances. Experimentation with this representation has yielded promising preliminary results.[52]

The similarity probe and trend vector analysis programs have been used in about 2 dozen studies at Lederle over the past few years. In those cases for which statistics are available on the testing of randomly selected compounds, we have found with few exceptions that compounds selected by these methods are substantially enriched in activity, with the new active structures showing good diversity and novelty compared to structures in the training set. These techniques are becoming a standard feature of our discovery process. Much of our current effort is toward placing these tools in the hands of research chemists and biologists.

REFERENCES AND NOTES

(1) Blankley, C. J. "Introduction: A Review of QSAR Methodology". *Med. Chem. (Academic)* **1983**, *19*.
(2) Varmuza, K. "Pattern Recognition in Chemistry". *Lect. Notes Chem.* **1980**, *21*.
(3) Jurs, P. C.; Isenhour, T. L. "Chemical Applications of Pattern Recognition"; Wiley: New York, 1975.
(4) Kirschner, G. L.; Kowalski, B. R. "The Application of Pattern Recognition to Drug Design". *Med. Chem. (Academic)* **1979**, *8*.
(5) See, e.g., Tabachnick, B. G.; Fidell, L. S. "Using Multivariate Statistics"; Harper and Row: Philadelphia, 1983.
(6) Dunn, W. J., III; Wold, S. "A Structure–Carcinogenicity Study of 4-Nitroquinoline 1-Oxides Using the SIMCA Method of Pattern Recognition". *J. Med. Chem.* **1978**, *21*, 1001–1007.
(7) Hammett, L. P. "Physical Organic Chemistry", 2nd ed.; McGraw-Hill: New York, 1970; pp 355–362.
(8) Iwasa, J.; Fujita, T.; Hansch, C. "Substituent Constants for Aliphatic Functions Obtained from Partition Coefficients". *J. Med. Chem.* **1965**, *8*, 150–153.
(9) Hansch, C. "Quantitative Approaches to Pharmacological Structure-Activity Relationships". In "Structure Activity Relationships"; Cavallito, C. J., Ed.; Pergamon Press: New York, 1973; Vol. 1, Chapter 3.
(10) Kier, L. B.; Hall, L. H. "Molecular Connectivity in Chemistry and Drug Research". *Med. Chem.* **1976**, *14*.
(11) Randic, M.; Brissey, G. M.; Spencer, R. B.; Wilkins, C. L. "Search for All Self-Avoiding Paths for Molecular Graphs". *Comput. Chem.* **1979**, *3*, 5–13.
(12) Randic, M.; Brissey, G. M.; Spencer, R. B.; Wilkins, C. L. "Use of Self-Avoiding Paths for Characterization of Molecular Graphs with Multiple Bonds". *Comput. Chem.* **1980**, *4*, 27–43.
(13) Moreau, G.; Broto, P. "The Autocorrelation of a Topological Structure: A New Molecular Descriptor". *Nouv. J. Chim.* **1980**, *4*, 359–360.
(14) Hopfinger, A. J. "A QSAR Investigation of Dihydrofolate Reductase Inhibition by Baker Triazimes Based upon Molecular Shape Analysis". *J. Am. Chem. Soc.* **1980**, *102*, 7196–7206.
(15) Hopfinger, A. J. "Theory and Application of Molecular Potential Energy Fields in Molecular Shape Analysis: A Quantitative Structure–Activity Relationship of 2,4-Diamino-5-benzylpyrimidines as Dihydrofolate Reductase Inhibitors". *J. Med. Chem.* **1983**, *26*, 990–996.
(16) Gray, N. A. B. "Constraints on 'Learning Machine' Classification Methods". *Anal. Chem.* **1976**, *48*, 2265–2268.
(17) See, e.g., Varmuza, K. "Pattern Recognition in Chemistry". *Lect. Notes Chem.* **1980**, *21*, 106–113.
(18) Topliss, J. G.; Edwards, R. P. "Chance Factors in Studies of Quantitative Structure–Activity Relationships". *J. Med. Chem.* **1979**, *22*, 1238–1244.
