Design of Molecules from Quantitative Structure-Activity Relationship Models. 1. Information Transfer between Path and Vertex Degree Counts

Lemont B. Kier*

Department of Medicinal Chemistry, School of Pharmacy, Virginia Commonwealth University, Richmond, Virginia 23298

Lowell H. Hall

Department of Chemistry, Eastern Nazarene College, Quincy, Massachusetts 02170

Jack W. Frazer

Scientific Computing Group, Sterling Winthrop Research, Malvern, Pennsylvania 19355

Received September 21, 1992

The concept is presented for obtaining a set of graphs of molecular structures from QSAR equations and a target property or activity value. Background for this inverse imaging problem is presented. The outline is developed for a general method in which molecular connectivity indexes from QSAR equations are transformed into path counts and then into graph vertex degrees. This conversion is carried out by exact relating equations. The further construction of graphs corresponding to the degree sets is outlined. Examples are given for each aspect of the process. Subsequent papers will present derivation and proofs and a detailed example of the inverse imaging.

INTRODUCTION

The past 15 years has witnessed the burgeoning of a new paradigm in quantitative structure-activity relationships (QSAR). This paradigm is built upon the representation of molecular structure by nonempirical molecular indexes based upon chemical graph theory. In particular, the most widely used structure description method is molecular connectivity.[1,2] This method, developed by Kier and Hall[3,4] is an approach for encoding the structure of covalently bonded molecules with varying unsaturation, heteroatom content, and degrees of branching and cyclization. Using this and similar methods, QSAR equations of high statistical quality have been developed for a wide range of physical properties and biological activities.[5-9]

The most important value of these QSAR equations lies in their potential ability to predict new candidate molecules with desired physical or biological properties. From a sound equation, or a set of equations, we can approach the process of compound design with confidence, replacing the expensive and slow Edisonian-like approach of trial and error.

A desirable characteristic of high-quality QSAR equations, built up from nonempirical structure indexes, is the ability to generate a molecular structure from a preselected value (called the target value) for a property or activity, P. Thus, if X and Y are nonempirical structure indexes which truly encode the basic elements of molecular structure, then a relation between P and X,Y may be developed by statistical methods as

$$P = aX + bY + c$$

Starting with a preselected (target) value of the property P, it should ultimately be possible to progress through several well-defined steps to develop a list of target molecules which have that desired property value. This strategy is not a database-searching technique but a method based upon precise mathematical relationships among structure descriptors as used in QSAR equations and the predicted structures of interest. The selection process is bounded by relations based upon the nature of molecular graphs and expressed in certain mathematical relations. This process, molecule building from QSAR equations, has been called by several names: graph reconstruction, inverse structure generation, inverse imaging, and so forth. We have generally referred to this problem as that of inverse imaging.

A few studies have been reported on this difficult but extremely important problem. The Zefirov group in Moscow has made some progress by analyzing single-index QSAR equations.[10,11] Their approach focuses on edge types in hydrogen-suppressed graphs. From edge types, vertex types are encoded through the use of restricting relations, and then on to the construction of graphs via the Faradzhev algorithm.[12] In this approach, a $^1\chi$ index of even moderate magnitude leads to a very large number of graph possibilities, even for a narrow range of property value. A group of investigators in Czechoslovakia has laid some ground work for the study of the inversing problem by erecting a series of theorems involving edge and vertex enumeration.[13] Klopman and his group developed a vertex index and gave an example for inverse imaging.[14] Contreras et al.[15] have developed algorithms for the exhaustive generation of acyclic isomers, which might be extended to more general structures.

Over the past 5 years, we have addressed this very important problem and have made some significant progress toward a solution. In this paper we report the development of a concept, the general outline of our approach, in order to stimulate others to advance the status of this inverse imaging problem. This paper includes the presentation of a set of equations which relate low-order path counts to vertex degrees. A simple example is included in this paper to illustrate the various steps in the inversing process.

In the second paper, these relating equations are derived and proven so as to demonstrate their generality and develop insight into structure representation. The third paper extends the formalism to include paths of length three, describes ways to construct the graphs from a set of vertex degrees, and

* Author to whom correspondence should be addressed.

