A GRAPH THEORY-BASED "EXPERT SYSTEM" METHODOLOGY FOR STRUCTURE-ACTIVITY STUDIES

Gilles KLOPMAN and Robert V. HENDERSON
Department of Chemistry, Case Western Reserve University, Cleveland, OH 44106, USA

Abstract

A new graph theory-based methodology is proposed for the statistical evaluation of the link between the structure and the chemical/biological activity of organic molecules. The computer-aided analysis involves the heuristic processing of molecules as chemical graphs which are decomposed into their component subgraphs. Common topological features among the subgraphs are then statistically isolated, and a set of rules is developed that can be used to explain the activities of the analyzed compounds as well as predict the activities of new compounds. The validity of the methodology is demonstrated by its application to actual experimental data.

1. Introduction

The formulation of topological molecular descriptors by the application of graph theory to chemical problems is a topic of considerable interest to many researchers working in the area of medicinal chemistry. Previous investigations have provided evidence that biological properties (mutagenicity, toxicity, etc.) can be viewed as functions of topology-based parameters [1]. The goal of such an approach is to identify the appropriate submolecular features that are useful for the recognition of the reactive regions of biologically-important molecules. The challenge to the researcher is to find an adequate method with which molecular entities can be represented mathematically by the appropriate treatment of an input chemical code.

In this work, we propose that the union of two independent disciplines – graph theory and artificial intelligence – can serve to provide a rather unique approach to the problems confronted by workers in the area of QSAR – quantitative structure/activity relationships. Graph theory offers a simplicity in the representation of otherwise often complex molecular structures in addition to its primary function – the numerical quantification of submolecular features. On the other hand, expert system analysis provides the tools for taking this graph-theoretical information as input data and processing it heuristically, with the final product being one or more logical statements (in the finite mathematical sense) that explain data that was previously unmanageable.

© J.C. Baltzer AG, Scientific Publishing Company

The usefulness of the proposed methodology will be demonstrated by its application to the analysis of two different data bases; a set of 122 compounds classified as acids and nonacids, and a set of 64 nitrosoamines classified as rat carcinogens and noncarcinogens.

2. Graph-theoretical representation of data

There are a multitude of methods provided by graph theory which have been employed to represent or numerically quantitate the chemical graph [2]. A common procedure is to assign a weight to each vertex (atom) based on the structural environment of the atom, perform a given mathematical operation on each weight or group of weights, and then to perform a summation over the entire molecule [3]. The index that results is then assumed to represent that molecule as a correlation is sought between structure (as reflected by the magnitude of the index) and a particular activity [4]. The activity can range from something as simple to measure as pH to something as complicated as "lethal dose", where the determination is costly and may involve years of extensive animal testing.

The hypothesis behind the following methodology is that current graph-theoretical methods over-emphasize the compression of the molecular graph into a single numerical entity (index). Further, it is hereby proposed that the reduction of many different structural features into a single representative value for an entire molecule is an unrealistic goal. The individual and possibly unique structural environment of each atom becomes a term in a summation that mixes and, in a sense, averages everything together. Methodologies that attempt to address this issue can be found in the current literature [5,6]. The alternative approach offered here is that the particular structural environment of each non-hydrogen atom within the molecule be characterized by a parameter of dimension greater than one, a k-dimensional vector ($1 < k <$ number of atoms in the molecule) for our purposes. Each molecule is now represented by a matrix, wherein each row characterizes the structural environment of a heavy atom (non-hydrogen atom) within the molecule, as demonstrated below for compound 1 of the acid/nonacid data base.

The input information for the formulation of these graph-theoretical entities is in the form of a linear code which is translated into a connection table (connectivity matrix) which identifies the atoms, their connectivity, and their bond multiplicity. This modified connectivity matrix $[\mathbf{C}_{ij}]$ is then transformed into a distance matrix $[\mathbf{D}_{ij}]$ which contains the topological information required to construct the vectors to be used to describe the structural environment of each atom. The structure and atom numbering for compound 1 of table 1 is given in fig. 1, and the resulting distance matrix is given in table 2(a). As can be seen, these distances are not based on actual bond lengths but are instead distances in the graph-theoretical sense.

Table 2(b) shows the vector representation of each heavy atom of compound 1 obtained by counting the number of equivalent elements in each row (atom)

Table 1

The training set (see text for explanation) Scale: acidity = −10 pka + 90

Acidity	KLN code	Chemical formula	Compound no.
90	XDXX	$CH(NO_2)_3$	1
90	DYMYMYM	$CH(SO_2CH_3)_3$	2
90	DC3NC3NC3N	$CH(CN)_3$	3
88	FCTKFF	F_3CCOOH	4
83	GCTKGG	Cl_3CCOOH	5
77	KTTK	HOOCCOOH	6
77	GDTKG	Cl_2CHCO_2H	7
73	RXTK	CH_2NO_2COOH	8
64	FRTK	FCH_2COOH	9
62	KTRTK	$HOOC\ CH_2COOH$	10
61	GRTK	$Cl\ CH_2CO_2H$	11
61	MRDGTK	$CH_3CH_2CH\ Cl\ COOH$	12
56	X*TK	NO_2-Ph-COOH	13
52	KRTK	$HOCH_2COOH$	14
51	MC2DD2DD2CTK/	o-toluic acid	15
50	MDGRTK	$CH_3CH\ Cl\ CH_2COOH$	16
50	XRX	$CH_2(NO_2)_2$	17
50	G*TK	Cl-Ph-COOH	18
50	B*TK	Br-Ph-COOH	19
48	KT=2TK	$HOOC(CH_2)_2COOH$	20
48	*TK	Ph-COOH	21
47	KT=3TK	$HOOC(CH_2)_3COOH$	22
46	M*TK	p-toluic acid	23
46	KT=4TK	$HOOC(CH_2)_4COOH$	24
45	MO*TK	MeO-Ph-COOH	25
45	KT=5TK	$HOOC(CH_2)_5COOH$	26
45	KT=6TK	$HOOC(CH_2)_6COOH$	27
45	KT=7TK	$HOOC(CH_2)_7COOH$	28
44	KT=8TK	$HOOC(CH_2)_8COOH$	29
44	G=3TK	$Cl(CH_2)_3COOH$	30
42	MTK	CH_3CO_2H	31
42	M=2TK	$CH_3(CH_2)_2COOH$	32
41	MRTK	CH_3CH_2COOH	33
41	M=3TK	$CH_3(CH_2)_3COOH$	34
30	DTMTMTM	$CH(C{=}O\ CH_3)_3$	35
10	*J	Ph-SH	36
10	MTRTM	$CH_3C{=}O\ CH_2C{=}O\ CH_3$	37
10	*K	Phenol	38
10	RC3NC3N	$CH_2(CN)_2$	39
10	RYMYM	$CH_2(SO_2CH_3)_2$	40

. . . continued

Table 1 (continued)

Acidity	KLN code	Chemical formula	Compound no.
10	MRK	CH_3CH_2OH	41
10	RD2DD2D/	Cyclopentadiene	42
10	X*A	NO_2-Ph-NH_2	43
10	D*C2DD2DD2C)C2DD2DD2C)/	9-phenyl fluorene	44
10	RD2DC2DD2DD2C)/	Indene	45
10	MTM	$CH_3C{=}O\ CH_3$	46
10	RC2DD2DD2C)C2DD2DD2C)/	Fluorene	47
10	*C3D	Phenylacetylene	48
10	C2DR***	1,1,3-triphenylpropene	49
10	*A	Ph-NH_2	50
10	M*A	Me-Ph-NH_2	51
10	MYM	$CH_3SO_2CH_3$	52
10	D***	$(Ph)_3CH$	53
10	*R*	$(Ph)_2CH_2$	54
10	M*	Toluene	55
10	D2DD2DD2D/	Benzene	56
10	RRR/	Cyclopropane	57
10	MDM*	Cumene	58
10	DC2DD2DD2C)DC2DD2DD2C)/C2D	Tripticene	59
10	DC2C/***	Triphenylcyclopropene	60
10	RRRR/	Cyclobutane	61
10	R=4/	Cyclopentane	62
10	R=5/	Cyclohexane	63
10	=4	Butane	64
10	=5	Pentane	65
10	=6	Hexane	66
10	=7	Heptane	67
10	=8	Octane	68
10	=9	Nonane	69
10	=9=6	$C_{15}H_{32}$	70
10	=9=9=2	$C_{20}H_{42}$	71
10	=3TH	Butanal	72
10	=4TH	Pentanal	73
10	=5TH	Hexanal	74
10	=6TH	Heptanal	75
10	=7TH	Octanal	76
10	=8TH	Nonanal	77
10	MORM	CH_3-O-CH_2CH_3	78
10	MRORM	$(CH_3CH_2)_2O$	79
10	MRRORRM	$(CH_3CH_2CH_2)_2O$	80