(19) Lynch, M. F.; Harrison, J. M.; Town, W. G.; Ash, J. E. "Computer Handling of Chemical Structural Information"; American Elsevier: New York, 1972; pp 67–95.
(20) Craig, P. N.; Ebert, H. M. "Eleven Years of Structure Retrieval Using the SK&F Fragment Codes". *J. Chem. Doc.* **1969**, *9*, 141–146.
(21) Milne, M.; Hazard, G. F. "National Cancer Institute's Drug Research and Development Chemical Information System: Design of, and User Experience in the Interactive Inquiry System". "Abstracts of Papers", 169th National Meeting of the American Chemical Society, Philadelphia, PA, Apr. 1975; American Chemical Society: Washington, DC, 1975; CHLT 14.
(22) Adamson, G. W.; Lynch, M. F.; Town, W. G. "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-based File. Part II. Atom-Centered Fragments". *J. Chem. Soc. C* **1971**, 3702–3706.
(23) Randic, M.; Wilkins, C. L. "Graph Theoretical Approach to Recognition of Structural Similarity in Molecules". *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 31–37.
(24) Varkony, T. H.; Shiloach, Y.; Smith, D. H. "Computer-Assisted Examination of Chemical Compounds for Structural Similarities". *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 104–111.
(25) Raevskii, O. A.; Avidon, V. V.; Novikov, V. P. "Use of a Unified Scale of Donor-Acceptor Interactions for the Analysis of the Similarity of Biologically Active Compounds". *Pharm. Chem. J. (Engl. Transl.)* **1982**, *16*, 633–636 (transl. from *Khim.-Farm. Zh.* **1982**, *16*, 968–971).
(26) Morita, K.; Oka, Y. "Synthetic Drugs Containing Nitrogen, with Special Reference to Classification according to Functional Group Pair". *Kagaku, Zokan (Kyoto)* **1979**, *79*, 141–175.
(27) Dreyfus, S. E. "An Appraisal of Some Shortest-Path Algorithms". *Oper. Res.* **1969**, *17*, 395–412.
(28) For a discussion of the normalized representation of tautomeric structures, see Mockus, J.; Stobaugh, R. E. "The Chemical Abstracts Service Chemical Registry System. VII. Tautomerism and Alternating Bonds". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 18–22.
(29) See, e.g., "Computer Representation and Manipulation of Chemical Information"; Wipke, W. T.; Heller, S. R.; Feldman, R. J.; Hyde, E., Eds.; Wiley: New York, 1974.
(30) Wilkins, C. L.; Randic, M.; Schuster, S. M.; Markin, R. S.; Steiner, S.; Dorgan, L. "A Graph-Theoretic Approach to Quantitative Structure-Activity/Reactivity Studies". *Anal. Chim. Acta* **1981**, *133*, 637–645.
(31) Varmuza, K. "Pattern Recognition in Chemistry". *Lect. Notes Chem.* **1980**, *21*, 25.
(32) The SANSS file of connection tables was obtained through the National Technical Information Service (NTIS), U.S. Department of Commerce, Springfield, VA 22161.
(33) Milne, G. W. A.; Fisk, C. L.; Heller, S. R.; Potenzone, R. "Environmental Uses of the NIH-EPA Chemical Information System". *Science (Washington, D.C.)* **1982**, *215*, 371–375.
(34) Heller, S. R.; Milne, G. W. A.; Feldmann, R. J. "A Computer-Based Chemical Information System". *Science (Washington, D.C.)* **1977**, *195*, 253–259.
(35) Gatta, F.; Tomassetti, M.; Zaccari, V.; Landri-Vittory, R. "Sur quelques derives de 1a-phenyl-1-methyl-2-dihydro-4,5-6*H*-pyrrolo-(3,2,1-*ij*)-quinoleinone-6 (2e memoire). Synthese de pyrido-(3,2,1-*jk*)-benzodiazepines-1,4". *Eur. J. Med. Chem.–Chim. Ther.* **1974**, *9*, 133–135.
(36) Almirante, L.; Mugnaini, A.; Rugarli, P.; Gamba, A.; Zefelippo, E.; DeToma, N.; Murmann, W. "Derivatives of Imidazole. III. Synthesis and Pharmacological Activities of Nitriles, Amides, and Carboxylic