0095-2338/93/1633-0143$04.00/0

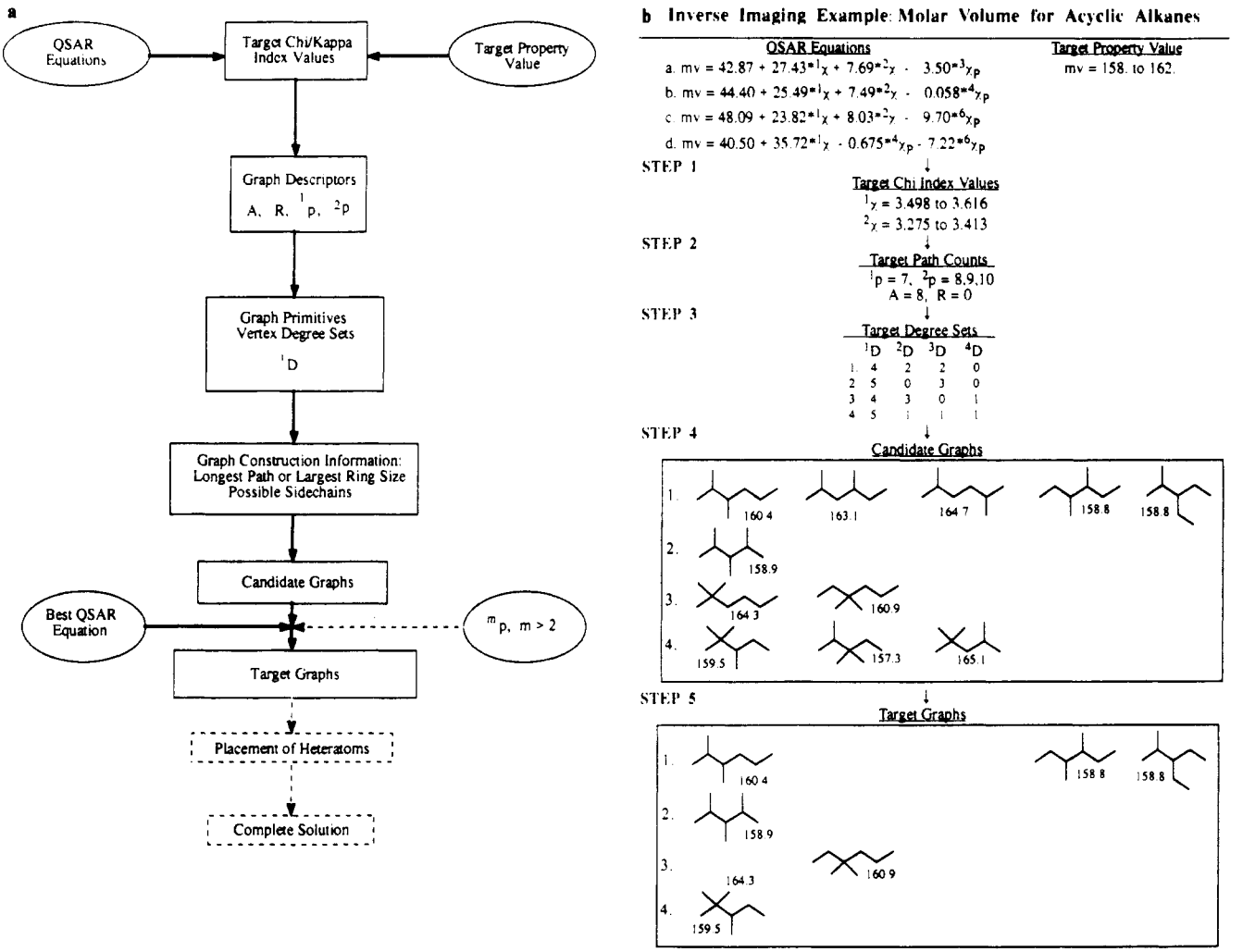


Figure 1. (a) Flow chart proposed for the development of inverse QSAR. (b) Examples to illustrate each step in the inversing process, based on the experimental molar volume (MW/density) for liquid acyclic alkanes.

illustrates inverse imaging with a detailed example. Subsequent papers will describe our work on higher-order path counts and will extend the overall generality of our approach.

