INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

U·M·I

University Microfilms International
A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
313/761-4700 800/521-0600

Order Number 9220024

**A graph theory-based 'expert system' methodology for
structure-activity studies**

Henderson, Robert Vann, Ph.D.

Case Western Reserve University, 1992

Copyright ©1991 by Henderson, Robert Vann. All rights reserved.

U·M·I

**300 N. Zeeb Rd.
Ann Arbor, MI 48106**

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

by

ROBERT V. HENDERSON

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

Department of Chemistry
CASE WESTERN RESERVE UNIVERSITY

January, 1992

Copyright c 1991 by

Robert Vann Henderson

CASE WESTERN RESERVE UNIVERSITY

GRADUATE STUDIES

We hereby approve the thesis of

Robert V. Henderson

candidate for the Ph. D.

degree.*

Signed: A. B. C. L. L. L.
(Chairman)

F. L. Urbach

[Signature]

[Signature]

Date 10/10/91

*We also certify that written approval has
been obtained for any proprietary material
contained therein.

Abstract

by

ROBERT V. HENDERSON

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 in order to reveal embedded substructural features, pharmacophores, which may play an important role in dictating the activity of a molecule. Common topological features among the subgraphs are represented as distance-based vector descriptors. These descriptors are then statistically ordered and a set of hierarchical rules is developed that can be used to explain the activities of the analyzed compounds as well as predict the activities of new compounds. Both qualitative and semi-quantitative conclusions regarding the structure-activity relationship are possible, depending on the nature of the original data. The validity of the methodology is demonstrated by its application to experimental data. The analysis of four vastly different data bases, and the results obtained, are presented in detail. In addition, formal validation is presented which involves randomization of the original data sets followed by the application of the t-statistic in order to establish the statistical significance of the descriptors employed by the methodology.

Dedication

This work, be it of any value, is dedicated entirely to Eva M. and Robert V. Henderson Sr. whose love and support over the years I cannot express in words.

Acknowledgements

I wish to thank every one of my research colleagues, past and recent, for the pleasure of working with them. In addition to their research talents, they were an exceptional group of individuals who constantly demonstrated concern for each other.

I would like to offer a special thanks to Shaomeng Wang and Sanjay Srivastava for their help, often late at night, in the preparation and printing of the final manuscript. Their caring and encouragement made this dissertation a reality. A special thanks also to Mario Dimayuga for his invaluable help and occasionally performed miracles in data recovery from whacky disks.

I thank former colleagues Ruben Venegas and Istvan Kolossvary for their continued friendship and encouragement.

Thanks also to the following individuals for their specific contributions:

Chandan Raychaudhury for collaboration on 'vertex indices'

Orest Macina for assemblage of the phenylalkylamine data base

Rajiv Bendale for assemblage of the peptide data base

Rebecca Zaniwski for help in preparation of the 'randomized' data

Mario Dimayuga (again?) for the printing of structural formulas

Abstract	ii
Dedication	iii
Acknowledgements	iv
1. Introduction	
1.1 QSAR Historical Development	2
1.2 Chemical Graph Theory	7
References	14
2. Vertex Indices - An Information Theory Approach to SAR using a Distance-based Topological Index	
2.1 Introduction	18
2.2 Computation of Vertex Index	20
2.3 Application of Methodology	23
2.4 Data	25
2.5 Results	26
2.6 Conclusion	28
References	36
3. A Graph Theory-based Expert System Methodology for Structure-Activity Studies	
3.1 Introduction	39
3.2 Graph Theoretical Representation of Molecular Structure .	41
3.3 Expert System Analysis	46
3.4 Statistical Parameters	48
3.5 Substructural Descriptors and Data Treatment	
3.5.1 SEV descriptors	52
3.5.2 Common and 'activity-specific' descriptors	58
3.5.3 Descriptor selection	60
References	65
4. Application of Expert System Methodology to the Analysis of Experimental Data	
4.1 Organic Acids and Nonacids - pKa data	67
References	90
4.2 N-nitroso Compounds - carcinogenic activity data	91
References	113
4.3 Phenylalkylamines - hallucinogenic activity	114
References	137
4.4 Peptides - thermolysin enzyme inhibition	138
References	164
5. Validation of Expert System Methodology	
5.1 Introduction	166
5.2 Real/random distribution profiles	169
5.3 Calculation of t-statistic	172
5.4 Conclusion	176
References	223
Bibliography	224

Chapter 1

INTRODUCTION

1.1 QSAR Historical Development

Historically, the first structure - activity relationship (SAR) is attributed to Crum-Brown and Fraser¹ who in 1868, after studying the biological effect of a series of strychnine derivatives, proposed that their physiological activity was dependent upon their structure. Physiological or biological activity was taken to mean potency and was defined as the reciprocal of concentration required to elicit a particular response whereas "structure" - of which little was known at this time - consisted basically of elemental composition. The hypothesis was the existence of a semi-quantitative mathematical relationship between physiological activity and chemical structure. Support of this hypothesis came the following year in the work of Richardson² which showed that the severity of narcosis induced by primary alcohols was proportional to their molecular weight. Richet³ provided further verification in 1893 by reporting an inverse relationship between the toxicity and aqueous solubility of various simple organic compounds.

By the beginning of the twentieth century, a correlation between narcosis and the chain length of selected organic compounds had been established. Since the chain length of organic compounds was closely related to their oil/water partition coefficients, Overton⁴ proposed that narcotic potency was a consequence of lipid solubility at the cellular level. This work was a significant contribution in QSAR since it revealed an important connection between drug activity and the transport system within a body. The hypothesis was that the delivery of the drug was regulated by its solubility at specific lipid/aqueous

interfaces. Furthermore, the use of the partition coefficient (P), an experimentally measurable physical property, allowed the establishment of a quantitative linear relationship with narcotic potency⁵, and numerous other biological properties in the following years⁶.

$$\log A = a \log P + b \quad (1)$$

where

A = activity (narcotic potency)
P = partition coefficient
a,b = empirically determined constants

Throughout the 1930s and 40s, considerable advances were made by various researchers which incorporated thermodynamic, molecular size and shape, and ionization descriptors into the correlation with a wide variety of different biological properties⁷. A particularly important advance was made by Albert⁸ who correlated the antibacterial activity (and inactivity) of more than 100 acridines, quinolines and pyridines with molecular charge and planarity. From this successful study, predictions were made which produced potent derivatives from some of the formerly inactive parent compounds. This work is still relevant today in that many of these same compounds, in more recent studies⁹, have been found to exhibit antitumor activity as well. With the availability of several different structure-based parameters, more accurate correlations with activity were obtained. Exemplary results in this direction are found in the 1942 collaboration of Bell and Roblin¹⁰ where the E.coli inhibitory potency of sulfonamides was plotted as a function of pK_a to reveal a parabolic curve of the form :

$$A = - (pK_a - a)^2 + b \quad (2)$$

A=activity, expressed as $\log(1/C)$, [C=minimal inhibitory concentration]
a and b are the 'x' and 'y' coordinates of the maxima of the curve

In addition to providing useful quantitative results, this relationship established an important link between biological activity and the electronic effect of substituents (electronic releasing or withdrawing) as expressed by pK_a values. During this same period, physical chemists were making significant contributions that would also greatly advance progress in QSAR. Hammett's contribution¹¹ defined substituent constants σ and ρ which together helped quantitatively characterize the electronic properties of many organic functionalities. Taft¹² offered modifications with the addition of steric and inductive substituent parameters.

In 1962 Hansch¹³ and coworkers incorporated these Hammett-Taft physicochemical parameters, along with a lipophilic (solubility) term, into several multi-parameter equation¹⁴ based on the additivity of these effects, partially justified by a linear free energy argument.

$$\log (1/C) = c_1 \log P - c_2 (\log P)^2 + c_3 \sigma + c_4 \quad (3)$$

where

C = physiological concentration (moles/liter)

P = octanol/water partition coefficient

σ = Hammett electronic parameter

c_i = least squares multiple-regression coefficient

Equation(3) is typical of the form used to obtain correlations between biological activity and physicochemical variables. Although intuitive in its origin, a theoretical derivation¹⁵ has been given based on a kinetic model, represented in a simplified form by equation (4).

$$dA/dt = k \text{Pr}(p) [C] \quad (4)$$

A = activity (biological response)
 k = rate constant
 Pr = probability of chemical reaching reaction site
 (a gaussian function of the partition coefficient)
 C = extra-cellular concentration (moles/liter)

However, the validity of any QSAR equation is ultimately to be judged by its ability to successfully predict the activity of compounds not included in the original data set. Through the empirical success of equation (3) in predicting many different types of physiological activity, the Hansch method gained universal acceptance and continues to be as a standard by which other methodologies are compared.

Shortly after the introduction of the Hansch model, Free and Wilson¹⁶ proposed a model of their own which was also based on the 'additivity' of substituent effects, but did not require any physicochemical parameters.

$$A = a + \sum b_{ij} S_{ij} \quad (5)$$

A = biological activity, usually $\log(1/C)$
 $S_{ij} = 1$, if substituent i is present at position j
 0, if " " not " " " "
 b_{ij} = multiple-regression coefficients
 a = average activity for series of compounds

Coefficient b_{ij} is interpreted as the contribution of substituent S_{ij} to the biological activity. Although having the benefit of being free of physicochemical parameters, this methodology is generally limited in its application to only congeneric data sets. However, if the definition of S_{ij} is broadened to include common substructures, rather than position-specific substituents, noncongeneric data may be analyzed.

The Hansch and Free-Wilson methodologies of the mid-1960s are considered to be the cornerstones of modern QSAR. These techniques are still in widespread use and have also influenced the development of various other models¹⁷ that have followed. In fact, numerous contemporary approaches can be viewed as modifications or even hybrids of both methods.¹⁸ Common to almost all of the currently employed methodologies is their dependence on computer implementation - specifically the use of multiple linear regression or discriminant analysis as a technique for optimizing equation parameters. Thus, the technological advances in the development of both hardware and software since the 1960s have been a great asset to the structure-activity practitioner.

In addition to electronic and hydrophobic parameters, steric,¹⁹ various physicochemical,²⁰ molecular orbital,²¹ molecular shape,²² and topological²³ descriptors have all been included as independent variables in numerous QSAR studies. At this point, focus will be placed on the development of topological descriptors and their application in structure-activity relationships.

1.2 Chemical Graph Theory

Graph theory was first introduced in 1736 by Leonhard Euler²⁴, a Swiss mathematician whose interest was in the systematic solution of some of the classical mathematics problems of the day. Today, graph theory is an established branch of mathematics dealing with the way in which points, or objects, are connected to each other²⁵. Thus, chemical graph theory is the study and subsequent representation of the way in which the atoms of molecules are connected. This area of study has expanded over the years and now involves the use of other related disciplines such as matrix theory, probability, numerical analysis, and topology. The combined application of these fields to the solution of chemical problems represents an interesting and exciting prospect for future development. Previous successful applications of chemical graph theory make one optimistic - yet current challenges provide the incentive to further explore its use in new and different ways. Such further exploration is the focus of this thesis.

The formal introduction of graph theoretical methods to chemistry can be traced to the mid-nineteenth century in the independent works of two researchers, Brown and Cayley. Brown²⁶ defined a graphic system for use in representing molecular compounds and Cayley²⁷ introduced an algorithm for calculating the number of structural isomers for a given alkane series by using graph representations and applying the mathematics of combinatorics. In fact, the entire development of organic chemistry has been aided by the implementation of graph-based

systems for the two dimensional representation of molecules. It is through such applications as these that the chemical graph became a means of defining molecular structure simply by showing the internal connectivity of the atoms. Thus, the graph used was completely compatible with and complementary to the valence bond theory.

This convenient method - where lines represent electron pairs (covalent bonds), points of intersection designate carbon atoms, and hydrogens are implied by valence - initially implied no mathematically quantitative interpretations in terms of physical or chemical properties. However, with the advent of quantum mechanics came the interest in generating numerical data which could characterize the molecule (graph) and its atoms (vertices). One example of a consequence of this interest is the calculation of orbital energies directly from the graph, wherein the Huckel matrix $[\mathbf{H}-\mathbf{ES}]$ is equivalent to a modified adjacency matrix - in graph theoretical terms. The determinant of this 'modified' adjacency matrix is the characteristic polynomial²⁸ of the graph. Thus, it can be shown that:

$$\begin{aligned} [\mathbf{H}] &= \alpha[\mathbf{I}] + \beta[\mathbf{A}] \\ [\mathbf{S}] &= [\mathbf{I}] \end{aligned} \quad (1)$$

where $\mathbf{H}$ = Hamiltonian matrix
 $\mathbf{S}$ = overlap matrix
 $\mathbf{I}$ = identity matrix
 $\mathbf{A}$ = adjacency matrix
 α = Coulomb integral
 β = resonance integral

According to MO theory, the Huckel (secular) determinant must vanish :

$$\det | \mathbf{H} - \epsilon \mathbf{S} | = 0 \quad (2)$$

which produces a set of eigenfunctions (MO's) and their eigenvalues

(energies ϵ_i). Since $\mathbf{H}$ and $\mathbf{A}$ commute, ie. $[\mathbf{H}, \mathbf{A}] = 0$, the eigenvectors of the adjacency matrix $\mathbf{A}$ are the same as the Hückel molecular orbitals. Therefore, HMO's are obtainable from graph theoretical arguments since they are topological orbitals dependent only on the connectivity.

One of the first topological indexes to be quantitatively related to chemical and physical properties is the Wiener index. In 1947, in the interest of quantifying the structural features of paraffins, Wiener²⁹ introduced a graph-based measure of "compactness"; defined as the sum of the distances between all pairs of carbon atoms within the molecule. This can be expressed simply as

$$W = 1/2 \sum d_{ij} \quad (3)$$

wherein d_{ij} represents the topological distance (shortest path) between two atoms, i and j , and the summation is carried-out over every pair of atoms in the molecule. In general, the more compact (highly branched) a molecule, the smaller the W number. The physical property most closely associated with the index is molecular volume.³⁰

With just two topological indices Wiener was able to obtain an equation which was used to calculate the boiling points of 94 paraffins (up through the decane series) with an average deviation of less than one percent. Correlations of this newly formed index with other physical properties, such as heat of vaporization³¹ and surface tension³², followed shortly. Although the relationships between physical property and molecular topology were restricted to a rather limited class of acyclic alkanes, this work nonetheless pioneered the way for the use of chemical graph theory in structure-property and structure-activity studies. More recent applications of the Wiener

index include correlations with such diverse properties as chromatographic retention indices³³ and carcinogenic potencies of PAHs³⁴.

Milan Randic is a familiar name in the current literature of chemical graph theory since he has personally been responsible for numerous methods for numerically encoding the topology of organic molecules for use in molecular identification as well as in structure-property correlations. Since the mid-1970s, his work has introduced such concepts as binary labeling³⁵ (for compact molecular codes) and connectivity (chi) indices³⁶. As originally proposed, the chi indices

$$\chi = \sum (v_i v_j)^{-1/2} \quad (4)$$

were calculated by eq.(4) where v_i and v_j are the graph valences of atoms i and j with the summation performed over all pairs of adjacent atoms. The specific form of this expression was designed to characterize and distinguish different bond types within a chemical structure. The summation was then taken to represent the molecule as a function of the different bonds found within the molecule. Upon introduction of the index, Randic demonstrated its structural sensitivity through correlations with a number of properties such as enthalpy of formation and molecular surface area.³⁶ Other papers have followed showing significant correlations with molar refractivity, density, diamagnetic susceptibility, aqueous/lipid partitioning and solubility.³⁷ Similar to the Wiener index, chi values can also be interpreted as a measure of molecular branching, but are considered to be more sensitive to molecular shape than volume or size.³⁸

Kier and Hall³⁸ have greatly popularized the chi index by expanding the original definition to produce a 'set' of indexes which have been correlated with a wide variety of biological properties exhibited by molecules far more structurally diverse than the hydrocarbons initially investigated. Equations have been derived for biological activities as a function of chi index(es) through least squares regression techniques.³⁹ Once the equation has been developed from compounds of known activity, the activity of any other compound - even a speculative, yet currently unsynthesized molecule - can be readily calculated. Whether or not the calculated, or predicted, activity is a reasonable approximation for the actual (unknown) activity is a numerical issue of interpolation verses extrapolation and can also be addressed during the analysis. A few of the many biological activities successfully subjected to chi-analysis by various researchers are narcotic, hallucinogenic, and anesthetic potencies; mutagenicity and toxicity; antimicrobial, antitumor and antiviral activity; and the inhibition of various microorganisms, to mention a few.^{37,38}

The chi indices introduced by Randic and developed by Kier and Hall are by far the most frequently employed topological descriptors used in QSAR. One of the major features contributing to the success of this index is the relatively low number of degeneracies (different graphs having the same index value). Just as it is essential that a meaningful index produce similar values for structurally similar molecules, likewise it is undesirable that an index produce similar or identical values for structurally diverse molecules. One such index

designed for its ability to discriminate structures and minimize degeneracies is Balaban's J index⁴⁰ :

$$J = q / u \sum (D_i D_j)^{-0.5} \quad (5)$$

where

q = total number of edges (bonds)
u = number of cycles (rings) contained in graph
D_i = total of row i elements of distance matrix

It is noted that the operation of summation over the inverse square root of the product of adjacent invariants is borrowed from the Randic approach. Although this particular operation has been used in the formulation of other indexes as well, its development is empirical rather than through theoretical derivation and is justified by its resultant reduction in degeneracies. Although the J index has been shown to be one of the most discriminating indexes,⁴¹ its ability to correlate with chemical and biological properties has thus far not been investigated to the extent that the chi index has been studied.

Another index scheme that can be exploited for its discriminating property is derived from information theory⁴². In general, for any N elements distributed into separate classes (N₁, N₂, ... N_i ...) a probability distribution P{p₁, p₂, ... p_i ...} can be generated and the information content of such a distribution can be calculated by Shannon's formula⁴³ :

$$\bar{I} = - \sum p_i \log p_i \quad (6)$$

where

$\bar{I}$ = mean information content (entropy)
p_i = probability of an element belonging to class i
p_i = N_i / N

When applied to the molecular graph, equation (6) produces an entire family of indices depending on the subgraph invariant chosen for representation in the frequency distribution. The logarithm is usually taken as base 2 thereby expressing the information content in 'bits'.

Various information indices have been explored in some detail by Bonchev and Trinajstić.⁴⁴ Of particular interest and to be developed further in this thesis is the use of the distance matrix⁴⁵ as the source of the elements which can be partitioned and converted to probabilities for use in equation (6). Such a procedure has previously been employed for the calculation of molecular information indices - a single index value calculated to represent the entire molecule.⁴⁶ A somewhat unique procedure will be described which uses equation (6) to calculate atomic information indices, thereby numerically characterizing the structural environment (with respect to distance distributions) of each atom within a molecule.⁴⁷ The hypothesis is that such a scheme would allow biological properties to be correlated with frequently occurring substructural features.

References

1. Crum-Brown, A. and Fraser, T.R., Trans. Royal Soc. Edinburgh, **25**, 151 (1868).
2. Richardson, B.J., Medical Times and Gazette, **2**, 703 (1869).
3. Richet, C.R., Seances Soc. Biol. Ses Fil., **9**, 775 (1893).
4. Overton, E., Vierteljahresschr. Naturforsch. Ges. Zuerich, **44**, 88 (1899).
5. Dunn, W.J., "Annual Reports in Medicinal Chemistry", Vol. 8, Quantitative Structure-Activity Relationships, R.V. Heinzelman (Ed.), Academic, New York, 1973.
6. Hansch, C. and Dunn, W.J., J. Pharm. Sci., **61**, 1 (1972).
7. Tute, M.S., "Comprehensive Medicinal Chemistry", Vol. 4, Quantitative Drug Design, C.A. Ramsden (Ed.), Pergamon, Oxford, 1990.
8. Albert, A., Rubbo, S.D. and Burvill, M.I., Br. J. Exp. Pathol., **30**, 159 (1949).
9. Denny, W.A., Cain, B.F., Atwell, G.J., Hansch, C., Panthananicakal, A. and Leo, A., J. Med. Chem., **25**, 276 (1982).
10. Bell, P.H. and Roblin, R.O., J. Am. Chem. Soc., **64**, 2905 (1942).
11. Hammett, L.P., "Physical Organic Chemistry", McGraw-Hill, New York, 1970.
12. Taft, R.W., "Steric Effects in Organic Chemistry", M.S. Newman (Ed.), Wiley, New York, 1956.
13. Hansch, C., Maloney, P.P., Fujita, T. and Muir, R.M., Nature (London), **194**, 178 (1962).
14. Hansch, C., Acc. Chem. Res., **2**, 232 (1969).
15. Penniston, J.T., Beckett, L., Bentley, D.L. and Hansch, C., Mol. Pharmacol., **5**, 333 (1969).
16. Free, S.M. and Wilson, J.W., J. Med. Chem., **7**, 395 (1964).
17. Martin, Y.C., "Quantitative Drug Design", Dekker, New York, 1978.
18. Martin, Y.C., J. Med. Chem., **24**, 229 (1981).

19. Verloop, A., Hoogenstraaten, W. and Tipker, J., "Drug Design", Vol. 7, E.J. Ariens (Ed.), Academic, New York, 1976.
20. Franke, R., "Theoretical Drug Design Methods", Elsevier, Amsterdam, 1984.
21. Kollman, P., Acc. Chem. Res., **18**, 105 (1985).
22. Mabilia, M., Pearlstein, R.A. and Hopfinger, A.J., "Topics in Molecular Pharmacology", Vol. 3, A. Burgen, G. Roberts and M. Tute (Eds.), Elsevier, Amsterdam, 1986.
23. Kier, L.B. and Hall, L.H., "Molecular Connectivity in Structure - Activity Analysis", Research Studies Press, Letchworth, 1986.
24. Neuman, J.R., Scientific American, **189**, 66 (1953).
25. Wilson, R.J., "Introduction to Graph Theory", Oliver and Boyd, Edinburgh, 1972.
26. Biggs, N., Lloyd, E. and Wilson, R., "Graph Theory 1736-1936", Clarendon, Oxford, 1976.
27. Cayley, A., Phil. Mag., **13**, 172, (1857).
28. Trinajstić, N., "Chemical Graph Theory", Vol. 1, CRC, Boca Raton, 1983.
29. Wiener, H., J. Am. Chem. Soc., **69**, 17 (1947).
30. Rouvray, D.H., Sci. Am., **255**, 40 (1986).
31. Wiener, H., J. Am. Chem. Soc., **69**, 2636, (1947).
32. Wiener, H., J. Phys. Chem., **52**, 1082, (1948).
33. Bonchev, D., Mekenjan, P., Protic, G. and Trinajstić, N., J. Chromatogr., **176**, 149 (1979).
34. Seybold, P.G., Int. J. Quantum Chem., **10**, 95 (1983).
35. Randić, M., J. Chem. Phys., **60**, 3920 (1974);
J. Chem. Inf. Comput. Sci., **26**, 134-148, (1986).
36. Randić, M., J. Am. Chem. Soc., **97**, 6609 (1975).
37. Kier, L.B. and Hall, L.H., "Molecular Connectivity in Chemistry and Drug Research", Academic, New York, 1976.
38. Kier, L.B. and Hall, L.H., "Molecular Connectivity in Structure - Activity Analysis", Research Studies Press, Letchworth, 1986.

39. Kleinbaum, D. and Kupper, L., "Applied Regression Analysis and Other Multivariable Methods", Duxbury, Belmont, 1978.
40. Balaban, A.T., Chem. Phys. Lett., **89**, 399 (1982).
41. Balaban, A.T., J. Chem. Inf. Comput. Sci., **25**, 334 (1985).
42. Rashevsky, N., Bull. Math. Biophysics, **17**, 229, (1955);
Bull. Math. Biophysics, **16**, 317, (1954).
43. Shannon, C. and Weaver, W., "Mathematical Theory of Communication",
University of Illinois, Urbana, 1949.
44. Bonchev, D. and Trinajstić, N., Int. J. Quantum Chem., **16**, 463
(1982).
45. Harary, F., "Graph Theory", Addison Wesley, Reading, 1971.
46. Bonchev, D. and Trinajstić, N., J. Chem. Phys., **67**, 4517 (1977).
47. Klopman, G., Raychaudhury, C. and Henderson, R.V., Math. Comput.
Modell., **11**, 635 (1988).

Chapter II

VERTEX INDICES - AN INFORMATION THEORY APPROACH TO SAR USING A DISTANCE-BASED TOPOLOGICAL INDEX

2.1 Introduction

The assumption that chemical compounds must possess some structural specificity in order to exhibit certain biological activity has led some researchers to propose that topologically similar structures could have similar biological activities.¹ The aim of the present approach is to use graph theory to relate the structural components of the molecular graph to the resulting biological activity. The unique aspect of this approach is in the use of 'vertex (atomic) indices' as opposed to traditional approaches which emphasize 'molecular indices'.² The motivation for this alternate method is in its ability to recognize and discriminate embedded substructural features that may be relevant to biological activity. Such features generally go undetected when employing a single molecular index.

In this study, molecular topology is represented by a set of indices - one index for each non-hydrogen atom in the molecule. Each vertex index carries information about the specific structural environment of a given atom within the molecule and the collection of these indices characterize the structure of the entire molecule. Application of an appropriate index should allow the atoms of even diversely different molecules to be classified together when found in structurally similar environments and should discriminate atoms found in dissimilar structural surroundings.

The methodology is implemented by using an information theoretical distance-based topological index for each vertex of the

molecular graph as computed by the distance of a given vertex from all the other vertices of the graph.³ This computation is performed for each atom of every molecule in the data base. The vertex indices are then ordered so as to group atoms together according to their topological environments. The underlying hypothesis is that similar structural environments produce similar vertex indices, $I(v)$, which can then be correlated with similar biological activities.

In the initial investigation offered here, the described vertex index has been used for the qualitative classification (active/inactive) of hallucinogenic potency of 44 phenylalkylamines. The $I(v)$ index was calculated for all non-hydrogen atoms of each molecular graph and these values were then used to establish correlations with activity and ultimately to classify the molecules as either active or inactive with respect to hallucinogenic potency.

2.2 Computation of Vertex Index I(v)

A molecular graph (G) is derived from a structural formula by replacing the atoms with vertices (points) and the sigma bonds with edges (lines). Such a transformation is illustrated in the figures below. Although pi bonds are not explicitly included in the graph, they are represented implicitly by the retention of hydrogens in the molecular graph, as shown in the graphs of benzene and cyclohexane.

A graph (G) can be defined as an ordered pair composed of a two sets
- a set of vertices and a set of edges.⁴

$$G = [V(G), E(G)]$$

A path within G is defined as a sequence of incident vertices that are distinct. The length of such a path is simply a count of the number of edges connecting the vertices. The distance between two vertices is the length of the shortest path between the vertices. (There is often more than one path between vertices in molecular graphs). For computational convenience, the distances between all

pairs of vertices within a graph is arranged in a matrix - the distance matrix $D(G)$ composed of elements d_{ij} .

$$\begin{aligned} d_{ij} &= \text{length of shortest path connecting atoms } i \text{ and } j \\ d_{ii} &= 0 \end{aligned}$$

$D(G)$ is symmetrical about the main diagonal which consists of zeros.

The index calculated for each atom, as described above, is that of information content. This index can be formulated by considering the partitioning of distances associated with atom/vertex i ,

$$d(v_i) = D_i(d_{i1}, d_{i2}, \dots, d_{in})$$

where d_{ij} is the j th element of row i of the distance matrix and D_i is the sum over the elements of row i for a graph of n vertices. Using the elements of the distance matrix as the structural descriptors, Shannon's formula⁵ takes the following form:

$$I(v_i) = \sum_{j \neq i}^n d_{ij}/D_i \text{ lb } (D_i/d_{ij}) \quad (1)$$

where

$$D_i = \sum_{j=1}^n d_{ij}$$

and

$$\begin{aligned} I(v_i) &= \text{mean information content of atom } i \\ d_{ij} &= \text{topological distance between atoms } i \text{ and } j \\ D_i &= \text{sum of the distance elements of the } i\text{th row} \\ n &= \text{number of atoms in the molecule} \\ \text{lb} &= \log_2 \end{aligned}$$

For efficient computations, equation (1) can be rearranged into the following form:

$$I(v_i) = \ln D_i / \ln 2 - 1/(D_i \ln 2) \sum_{j \neq i}^n d_{ij} \ln d_{ij} \quad (2)$$

In terms of computer implementation, the molecular graph is taken as the sigma bond skeleton in which each atom (including hydrogens) is represented by a vertex and each o bond is represented as an edge. This molecular graph is entered into the computer program as a linear code from which the adjacency matrix is generated. The adjacency matrix is then transformed into a distance matrix. From the elements of the distance matrix, equation (2) is used to calculate vertex indices for each nonhydrogen atom of every molecule in the data base. (See figure 1 for a sample calculation) The described procedure is carried out entirely by a computer program (CPATH).

2.3 Application of Methodology

Typical of most QSAR techniques, this methodology requires a learning or training data set consisting of molecules for which the biological activity is already known. This data is needed in order to learn the pertinent parameters that are relevant to activity/inactivity. The described vertex index is then calculated for each non-hydrogen atom of all the molecules of the learning data set. The I_v indices are then arranged on a number line in increasing order. Each index value also carries an activity label (active or inactive) which is the same as the activity label on the molecule of origin. Ranges of values on the number line associated solely with either active or inactive molecules are then identified and retained for the prediction of the activity of molecules not included in the original training set. The success of the methodology rests on the hypothesis that the atoms of indices falling in these active and inactive 'ranges' share common structural environments that are relevant to the activity of their respective molecules.

The following criteria are adopted for the selection of a range:

a) Three or more consecutive values contributed exclusively by active molecules define an active range on the $I(v)$ number line. Alternatively, three or more consecutive values contributed exclusively by inactive molecules define an inactive range.

b) Identical values originating from the same molecule only contribute a single entry on the number line. This is necessary because molecular symmetry produces topologically equivalent vertices which have the same $I(v)$ index. Obviously, it is not meaningful to

have a range formed from atoms all belonging to the same molecule.

c) $I(v)$ indices not falling in either an active or inactive range are not considered to be relevant to activity and play no role in activity predictions.

