

Design of Molecules from Quantitative Structure-Activity Relationship Models.

3. Role of Higher Order Path Counts: Path 3

Lowell H. Hall* and Robert S. Dailey

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

Lemont B. Kier

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

Received December 29, 1992

The development of the inverse imaging process is continued here with attention given to the role of the count of paths of length 3 in molecular graphs. The relation between path counts and edge type counts is developed for paths of length 0, 1, 2, and 3. A general relation is developed among the edge type counts which places significant restrictions upon the edge types which can constitute a molecular graph. Independent equations are derived for relations among the edge types and path counts. A method for generating and validating edge type counts is presented for the nine edge types encountered in molecular graphs, including those for acyclic and cyclic molecules. These relations are applied to the generation of molecular graphs which represent molecules with molar volumes in a specified range of values as was done in part 1 using only path counts 0, 1, and 2. This development extends the inverse process to include the path 3 count.

INTRODUCTION

In parts 1 and 2 in this series,^{1,2} a method was described in which structure information, represented by molecular connectivity indexes in quantitative structure-activity relationship (QSAR) equations, is transformed into graph theory information from which molecules are constructed to have a property value in a desired range. In graph theory a molecule is envisioned as a collection of points and lines called vertices and edges. Terms and symbols used in these papers are defined and illustrated in part 1.

The several steps in this molecule construction process, called inverse imaging, are presented in Figure 1a of part 1 along with a summary of an example of inverse imaging in Figure 1b. This inverting process involves passage through two echelons of information. The higher level involves *graph descriptors* which characterize the whole molecular graph; the other and lower level is the *graph primitives* which describe individual skeletal groups (vertices), as well as the connections between skeletal groups (edges).

The whole molecule descriptors used here are the path counts. The edges of a chemical graph, the connections between skeletal groups, are called paths of length 1 and their count in a graph is 1p . Two contiguous edges are called a path of length 2; their count is symbolized 2p . Three contiguous edges, such as the skeleton of butane, constitute a path of length 3; their count in a graph is 3p . The atom count, A , can also be called the count of paths of length 0, $^0p = A$. The set of path counts is part of the structure information which characterizes a molecular graph and, in turn, represents the structure of the corresponding molecule.

The second and lower level echelon of graph information, called graph primitives, includes the vertex degree counts and edge type counts. In the hydrogen-suppressed graph, each vertex (skeletal group), such as a $-\text{CH}_3$, $-\text{NH}-$, $=\text{O}$, $-\text{O}-$, etc., is characterized, in part, by the count of its neighboring vertices, called the vertex degree, δ . In a graph the number of vertices with a given degree is called the vertex degree count, iD . For a $-\text{CH}_2-$, there are two neighbors, and for the $=\text{O}$ there is only one, whereas for the $>\text{N}-$ there are three

neighbors. For our purposes we will consider degrees of 1, 2, 3, and 4. The counts of the vertex degrees for a graph are called the vertex degree set. It has been shown³⁻⁷ that the graph of a molecule can be constructed from its vertex degree set. Similar information can be developed for counts of edge types, as will be shown in this paper.

The inverse imaging process requires relations between the graph descriptors (path counts) and the graph primitives (the vertex degree and edge type sets). The set of equations which provides these relations is called the relating equations. They were presented in part 1 and derived in part 2 for vertex degree sets. These equations are the link between the realm of QSAR equations for the data set, on the one hand, and molecule construction, on the other.

The inverse imaging process is closely related to the problem of graph reconstruction⁵⁻⁷ in which a set of graphs is to be (re)constructed from the set of constituent primitives, such as vertexes (with their association degrees) or edges (including the degrees of the associated vertices). The reconstruction is required to be both exhaustive and nonredundant. The work of Faradzhev⁵ has produced a general method for nonredundant generation of graphs based upon a canonical indexing of graphs. The works of Pospichal and Kvasnicka⁶ and Contreras, Valdivia, and Rozas⁷ also show methods for nonredundant generation based upon canonical indexing methods.