GENERAL APPROACH

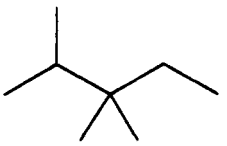
A. Information Modules. We have approached the problem of deriving a set of molecular structures by basing our QSAR equations on molecular connectivity indexes. The information necessary for the inverse imaging lies in a series of modules, connected by a sequence of steps in the inversing process, as shown in Figure 1. These modules include the basic QSAR equations, the graph index values, the path counts, and finally, the graph primitives information, the vertex degrees. The candidate graphs can be constructed from adjacency matrices obtained from the vertex degree counts. Higher-order path counts supply information which act to bound the number of constructed graphs by acting as restricting relations. The flow of information is as follows: step 1, QSAR equations plus target property value → step 2, targe χ and κ index values → step 3, path counts for orders one and two → step 4, sets of vertex degree counts → step 5, graphs which are candidates for molecules with the desired property value → target graphs obtained by the filtering through the best QSAR equation and higher-order path count information.

The information modules in this model of structure–activity interrelation are illustrated in Figure 1a. The basic elements are defined in Table I along with the symbols which we will use. To illustrate each step in the process, we have developed a simple example. A set of 54 acyclic alkanes was selected along with molar volumes computed from experimental liquid density. QSAR equations were developed from this data set and computed molecular connectivity χ indexes.[16] The inversing process was carried out as outlined in the flow chart of Figure 1a. The results of each step are shown in Figure 1b. This simple example, limited only to small acyclic alkanes, is meant to illustrate each step in the process rather than the efficacy of the overall inversing strategy. A subsequent paper in this series will more fully develop a complete example. A further example is given in Figure 2 to illustrate the terms and symbols used.

Figure 1a illustrates the flow of information between the modules. We begin with a set of QSAR equations developed from a data set containing a property (or activity) as a dependent variable together with several nonempirical topo-

Table I. Definition of Terms and Symbols Used in Inverse Imaging

symbol	definition
A	count of graph atoms (vertexes; skeletal groups such as $-CH_2-$, –OH, >CH–, etc.)
^1p	count of paths of order 1 (graph edges, skeletal σ bonds)
^2p	count of paths of order 2 (two contiguous edges)
^mp	count of paths of order m (path of m contiguous edges)
i	vertex degrees, no. of neighbor vertexes; also called the molecular connectivity simple δ value
^iD	count of atoms (vertexes) with degree i
R	count of rings in graph $[R = {}^1p - (A - 1)]$



R = 0, A = 8, ^1p = 7, ^2p = 10, ^1D = 5, ^2D = 1, ^3D = 1, ^4D = 1

Figure 2. Hydrogen-suppressed graph for the 2,3,3-trimethylpentane molecule, showing the values for the some of the quantities used in this paper.

logical indexes as independent variables. To pass from the QSAR equation(s) to the desired (target) molecular structures which are predicted to have a specific property value, we must acquire information within several intermediate modules. From the QSAR equation(s) and the target property value, we must obtain values for the topological indexes which appear in the QSAR equations. This information must then be converted into graph descriptors, that is, values for the atom count and path counts. These quantities are more closely related to the desired graphs than the original topological indexes.

From these graph descriptors, a critical step is necessary if we are to complete the inversing process without the need for massive search techniques. That step is the derivation of information on the graph primitives, the count of atoms with various degrees, that is, the set of vertex degrees, ^iD. From the ^iD we can deduce information about the graphs which can be constructed. This information includes the longest path(s) in the graph or the largest ring size and the nature of side chains. When this level of graphical information is reached, the Faradzhev algorithm or an analogous procedure may be used to construct graphs from sets of vertex degrees.[12,13,15] Finally, beyond these steps lies the more complex problem of dealing with heteroatoms and various bonding schemes. Dashed lines in Figure 2 represent aspects of the problem yet to be described.

B. Extraction of Target χ Index Values from QSAR Equations. The first step in the inverse imaging process is to obtain a set of QSAR equations based on the property under study and a wide range of graph-based indexes. The indexes could include the number of atoms, path counts, and molecular connectivity indexes. For a given problem to be investigated, a specific property value is selected as the target value to meet the goal of the investigation.

Beginning with the data set, QSAR equations are developed in the usual manner using multiple linear regression. The target property value, together with the set of QSAR equations, are then solved simultaneously (inverted) to obtain the set of target index values. There must be at least as many QSAR equations as there are indexes in the set of equations. See Figure 1b for a specific example.

C. Determination of Path Counts from Target χ Index Values. Each χ index is a weighted count of the corresponding path count, a typical quantity in chemical graph theory. A path of length 1 is simply a graph edge; the path-1 count (edge count) is ^1p. A path is, in general, a sequence of edges in which no vertex is repeated. Two consecutive edges are called a path of length 2, and their count is given as ^2p. Examples are given in Figure 2.