. . . continued

Table 1 (continued)

Acidity	KLN code	Chemical formula	Compound no.
10	MRA	$CH_3CH_2NH_2$	81
10	MRRA	$CH_3CH_2CH_2NH_2$	82
10	MRRRA	$CH_3CH_3CH_2CH_2NH_2$	83
10	*RA	Ph-CH_2-NH_2	84
10	MR*A	CH_3CH_2-Ph-NH_2	85
10	G*RA	Cl-Ph-CH_2-NH_2	86
10	R2DRA	CH_2=$CHCH_2NH_2$	87
10	MRRK	$CH_3CH_2CH_2OH$	88
10	MRRRK	$CH_3(CH_2)_3OH$	89
10	MRRDKM	$CH_3(CH_2)_2CHOHCH_3$	90
10	D2DRRR)	Cyclopentene	91
10	D2DRRRR)	Cyclohexene	92
10	MTRM	$CH_3COCH_2CH_3$	93
10	MRTRM	$(CH_3CH_2)_2CO$	94
10	*TM	Ph-CO-CH_3	95
10	*RTM	Ph-CH_2-CO-CH_3	96
10	R2DRM	CH_2=$CHCH_2CH_3$	97
10	MD2DRM	CH_3CH=$CHCH_2CH_3$	98
10	FCFFM	F_3CCH_3	99
10	GCGGRK	Cl_3CCH_2OH	100
10	ARTK	Glycine	101
10	MDATK	Alanine	102
10	MDMDATK	Valine	102
10	MDMRDATK	Leucine	104
10	MRDMDATK	Isoleucine	105
10	KRDATK	Serine	106
10	MDKDATK	Threonine	107
51	KTRDATK	Aspartic acid	108
10	ATRDATK	Asparagine	109
48	KTRRDATK	Glutamic acid	110
10	ATRRDATK	Glutamine	111
10	A=4DATK	Lysine	112
10	ARDKRRDATK	Hydroxylysine	113
30	C2DN2DE/RDATK	Histidine	114
10	AC2EERRRDATK	Arginine	115
10	*RDATK	Phenylalanine	116
10	K*RDATK	Tyrosine	117
10	D2DD2DC2C/ED2C/RDATK	Tryptophan	118
10	JRDATK	Cysteine	119
10	MSRRDATK	Methionine	120
10	DRRRE/TK	Proline	121
10	DRDKRE/TK	Hydroxyproline	122

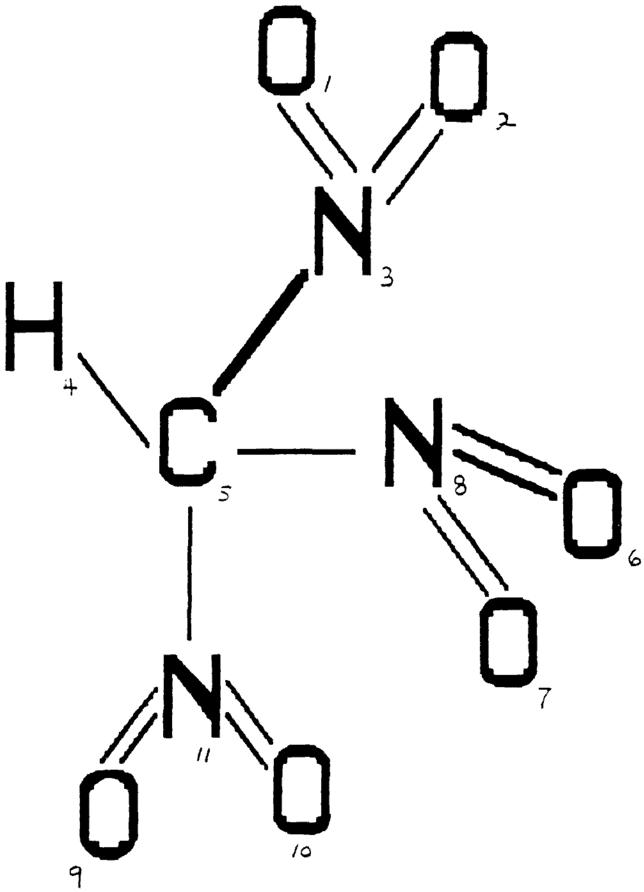


Fig. 1.

of the distance matrix in table 2(b). With this notation, each vector signifies a rooted tree in which the entire molecule is viewed from the perspective of a single reference atom. Due to the high degree of symmetry, compound 1 has only three different types of heavy atoms – each represented by its own vector descriptor (table 2(c)). Specifically, the six oxygen atoms are topologically equivalent, the three nitrogens are equivalent, and the carbon atom is unique within the molecular environment of compound 1.

By sorting and combining the elements of the distance matrix, the information contained therein is compressed into the path matrix [$\mathbf{P}_{ij}$]. This procedure is repeated for each molecule of the data set. Next, all the rows of [$\mathbf{P}_{ij}$] are combined into one large cumulative matrix [$\mathbf{P}_{IJ}$] (see table 2 for an actual example based on compound 1). Finally, each row (vector) of [$\mathbf{P}_{IJ}$] is decomposed into all possible subunits, Table 2(a)

Distance matrix $[\mathbf{D}_{ij}]$

Atom	1	2	3	4	5	6	7	8	9	10	11
1	0	2	1	3	2	4	4	3	4	4	3
2	2	0	1	3	2	4	4	3	4	4	3
3	1	1	0	2	1	3	3	2	3	3	2
4	3	3	2	0	1	3	3	2	3	3	2
5	2	2	1	1	0	2	2	1	2	2	1
6	4	4	3	3	2	0	2	1	4	4	3
7	4	4	3	3	2	2	0	1	4	4	3
8	3	3	2	2	1	1	1	0	3	3	2
9	4	4	3	3	2	4	4	3	0	2	1
10	4	4	3	3	2	4	4	3	2	0	1
11	3	3	2	2	1	3	3	2	1	1	0

Table 2(b)

Path matrix $[\mathbf{P}_{ij}]$

Atom	Vector descriptor								
1	1	2	3	4	0	0	0	0	0
2	1	2	3	4	0	0	0	0	0
3	3	3	4	0	0	0	0	0	0
4	hydrogen atom								
5	4	6	0	0	0	0	0	0	0
6	1	2	3	4	0	0	0	0	0
7	1	2	3	4	0	0	0	0	0
8	3	3	4	0	0	0	0	0	0
9	1	2	3	4	0	0	0	0	0
10	1	2	3	4	0	0	0	0	0
11	3	3	4	0	0	0	0	0	0

Table 2(c)

Path matrix, cumulative $[\mathbf{P}_{IJ}]$
(compound 1 only)

1	2	3	4
3	3	4	0
4	6	0	0

Table 2(d)

Potential descriptors
(from compound 1)

1			
3			
4			
1	2		
3	3		
4	6		
1	2	3	
3	3	4	
1	2	3	4

each representative of one or more chemical substructures. The assemblage of these potential descriptors will then serve as an input file for the "expert system" analysis to follow.