The process described in part 1 leads to a set of graphs, which represent molecules, called the candidate set. The candidate set contains some members which are not consistent with the desired property value range. This condition arises because there are only three equations relating the four vertex degree counts. The candidate set can be filtered in several ways to remove the undesired members, yielding the target set. In part 1 we used the best QSAR equation developed for the data set. Property values were computed from the best QSAR equation. Those candidate molecular graphs with computed values outside the desired range (target range) were filtered out; the remaining candidates constitute the target set. It is expected that there is a high probability that the

Table I. Graphs To Illustrate Path Counts, Degree Counts, and Edge Counts

graph	path counts				degree counts				edge type counts									
	0p	1p	2p	3p	1D	2D	3D	4D	11	12	13	14	22	23	24	33	34	44
	6	5	4	3	2	4	0	0	0	2	0	0	4	0	0	0	0	0
	6	5	5	3	3	2	1	0	0	1	2	0	1	1	0	0	0	0
	6	5	5	4	3	2	1	0	0	2	1	0	0	2	0	0	0	0
	6	5	6	4	4	0	2	0	0	0	4	0	0	0	0	1	0	0
	6	5	7	3	4	1	0	1	0	1	0	3	0	0	1	0	0	0
	3	3	3	0	0	3	0	0	0	0	0	0	3	0	0	0	0	0
	4	4	5	2	1	2	1	0	0	0	1	0	1	2	0	0	0	0
	5	5	7	4	2	1	2	0	0	0	2	0	0	2	0	1	0	0
	5	6	10	8	0	4	0	1	0	0	0	0	2	0	4	0	0	0

target set possesses property values in the target range. This probability is limited by the experimental error in the original data set upon which the equations are based.

It was indicated in part 1 that information from higher order path counts can also be used in this screening process. This is the topic of the present paper. We will focus our attention here on the count of paths of length 3. Future papers will generalize and apply this approach. It should be noted here that the first chemist to point out the importance of paths of length 3 was Harry Wiener in what he called the polarity number in the development of what is now called the Wiener Index.^{8,9}

DEVELOPMENT

The set of path counts from zero order up to third and higher orders contain an increasing level of complexity in molecular structure information. Molecular structure is described with an increasing level of detail as the description moves from 0p to 3p . Atom count is a very crude yet important expression of structure information. The edge count, 1p , is somewhat higher in information content but lacks specificity such as adjacency information. For example, all geometric isomers possess the same 1p value. The 2p count encodes more specific structure information. For example, for normal hexane $^2p = 4$, whereas for 2-methylpentane $^2p = 5$. Table I indicates the greater discriminating power of 2p ; however, 2-methylpentane and 3-methylpentane are not distinguished by 2p alone.¹⁰ The structure information in 3p provides that discrimination, for example, $^3p = 3$ for 2-methylpentane and $^3p = 4$ for 3-methylpentane. The set of path counts 0p , 1p , 2p , and 3p provide a discriminating profile for these isomers. For larger molecules, higher order paths are necessary for isomer discrimination.

At the level of graph primitives such as vertex degree and edge type counts, there is a similar progression of structure information. The atom count may be considered the simplest of the graph primitives. The set of vertex degrees possesses more information. In the hexane isomers the vertex degree set of normal hexane is different from the other isomers, as shown in Table I. However, 2- and 3-methylpentane have the same vertex degree sets. More detailed information is required to distinguish isomers such as the 2- and 3-methylpentanes.

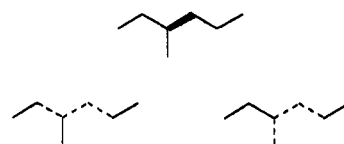


Figure 1. Hydrogen-suppressed graph for 3-methylhexane indicating the "23" edge in boldface along with the two subgraphs associated with that edge, shown in dashed lines in the two additional graphs.

Edge types contain the necessary information for discrimination. A graph edge is in part characterized by the vertex degree values of its two defining vertices. The edge is described by stating the two vertex degrees. For example, the $-\text{CH}_2-\text{CH}_2-$ edge in an alkane is called a "22" edge. The count of edges of a given type ij is symbolized ne_{ij} . The 2-methyl isomer has two "13" edge types ($ne_{13} = 2$) but the 3-methyl isomer possesses only one ($ne_{13} = 1$). This ability to distinguish among isomers using edge types, as compared to the degree types, is similar to the ability to distinguish among isomers possessed by the 3p path count compared to the 2p count. As will be shown in this paper, 3p information assists in the computation of edge type counts.