The relation between a χ index and the path count upon which it depends is not unique. A given value of χ index can arise from more than one path count value. A typical relation is illustrated for the dependence of $^2\chi$ on ^2p count as shown in Figure 3 for 848 acyclic and monocyclic alkanes. Whether or not this relation may be expressed in some simple analytical form, a graphical solution will produce the desired possible

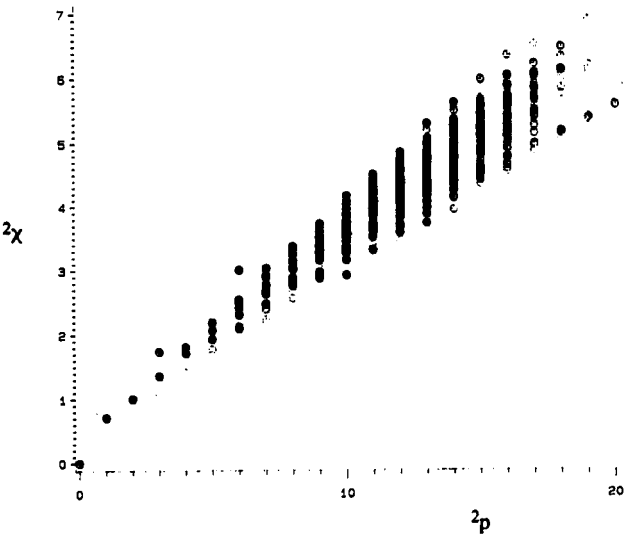
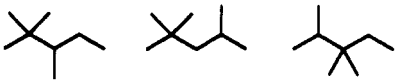


Figure 3. Plot of the $^2\chi$ index versus the count of paths of length 2 (^2p) for 672 alkanes, both acyclic and monocyclic.

Table II. Example Showing Vertex Degree Set[a]

Input Basic Graph Descriptors	
R = no. of rings = 0	A = no. of atoms = 8
^1p = no. of paths of length 1 = 7	^2p = no. of paths of length 2 = 10
Output Graph Primitive Information: Vertex Degree Set	
^1D = no. of vertexes with degree 1 = 5	^2D = no. of vertexes with degree 2 = 1
^3D = no. of vertexes with degree 1 = 5	^4D = no. of vertexes with degree 1 = 1

Graphs Generated from the Above Set of Vertex Degrees



[a] Generated from a count of path-1, path-2, and number of rings for an acyclic graph set along with the corresponding graphs.

values of ^2p which correspond to a target value for $^2\chi$. For example, from Figure 3 it can be seen that the value for $^2\chi$ of 3.8 corresponds to the values of 10, 11, 12, and 13 for ^2p. See Figure 1b for a specific example in this case only acyclic graphs were considered in obtaining the ^1p and ^2p counts.

D. Transition from Path Counts to Vertex Degrees. This crucial step in the information transfer was developed in the early stages of our work. Three equations were conceived for acyclic graphs, relating path counts of orders 1 and 2 to the set of vertex degrees:

$$^3D + 3\,^4D = {}^2p - {}^1p + 1 \quad (1)$$

$$^1D + {}^4D = {}^2p - {}^1p + 3 \quad (2)$$

$$^2D - 3\,^4D = -2\,^2p + 3\,^1p - 3 \quad (3)$$

These equations were manipulated to obtain an equation involving only the vertex degrees:

$$^1D - {}^3D - 2\,^4D = 2 \quad (4)$$

From these three equations, several other relations can be obtained which may be useful in various situations.

The equations above apply to acyclic graphs. For a graph with rings, there is a simple relation between the number of edges, ^1p, the number of vertexes, and the number of rings. In acyclic graphs the number of edges is one less than the number of vertexes, A. For each ring there is one additional

edge; hence, the number of rings R is $R = {}^1p - (A - 1)$. The relations given above can be extended to cyclic graphs with the following results:

$$ {}^3D + 3\,{}^4D = {}^2p - {}^1p + 1 - R \quad (5) $$

$$ {}^1D + {}^4D = {}^2p - {}^1p + 3 - 3R \quad (6) $$

$$ {}^2D - 3\,{}^4D = -2\,{}^2p + 3\,{}^1p - 3 + 3R \quad (7) $$

These relations will be derived and proven in the second paper in this series along with several other forms of these relations. Further, these relations will be extended to vertex degrees of fifth and sixth orders.