In table 1, the range (92-96) is composed of five indices contributed by four active molecules and is thereby labeled active. The range (123-128) includes six indices of which five are distinct and have come from five inactive molecules to form an inactive range. The non-range (141-148) contains indices that are dismissed as irrelevant to activity.

The activity of a compound is predicted on the basis of the occurrence of its vertex indices in active and inactive ranges. The compound is predicted to be active if the $I(v)$ value of all, or some of its atoms fall:

- (a) only in the active range(s), or
- (b) in both active and inactive ranges with the number of $I(v)$ values in the active range(s) being larger than that in the inactive range(s).

Otherwise the compound is predicted to be inactive.

The prediction of the activity of a compound submitted for analysis is illustrated in the table 2. The procedure described for activity predictions has also been automated and is carried-out entirely by computer.

2.4 Data

The biological activity of the compounds investigated in this study is hallucinogenic potency which is defined as the ratio of the effective dose of mescaline (3.75 mg/kg) to the effective dose of the experimental compound⁶. The effective dose is determined as the mean of the human threshold dose and the dosage that is maximally effective. The hallucinogenic potency is expressed in mescaline units (MU). For precision in the analysis, the potency has been converted to molar units and is taken as $\log(1/C)$ where C is the concentration in mmol/kg. The data has been collected from the literature.^{6,7} In order to obtain a rather even distribution of active and inactive molecules, compounds having hallucinogenic potency, $\log(1/C)$, larger than 2.60 were labeled as active. From a series of 44 compounds, a 'training set' of 30 compounds, 16 actives and 14 inactives, was formed. A 'test set' of 14 compounds, 7 actives and 7 inactives, was retained for the purpose of evaluating the success of the methodology in predicting the activity of compounds not included in the training set. The compounds of the training set along with their experimental and predicted activity (active = + , inactive = -) are shown in table 3 and those for the test set are given in table 4.

2.5 Results

As mentioned earlier, the compounds of the training set were used to elucidate the active and inactive ranges of index values believed to be useful for predictive purposes. As a preliminary test, these same compounds were then submitted for analysis to verify that the method is self-consistent by its ability to predict these compounds correctly. This procedure is referred to as a retrofit and does not validate a method but does serve as a minimum requirement for a successful analysis. As anticipated, all 30 compounds of the training set were classified correctly.

Fourteen compounds not included in the original training set were used as a test set. These compounds, for which the experimental activity is known, were then submitted to the program for evaluation. The purpose being to compare experimental and predicted activities. This procedure does play a role in validation since the data of the test set has not been previously encountered by the program and is therefore independent of the parameter optimization (range selection) process. Eleven of the fourteen compounds of the test set were predicted correctly - which included correct evaluation of all highly potent compounds. Of the three misclassified compounds, two were experimentally inactive and one was active. Seventy-eight percent correct classification; significantly better than the fifty percent correctness expected from random guessing, is judged to be supporting evidence for the validity of the proposed methodology. The χ^2 value (4.67) associated with the test set predictions verifies with a 95% confidence that this result is not due to chance.⁸

Although still far from conclusive, this initial study does provide optimism for the useful role of these topological parameters as descriptors of embedded substructure.

2.6 Conclusion

In addition to this initial investigation, this same methodology has since been applied to several other data bases and has produced encouraging results that have subsequently been published⁹.

A critical analysis of the methodology, however, does reveal some shortcomings that merit attention. Although the described active and inactive index ranges may be useful for the prediction of a compound's activity, there appears to be no way to use this information to 'design' a target compound for a specific activity in a chemically meaningful way. In fact, the numerical magnitude of an isolated index, or even an entire range of indices, reveals nothing about molecular structure in and of itself. That is, given an index value associated with one atom of a molecule, it is generally not possible to deduce from which atom the index originated. The complex nature of distance distributions makes this humanly impractical. This reduces the methodology to that of a black box which may be capable of correct activity predictions but reveals nothing about the specific structural features that played a role in those predictions.

Even more serious than the aforementioned concerns is the existence of evidence that refutes the original hypothesis, ie., very similar indices, even occurring within the same activity range, can arise from structurally diverse atomic environments. For example, in the phenylalkylamine analysis, an amine nitrogen, a methoxy oxygen and an aromatic carbon (each from differing regions of different molecules) all contribute indices to the same specific active range.

Even if the distance distributions about these three vertices is similar, clearly their chemical properties must differ. Such a range must therefore be viewed as meaningless and any correct predictions derived from it are merely due to chance. The current inability of the program to recognize and delete such meaningless ranges is a consequence of the arbitrarily devised rules of range selection and represents a serious inadequacy detracting from the viability of the methodology. Clearly a statistically sound procedure for weeding-out insignificant activity ranges is required.

The concerns expressed above do not represent insurmountable problems and do not challenge the use of the method as a black box activity predictor. They do, however, represent areas in which improvements can and must be made in order to make the analysis more robust. The very motivation for the vertex index was the observation that traditional molecular indexes compressed too much structural information into a single value. Although the $I(v)$ index offers an improvement in this situation, it too undergoes some compression of information which may be the source of some of the mentioned deficiencies. One possible solution is to segregate heteroatoms so that their indices are plotted on separate number lines. In this way only identical atom types could contribute to an activity range. This procedure would eliminate the meaningless range cited above. Another possibility is to introduce an additional topological criterion (a second index) which would lead to 'activity areas' (2-dimensional clusters) as opposed to the one-dimensional activity ranges.

Another alternative, which will be pursued in the next chapter, is to abandon the index approach completely and to work directly with the distance partitions. This offers the advantage of retaining more structural information about each vertex which would allow better discrimination of vertices. The distance partitions also permit greater chemical interpretation since at least partial graph reconstruction (regenerating the graph from its mathematical descriptors) is possible, thus allowing the direct display of structural features relevant to activity - an obstacle that could not be overcome with the index method. The only disadvantage to this alternative technique is the greater complexity introduced into the analysis since each vertex/atom is now associated with a vector (distance partition) as compared to a single value index. Correspondingly, each molecule is now described by a matrix rather than a vector. The details of this investigation are presented in the following chapter.

Figure 1

4-methoxyphenylethylamine and its vertex labeled molecular graph G_1 .

The vertices labeled 1-11 in G_1 correspond to the non-hydrogen atoms of the molecule and therefore each is associated with an information content vertex index, I_v , as shown for atom 1.

$$\begin{aligned}
 I(v_1) &= \ln 106 / \ln 2 - 1 / (106 * \ln 2) * [3*(1)\ln(1) + 3*(2)\ln(2) \\
 &\quad + 3*(3)\ln(3) + 2*(4)\ln(4) + 4*(5)\ln(5) + 3*(6)\ln(6) \\
 &\quad + 1*(7)\ln(7) + 1*(8)\ln(8) + 3*(9)\ln(9)] \\
 &= 4.2895 \text{ bits.}
 \end{aligned}$$

The I_v value of the vertices labeled 1-11 in G_1 are computed and listed below, to five significant figures.

$$\begin{aligned}
 I(v_1) &= 4.2895 \\
 I(v_2) &= 4.2606 \\
 I(v_3) &= 4.2926 \\
 I(v_4) &= 4.3483 \\
 I(v_5) &= 4.3908 \\
 I(v_6) &= 4.3768 \\
 I(v_7) &= 4.3186 \\
 I(v_8) &= 4.3768 \\
 I(v_9) &= 4.3908 \\
 I(v_{10}) &= 4.3130 \\
 I(v_{11}) &= 4.2907
 \end{aligned}$$

Table 1

Serial no.	I_v	Compound no.	Activity of compound	Range label
92	4.7269	8	+	Active
93	4.7280	22	+	
94	4.7281	3	+	
95	4.7297	22	+	
96	4.7339	9	+	
123	4.7622	28	-	Inactive
124	4.7627	1	-	
125	4.7633	28	-	
126	4.7637	16	-	
127	4.7676	23	-	
128	4.7676	29	-	
141	4.7835	22	+	Not a range
142	4.7839	1	-	
143	4.7849	22	+	
144	4.7903	29	-	
145	4.7950	3	+	
146	4.7950	28	-	
147	4.7952	29	-	
148	4.7960	3	+	

Table 1: A sample of 2 ranges and a non-range that result from the ordering of $I(v)$ values on a number line. The serial no. corresponds to the relative position of an I_v on the number line. Column four gives the activity label derived from that of the compound ('+' = active and '-' = inactive)

Table 2

Ser No.	I(v) Index	Falling in the range	Distribution of range	
			Act	Inact
1	4.5694	4.5503-4.5694	3	0
2	4.5503	4.5503-4.5694	3	0
3	4.5835	4.5780-4.6237	8	0
4	4.6347	Not falling in any range		
5	4.6089	4.5780-4.6237	8	0
6	4.6237	4.5780-4.6237	8	0
7	4.5951	4.5780-4.6237	8	0
8	4.5946	4.5780-4.6237	8	0
9	4.6405	4.6405-4.6523	6	0
10	4.5868	4.5780-4.6237	8	0
11	4.5658	4.5503-4.5694	3	0
12	4.6435	4.6405-4.6523	6	0
13	4.6018	4.5780-4.6237	8	0
14	4.5780	4.5780-4.6237	8	0

PREDICTION : The molecule is ACTIVE

Table 2: Activity prediction of 2,5-dimethoxy-4-bromophenylethylamine.

Training set: Substituted phenylalkylamines
along with their experimentally-determined
and predicted hallucinogenic activities

Compd no.	Substituents R	X	Activity*	
			Expt	Pred
1.	3,4,5-(OCH ₃) ₃	H	-	-
2.	4-OCH ₃	CH ₃	-	-
3.	2,5-(OCH ₃) ₂	CH ₃	+	+
4.	3,4,5-(OCH ₃) ₃	CH ₃	-	-
5.	2,4,5-(OCH ₃) ₃	CH ₃	+	+
6.	-3,4-(OCH ₂ O)	CH ₃	-	-
7.	3-OCH ₃ -4,5-(OCH ₂ O)	CH ₃	-	-
8.	2-OCH ₃ -4,5-(OCH ₂ O)	CH ₃	+	+
9.	2-OCH ₃ -3,4-(OCH ₂ O)	CH ₃	+	+
10.	2,5-(OCH ₃) ₂ -3,4-(OCH ₂ O)	CH ₃	+	+
11.	2,3,4,5-(OCH ₃) ₄	CH ₃	+	+
12.	2,5-(OCH ₃) ₂ -4-C ₂ H ₅	CH ₃	+	+
13.	2,5-(OCH ₃) ₂ -4-(n-C ₃ H ₇)	CH ₃	+	+
14.	2,5-(OCH ₃) ₂ -4-(n-C ₄ H ₉)	CH ₃	+	+
15.	2,3,4-(OCH ₃) ₃	CH ₃	-	-
16.	3-OCH ₃ -4,5-(OC ₂ H ₄ O)	CH ₃	-	-
17.	2,4-(OCH ₃) ₂ -5-OC ₂ H ₅	CH ₃	+	+
18.	3,4,5-(OCH ₃) ₃	C ₂ H ₅	-	-
19.	4-OCH ₃	H	-	-
20.	3,5-(OCH ₃) ₂ -4-OC ₂ H ₅	H	+	+
21.	3,5-(OCH ₃) ₂ -4-OC ₃ H ₇	H	-	-
22.	2,5-(OCH ₃) ₂ -4-CH ₃	H	+	+
23.	2,3,4-(OCH ₃) ₃	H	-	-
24.	2,5-(OCH ₃) ₂ -4-I	H	+	+
25.	2,5-(OCH ₃) ₂ -4-C ₂ H ₅	H	+	+
26.	2,3,5-(OCH ₃) ₃	CH ₃	-	-
27.	2,5-(OCH ₃) ₂ -4-(n-C ₅ H ₁₁)	CH ₃	+	+
28.	3,4-(OCH ₃) ₂	CH ₃	-	-
29.	2,4,5-(OCH ₃) ₃	H	-	-
30.	2,5-(OCH ₃) ₂ -4-(OC ₂ H ₅)	CH ₃	+	+

* Experimentally-determined potency in
log(1/C), see ref. 37

+ = active ; log(1/C) > 2.60

- = inactive ; log(1/C) < 2.60

Test set : Substituted phenylalkylamines
along with their experimentally-determined
and predicted hallucinogenic activities

Compd No.	Substituents		Activity*	
	R	X	Expt	Pred
1.	3-OCH ₃ -4,5-(OCH ₂ O)	H	-	-
2.	2,4-(OCH ₃) ₂	CH ₃	-	-
3.	2,3,6-(OCH ₃) ₃	CH ₃	+	+
4.	-2,3-(OCH ₂ O)-4-OCH ₃	CH ₃	-	+
5.	2,5-(OCH ₃) ₂ -4-CH ₃	CH ₃	+	+
6.	2,5-(OCH ₃) ₂ -4-Br	CH ₃	+	+
7.	2-OC ₂ H ₅ -4,5-(OCH ₃) ₂	CH ₃	+	+
8.	2-OCH ₃ -3,4-(OCH ₂ O)	H	-	-
9.	2,5-(OCH ₃) ₂ -4-Br	H	+	+
10.	3,5-(OCH ₃) ₂ -4-SCH ₃	H	+	-
11.	2,3-(OCH ₃) ₂ -4,5-(OCH ₂ O)	CH ₃	-	+
12.	3,4-(OCH ₃) ₂	H	-	-
13.	2,4,6-(OCH ₃) ₃	CH ₃	+	+
14.	-3,4-(OCH ₂ O)	H	-	-

* Experimentally-determined potency in

log(1/C), see ref. 37

+ = active ; log(1/C) > 2.60

- = inactive ; log(1/C) < 2.60

References

1. Norrington, F.E., Hyde, R.M., Williams, S.G. and Wooten, R., J. Med. Chem., **18**, 604 (1975).
2. Kier, L.B. and Hall, L.H., "Molecular Connectivity in Structure - Activity Analysis", Research Studies Press, Letchworth, 1986.
3. Klopman, G., Raychaudhury, C. and Henderson, R.V., Math. Comput. Modell., **11**, 635 (1988).
4. Trinajstić, N., "Chemical Graph Theory", Vol. 1, CRC, Boca Raton, 1983.
5. Shannon, C. and Weaver, W., "Mathematical Theory of Communication", University of Illinois, 1949.
6. Shulgin, A., Sargent, J. and Narenjo, C., Nature, **221**, 537 (1969).
7. Gupta, S.P., Singh, P. and Bindal, M.C., Chem. Rev., **83**, 633 (1983).
8. Baggaley, A.R., "Intermediate Correlational Methods", Wiley, New York, 1964.
9. Klopman, G. and Raychaudhury, C., J. Comput. Chem., **9**, 232 (1988); J. Chem. Inf. Comput. Sci., **30**, 12 (1990).

Chapter III

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

3.0 Abstract

A new 'expert system' approach employing graph theoretical descriptors is proposed for the statistical evaluation of the link between the structure and the chemical/biological activity of organic molecules. The computer program developed for the analysis serves as a 'learning machine' for the heuristic treatment of input data concerning the particular activity of interest. The molecules to be analyzed are represented as chemical graphs which are decomposed into their component subgraphs through manipulation of the topological distance matrix. Each subgraph is then statistically evaluated, by calculating its binomial cumulative probability and a newly introduced index of skewness, in order to isolate common substructural features that are correlated with activity. In this way, the selection of an appropriate set of topological descriptors is determined from the data itself. Upon complete evaluation of the input data, a set of hierarchical rules is developed that can be used to explain the activities of the previously submitted compounds and also to predict the activity of compounds of unknown activity. The validity of the methodology is demonstrated through its application to several different data types as well as a randomization study that will be presented in the chapters that follow.

3.1 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 parameters¹. 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 compounds can be represented mathematically by the appropriate treatment of an input chemical code.

In this work, it is proposed that the union of 2 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.

The usefulness of the proposed methodology will be demonstrated by its application to the analysis of several different types of data. The goal in each case is to obtain a set of descriptors - submolecular units - which correlate with the experimentally observed activities. These descriptors can then be used to predict the activity of compounds yet untested. Such procedures are of pharmacological interest as part of the screening process in the design of new drugs. In addition to this, another more general long-term goal seems quite achievable - the formation of a library of 'biological functional groups'. Organic chemists have had great success in explaining reactivity and designing synthetic pathways by recognizing and focusing on 'chemical functionalities' which are usually two, three, or four atom clusters that serve as reactive sites within the molecule. Analogously, it is interesting to hypothesize the existence of 'biological functionalities' that are expected to be considerably more complex than their chemical counterparts. By examining the increasingly vast and diverse amount of biological data currently available in the literature, this proposed methodology could be employed to tersely investigate such an hypothesis.

3.2 Graph Theoretical Representation of Molecular Structure

There are a multitude of methods provided by graph theory which have been employed to represent or numerically quantitate the chemical graph². 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 molecule³. 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 activity⁴. The activity can range from well-defined and readily measurable physicochemical properties (e.g. pH) to complex pharmacological properties (e.g. 'lethal dose') which can be difficult to determine 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 (molecular index). This issue has also been addressed in the current literature⁵ and was discussed in chapter II. In fact, it was a motivating factor for proposing the vertex index⁶, as described in detail (2.6). A consideration of both the advantages and disadvantages of the vertex index has contributed to the development of a markedly different methodology. Although the vertex index itself is abandoned, the concept of representing the structural environment of each atom by distance-based topological descriptors is retained. 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 ($1 < k < \#$ of atoms in the molecule) as opposed to a single index. 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.

To illustrate the application of such a procedure, the following terms are defined:

vertex (v) - a point (atom) in the graph
 edge (e) - a line connecting 2 points
 (bond connecting 2 atoms)
 length (l) - the length of an edge
 (taken as unity unless a
 weighing factor is introduced)
 path (p) - a sequence of distinct vertices
 in which adjacent vertices are
 incident (a linear string of
 atoms within the molecule)
 distance(d) - path of shortest length between
 2 vertices

For example, below is the simple graph G_1 :

G_1

Consider the vertex labeled X. Starting from this atom, there is one path of length 1, two paths of $l=2$, three paths of $l=3$, and four paths of $l=4$. Thus the structural environment of atom X could be characterized by the vector [1 2 3 4]. Now, if such a vector is formed for each of the n atoms of the molecule, an $n \times k$ matrix can be generated which we shall call the path matrix $[P_{ij}]$. (k = number of edges between vertices of maximum distance within the molecule).

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 $[C_{ij}]$ is then transformed into a distance matrix $[D_{ij}]$ ⁷ which contains the topological information required to construct the vectors to be used to describe the structural environment of each atom. The example, G_1 is a graph representation of compound 1 (4.1, table 2). The structure and atom numbering of compound 1 is given in fig(1) and the resulting distance matrix is given in table (2a). As can be seen, these distances are not based on actual bond lengths but are instead distances in the graph theoretical sense - as defined previously.

The distance matrix is symmetrical and each element is the shortest path length between 2 atoms. For example, the distance between atom 2 and atom 7 is a path of length 4 :

$$d_{2,7} = d_{7,2} = 4$$

For our purposes, each row (i) will be viewed as a distribution of distances with respect to atom (i). For each heavy atom in the molecule, this information is compressed into a vector descriptor that represents the structural environment of that atom as determined by the arrangement of the other atoms around it at varying distances. The j^{th} element of the vector descriptor is a count of the total number of atoms (hydrogen and non-hydrogen) at a distance j from the reference atom. Table(2b) shows the vector representation of each heavy atom of compound 1 obtained by counting the number of equivalent elements in each row (atom) of the distance matrix in table(2a). 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 3 different types of heavy atoms - each represented by its own vector descriptor (table 2c). 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 $[P_{ij}]$. This procedure is repeated for each molecule of the data set. Next, all the rows of $[P_{ij}]$ are combined into one large cumulative matrix, $[P_{i,j}]$ (see table 2 for an actual example based on compound 1). Finally, each row (vector) of $[P_{i,j}]$ is decomposed into all possible subunits, 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.

3.3 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 observations⁸. 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 experimental data available is that the compounds involved are usually quite complex and the number of potentially important structural features is overwhelming. It is under these conditions that computer implemented expert system analysis is ideal⁹. 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 sections). 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.

The final product of the analysis is a set of rules which express activity as a consequence of conditions represented by logical statements based on the presence and/or absence of certain embedded chemical substructures (functionalities). This information can then be used as a tool with the potential of providing fundamental insights into a compounds' chemical/biological reactivity.

Another useful application of these results is in the prediction of the activity of compounds not previously tested. When used in this mode the results of an analysis can be treated simply as a 'black box' predictor of activity which does not require an expertise on the part of the user in order to obtain useful activity predictions. In the absence of any mechanistic knowledge regarding the pharmacological pathway, such crude predictions may be the best that are available. In terms of drug design, features found to enhance potency can be optimized while other features linked to undesirable properties (e.g. toxicity) can possibly be eliminated or modified in the new target compound.

3.4 Statistical Parameters

In order to investigate the link between structure and activity, each subgraph is associated with an activity distribution which is simply a count of the number of active and inactive molecules which contain the given subgraph. This is essentially a 'null hypothesis' approach in which the subgraph distribution is assumed to be the same as a distribution that would result from a random sampling of the binomial parent distribution (entire data base). In other words,

- H_0 = no difference between subgraph distribution and that of a random sampling
 H_1 = subgraph distribution is not the same as a random sampling

The null hypothesis (H_0) is accepted when the binomial probability of the subgraph distribution is larger than a given cutoff value (e.g. $p > .15$) which carries the chemical interpretation that the subgraph is not related to activity. Conversely, H_0 is rejected and H_1 accepted when $p < .15$, which indicates a correlation between the presence of the given substructure and molecular activity/inactivity. Obviously, it is the later case that is of interest in establishing a structure-activity relationship. However, since H_0 will be accepted for the majority of subgraph distributions, this process serves to weed-out inappropriate topological descriptors.

in order to determine the significance and usefulness of a subgraph distribution, two independent measures are employed for the statistical evaluation of data - one for the selection of the optimum parameter and the other to insure the statistical significance of that parameter. The selection of parameters is based on a formulation by Laplace¹⁰ and the establishment of statistical significance is based on binomial cumulative probabilities¹¹.

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, a modification of Laplace's expression is implemented simply to take advantage of its favorable properties. The previous controversy is circumvented by viewing this parameter merely as an "index of skewness" as opposed to a formal probability. The modified expression used here defines a normalized index identified as $I(m)$:

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

a = number of actives
i = number of inactives
4 = normalization constant
m = adjustable parameter

The $I(m)$ function is an attempt to quantify a rather qualitative and arbitrary entity, i.e. the amount of skewness which indicates a link between a structural feature (subgraph) 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 .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 subgraph distributions of actives and inactives as obtained from the data set (parent) distribution. The expression employed is a summation over the relevant binomial probabilities - directly analogous to the probabilities calculated for the 'coin toss' experiment. The probability calculated for some subgraph (i) is :

$$P_i (x \geq x_0) = \sum_{x=x_0}^n \frac{n!}{x! (n-x)!} p^x q^{n-x} \quad (2)$$

x_0 = # of active* compounds in which subgraph i occurs
 n = total # of comps containing subgraph i
 p = (# of active comps) / (total # of comps 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 subgraphs

to illustrate its use, we use the coin toss experiment as an example and pose the question, 'What is the probability of getting 8 or more heads out of 10 tosses of a fair coin?'. For this problem, $x_0=8$, $n=10$, $p=q=.5$, and the above expression becomes:

$$P(x \geq 8) = \frac{10!}{8! 2!} .5^{10} + \frac{10!}{9! 1!} .5^{10} + \frac{10!}{10! 0!} .5^{10} = .0537$$

which means that the odds are approximately 19 to 1 against observing such an outcome. This small probability (5.37%) is complementary to a large confidence ($100 - 5.37 = 94.6\%$) that such an event is NOT due to chance. If this example had actually involved substructure activity rather than coin tosses, we would conclude, at a 94.6% confidence level, that the subgraph is correlated with activity since such a distribution of activities (8 actives, 2 inactives) is not to be expected from a random sampling of a data base with an equal number of active and inactive molecules (H_0 rejected, H_1 accepted).

An obvious restriction placed on the training data is that it must include inactive as well as active compounds. As can be seen from the previous discussion, the reason for this is that the statistical analysis is dependent on the existence of at least two activity classes. Because activity is a relative quantity, its definition is flexible and can often involve a certain degree of arbitrariness. However, care should be taken to define a meaningful cutoff point between actives and inactives since the statistical parameters are dependent on a 'real' distinction between activity classes.

3.5.1 SEV descriptors

The selection of relevant submolecular descriptors from the large data pool is now discussed. Matrix $[P_{ij}]$ serves as the source for the generation of subgraph vector descriptors. Each row of $[P_{ij}]$ is considered to be a vector which represents the molecular topology with respect to a particular atom within the molecule. Next, a list of potential descriptors is formed by decomposing each row of $[P_{ij}]$ - one element at a time. For example, the row of $[P_{ij}]$ that represents compound 1, atom 1 of the acid data base (4.1, table 2) is decomposed into the following subsets : [1], [1 2], [1 2 3] and [1 2 3 4] ; each representing an increasing portion of the molecular graph with the last vector representing the entire molecular graph, from the perspective of atom 1. Each of these 'vector subsets' is identified as a **SEV** (structural environment vector) since it characterizes the substructural environment of a given atom. An activity distribution is associated with each by counting the number of molecules in which the various **SEVs** are found. The purpose of this is to establish a link between activity and commonly occurring vector elements. A meaningful structure-activity link is assumed if a statistically significant number of molecules with similar activity all possess the same **SEV**. To illustrate this process, the example vector descriptors given above have the following distributions between active and inactive molecules:

SEV descriptor	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 inactives (non-acids). Thus, every SEV is associated with a distribution that can be tested for statistical significance and evaluated for its potential usefulness as a predictive parameter. Calculating the index of skewness I(m) and the binomial probability (BiP) according to eqs.(1 & 2) for these 4 subgraphs, we obtain the following values :

SEV descriptor	I(m) (m=.1)	BiP
[1 2 3 4]	.9518	.0001
[1 2 3]	.0845	.0045
[1 2]	.0019	.0002
[1]	.0000	.0005

Thus, in comparing these four descriptors, [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, 8 compounds are successfully accounted for and zero misclassifications result. In contrast, if any of the other 3 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 ($BiP < .05$) 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 3 fragments are also statistically significant, but of questionable use as predictive parameters.

Although real numerical data was used for this example, it should be noted that in the actual analysis of the entire data set [1 2 3 4] is not selected as a descriptor because better descriptors exist and will be discussed in chapter 4.1. The fact that a descriptor exhibits statistical significance in no way implies its selection as an important structural feature. It is merely a minimum requirement that must be satisfied. Furthermore, the $I(m)$ criterion associated with a vector descriptor is employed as a relative measure which is useful only when compared and found superior to other $I(m)$ values. It should also be mentioned that in the complete analysis of all the data (see 4.1), compounds are selectively removed from the data set after a substructural feature (**SEV**) is revealed that accounts for their activity. It is this 'selective removal' process that modifies the original data and thereby changes

the relative importance of the remaining descriptors.

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 (# actives vs. # inactives) and thus identical $I(m)$ values. Such cases are encountered so frequently that the following algorithm 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 labelling of the various alternatives (i.e., case i,ii,...,ix) for convenient location of the appropriate path of logic. The rationale behind the flowchart below - where not intuitively obvious - is further explained when it is applied to actual cases that arise in the analysis of data (see chapter 4, 'Application of Methodology').

Since the algorithm is a means of evaluating the interdependence of descriptors, it is applied only to potential descriptors of like activity. If two descriptors are of opposite activity, the flowchart analysis is inappropriate since they are already deemed independent. Furthermore, although only two-descriptor interactions are represented in the Venn diagrams, multiple descriptor evaluation is possible by repeated application of the algorithm - two descriptors at a time - as will be demonstrated in treatment of actual data in chapter 4.

For the purpose of example, [A] and [B] will symbolize two vector descriptors that are both statistically significant and of acceptable rank according to the I(m)-value ordering. The following flowchart represents the possible relationships between descriptors [A] and [B], along with the resulting logical statements. ([A] and [B] are assumed to represent activating subgraphs - for inactive subgraphs, simply interchange c,x,e with d,y,f). The symbols used in the flowchart are defined in the following table.

$\subset$	subset;
$-, '$	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)

3.5.2 Common and 'activity-specific' descriptors

Upon formation of all SEV descriptors, a search is made for common descriptors, i.e. SEVs found in every molecule in the data base. This allows a logical statement to be made which defines the limitations imposed by the congeneric nature of the data. Such a statement then becomes the first or 'topmost' rule in the hierarchy of rules (logic statements) to follow. As an example, suppose 'a' and 'b' represent submolecular functionalities common to all the compounds of the data set. From this it is stated :

$$\sim (a \wedge b) \rightarrow \text{indeterminate}$$

meaning that any attempt to predict the activity of a new compound not containing both 'a' and 'b' is an extrapolation beyond the limitations imposed by the original training set. (The nitroso data set, chapter 4.2, illustrates this point.)

Another useful type of descriptor to identify is the 'activity-specific' common descriptor. These are descriptors found not in the entire data set, but rather are found to occur uniquely to a specific activity class. For example, suppose all of the active molecules contain submolecular moieties 'c', 'd' and 'e' and only a small percentage of the inactive compounds contain these same features. Under such circumstances, the data set can then be considerably reduced by the statement :

$$\sim (c \wedge d \wedge e) \rightarrow \text{inactive}$$

which correctly assigns the 'inactive' label to molecules not containing these descriptors. The analysis of the acid/nonacid data set (chapter 4.1) illustrates just such a condition.