Relation of Path Counts to Edge Type Counts. In part 2,² relations were presented and derived for 1p and 2p as follows:

$$^1p = ne_{11} + ne_{12} + ne_{13} + ne_{14} + ne_{22} + ne_{23} + ne_{24} + ne_{33} + ne_{34} + ne_{44} \quad (1)$$

$$^2p = \frac{1}{2}ne_{12} + ne_{13} + \frac{3}{2}ne_{14} + ne_{22} + \frac{3}{2}ne_{23} + 2ne_{24} + 2ne_{33} + \frac{5}{2}ne_{34} + 3ne_{44} \quad (2)$$

In these equations, ne_{ij} stands for the count of edges between vertices with degrees i and j . Each edge is defined by two vertices; the edge type specifies the vertex types (degrees).

A relation can be developed between atom count, A (which may also be symbolized as 0p), and the edge type counts, ne_{ij} . Each vertex of degree i is associated with i edges of type ij and each vertex of degree j is associated with j edges of type ij . For example, in the hydrogen-suppressed graph of 3-methylhexane, the $>\text{CH}-$ skeletal group is represented by a vertex of degree 3. See Figure 1. It is associated with

Table II. Contributions of Graph Edge Types, ne_{ij} , to the Path 3 Count, 3p

edge type ne_{ij}	embedded edge	3p contribution
11		0
12		0
13		0
14		0
22		1
23		2
24		3
33		4
34		6
44		9

(connected to) three edges of the types "13", "13", and "23": one atom (vertex) is found in three edges. Hence, the contribution of these edges to atom count A is $(ne_{13} + ne_{23})/3 = (2 + 1)/3 = 1$. In general, then, the contribution of a vertex to the atom count is given by $ne_{ij}/i + ne_{ij}/j$. Thus, we can write

$${}^0p = A = \sum_{i=1}^4 \sum_{j=i}^4 (ne_{ij}/i + ne_{ij}/j) \quad (3)$$

When terms are collected, the following relation results:

$${}^0p = A = 2ne_{11} + \frac{3}{2}ne_{12} + \frac{4}{3}ne_{13} + \frac{5}{4}ne_{14} + ne_{22} + \frac{5}{6}ne_{23} + \frac{3}{4}ne_{24} + \frac{2}{3}ne_{33} + \frac{7}{12}ne_{34} + \frac{1}{2}ne_{44} \quad (4)$$

Now consider the relation between the path 3 count, 3p , and the edge type counts, ne_{ij} . Consider an edge embedded in a graph to be part of a path 3 subgraph, such as is shown in Figure 1. A 23 edge is indicated in boldface in the graph. This edge is part of two path 3 subgraphs, indicated by the paths (as dashed lines) in the two graphs below the graph of 3-methylhexane. This 23 edge contributes two paths of length 3 to the total of 3p . The contribution from this edge depends only on the edge type, 23 in this case, and, further, this contribution is independent of the contribution of the other edges. That is, one may obtain the 3p value by summing the contributions from each edge, independent of the contributions from other edges. It is observed that terminal edges correspond to subgraphs which are only two edges in length and, therefore, make a zero contribution to the total 3p . These considerations are summarized in Table II.¹⁰ As a result, it is possible to write the following relation for 3p and edge type counts:

$${}^3p = ne_{22} + 2ne_{23} + 3ne_{24} + 4ne_{33} + 6ne_{34} + 9ne_{44} = \sum_{i=1}^4 \sum_{j=i}^4 (i-1)(j-1)ne_{ij} \quad (5)$$

This relation was obtained for acyclic systems. One modification is required when the relation is extended to include cyclic graphs. The required modification applies only to three-membered rings, as is shown below.