The set of independent relations given as eqs 5–7 relate the four ${}^iD$ values found in typical organic molecules to path and ring counts. We can rearrange them for explicit solution of ${}^1D$, ${}^2D$, and ${}^3D$, as follows:

$$ {}^1D = -{}^4D + {}^2p - {}^1p + 3 - 3R \quad (6a) $$

$$ {}^2D = 3\,{}^4D + -2\,{}^2p + 3\,{}^1p - 3 + 3R \quad (7a) $$

$$ {}^3D = -3\,{}^4D + {}^2p - {}^1p + 1 - R \quad (5a) $$

From a set of R, ${}^1p$, and ${}^2p$ values, there is no unique solution for the four ${}^iD$ values. However, it is possible to select values for ${}^4D$ and solve for the other three. The values of ${}^4D$ can take on values from 0 to a maximum value determined by the number of atoms A and the number of rings R. Then it is possible to determine whether such a vertex degree set ${}^iD$ corresponds to an actual graph. There are three relations which can be used to determine the suitability of each ${}^iD$ set. First, the vertex degree count is a positive integer: ${}^iD \leq 0$. Second, the total count of vertexes of each degree is simply the total number of atoms, A; thus, $\sum {}^iD = A$. Third, the total sum of degrees in the graph must be even. This is known as the Handshake Lemma:[17] $\sum {}^iD(i) = 2n$. (It can be shown that $n = {}^1p$: the number of hands involved is twice the number of handshakes.) Thus, a practical solution is possible for converting the three graph descriptors, R, ${}^1p$, and ${}^2p$, into one or more sets of four vertex degrees.

Tables II and III give examples of the vertex degree sets computed from a set of R, ${}^1p$, and ${}^2p$ values. Table II gives an example of the vertex degree set which is obtained from the input set: $R = 0$, ${}^1p = 8$, and ${}^2p = 10$ values. Only one vertex set results in this case, and this set corresponds to three graphs, acyclic octanes. For example 2 in Table III, $R = 1$, ${}^1p = 8$, and ${}^2p = 10$, the same values as were used for path-1 and path 2 counts in example. Again only one vertex degree set results, but this set corresponds to 10 graphs with ring sizes ranging from three to eight. In general, a set of R, ${}^1p$, and ${}^2p$ values will result in more than one set of vertex degrees. A degree set may be written as four numbers representing the four degree counts: $({}^1D\,{}^2D\,{}^3D\,{}^4D)$. For example, for $R = 0$, ${}^1p = 7$, and ${}^2p = 9$, there are two degrees sets generated: (5 0 3 0) and (4 3 0 1). For the set $R = 1$, ${}^1p = 8$, and ${}^2p = 11$, two degree sets result: (3 2 3 0) and (2 5 0 1).

See Figure 1b for a specific example. More detailed and complete examples of the inversion process will be given in later papers in this series.

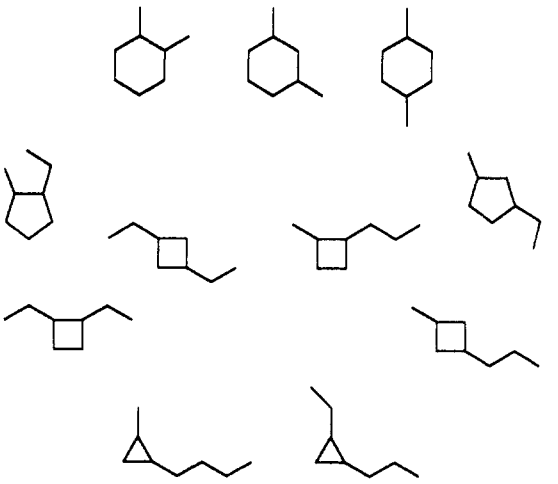
E. Construction of Connection Matrix from Degree Counts. With a set of degree counts, ${}^iD$, in hand from the R, ${}^1p$, and ${}^2p$ analysis, it is possible to translate this information into graphs (molecular structures), represented as an adjacency matrix or connection table for each graph. One approach is to make a direct comparison between R, ${}^1p$, and ${}^2p$ values and the list of alkanes with known values. This approach has limited utility in view of the complexity of molecules that we ultimately want to consider.