Acid Derivatives of Imidazole[1,2-*a*]pyridine". *J. Med. Chem.* **1969**, *12*, 122–126.
(37) From the abstract [*Chemical Abstracts* CA90-66721(9)] of Tsurumi, K.; Nakano, M.; Hasegawa, J.; Fujimura, H. "General Pharmacological Actions of 1-Isopropyl-4-phenyl-7-methyl-2(1*H*)-quinazolinone (Proquazone)". *Oyo Yakuri* **1978**, *16*, 115–123.
(38) Van der Stelt, C.; Hofman, P. S.; Funcke, A. B. H.; Timmerman, H. "5*H*-Benzo[4,5]cyclohepta[1,2-*b*]pyridine and 11*H*-Benzo[5,6]cyclohepta[1,2-*c*]pyridine. III. Synthesis and Pharmacological Properties of Some Derivatives of 5*H*-Benzo[4,5]cyclohepta[1,2-*b*]pyridine and of 11*H*-Benzo[5,6]cyclohepta[1,2-*c*]pyridine". *Arzneim.-Forsch.* **1972**, *22*, 133–137.
(39) Cramer, R. D., III; Redl, G.; Berkoff, C. E. "Substructure Analysis. A Novel Approach to the Problem of Drug Design". *J. Med. Chem.* **1974**, *17*, 533–535.
(40) Hodes, L.; Hazard, G. F.; Geran, R. I.; Richman, S. "A Statistical-Heuristic Method for Automated Selection of Drugs for Screening". *J. Med. Chem.* **1977**, *20*, 469–475.
(41) Hodes, L. "Computer-Aided Selection of Novel Antitumor Drugs for Animal Screening". *ACS Symp. Ser.* **1979**, *112*, 583–602.
(42) Hodes. L. "Computer-Aided Selection of Compounds for Animal Screening: Validation of a Statistical-Heuristic Method". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 128–132.
(43) Hodes, L. "Selection of Molecular Fragment Features for Structure–Activity Studies in Antitumor Screening". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 132–136.
(44) Tinker, J. "Relating Mutagenicity to Chemical Structure". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 3–7.
(45) Varmuza, K. "Pattern Recognition in Chemistry". *Lect. Notes Chem.* **1980**, *21*, 21–22.
(46) Henry, D. R.; Jurs, P. C.; Denny, W. A. "Structure–Antitumor Activity Relationships of 9-Anilinoacridines Using Pattern Recognition". *J. Med. Chem.* **1982**, *25*, 899–908.
(47) Denny, W. A.; Cain, B. F.; Atwell, G. J.; Hansch, C.; Panthanaickal, A.; Leo, A. "Potential Antitumor Agents. 36. Quantiative Relationships between Experimental Antitumor Activity, Toxicity, and Structure for the General Class of 9-Anilinoacridine Antitumor Agents". *J. Med. Chem.* **1982**, *25*, 276–315.
(48) See, e.g., Zar, J. H. "Biostatistical Analysis"; Prentice-Hall: Englewood Cliffs, NJ, 1974; pp 60–67.
(49) Based upon the relative importance of each descriptor listed in Table IV of reference 47.
(50) Miller, K. J.; Savchik, J. A. "A New Empirical Method to Calculate Molecular Polarizabilities". *J. Am. Chem. Soc.* **1979**, *101*, 7206–7213.
(51) Gasteiger, J.; Marsili, M. "Iterative Partial Equalization of Orbital Electronegativity - A Rapid Access to Atomic Charges". *Tetrahedron* **1980**, *36*, 3219–3228.
(52) Smith, D. H.; Carhart, R. E.; Crandell, C. W.; Venkataraghavan, R. "Constructive Perception of Shared Three-Dimensional Substructures". "Abstracts of Papers", 186th National Meeting of the American Chemical Society, Washington, DC, Aug 1983; American Chemical Society: Washington, DC, 1983; CINF 3.