3. Expert system analysis

The expert system analysis is a useful procedure when there exists a general lack of knowledge regarding causality yet a wealth of data detailing experimental observations [8]. Since the basis of chemistry itself is the explanation of molecular properties in terms of structure, it is assumed that the experimental data (biological properties) are also the consequence of specific structural features. The problem in processing the large amount of available experimental data is that the compounds involved are usually quite compex and the number of potentially important structural features is overwhelming. It is under these conditions that computer implemented expert system analysis is ideal [9]. The general procedure is to examine *all possible* structural features – a feat that is not really possible without the use of a computer due to the sheer amount of data generated. The potential descriptors are then sorted and ordered according to importance on the basis of one or more statistical parameters (described in the following section). A heuristic algorithm is then followed which selects the optimal topological structural feature or features. Having isolated the structural feature that apparently accounts for the activity of compounds in which it occurs, these compounds are then eliminated from the data set and the remaining compounds are then submitted to the same analysis. This procedure is continued repeatedly until either:

(1) the entire data set is eliminated (i.e. various structural features have been revealed that account for the experimentally observed activities of all the compounds), or

(2) all statistically relevant fragments have been isolated and the remaining data set (usually only a few compounds) is too small to provide any significant information.

4. Methodology

4.1. STATISTICAL PARAMETERS

Two independent measures are employed for the statistical evaluation of data – one for the selection of the optimum parameter and the other to ensure the statistical significance of that parameter. The selection of parameters is based on a formulation by Laplace [10] and the establishment of statistical significance is based on binomial probabilities [11].

The optimum parameter selection is determined by an expression evolved from Laplace's work concerning the probability of future events. Laplace's "inverse probability" or "probability of causes" is controversial in the sense that it does not satisfy the formal definition of probability and is arrived at through an intuitive argument as opposed to a rigorous derivation. For our purposes, we implement a modification of Laplace's expression simply to take advantage of its favorable properties. We circumvent the previous controversy by merely viewing is as an "index of likeliness" as opposed to a formal probability. The modified expression we use defines a normalized index identified as $I(m)$:

$$I(m) = 1 - 4\,\frac{(a+m)\,(i+m)}{(a+i+2m)^2}\,, \tag{1}$$

where a = number of actives, i = number of inactives, 4 = normalization constant, and m = adjustable parameter. The $I(m)$ function is an attempt to quantify a rather qualitative and arbitrary entity, i.e. the amount of skewedness which indicates a link between a structural feature (fragment) and consequent activity, thus serving as a precursor to the establishment of causality. Thus, the larger the $I(m)$ value for a descriptor, the stronger its association with either activity or inactivity. Adjustable parameter (m), also of an arbitrary nature, is viewed as a "rate of learning" constant whose optimal value was determined to be in the range from 0.01 to 10.

Although $I(m)$ serves to evaluate the "predictive potential" of a fragment, it imparts no information concerning the statistical significance of such a prediction. To establish statistical significance, binomial probabilities are calculated based on the fragment distributions of actives and inactives as obtained from the data set (parent) distribution. The expression we employ is a summation over the relevant binomial probabilities – directly analogous to the probabilities calculated for the "coin toss" experiment. The probability calculated for some fragment (i) is:

$$P_i(x > x_0) = \sum_{x=x_0}^{n} \frac{n!}{x!(n-x)!}\, p^x q^{n-x}, \tag{2}$$

with

x_0 = number of active* compounds in which fragment i occurs,

n = total number of compounds containing fragment i,

p = (number of active compounds)/(total number of compounds in entire data set),

q = $1 - p$.

*The term "active" in the definition of x_0 and p can be replaced by the label "inactive" to evaluate inactive fragments.

4.2. DATA TREATMENT

The selection of relevant submolecular descriptors from the large data pool is now discussed. Each row of matrix $[\mathbf{P}_{IJ}]$ is identified as an **SEV** (structural environment vector) which is used to characterize each atom within a molecule. Each atom has also been assigned the activity of the molecule to which its belongs. We now follow a previously employed artificial intelligence algorithm [9] whereby the **SEV** elements are broken into their various subsets and each then tested for significance. For example, the row of $[\mathbf{P}_{IJ}]$ that represents compound 1, atom 1 of the acid data base (table 1) is decomposed into the following **SEV** subsets – given together with the number of molecules that contain the same fragment:

Fragment vector	Compounds		
	total	active	inactive
[1]	76	38	38
[1 2]	69	36	33
[1 2 3]	17	11	6
[1 2 3 4]	8	8	0
Entire data	122	38	84

This means that vector [1] occurred in 76 molecules, 38 of which were active (acids) and 38 inactive (non-acids). Thus, every **SEV** and its component subsets form a distribution that can be tested for statistical significance and evaluated for its potential usefulness as a predictive parameter. Calculating the index of skewedness $I(m)$ and the binomial probability (BiP) according to eqs. (1) and (2) for these four fragments, we obtain the following values:

Fragment vector	$I(m)$ ($m = 0.01$)	BiP
[1 2 3 4]	0.9950	0.0001
[1 2 3]	0.0863	0.0045
[1 2]	0.0019	0.0002
[1]	0.0000	0.0005

Thus, in comparing these four fragments, [1 2 3 4] has the largest $I(m)$ and is therefore assumed to have greater predictive potential than the other three. The reason for this is that the distribution (8 : 0) indicates that if [1 2 3 4] is considered the sole criterion for activity, eight compounds are successfully accounted for and zero misclassifications result. In contrast, if any of the other three were the sole criterion, a high number of misclassifications would result. It is further noted that $I(m)$ greatly distinguishes

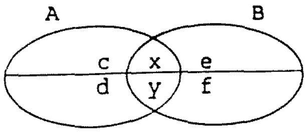
[1 2 3 4] from the others, whereas *BiP* does not. In fact, according to *BiP* values, fragments [1] and [1 2] are comparable to [1 2 3 4] in terms of statistical significance (smaller *BiP* ⇒ greater significance). The important feature of the *BiP* values is that all four fragments are significant at a 95% confidence level ($BiP < 0.05$) and any small difference between them is of little consequence. If these in fact were our fragments of interest, we would conclude that [1 2 3 4] has the greatest predictive potential and also has an acceptable statistical significance. The other three fragments are also statistically significant, but of questionable use as predictive parameters.

In the previous example, the choice between descriptors was straightforward due to the extreme variance in $I(m)$ values. However, often the selection of optimal descriptors is not so well defined. In fact, multiple descriptors may even have the same activity distributions (number of actives versus number of inactives) and thus identical $I(m)$ values. Such cases are encountered so frequently that the algorithm below is systematically applied whenever multiple potential descriptors are found to be significant yet are not selectively discriminated by the $I(m)$ and *BiP* functions. The extensive use made of this flowchart as a reference in the selection and rejection of potential descriptors merits the labeling of the various alternatives (i.e. cases (i), (ii), . . . , (ix)) for convenient location of the appropriate path of logic. The rationale behind the flowchart – where not intuitively obvious – is further explained when it is applied to actual cases that arise in the analysis of data (see section 5).