After the identification of 'common' and 'activity-specific' descriptors (if they exist), the SEVs are arranged by rank in descending order according to their $I(m)$ index values. This puts the most favorable potential descriptors at the top of the list. An additional criterion also imposed before allowing the further consideration of a given SEV descriptor is that it also be significant according to the cumulative binomial probability (BiP) calculated from the partitioning of molecules between the active and inactive classes. Provided the BiP is acceptable, the top 'n' $I(m)$ ranked SEVs are then considered for further evaluation. The cutoff value for $I(m)$ is obtained by selection of the smallest 'n' for which the following expression is true :

$$\frac{|\Delta I(m)|}{|\Delta r|_n} >> \frac{\sum_{j=1}^{n-1} |\Delta I(m)|}{n-1}$$

Analytically, if $I(m)$ is considered to be a continuous function of r (rank), then we are interested in finding a point of inflection (large negative derivative). The n highest $I(m)$ -valued descriptors preceding this point are the ones of interest. The above expression can be used to determine the smallest n for which this condition occurs. As can be seen, the right side of the above expression is simply an averaging of the slopes ($\Delta I(m)/\Delta r$) of the preceding $n-1$ descriptors.

In the example graph, there is a considerably larger $\Delta I(m)$ between the 5th and 6th ranked SEVs than between any of those preceding this pair. Therefore the cutoff is at this point ($n=5$) and the first five SEVs are considered for further evaluation.

The **SEVs** considered for further evaluation are first segregated into active and inactive classes to be compared separately. Any vector descriptors which are explicitly subsets of other potential descriptors are grouped together and evaluated pairwise according to the Venn diagram analysis presented in the previous flowchart wherein the optimal vector descriptor is selected from the group. By following this procedure, each group of related descriptors is reduced to a single representative to be considered in the formation of a logic statement.

When in the learning mode of the algorithm, the above procedure results in the elimination of all compounds which satisfy the logic expression. This process is repeated over and over until either all the compounds in the training set have been eliminated or no more statistically significant substructural features remain. Contrastingly, when in the predicting mode the resulting logic statements are applied in a hierarchic order to predict the activity of test compounds.

Summary of Substructure Analysis Algorithm

Common descriptor analysis

- identify all common SEV descriptors
- identify all 'activity-specific' SEVs

SEV indexing and selection

- descriptors arranged in descending order according to $I(m)$ index
- binomial probability calculated to assure statistical significance
- top n SEV fragments are retained as determined by equation (1)

Evaluation of descriptors / activity statements

- separate active and inactive classes
- within a given class, evaluate explicit subsets and eliminate redundancies
- identify interdependent descriptors (implicit subsets) by pairwise comparison and formulation of appropriate intersections
- evaluate reduced set of SEVs by application of flowchart to pairwise Venn diagrams

Table 2

Distance matrix [D_{ij}]												Path matrix [P_{ij}]									
atom	1	2	3	4	5	6	7	8	9	10	11	atom	Vector Descriptor								
1	0	2	1	3	2	4	4	3	4	4	3	1	1	2	3	4	0	0	0	0	0
2	2	0	1	3	2	4	4	3	4	4	3	2	1	2	3	4	0	0	0	0	0
3	1	1	0	2	1	3	3	2	3	3	2	3	3	4	0	0	0	0	0	0	0
4	3	3	2	0	1	3	3	2	3	3	2	4	hydrogen	a	t	o	m				
5	2	2	1	1	0	2	2	1	2	2	1	5	4	6	0	0	0	0	0	0	0
6	4	4	3	3	2	0	2	1	4	4	3	6	1	2	3	4	0	0	0	0	0
7	4	4	3	3	2	2	0	1	4	4	3	7	1	2	3	4	0	0	0	0	0
8	3	3	2	2	1	1	1	0	3	3	2	8	3	3	4	0	0	0	0	0	0
9	4	4	3	3	2	4	4	3	0	2	1	9	1	2	3	4	0	0	0	0	0
10	4	4	3	3	2	4	4	3	2	0	1	10	1	2	3	4	0	0	0	0	0
11	3	3	2	2	1	3	3	2	1	1	0	11	3	3	4	0	0	0	0	0	0

Table 2a

Table 2b

Path matrix, cumulative [P_{ij}]
(comp. 1 only)

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

Table 2c

potential descriptors
(from comp. 1)

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

Table 2d

Above matrices and descriptors are an example of the processing of the information generated by only one compound in a training data set (see figure 1 for the molecular structure)

Figure 1

References

1. Kier, L. and Hall, L., "Molecular Connectivity in Chemistry and Drug Research", Academic Press, N.Y., 1976.
2. Trinajstić, N., "Chemical Graph Theory" Vol.I&II, CRC Press, Inc., 1983.
3. Randić, M., J. Am. Chem. Soc., **97**, 6609 (1975).
4. Kier, L., J. Pharmaceutical Sciences, **69**, 807 (1980).
5. Grossman, S., Dzonova, B., Randić, M., Int. J. Quantum Chem., **12**, 123 (1986).
6. Klopman, G., Raychaudhury, C., and Henderson, R., Mathl Comput. Modelling, **2**, 635 (1988).
7. Muller, W.R., Szymanski, K., Knop, K., and Trinajstić, N., J. Comput. Chem., **8**, 170 (1987).
8. Shafer, G., Statistical Science, **2**, 3 (1987).
9. Davis, R., "The AI Business", 4th ed., P. Winston and K. Prendergast (Eds.), MIT Press, Cambridge, 1986.
10. Laplace, "A Philosophical Essay on Probabilities", Dover Publications, N.Y. 1951.
11. Bailey, D.E., "Probability and Statistics - Models for Research", John Wiley & Sons, N.Y., 1971.

Chapter IV

APPLICATION OF EXPERT SYSTEM METHODOLOGY TO THE ANALYSIS OF EXPERIMENTAL DATA

In order to illustrate the methodology, the analysis is applied 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. The 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.

The process is begun 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(2) and contains the following, from left to right:

- acidity - the relative acidities are based on pKa data^{1,2} (extrapolated where necessary) and are rescaled from 10 through 90 for the program (90 = most active) by the expression: $\text{acidity} = -10 \text{ pKa} + 90$
- KLN code - a linear coding system³ in which different symbols represent atoms and/or functional groups (see table 3)
- comp. 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 analysis of this data and the interpretation of the results take place in the five stages that follow.

4.1.1 Stage I - identification of nonacids

The input data shown in table(2) produces a 1014×21 $[P_{IJ}]$ 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 subgraphs common to all the compounds or subgraphs common to all the active compounds of the data set. Vector [1], as indicated below, represents just such a subgraph.

	<u>total</u>	<u>active</u>	<u>inactive</u>
data	122	38	84
[1]	76	38	38

The fact that all active molecules are characterized by a common substructure represented by this simple vector leads to the conclusion 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 :

$$\sim [1] \implies \text{inactive} \quad (1)$$

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.

4.1.1.2 Stage II - aliphatic carboxylic acid descriptors

Following the elimination of 46 inactive compounds, the remaining 76 molecules are subject to the previously described analysis wherein the 636 rows of the reduced $[P_{IJ}]$ provide 256 potential descriptors to be evaluated in this second stage of the analysis. These subgraph descriptors are sorted and ordered according to relevance to activity by the $I(m)$ parameter. The **SEV** descriptors that emerge as the statistically best, according to the prescribed cutoff determination, are given below:

	<u>total</u>	<u>active</u>	<u>inactive</u>
	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 explicitly a subset of [2 2 3 3], as demonstrated by the following Venn diagram, and thereby provides no useful information (see 3.5.1, Flowchart, case ii).

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 [1243] accounts for more of the data, but makes one false classification (ie, 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.

	<u>total</u>	<u>active</u>	<u>inactive</u>	
	76	38	38	biP
[1 2 4 3]	18	17	1	
[1 2 4 3 3]	12	12	0	

[1 2 4 3] ^ [1 2 4 3 3]'	6	5	1	.109

The last line of the table above 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=.109) and we conclude that [1 2 4 3] is statistically significant even in the absence of [1 2 4 3 3] (3.5.1, flowchart, case iii). 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 the two remaining SEVs, [2 2 3 3] and [1 2 4 3] is investigated.

[1 2 4 3] is rejected 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, (ie., 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 the most reliable classification to be made. (3.5.1, flowchart, case iv) The conclusion is represented as follows:

[2 2 3 3] ^ [1 2 4 3] --> active (activity= 46 +/- 5)
statement (2)

Since the average activity of the compounds containing these 2 descriptors is 46.5 with a standard deviation of less than 5 activity units, the expression above includes this information. This case shows that although the predictions are generally qualitative (active or inactive), it is possible to obtain semi-quantitative results.

chemical interpretation of statement 2

The chemical functionality associated with this statement is the carboxylic group of an aliphatic substrate:

I

Although the identification of the COOH group as an activating feature was well anticipated, the substructure also requires the presence of two terminal atoms (generally hydrogens) at the alpha position. The numerous amino acids in the data base are the reason for this additional criterion. These compounds, in general, are nonacidic yet contain the COOH moiety. Therefore, the 'terminal' alpha position requirement excludes the amino functionality from the defined substructure thus allowing the acidic (active) aliphatic carboxylic acids to be discriminated from the nonacidic (inactive) amino acids.

4.1.3 Stage III - inactivity descriptors

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 all associated with inactivity.

	<u>total</u>	<u>active</u>	<u>inactive</u>
	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. (see 3.5.1, Flowchart, case iii, inactive)

	<u>total</u>	<u>active</u>	<u>inactive</u>	
	61	23	38	biP
[4 5]	19	1	18	
[4 5 3]	9	0	9	

[4 5] ^ [4 5 3]'	10	1	9	.062

The results of the pairwise Venn diagram comparison of the remaining descriptors (see fig. 1) are arranged in the following matrix:

	a	b	c	d	e
a = [4 5]	a	...	aUc	aUd	aUe
b = [4 6 5]	b	...	bUc	bUd	bUe
c = [4 7]	c		...	c^d	c^e
d = [1 2 4 5]	d			...	d^e
e = [2 2 3 5]	e				...

Since this 'logic matrix' is symmetrical about the main diagonal, the upper triangle displays all the various relationships. According to row a, column b; a comparison of descriptors a and b results in a being retained and b being rejected (3.5.1, flowchart, case vi).

In addition, the logic matrix indicates c, d and e to be interdependent. Venn diagram analysis reveals that the 3 subgraphs represented by these vectors are found in the same set of compounds.

(3.5.1, flowchart, case i, inactive)

In other words, the same 19 compounds contain these same 3 substructural features.

The logic matrix also reveals **a** to be independent of **c**, **d** and **e**, as demonstrated below:

Since both **A** and **a** are significant as independent descriptors, (3.5.1, flowchart, case v, inactive), stage 3 of the analysis concludes with statement(3)

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

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

chemical interpretation of statement 3

Two independent chemical functionalities are associated with the logical statement above, as indicated by the union operator. The compounds containing $([1245] \wedge [2235] \wedge [47])$ are all α -amino acids. The substructural unit defined by these intersecting descriptors is:

II

Of the nineteen molecules which contain II, 18 are nonacidic and we therefore associate this substructure with 'inactivity'.

The fragment substructures associated with descriptor [45], aside from the amino acids which have already been explained, are the aldehyde and ketone functionalities, as shown:

III

All eight nonamino-compounds containing this structure are inactive. The interpretation is that the 'inactive' label applied to a chemical functionality simply means that this class of compounds lacks an activating feature.

At stage 3 of the analysis, the two chemical species above were linked to inactivity. The first structure is quite reasonable considering the presence of the basic amine functionality. Not only are these compounds generally nonacidic, but they obviously have the potential to exhibit basicity. Although the second structure is equally linked to 'nonacidity', the situation is quite different. It exhibits inactivity only by association with the aldehyde and ketone functionalities, which are, by our defined pKa standards, nonacidic. Thus, the amino compounds are inactive because they possess a 'deactivating' feature whereas the aldehydes and ketones are inactive simply due to the lack of an activating feature (structure II is causally related to inactivity, structure III is not). Although there is a fundamental difference in these two types of inactivity, it is not discernible from the logic statement alone.

4.1.4 Stage IV - carboxylic group descriptor

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

	[123]		[22]	
	<u>total</u>	<u>active</u>	<u>inactive</u>	
	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 (ie. $[1\ 2\ 3] \wedge [2\ 2]$) is not considered significant (BiP = .42) whereas [2 2] is considered independently significant (BiP = .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. (3.5.1, flowchart, case vii) The qualitative conclusion is given in the expression below in the usual manner.

[2 2] --> active (act.= 56 +/- 18) (4)

chemical interpretation of statement 4

The chemical entity associated with [2 2] is the familiar carboxylic group :

IV

Note that this substructure differs from structure (I) in that it poses no restriction beyond the carboxylic group itself. The alpha position constraint imposed by (I) was necessary to discriminate the α -amino compounds from the aliphatic carboxylic acids. At this stage of the analysis, both these classes of compounds have previously been successfully evaluated and systematically eliminated from the data base. The remaining carboxylic acids in the data set (17 active and 1 inactive) are now identified and properly classified by (IV). This group of compounds includes both aromatic and low molecular weight aliphatic carboxylic acids - neither of which is represented by structure (I).

4.1.5 Stage V - noncarboxylic acid descriptors

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 (4) 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 towards inactivity. This situation is indicative of the final stages of analysis, for at this point most of the data has been explained with considerable rigor and 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". (3.5.1, flowchart, case viii)

	<u>total</u>	<u>active</u>	<u>inactive</u>	
	16	5	11	biP
[1 2 3]	2	2	0	.098
[4 6]	5	3	2	.180
[4 6] ^ [1 2 3]	1	1	0	.312
[4 6] U [1 2 3]	6	4	2	.081 *

Stage 5 thus concludes the analysis with the statement :

[1 2 3] V [4 6] --> active (5)

chemical interpretation of statement 5

The acidic compounds classified by statement (5) are all of the form :

where the Xs represent electron withdrawing groups such as NO₂ and SO₂. Although relatively few of these noncarboxylic acids exist in the data, they are generally of high acidity (low pKa). The previous activating descriptors were unable to identify these molecules due to the much greater statistical significance of the classical carboxylic acids - which at this point have been eliminated from the data. Although statement (5) is not of the same order of statistical significance as those previously, it is useful in identifying some of the most active compounds in the data base.

4.1.6 Conclusion - Acidity data analysis

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 along with their chemical interpretations were explained in detail and a summary is given in table (1). The familiar chemical functionalities (eg., activating carboxylic group, deactivating amino moiety) serve to help validate the proposed methodology.

Table 1

Chemical Interpretation of Acid Data Analysis

<u>logical statement</u>	<u>classification</u>		<u>chemical interpretation</u>
	true	false	
~[1] -> inactive	.46	-	all active molecules contain a terminal heavy (non-hydrogen) atom
[2233] ^ [1243] -> active	15	+	$\begin{array}{c} \quad \\ -C - C - C \begin{array}{l} \nearrow O \\ \searrow O-H \end{array} \\ \quad \end{array}$ <p>all carboxylic-containing molecules with terminal α - atoms</p>
[[1245]^[2235]^[47]] V [4 5] -> inactive	27	- 1 -	$\begin{array}{c} \quad \\ -C - C - C \begin{array}{l} \nearrow O \\ \searrow O-H \end{array} \\ \quad \\ \quad N \\ \quad / \quad \backslash \\ \quad H \quad H \end{array}$ <p>α - amino group deactivates the carboxylic group</p>
[2 2] -> active	17	+	$C - C \begin{array}{l} \nearrow O \\ \searrow O-H \end{array}$
[123] V [46] -> active	4	+	$\begin{array}{c} \quad \quad X \\ \quad \quad / \\ H - C - X \\ \quad \quad \backslash \\ \quad \quad X \end{array}$

Table 2

<u>acidity</u>	<u>KLN code</u>	<u>chemical formula</u>	<u>comp #</u>
90	XDXX	CH (NO ₂) ₃	1
90	DYMYMYM	CH (SO ₂ CH ₃) ₃	2
90	DC3NC3NC3N	CH (CN) ₃	3
88	FCTKFF	F ₃ CCOOH	4
83	GCTKGG	Cl ₃ CCOOH	5
77	KTTK	HOCCO ₀ H	6
77	GDTKG	Cl ₂ CHCO ₂ H	7
73	RXTK	CH ₂ NO ₂ COOH	8
64	FRTK	F CH ₂ COOH	9
62	KTRTK	HOOC CH ₂ COOH	10
61	GRTK	Cl CH ₂ CO ₂ H	11
61	MRDGTK	CH ₃ CH ₂ CH Cl COOH	12
56	X*TK	NO ₂ -Ph-COOH	13
52	KRTK	HOCH ₂ CO ₀ H	14
51	MC2DD2DD2CTK/	o-TOLUIC ACID	15
50	MDGRTK	CH ₃ CH Cl CH ₂ COOH	16
50	XRK	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) ₂ COOH	20
48	*TK	Ph-COOH	21
47	KT=3TK	HOOC (CH ₂) ₃ COOH	22
46	M*TK	p-TOLUIC ACID	23
46	KT=4TK	HOOC (CH ₂) ₄ COOH	24
45	MO*TK	MeO-Ph-COOH	25
45	KT=5TK	HOOC (CH ₂) ₅ COOH	26
45	KT=6TK	HOOC (CH ₂) ₆ COOH	27
45	KT=7TK	HOOC (CH ₂) ₇ COOH	28
44	KT=8TK	HOOC (CH ₂) ₈ COOH	29
44	G=3TK	Cl (CH ₂) ₃ COOH	30
42	MTK	CH ₃ CO ₂ H	31
42	M=2TK	CH ₃ (CH ₂) ₂ COOH	32
41	MRTK	CH ₃ CH ₂ COOH	33
41	M=3TK	CH ₃ (CH ₂) ₃ COOH	34
30	DTMTMTM	CH (C=O CH ₃) ₃	35
10	*J	Ph-SH	36
10	MTRTM	CH ₃ C=O CH ₂ C=O CH ₃	37
10	*K	PHENOL	38
10	RC3NC3N	CH ₂ (CN) ₂	39
10	RYMYM	CH ₂ (SO ₂ CH ₃) ₂	40
10	MRK	CH ₃ CH ₂ OH	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	CH3C=O CH3	46
10	RC2DD2DD2C) C2DD2DD2C) /	FLUORENE	47
10	*C3D	PHENYLACETYLENE	48
10	C2DR* * *	1,1,3-TRIPHENYLPROPENE	49
10	*A	Ph-NH2	50
10	M*A	Me-Ph-NH2	51
10	MYM	CH3SO2 CH3	52
10	D* * *	(Ph) 3 CH	53
10	*R*	(Ph) 2 CH2	54
10	M*	TOLUENE	55
10	D2DD2DD2D/	BENZENE	56
10	RRR/	CYCLOPROPANE	57
10	MDM*	CUMENE	58
10	DC2DD2DD2C) DC2DD2DD2C) /C2D	D2DD2C/) 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	CH3-O-CH2CH3	78
10	MRORM	(CH3CH2) 2 O	79
10	MRRORM	(CH3CH2CH2) 2 O	80
10	MRA	CH3CH2 NH2	81
10	MRRRA	CH3CH2CH2 NH2	82
10	MRRRA	CH3CH2CH2CH2 NH2	83
10	*RA	Ph-CH2-NH2	84
10	MR*A	CH3CH2-Ph-NH2	85
10	G*RA	Cl-Ph-CH2-NH2	86
10	R2DRA	CH2=CHCH2 NH2	87
10	MRRK	CH3CH2CH2 OH	88
10	MRRRK	CH3(CH2) 3 OH	89
10	MRRDKM	CH3(CH2) 2CHOH CH3	90
10	D2DRRR)	CYCLOPENTENE	91
10	D2DRRRR)	CYCLOHEXENE	92
10	MTRM	CH3 CO CH2CH3	93
10	MRTRM	(CH3CH2) 2 CO	94
10	*TM	Ph-CO-CH3	95

10	*RTM	Ph-CH ₂ -CO-CH ₃	96
10	R2DRM	CH ₂ =CHCH ₂ CH ₃	97
10	MD2DRM	CH ₃ CH=CHCH ₂ CH ₃	98
10	FCFFM	F ₃ CCH ₃	99
10	GCGGRK	Cl ₃ CCH ₂ OH	100
10	ARTK	glycine	101
10	MDATK	alanine	102
10	MDMDATK	valine	103
10	MDMRDATK	leucine	104
10	MRMDATK	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

scale: acidity = -10 pka + 90

Table 3

KLN code

ATOM	+ 1 H-	+ 2 H-	+ 3 H-	+ 1 O=	+ 2 O=
----	-----	-----	-----	-----	-----
H -H					
F -F					
G -Cl					
B -Br					
I -I					
C -C	D -CH	R -CH2	M -CH3	T -CO	
N -N(3)	E -NH	A -NH2		U -NO	X -NO2
Q -N+					
O -O	K -OH				
P -P(3)				W -PO	
Z -P(5)				L -PO	
S -S					Y -SO2

code	definition
/	close ring with first open valence
)	close ring with last open valence
*	monosubstituted phenyl ring
(n	monosubstituted aliphatic ring of size n
=n	linking aliphatic chain of n carbon atoms

Table 4

BiP ordering of descriptors - stage V

Total = 16 Actives = 5 Inactives = 11

total	act	inact	BiP	SEV
2	2	0	.0977	1 2 3
5	3	2	.1800	4 6
4	0	4	.2234	3 4
3	2	1	.2319	3 3
3	0	3	.3250	3 4 3
3	0	3	.3250	3 4 3 1
6	1	5	.3936	3 6
6	1	5	.3936	4 2
2	0	2	.4727	3 7
2	0	2	.4727	1 3 3
2	0	2	.4727	3 4 5
2	0	2	.4727	3 4 6

Figure 1

descriptors	Venn analysis	result
a,b		a
a,c		aUc
a,d		aUd
a,e		aUe
b,c		bUc
b,d		bUd
b,e		bUe
c,d		c^d
c,e		c^e
d,e		d^e

References

1. Lowry, T.H. and Richardson, K.S., "Mechanism and Theory in Organic Chemistry", Harper & Row, N.Y., 1976.
2. Mahler, H. and Cordes, E., "Biological Chemistry", 2nd ed., Harper & Row, N.Y., 1971.
3. Klopman, G. and McGonigal, M., J. Chem. Inf. Comput. Sci, **21**, 48 (1981).

4.2 N-Nitroso Compounds - carcinogenic activity data

The N-nitroso compounds are of particular pharmacological interest because of their suspected active role in carcinogenesis.¹ Although nitrosamine synthesis dates back to the late 19th century, investigation of the compounds' toxicity did not begin until the 1950s - initiated as a consequence of an accidental poisoning. Testing of nitrosodimethylamine in rats soon revealed this compound to be linked to liver necrosis, with a nearly 100% incidence rate², and the scope of testing broadened to include the toxicological examination of the entire class of N-nitroso compounds in general.¹ These studies continue in full swing today. Even though the detailed mechanisms of toxicology and carcinogenesis are not yet well-understood, the ever-increasing number of compounds and classes of compounds being tested allows reason for optimism regarding the eventual understanding of the mechanistic function of these compounds as well as the mechanism(s) of carcinogenesis in general.

The N-nitroso compounds currently under extensive investigation include nitrosoalkylamides as well as nitrosoalkylamines. Although it was originally proposed that these classes of compounds were carcinogenic through a diazoalkane ($RR'C=N=N$)⁻ alkylating intermediate, current evidence suggests an α -hydroxynitrosamine pathway leading to alkyl diazohydroxide which decomposes to form the carbocation intermediate responsible for alkylation of DNA.^{3,4}

N-nitrosoalkylamines, the compounds studied here, are believed to undergo α -hydroxylation by microsomal mixed-function oxidases⁵, although the details of this process are currently lacking. Further oxidation by enzymes results in monodealkylation, thereby generating a relatively inert aldehyde which is ultimately oxidized to CO₂. The remaining monoalkylnitrosamine is proposed to be the biologically reactive entity through its decomposition to alkylating intermediates which bind to DNA. The main site of attack of these alkylating agents is the N-7 position of guanine, although numerous other sites have also been reportedly attacked.⁵ The relative importance of the position of attack with regard to its mutagenic effect is currently the target of considerable investigation by many researchers.

The principal source of exposure to N-nitroso compounds for humans is believed to be the gastrointestinal tract - mainly the stomach¹. It is here that nitrites (considered the most common nitrating agent) from foods react with secondary and tertiary amines to form nitroso compounds. In fact, a large number of amines used as drugs and agricultural chemicals have been shown to form N-nitroso derivatives that are found to be mutagenic.⁶

The following amines are examples of compounds that have been shown to form N-nitroso derivatives, when fed to rats with nitrite, and are hypothesized to follow the previously discussed α -hydroxylation/alkylation pathway resulting in carcinogenesis:

<u>compound</u>	<u>mutagenic</u>	<u>tumor type</u>
dimethyldodecylamine	+	bladder , stomach
morpholine	++	liver
piperine	+	stomach
thiram	+	nasal cavity , stomach

<u>drug</u>		
allantoin	+	liver , stomach
aminopyrine	+++	liver
chlorpheniramine	+	liver
disulfiram	+	esophagus
oxytetracycline	+	liver
tolbutamide	+	liver , stomach

4.2.1 Stage I - congeneric data recognition

The data in table(2) was taken from the literature⁷ and serves as the learning data for the analysis. The data contains 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 subgraphs [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 subgraphs leads to the default label of "indeterminate", as indicated by expression (1a).

$$\sim [1 \ 1 \ 2] \vee \sim [2 \ 2] \quad \text{--->} \quad \text{indeterminate} \quad (1a)$$

Upon further examination, it can be shown that [1 1 2] is a topological subset of [2 2]. Therefore, the presence of [112] implies the presence of [2 2] ([112] can not occur in the absence of [2 2]). In such a case, the implied descriptor carries no useful information and is eliminated as redundant. The simplified statement resulting from this observation is (1b):

$$\sim [1 \ 1 \ 2] \quad \text{--->} \quad \text{indeterminate} \quad (1b)$$

which is every bit as informative as (1a). In general, the detection of such 'descriptor redundancies' is convenient only in terms of simplifying the logic expressions and does not affect in any way the conclusions drawn from the statements (i.e. statements(1a) and (1b) are mathematically equivalent for this data set).

chemical interpretation of statement 1

The descriptor(s) of statement (1) specify the nitroso functionality itself :

which is present in every compound in the data base. Expression (1) can thereby be viewed as a statement concerning the congeneric nature of the data. Structure (I) may seem trivial and of little value - especially since it is easily observed by inspection of the structural formulas of the data set compounds. However, the ability to identify such conditions is nonetheless important in a predictive methodology so as to prevent the risky situation of extrapolation beyond the limitations imposed by the training set itself.

4.2.2 Stage II - α -methyl and halogen activation

The ordering of the statistically significant SEV descriptors according to their $I(m)$ values produces the following set of best potential descriptors:

SEV	<u>total</u>	<u>active</u>	<u>inactive</u>
[4 2 4]	7	7	0
[1 3]	5	5	0
[1 2 4]	5	0	5
[3 7 3]	5	5	0
[2 2 6 3]	5	5	0
[4 2 4 3]	5	5	0
[1 1 2 6 3]	5	5	0
all data	64	45	19

Although the data is highly skewed toward activity, the BiP calculation (see 3.4, eq.2) involves terms (p and q) that allow for the fair evaluation of statistical significance even under such unfavorable circumstances.

Following the prescribed method of analysis, the active and inactive descriptors are segregated into separate classes for further analysis. SEV [1 2 4], being the only inactive descriptor, is temporarily put aside while the active ones are evaluated. To simplify the logic table evaluation of the six potential active descriptors, one SEV is eliminated as a subset ($[4243] \subset [424]$) and three others are observed to be interdependent, as shown below:

The results of the pairwise Venn diagram analysis (see fig.1) of the remaining descriptors is shown in the 3x3 matrix below.

a = [4 2 4]
b = [1 3]
C = [373] ^ [2263] ^ [11263]

	a	b	C
a	.	aUb	a
b		.	bUC
C			.

As shown, **a** and **b** are independent and **C** is eliminated as a descriptor since the compounds identified by **C** are a subset of the compounds identified by **a**. Expression (2) states this conclusion.

[4 2 4] V [1 3] ----> active (2)

[4 2 4] ----> 48.4 +/- 6.5 (2a)

[1 3] ----> 53.0 +/- 4.5 (2b)

It is important at this point to emphasize the hierarchic nature of these logical statements. The descriptor of expression (1) ([1 1 2]) from stage 1 is considered to be essential for activity classification due to the limitations imposed by the congeneric nature of the training data set. Thus expression (2) could be more rigorously written as (2c):

([4 2 4] V [1 3]) ^ [1 1 2] ----> active (2c)

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 (1b) 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.

[1 2 4], the lone inactive descriptor of stage 2, is obviously independent of the other descriptors (the point can be proven through Venn diagram analysis) and is accordingly linked to inactivity by the following statement:

$$[1\ 2\ 4] \quad \text{--->} \quad \text{inactive} \quad (3)$$

Since all three SEVs are at the same level within the hierarchy, we have no statistical means of resolving the conflicting conclusions produced by statements (2) and (3) when a test compound is found to contain fragments that satisfy both conditions simultaneously.

$$([4\ 2\ 4] \vee [1\ 3]) \wedge [1\ 2\ 4] \quad \text{--->} \quad ?$$

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 the above expression with an 'active' label. Thus, statement (2) is considered to be above (3) in the hierarchy.

chemical interpretation of statements 2 & 3

There are three activity descriptors found during stage II of the analysis, two which are associated with activity ([4 2 4] and [1 3]) and one that is associated with inactivity ([1 2 4]).

[424] -> 48.4 +/- 6.5 [13] -> 53.0 +/- 4.5 [124] -> inactive

I

II

III

Structure (I) is the most statistically significant and indicates that either a methyl or methylene group alpha to the N-nitroso plays a role in the activity of these compounds. This finding nicely complements the proposed mechanism which requires hydrogens on the α -carbons in order for α -hydroxylation to take place. In fact, from a probabilistic point of view, the presence of the α -methyl group in (I) maximizes the potential for α -hydroxylation (3 hydrogens available for substitution by a hydroxyl group). The actual structural formulas of the molecules containing this substructure are shown in table (3a).