Table III. Example Showing Vertex Degree Set[a]

Input Basic Graph Descriptors	
R = no. of rings = 1	A = no. of atoms = 8
${}^1p$ = no. of paths of length 1 = 8	${}^2p$ = no. of paths of length 2 = 10

Output Graph Primitive Information: Vertex Degree Set	
${}^1D$ = no. of vertexes with degree 1 = 2	${}^2D$ = no. of vertexes with degree 2 = 4
${}^3D$ = no. of vertexes with degree 1 = 2	${}^4D$ = no. of vertexes with degree 1 = 0

Graphs Generated from the Above Set of Vertex Degrees



[a] Generated from a count of path-1, path-2, and number of rings for cyclic graphs along with the corresponding graphs

The matrix generation process, in our studies, begins with the deduction of the maximum path length (in an acyclic graph) or the maximum ring size (for cyclic graphs) from the ${}^iD$ values. This result corresponds to the number of diagonal elements in the matrix which can be written with as '1 0 1'. In a set of graphs with the same number of atoms and the same number of rings, an isomeric series, there may be several values for the 'longest path'. For example, in the acyclic heptanes, n-heptane has a 'longest path' of 7 whereas 2,2,3-trimethylbutane has the smallest value, 4. There are graphs among the heptanes with all the intervening values, 5 and 6. Hence, the number of vertexes in the longest path, N_1, runs from 4 to 7 in this example.

The equation for the maximum path length, ${}^{max}N_1$, is given as

$$ {}^{max}N_1 = {}^2D + {}^3D + {}^4D + 2 \quad (8) $$

This equation is based upon the idea that only branch point vertexes may be in the 'longest path' and that its length is that number plus the two terminal groups. Expansion of this work leads to another relation on the upper bound on the longest path, ${}^{max}n_{LP}$, in the graph as

$$ {}^{max}N_1 = A - (n_{\text{methyl groups}} - 2) \quad (9) $$

in which $n_{\text{methyl groups}}$ is the number of methyl groups (or terminal groups) in the graph. This 'longest path' corresponds to the least branched of the isomeric series. The lower bound on the length of the 'longest path' in a graph, ${}^{min}N_1$, is given as

$$ {}^{min}N_1 = A - {}^{max}(n_{\text{side chain atoms}}) \quad (10) $$

in which ${}^{max}(n_{\text{side chain atoms}})$ is the count of the maximum number of atoms which can be placed in side chains.

In the same manner for cyclic structures, it can be seen that the size of the largest ring possible can also be determined. Terminal vertexes may not be part of a ring system. Hence, the number of vertexes in the largest possible ring for a given vertex degree set, ${}^{max}N_r$, is simply the sum of the nonterminal vertexes:

$$ {}^{max}N_r = {}^2D + {}^3D + {}^4D \qquad (11) $$

If there is more than one ring, the ring vertexes (${}^{max}N_r$ in number) are partitioned among all the rings. The possible ring sizes range from three up to ${}^{max}N_r$. As shown in example 2 (Table III) the value of ${}^{max}N_r$ is eight, and the ring sizes vary from three to eight.

From these analyses, a set of molecular structures can be derived for a preselected set of atom degrees. See Figure 1b for a specific example in which the inversing process leads to 11 candidate graphs from step 4. The experimental molar volume is included for each graph.

Finally, the best QSAR equation (a in Figure 1b) is used to compute the molar volume for each of these candidate graphs. In this case, five of the graphs have predicted values which lie outside the target range, 160.0 ± 2. This best QSAR equation is used to screen out these graphs. The remaining six graphs are then the target graphs, the result of step 5, and the outcome of the inversing process. It is to be noted that all six of these graphs lie within the target range.

The flow chart in Figure 1a indicates that future work will include the role of higher-order paths in the inversing process. That role may include filtering of candidate structures or an influence in the graph construction process.

CONCLUSIONS

A general plan has been outlined and briefly described for the generation of molecular graphs corresponding to a preselected or target property value. This target property value is based upon a QSAR equation relating a property to molecular connectivity indices. To accomplish this generation, a series of equations has been conceived to relate path counts, ${}^ip$, in graphs to the count of atom (vertex) degrrees, ${}^iD$, and ring count, R. These relationships permit the flow of information from QSAR equations together with a particular property value (or narrow range of property value) to the corresponding adjacency matrices representing target molecules. The relating equations may be used as tools to assist the chemist in quantitative thinking about molecular structure variation. Ideas about possible molecules, based upon suggested counts of vertexes with given degrees, may be implemented by the process outlined in this paper.