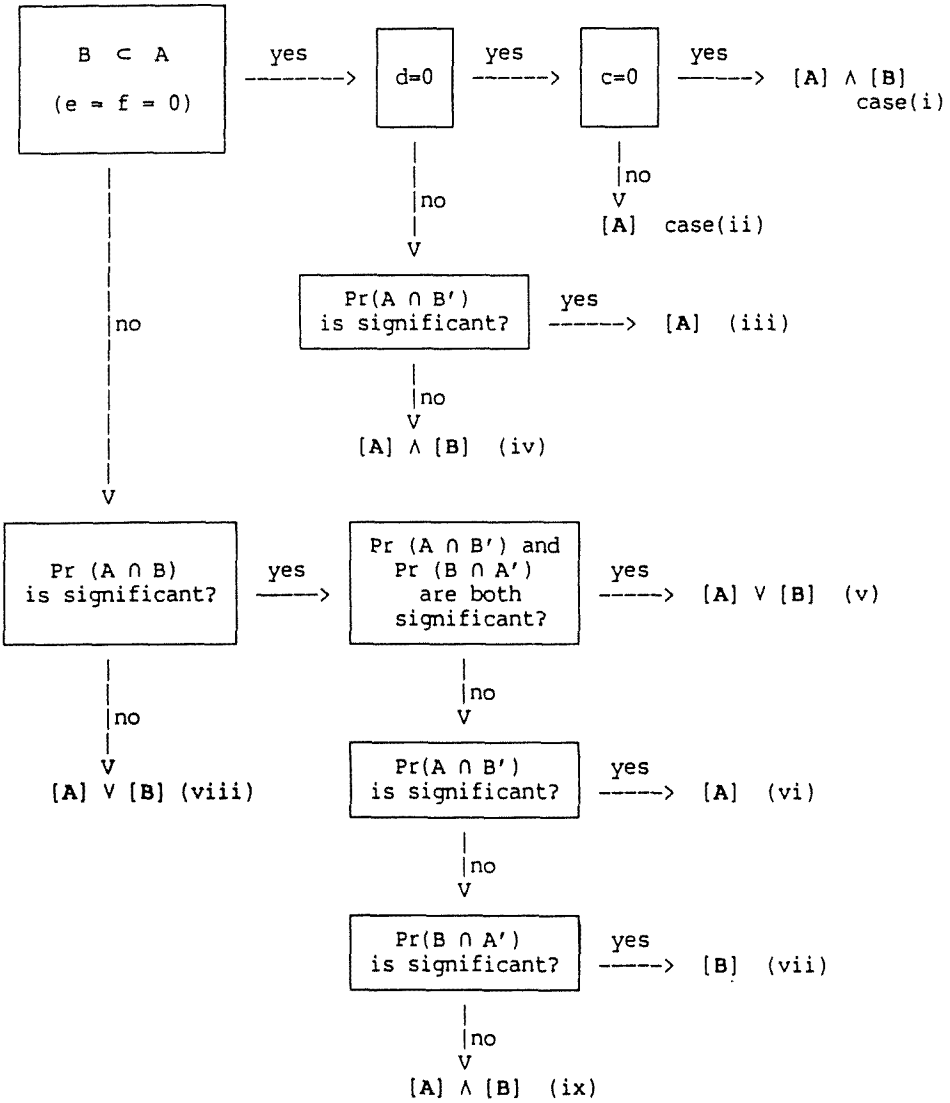
The flowchart represents the possible relationships between fragments [**A**] and [**B**], together with the resulting logical statements. ([**A**] and [**B**] are assumed to be activating fragments – for inactives, simply interchange c, x, e with d, y, f.) The symbols used in the flowchart are defined below.

$\subset$	subset;
$\sim$, $'$	negation, complement (not);
$\wedge$, $\cap$	conjunction, intersection (and);
$\vee$, $\cup$	disjunction, union (or);
[**A**], [**B**]	2 **SEV** descriptors;
A	set of molecules containing [**A**];
B	set of molecules containing [**B**];
c	number of active molecules containing [A], but not [B];
d	number of inactive molecules containing [A], but not [B];
e	number of active molecules containing [B], but not [A];
f	number of inactive molecules containing [B], but not [A];
x	number of active molecules containing both [A] and [B];
y	number of inactive molecules containing both [A] and [B].

Flowchart



Assume [A] and [B] are activating fragments found to be statistically significant (same argument applies to inactive fragments, simply switch c:d, e:f, and x:y)



5. Application of methodology

5.1. ACIDITY DATA

In order to illustrate the methodology, we applied our analysis to the rather trivial task of recognizing and predicting simple Lowry–Bronsted acidity of 122 randomly selected organic compounds. (Trivial in the sense that the mechanism and factors influencing acidity – electron withdrawal, resonance, etc. – are well understood.) However, such a well-known property is ideal for illustrating and validating a new methodology. Our goal here is to build-up a heuristic table (set of rules) that employ graph-theoretical descriptors for the purpose of determining the factors responsible for any particular activity of interest.

We start by submitting a data file that will serve as a training set from which the program will "learn" the factors that contribute to acidity and those that do not. In order to accomplish this, the input data must contain "active" as well as "inactive" compounds, with the ideal data base containing 50% of each. The training set is shown in table 1 and contains the following, from left to right:

acidity – the relative acidities are based on Ka data [12] (extrapolated where necessary) and are rescaled from 10 through 90 for the program (90 = most active) by the expression: acidity = -10 pKa + 90;

KLN code – this is a linear coding system [13] in which different symbols represent atoms and/or functional groups;

compound name – this column is not used by the program and merely serves as a convenience by which the user can represent a compound by any character sequence desired.

The input data shown in table 1 produces a 1014 × 21 [$\mathbf{P}_{IJ}$] matrix, which in turn generates a list of 363 potential descriptors. At the first stage of the analysis, a search is made to find either fragments common to all the compounds or fragments common to all the active compounds of the data set. Vector [1], as indicated below, is just such a fragment.

	total	active	inactive
data	**122**	**38**	**84**
[1]	76	38	38

The fact that all active molecules contain this simple vector allows us to conclude that a terminal heavy (non-hydrogen) atom is necessary for activity and any molecule not meeting this criterion is classified as inactive, regardless of any other factors. This is represented by the logical statement:

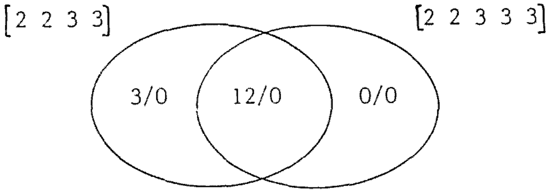
$\tilde{}$ [1] ⇒ inactive,

which is interpreted as "the absence of [1] implies inactivity". This statement alone eliminates 46 inactive compounds from the data set, having correctly classified them as inactive.

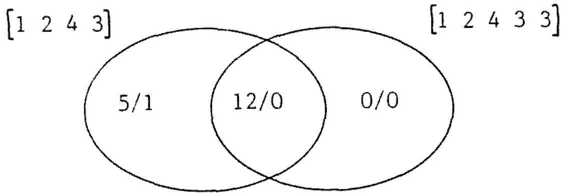
Following the elimination of 46 inactive compounds, the remaing 76 molecules are subjected to the previously described analysis wherein the 636 rows of the reduced [$\mathbf{P}_{IJ}$] provide 256 potential descriptors to be evaluated in the second stage of the analysis. These fragment descriptors are sorted and ordered according to relevance to activity by the $I(m)$ parameters. The **SEV** descriptors that emerge as the statistically best are given below:

	total	active	inactive
	76	**38**	**38**
[2 2 3 3]	15	15	0
[1 2 4 3 3]	12	12	0
[2 2 3 3 3]	12	12	0
[1 2 4 3]	18	17	1

[2 2 3 3 3] is rejected since it is a subset of [2 2 3 3], as demonstrated by the following Venn diagram, and thereby provides no useful information (see case (ii) of flowchart).



The same procedure is used to examine the relationship between descriptor [1 2 4 3] and its subset [1 2 4 3 3].

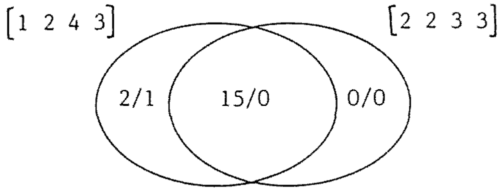


Here, the choice between descriptors is not as obvious as in the previous case, and we again rely on a statistical solution to the problem. Displayed in the table below is the information that [1 2 4 3] accounts for more of the data, but makes one false classification (i.e. one inactive compound would incorrectly be classified as active). On the other hand, subset [1 2 4 3 3], by being more restrictive, can account for a smaller percentage of the data, but makes no false classifications.

	total	active	inactive	*BiP*
	76	**38**	**38**	
[1 2 4 3]	18	17	1	
[1 2 4 3 3]	12	12	0	
[1 2 4 3] $\cap$ [1 2 4 3 3]′	6	5	1	0.109

The last line of the above table is the evaluation of [1 2 4 3] as an independent descriptor. In other words, there are six molecules that contain [1 2 4 3] but do not contain [1 2 4 3 3], of which five are active and one is inactive. If the probability of this distribution (6:5 to 1) is high enough to be considered random, we would attribute the cause of activity to [1 2 4 3 3]. However, the probability is low (*BiP* = 0.109) and we conclude that [1 2 4 3] is statistically significant even in the absence of [1 2 4 3 3] (see case (iii) of flowchart). At this point, two vectors ([2 2 3 3] and [1 2 4 3]) have been linked to activity and the two subset vectors have been rejected, in accordance with the algorithm and statistical arguments.

Next, the relationship between [2 2 3 3] and [1 2 4 3] is investigated.