The second activating descriptor, SEV [1 3], indicates that the halogens chlorine and bromine are the root atoms of substructure (II) and thereby correlated with activity. Considering the position of (II) with respect to the N-nitroso group (see table (3b)), the

following resonance argument can be introduced to explain the possible role of the halogens in stabilization of the alpha-carbonium ion formed upon decomposition of the diazoalkane.

Along with the carbonium ion stabilization argument it is also noted that these compounds have the highest average activity and lowest deviation.

Structure (III), the chemical functionality represented by SEV descriptor [1 2 4], is the only substructure associated with inactivity. It is also found only in molecules (see table 3c) which contain neither activating structure (I) nor (II). Due to this independence, there is not enough information contained in the data base to investigate the possible role of (III) as a deactivator of otherwise active molecules. Its correlation with inactivity may be due to transport phenomena, thereby allowing molecules containing (III) to follow a different metabolic pathway. Considering the variety of physiochemical properties that can be modified by the presence of the carboxylate group of (III) - such as acidity, solubility, hydrophobicity, polarity, etc. - it seems quite reasonable to anticipate that these compounds could behave in a physiologically different way than the rest of the compounds.

4.2.3 Stage III - cyclic, α -methylene activation

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 (4)

Considering the hierarchy established so far, statement (4) is valid if and only if the conditions of the previous statements (1-3) are false.

chemical interpretation of statement 4

[464] --> 48.5 +/- 5.8

I

The molecules containing structure (I) above are all active and are shown in table (3d). All are cyclic and possess at least one α -methylene - as required by the proposed mechanism. These structures are presumed to react by way of ring opening during

formation of the alkyldiazohydroxide. The activities of these compounds vary between 39 to 55 standardized log units and show reasonable correlation with a simple count of available α -carbon hydrogens. Since there are only five molecules in this class, more data is needed before this correlation can be considered significant. However, it is worth noting that this somewhat speculative relationship is consistent with the kinetic aspect of the α -hydroxylation mechanism and is in agreement with the results of stage (2) regarding acyclic compounds.

<u>activity</u>	<u>H count</u> *
45	2
39	1
45	2
45	2
55	3

* hydrogen count is log scaled because original data was log scaled

4.2.4 Stage IV - N-nitroso group, moderate activity

Upon removal of the five active compounds with [4 6 4], no subgraph descriptor occurring more than twice is found to be statistically relevant (BiP < .20). Although many of the substructures occurring only twice are 'technically' significant (Bip = .11), this is undoubtedly more an artifact 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 is also based on the justifiable criticism⁸ 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 subgraphs common to the entire data set (see expression(1)) are also common to the remaining data. The concluding statement is :

$$[1 \ 1 \ 2] \quad \text{----->} \quad 33.8 \ +/- \ 18.0 \quad (5)$$

chemical interpretation of statement 5

Alternatively stated, in the absence of any other information (ie., none of the conditions of statements (3-6) are met by the test

compound), the presence of the nitroso group alone, structure (I),

is sufficient for the active label and the numerical prediction of 33.8 +/- 18.0 . A summary of the chemical interpretation of the **SEV** descriptors of statements (1-5) is given in table (1) and the compounds containing these descriptors are shown in tables (3.a-d) with the specific active/inactive substructural features appropriately outlined.

Chemical Interpretation of Nitroso Data Analysis

<u>logical statement</u>	<u>chem. substructure</u>	<u>chemical interpretation</u>
~[112] -> indeterminate	$\begin{array}{c} \backslash \\ \text{N} - \text{N} = \text{O} \\ / \end{array}$	all training compounds contain nitroso group, the presence of this functionality is essential for activity predictions
[424] -> 48.4 +/-6.5	$\begin{array}{c} \text{CH}_3 \\ \\ \text{O}=\text{N}-\text{N}-\text{CH}_2-(\text{C},\text{H}) \end{array}$	SEV originates from methyl substituent to nitroso N, activating
[1 3] -> 53.0 +/-4.5	$\begin{array}{c} (\text{C},\text{H}) \\ \\ \text{C}--\text{CH}--(\text{Cl},\text{Br}) \end{array}$	see table (3.b)
[124] -> inactive	$\begin{array}{c} \text{C} \\ \\ (\text{N},\text{C})--\text{CH}--\text{C} \begin{array}{l} \nearrow \text{O}-(\text{C},\text{H}) \\ \searrow \text{O} \end{array} \end{array}$	see table (3.c)
[464] -> 48.5 +/-5.8	$\begin{array}{c} \text{CH}_2--\text{CH}_2 \\ / \quad \backslash \\ \text{N}--\text{N} \quad \text{CH}_2 \\ \backslash \quad / \\ \text{HC} \quad (\text{C},\text{N}) \\ \quad \quad / \quad \backslash \end{array}$	see table (3.d)
[112] -> 33.8 +/- 18.0		see first statement at top of table

<u>act.</u>	<u>KLN code</u>	<u>compound name</u>
45	RRRRRN/U	NITROSO-PIPERIDINE
39	MDRRRRN/U	2-METHYL-NITROSO-PIPERIDINE
45	RDMDRRN/U	3-METHYL-NITROSO-PIPERIDINE
45	RRDMRRN/U	4-METHYL-NITROSO-PIPERIDINE
10	MDRRDDMN/U	2,6-DIMETHYL-NITROSO-PIPERIDINE
25	RDMDRMRN/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	RRDCMMRRN/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	RDGDGRN/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)TOMRRRN/U	NITROSO-METHYLPHENIDATE
39	RRRRN/U	NITROSO-PYRROLIDINE
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	RDMDMRN/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-DIMETHYL-DINITROSO-PIPERAZINE
10	MDDMNUDMDMN/U	2,3,5,6-TETRAMETHYL-DINITROSO-PIPERASINE
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
55	RRRRRRRN/U	NITROSO-HEPTAMETHYLENEIMINE
45	RRRRRRRRN/U	NITROSO-OCTAMETHYLENEIMINE
39	RRRRRRRRRRRN/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	MDMNUDDMM	DI-isoPROPYL-NITROSAMINE
10	MRDMNUDDMRM	DI-sec.BUTYL-NITROSAMINE
55	MRRNURRM	DI-n.PROPYL-NITROSAMINE
39	RKRNUURK	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

Table 3.a

compounds containing
descriptor [4 2 4]

atom of SEV origin
is circled, complete
subgraph is outlined

Table 3.b

(55) 4-CHLORO-NITROSO-PIPERIDINE

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

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

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

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

compounds containing
descriptor [1 3]

atom of **SEV** origin
is circled, complete
subgraph is outlined

(10) 2-CARBOXY-NITROSO-PIPERIDINE

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

(10) 4-CARBOXY-NITROSO-PIPERIDINE

(10) NITROSO-METHYLPHENIDATE

(10) 2-CARBOXY-NITROSO-PYRROLIDINE

compounds containing
descriptor [1 2 4]

atom of SEV origin
is circled, complete
subgraph is outlined

(45) NITROSO-PIPERIDINE

(55) DINITROSO-HOMOPIPERAZINE

(39) 2-METHYL-NITROSO-PIPERIDINE

(45) 3-METHYL-NITROSO-PIPERIDINE

(45) 3-HYDROXY-NITROSO-PIPERIDINE

compounds containing
descriptor [4 6 4]

atom of SEV origin
is circled, complete
subgraph is outlined

Figure 1

a = [4 2 4]
b = [1 3]
c = [3 7 3] ^ [2 2 6 3] ^ [1 1 2 6 3]

Venn comparison

result

a v b -> active

a --> active

b v c -> active

References

1. Rao, T.K., Lijinsky, W. and Epler, J.L., "Genotoxicology of N-Nitroso Compounds", Plenum Press, N.Y., 1984.
2. Magee, P.N. and Barnes, J.M., Brit. J. Cancer, **10**, 114, (1956).
3. Lijinsky, W., Loo, J. and Ross, A., Nature, **218**, 1174, (1968).
4. Keefer, L.K., Lijinsky, W. and Garcia, H., J. Natl. Cancer Inst., **51**, 299, (1973).
5. Montesano, R. and Bartsch, H., Mutat. Res., **32**, 179, (1976).
6. Singer, G.M. and Lijinsky, W., J. Agric. Food Chem., **24**, 550, (1976).
7. Lijinsky, W. and Taylor, H.W., Cancer Res., **36**, 1988, (1976).
8. Wold, S. and Dunn, W., J. Chem. Inf. Comput. Sci., **23**, 6 (1983).

4.3 Phenylalkylamines - hallucinogenic activity

Phenylalkylamines (I) are a class of compounds that have been of pharmacological interest due mainly to their well-documented property of producing hallucinations and sensory illusions within humans through their action on the central nervous system. Hallucinogens belong to a wider class of compounds, termed psychotomimetics which induce behavior mimicking that of naturally occurring psychoses. With such drugs, the alteration of sensory perceptions is considered to be their primary effect, as opposed to other drugs in which sensory distortion is merely an undesirable side effect. Despite much research, the mechanism(s) of hallucinogenic activity is still uncertain. Continued research into the properties and effects of such substances may provide greater understanding of the biochemistry and neurophysiology of the central nervous system. Advances in this area could have therapeutic applications in the treatment of mental disorders as well. An additional motivation for obtaining more insight into the activity of these compounds stems from the popularity of their nonmedical misuse in today's society.

Phenylalkylamines most closely resemble the neurotransmitter dopamine (II); the basis on which they are often classified as dopamine-like (or norepinephrine-like) hallucinogens, which may infer their role (agonist/antagonist) at the dopamine and norepinephrine receptor sites. Although the mode of action is unknown, all dopamine-like hallucinogens have the common feature of altering the serotonergic transmission, which normally functions to inhibit sensory perceptions.^{1,2}

II

The data set investigated here is composed of 40 phenylalkylamines of variable hallucinogenic activity, as measured in humans by Shulgin³. This same data set, or subsets thereof, has been studied with previous QSAR techniques^{4,5} including Kier and Hall's molecular connectivity (χ) indices⁶. Activity originally was measured in mescaline units (MU) in which hallucinogenic potency is directly proportional to the MU value². For this study, compounds with MU values greater than 3 ($\log(1/C) > 2.16$) were labelled as active and compounds with MU < 3 were correspondingly labelled inactive. The activity values listed in table (2) are calculated from mmol concentrations by the following transformation equation:

$$\text{Activity} = 21.6 [\log(1/C)] - 26.6$$

to provide a convenient rescaling of the data in which the activities now range between 10 and 70 with values greater than 20 labelled as active. This classification scheme produced 27 active and 13 inactive compounds and, as a consequence, provided a highly skewed data set with which to challenge the methodology. The results of the analysis of these compounds is described in the following sections.

4.3.1 Stage I - congeneric data recognition

The data set of phenylalkylamines consists of 40 compounds and their experimentally determined hallucinogenic potency. (see table 2) This data, like the N-nitroso compounds, is an example of a congeneric series. In this case, all the compounds contain the phenylalkylamine backbone, with multiple substituents at various positions. The structural formulas are shown in table (3). It is important with data of this nature to recognize the limitations within which useful predictions can be made. The described methodology addresses this situation during the first stage of analysis wherein all common descriptors are generated for the very purpose of defining these limiting conditions. A list of these commonly occurring SEVs and their distributions is given below:

Total = 40 Actives = 27 Inactives = 13

<u>total</u>	<u>act</u>	<u>inact</u>	<u>SEV</u>
40	27	13	2
40	27	13	3
40	27	13	4
40	27	13	2 5
40	27	13	3 3
40	27	13	3 4
40	27	13	3 5
40	27	13	3 7
40	27	13	4 5

It is noted that since the 2-element SEVs ([25], [33], ...[45]) are subsets of the one-element SEVs ([2], [3], and [4]), the one-element SEVs are unnecessary and in fact are redundant since the subset descriptors already include all the compounds in the data

base. Therefore, we omit the one-element vector descriptors and define a symbol (C) to represent the conditions under which a compound belongs to this same congeneric series.

$$C = [25] \wedge [33] \wedge [34] \wedge [35] \wedge [37] \wedge [45]$$

The presence of the above descriptors in all of the 40 molecules (27 actives and 13 inactives) of the training set results in logical statement (1):

$$\sim C \rightarrow \text{indeterminate} \quad (1)$$

These six common features graph theoretically define the congeneric nature of the data. The absence of any one of these attributes in a test molecule will result in the prediction of 'indeterminate' since this molecule does not belong to the same class of compounds and therefore any further analysis is inappropriate. Thus, a compound must contain all 6 of the above descriptors in order for the predicted activity to be meaningful. It is noted that it is possible for a large and complex structure to contain all 6 descriptors and still not belong to this congeneric series. However, even if a non-congeneric compound does satisfy this condition, the probability is small that this compound also has the appropriate subgraph/subgraphs to be predicted active. In addition to this, test compounds to undergo prediction should always be subjected to visual inspection of their structural formulas when the training set is a congeneric series since non-congeners are easily identifiable.

After isolation of the above subset descriptors, the analysis proceeds in the usual fashion with the result being a series of logical statements that are arranged in hierarchical order and contain substructural descriptors that are statistically associated with hallucinogenic potency.

4.3.2 Stage II - 2-methoxy activation

The second step of the analysis involves the comparison of the following 3 descriptors which appeared at the top of the I(m) value ordered list :

<u>SEV</u>	<u>total</u>	<u>active</u>	<u>inactive</u>
	40	27	13
[3 5 9]	17	17	0
[2 5 4 6]	16	16	0
[4 1 2 4 6]	16	16	0

The interrelationship of these descriptors is given in the Venn diagram analysis below.

As indicated above, the same set of molecules contain both subgraph descriptors **b** and **c**. In addition, these molecules are a subset of the set of molecules that contain subgraph **a**. The conclusion drawn is expressed in logic statement (2).

$$[3 \ 5 \ 9] \quad \text{-----} \rightarrow \quad \text{active} \quad (2)$$

This first activating fragment, [3 5 9], is found only in active compounds and this descriptor alone explains 63% (17/27) of the actives.

Chemical interpretation

The chemical functionality associated with [3 5 9] of statement (2) is given below:

I

A list of the molecules containing (I) is given in table (4). There are 2 features in particular that are associated with this substructural fragment and therefore of use in identifying active molecules simply by visual inspection of the structural formula :

- 1) methoxy at 2 position
- 2) hydrogen at 3 position

Therefore, after verification that a compound belongs to this particular congeneric class (phenylalkylamines, statement 1), fulfillment of the two conditions above are the only requirement for the classification of a test molecule as 'active'. In fact, the presence of these two functionalities alone allows for the correct classification of most of the active molecules in the learning set. This type of simple deduction illustrates one of the main advantages of this type of methodology. Vector descriptors are readily translated into chemical functionalities which provide a reasonable

classification of a molecule's activity much like the way in which organic chemists make judgments about reactivity of a molecule based on the presence of simple familiar functionalities and their position within the overall structure. A useful consequence of the simplicity of the topological descriptors is that the conclusions (ie., logical statements) are also rather simple and do not require computational equipment to apply and can thereby be of use to the chemist working at the bench level.

4.3.3 Stage III - activity attenuating features

The elimination of the 17 active compounds containing [3 5 9] reduces the remaining data set to 23 compounds; 10 actives and 13 inactives. Only one SEV descriptor at the top of the I(m) priority-ordered list of potential descriptors meets the prescribed conditions for acceptance and is shown in statement (3) below.

[3 4 4 8] ---> inactive (3)

This descriptor represents a common substructure which is found within 5 'inactive' (low hallucinogenic potency) molecules and is thus associated with inactivity, as indicated.

chemical interpretation

The chemical equivalence of [3 4 4 8] is given below:

There are 3 easily identifiable chemical features associated with all five of the inactive molecules which contain the above subgraph.

- 1) O-CH₂- (alkoxy) at position 3
- 2) methoxy at position 4
- 3) H at positions 5 and 6

Table(4) gives a list of the molecules containing this substructure.

4.3.4 Stage IV - 2-methoxy-3-alkoxy activating features

Stage IV begins with the reduced data set composed of 10 active and 8 inactive molecules. Upon analysis, four different SEV descriptors are at the top of the I(m) ordering - all with the same I(m) value.

<u>SEV</u>	<u>total</u> 18	<u>active</u> 10	<u>inactive</u> 8
[3 5 10]	6	6	0
[3 7 7]	6	6	0
[2 5 4 7]	6	6	0
[4 1 2 4 7]	6	6	0

As is frequently the case in such situations where activity distributions are identical, these four SEVs are structurally related and thus interdependent, as indicated below:

a = [3 5 10]
b = [3 7 7]
c = [2 5 4 7]
d = [4 1 2 4 7]

From this relationship, statement (4) follows :

$$\mathbf{a} \wedge \mathbf{b} \wedge \mathbf{c} \wedge \mathbf{d} \quad \text{--->} \quad \text{active} \quad (4)$$

chemical interpretation

The chemical interpretation of the four descriptors of statement (4) is shown below :

The four atoms on which the SEVs are centered are shown in bold and can be seen to be adjacent to each other thereby forming a linear substructure. The prominent chemical features common to the six active compounds represented by the vector descriptors are :

- 1) methoxy at position 2
- 1) O-CH₂- (alkoxy) at position 3
- 3) H at positions 6

Table(4) gives a list of the molecules containing this substructure.

4.3.5 Stage V - 3,5-dimethoxy activation

At stage V of the analysis 12 compounds remain, with only 4 labeled active. When the SEV descriptors are I(m)-value ordered, the top descriptors do not satisfy the condition of being statistically significant. At this point the procedure, as illustrated previously, is to discard the I(m) ordering and a BiP (binomial probability) ordering is generated in a final attempt to find statistically significant descriptors (see table 5). Since activity is of greater interest than inactivity, in such a data base, the final descriptor selected is one that occurs in the most (remaining) active molecules and yet is also statistically significant. [3 4 7 8 6], found in three active and one inactive compound (BiP=.11) is just such a descriptor. The resultant statement :

[3 4 7 8 6] -----> active (5)

classifies 3 active compounds correctly but makes one false positive classification.

Chemical interpretation

The chemical equivalent of [3 4 7 8 6] is given below:

The general features of the compounds containing this substructure are:

- 1) H at 2 and 6 positions
- 2) methoxy at 3,5
- 3) all belong to mescaline class

Table(4) gives a list of the molecules containing this substructure.

4.3.6 Conclusion - hallucinogenicity of phenylalkylamines

Upon conclusion of the analysis of the hallucinogenic potency of this class of compounds, substructural descriptors and their corresponding chemical functionalities (pharmacophores) were obtained which properly classified 97% (31/32) of the data to which they were applied. There remain 8 compounds, 1 active and 7 inactive, for which no statistically significant descriptor was produced. Under such conditions, the default classification is 'inactive' which simply means that a test compound belonging to this congeneric series yet not containing any of the appropriate activity-correlated substructures is assumed inactive. In the learning set, this resulted in one false negative (active molecule predicted inactive). The overall classification for the entire data set is 95% (38/40). A summary of the logical statements (applied in hierarchical order) derived from the analysis is given in table (1) (see previous text for chemical interpretations).

Table 1

Vector substructure descriptors for hallucinogenic activity

$$C = [25] \wedge [33] \wedge [34] \wedge [35] \wedge [37] \wedge [45]$$

$$\sim C \longrightarrow \text{indeterminate} \quad (1)$$

$$[3 \ 5 \ 9] \longrightarrow \text{active} \quad (2)$$

$$[3 \ 4 \ 4 \ 8] \longrightarrow \text{inactive} \quad (3)$$

$$\begin{aligned} a &= [3 \ 5 \ 10] \\ b &= [3 \ 7 \ 7] \\ c &= [2 \ 5 \ 4 \ 7] \\ d &= [4 \ 1 \ 2 \ 4 \ 7] \end{aligned}$$

$$a \wedge b \wedge c \wedge d \longrightarrow \text{active} \quad (4)$$

$$[3 \ 4 \ 7 \ 8 \ 6] \longrightarrow \text{active} \quad (5)$$

Table 2

<u>act</u>	<u>KLN code</u>	<u>compound name</u>
10	ARRC2DC2CC2D/OMOMOM	3,4,5-TRIMETHOXYPHENYLETHYLAMINE
10	OROC2C/COM2DC2D/RRA	3-METHOXY-4,5-ORO-PHENYLETHYLAMINE
23	ADMRC2DD2CD2D/OM	4-METHOXYAMPHETAMINE
25	ADMRC2CD2CD2D/OMOM	2,4-DIMETHOXYAMPHETAMINE
30	ADMRC2COMD2DC2D/OM	2,5-DIMETHOXYAMPHETAMINE
18	ADMRC2DC2CC2D/OMOMOM	3,4,5-TRIMETHOXYAMPHETAMINE
38	ADMRC2COMD2COMC2D/OM	2,4,5-TRIMETHOXYAMPHETAMINE
35	ADMRC2CC2DD2C/OMOMOM	2,3,6-TRIMETHOXYAMPHETAMINE
20	OROC2C/D2CD2D/RDMA	3,4-ORO-AMPHETAMINE
20	OROC2C/COM2DC2D/RDMA	3-METHOXY-4,5-ORO-AMPHETAMINE
34	OROC2C/D2COMC2D/RDMA	2-METHOXY-4,5-ORO-AMPHETAMINE
32	OROC2C/COM2CD2D/RDMA	2-METHOXY-3,4-ORO-AMPHETAMINE
20	OROC2C/CRDMA2DD2C/OM	4-METHOXY-2,3-ORO-AMPHETAMINE
35	OROC2C/C2CD2C/OMRDMAOM	2,5-DIMETHOXY-3,4-ORO-AMPHETAMINE
30	ADMRC2CC2CC2D/OMOMOMOM	2,3,4,5-TETRAMETHOXYAMPHETAMINE
52	ADMRC2CD2CC2D/OMMOM	2,5-DIMETHOXY-4-METHYLAMPHETAMINE
54	ADMRC2CD2CC2D/OMRMOM	2,5-DIMETHOXY-4-ETHYLAMPHETAMINE
53	ADMRC2CD2CC2D/OMRRMOM	2,5-DIMETHOXY-4-PROPYLAMPHETAMINE
46	ADMRC2CD2CC2D/OMRRRMOM	2,5-DIMETHOXY-4-BUTYLAMPHETAMINE
70	ADMRC2CD2CC2D/OMBOM	2,5-DIMETHOXY-4-BROMOAMPHETAMINE
14	ADMRC2CC2CD2D/OMOMOM	2,3,4-TRIMETHOXYAMPHETAMINE
10	OROC2C/C2DC2D/RDMAOM	3-METHOXY-4,5-ORRO-AMPHETAMINE
30	ADMRC2CD2CC2D/OMOMORM	4,5-DIMETHOXY-2-ETHOXYAMPHETAMINE
30	ADMRC2CD2CC2D/ORMOMOM	2,4-DIMETHOXY-5-ETHOXYAMPHETAMINE
15	ADRMRC2DC2CC2D/OMOMOM	3,4,5-TRIMETHOXYPHENYL-SEC-BUTYLAMINE
10	ARRC2DD2COMD2D/	4-METHOXYPHENYLETHYLAMINE
24	OROC2C/COM2CRRAD2D/	2-METHOXY-3,4-ORO-PHENYLETHYLAMINE
30	ARRC2DC2CC2D/OMORMOM	4-ETHOXY-3,5-DIMETHOXYPHENYLETHYLAMINE
28	ARRC2DC2CC2D/OMORRMOM	4-PROPOXY-3,5-DIMETHOXYPHENYLETHYLAMINE
38	ARRC2CD2CC2D/OMMOM	2,5-DIMETHOXY-4-METHYLPHENYLETHYLAMINE
46	ARRC2CD2CC2D/OMBOM	2,5-DIMETHOXY-4-BROMOPHENYLETHYLAMINE
10	ARRC2CC2CD2D/OMOMOM	2,3,4-TRIMETHOXYPHENYLETHYLAMINE
50	ARRC2CD2CC2D/OMIOM	4-iodo-2,5-DIMETHOXYPHENYLETHYLAMINE
37	ARRC2CD2CC2D/OMRMOM	4-ETHYL-2,5-DIMETHOXYPHENYLETHYLAMINE
34	ARRC2DC2CC2D/OMSMOM	3,5-DIMETHOXY-4-METHYLTHIOETHER-PEA
24	ADMRC2COMCOM2DC2D/OM	2,3,5-TRIMETHOXYAMPHETAMINE
26	OROC2C/C2CC2D/RDMAOMOM	2,3-DIMETHOXY-4,5-ORO-AMPHETAMINE
34	ADMRC2CD2CC2D/OMRRRRMOM	2,5-DIMETHOXY-4-PENTYLAMPHETAMINE
10	ADMRC2DC2CD2D/OMOM	3,4-DIMETHOXYAMPHETAMINE
10	ARRC2DC2CD2D/OMOM	3,4-DIMETHOXYPHENYLETHYLAMINE

(10) 3,4,5-TRIMETHOXYPHENYLETHYLAMINE

c1

(10) 3-METHOXY-4,5-DIOXY-PHENYLETHYLAMINE

(23) 4-METHOXYAMPHETAMINE

c3

(25) 2,4-DIMETHOXYAMPHETAMINE

(30) 2,5-DIMETHOXYAMPHETAMINE

c5

(18) 3,4,5-TRIMETHOXYAMPHETAMINE

(38) 2,4,5-TRIMETHOXYAMPHETAMINE

c7

(35) 2,3,6-TRIMETHOXYAMPHETAMINE

(20) -3,4-DIOXY-AMPHETAMINE

c9

(20) 3-METHOXY-4,5-DIOXY-AMPHETAMINE

(34) 2-METHOXY-4,5-ORO-AMPHETAMINE

(20) 4-METHOXY-2,3-ORO-AMPHETAMINE

(30) 2,3,4,5-TETRAMETHOXYAMPHETAMINE

(54) 2,5-DIMETHOXY-4-ETHYLAMPHETAMINE

(46) 2,5-DIMETHOXY-4-BUTYLAMPHETAMINE

c11

(32) 2-METHOXY-3,4-ORO-AMPHETAMINE

c13

(35) 2,5-DIMETHOXY-3,4-ORO-AMPHETAMINE

c15

(52) 2,5-DIMETHOXY-4-METHYLAMPHETAMINE

c17

(53) 2,5-DIMETHOXY-4-PROPYLAMPHETAMINE

c19

(70) 2,5-DIMETHOXY-4-BROMOAMPHETAMINE

(14) 2,3,4-TRIMETHOXYAMPHETAMINE

c21

(10) 3-METHOXY-4,5-DIPRO-AMPHETAMINE

(30) 4,5-DIMETHOXY-2-ETHOXYAMPHETAMINE

c23

(30) 2,4-DIMETHOXY-5-ETHOXYAMPHETAMINE

(15) 3,4,5-TRIMETHOXYPHENYL-SEC-BUTYLAMINE

c25

(10) 4-METHOXYPHENYLETHYLAMINE

(24) 2-METHOXY-3,4-DIPRO-PHENYLETHYLAMINE

c27

(30) 4-ETHOXY-3,5-DIMETHOXYPHENYLETHYL

(28) 4-PROPOXY-3,5-DIMETHOXYPHENYLETHYLAMINE

c29

(38) 2,5-DIMETHOXY-4-METHYLPHENYLETHYL

(46) 2, 5-DIMETHOXY-4-BROMOPHENYLETHYLAMINE c31

(10) 2, 3, 4-TRIMETHOXYPHENYLETHYLAMINE

(50) 4-iodo-2, 5-DIMETHOXYPHENYLETHYLAMINE c33

(37) 4-ETHYL-2, 5-DIMETHOXYPHENYLETHYLAMINE

(34) 3, 5-DIMETHOXY-4-METHYLTHIOETHER-PEA c35

(24) 2, 3, 5-TRIMETHOXYAMPHETAMINE

(26) 2, 3-DIMETHOXY-4, 5-DIPRO-AMPHETAMINE c37

(34) 2, 5-DIMETHOXY-4-PENTYLAMPHETAMINE

(10) 3, 4-DIMETHOXYAMPHETAMINE c39

(10) 3, 4-DIMETHOXYPHENYLETHYLAMINE

Table 4

<u>activity statement</u>	<u>molecules classified</u>
[3 5 9] -> active	4, 5, 7, 8,11,16, 17,18,19,20,23,24, 30,31,33,34,38
[3 4 4 8] -> inactive	13,21,32,39,40
a ^ b ^ c ^ d -> active	12,14,15,27,36,37
a=[3 5 10] b=[3 7 7] c=[2 5 4 7] d=[4 1 2 4 7]	
[3 4 7 8 6] -> active	1,28,29,35

Table 5

Total = 12 Actives = 4 Inactives = 8

<u>total</u>	<u>act</u>	<u>inact</u>	<u>Bip</u>	<u>SEV</u>
2	2	0	.1111	2 5 7
2	2	0	.1111	3 5 8 10
4	3	1	.1111	3 4 7 8 6
2	2	0	.1111	3 5 7 8 9
4	3	1	.1111	3 7 6 5 7
2	2	0	.1111	2 5 4 4 8 9
2	2	0	.1111	3 4 7 8 6 6
4	3	1	.1111	3 7 6 5 7 3
4	3	1	.1111	4 5 6 3 3 7
2	2	0	.1111	3 7 6 5 7 3 3
2	2	0	.1111	4 1 2 4 4 8 9
4	3	1	.1111	4 5 2 4 3 3 7
4	3	1	.1111	4 5 6 3 3 7 3
4	3	1	.1111	3 3 3 2 4 3 3 7
4	3	1	.1111	4 5 2 4 3 3 7 3
2	2	0	.1111	4 5 6 3 3 7 3 3
4	3	1	.1111	3 3 3 2 4 3 3 7 3
2	2	0	.1111	4 5 2 4 3 3 7 3 3
2	2	0	.1111	3 3 3 2 4 3 3 7 3 3

References

1. Katzung, B.G., "Basic and Clinical Pharmacology", 4th ed., Appleton & Lange, Norwalk, 1989.
2. Gerald, M.C., "Pharmacology: An Introduction to Drugs", 2nd ed., Prentice-Hall, Englewood Cliffs, 1981.
3. Shulgin, A., Sargent, T. and Naranjo, C., Nature (London), **221**, 537, (1969).
4. Barfknecht, C., Nichols, D. and Dunn, W., J. Med. Chem., **18**, 208 (1975).
5. Klopman, G. and Macina, O., J. Theor. Biol., **113**, 637 (1985).
6. Kier, L. and Hall, L., J. Med. Chem., **20**, 1631 (1977).