The definition of path is "consecutive edges with distinct vertices"; that is, the path must not cross itself or terminate on the initial vertex of the sequence. In a three-membered ring there are no paths of length 3 because, in such a sequence of edges, the path terminates on itself, forming the ring. However, the unsubstituted three-membered ring does contain

edges of the type "22". In fact, $ne_{22} = 3$ and these edges contribute three to the 3p count as reckoned in eq 5. For this reason, eq 5 must be modified to exclude the 3p count generated by the 22 edges in three-membered rings:

$${}^3p = ne_{22} + 2ne_{23} + 3ne_{24} + 4ne_{33} + 6ne_{34} + 9ne_{44} - 3R_3 \quad (6)$$

in which R_3 is the count of three-membered rings.

To this point we have developed four relations between path counts and edge counts. It is also of interest to consider relations between degree counts and edge counts. In part 2, it was shown that there are also relations among the degree counts and the edge type counts:

$${}^1D = 2ne_{11} + ne_{12} + ne_{13} + ne_{14} \quad (7)$$

$${}^2D = \frac{1}{2}[ne_{12} + 2ne_{22} + ne_{23} + ne_{24}] \quad (8)$$

$${}^3D = \frac{1}{3}[ne_{13} + ne_{23} + 2ne_{33} + ne_{34}] \quad (9)$$

$${}^4D = \frac{1}{4}[ne_{14} + ne_{24} + ne_{34} + 2ne_{44}] \quad (10)$$

These relations indicate that there are restrictions upon the edge counts; they are not independent quantities. This consideration becomes important when we attempt to compute edge counts from the path counts, as will be considered below.

In part 1,¹ it was also shown that there is a relation among the degree counts, including the number of rings R :

$${}^1D - {}^3D - 2{}^4D = 2 - 2R \quad (11)$$

This is also a restricting relation among degree counts. In order for a set of vertices to constitute a graph, this equation places restrictions upon that set. Note that the 2D term is absent from eq 11. In alkanes the vertex degree of the $-\text{CH}_2-$ skeletal group is 2, and this vertex may be considered a spacer, providing size for the graph. Its count is not involved in the restricting equation, eq 11. To put it another way, for a graph with a given set of vertices of orders 1, 3, and 4, one may add an arbitrary number of vertices of order 2 without affecting the count of the other vertices. On the other hand, the number of rings has a direct effect on the count of vertices of order 1, 3, and 4.

It is possible to develop a relationship among the edge counts, which also acts as a restricting relation upon the sets of edge counts which are possible for a molecular graph. To obtain this relation, eqs 7–10 are combined with eq 11. The following relation is obtained:

$$ne_{11} + \frac{1}{2}ne_{12} + \frac{1}{3}ne_{13} + \frac{1}{4}ne_{14} - \frac{1}{6}ne_{23} - \frac{1}{4}ne_{24} - \frac{1}{3}ne_{33} - \frac{5}{12}ne_{34} - \frac{1}{2}ne_{44} = 1 - R \quad (12)$$

This dependence relation expresses the basic concept that when certain edges are present, certain others are required to make a graph. For example, if a 23 edge is present, there must be edges to connect with vertices with degrees of "2" and "3", such as 12, 23, 24, or 13, etc. Further, a graph cannot consist of only 22 edges unless there is only one ring ($R = 1$ with no branch points). If there are two rings present, with no terminal branch points (as in naphthalene), then the relation between the 23 and 33 edge counts must be $(1/6)ne_{23} + (1/3)ne_{33} = R - 1$. That is, there must be four times as many 23 edges as 33 edges. This relation holds for such graphs as anthracene, tetracene, and homologues.

It should be noted that in eq 12 the count of "edge spacers", ne_{22} , is missing. Just as in eq 11, 2D , the count of "vertex

spacers" is absent. This restricting relation is not primarily size dependent. The combination of eqs 7–12 summarizes rules concerning the types of skeletal groups and edges which can form the graph of an actual molecule.

Generation of Edge Type Counts. We are now in a position to consider the determination of edge type counts from path count information. Four relations have been developed between path counts and edge type counts: eqs 1, 2, 4, and 6. These path counts can be used to determine the degree counts, as shown in parts 1 and 2.^{1,2} Further, we have shown the relation between degree counts and edge type counts, eqs 7–10, along with two restricting relations, eqs 11 and 12, for a total of 10 equations which are not independent.