We now reject [1 2 4 3] as an "independent" descriptor since it is not significant in the absence of [2 2 3 3]. However, all 15 compounds containing [2 2 3 3] also contain [1 2 4 3]. Thus, since we have no information on the statistical validity of [2 2 3 3] as an independent parameter (i.e. in the absence of [1 2 4 3]), we recognize the limitation imposed by the data and simply conclude that both descriptors must be present for a statistically reliable classification to be made (case (iv) of flowchart). This conclusion is represented as follows:

[2 2 3 3] $\wedge$ [1 2 4 3] $\rightarrow$ "active" (activity = 46 ± 5)

Table 3

BiP ordering; total = 16, actives = 5, inactives = 11

Total	Active	Inactive	*BiP*	Vector
2	2	0	0.0977	1 2 3
5	3	2	0.1800	4 6
4	0	4	0.2234	3 4
3	2	1	0.2319	3 3
3	0	3	0.3250	3 4 3
3	0	3	0.3250	3 4 3 1
6	1	5	0.3936	3 6
6	1	5	0.3936	4 2
2	0	2	0.4727	3 7
2	0	2	0.4727	1 3 3
2	0	2	0.4727	3 4 5
2	0	2	0.4727	3 4 6

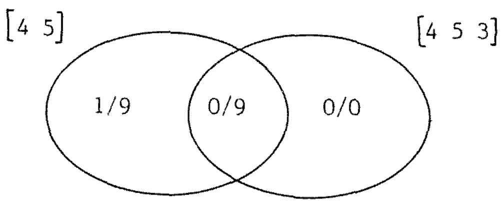
(see table 3 for a chemical interpretation of the vector descriptors). Since the average activity of the compounds containing these two descriptors is 46.5 with a standard deviation of less than 5 activity units, the above expression includes this information. This case shows that although the predictions are generally qualitative (active or inactive), it is possible to obtain semi-quantitative results.

At stage 3 of the analysis, the following vectors are found to be the best potential descriptors and thereby qualify for further consideration. It is noted that these vectors are associated with inactivity.

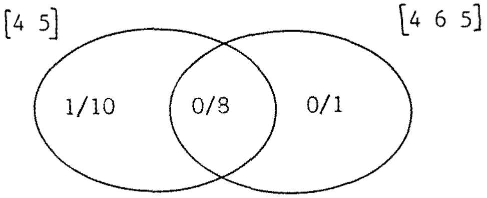
	total	active	inactive
	61	**23**	**38**
[4 5 3]	9	0	9
[4 6 5]	9	0	9
[4 5]	19	1	18
[4 7]	19	1	18
[1 2 4 5]	19	1	18
[2 2 3 5]	19	1	18

Since [4 5 3] is a subset of [4 5], a comparison between them results in the rejection of [4 5 3] on statistical grounds (case (iii), inactive).

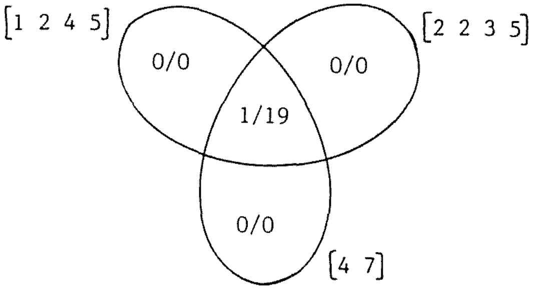
	total	active	inactive	*BiP*
	61	**23**	**38**	
[4 5]	19	1	18	
[4 5 3]	9	0	9	
$[4\ 5] \cap [4\ 5\ 3]'$	10	1	9	0.062



[4 6 5] is now compared to [4 5]:



As is easily concluded from the this diagram, [4 6 5] is also rejected as a descriptor and [4 5] is retained (case (vi), inactive). [4 5] will next be compared to the other potential descriptors. First, the relationship between the three remaining vectors is diagrammed below.

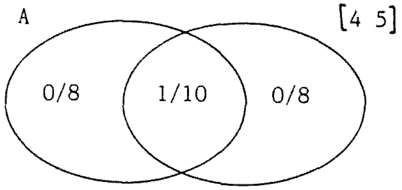


From this, we readily conclude that:

$$[1\ 2\ 4\ 5] \wedge [2\ 2\ 3\ 5] \wedge [4\ 7] \rightarrow \text{inactive}$$

(case (i), inactive – applied twice successively)

Secondly, we examine the relationship of [4 5] to the intersection of the other three descriptors above.



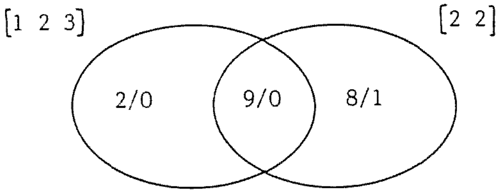
where $A = [1\ 2\ 4\ 5] \cap [2\ 2\ 3\ 5] \cap [4\ 7]$. Since both [A] and [4 5] are significant as independent parameters (case (v), inactive), we conclude stage 3 of the analysis with the statement:

$$([1\ 2\ 4\ 5] \wedge [2\ 2\ 3\ 5] \wedge [4\ 7]) \vee [4\ 5] \rightarrow \text{inactive},$$

which eliminates 28 compounds (27 inactives, one active) from the data set.

Stage 4 involves the evaluation of only two potential vector descriptors.

	total	active	inactive
	34	**22**	**12**
[1 2 3]	11	11	0
[2 2]	18	17	1



[1 2 3] is rejected since its binomial probability as an "independent" parameter (i.e. [1 2 3] $\cap$ [2 2]′] is not considered significant (*BiP* = 0.42), whereas [2 2] is considered independently significant (*BiP* = 0.12). Therefore, [2 2] is the substructural feature selected at this fourth stage of the analysis. As seen, it occurs in 18 molecules, 17 of which are active (case (vii)). The qualitative conclusion is given in the expression below in the usual manner.

[2 2] → active (act. = 56 ± 18).

At stage 5 of the analysis, we find that none of the top $I(m)$ value ordered vectors are statistically significant (at an 80% confidence level) and therefore are all rejected. Under these conditions, the methodology departs from the regular procedure and selects the vectors of lowest binomial probability provided they also meet the criterion of being statistically significant. The top vector descriptors are listed in table 3 by order of statistical significance. As is evident, only the first two potential descriptors ([1 2 3] and [4 6]) are significant according to our standard. Thus, [1 2 3], although only present in two active molecules, is the most significant because the small remaining data set is skewed toward inactivity. This situation is indicative of the final stages of analysis, for at this point most of the data have been explained with considerable rigor. Now the last few compounds remain, and any additional descriptors selected will be of questionable validity due to the depleted data pool from which they are chosen. Be this as it may, [1 2 3] and [4 6] are selected in an attempt to correctly classify several active compounds that would otherwise be misclassified. The following table reveals the union of these two descriptors to be optimal, as indicated by an asterisk (case (viii)).

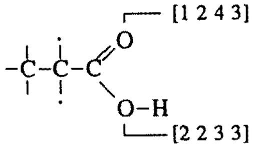
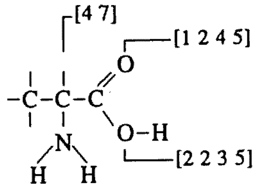
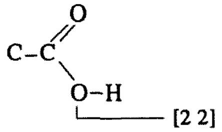
	total	active	inactive	*BiP*
	16	**5**	**11**	
[1 2 3]	2	2	0	0.098
[4 6]	5	3	2	0.180
[4 6] $\cap$ [1 2 3]	1	1	0	0.312
[4 6] $\cup$ [1 2 3]	6	4	2	0.081*

Stage 5 thus concludes the analysis with the statement:

[1 2 3] ∨ [4 6] → active.

The combined use of the five logical statements from each stage of the analysis provides correct classification of 36 of the total 38 active compounds and 81 of the total 84 inactives, for a 96% overall success rate. The most relevant logical statements, together with their chemical interpretations, are given in table 4. The familiar chemical functionalities (e.g. activating carboxylic group, deactivating amino moiety) serve to help validate the proposed methodology.