4.4 Peptides - thermolysin enzyme inhibitors

Thermolysin is a proteolytic enzyme which acts specifically to hydrolyze peptide bonds of hydrophobic amino acid residues¹. The X-ray crystallographic structure of this enzyme has been determined as well as the crystal structure of a number of thermolysin-inhibitor complexes². In addition, crystallographic studies reveal the enzyme to contain a zinc ion at the catalytic (active) site³. Considerable experimental data has been gathered concerning thermolysin's reactivity through in-depth kinetic study of its inhibition by more than 70 peptides. Due to the detailed information available, thermolysin has been studied as a prototypical enzyme of the endopeptidase class which are known to be involved in very important physiological processes, such as digestion.

Since the regulation of this enzyme is achieved through its interaction with inhibitor compounds, the structural features of these compounds is also of interest. The inhibitors studied here are proposed to mimic natural substrates in their ability to bind with an active site on the thermolysin enzyme, thereby deactivating it.³ The inhibitor compounds in this data set consist of 73 peptides and their derivatives. The activity of these compounds is based on the inhibition constant, K_I , obtained from kinetic studies.³⁻⁶ K_I values range from 19.0×10^{-6} mmol to 250 mmol with inhibition activity being inversely related to the magnitude of the inhibition constant. K_I values less than 1.1 mmol are considered to be 'active'

inhibitors. These activities ($1/K_i$) are then log scaled⁷ to be between 0 - 100 by the following transformation equation :

$$\text{Activity} = 20.0 \log(1/K_i) + 30.4 \quad (\text{eq.1})$$

A list of the compounds, along with their experimentally determined activities, is given in table (2).

4.4.1 Stage I - identification of common substructures

The data set of 73 poly-peptides is comprised of 32 active thermolysin inhibitors and 41 inactive noninhibitor compounds. The average activity of a compound, as calculated by equation (1), is 32 with a standard deviation of 29 activity units. This data provides an initial P_{ij} matrix of 1499 rows, one for each atom of the training data set, which will serve as the source of potential vector descriptors.

The following SEV vector descriptors are found to be generated by every molecule of the data set.

Total = 73 Actives = 32 Inactives = 41

<u>total</u>	<u>act</u>	<u>inact</u>	<u>SEV</u>	<u>chemical interpretation</u>
73	32	41	[1]	terminal nonhydrogen atom
73	32	41	[3]	trivalent atom
73	32	41	[4]	tetravalent atom

The absence of any one of the above common features indicates that a molecule is outside the knowledge realm of the learning set and therefore can be classified (predicted) to be neither active nor inactive.

$$\sim ([1] \wedge [3] \wedge [4]) \text{ ----> indeterminate} \quad (1)$$

The priority rearrangement of SEVs, in the usual manner, reveals two related descriptors at the top of the I(m)-value ordered list.

<u>SEV</u>	<u>total</u>	<u>active</u>	<u>inactive</u>
[2 2 2]	11	11	0
[2 2 2 3]	11	11	0

Since all molecules containing a substructure characterized by [2 2 2 3] are a subset of the universal set of molecules containing substructure [2 2 2], the following statement results:

[2 2 2 3] -----> active (2)

thereby correctly identifying the eleven active compounds shown in table (3). The substructure descriptors of statement (2) can be expanded by one element to obtain a more quantitative classification, as shown in (2a) and (2b),

[2 2 2 3 5] -----> active (activity = 75 +/- 18) (2a)

[2 2 2 3 2] -----> active (activity = 30 +/- 1) (2b)

in which (2a) applies to 9 of the 11 active compounds and (2b) applies to the remaining two, which are both of considerably lower activity.

chemical interpretation of statement 2

The embedded substructure associated with [2 2 2 3] is shown below.

II

It has previously been suggested in the biochemical literature that such a species is a bidentate ligand, through the two oxygen atoms, to a Zn^{2+} active site within the enzyme⁶. A simple Huckel MO representation of the ligand reveals the orbitals of the HOMO to be of the proper symmetry to interact with the empty (LUMO) sp^3 orbitals of the zinc ion.

The chemical substituents responsible for statements (2a) and (2b) are shown in the following table:

position		compound	activity
1	2	number	
N	C	6	45
		7	70
		9	68
		12	82
		14	59
C	C	15	64
		16	94
		17	94
		18	97
N	H	13	31
C	H	19	30

in which it is revealed that those compounds possessing carbons at both positions 1 and 2 are among the highest in inhibitory activity while the compounds with a hydrogen at either position are of low inhibitory activity. This information is also consistent with the proposed models which consider a hydrocarbon side chain adjacent to the hydroxamic acid functionality to be important for binding the substrate to a hydrophobic pocket on the enzyme³.

4.4.3 Stage III - carboxylic group inactivity

Sixty-two compounds, 21 active and 41 inactive, remain after the elimination of the 11 active inhibitors found in stage II. The seven SEVs listed below are at the top of the I(m)-ordered list of potential descriptors and are all associated with inhibitory inactivity (inability to inhibit the thermolysin enzyme).

	SEV	<u>total</u>	<u>active</u>	<u>inactive</u>
1	[3 4 2]	11	0	11
2	[1 2 4 2]	11	0	11
3	[2 2 3 2]	11	0	11
4	[3 4 2 2]	11	0	11
5	[4 4 3 3]	11	0	11
6	[1 2 4 2 2]	11	0	11
7	[2 2 3 2 2]	11	0	11
	all data	62	21	41

The first three descriptors are eliminated from the list because the substructures they represent are also represented by 'subset' vector descriptors also contained in the list - descriptors 4, 6 and 7, respectively. The Venn diagram analysis of the four remaining vector descriptors is given in figure(1) and the results are reported in the logic matrix below.

results of pairwise Venn comparison

a = [3 4 2 2]
b = [4 4 3 3]
c = [1 2 4 2 2]
d = [2 2 3 2 2]

	a	b	c	d
a	...	a^b	a^c	a^d
b		...	b^c	b^d
c			...	c^d
d				...

Thus. the four SEVs are interdependent with none of them significant in the absence of the others. Statement (3) follows.

$$\mathbf{a} \cdot \mathbf{b} \cdot \mathbf{c} \cdot \mathbf{d} \rightarrow \text{inactive} \quad (3)$$

where

$$\mathbf{a} = [3 \ 4 \ 2 \ 2]$$

$$\mathbf{b} = [4 \ 4 \ 3 \ 3]$$

$$\mathbf{c} = [1 \ 2 \ 4 \ 2 \ 2]$$

$$\mathbf{d} = [2 \ 2 \ 3 \ 2 \ 2]$$

chemical interpretation

The ten inactive compounds of statement (3) are all carboxylic acids containing embedded substructure (III).

III

It is interesting to note that the carboxylate anion, a potential ligand to Zn^{2+} , does **not** have HOMO orbitals of the proper symmetry to bond to zinc, which supports their classification as inactive. The structure of the molecules containing (III) are shown in table (4).

4.4.4 Stage IV - phosphoramidate activation

A search for additional activity descriptors is conducted on the reduced data set of 21 active and 31 inactive compounds. These 52 compounds remain yet unclassified while 21 others have been correctly classified by the preceding three stages of the analysis. A list of the best remaining potential descriptors is given below. This list includes three inactive and three active descriptors, as seen by their corresponding distributions.

	<u>SEV</u>	<u>total</u>	<u>active</u>	<u>inactive</u>
1	[3 3 2]	7	0	7
2	[4 4 2]	7	0	7
3	[3 3 2 2]	7	0	7
4	[4 4 3 5]	7	7	0
5	[1 3 4 3 5]	7	7	0
6	[2 3 3 3 5]	7	7	0
all data		52	21	31

Statements (4a) and (4b) result from the Venn diagram analysis that is presented in figure(2).

$$a \wedge b \text{ -----> inactive} \quad (4a)$$

$$c \wedge d \wedge 2e \text{ -----> active} \quad (4b)$$

(73 +/- 21)

where

$$a = [4 \ 4 \ 2]$$

$$b = [3 \ 3 \ 2 \ 2]$$

$$c = [4 \ 4 \ 3 \ 5]$$

$$d = [1 \ 3 \ 4 \ 3 \ 5]$$

$$e = [2 \ 3 \ 3 \ 3 \ 5]$$

chemical interpretation

The following chemical substructure is common to seven inactive compounds and corresponds to the topological descriptors of statement (4a).

Embedded substructure (IVb) is common to the seven active compounds of stage IV, as expressed in statement (4b).

Phosphoramidate derivatives, which contain substructure (IVb), are among the classes of compounds for which x-ray crystal structures have been determined for enzyme-inhibitor complexes⁸. Such data reveals coordination of the thermolysin Zn^{2+} with a phosphoric oxygen. A hydrophobic side chain is also proposed to play a role in the activity of these compounds while the NH of (IVb) serves to hydrogen bond the substrate to a carbonyl of the enzyme. All of this information is consistent with the structures which contain (IVb). The structures of the molecules containing (IVb) are given in table (5).

4.4.5 Stage V - hydrophobic group activation

At this stage of the analysis, there remain 38 compounds in the data set of which only 14 are active inhibitors of thermolysin. The top vector descriptors, arranged according to I(m) value priority, do not meet the criterion for statistical significance ($\text{BiP} < .15$); therefore the I(m) ordering is abandoned and replaced by a binomial probability ordering of potential descriptors. The three most statistically significant descriptors are listed below, along with their binomial probabilities (probability that distribution is due to chance).

<u>SEV</u>	<u>total</u>	<u>active</u>	<u>inactive</u>	<u>BiP</u>
[3 5 8 8]	6	5	1	.028
[4 5 8 7]	6	5	1	.028
[1 2 5 8]	18	11	7	.032
all data	38	14	24	

Although all of the above are significant, only the first two will be considered for further analysis since the third descriptor, [1 2 5 8], would produce 7 false positive classifications (nearly as many incorrect as correct). The comparison of the first two potential descriptors is as follows:

a = [3 5 8 8]

b = [4 5 8 7]

resulting in statement (5).

[3 5 8 8] V [4 5 8 7] -----> active (5)

Statement (5) produces two false positives (inactive compounds falsely classified as active).

chemical interpretation

The embedded substructures below are common to the compounds classified by statement (5).

Va

Vb

By comparing (Va) and (Vb), which both contain a peptide linkage and a hydrophobic substituent alpha to the peptide amide, the following composite substructure can be formed.

Vc

R= hydrophobic group

The presence of a bulky 'R' group in (Vc) is consistent with the proposal of a hydrophobic pocket on the enzyme which serves as a primary binding site³. Generalized substructure (Vc) shows that (Va) and (Vb) have similarities which may mean they form similar substrate-enzyme complexes, although such speculation is currently unsupported by experimental evidence. The structures of the molecules containing (Va) and (Vb) are presented in table (6).

4.4.6 Stage VI - inactivity classification by default

After removal of the compounds classified by statement (5), only five active compounds remain in the reduced data set. Although numerous statistically significant descriptors are revealed in the BiP ordering (shown below), most of these 'active' descriptors make more incorrect than correct classifications.

Total = 27 Actives = 5 Inactives = 22

<u>total</u>	<u>act</u>	<u>inact</u>	<u>BiP</u>	<u>SEV</u>
9	4	5	.0673	4 610 4
9	4	5	.0673	4 9 3 4 4
9	4	5	.0673	4 3 6 3 4 4
3	2	1	.0902	2 2 5
3	2	1	.0902	3 4 3 5 5 9
10	4	6	.0962	3 5 8 7
15	5	10	.1282	4 9
15	5	10	.1282	4 610
15	5	10	.1282	4 9 3
15	5	10	.1282	4 3 6 3
15	5	10	.1282	4 9 3 4
15	5	10	.1282	4 3 6 3 4

This is to be expected in the final stage of the analysis since most the data has already been explained and the relatively few compounds remaining tend to be skewed in activity. For this reason, the five active molecules will remain unexplained. With no more useful descriptors available, the remaining compounds will be classified as inactive; resulting in five false negative classifications.

Statement (6) expresses the situation when a compound is found to contain no statistically significant activating substructures and is thereby assumed inactive, by default.

$$\sim (a \vee (b \wedge c \wedge d) \vee e \vee f) \text{ ----> inactive} \quad (6)$$

where

a = [2 2 2 3]
b = [4 4 3 5]
c = [1 3 4 3 5]
d = [2 3 3 3 5]
e = [3 5 8 8]
f = [4 5 8 7]

4.4.7 Conclusion - thermolysin inhibitor analysis

A data base of thermolysin inhibitor compounds has been analyzed with topology-based activity and inactivity descriptors isolated. The chemical analogs of the isolated descriptors were also given. Of the original 73 compounds, 66 were correctly classified resulting in a 90% success rate with five false negatives and two false positives. Comparisons of the reported substructures with experimental findings reported in the biochemical literature were presented and used to validate the descriptors selected. The result of the analysis is a set of logical statements which correctly classify all but seven compounds of the learning data set. A summary of the logical statements developed from the analysis is given in table (1).

Table 1

$$\sim ([1] \wedge [3] \wedge [4]) \text{ ----> indeterminate} \quad (1)$$

$$[2 \ 2 \ 2 \ 3] \text{ ----> active} \quad (2)$$

$$[2 \ 2 \ 2 \ 3 \ 5] \text{ ----> active} \quad (\text{activity} = 75 \pm 18) \quad (2a)$$

$$[2 \ 2 \ 2 \ 3 \ 2] \text{ ----> active} \quad (\text{activity} = 30 \pm 1) \quad (2b)$$

$$a \wedge b \wedge c \wedge d \text{ ----> inactive} \quad (3)$$

where

$$a = [3 \ 4 \ 2 \ 2]$$

$$b = [4 \ 4 \ 3 \ 3]$$

$$c = [1 \ 2 \ 4 \ 2 \ 2]$$

$$d = [2 \ 2 \ 3 \ 2 \ 2]$$

$$a \wedge b \text{ ----> inactive} \quad (4a)$$

$$c \wedge d \wedge 2e \text{ ----> active} \quad (4b)$$

$$(\text{activity} = 73 \pm 21)$$

where

$$a = [4 \ 4 \ 2 \ 2]$$

$$b = [3 \ 3 \ 2 \ 2]$$

$$c = [4 \ 4 \ 3 \ 5]$$

$$d = [1 \ 3 \ 4 \ 3 \ 5]$$

$$e = [2 \ 3 \ 3 \ 3 \ 5]$$

$$[3 \ 5 \ 8 \ 8] \vee [4 \ 5 \ 8 \ 7] \text{ ----> active} \quad (5)$$

$$\text{none of the above} \text{ ----> inactive} \quad (6)$$

Table 2

<u>activity</u>	<u>KLN code</u>	<u>peptide sequence</u>
14	C2DD2DD2D) ROTEEC2SEDTEARDMM	Z-NHNH-CS-LEU-NHNH2
39	C2DD2DD2D) ROTEEETEDTEARDMM	Z-AGLY-LEU-NHNH2
30	C2DD2DD2D) ROTERTEDTEARDMM	Z-GLY-LEU-NHNH2
14	C2DD2DD2D) RNTEDTEARDMMETDMETM	AC-ALA-APHE-LEU-NHNH2
12	MTEDTEDTENTEDTEARDMMMM	AC-ALA-ALA-Aala-LEU-NHNH2
45	ADTEKRDMM	L-LEU-NHOH
70	C2DD2DD2D) ROTEDTEKRDMM	Z-L-LEU-NHOH
10	C2DD2DD2D) ROTERTEDTARDMM	Z-GLY-L-LEU-NH2
68	C2DD2DD2D) ROTERTEDTEKRDMM	Z-GLY-L-LEU-NHOH
23	C2DD2DD2D) ROTERTEDTNMKRDMM	Z-GLY-L-LEU-N(CH3)OH
10	C2DD2DD2D) ROTERTEDTEOMRDMM	Z-GLY-L-LEU-NHOCH3.n.i.
82	C2DD2DD2D) ROTEEETEDTEKRDMM	Z-AGLY-L-LEU-NHOH
31	C2DD2DD2D) ROTERTERTEK	Z-GLY-GLY-NHOH
59	C2DD2DD2D) ROTERTERTEDTEKRDMM	Z-GLY-GLY-L-LEU-NHOH
64	C2DD2DD2D) RDTORMTEK	HONH-B2M-OET
94	C2DD2DD2D) RDTEDTERTAMTEK	HONH-B2M-L-ALA-GLY-NH2
94	C2DD2DD2D) RDTEDTERTKMTEK	HONH-B2M-L-ALA-GLY-OH
97	MDRDTEDTERTAMTEKM	HONH-IBM-L-ALA-GLY-NH2
30	KETRTEDTERTAM	HONH-MAL-L-ALA-GLY-NH2
38	C2DD2DD2D) RDTEDTERTAMTK	HO-B2M-L-ALA-GLY-NH2
93	C2DD2DD2D) RDTEDTERTAMRJ	HSCH2CH(CH2C6H5)CO-L-ALA-GLY-NH2
21	C2DD2DD2D) RDTEDTERTAMETRK	HOCH2CO-L-PHE-L-ALA-GLY-NH2
21	C2DD2DD2D) RDTEDTERTAMETRTM	CH3COCH2CO-L-PHE-L-ALA-GLY-NH2
82	C2DD2DD2D) RDTEDTERTAMEWKK	(-O)2(O=)P-L-PHE-L-ALA-GLY-NH2
73	C2DD2DD2D) RDTEDTERTAMEWKJ	-O(-S)(O=)P-L-PHE-L-ALA-GLY-NH2
18	C2DD2DD2D) RDTEDTERTAMEY1K	-O(O=)2S-L-PHE-L-ALA-GLY-NH2
10	C2DD2DD2D) RDTEDTERTAMEY1M	CH3SO2-L-PHE-L-ALA-GLY-NH2
53	WEDRDMMTAKK	P-LEU-NH2
10	MREWKK	P-NH-ET
10	WEDRDMMTAO1MO1C2DD2DD2D)	MEOP0OC6H5-LEU-NH2
53	C2DD2DD2D) RDTKEWKK	P-PHE-OH
51	MDTEDTKMEWKK	P-ALA-ALA-OH
99	MRDDTEDTKMEWKKM	P-ILE-ALA-OH
99	C2DD2DD2D) RDETDRDMMEWKKT	P-LEU-PHE-OH*
99	MD1O1DDDD/KKK1O1WKEDRDMMTEDTKRC2DEC2C/D2DD2D)	PHOSPHORAMIDON*
79	HTN1K1DTEdTERTAMRDMM	CHO-OHLEU-L-ALA-GLY-NH2
20	MTN1K1DTEdTERTAMRDMM	AC-OHLEU-L-ALA-GLY-NH2
39	C2DD2DD2D) ROTEDTERTARC2DD2DD2D)	Z-PHE-GLY-NH2
17	C2DD2DD2D) ROTEDTERTKRC2DD2DD2D)	Z-PHE-GLY
10	C2DD2DD2D) RDTERTAA	PHE-GLY-NH2
10	C2DD2DD2D) RDTERTKA	PHE-GLY
21	C2DD2DD2D) ROTEDRDMMTERTA	Z-LEU-GLY-NH2
18	C2DD2DD2D) ROTEDRDMMTERTK	Z-LEU-GLY
12	ADRDMMTERTA	LEU-GLY-NH2
10	ADRDMMTERTK	LEU-GLY

10	ADRDMMTERTERTK	LEU-GLY-GLY
10	C2DD2DD2D) ROTERTEDTKRC2DD2DD2D)	Z-GLY-PHE
10	C2DD2DD2D) TERTEDTKRC2DD2DD2D)	BZ-GLY-PHE
10	C2DD2DD2D) RDETRATA	GLY-PHE-NH2
10	C2DD2DD2D) RDETRATK	GLY-PHE
15	C2DD2DD2D) ROTERTEDRDMMTK	Z-GLY-LEU
10	ARTEDRDMMTA	GLY-LEU-NH2
10	ARTEDRDMMTK	GLY-LEU
14	C2DD2DD2D) ROTEDTERTKRC2DD2CD2D/K	Z-TYR-GLY
44	C2DD2DD2D) ROTEDTERTARC2DD2CD2D/K	Z-TYR-GLY-NH2
10	C2DD2DD2D) ROTERTEDTKRC2DD2CD2D/K	Z-GLY-TYR
10	ARTEDTARC2DD2CD2D/K	GLY-TYR-NH2
17	C2DD2DD2D) ROTEDTERTKRC2DEC2C/D2DD2D)	Z-TRP-GLY
10	C2DD2DD2D) ROTERTEDTKRC2DEC2C/D2DD2D)	Z-GLY-TRP
10	C2DD2DD2D) ROTERTEDMTK	Z-GLY-ALA
13	C2DD2DD2D) ROTEDMTERTK	Z-ALA-GLY
32	ADRDMMTEDRDMMTK	LEU-LEU
24	C2DD2DD2D) RDATEDTKRC2DD2DD2D)	PHE-PHE
10	KTRDATEDTOMRC2DD2DD2D)	ASP-PHE-OME
10	ARTNDTKRRR)	GLY-PRO
10	ARTERTK	GLY-GLY
10	C2DD2DD2D) TERTEDTKRRRRA	BZ-GLY-LYS
10	C2DD2DD2D) ROTERTEDTKRREC2EA	Z-GLY-ARG
10	C2DD2DD2D) ROTERTK	Z-GLY
10	MTERTK	AC-GLY
10	ARTA	GLY-NH2
10	ADRDMMTA	LEU-NH2
10	C2DD2DD2D) RDTKA	L-PHE

Table 3

(45) L-LEU-NHOH

(70) Z-L-LEU-NHOH

(66) Z-GLY-L-LEU-NHOH

(82) Z-GLY-L-LEU-NHOH

(31) Z-GLY-GLY-NHOH

(59) Z-GLY-GLY-L-LEU-NHOH

(84) HONH-BZM-OET

(94) HONH-BZM-L-ALA-GLY-NH2

(94) HONH-BZM-L-ALA-GLY-OH

(97) HONH-IBM-L-ALA-GLY-NH2

Table 4

Table 5

(10) ASP-PHE-ONE

Figure 1

Venn analysis of stage III vector descriptors

a = [3 4 2 2]
b = [4 4 3 3]
c = [1 2 4 2 2]
d = [2 2 3 2 2]
A = **a** ^ **c** ^ **d**

Venn comparison

result

a ^ **c** ^ **d** -> inactive

A ^ **b** --> inactive

Figure 2

Venn analysis of stage IV vector descriptors

a = [4 4 2]

b = [3 3 2 2]

c = [4 4 3 5]

d = [1 3 4 3 5]

e = [2 3 3 3 5]

Venn comparison

result

a ^ b -> inactive

c ^ d ^ e --> active

References

1. Morihara, K., Biochem. Biophys. Res. Commun., **26**, 656 (1967).
2. Holmes, M. and Matthews, B., J. Molec. Biol., **160**, 623 (1982).
3. Nishino, N. and Powers, J., Biochemistry, **17**, 2846 (1978).
4. Feder, J., Brougham, L. and Wildi, B., Biochemistry, **6**, 1186 (1974).
5. Kam, C., Nishino, N. and Powers, J., Biochemistry, **18**, 3032 (1979).
6. Nishino, N. and Powers, J., Biochemistry, **18**, 4340 (1979).
7. Klopman, G. and Bendale, R.D., J. Theor. Biol., **136**, 67 (1987).
8. Weaver, L., Kester, W. and Matthews, B. J. Molec. Biol., **114** 119 (1977).

Chapter V

VALIDATION

5.1 Validation - Introduction

The proposed SEV methodology has been demonstrated through its application to four separate data bases composed of molecules of distinctly different structure and biological/chemical activity. In each case, an attempt was made to establish the legitimacy of the computer-generated substructural descriptors based on chemical arguments. However, although such a procedure may demonstrate the apparent success of the methodology when applied to a specific data base, it does not fulfill the need for a generalized validation procedure that can routinely be employed as part of the analysis. The validation procedure should be standardized so that it can be conducted in the same way each time, regardless of the nature of the data being analyzed.

One such procedure is the establishment of appropriate statistical criteria for the discrimination of an actual (real) data base from a randomized data set¹. The randomized set is comprised of the same compounds as the original data, but with the activities scrambled in order to insure that they are randomly redistributed among different compounds within the data set². Since the compounds of the randomized data set are the same as those of the original data set, the substructures generated by the methodology are identical. The difference will be in the distribution (the number of active and inactive molecules) associated with each potential descriptor. For a real data set, the activity distributions are assumed to convey meaningful information regarding the relevance of each potential

descriptor in establishing a structure-activity correlation. For a truly random data set, on the other hand, there no longer exists a correlation between structure and activity. Therefore, the activity distributions of the random set are meaningless and should be discerned as such by some appropriate sample statistic. The inability to find such a statistical parameter to distinguish real from random data would raise a serious question regarding the stability of the methodology. If the analysis of randomly generated data appears to proceed as successfully as that of real data, the method would then have the dubious distinction of being able to generate 'virtual' descriptors for activity/inactivity even where there exists no real or theoretical grounds³.

The validation procedure is as follows. A 'real' data set will be assigned random activity values obtained by scrambling the original activities and then reassigning them to different compounds within the set. The randomized data set will then undergo stage 1 of the analysis in which an initial list of potential descriptors will be generated. Under ideal conditions, the binomial probability (BiP) of the activity distribution associated with each descriptor should be statistically insignificant, thereby revealing the descriptor to be 'random'. However, since such a large number of descriptors are generated, it is anticipated that a few random descriptors will have distributions that are statistically significant even though they are merely due to chance. The expected difference between the real and random data is in the relative number of statistically significant

descriptors generated by the data sets. The real data is expected to produce far more statistically significant descriptors than the random data set. By contrasting the relative number of significant descriptors generated, a discrimination between real and random data should become apparent.

An initial investigation will be carried-out in which frequency distribution data is created by counting the number of potential descriptors that fall within given binomial probability ranges. The relevance of this count is then accessed by considering it to be a function of the BiP. This procedure will allow a direct contrast to be made between the real and random data sets through the frequency distributions they generate. The data bases employed for the real/random comparisons involve the same compounds that were introduced earlier (chapter iv). For convenience, the following notation will be used to represent the different data types:

- acids** - pKa acidity data, assorted organic acids and nonacids
- nitroso** - N-nitroso compounds, carcinogenicity activity
- halluc** - hallucinogenic activity data, phenylalkylamines
- thermo** - thermolysin inhibition data, peptides and derivatives

5.2 Validation - real/random distribution profiles

To begin the study, the original acid data file of 122 acids/nonacids (4.1, table 2) is randomized, as described previously, using a VAX random number generator subroutine (Call RANDU). The result is a scrambled data file (table 2) in which activities (pKa values) are randomly assigned to different compounds within the data file. Thus, the total, parent, distribution (number of active and number of inactive molecules) remains the same while the distributions associated with each individual descriptor are changed. The vector descriptors generated, along with their distributions, are ordered according to their statistical significance and are listed in tables (3&4) for both the real and random data sets, respectively. A count of the number (frequency) of descriptors in each BiP range for both the real and randomized data sets is given in the table below :

<u>BiP range</u>	<u>NUMBER OF DESCRIPTORS IN BIP RANGE</u>	
	<u>random data</u>	<u>real data</u>
.00 - .10	6	96
.10 - .20	20	26
.20 - .30	33	21
.30 - .40	50	61
.40 - .50	110	143

A graph of the histogram of this data is given in fig.(1). The one-tailed, 'normal' aspect of the random data set is as anticipated. This is the graph to be expected from data generated by "chance". As shown, the population of descriptors with low BiP is small and increases with each range of higher probability. It is noted from the

table above, as well as the frequency distribution of fig. (1), that there are six low probability descriptors generated from the random data that seem to be statistically significant ($\text{BiP} < .10$). These six are simply due to chance and indicate that it is possible for descriptors in the real data set to be due to chance as well - even low probability descriptors. The important feature to recognize, however, is the vast difference in the population within the low probability range ($\text{BiP} = .00 - .10$) between the real and random sets. Whereas only six low probability descriptors were generated by the random set, 96 were generated by the real data set within the same BiP range. It is this low probability region that allows easy discrimination between real and random data sets. This is a necessary condition considering that there is a causal link between structure and activity within 'real' data that will lead to far more structure-activity correlations than expected from chance. With the random data set, of course, there exists no structure-activity link.

The same relationship that was shown for the descriptor distribution profile of the acid/random data sets in fig.(1) was also found to be true for the three other data sets upon comparing their real and random distribution profiles. In each case, it is the relative population of the low probability region of the BiP scale that readily discriminates real from random data. Specifically, it is the relative number of descriptors with activity distributions in the $\text{BiP} < .10$ and $\text{BiP} < .05$ ranges that reveals a considerable difference

between real and randomized data. The frequency distribution information for both these low probability ranges is given below:

NUMBER OF DESCRIPTORS IN BIP RANGE		
	BiP < .10	BiP < .05

acid data		
real	96	81
random	6	4
nitroso data		
real	15	9
random	2	0
halluc. data		
real	40	30
random	16	9
thermo. data		
real	129	70
random	29	4

Although the same trend in each real/random data comparison is clearly evident, such a comparison remains only qualitative in the absence of an appropriate statistic.

5.3 Validation - calculation of t-statistic

In order to quantify the discrimination between real and random data, several additional randomizations are performed on each different data type. This is to provide a sample of the random population associated with each of the four different data types. The SEV descriptors of BiP<.10 and their distributions are given in tables (5-8) for the real and five random sets associated with each data type. Both the real and random data sets are numerically characterized by the number of low probability descriptors which they contain. Each random set is expected to be similar to the other random sets, yet considerably different from the real set. This data is shown in table(1) and plotted in figs.(2a-d). In every case the real data is easily distinguished from the random data.