Computer Code for Producing Eh–pH Plots of Equilibrium Chemical Systems†

DENNIS R. DREWES*

Rockwell Hanford Operations, Basalt Waste Isolation Project, Richland, Washington 99352

Received February 14, 1984

This paper describes a computer code that produces high-quality potential (Eh)–pH diagrams by coupling the power of the computer for handling extensive calculations with modern digital graphics hardware. The code, called EHPH, is written in standard FORTRAN 77 and has been designed to offer the user considerable flexibility as well as ease of use. The code produces three Eh vs. pH plots for each problem submitted: (1) a plot of the regions of stability for the stable solid species, (2) a plot of the regions of dominance of the dominant aqueous species, and (3) a contour plot of the total solubility of all species. The DISSPLA graphics package is employed in production of the plots.[1] A user's guide is available.[2]

INTRODUCTION

Potential–pH (Eh–pH) diagrams are useful analytical tools for understanding the thermodynamic relationships between species in chemical systems.[3,4] In engineering applications, for instance, Eh–pH analysis may be used to predict how corrosion product formation depends on the electrochemical environment. Geochemical studies use Eh–pH analysis to infer conditions under which various geologic units were formed or to determine stable mineral assemblages under given conditions. The thermodynamic basis upon which Eh–pH analysis rests is straightforward, but the calculations required to plot an Eh–pH diagram are extremely tedious and time consuming. These two qualities make Eh–pH analysis an ideal application for computers, which, in addition to making calculations rapidly, thrive on tedium.

The computer code described in this paper produces high-quality Eh–pH diagrams by coupling the speed of the computer with modern digital graphics hardward. The EHPH code is written in standard FORTRAN 77 and has been designed to offer the user considerable flexibility as well as ease of use. The code produces three plots for each problem submitted: (1) a plot of the regions of stability for the stable solid species, (2) a plot of the regions of dominance of the dominant aqueous species, and (3) a contour plot of the total solubility of all species.

The code is simple to use, requiring as input only a list of the species formed from the element under investigation, corresponding thermodynamic data, and concentrations of complexing species. The necessary thermodynamic data may exist in one of several forms, such as free energies of formation or reaction, standard electrode potentials, or equilibrium constants. The code includes modules for interactive entry of the necessary data to simplify input.

The program includes a mechanism for relating the thermodynamic data to nonstandard temperatures using entropy data. In addition, the code allows the user to include the speciation of sulfur-bearing complexing species; if this option is ignored, S(VI) is assumed throughout the Eh range.

The user has full control over the pH and Eh ranges to be covered in the analysis and subsequently plotted. Default values are included for most of the input requested from the user, and several calculation and plotting options are included. In addition to graphical output, the program also produces a log file to preserve a record of the details of the calculation.

The code was developed on a PRIME 750 computer using the PRIMOS operating system. System routines have been avoided except for a few that have analogs in virtually all systems. Thus, conversion to other FORTRAN 77 systems should be reasonably simple. A user's guide has been written

† This work was prepared for the U.S. Department of Energy under Contract DE-AC06-77RL01030.

* Address correspondence to the author at Boeing Computer Services, Inc., Seattle, WA 98124.

0095-2338/85/1625-0073$01.50/0 © 1985 American Chemical Society