Table 4
Chemical interpretation of acid data analysis

Logical statement	Classification		Chemical interpretation
	true	false	
~[1] → inactive	46–		all active molecules contain a terminal heavy (non-hydrogen) atom
[2 2 3 3] ∧ [1 2 4 3] → active	15+		 all carboxylic-containing molecules with terminal α-atoms
([1 2 4 5] ∧ [2 2 3 5] ∧ [4 4]) ∨ [4 5] → inactive	27–	1–	 α-amino group deactivates the carboxylic group
[2 2] ⇒ active	17+	1+	

5.2. N-NITROSO COMPUNDS – CARCINOGENIC ACTIVITY DATA

The N-nitroso compounds are of particular pharmacological interest because of their suspected active role in carcinogenesis [14]. The data in table 5 were taken from the literature [15] and serve as the learning data for the analysis. The data contain 45 active and 19 inactive compounds. The average activity is 36 – a value considered to represent moderate potency.

The first step in the analysis identifies fragments [1 1 2] and [2 2] as being common to all the data molecules. Although neither descriptor is of statistical significance, the absence of either of these fragments leads to the default label of "indeterminate", as indicated by expression (3):

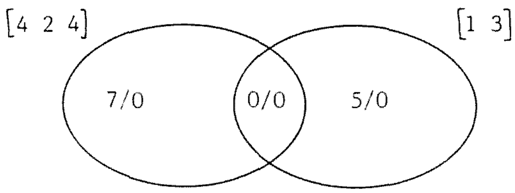
$$\tilde{}[1\ 1\ 2] \wedge \tilde{}[2\ 2] \rightarrow \text{indeterminate.} \tag{3}$$

Both these descriptors are associated with the nitroso functionality, as shown in table 6. Expression (3) can be viewed as a representation of the congeneric nature of the data.

Ordering of the statistically significant ($BiP < 0.15$) SEV descriptors according to their $I(m)$ values produces the following:

SEV	total	active	inactive
[4 2 4]	7	7	0
[1 3]	5	5	0
[1 2 4]	5	0	5
All data	64	45	19

Although the data are highly skewed toward activity, the *BiP* calculation (see section 4.1, eq. (2)) involves terms (p and q) that allow for the fair evaluation of statistical significance even under such unfavorable circumstances. The relationship between the first two SEVs is shown:



The independence of these descriptors is clearly indicated (case (viii)). Expression (4) is the obvious conclusion (see table 6 for actual chemical functionalities).

Table 5

Learning data set for the analysis of N-nitroso compounds

Active	KLN code	Compound name
45	RRRRRN/U	Nitroso-piperidine
39	MDRRRRN/U	2-methyl-nitroso-piperidine
45	RDMRRRN/U	3-methyl-nitroso-piperidine
45	RRDMRRN/U	4-methyl-nitroso-piperidine
10	MDRRRDMN/U	2,6-dimethyl-nitroso-piperidine
25	RDMRDMRN/U	3,5-dimethyl-nitroso-piperidine
10	MCMRRRCMMN/U	2,2,6,6-tetramethyl-nitroso-piperidine
39	RRDC2DD2DD2D)RRN/U	4-phenyl-nitroso-piperidine
10	RRDCMMMRRN/U	4-tert.butyl-nitroso-piperidine
45	RDKRRRN/U	3-hydroxy-nitroso-piperidine
45	RRDKRRN/U	4-hydroxy-nitroso-piperidine
45	RRTRRN/U	4-keto-nitroso-piperidine
10	KTDRRRRN/U	2-carboxy-nitroso-piperidine
10	RRDTKRRN/U	4-carboxy-nitroso-piperidine
55	RRDGRRN/U	4-chloro-nitroso-piperidine
55	RDGDGRRN/U	3,4-dichloro-nitroso-piperidine
55	RDBDBRRN/U	3,4-dibromo-nitroso-piperidine
55	RRD2DRN/U	Nitroso-1,2,3,6-tetrahydropyridine
10	RCTOM2DRRN/U	Nitroso-guvacoline
10	DDC2DD2DD2D)TOMRRRRN/U	Nitroso-methylphenidate
39	RRRRN/U	Nitroso-pyrrolodine
10	MDRRDMN/U	2,5-dimethyl-nitroso-pyrrolidine
55	RDGDGRN/U	3,4-dichloro-nitroso-pyrrolidine
10	KTDRRRN/U	2-carboxy-nitroso-pyrrolidine
10	KTDRDKRN/U	2-carboxy-4-hydroxy-nitroso-pyrrolidine
39	RD2DRN/U	Nitroso-3-pyrroline
45	RRORRN/U	Nitroso-morpholine
55	RDMODMRN/U	2,6-dimethyl-morpholine
39	RRSRRN/U	Nitroso-thiomorpholine
10	MDDC2DD2DD2D)ORRN/U	Nitroso-phenmetrazine
45	RRNURRN/U	Dinitroso-piperazine
55	MDRNURRN/U	2-methyl-dinitroso-piperazine
45	MDRNUDMRN/U	2,5-dimethyl-dinitroso-piperazine
55	MDRNURDMN/U	2,6-dimethly-dinitroso-piperazine
10	MDDMNUDMDMN/U	2,3,5,6-tetramethyl-dinitroso-piperazine
55	NRRNURRR/U	Dinitroso-homopiperazine
10	RRNHRRN/U	Nitroso-piperazine
10	RRNMRRN/U	4-methyl-nitroso-piperazine
45	RRRN/U	Nitroso-azetidine
55	RRRRRRN/U	Nitroso-hexamethyleneimine

. . .continued

Table 5 (continued)

Active	KLN code	Compound name
55	RRRRRRRN/U	Nitroso-heptamethyleneimine
45	RRRRRRRRN/U	Nitroso-octamethyleneimine
39	RRRRRRRRRRRRN/U	Nitroso-dodecamethyleneimine
55	MNUM	Dimethyl-nitrosamine
55	MRNURM	Diethyl-nitrosamine
45	GRRNURRG	bis-(2-chloro)-diethyl-nitrosamine
10	N3CRRNURRC3N	bis-(2-cyano)-diethyl-nitrosamine
45	MORRNURROM	bis-(2-methoxy)-diethyl-nitrosamine
45	MRORRNURRORM	bis-(2-ethoxy)-diethyl-nitrosamine
10	MRODORMRNURDORMORM	bis-(2,2-diethoxy)-diethyl-nitrosamine
39	MDMNUDMM	Di-isopropyl-nitrosamine
10	MRDMNUDMRM	Di-sec.butyl-nitrosamine
55	MRRNURRM	Di-*n*.propyl-nitrosamine
39	RKRNURRK	bis-(2-hydroxy)-*n*.propyl-nitrosamine
45	MTRNURTM	bis-(2-oxo)-*n*.propyl-nitrosamine
10	MONMU	Nitroso-methoxy-methylamine
39	MNURM	Nitroso-methyl-ethylamine
45	MNURRRRRRRRRRM	Nitroso-methyl-undecylamine
45	MNURRRRRRRRRRRM	Nitroso-methyl-dodecylamine
10	MRRRRRRRNURRRRRRRM	Nitroso-di-*n*.octylamine
55	MNURRC2DD2DD2D)	Nitroso-methyl-2-phenyl-ethylamine
55	MNURCMMM	Nitroso-methyl-neopentylamine
45	C2DD2DD2D)NUM	Nitroso-methyl-phenylamine
45	MNUDRRRRR/	Nitroso-methyl-cyclohexylamine

$$[4\ 2\ 4] \vee [1\ 3] \rightarrow \text{active}, \tag{4}$$

$$[4\ 2\ 4] \rightarrow 48.4 \pm 6.5, \tag{4a}$$

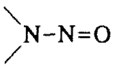
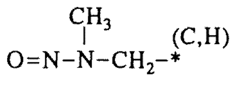
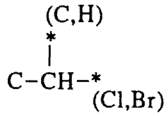
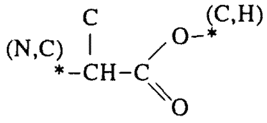
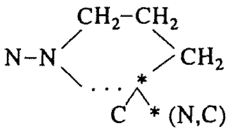
$$[1\ 3] \rightarrow 53.0 \pm 4.5. \tag{4b}$$

It is important at this point to emphasize the hierarchical nature of these logical statements. The descriptors of expression (3) ([1 1 2] and [2 2]) are both considered to be essential for activity classification due to the limitations imposed by the congeneric nature of the training data set. Thus, expression (4) could be more rigorously written as (4c):

$$([4\ 2\ 4] \vee [1\ 3]) \wedge [1\ 1\ 2] \wedge [2\ 2] \rightarrow \text{active}. \tag{4c}$$