Each activity class - acid, nitroso, halluc and thermo - now has five random sets created from the original data. The mean and standard deviation of these five will serve as a sample of the "random population" - the frequency distribution that would result from generating all possible random sets. The hypothesis (H_0) will then be posed that the real data also belongs to the random set. Obviously, the rejection of H_0 is desired to establish validation. The confidence level at which H_0 can be rejected will then be calculated based on a one-tailed t-test⁴. The alternative hypothesis (H_1) is that the real data contains significantly more descriptors and does not belong to the random set. The desired acceptance of H_1 can only be established through the rejection of H_0 . This described

procedure is very similar to that employed in analytical chemistry for the rejection of questionable experimental data.^{5,6} In both situations, the t-test is appropriate when considering samples that are small in size⁷. In its application here, the one-tailed t-test is appropriate since the acceptance of H_1 is only of interest when the real data is considerably "better" (greater number of descriptors) than expected from a random set. The implementation of the one-tailed test, as dictated by the nature of the data, allows a confidence level⁸ to be established for the rejection of H_0 .

The t-distribution for the random samples is given by the function Y in equation(1).

$$Y = Y_0 [1 + t^2/(n-1)]^{-n/2} \quad (1)$$

where

$$t = (x - x_0)/s \sqrt{n} \quad (2)$$

and

Y_0 = normalization constant

n = size of sample

x_0 = mean of random sample (average # of descriptors with BiP < .05)

x = # of descriptors (BiP<.05) being tested for significance

s = standard deviation of random sample

The t statistic, in this application, is a measure of the distance on the x axis (x = number of potential descriptors) between the real data point and the mean of the random sample. This distance is expressed in units of standard deviations and is related to a probability (confidence level) through the function Y.

From eq(2), the t statistic can be calculated for each real data set, in which case x and x_0 equal the number of potential descriptors

with $\text{BiP} < .05$ for the real and random data, respectively (other cutoff values can be used quite successfully, as well). A lower limit of acceptability for t can be determined by assuring that H_0 is accepted for all 20 random data sets upon submitting them to the t -test. This is a reasonable procedure because all the random sets are in fact sampled from the random population; therefore H_0 is already known to be true. For all the random sets generated, H_0 will be accepted with $t=4.60$ (99.5% confidence level). This value will thus be taken as the lower limit which must be exceeded by the real data for rejection of H_0 . In summary,

$t < 4.60 \quad \rightarrow \quad H_0 \text{ accepted, random data}$

$t > 4.60 \quad \rightarrow \quad H_1 \text{ accepted, nonrandom data}$

The calculated values of t for each real data set are given below along with the confidence level at which H_0 is rejected and H_1 is accepted.

<u>real data set</u>	<u>t statistic</u>	<u>confidence level at which H_0 rejected / H_1 accepted</u>
acid	67.3	> 99.9 %
nitroso	10.6	> 99.9 %
halluc	12.0	> 99.9 %
thermo	26.6	> 99.9 %

Obviously, these t values have greatly exceeded the imposed lower limit of acceptability. Considering that $t=7.17$ is required for a 99.9% confidence level, the fact that all the calculated t values are greater means they have exceeded this cutoff as well and are therefore significant even at a greater than 99.9% level⁸. In all cases H_0 is clearly rejected and H_1 accepted. We conclude that the

real data is successfully discriminated from the random data by showing quantitatively that the number of descriptors generated by the real data is far larger than what can be expected from chance. It can now be stated, at a greater than 99.9% confidence level, that the number of potential descriptors generated by the real data can not be due to random structure-activity correlations.

As a further test of this validation procedure, an additional (6th) random set was generated for each of the activity classes and tested in the same way the real data set was tested. The point of this exercise is to demonstrate that the implementation of the one-tailed t-test will correctly lead to the acceptance of H_0 when the tested data is indeed random. The results are shown below:

<u>random data set</u>	<u># of descriptors</u> (BiP<.05)	<u>t statistic</u>
acid	9	3.2
nitroso	2	1.1
halluc	1	---
thermo	9	---

The lower limit of t with which H_0 may be rejected was established earlier; $t=4.60$ (99.5% conf. level). Thus, for each of the additional random sets, H_0 is accepted thereby correctly identifying the randomized data as belonging to the random population distribution.

5.4 Conclusion

The integrity of the substructural descriptors (SEVs) employed during the expert system analysis is established through the validation methodology introduced in this chapter. The procedure first involved the examination of the frequency distribution data in which the number of descriptors initially generated is taken as a function of their binomial probability. Real and random frequency distributions were then compared. This vividly demonstrated that the low probability region was useful for the discrimination between real and random data. Five random data sets were then created for each of the four real data bases studied. Both the real and random data sets were characterized by the number of low probability descriptors ($BiP < .05$) which are initially generated. Each random set provides a sample of the corresponding random population from which it was selected. The mean and standard deviation of the random samples were calculated and the t test was applied to establish the confidence level at which the real data is successfully discriminated from the random population. Both real and additional random sets were subjected to this test.

All four real data sets, produced t statistics which established that the potential descriptors employed are not due to random structure-activity correlations (H_0 rejected at a confidence level greater than 99.9%). For each random set, H_0 was accepted thereby verifying that the t test is selective in that it correctly classified both the real and random data tested. This successful

classification of both real and random data is sighted as a quantitative validation of the expert system analysis introduced in Chapter III.

In addition, the validation procedure is general and can therefore be carried-out on any data base being considered for analysis. The failure of a 'real' data set to pass the t-test (H_0 accepted, type II error) would indicate that either -

1) there are excessive experimental errors or inconsistencies within the data which obscure the structure-activity relationship

or

2) the descriptors employed by the method of analysis are inappropriate for the data being studied. (i.e. the factors responsible for activity are unrelated to the substructures being represented)

In either case, failure to meet the validation criteria would indicate that the analysis will be unstable⁹ and should not proceed without modification of the data set being analyzed.

acidity data - "random"

acidity data - "real"

Descriptors of Bip < .05

Descriptors of Bip < .10

Descriptors of Bip < .05

Descriptors of Bip < .10

Descriptors of Bip < .05

Descriptors of Bip < .10

Thermo Data - real/random comparison

Descriptors of Bip < .05

Descriptors of Bip < .10

Table 1

NUMBER OF DESCRIPTORS IN BIP RANGE		
	Bip < .10	Bip < .05

acid data		
real	96	81
random 1	6	4
random 2	14	9
random 3	13	3
random 4	20	7
random 5	13	4

nitroso data		
real	15	9
random 1	2	0
random 2	7	0
random 3	5	3
random 4	1	0
random 5	5	3

halluc data		
real	40	30
random 1	16	9
random 2	0	0
random 3	11	0
random 4	13	7
random 5	15	9

thermo data		
real	129	70
random 1	29	4
random 2	32	16
random 3	53	8
random 4	39	15
random 5	35	11

Table 2

Randomized Acidity (pKa) data
 [for original 'real' data, see 4.1, table 2]
 122 activities randomized

<u>act</u>	<u>KLN code</u>	<u>formula</u>
10	XDXX	CH(N02) 3
10	DYMYMYM	CH(SO2 CH3) 3
44	DC3NC3NC3N	CH(CN) 3
41	FCTKFF	F3 CCOOH
10	GCTKGG	C13 CCOOH
10	KTK	HOCC00H
48	GDTKG	C12 CHCO2H
10	RXTK	CH2NO2 COOH
10	FRTK	F CH2COOH
10	KTRTK	HOOC CH2 COOH
45	GRTK	C1 CH2CO2H
10	MRDGTK	CH3CH2CH C1 COOH
10	X*TK	NO2-Ph-COOH
10	KRTK	HOCH2C00H
10	MC2DD2DD2CTK/	o-TOLUIC ACID
10	MDGRTK	CH3CH C1 CH2COOH
10	XRK	CH2(N02) 2
10	G*TK	C1-Ph-COOH
56	B*TK	Br-Ph-COOH
48	KT=2TK	HOOC(CH2) 2 COOH
10	*TK	Ph-COOH
10	KT=3TK	HOOC(CH2) 3 COOH
44	M*TK	p-TOLUIC ACID
10	KT=4TK	HOOC(CH2) 4 COOH
10	MO*TK	MeO-Ph-COOH
10	KT=5TK	HOOC(CH2) 5 COOH
10	KT=6TK	HOOC(CH2) 6 COOH
10	KT=7TK	HOOC(CH2) 7 COOH
10	KT=8TK	HOOC(CH2) 8 COOH
50	G=3TK	C1 (CH2) 3 COOH
10	MTK	CH3CO2H
10	M=2TK	CH3 (CH2) 2COOH
42	MRTK	CH3CH2COOH
50	M=3TK	CH3 (CH2) 3 COOH
10	DTMTMTM	CH(C=O CH3) 3
77	*J	Ph-SH
50	MTRTM	CH3C=O CH2C=O CH3
10	*K	PHENOL
10	RC3NC3N	CH2 (CN) 2
10	RYMYM	CH2 (SO2 CH3) 2
10	MRK	CH3CH2OH
10	RD2DD2D/	CYCLOPENTADIENE
10	X*A	NO2-Ph-NH2

10	D* C2DD2DD2C)C2DD2DD2C) /	9-PHENYL FLUORENE
10	RD2DC2DD2DD2C) /	INDENE
42	MTM	CH3C=O CH3
10	RC2DD2DD2C)C2DD2DD2C) /	FLUORENE
77	*C3D	PHENYLACETYLENE
10	C2DR* * *	1,1,3-TRIPHENYLPROPENE
10	*A	Ph-NH2
10	M*A	Me-Ph-NH2
10	MYM	CH3SO2 CH3
61	D* * *	(Ph) 3 CH
10	*R*	(Ph) 2 CH2
10	M*	TOLUENE
10	D2DD2DD2D/	BENZENE
45	RRR/	CYCLOPROPANE
45	MDM*	CUMENE
46	DC2DD2DD2C)DC2DD2DD2C) /C2DD2DD2C/)	TRIPTICENE
10	DC2C/* * *	TRIPHENYLCYCLOPROPENE
10	RRRR/	CYCLOBUTANE
10	R=4/	CYCLOPENTANE
10	R=5/	CYCLOHEXANE
30	=4	BUTANE
10	=5	PENTANE
10	=6	HEXANE
10	=7	HEPTANE
10	=8	OCTANE
10	=9	NONANE
10	=9=6	C 15 H 32
50	=9=9=2	C 20 H 42
51	=3TH	BUTANAL
48	=4TH	PENTANAL
10	=5TH	HEXANAL
88	=6TH	HEPTANAL
10	=7TH	OCTANAL
10	=8TH	NONANAL
10	MORM	CH3-O-CH2CH3
10	MRORM	(CH3CH2) 2 O
10	MRRORM	(CH3CH2CH2) 2 O
10	MRA	CH3CH2 NH2
52	MRRRA	CH3CH2CH2 NH2
10	MRRRA	CH3CH2CH2CH2 NH2
45	*RA	Ph-CH2-NH2
10	MR*A	CH3CH2-Ph-NH2
10	G*RA	C1-Ph-CH2-NH2
83	R2DRA	CH2=CHCH2 NH2
64	MRRK	CH3CH2CH2 OH
10	MRRRK	CH3 (CH2) 3 OH
73	MRRDKM	CH3 (CH2) 2CHOH CH3
41	D2DRRR)	CYCLOPENTENE
47	D2DRRRR)	CYCLOHEXENE
90	MTRM	CH3 CO CH2CH3
10	MRTRM	(CH3CH2) 2 CO

51	*TM	Ph-CO-CH3
10	*RTM	Ph-CH2-CO-CH3
10	R2DRM	CH2=CHCH2CH3
10	MD2DRM	CH3CH=CHCH2CH3
10	FCFFM	F3 CCH3
10	GCGGRK	C13 CCH2 OH
10	ARTK	glycine
10	MDATK	alanine
10	MDMDATK	valine
10	MDMRDATK	leucine
10	MRMDATK	isoleucine
10	KRDATK	serine
10	MDKDATK	threonine
62	KTRDATK	aspartic acid
10	ATRDATK	asparagine
90	KTRRDATK	glutamic acid
10	ATRRDATK	glutamine
46	A=4DATK	lysine
10	ARDKRRDATK	hydroxylysine
61	C2DN2DE/RDATK	histidine
30	AC2EERRRDATK	arginine
10	*RDATK	phenylalanine
10	K*RDATK	tyrosine
10	D2DD2DC2C/ED2C/RDATK	tryptophan
90	JRDATK	cysteine
10	MSRRDATK	methionine
10	DRRRE/TK	proline
10	DRDKRE/TK	hydroxyproline

Table 3

Acid data - real BiP<.50

Total = 122 Actives = 38 Inactives = 84

total act inact Bip SEV descriptors

54	33	21	.0000	2 2
18	17	1	.0000	1 2 4 3
15	15	0	.0000	2 2 3 3
12	12	0	.0000	1 2 4 3 3
12	12	0	.0000	2 2 3 3 3
8	8	0	.0001	2 2 2
8	8	0	.0001	3 3 4
8	8	0	.0001	1 2 3 4
69	36	33	.0002	1 2
47	27	20	.0002	1 2 4
65	34	31	.0003	2
7	7	0	.0003	3 6 4
7	7	0	.0003	2 2 2 4
44	25	19	.0004	2 2 3
16	12	4	.0004	4 5 4
24	16	8	.0004	3 4 3 3
76	38	38	.0005	1
6	6	0	.0009	3 4 0
6	6	0	.0009	4 2 1
6	6	0	.0009	1 2 4 0
6	6	0	.0009	2 2 3 0
6	6	0	.0009	3 3 4 3
6	6	0	.0009	3 6 4 1
6	6	0	.0009	4 2 1 0
6	6	0	.0009	1 2 3 4 3
6	6	0	.0009	2 2 2 4 3
6	6	0	.0009	3 3 4 3 1
6	6	0	.0009	1 2 3 4 3 1
6	6	0	.0009	1 2 4 3 3 3
6	6	0	.0009	2 2 2 4 3 1
6	6	0	.0009	2 2 3 3 3 3
5	5	0	.0029	1 3 2
5	5	0	.0029	1 3 2 1
5	5	0	.0029	1 3 2 1 0
17	11	6	.0045	1 2 3
13	9	4	.0053	1 3
19	1	18	.0080	4 6 3
4	4	0	.0094	4 5 4 0
4	4	0	.0094	4 6 2 1

4	4	0	.0094	4 3 3 2 1
4	4	0	.0094	4 6 2 1 0
4	4	0	.0094	1 2 4 3 3 0
4	4	0	.0094	2 2 3 3 3 0
4	4	0	.0094	3 4 3 3 3 3
4	4	0	.0094	4 3 3 2 1 0
4	4	0	.0094	1 2 4 3 3 3 3
4	4	0	.0094	2 2 3 3 3 3 3
18	1	17	.0111	3 5
18	1	17	.0111	4 3 3 3
12	0	12	.0114	4 6 3 3
12	0	12	.0114	4 3 3 3 3
8	6	2	.0138	4 5 4 3
11	0	11	.0165	3 3 3
11	0	11	.0165	3 4 3 1 3
74	32	42	.0189	3 4
26	3	23	.0195	4 4
10	0	10	.0239	4 6 6 3
41	19	22	.0294	3 4 3
3	3	0	.0302	3 4 3 3 1
3	3	0	.0302	3 4 5 2 0
3	3	0	.0302	3 6 4 1 0
3	3	0	.0302	4 6 5 4 3
3	3	0	.0302	3 3 4 3 1 0
3	3	0	.0302	3 4 3 1 2 1
3	3	0	.0302	3 4 3 3 1 0
3	3	0	.0302	4 5 4 3 3 3
3	3	0	.0302	1 2 3 4 3 1 0
3	3	0	.0302	2 2 2 4 3 1 0
3	3	0	.0302	3 4 3 1 2 1 0
3	3	0	.0302	3 4 3 3 3 3 3
3	3	0	.0302	1 2 4 3 3 3 3 3
3	3	0	.0302	2 2 3 3 3 3 3 3
9	6	3	.0305	3 4 3 3 3
9	0	9	.0348	3 3 3 3
9	0	9	.0348	4 6 3 3 3
9	0	9	.0348	4 3 3 3 3 3
5	4	1	.0353	4 6 5 4
5	4	1	.0353	3 4 3 3 0
5	4	1	.0353	4 5 4 3 3
19	2	17	.0371	3 4 6
48	9	39	.0403	4 3
8	0	8	.0505	3 5 3
8	0	8	.0505	4 5 3 3
13	1	12	.0538	4 5 3
22	3	19	.0546	4 7
21	3	18	.0697	3 7
21	3	18	.0697	1 2 4 5
21	3	18	.0697	2 2 3 5
7	0	7	.0734	3 3 3 3 3
7	0	7	.0734	4 6 6 3 3

7	0	7	.0734	4 6 3 3 3 3
7	0	7	.0734	4 3 3 3 3 3 3
18	9	9	.0740	4 2
6	4	2	.0800	4 3 3 2
16	2	14	.0837	3 4 3 4
20	3	17	.0886	3 4 3 1

6	0	6	.1065	2 3 3
6	0	6	.1065	3 6 6
6	0	6	.1065	1 2 3 3
6	0	6	.1065	1 2 3 3 3
6	0	6	.1065	3 4 3 1 3 2
96	36	60	.1098	3
19	3	16	.1118	3 3 5
9	5	4	.1135	3 4 5 2
10	1	9	.1323	3 2
14	2	12	.1397	2 3
7	4	3	.1414	4 6 2
29	6	23	.1545	4 3 3
5	0	5	.1547	3 2 3
5	0	5	.1547	4 4 4
5	0	5	.1547	4 4 6
5	0	5	.1547	3 4 5 6
5	0	5	.1547	4 5 3 3 3
5	0	5	.1547	4 6 6 6 3
5	0	5	.1547	1 2 3 3 3 3
5	0	5	.1547	3 3 3 3 3 3
5	0	5	.1547	4 6 6 3 3 3
5	0	5	.1547	4 6 3 3 3 3 3
5	0	5	.1547	4 3 3 3 3 3 3 3
9	1	8	.1764	3 7 5
9	1	8	.1764	3 4 5 4
9	1	8	.1764	3 4 3 4 2

4	0	4	.2247	2 6
4	0	4	.2247	2 2 4
4	0	4	.2247	2 3 6
4	0	4	.2247	4 4 3
4	0	4	.2247	4 6 7
4	0	4	.2247	4 7 5
4	0	4	.2247	2 2 4 3
4	0	4	.2247	3 2 3 3
4	0	4	.2247	4 5 3 0
4	0	4	.2247	3 4 6 3 0
4	0	4	.2247	1 2 3 3 3 3 3
3	2	1	.2306	3 4 2
3	2	1	.2306	1 2 4 2
3	2	1	.2306	2 2 3 2
3	2	1	.2306	3 4 5 5
3	2	1	.2306	1 2 4 3 1
3	2	1	.2306	3 4 3 3 4

8	1	7	.2333	4 7 4
43	16	27	.2407	4 5
6	3	3	.2772	4 6 6 5
11	2	9	.2841	3 4 6 3
<hr/>				
7	1	6	.3057	4 3 2
7	1	6	.3057	3 3 5 4
7	1	6	.3057	1 2 4 5 3
7	1	6	.3057	2 2 3 5 3
3	0	3	.3264	1 2 5
3	0	3	.3264	3 5 2
3	0	3	.3264	3 5 4
3	0	3	.3264	3 6 8
3	0	3	.3264	3 7 6
3	0	3	.3264	3 7 7
3	0	3	.3264	4 3 4
3	0	3	.3264	4 3 6
3	0	3	.3264	4 5 2
3	0	3	.3264	4 6 10
3	0	3	.3264	3 4 5 1
3	0	3	.3264	3 4 6 5
3	0	3	.3264	3 5 3 1
3	0	3	.3264	3 5 3 3
3	0	3	.3264	3 6 6 1
3	0	3	.3264	3 7 5 1
3	0	3	.3264	3 7 5 5
3	0	3	.3264	4 4 4 3
3	0	3	.3264	4 6 2 0
3	0	3	.3264	4 7 4 3
3	0	3	.3264	1 2 4 5 4
3	0	3	.3264	2 2 3 5 4
3	0	3	.3264	2 2 4 3 1
3	0	3	.3264	3 3 5 4 3
3	0	3	.3264	3 4 3 1 1
3	0	3	.3264	3 4 3 3 5
3	0	3	.3264	3 4 5 1 0
3	0	3	.3264	3 4 5 6 3
3	0	3	.3264	3 4 6 3 4
3	0	3	.3264	3 6 6 1 0
3	0	3	.3264	3 7 5 1 0
3	0	3	.3264	4 4 4 3 1
3	0	3	.3264	4 5 3 3 0
3	0	3	.3264	4 6 5 3 3
3	0	3	.3264	4 6 6 6 6
3	0	3	.3264	1 2 4 5 3 3
3	0	3	.3264	2 2 3 5 3 3
3	0	3	.3264	3 4 3 4 2 0
3	0	3	.3264	3 4 3 4 2 4
3	0	3	.3264	4 4 4 3 1 0
3	0	3	.3264	4 5 3 3 3 3
3	0	3	.3264	4 6 6 6 3 3

3	0	3	.3264	3 3 3 3 3 3 3
3	0	3	.3264	4 6 6 3 3 3 3
3	0	3	.3264	1 2 3 3 3 3 3 3
3	0	3	.3264	4 6 3 3 3 3 3 3
3	0	3	.3264	4 3 3 3 3 3 3 3 3
38	10	28	.3265	3 4 5
4	2	2	.3686	1 3 6
4	2	2	.3686	4 2 4
7	3	4	.3791	3 3 2
7	3	4	.3791	3 4 3 1 2
53	15	38	.3892	4 6
6	1	5	.3957	1 2 6
6	1	5	.3957	3 6 3
6	1	5	.3957	4 6 5 3
6	1	5	.3957	3 4 5 4 0

111	33	78	.4178	4
17	6	11	.4443	4 6 5
22	6	16	.4472	4 6 6
12	3	9	.4575	3 4 5 3
2	0	2	.4741	2 6 3
2	0	2	.4741	2 6 6
2	0	2	.4741	3 6 2
2	0	2	.4741	4 1 3
2	0	2	.4741	4 2 2
2	0	2	.4741	4 4 8
2	0	2	.4741	4 6 1
2	0	2	.4741	4 7 2
2	0	2	.4741	4 7 7
2	0	2	.4741	1 2 5 3
2	0	2	.4741	1 2 6 2
2	0	2	.4741	2 3 3 0
2	0	2	.4741	2 3 3 3
2	0	2	.4741	2 3 3 4
2	0	2	.4741	2 3 6 3
2	0	2	.4741	3 3 2 4
2	0	2	.4741	3 3 5 2
2	0	2	.4741	3 3 5 7
2	0	2	.4741	3 4 4 1
2	0	2	.4741	3 4 6 6
2	0	2	.4741	3 5 2 0
2	0	2	.4741	3 6 8 4
2	0	2	.4741	3 7 6 5
2	0	2	.4741	3 7 7 7
2	0	2	.4741	3 7 7 9
2	0	2	.4741	4 1 3 3
2	0	2	.4741	4 2 3 2
2	0	2	.4741	4 3 1 3
2	0	2	.4741	4 3 2 2
2	0	2	.4741	4 3 2 3
2	0	2	.4741	4 3 3 1

2	0	2	.4741	4 4 3 0
2	0	2	.4741	4 4 3 3
2	0	2	.4741	4 4 4 1
2	0	2	.4741	4 5 2 0
2	0	2	.4741	4 5 8 4
2	0	2	.4741	4 6 4 0
2	0	2	.4741	4 6 5 0
2	0	2	.4741	4 6 6 0
2	0	2	.4741	4 7 2 0
2	0	2	.4741	4 7 3 2
2	0	2	.4741	4 7 4 1
2	0	2	.4741	4 7 5 0
2	0	2	.4741	1 2 4 5 1
2	0	2	.4741	1 2 4 5 6
2	0	2	.4741	2 2 3 5 1
2	0	2	.4741	2 2 3 5 6
2	0	2	.4741	2 3 3 4 1
2	0	2	.4741	3 2 3 3 3
2	0	2	.4741	3 3 2 4 3
2	0	2	.4741	3 3 3 3 0
2	0	2	.4741	3 3 5 2 0
2	0	2	.4741	3 4 3 2 0
2	0	2	.4741	3 4 3 4 4
2	0	2	.4741	3 4 4 1 0
2	0	2	.4741	3 4 6 5 8
2	0	2	.4741	3 5 3 1 0
2	0	2	.4741	3 6 3 1 3
2	0	2	.4741	4 3 3 2 0
2	0	2	.4741	4 4 3 3 0
2	0	2	.4741	4 4 4 1 0
2	0	2	.4741	4 5 8 4 1
2	0	2	.4741	4 6 3 2 0
2	0	2	.4741	4 6 3 3 0
2	0	2	.4741	4 6 6 3 0
2	0	2	.4741	4 7 3 4 3
2	0	2	.4741	1 2 4 5 1 0
2	0	2	.4741	1 2 4 5 4 0
2	0	2	.4741	2 2 3 5 1 0
2	0	2	.4741	2 2 3 5 4 0
2	0	2	.4741	2 2 4 3 1 0
2	0	2	.4741	2 3 3 4 1 0
2	0	2	.4741	3 2 4 3 1 3
2	0	2	.4741	3 3 2 4 3 1
2	0	2	.4741	3 3 3 3 3 0
2	0	2	.4741	3 3 5 3 4 3
2	0	2	.4741	3 4 3 1 1 0
2	0	2	.4741	3 4 3 1 2 3
2	0	2	.4741	3 4 3 1 3 4
2	0	2	.4741	3 4 3 3 5 3
2	0	2	.4741	3 4 3 4 4 8
2	0	2	.4741	3 4 5 2 4 3

2	0	2	.4741	3	4	5	3	3	3
2	0	2	.4741	4	3	3	3	2	0
2	0	2	.4741	4	3	3	3	3	0
2	0	2	.4741	4	6	5	3	3	3
2	0	2	.4741	4	6	6	6	6	3
2	0	2	.4741	4	6	6	6	6	6
2	0	2	.4741	4	7	3	4	3	1
2	0	2	.4741	1	2	4	5	2	4
2	0	2	.4741	2	2	3	5	2	4
2	0	2	.4741	3	3	2	4	3	1
2	0	2	.4741	3	3	5	3	4	3
2	0	2	.4741	3	4	3	1	3	2
2	0	2	.4741	3	4	3	1	3	2
2	0	2	.4741	3	4	5	2	4	3
2	0	2	.4741	4	5	3	3	3	3
2	0	2	.4741	4	6	6	6	3	3
2	0	2	.4741	4	6	6	6	6	3
2	0	2	.4741	4	6	6	6	6	3
2	0	2	.4741	4	6	6	6	6	6
2	0	2	.4741	1	2	4	5	2	4
2	0	2	.4741	2	2	3	5	2	4
2	0	2	.4741	3	3	3	3	3	3
2	0	2	.4741	4	6	6	3	3	3
2	0	2	.4741	4	6	6	6	3	3
2	0	2	.4741	4	6	6	6	6	3
2	0	2	.4741	4	6	6	6	6	6
2	0	2	.4741	1	2	3	3	3	3
2	0	2	.4741	4	6	3	3	3	3
2	0	2	.4741	4	6	6	3	3	3
2	0	2	.4741	4	6	6	6	3	3
2	0	2	.4741	4	6	6	6	6	3
2	0	2	.4741	4	6	6	6	6	6
2	0	2	.4741	1	2	3	3	3	3
2	0	2	.4741	4	6	3	3	3	3
2	0	2	.4741	4	6	6	3	3	3
2	0	2	.4741	4	6	6	6	3	3
2	0	2	.4741	4	6	6	6	6	3
2	0	2	.4741	4	6	6	6	6	6
2	0	2	.4741	4	6	6	6	6	6
2	0	2	.4741	1	2	3	3	3	3
2	0	2	.4741	4	6	3	3	3	3
2	0	2	.4741	4	6	6	3	3	3
2	0	2	.4741	4	6	6	6	3	3
2	0	2	.4741	4	6	6	6	6	3
2	0	2	.4741	4	6	6	6	6	6
2	0	2	.4741	4	6	6	6	6	6
2	0	2	.4741	1	2	3	3	3	3
2	0	2	.4741	4	6	3	3	3	3
2	0	2	.4741	4	6	6	3	3	3
2	0	2	.4741	4	6	6	6	3	3
2	0	2	.4741	4	6	6	6	6	3
2	0	2	.4741	4	6	6	6	6	6
2	0	2	.4741	4	6	6	6	6	6
2	0	2	.4741	1	2	3	3	3	3
2	0	2	.4741	4	6	3	3	3	3
2	0	2	.4741	4	6	6	3	3	3
2	0	2	.4741	4	6	6	6	3	3
2	0	2	.4741	4	6	6	6	6	3
2	0	2	.4741	4	6	6	6	6	6
2	0	2	.4741	4	6	6	6	6	6
2	0	2	.4741	1	2	3	3	3	3
2									

Table 4

Acid data - random BiP<.50

Total = 122 Actives = 38 Inactives = 84

total	act	inact	BiP	SEV descriptors
4	4	0	.0094	4 5 3 0
13	8	5	.0230	4 5 3
3	3	0	.0302	3 5 4
3	3	0	.0302	4 6 2 0
4	3	1	.0926	4 6 7
4	3	1	.0926	3 4 3 2
9	5	4	.1135	3 3 3 3
22	4	18	.1378	4 6 6
7	4	3	.1414	4 6 2
5	0	5	.1547	3 2 4
5	0	5	.1547	4 3 1
5	0	5	.1547	3 2 4 3
5	0	5	.1547	3 4 5 6
5	0	5	.1547	3 6 3 1
5	0	5	.1547	4 6 5 4
5	0	5	.1547	3 2 4 3 1
5	0	5	.1547	4 5 4 3 3
18	8	10	.1668	4 2
9	1	8	.1764	3 7 5
9	1	8	.1764	3 4 5 4
9	1	8	.1764	3 4 3 4 2
5	3	2	.1786	1 3 2
5	3	2	.1786	4 2 3
5	3	2	.1786	1 3 2 1
5	3	2	.1786	1 3 2 1 0
24	5	19	.1940	3 4 3 3
38	9	29	.2089	3 4 5
4	0	4	.2247	2 6
4	0	4	.2247	4 9
4	0	4	.2247	1 3 6
4	0	4	.2247	4 7 5
4	0	4	.2247	3 4 3 3 3 3
4	0	4	.2247	1 2 4 3 3 3 3
4	0	4	.2247	2 2 3 3 3 3 3
12	2	10	.2263	4 6 3 3
12	2	10	.2263	4 3 3 3 3
3	2	1	.2306	4 7 4 3
3	2	1	.2306	3 3 5 4 3
3	2	1	.2306	3 4 3 1 1
3	2	1	.2306	4 3 3 3 2