This general scheme of analysis and prediction is offered as a viable alternative to large database searches. This scheme is intended to avoid a combinatorial explosion in the search for structures corresponding to a preselected set of QSAR indices. The ultimate value of a QSAR equation, namely, structure prediction, can now be approached on a more rational basis using this general plan. Subsequent papers in this series will present proof-of-concept information and examples of inverse imaging and extension to other graph quantities.

ACKNOWLEDGMENT

We wish to express appreciation for support of this work by Sterling-Winthrop Pharmaceutical Research Division, Malvern, PA, and Allied-Signal Corp., Engineered Materials Research, Des Plaines, IA.

REFERENCES AND NOTES

(1) Trinajstić, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1983; Vols. I and II.
(2) Rouvray, D. H. Predicting Chemistry from Topology. *Sci. Am.* **1986**, *255*, 40.
(3) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.
(4) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure–Activity Analysis*; John Wiley: London, 1986.
(5) Hall, L. H. Computational Aspects of Molecular Connectivity and its Role in Structure–Property Modeling. In *Computational Chemical Graph Theory*; Rouvray, D. H., Ed., Nova Press: New York, 1990; Chapter 8, pp 202–233.
(6) Kier, L. B. Indexes of Molecular Shape from Chemical Graphs. *Med. Res. Rev.* **1987**, *7*, 417–440.
(7) Hall, L. H.; Kier, L. B. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure–Property Relations. In *Reviews of Computational Chemistry*; Boyd, D., Lipkowitz, K., Eds.; VCH Publishers, Inc.: New York, 1991; Chapter 9, pp 367–422.
(8) Hall, L. H.; Kier, L. B. Enumeration, Topological Indexes and Molecular Properties in Alkanes. In *Alkanes and Cycloalkanes, Chemistry of the Functional Groups*; Patei, S., Rapoport, Z., Eds.; John Wiley: Chichester, 1992; Chapter 5, pp 186–213.
(9) Kier, L. B.; Hall, L. H. An Atom-Centered Index for Drug QSAR Models. In *Advances in Drug Design*; Testa, B., Ed.; Academic Press: New York, 1992; Vol. 22.
(10) Baskin, I. I.; Gordeeva, E. V.; Devdariani, R. O.; Zefirov, N. S.; Palyulin, V. A.; Stankevitch, I. V. Solving the Inverse Problem of Structure–Property Relations for the Case of Topological Indexes. *Dokl. Akad. Nauk. USSR* **1989**, *307*, 613–617.
(11) Skvortsova, M. I.; Baskin, I. I.; Slovokhotova, O. L.; Palyulin, V. A.; Zefirov, N. The Inverse Problem in QSAR/QSPR Studies for the Case of Topological Indices Characterizing Molecular Shape (Kier Indices). *J. Chem. Inf. Comput. Sci.*, submitted for publication.
(12) Faradzhev, I. Generation of Nonisomorphic Graphs with Given Partition of Vertex Degrees. In *Algorithmic Investigations in Combinatorics*; Faradzhev, I. A., Ed.; Nauka: Moscow, 1978; pp 11–19 (in Russia).
(13) Kvasnicka, V.; Poshipal, J. Canonical Indexing and Constructive Enumeration of Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 105.
(14) Klopman, G.; Raychandhury, C. J. Vertex Indexes of Molecular Graphs in Structure–Activity Relationships: A Study of the Convulsant–Anticonvulsant Activity of Barbiturates and the Carcinogenicity of Unsubstituted Polycylic Aromatic Hydrocarbons. *J. Chem. Inf. Comput. Sci.* **1988**, *30*, 12–19.
(15) Contreras, M. L.; Valdivia, R.; Rozas, R. Exhaustive Generation of Organic Isomers. 1. Acyclic Structures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 323–330.
(16) The computation of molecular connectivity χ indexes path and vertex degree counts was carried out using the software package Molconn-X, version 2.0, from Hall Associates Consulting; for information contact L. H. Hall.
(17) Wilson, J. R. *Introduction to Graph Theory*; Oliver & Boyd: Edinburgh, 1972; Chapter 2.