Table 6

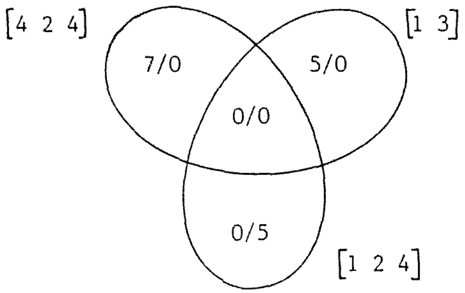
Chemical interpretation of nitroso data analysis

Logical statement	Chemical substructure[a]	Chemical interpretation
~[1 1 2] ∨ ~[2 2] ⇒ indeterminate		all training compounds contain nitroso group, the presence of this functionality is essential for activity predictions
[4 2 4] → 48.4 ± 6.5		**SEV** originates from methyl substituent to nitroso N, activating
[1 3] → 53.0 ± 4.5		see table 7(b)
[1 2 4] → inactive		see table 7(c)
[4 6 4] → 48.5 ± 5.8		see table 7(d)
[1 1 2] ∧ [2 2] → 33.8 ± 18.0		see first statement at top of table

[a] * indicates that more than one type of atom is found at the position.

However, rather than adhere to this cumbersome notation, the same result can be achieved by considering the order of selection of the logical statements to also be their hierarchic order. In other words, if the conditions of expression (3) are satisfied (i.e. either [1 1 2] or [2 2] is absent), then the label "indeterminate" will be assigned to the molecule irrespective of the conclusions drawn from any of the other logical statements occurring lower in the hierarchy.

The third remaining descriptor [1 2 4] is associated with inactivity and is also independent, as shown below:



Statement (5) follows:

[1 2 4] → inactive. (5)

Since all three **SEV**s are at the same level within the hierarchy, we have no statistical means of resolving the conflicting conclusions produced by statements (4) and (5) when a test compound is found to contain fragments that satisfy both conditions simultaneously (5a).

$$([4\ 2\ 4] \vee [1\ 3]) \wedge [1\ 2\ 4] \rightarrow\ ? \qquad (5a)$$

Although a rigorous solution requires more data, when such situations are occasionally encountered, the nature of the activity is considered in order to arrive at an "intuitive conclusion". Since carcinogenesis is the activity, in this case false positives are considered more tolerable than false negatives and the prudent resolution is therefore to replace the "?" in expression (5a) with an "active" label. Thus, statement (4) is considered to be above (5) in the hierarchy.

Thus far, descriptors have been selected and associated with the activity/inactivity of 17 molecules of the data set. After removal of these 17 compounds, the next stage of the analysis produces only one **SEV** descriptor of statistical significance. [4 6 4] occurs in 5 molecules, all of which are active.

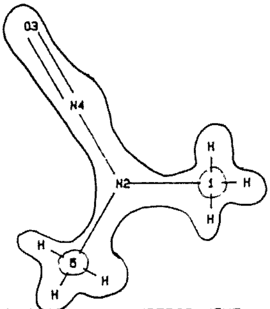
[4 6 4] → active 45.8 ± 5.8 (6)

Considering the hierarchy established so far, statement (6) is valid if and only if the conditions of the previous statements (3)–(5) are false.

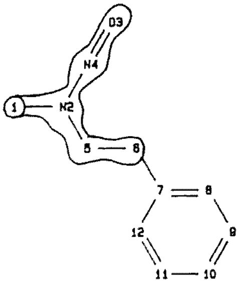
Upon removal of the 5 active compounds with [4 6 4], no fragment descriptor occurring more than twice is found to be statistically relevent (*BiP* < 0.20, 80% confidence). Although numerous of the fragments occurring only twice are "technically" significant (*BiP* = 0.11), this is undoubtedly more an arifact of the depleted and highly skewed data pool than of any meaningful link with activity. For this reason, the methodology ignores only singly and doubly occurring fragments. This decision

Table 7(a)

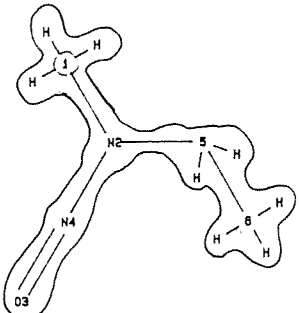
Compounds containing descriptor [4 2 4]. The atom of **SEV** origin is circled and the complete fragment is outlined. Activity is in brackets.



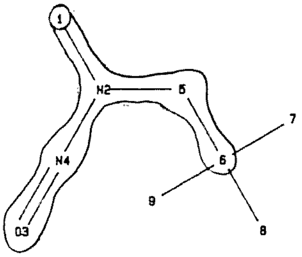
(55) DIMETHYL-NITROSAMINE



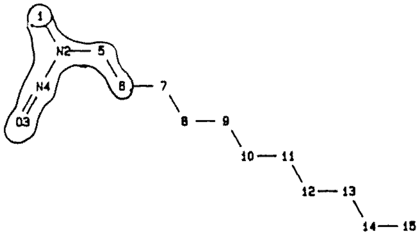
(55) NITROSO-METHYL-2-PHENYL-ETHYLAMINE



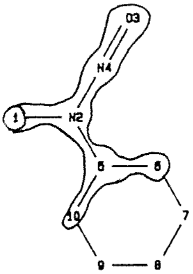
(39) NITROSO-METHYL-ETHYLAMINE



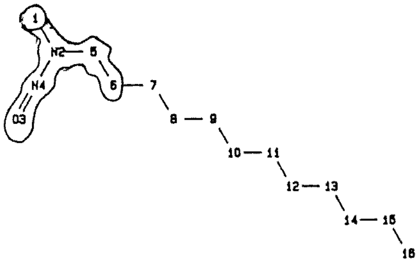
(55) NITROSO-METHYL-NEOPENTYLAMINE



(45) NITROSO-METHYL-UNDECYLAMINE



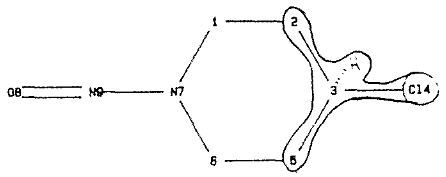
(45) NITROSO-METHYL-CYCLOHEXYLAMINE



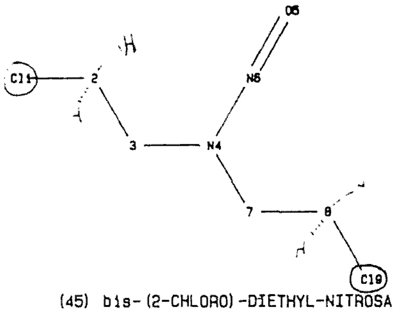
(45) NITROSO-METHYL-DODECYLAMINE

Table 7(b)

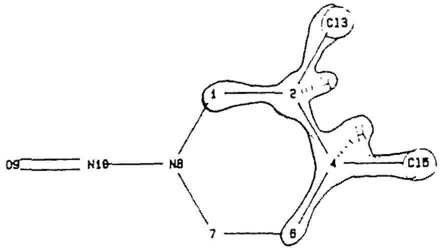
Compounds containing descriptor [1 3]. The atom of **SEV** origin is circled and the complete fragment is outlined. Activity is in brackets.