3	2	1	.2306	1 2 4 5 3 3
3	2	1	.2306	2 2 3 5 3 3
3	2	1	.2306	3 4 3 3 3 0
8	1	7	.2333	4 5 4 3
11	5	6	.2362	3 3 3
43	16	27	.2407	4 5
26	6	20	.2544	3 6
26	6	20	.2544	4 4
6	3	3	.2772	1 2 6
6	3	3	.2772	3 4 0
6	3	3	.2772	4 2 1
6	3	3	.2772	1 2 3 3
6	3	3	.2772	1 2 4 0
6	3	3	.2772	2 2 3 0
6	3	3	.2772	4 2 1 0
6	3	3	.2772	4 3 3 2
6	3	3	.2772	1 2 3 3 3
11	2	9	.2841	3 4 6 3
11	2	9	.2841	4 6 6 6

7	1	6	.3057	4 6 6 3 3
7	1	6	.3057	4 6 3 3 3 3
7	1	6	.3057	4 3 3 3 3 3 3
15	6	9	.3129	2 2 3 3
18	7	11	.3160	3 5
18	7	11	.3160	1 2 4 3
3	0	3	.3264	4 1
3	0	3	.3264	3 4 2
3	0	3	.3264	3 6 8
3	0	3	.3264	3 7 6
3	0	3	.3264	4 3 6
3	0	3	.3264	4 5 2
3	0	3	.3264	1 2 4 2
3	0	3	.3264	2 2 3 2
3	0	3	.3264	3 7 5 5
3	0	3	.3264	4 3 1 0
3	0	3	.3264	1 2 4 5 4
3	0	3	.3264	2 2 3 5 4
3	0	3	.3264	3 4 3 3 5
3	0	3	.3264	3 4 5 3 1
3	0	3	.3264	3 4 5 6 3
3	0	3	.3264	3 4 6 3 4
3	0	3	.3264	4 6 5 4 3
3	0	3	.3264	4 6 6 5 3
3	0	3	.3264	3 4 3 4 2 4
3	0	3	.3264	4 5 4 3 3 3
3	0	3	.3264	3 4 3 3 3 3 3
3	0	3	.3264	1 2 4 3 3 3 3 3
3	0	3	.3264	2 2 3 3 3 3 3 3
10	2	8	.3528	3 2
10	2	8	.3528	4 6 6 3

96	32	64	.3579	3
4	2	2	.3686	2 2 4
4	2	2	.3686	4 5 5
4	2	2	.3686	2 2 4 3
4	2	2	.3686	4 6 3 2
7	3	4	.3791	3 3 2
7	3	4	.3791	3 4 4
7	3	4	.3791	3 3 5 4
7	3	4	.3791	1 2 4 5 3
7	3	4	.3791	2 2 3 5 3
7	3	4	.3791	3 3 3 3 3
41	14	27	.3953	3 4 3
6	1	5	.3957	3 6 3
6	1	5	.3957	3 6 6
6	1	5	.3957	4 6 6 5
6	1	5	.3957	3 4 5 4 0
6	1	5	.3957	1 2 4 3 3 3
6	1	5	.3957	2 2 3 3 3 3
6	1	5	.3957	3 4 3 1 3 2

16	4	12	.4104	4 5 4
29	10	19	.4163	4 3 3
19	5	14	.4304	3 4 6
9	2	7	.4326	3 4 3 3 3
9	2	7	.4326	4 6 3 3 3
9	2	7	.4326	4 3 3 3 3 3
20	7	13	.4362	3 4 3 1
74	22	52	.4514	3 4
48	14	34	.4519	4 3
14	5	9	.4534	2 3
12	3	9	.4575	3 4 5 3
12	3	9	.4575	1 2 4 3 3
12	3	9	.4575	2 2 3 3 3
54	16	38	.4700	2 2
2	0	2	.4741	1 2 2
2	0	2	.4741	2 6 3
2	0	2	.4741	2 6 6
2	0	2	.4741	3 6 7
2	0	2	.4741	4 1 3
2	0	2	.4741	4 4 1
2	0	2	.4741	4 4 8
2	0	2	.4741	4 6 0
2	0	2	.4741	4 7 7
2	0	2	.4741	1 2 2 4
2	0	2	.4741	1 2 3 2
2	0	2	.4741	1 2 5 3
2	0	2	.4741	2 3 3 0
2	0	2	.4741	3 3 2 1
2	0	2	.4741	3 3 5 7
2	0	2	.4741	3 4 2 0
2	0	2	.4741	3 4 4 6

2	0	2	.4741	3 6 8 4
2	0	2	.4741	3 7 6 5
2	0	2	.4741	3 7 7 7
2	0	2	.4741	4 1 3 3
2	0	2	.4741	4 3 1 3
2	0	2	.4741	4 3 2 2
2	0	2	.4741	4 3 3 6
2	0	2	.4741	4 4 1 0
2	0	2	.4741	4 4 3 3
2	0	2	.4741	4 5 2 0
2	0	2	.4741	4 5 8 4
2	0	2	.4741	4 6 4 0
2	0	2	.4741	4 6 6 0
2	0	2	.4741	4 6 6 4
2	0	2	.4741	4 7 3 2
2	0	2	.4741	4 7 4 1
2	0	2	.4741	4 7 5 0
2	0	2	.4741	1 2 2 4 3
2	0	2	.4741	1 2 4 2 0
2	0	2	.4741	1 2 4 5 6
2	0	2	.4741	2 2 3 2 0
2	0	2	.4741	2 2 3 5 6
2	0	2	.4741	3 3 2 1 0
2	0	2	.4741	3 4 3 3 2
2	0	2	.4741	3 6 3 1 2
2	0	2	.4741	3 6 3 1 3
2	0	2	.4741	4 4 3 3 0
2	0	2	.4741	4 5 4 2 1
2	0	2	.4741	4 5 8 4 1
2	0	2	.4741	4 6 3 3 0
2	0	2	.4741	4 6 6 3 0
2	0	2	.4741	4 6 6 5 4
2	0	2	.4741	4 6 6 6 0
2	0	2	.4741	4 6 6 6 5
2	0	2	.4741	4 7 3 4 3
2	0	2	.4741	1 2 2 4 3 1
2	0	2	.4741	1 2 4 5 4 0
2	0	2	.4741	2 2 3 5 4 0
2	0	2	.4741	3 2 4 3 1 2
2	0	2	.4741	3 2 4 3 1 3
2	0	2	.4741	3 3 5 3 4 3
2	0	2	.4741	3 4 3 3 5 3
2	0	2	.4741	3 4 5 2 4 3
2	0	2	.4741	3 4 5 3 1 0
2	0	2	.4741	4 3 3 3 3 0
2	0	2	.4741	4 5 4 2 1 0
2	0	2	.4741	4 6 5 3 3 3
2	0	2	.4741	4 6 5 4 3 3
2	0	2	.4741	4 7 3 4 3 1
2	0	2	.4741	1 2 2 4 3 1 2
2	0	2	.4741	1 2 4 5 2 4 3

2	0	2	.4741	2 2 3 5 2 4 3
2	0	2	.4741	3 3 5 3 4 3 1
2	0	2	.4741	3 4 3 1 3 2 4
2	0	2	.4741	3 4 5 2 4 3 1
2	0	2	.4741	4 5 3 3 3 3 3
2	0	2	.4741	4 5 4 3 3 3 3
2	0	2	.4741	1 2 4 5 2 4 3 1
2	0	2	.4741	2 2 3 5 2 4 3 1
2	0	2	.4741	3 3 3 3 3 3 3 3
2	0	2	.4741	3 4 3 3 3 3 3 3
2	0	2	.4741	1 2 3 3 3 3 3 3 3
2	0	2	.4741	1 2 4 3 3 3 3 3 3
2	0	2	.4741	2 2 3 3 3 3 3 3 3
8	3	5	.4773	4 7 3
8	3	5	.4773	4 7 4
8	3	5	.4773	4 5 3 3
43	14	29	.4777	3 3
44	13	31	.4817	2 2 3
76	23	53	.4893	1.
47	14	33	.4905	1 2 4
69	22	47	.4926	1 2
53	17	36	.4935	4 6
5	2	3	.4953	3 2 3
5	2	3	.4953	4 4 4
5	2	3	.4953	3 4 5 3 3
5	2	3	.4953	4 5 3 3 3
5	2	3	.4953	1 2 3 3 3 3
5	2	3	.4953	3 3 3 3 3 3

Table 5a

Acid data - real BiP<.10

Total = 122

Actives = 38

Inactives = 84

total	act	inact	BiP	SEV descriptors
54	33	21	.0000	2 2
18	17	1	.0000	1 2 4 3
15	15	0	.0000	2 2 3 3
12	12	0	.0000	1 2 4 3 3
12	12	0	.0000	2 2 3 3 3
8	8	0	.0001	2 2 2
8	8	0	.0001	3 3 4
8	8	0	.0001	1 2 3 4
69	36	33	.0002	1 2
47	27	20	.0002	1 2 4
65	34	31	.0003	2
7	7	0	.0003	3 6 4
7	7	0	.0003	2 2 2 4
44	25	19	.0004	2 2 3
16	12	4	.0004	4 5 4
24	16	8	.0004	3 4 3 3
76	38	38	.0005	1
6	6	0	.0009	3 4 0
6	6	0	.0009	4 2 1
6	6	0	.0009	1 2 4 0
6	6	0	.0009	2 2 3 0
6	6	0	.0009	3 3 4 3
6	6	0	.0009	3 6 4 1
6	6	0	.0009	4 2 1 0
6	6	0	.0009	1 2 3 4 3
6	6	0	.0009	2 2 2 4 3
6	6	0	.0009	3 3 4 3 1
6	6	0	.0009	1 2 3 4 3 1
6	6	0	.0009	1 2 4 3 3 3
6	6	0	.0009	2 2 2 4 3 1
6	6	0	.0009	2 2 3 3 3 3
5	5	0	.0029	1 3 2
5	5	0	.0029	1 3 2 1
5	5	0	.0029	1 3 2 1 0
17	11	6	.0045	1 2 3
13	9	4	.0053	1 3
19	1	18	.0080	4 6 3
4	4	0	.0094	4 5 4 0
4	4	0	.0094	4 6 2 1
4	4	0	.0094	4 3 3 2 1
4	4	0	.0094	4 6 2 1 0
4	4	0	.0094	1 2 4 3 3 0

4	4	0	.0094	2 2 3 3 3 0
4	4	0	.0094	3 4 3 3 3 3
4	4	0	.0094	4 3 3 2 1 0
4	4	0	.0094	1 2 4 3 3 3 3
4	4	0	.0094	2 2 3 3 3 3 3
18	1	17	.0111	3 5
18	1	17	.0111	4 3 3 3
12	0	12	.0114	4 6 3 3
12	0	12	.0114	4 3 3 3 3
8	6	2	.0138	4 5 4 3
11	0	11	.0165	3 3 3
11	0	11	.0165	3 4 3 1 3
74	32	42	.0189	3 4
26	3	23	.0195	4 4
10	0	10	.0239	4 6 6 3
41	19	22	.0294	3 4 3
3	3	0	.0302	3 4 3 3 1
3	3	0	.0302	3 4 5 2 0
3	3	0	.0302	3 6 4 1 0
3	3	0	.0302	4 6 5 4 3
3	3	0	.0302	3 3 4 3 1 0
3	3	0	.0302	3 4 3 1 2 1
3	3	0	.0302	3 4 3 3 1 0
3	3	0	.0302	4 5 4 3 3 3
3	3	0	.0302	1 2 3 4 3 1 0
3	3	0	.0302	2 2 2 4 3 1 0
3	3	0	.0302	3 4 3 1 2 1 0
3	3	0	.0302	3 4 3 3 3 3 3
3	3	0	.0302	1 2 4 3 3 3 3 3
3	3	0	.0302	2 2 3 3 3 3 3 3
9	6	3	.0305	3 4 3 3 3
9	0	9	.0348	3 3 3 3
9	0	9	.0348	4 6 3 3 3
9	0	9	.0348	4 3 3 3 3 3
5	4	1	.0353	4 6 5 4
5	4	1	.0353	3 4 3 3 0
5	4	1	.0353	4 5 4 3 3
19	2	17	.0371	3 4 6
48	9	39	.0403	4 3

8	0	8	.0505	3 5 3
8	0	8	.0505	4 5 3 3
13	1	12	.0538	4 5 3
22	3	19	.0546	4 7
21	3	18	.0697	3 7
21	3	18	.0697	1 2 4 5
21	3	18	.0697	2 2 3 5
7	0	7	.0734	3 3 3 3 3
7	0	7	.0734	4 6 6 3 3
7	0	7	.0734	4 6 3 3 3 3
7	0	7	.0734	4 3 3 3 3 3 3

18	9	9	.0740	4 2
6	4	2	.0800	4 3 3 2
16	2	14	.0837	3 4 3 4
20	3	17	.0886	3 4 3 1

Table 5b

Randomized acid data BiP<.10

Acid data - random 1 BiP<.10

Total = 122 Actives = 38 Inactives = 84

total	act	inact	BiP	SEV descriptors
4	4	0	.0094	4 5 3 0
13	8	5	.0230	4 5 3
3	3	0	.0302	3 5 4
3	3	0	.0302	4 6 2 0
4	3	1	.0926	4 6 7
4	3	1	.0926	3 4 3 2

Acid data - random 2

Total = 122 Actives = 38 Inactives = 84

total	act	inact	BiP	SEV descriptors
6	5	1	.0130	1 2 3 3
6	5	1	.0130	1 2 3 3 3
17	10	7	.0167	1 2 3
43	20	23	.0248	3 3
11	7	4	.0267	3 3 3
3	3	0	.0302	4 6 6 5 3
9	6	3	.0305	3 3 3 3
7	5	2	.0339	3 3 3 3 3
5	4	1	.0353	1 2 3 3 3 3
8	5	3	.0674	4 5 3 3
13	7	6	.0750	4 5 3
16	8	8	.0903	3 4 3 4
4	3	1	.0926	4 4 0
4	3	1	.0926	1 2 3 3 3 3 3

Acid data - random 3

Total = 122 Actives = 38 Inactives = 84

total	act	inact	Bip	SEV descriptors
3	3	0	.0302	4 5 2
5	4	1	.0353	4 2 3
47	9	38	.0484	1 2 4
65	14	51	.0585	2
8	5	3	.0674	3 5 3
21	3	18	.0697	1 2 4 5
21	3	18	.0697	2 2 3 5
7	0	7	.0734	3 4 3 1 2
6	4	2	.0800	1 2 6
44	9	35	.0820	2 2 3
20	3	17	.0886	3 4 3 1
4	3	1	.0926	1 3 6
11	1	10	.0985	3 4 3 1 3

acid data - random 4

Total = 122 Actives = 38 Inactives = 84

total	act	inact	Bip	SEV descriptors
11	0	11	.0165	4 6 6 6
10	0	10	.0239	4 6 6 3
3	3	0	.0302	4 5 4 2
3	3	0	.0302	4 6 2 0
9	0	9	.0348	4 6 3 3 3
9	0	9	.0348	4 3 3 3 3 3
14	1	13	.0395	2 3
17	9	8	.0506	1 2 3
22	3	19	.0546	4 6 6
8	5	3	.0674	2 2 2
8	5	3	.0674	3 3 4
8	5	3	.0674	1 2 3 4
7	0	7	.0734	3 3 2
7	0	7	.0734	4 6 6 3 3
7	0	7	.0734	4 6 3 3 3 3
7	0	7	.0734	4 3 3 3 3 3 3
13	7	6	.0750	4 5 3
4	3	1	.0926	4 6 4
4	3	1	.0926	4 5 3 0
4	3	1	.0926	4 6 3 2

Total = 122 Actives = 38 Inactives = 84

total	act	inact	Bip	SEV descriptors
12	0	12	.0114	3 4 5 3
38	6	32	.0258	3 4 5
3	3	0	.0302	3 4 3 1 1
19	2	17	.0371	3 3 5
8	0	8	.0505	4 7 4
21	3	18	.0697	1 2 4 5
21	3	18	.0697	2 2 3 5
7	0	7	.0734	3 3 5 4
7	0	7	.0734	1 2 4 5 3
7	0	7	.0734	2 2 3 5 3
6	4	2	.0800	1 2 6
4	3	1	.0926	3 4 3 2
43	9	34	.0970	3 3

random 6 - submitted as a test

Total = 122 Actives = 38 Inactives = 84

total	act	inact	Bip	SEV descriptors
15	10	5	.0050	2 2 3 3
12	8	4	.0121	1 2 4 3 3
12	8	4	.0121	2 2 3 3 3
6	5	1	.0130	1 2 4 3 3 3
6	5	1	.0130	2 2 3 3 3 3
18	10	8	.0273	1 2 4 3
3	3	0	.0302	4 5 2
9	6	3	.0305	3 4 3 3 3
18	2	16	.0489	4 2
26	4	22	.0578	3 6
8	5	3	.0674	4 5 4 3
7	0	7	.0734	3 3 5 4
7	0	7	.0734	1 2 4 5 3
7	0	7	.0734	2 2 3 5 3
7	0	7	.0734	3 4 3 1 2
6	4	2	.0800	4 6 6 5
6	4	2	.0800	3 4 5 4 0
48	20	28	.0804	4 3
20	3	17	.0886	3 4 3 1
16	8	8	.0903	4 5 4
4	3	1	.0926	4 6 4
4	3	1	.0926	3 4 3 3 3 3
4	3	1	.0926	1 2 4 3 3 3 3
4	3	1	.0926	2 2 3 3 3 3 3

Table 6a

Nitroso data - real Bip<.10

Total = 64		Actives = 45		Inactives = 19	
total	act	inact	Bip	SEV descriptors	
5	0	5	.0023	1 2 4	
4	0	4	.0078	2 2 3	
8	2	6	.0107	1 2	
9	3	6	.0240	3 4	
3	0	3	.0262	2 5	
3	0	3	.0262	4 1 2	
3	0	3	.0262	1 2 4 5	
3	0	3	.0262	2 2 3 5	
3	0	3	.0262	4 6 6 4 0	
4	1	3	.0814	4 1	
4	1	3	.0814	3 4 5	
4	1	3	.0814	4 8 6	
4	1	3	.0814	4 6 6 4	
7	7	0	.0850	4 2 4	
14	7	7	.0888	4 8	

Table 6b

Randomized Nitroso data BiP<.10

Nitroso data - random 1

Total = 64		Actives = 45		Inactives = 19	
total	act	inact	BiP	SEV descriptors	
4	1	3	.0814	1	3 6
4	1	3	.0814	4	6 6 4

Nitroso data - random 2

Total = 64		Actives = 45		Inactives = 19	
total	act	inact	BiP	SEV descriptors	
18	9	9	.0560	4	5 6
21	11	10	.0634	2	2 6 5
21	11	10	.0634	1	1 2 6 5
6	2	4	.0680	4	6 7
4	1	3	.0814	4	9
4	1	3	.0814	4	5 8
22	12	10	.0862	3	7 5

Nitroso data - random 3

Total = 64		Actives = 45		Inactives = 19	
total	act	inact	BiP	SEV descriptors	
3	0	3	.0262	3	711
3	0	3	.0262	2	2 611
3	0	3	.0262	1	1 2 611
6	2	4	.0680	4	6 5 7
4	1	3	.0814	4	1

Nitroso data - random 4

Total = 64		Actives = 45		Inactives = 19	
total	act	inact	Bip	SEV descriptors	
11	5	6	.0747	4 2	

Nitroso - random 5

Total = 64		Actives = 45		Inactives = 19	
total	act	inact	Bip	SEV descriptors	
3	0	3	.0262	3 4 5 5	
21	19	2	.0290	2 2 6 5	
21	19	2	.0290	1 1 2 6 5	
22	19	3	.0722	3 7 5	
4	1	3	.0814	3 4 5	

Nitroso data - random 6

Total = 64		Actives = 45		Inactives = 19	
total	act	inact	Bip	SEV descriptors	
14	14	0	.0072	4 8	
10	10	0	.0295	4 3 5	
13	12	1	.0666	4 6 6	
4	1	3	.0814	4 9	
9	4	5	.0950	3 7 5 2	
9	4	5	.0950	2 2 6 5 2	
9	4	5	.0950	1 1 2 6 5 2	

Table 7a

Halluc. data - real Bip<.10

Total = 40		Actives = 27		Inactives = 13			
total	act	inact	Bip	SEV descriptors			
17	17	0	.0013	3	5	9	
16	16	0	.0019	2	5	4	6
16	16	0	.0019	4	1	2	4 6
14	4	10	.0033	3	7	6	
4	0	4	.0112	3	7	6	8
6	1	5	.0159	4	5	9	3
6	1	5	.0159	4	8	2	4 3
6	1	5	.0159	3	3	6	2 4 3
6	1	5	.0159	4	3	5	2 4 3
25	22	3	.0183	3	7	7	
19	17	2	.0284	4	5	9	4
19	17	2	.0284	4	8	2	4 4
19	17	2	.0284	3	3	6	2 4 4
19	17	2	.0284	4	3	5	2 4 4
11	4	7	.0337	3	4	6	
3	0	3	.0343	3	4	7	6
3	0	3	.0343	3	5	7	6 3
3	0	3	.0343	4	5	9	3 3
3	0	3	.0343	4	8	2	4 3 3
3	0	3	.0343	3	3	6	2 4 3 3
3	0	3	.0343	4	3	5	2 4 3 3
5	1	4	.0413	3	4	6	5
13	12	1	.0438	3	7	7	10
13	12	1	.0438	4	5	9	4 5
13	12	1	.0438	4	8	2	4 4 5
13	12	1	.0438	3	3	6	2 4 4 5
13	12	1	.0438	4	3	5	2 4 4 5
21	10	11	.0468	3	5	7	
21	10	11	.0468	2	5	4	4
21	10	11	.0468	4	1	2	4 4
7	7	0	.0638	3	4	7	11
7	7	0	.0638	3	5	9	7
7	7	0	.0638	2	5	4	6 7
7	7	0	.0638	4	1	2	4 6 7
8	3	5	.0797	2	5	4	5
8	3	5	.0797	4	1	2	4 5
6	6	0	.0946	3	4	5	
6	6	0	.0946	3	5	9	4
6	6	0	.0946	2	5	4	6 4
6	6	0	.0946	4	1	2	4 6 4

Table 7b

Randomized Halluc. data BiP<.10

Halluc. data - random 1

Total = 40		Actives = 27		Inactives = 13	
total	act	inact	BiP	SEV descriptors	
3	0	3	.0343	3	5 5 2
3	0	3	.0343	3	5 10 6
3	0	3	.0343	2	5 3 3 2
3	0	3	.0343	2	5 4 7 6
3	0	3	.0343	3	5 5 2 6
3	0	3	.0343	2	5 3 3 2 6
3	0	3	.0343	3	5 5 2 6 3
3	0	3	.0343	4	1 2 4 7 6
3	0	3	.0343	2	5 3 3 2 6 3
13	6	7	.0919	3	7 7 10
13	6	7	.0919	4	5 9 4 5
13	6	7	.0919	4	8 2 4 4 5
13	6	7	.0919	3	3 6 2 4 4 5
13	6	7	.0919	4	3 5 2 4 4 5
6	6	0	.0946	3	5 6 4
6	6	0	.0946	3	4 4 8 6

Halluc. data - random 2

Total = 40		Actives = 27		Inactives = 13	
<hr/>					
total	act	inact	Bip	SEV descriptors	
<hr/>					
no descriptors with Bip < .10					

Halluc. data - random 3

Total = 40		Actives = 27		Inactives = 13	
total	act	inact	Bip	SEV descriptors	
19	16	3	.0902	4	5 9 4
19	16	3	.0902	4	8 2 4 4
19	16	3	.0902	3	3 6 2 4 4
19	16	3	.0902	4	3 5 2 4 4
6	2	4	.0921	4	5 9 3
6	2	4	.0921	3	4 7 8 6
6	2	4	.0921	4	8 2 4 3
6	2	4	.0921	3	3 6 2 4 3
6	2	4	.0921	4	3 5 2 4 3
18	9	9	.0938	3	4 7 8
6	6	0	.0946	3	4 5

Halluc. data - random 4

Total = 40		Actives = 27		Inactives = 13	
total	act	inact	Bip	SEV descriptors	
3	0	3	.0343	3	510 6
3	0	3	.0343	2	5 4 7 6
3	0	3	.0343	4	1 2 4 7 6
5	1	4	.0413	4	6
5	1	4	.0413	4	3 3
8	8	0	.0431	2	5 4 3
8	8	0	.0431	4	1 2 4 3
6	2	4	.0921	3	5 7 7
6	2	4	.0921	2	5 4 4 7
6	2	4	.0921	4	1 2 4 4 7
6	6	0	.0946	3	5 9 4
6	6	0	.0946	2	5 4 6 4
6	6	0	.0946	4	1 2 4 6 4

Halluc. data - random 5

Total = 40		Actives = 27		Inactives = 13	
total	act	inact	Bip	SEV descriptors	
3	0	3	.0343	3 5 9 5 9	
3	0	3	.0343	2 5 4 6 5 9	
3	0	3	.0343	3 5 9 5 9 3	
3	0	3	.0343	2 5 4 6 5 9 3	
3	0	3	.0343	4 1 2 4 6 5 9	
3	0	3	.0343	4 1 2 4 6 5 9 3	
5	1	4	.0413	3 5 9 5	
5	1	4	.0413	2 5 4 6 5	
5	1	4	.0413	4 1 2 4 6 5	
7	7	0	.0638	3 5 5	
7	7	0	.0638	2 5 3 3	
8	3	5	.0797	2 5 4 3	
8	3	5	.0797	4 1 2 4 3	
11	10	1	.0835	4 2	
6	2	4	.0921	3 4 5	

Halluc. data - random 6

Total = 40		Actives = 27		Inactives = 13	
<hr/>					
total	act	inact	Bip	SEV descriptors	
<hr/>					
3	0	3	.0343	3 4 7 7	

Table 8a

Thermo. data - real Bip < .10

Total = 73			Actives = 32	Inactives = 41
total	act	inact	Bip	SEV descriptors
11	11	0	.0001	2 2 2
11	11	0	.0001	2 2 2 3
9	9	0	.0006	3 3 3 5
9	9	0	.0006	2 2 2 3 5
12	11	1	.0008	3 3 3
7	7	0	.0031	4 4 3 5
7	7	0	.0031	1 3 4 3 5
7	7	0	.0031	2 3 3 3 5
10	9	1	.0036	2 3 3
10	9	1	.0036	1 2 5 6
12	10	2	.0063	2 3
9	8	1	.0075	3 6 7
8	0	8	.0099	3 3 2
8	0	8	.0099	3 3 2 2
12	1	11	.0102	3 4 2
12	1	11	.0102	1 2 4 2
12	1	11	.0102	2 2 3 2
12	1	11	.0102	3 4 2 2
12	1	11	.0102	4 4 3 3
12	1	11	.0102	1 2 4 2 2
12	1	11	.0102	2 2 3 2 2
18	13	5	.0142	3 5 2 2
18	13	5	.0142	4 4 4 3
18	13	5	.0142	1 2 5 2 2
18	13	5	.0142	3 2 3 2 2
18	13	5	.0142	3 5 2 2 3
18	13	5	.0142	4 4 4 3 5
18	13	5	.0142	1 2 5 2 2 3
18	13	5	.0142	3 2 3 2 2 3
18	13	5	.0142	3 5 2 2 3 5
18	13	5	.0142	1 2 5 2 2 3 5
18	13	5	.0142	3 2 3 2 2 3 5
8	7	1	.0153	1 3 4 3
8	7	1	.0153	2 3 3 3
8	7	1	.0153	3 5 8 8
5	5	0	.0162	3 5 8 9
5	5	0	.0162	1 2 5 6 5
5	5	0	.0162	3 3 3 5 5
5	5	0	.0162	3 5 8 7 7
5	5	0	.0162	2 2 2 3 5 5

13	10	3	.0165	4 4 4 3 5 2 3
13	10	3	.0165	3 5 2 2 3 5 2 3
13	10	3	.0165	1 2 5 2 2 3 5 2 3
13	10	3	.0165	3 2 3 2 2 3 5 2 3
11	1	10	.0168	3 4 2 2 3
11	1	10	.0168	1 2 4 2 2 3
11	1	10	.0168	2 2 3 2 2 3
15	11	4	.0205	4 7 4 6
15	11	4	.0205	4 3 4 4 6
10	8	2	.0230	1 3 4
19	13	6	.0271	3 5 2
19	13	6	.0271	1 2 5 2
19	13	6	.0271	3 2 3 2
12	9	3	.0296	4 7 4 6 7
12	9	3	.0296	4 3 4 4 6 7
21	14	7	.0297	4 3 4
6	0	6	.0314	3 5 5 5
6	0	6	.0314	1 2 5 5 5
6	0	6	.0314	3 3 2 2 3
14	10	4	.0351	3 5 8 4 5
14	10	4	.0351	1 2 5 8 4 5
14	10	4	.0351	3 5 5 7 2 3
14	10	4	.0351	4 4 4 3 5 2
14	10	4	.0351	3 5 2 2 3 5 2
14	10	4	.0351	1 2 5 2 2 3 5 2
14	10	4	.0351	3 2 3 2 2 3 5 2
4	4	0	.0369	1 2 5 8 9
33	9	24	.0386	2 2 3
9	1	8	.0446	4 4 2
9	1	8	.0446	3 5 8 5