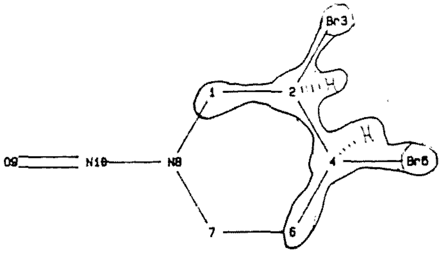
(55) 4-CHLORO-NITROSO-PIPERIDINE



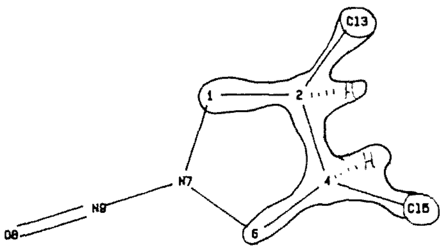
(45) bis-(2-CHLORO)-DIETHYL-NITROSAMINE



(55) 3, 4-DICHLORO-NITROSO-PIPERIDINE



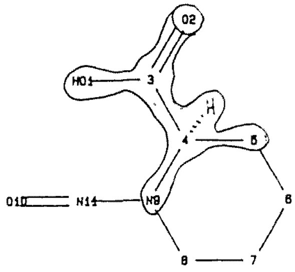
(55) 3, 4-DIBROMO-NITROSO-PIPERIDINE



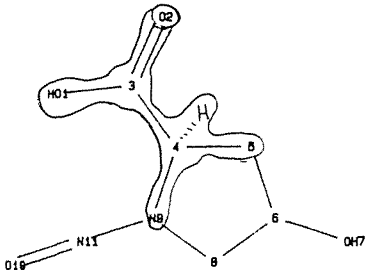
(55) 3, 4-DICHLORO-NITROSO-PYRROLIDINE

Table 7(c)

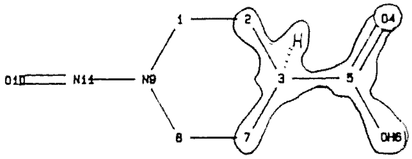
Compounds containing descriptor [1 2 4]. The atom of **SEV** origin is circled and the complete fragment is outlined. Activity is in brackets.



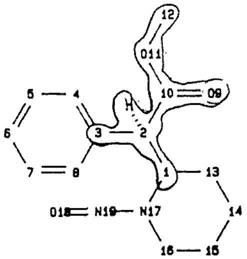
(10) 2-CARBOXY-NITROSO-PIPERIDINE



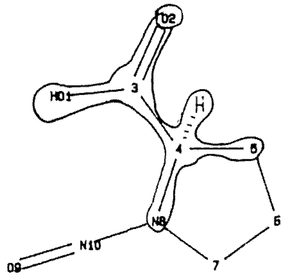
(10) 2-CARBOXY-4-HYDROXY-NITROSO-PYRROLI



(10) 4-CARBOXY-NITROSO-PIPERIDINE



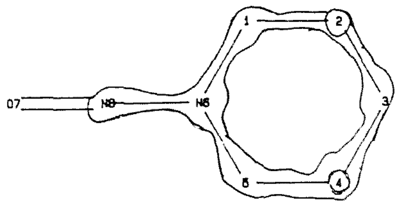
(10) NITROSO-METHYLPHENIDATE



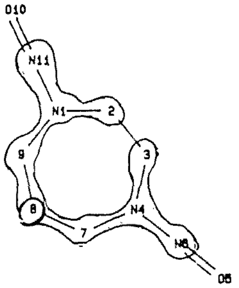
(10) 2-CARBOXY-NITROSO-PYRROLIDINE

Table 7(d)

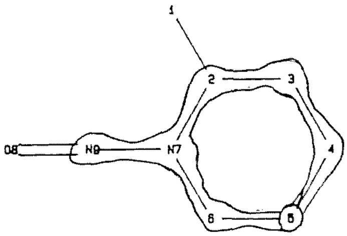
Compounds containing descriptor [4 6 4]. The atom of **SEV** origin is circled and the complete fragment is outlined. Activity is in brackets.



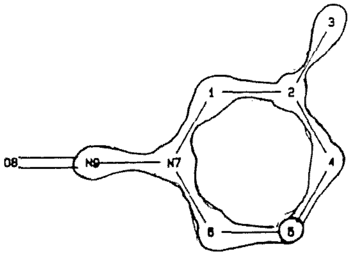
(45) NITROSO-PIPERIDINE



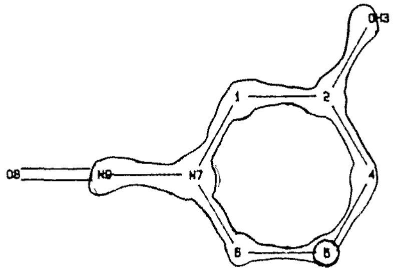
(55) DINITROSO-HOMOPIPERAZINE



(39) 2-METHYL-NITROSO-PIPERIDINE



(45) 3-METHYL-NITROSO-PIPERIDINE



(45) 3-HYDROXY-NITROSO-PIPERIDINE

is also based on the justifiable criticism [16] that many QSAR techniques lose credibility because they "overfit" the data by selecting too many parameters (descriptors) with respect to the number of data points (compounds). Therefore, at this final stage of the analysis, we attempt to extract what useful information there is remaining by isolating fragments common to all the remaining compounds and associating these descriptors with their average activity. Obviously, the fragments common to the entire data set (see expression (3)) are also common to the remaining data. The concluding statement is:

$$[1\ 1\ 2] \wedge [2\ 2] \quad \rightarrow \quad 33.8 \pm 18.0. \tag{7}$$

Alternatively stated, in the absence of any other information (i.e. none of the conditions of statements (3)–(6) are met by the test compound), the presence of the nitroso group alone ([1 1 2] ^ [2 2]) is sufficient for the active label and the numerical prediction of 33.8 ± 18.0.

The chemical interpretation of the SEV descriptors of statements (3)–(6) is given in table 6, and the compounds containing these descriptors are shown in tables 7(a)–7(d) with the specific active/inactive substructural features appropriately outlined.

6. Conclusion

A new graph theory-based methodology for use in SAR studies has been introduced and explained in detail. Starting with the input data that consists only of molecular activity and connectivity, submolecular descriptors are generated and statistically evaluated, with the final result being a collection of heuristic logical statements which relate chemical substructures with the activities that they presumably produce. Although the statistical treatment is based on qualitative (binary) activity data, frequently semi-quantitative results are obtained. A more rigorous QSAR for the development of "potency parameters" is currently being studied. In addition to its usefulness in probing the structure/activity link for compounds of known activity, the method is also capable of predicting the activity of untested compounds if provided with a training set from which it can "learn" the appropriate set of descriptors for the particular activity of interest.

A rather unique feature of this methodology is the ease with which the graph-theoretical descriptors are translatable into chemically meaningful submolecular fragments. In numerous other techniques, such a direct correspondence is not at all possible. Therefore, in addition to its use as a "black box" for correlating numerical descriptors with activity, this method offers the even more interesting possibility of relating specific chemical entities to activity.

Acknowledgements

We wish to thank the Pittsburgh Supercomputing Center (PSC) for the use of their CRAY X-MP/48. We also thank Mario Dimayuga for the development of computer graphics. We gratefull acknowledge N.I.H. for a fellowship, which helped to support this work.

References

[1] L. Kier and L. Hall, *Molecular Connectivity in Chemistry and Drug Research* (Academic Press, New York, 1976).

[2] N. Trinajstić, *Chemical Graph Theory*, Vols. 1 and 2 (CRC Press, 1983).

[3] M. Randić, J. Amer. Chem. Soc. 97(1975)6609.

[4] L. Kier, J. Pharm. Sci. 69(1980)807.

[5] G. Klopman, C. Raychaudhury and R. Henderson, Math. Comput. Modelling 2(1988)635.

[6] S. Grossman, B. Dzonova and M. Randić, Int. J. Quant. Chem. 12(1986)123.

[7] W.R. Muller, K. Szymanski, K. Knop and N. Trinajstić, J. Comput. Chem. 8(1987)170.

[8] G. Schafer, Statist. Sci. 2(1987)3.

[9] G. Klopman, J. Amer. Chem. Soc. 106(1984)7315.

[10] Laplace, *A Philosophical Essay on Probabilities* (Dover Publ., New York, 1951).

[11] D.E. Bailey, *Probability and Statistics – Models for Research* (Wiley, New York, 1971).

[12] T.H. Lowry and K.S. Richardson, *Mechanism and Theory in Organic Chemistry* (Harper and Row, New York, 1976).

[13] G. Klopman and M. McGonigal, J. Chem. Inf. Comput. Sci. 21(1981)48.

[14] J.S. Wishnok, ACS Symp. Ser. Amer. Chem. Soc., Washington, DC (1981), pp. 77–87.

[15] W. Lijinsky and H.W. Taylor, Cancer Res. 36(1976)1988.

[16] S. Wold and W.J. Dunn, J. Chem. Inf. Comput. Sci. 23(1983)6.