24	15	9	.0513	3 5 8 7
11	8	3	.0520	3 5 8 4 5 5
11	8	3	.0520	1 2 5 8 4 5 5
11	8	3	.0520	3 5 5 7 2 3 5
11	8	3	.0520	4 4 4 3 5 2 3 5
11	8	3	.0520	3 5 2 2 3 5 2 3 5
11	8	3	.0520	1 2 5 2 2 3 5 2 3 5
11	8	3	.0520	3 2 3 2 2 3 5 2 3 5
5	0	5	.0559	3 3 2 2 3 5
6	5	1	.0616	3 5 6 5
15	10	5	.0644	3 5 5 7 2
39	12	27	.0674	1 2 4
17	11	6	.0686	3 5 8 4
17	11	6	.0686	4 3 4 4
17	11	6	.0686	1 2 5 8 4
8	1	7	.0717	4 4 3 3 5
8	1	7	.0717	3 4 2 2 3 5
8	1	7	.0717	1 2 4 2 2 3 5

8	1	7	.0717	2 2 3 2 2 3 5
8	6	2	.0780	3 4 6 4 4 4
8	6	2	.0780	3 4 3 4 3 4 4
8	6	2	.0780	3 4 3 1 3 3 4 4
3	3	0	.0842	1 2 5 8 8
3	3	0	.0842	3 5 6 5 9
3	3	0	.0842	3 5 8 8 6
3	3	0	.0842	3 5 8 8 10
3	3	0	.0842	3 5 8 9 9
3	3	0	.0842	4 4 4 4 5
3	3	0	.0842	4 7 5 7 8
3	3	0	.0842	1 2 5 6 5 9
3	3	0	.0842	3 3 3 5 5 9
3	3	0	.0842	3 4 5 4 7 8
3	3	0	.0842	3 5 8 7 7 7
3	3	0	.0842	4 5 8 6 4 5
3	3	0	.0842	4 7 4 6 7 6
3	3	0	.0842	1 2 4 5 4 7 8
3	3	0	.0842	2 2 2 3 5 5 9
3	3	0	.0842	2 2 3 5 4 7 8
3	3	0	.0842	3 5 8 4 5 5 4
3	3	0	.0842	3 7 6 5 3 3 5
3	3	0	.0842	4 3 4 4 6 7 6
3	3	0	.0842	1 2 5 8 4 5 5 4
3	3	0	.0842	3 4 6 4 4 3 3 5
3	3	0	.0842	3 5 5 7 2 3 5 4
3	3	0	.0842	3 4 3 4 3 4 3 3 5
3	3	0	.0842	4 4 4 3 5 2 3 5 4
3	3	0	.0842	3 4 3 1 3 3 4 3 3 5
3	3	0	.0842	3 5 2 2 3 5 2 3 5 4
3	3	0	.0842	1 2 5 2 2 3 5 2 3 5 4
3	3	0	.0842	3 2 3 2 2 3 5 2 3 5 4
27	8	19	.0964	1 2 5 5
12	8	4	.0967	1 3
4	0	4	.0995	4 2 2
4	0	4	.0995	3 5 8 5 6
4	0	4	.0995	3 5 8 6 6
4	0	4	.0995	4 7 5 7 5
4	0	4	.0995	3 4 5 4 7 5
4	0	4	.0995	1 2 4 5 4 7 5
4	0	4	.0995	2 2 3 5 4 7 5

Table 8b

Randomized Thermo. data BiP < .10

Thermo. data - random 1

Total = 73		Actives = 32		Inactives = 41	
total	act	inact	BiP	SEV descriptors	
6	0	6	.0314	4 7 8	
6	0	6	.0314	4 610 5	
6	0	6	.0314	4 9 3 4 5	
6	0	6	.0314	4 3 6 3 4 5	
5	0	5	.0559	1 2 5 8 7	
6	5	1	.0616	3 3 5 5	
6	5	1	.0616	4 610 2	
6	5	1	.0616	4 9 3 4 2	
6	5	1	.0616	4 3 6 3 4 2	
3	3	0	.0842	4 3 4 3	
3	3	0	.0842	4 7 3 3	
3	3	0	.0842	1 2 5 8 8	
3	3	0	.0842	3 4 5 2 3	
3	3	0	.0842	4 3 4 3 3	
3	3	0	.0842	4 4 4 4 5	
3	3	0	.0842	4 610 4 3	
3	3	0	.0842	4 7 7 9 2	
3	3	0	.0842	1 2 4 5 2 3	
3	3	0	.0842	2 2 3 5 2 3	
3	3	0	.0842	4 610 4 3 2	
3	3	0	.0842	4 9 3 4 4 3	
3	3	0	.0842	4 3 6 3 4 4 3	
3	3	0	.0842	4 9 3 4 4 3 2	
3	3	0	.0842	4 3 6 3 4 4 3 2	
30	9	21	.0881	4 5 8	
35	11	24	.0940	4 5	
27	8	19	.0964	3 7 6 5	
4	0	4	.0995	3 5 811	
4	0	4	.0995	1 2 5 811	

Thermo. data - random 2

Total = 73 Actives = 32 Inactives = 41				
total	act	inact	Bip	SEV descriptors
7	7	0	.0031	3 5 8 7 8
5	5	0	.0162	3 5 810
5	5	0	.0162	1 2 5 810
5	5	0	.0162	4 7 4 6 7 7
5	5	0	.0162	3 5 8 4 5 5 5
5	5	0	.0162	4 3 4 4 6 7 7
5	5	0	.0162	1 2 5 8 4 5 5 5
5	5	0	.0162	3 5 5 7 2 3 5 5
5	5	0	.0162	4 4 4 3 5 2 3 5 5
5	5	0	.0162	3 5 2 2 3 5 2 3 5 5
5	5	0	.0162	1 2 5 2 2 3 5 2 3 5 5
5	5	0	.0162	3 2 3 2 2 3 5 2 3 5 5
4	4	0	.0369	3 5 6 8
4	4	0	.0369	3 4 6 4 4 5
4	4	0	.0369	3 4 3 4 3 4 5
4	4	0	.0369	3 4 3 1 3 3 4 5
6	5	1	.0616	4 7 710 5
6	5	1	.0616	4 610 4 4 5
6	5	1	.0616	4 9 3 4 4 4 5
6	5	1	.0616	4 3 6 3 4 4 4 5
3	3	0	.0842	3 3 6 7 9
3	3	0	.0842	1 2 3 6 7 9
3	3	0	.0842	2 5 4 7 8 6
3	3	0	.0842	4 3 6 5 4 5 5
3	3	0	.0842	3 7 4 3 2 3 5 5
3	3	0	.0842	3 4 6 2 2 2 3 5 5
3	3	0	.0842	3 4 3 4 1 2 2 3 5 5
3	3	0	.0842	3 4 3 1 3 1 2 2 3 5 5
12	8	4	.0967	4 7 4 6 7
12	8	4	.0967	4 3 4 4 6 7
4	0	4	.0995	3 4 5 6
4	0	4	.0995	3 4 5 6 3

Thermo. data - random 3

Total = 73 Actives = 32 Inactives = 41				
total	act	inact	Bip	SEV descriptors
9	1	8	.0446	3 3 6 7
9	1	8	.0446	1 2 3 6 7
9	1	8	.0446	2 5 4 7 8

9	1	8	.0446	4 3 6 5 4 5
9	1	8	.0446	3 7 4 3 2 3 5
9	1	8	.0446	3 4 6 2 2 2 3 5
9	1	8	.0446	3 4 3 4 1 2 2 3 5
9	1	8	.0446	3 4 3 1 3 1 2 2 3 5

18	4	14	.0505	4 7 6
5	0	5	.0559	3 4 3 3
5	0	5	.0559	3 4 5 3
5	0	5	.0559	3 5 6 7
5	0	5	.0559	3 6 5 4
5	0	5	.0559	3 3 6 7 8
5	0	5	.0559	3 4 3 3 2
5	0	5	.0559	3 4 5 3 3
5	0	5	.0559	3 5 6 7 9
5	0	5	.0559	1 2 3 6 7 8
5	0	5	.0559	2 5 4 7 8 5
5	0	5	.0559	3 3 6 7 8 10
5	0	5	.0559	3 4 3 3 2 3
5	0	5	.0559	1 2 3 6 7 8 10
5	0	5	.0559	2 5 4 7 8 5 7
5	0	5	.0559	4 3 6 5 4 5 4
5	0	5	.0559	3 7 4 3 2 3 5 4
5	0	5	.0559	4 3 6 5 4 5 4 7
5	0	5	.0559	3 4 6 2 2 2 3 5 4
5	0	5	.0559	3 7 4 3 2 3 5 4 7
5	0	5	.0559	3 4 3 4 1 2 2 3 5 4
5	0	5	.0559	3 4 6 2 2 2 3 5 4 7
5	0	5	.0559	3 4 3 1 3 1 2 2 3 5 4
5	0	5	.0559	3 4 3 4 1 2 2 3 5 4 7
5	0	5	.0559	3 4 3 1 3 1 2 2 3 5 4 7
6	5	1	.0616	4 2
21	13	8	.0742	3 5 3
3	3	0	.0842	4 1
3	3	0	.0842	3 5 3 2
3	3	0	.0842	1 2 5 3 2
3	3	0	.0842	3 5 5 6 2
3	3	0	.0842	4 4 3 3 5 2
3	3	0	.0842	3 4 2 2 3 5 2
3	3	0	.0842	1 2 4 2 2 3 5 2
3	3	0	.0842	2 2 3 2 2 3 5 2
4	0	4	.0995	3 5 8 6 6
4	0	4	.0995	3 5 6 7 9 9
4	0	4	.0995	3 3 6 7 8 10 6
4	0	4	.0995	1 2 3 6 7 8 10 6
4	0	4	.0995	2 5 4 7 8 5 7 5
4	0	4	.0995	4 3 6 5 4 5 4 7 5
4	0	4	.0995	3 7 4 3 2 3 5 4 7 5
4	0	4	.0995	3 4 6 2 2 2 3 5 4 7 5
4	0	4	.0995	3 4 3 4 1 2 2 3 5 4 7 5
4	0	4	.0995	3 4 3 1 3 1 2 2 3 5 4 7 5

Thermo. data - random 4

Total = 73 Actives = 32 Inactives = 41

total	act	inact	Bip	SEV descriptors
8	7	1	.0153	3 4 6 4 4 4
8	7	1	.0153	3 4 3 4 3 4 4
8	7	1	.0153	3 4 3 1 3 3 4 4
5	5	0	.0162	3 5 8 7 7
4	4	0	.0369	1 2 5 8 9
4	4	0	.0369	3 7 6 5 4 4
4	4	0	.0369	4 5 8 7 5 5
4	4	0	.0369	3 4 6 4 4 4 4
4	4	0	.0369	3 7 6 5 4 4 5
4	4	0	.0369	3 4 3 4 3 4 4 4
4	4	0	.0369	3 4 6 4 4 4 4 5
4	4	0	.0369	3 4 3 1 3 3 4 4 4
4	4	0	.0369	3 4 3 4 3 4 4 4 5
4	4	0	.0369	3 4 3 1 3 3 4 4 4 5
9	7	2	.0428	3 7 6 5 4
3	3	0	.0842	3 5 6 4
3	3	0	.0842	3 5 8 7 7 7
3	3	0	.0842	4 7 4 6 7 6
3	3	0	.0842	3 5 8 4 5 5 4
3	3	0	.0842	4 3 4 4 6 7 6
3	3	0	.0842	1 2 5 8 4 5 5 4
3	3	0	.0842	3 5 5 7 2 3 5 4
3	3	0	.0842	4 4 4 3 5 2 3 5 4
3	3	0	.0842	3 5 2 2 3 5 2 3 5 4
3	3	0	.0842	1 2 5 2 2 3 5 2 3 5 4
3	3	0	.0842	3 2 3 2 2 3 5 2 3 5 4
4	0	4	.0995	3 4 6 5
4	0	4	.0995	3 5 3 1
4	0	4	.0995	3 5 8 1 1
4	0	4	.0995	1 2 5 8 1 1
4	0	4	.0995	2 2 4 3 1
4	0	4	.0995	3 4 6 5 4
4	0	4	.0995	3 5 3 1 3
4	0	4	.0995	2 2 4 3 1 3
4	0	4	.0995	3 4 4 4 3 4
4	0	4	.0995	3 5 3 1 3 3
4	0	4	.0995	2 2 4 3 1 3 3
4	0	4	.0995	3 5 3 1 3 3 4
4	0	4	.0995	2 2 4 3 1 3 3 4

Thermo. data - random 5

Total = 73 Actives = 32 Inactives = 41

total	act	inact	Bip	SEV descriptors
8	0	8	.0099	1 3 4 3
8	0	8	.0099	2 3 3 3
12	1	11	.0102	2 3
7	0	7	.0176	4 4 3 5
7	0	7	.0176	1 3 4 3 5
7	0	7	.0176	2 3 3 3 5
20	4	16	.0241	3 6
10	1	9	.0275	1 3 4
10	1	9	.0275	2 3 3
9	1	8	.0446	3 6 7
12	2	10	.0498	1 3
5	0	5	.0559	3 4 3 3
5	0	5	.0559	3 4 5 3
5	0	5	.0559	3 6 5 4
5	0	5	.0559	3 4 3 3 2
5	0	5	.0559	3 4 5 3 3
5	0	5	.0559	3 4 3 3 2 3
6	5	1	.0616	4 5 8 7 5
6	5	1	.0616	4 7 6 8 8
15	10	5	.0644	4 7 4 6
15	10	5	.0644	4 3 4 4 6
17	11	6	.0686	3 5 8 4
17	11	6	.0686	4 3 4 4
17	11	6	.0686	1 2 5 8 4
21	13	8	.0742	4 7 4
8	6	2	.0780	3 4 4 4
3	3	0	.0842	3 5 6 4
3	3	0	.0842	3 5 5 6 2
3	3	0	.0842	3 5 8 8 6
3	3	0	.0842	3 5 8 7 7 7
3	3	0	.0842	4 4 3 3 5 2
3	3	0	.0842	3 4 2 2 3 5 2
3	3	0	.0842	1 2 4 2 2 3 5 2
3	3	0	.0842	2 2 3 2 2 3 5 2
4	0	4	.0995	3 5 4

Thermo. data - random 6 used as test set

Total = 73 Actives = 32 Inactives = 41

total	act	inact	Bip	SEV descriptors
<hr/>				
10	1	9	.0275	4 7 5 7
9	1	8	.0446	3 4 5 4
9	1	8	.0446	3 5 8 5
9	1	8	.0446	1 2 4 5 4
9	1	8	.0446	2 2 3 5 4
9	1	8	.0446	3 4 5 4 7
9	1	8	.0446	3 7 6 5 4
9	1	8	.0446	1 2 4 5 4 7
9	1	8	.0446	2 2 3 5 4 7
<hr/>				
18	4	14	.0505	4 7 5
6	5	1	.0616	4 2
6	5	1	.0616	4 7 710 5
6	5	1	.0616	4 610 4 4 5
6	5	1	.0616	4 9 3 4 4 4 5
6	5	1	.0616	4 3 6 3 4 4 4 5
23	6	17	.0637	3 4 6 4 4
23	6	17	.0637	3 4 3 4 3 4
23	6	17	.0637	3 4 3 1 3 3 4
8	1	7	.0717	3 4 6 4 4 4
8	1	7	.0717	3 4 3 4 3 4 4
8	1	7	.0717	3 4 3 1 3 3 4 4
3	3	0	.0842	4 610 4 3
3	3	0	.0842	4 7 7 9 2
3	3	0	.0842	3 7 6 5 5 4
3	3	0	.0842	4 610 4 3 2
3	3	0	.0842	4 7 710 5 2
3	3	0	.0842	4 9 3 4 4 3
3	3	0	.0842	3 7 6 5 5 4 5
3	3	0	.0842	4 3 6 3 4 4 3
3	3	0	.0842	4 610 4 4 5 2
3	3	0	.0842	4 9 3 4 4 3 2
3	3	0	.0842	4 3 6 3 4 4 3 2
3	3	0	.0842	4 9 3 4 4 4 5 2
3	3	0	.0842	4 3 6 3 4 4 4 5 2
30	9	21	.0881	4 5 8
27	8	19	.0964	3 7 6 5
4	0	4	.0995	3 5 8 5 6
4	0	4	.0995	4 7 5 7 5
4	0	4	.0995	4 7 710 2
4	0	4	.0995	3 4 5 4 7 5
4	0	4	.0995	3 7 6 5 4 4
4	0	4	.0995	4 5 8 7 5 5
4	0	4	.0995	4 610 4 4 2
4	0	4	.0995	4 7 710 2 2

4	0	4	.0995	1 2 4 5 4 7 5
4	0	4	.0995	2 2 3 5 4 7 5
4	0	4	.0995	3 4 6 4 4 4 4
4	0	4	.0995	3 7 6 5 4 4 5
4	0	4	.0995	4 610 4 4 2 2
4	0	4	.0995	4 9 3 4 4 4 2
4	0	4	.0995	3 4 3 4 3 4 4 4
4	0	4	.0995	3 4 6 4 4 4 4 5
4	0	4	.0995	4 3 6 3 4 4 4 2
4	0	4	.0995	4 9 3 4 4 4 2 2
4	0	4	.0995	3 4 3 1 3 3 4 4 4
4	0	4	.0995	3 4 3 4 3 4 4 4 5
4	0	4	.0995	4 3 6 3 4 4 4 2 2
4	0	4	.0995	3 4 3 1 3 3 4 4 4 5

References

1. Topliss, J. and Edwards, R., J. Med. Chem., **22**, 1238 (1979).
2. Klopman, G., and Kalos, A., J. Comput. Chem., **6**, 492 (1985).
3. Topliss, J. and Costello, R., J. Med. Chem., **15**, 1066 (1972).
4. Hamburg, M., "Statistical Analysis for Decision Making",
Harcourt, Brace & World Inc., NY, 1970.
5. Pietrzyk, D. and Frank, C., "Analytical Chemistry", 2nd ed.,
Acedemic Press, NY, 1979.
6. Shoemaker, D., Garland, C., Steinfeld, J. and Nibler, J.,
"Experiments in Physical Chemistry", 4th ed., McGraw-Hill Inc.,
1981.
7. Mandel, J., "The Statistical Analysis of Experimental Data",
Dover Publications Inc., NY, 1984.
8. Conover, W., "Practical Nonparametric Statistics", 2nd ed.,
Wiley & Sons, NY, 1980.
9. Wold, S. and Dunn, W., J. Chem. Inf. Comput. Sci., **23**, 6 (1983).

Bibliography

Chapter 1

1. Crum-Brown, A. and Fraser, T.R., Trans. Royal Soc. Edinburgh, **25**, 151 (1868).
2. Richardson, B.J., Medical Times and Gazette, **2**, 703 (1869).
3. Richet, C.R., Seances Soc. Biol. Ses Fil., **9**, 775 (1893).
4. Overton, E., Vierteljahresschr. Naturforsch. Ges. Zuerich, **44**, 88 (1899).
5. Dunn, W.J., "Annual Reports in Medicinal Chemistry", Vol. 8, Quantitative Structure-Activity Relationships, R.V. Heinzelman (Ed.), Academic, New York, 1973.
6. Hansch, C. and Dunn, W.J., J. Pharm. Sci., **61**, 1 (1972).
7. Tute, M.S., "Comprehensive Medicinal Chemistry", Vol. 4, Quantitative Drug Design, C.A. Ramsden (Ed.), Pergamon, Oxford, 1990.
8. Albert, A., Rubbo, S.D. and Burvill, M.I., Br. J. Exp. Pathol., **30**, 159 (1949).
9. Denny, W.A., Cain, B.F., Atwell, G.J., Hansch, C., Panthananickal, A. and Leo, A., J. Med. Chem., **25**, 276 (1982).
10. Bell, P.H. and Roblin, R.O., J. Am. Chem. Soc., **64**, 2905 (1942).
11. Hammett, L.P., "Physical Organic Chemistry", McGraw-Hill, New York, 1970.
12. Taft, R.W., "Steric Effects in Organic Chemistry", M.S. Newman (Ed.), Wiley, New York, 1956.
13. Hansch, C., Maloney, P.P., Fujita, T. and Muir, R.M., Nature (London), **194**, 178 (1962).
14. Hansch, C., Acc. Chem. Res., **2**, 232 (1969).
15. Penniston, J.T., Beckett, L., Bentley, D.L. and Hansch, C., Mol. Pharmacol., **5**, 333 (1969).
16. Free, S.M. and Wilson, J.W., J. Med. Chem., **7**, 395 (1964).
17. Martin, Y.C., "Quantitative Drug Design", Dekker, New York, 1978.
18. Martin, Y.C., J. Med. Chem., **24**, 229 (1981).

19. Verloop, A., Hoogenstraaten, W. and Tipker, J., "Drug Design", Vol. 7, E.J. Ariens (Ed.), Academic, New York, 1976.
20. Franke, R., "Theoretical Drug Design Methods", Elsevier, Amsterdam, 1984.
21. Kollman, P., Acc. Chem. Res., **18**, 105 (1985).
22. Mabilia, M., Pearlstein, R.A. and Hopfinger, A.J., "Topics in Molecular Pharmacology", Vol. 3, A. Burgen, G. Roberts and M. Tute (Eds.), Elsevier, Amsterdam, 1986.
23. Kier, L.B. and Hall, L.H., "Molecular Connectivity in Structure - Activity Analysis", Research Studies Press, Letchworth, 1986.
24. Neuman, J.R., Scientific American, **189**, 66 (1953).
25. Wilson, R.J., "Introduction to Graph Theory", Oliver and Boyd, Edinburgh, 1972.
26. Biggs, N., Lloyd, E. and Wilson, R., "Graph Theory 1736-1936", Clarendon, Oxford, 1976.
27. Cayley, A., Phil. Mag., **13**, 172, (1857).
28. Trinajstić, N., "Chemical Graph Theory", Vol. 1, CRC, Boca Raton, 1983.
29. Wiener, H., J. Am. Chem. Soc., **69**, 17 (1947).
30. Rouvray, D.H., Sci. Am., **255**, 40 (1986).
31. Wiener, H., J. Am. Chem. Soc., **69**, 2636, (1947).
32. Wiener, H., J. Phys. Chem., **52**, 1082, (1948).
33. Bonchev, D., Mekenjan, P., Protic, G. and Trinajstić, N., J. Chromatogr., **176**, 149 (1979).
34. Seybold, P.G., Int. J. Quantum Chem., **10**, 95 (1983).
35. Randić, M., J. Chem. Phys., **60**, 3920 (1974);
J. Chem. Inf. Comput. Sci., **26**, 134-148, (1986).
36. Randić, M., J. Am. Chem. Soc., **97**, 6609 (1975).
37. Kier, L.B. and Hall, L.H., "Molecular Connectivity in Chemistry and Drug Research", Academic, New York, 1976.

38. Kier, L.B. and Hall, L.H., "Molecular Connectivity in Structure - Activity Analysis, Research Studies Press, Letchworth, 1986.
39. Kleinbaum, D. and Kupper, L., "Applied Regression Analysis and Other Multivariable Methods", Duxbury, Belmont, 1978.
40. Balaban, A.T., Chem. Phys. Lett., **89**, 399 (1982).
41. Balaban, A.T., J. Chem. Inf. Comput. Sci., **25**, 334 (1985).
42. Rashevsky, N., Bull. Math. Biophysics, **17**, 229, (1955);
Bull. Math. Biophysics, **16**, 317, (1954).
43. Shannon, C. and Weaver, W., "Mathematical Theory of Communication", University of Illinois, Urbana, 1949.
44. Bonchev, D. and Trinajstić, N., Int. J. Quantum Chem., **16**, 463 (1982).
45. Harary, F., "Graph Theory", Addison Wesley, Reading, 1971.
46. Bonchev, D. and Trinajstić, N., J. Chem. Phys., **67**, 4517 (1977).
47. Klopman, G., Raychaudhury, C. and Henderson, R.V., Math. Comput. Modell., **11**, 635 (1988).

Chapter 2

1. Norrington, F.E., Hyde, R.M., Williams, S.G. and Wooten, R., J. Med. Chem., **18**, 604 (1975).
2. Kier, L.B. and Hall, L.H., "Molecular Connectivity in Structure - Activity Analysis", Research Studies Press, Letchworth, 1986.
3. Klopman, G., Raychaudhury, C. and Henderson, R.V., Math. Comput. Modell., **11**, 635 (1988).
4. Trinajstić, N., "Chemical Graph Theory", Vol. 1, CRC, Boca Raton, 1983.
5. Shannon, C. and Weaver, W., "Mathematical Theory of Communication", University of Illinois, 1949.
6. Shulgin, A., Sargent, J. and Narenjo, C., Nature, **221**, 537 (1969).
7. Gupta, S.P., Singh, P. and Bindal, M.C., Chem. Rev., **83**, 633 (1983).
8. Baggeley, A.R., "Intermediate Correlational Methods", Wiley, New York, 1964.
9. Klopman, G. and Raychaudhury, C., J. Comput. Chem., **9**, 232 (1988); J. Chem. Inf. Comput. Sci., **30**, 12 (1990).

Chapter 3

1. Kier, L. and Hall, L., "Molecular Connectivity in Chemistry and Drug Research", Academic Press, N.Y., 1976.
2. Trinajstić, N., "Chemical Graph Theory" Vol.I&II, CRC Press, Inc., 1983.
3. Randić, M., J. Am. Chem. Soc., **97**, 6609 (1975).
4. Kier, L., J. Pharmaceutical Sciences, **69**, 807 (1980).
5. Grossman, S., Dzonova, B., Randić, M., Int. J. Quantum Chem., **12**, 123 (1986).
6. Klopman, G., Raychaudhury, C., and Henderson, R., Mathl Comput. Modelling, **2**, 635 (1988).
7. Müller, W.R., Szymanski, K., Knop, K., and Trinajstić, N., J. Comput. Chem., **8**, 170 (1987).
8. Shafer, G., Statistical Science, **2**, 3 (1987).
9. Davis, R., "The AI Business", 4th ed., P. Winston and K. Prendergast (Eds.), MIT Press, Cambridge, 1986.
10. Laplace, "A Philosophical Essay on Probabilities", Dover Publications, N.Y. 1951.
11. Bailey, D.E., "Probability and Statistics - Models for Research", John Wiley & Sons, N.Y., 1971.

Chapter 4.1

1. Lowry, T.H. and Richardson, K.S., "Mechanism and Theory in Organic Chemistry", Harper & Row, N.Y., 1976.
2. Mahler, H. and Cordes, E., "Biological Chemistry", 2nd ed., Harper & Row, N.Y., 1971.
3. Klopman, G. and McGonigal, M., J. Chem. Inf. Comput. Sci, **21**, 48 (1981).

Chapter 4.2

1. Rao, T.K., Lijinsky, W. and Epler, J.L., "Genotoxicology of N-Nitroso Compounds", Plenum Press, N.Y., 1984.
2. Magee, P.N. and Barnes, J.M., Brit. J. Cancer, **10**, 114 (1956).
3. Lijinsky, W., Loo, J. and Ross, A., Nature, **218**, 1174 (1968).
4. Keefer, L.K., Lijinsky, W. and Garcia, H., J. Natl. Cancer Inst., **51**, 299 (1973).
5. Montesano, R. and Bartsch, H., Mutat. Res., **32**, 179 (1976).
6. Singer, G.M. and Lijinsky, W., J. Agric. Food Chem., **24**, 550 (1976).
7. Lijinsky, W. and Taylor, H.W., Cancer Res., **36**, 1988 (1976).
8. Wold, S. and Dunn, W., J. Chem. Inf. Comput. Sci., **23**, 6 (1983).

Chapter 4.3

1. Katzung, B.G., "Basic and Clinical Pharmacology", 4th ed., Appleton & Lange, Norwalk, 1989.
2. Gerald, M.C., "Pharmacology: An Introduction to Drugs", 2nd ed., Prentice-Hall, Englewood Cliffs, 1981.
3. Shulgin, A., Sargent, T. and Naranjo, C., Nature (London), **221**, 537 (1969).
4. Barfknecht, C., Nichols, D. and Dunn, W., J. Med. Chem., **18**, 208 (1975).
5. Klopman, G. and Macina, O., J. Theor. Biol., **113**, 637 (1985).
6. Kier, L. and Hall, L., J. Med. Chem., **20**, 1631 (1977).

Chapter 4.4

1. Morihara, K., Biochem. Biophys. Res. Commun., **26**, 656 (1967).
2. Holmes, M. and Matthews, B., J. Molec. Biol., **160**, 623 (1982).
3. Nishino, N. and Powers, J., Biochemistry, **17**, 2846 (1978).
4. Feder, J., Brougham, L. and Wildi, B., Biochemistry, **6**, 1186 (1974).
5. Kam, C., Nishino, N. and Powers, J., Biochemistry, **18**, 3032 (1979).
6. Nishino, N. and Powers, J., Biochemistry, **18**, 4340 (1979).
7. Klopman, G. and Bendale, R.D., J. Theor. Biol., **136**, 67 (1987).
8. Weaver, L., Kester, W. and Matthews, B. J. Molec. Biol., **114** 119 (1977).

Chapter 5

1. Topliss, J. and Edwards, R., J. Med. Chem., **22**, 1238 (1979).
2. Klopman, G., and Kalos, A., J. Comput. Chem., **6**, 492 (1985).
3. Topliss, J. and Costello, R., J. Med. Chem., **15**, 1066 (1972).
4. Hamburg, M., "Statistical Analysis for Decision Making",
Harcourt, Brace & World Inc., NY, 1970.
5. Pietrzyk, D. and Frank, C., "Analytical Chemistry", 2nd ed.,
Academic Press, NY, 1979.
6. Shoemaker, D., Garland, C., Steinfeld, J. and Nibler, J.,
"Experiments in Physical Chemistry", 4th ed., McGraw-Hill Inc.,
1981.
7. Mandel, J., "The Statistical Analysis of Experimental Data",
Dover Publications Inc., NY, 1984.
8. Conover, W., "Practical Nonparametric Statistics", 2nd ed.,
Wiley & Sons, NY, 1980.
9. Wold, S. and Dunn, W., J. Chem. Inf. Comput. Sci., **23**, 6 (1983).