We have used a symbolic mathematical manipulation program, POWER MATH,¹¹ to ascertain the degree of independence among these relations on the 10 edge type counts, ne_{ij} . (Since the 11 edge type count only appears in one case, a two-vertex graph such as ethane, its value will not be considered further.) When entered into the mathematical system, it was determined that only 5 of the above 10 relations are independent. We have elected to solve for the 12, 13, 14, 22, and 23 edge type counts in terms of the remaining types, 24, 33, 34, 44. The solution for the edge type counts is as follows:

$$ne_{12} = -3^1D - ne_{24} - ne_{33} - 3ne_{34} - 6ne_{44} - 3A + 2^2p + ^3p - 15R + 15 + 3R_3 \quad (13)$$

$$ne_{13} = 8^1D + 2ne_{24} + ne_{33} + 4ne_{34} + 8ne_{44} + 7A - 6^2p - ^3p + 31R - 31 - 3R_3 \quad (14)$$

$$ne_{14} = -4^1D - ne_{24} - ne_{34} - 2ne_{44} - 4A + 4^2p - 16R + 16 \quad (15)$$

$$ne_{22} = -2^1D + ne_{24} + 2ne_{33} + 4ne_{34} + 7ne_{44} + 2A - ^3p + 2R - 2 - 3R_3 \quad (16)$$

$$ne_{23} = ^1D - 2ne_{24} - 3ne_{33} - 5ne_{34} - 8ne_{44} - A + ^3p - R + 1 + 3R_3 \quad (17)$$

In these equations we have used the symbol A for number of atoms. Because there are fewer equations than edge count variables, it is necessary to generate the possible values for ne_{24} , ne_{33} , ne_{34} , and ne_{44} . These values range from 0 up to a maximum value. The maximum is equal to the number of atoms, A , minus the number of edges assigned to any other edge type. Any edge type corresponding to a terminal vertex cannot be part of the 24, 33, 34, and 44 edge types. Hence it is possible to assign an upper bound to this maximum value for these edge type counts as $A - ^1D$. The actual maximum may be less, but for our present purposes this reckoning will suffice. Under these conditions, in which there are five relations for the nine edge type counts, some of the generated edge sets do not correspond to molecular graphs. For example, an edge type count may be computed to be negative, an unacceptable result for a count. Several tests of the acceptability for graph construction have been included in our generation process.

First, we state the nonnegative requirement of all edge type counts:

$$ne_{ij} > 0 \quad \text{for all } ij, \quad i = 1, 2, 3, 4 \quad \text{and} \quad j = 1, 2, 3, 4 \quad (18)$$

When an edge of the type " kk " is present, then there must be at least two vertices with vertex degree k . This relation is

embodied in the following inequality:

$$\text{If } ne_{kk} > 0, \text{ then } ^kD \geq 2 \quad (19)$$

For an acyclic graph when there is a 22 edge present, there must be edges present with a vertex of degree 2. So the following inequality must hold:

$$\text{if } ne_{22} > 0, \text{ then } ne_{12} + ne_{23} + ne_{24} > 0 \quad \text{for } R = 0 \quad (20)$$

Likewise for the higher order edged types

$$\text{if } ne_{33} > 0, \text{ then } ne_{13} + ne_{23} + ne_{34} > 0 \quad \text{for } R = 0 \quad (21)$$

$$\text{if } ne_{44} > 0, \text{ then } ne_{14} + ne_{24} + ne_{34} > 0 \quad \text{for } R = 0 \quad (22)$$

These restrictions will suffice for the present to generate valid edge type counts for the cases we wish to consider. The relations in eqs 13–17 along with the restricting relations in eqs 18–22 were used to generate the list of edge type counts shown in Table III.

Application of Path 3 Count. In part 1¹ an example was given for the design of acyclic alkanes with molar volume in the range 158–162 mL. The inverse process resulted in a candidate set of graphs which contained 11 members. The best QSAR equation was used to filter out candidates whose predicted molar volume is outside the target range, 158–162 mL. Those results are shown in Figure 1b in part 1. Five candidates were filtered out by that process, leaving six members in the target set, all of whose molar experimental volumes lie in the desired target range.

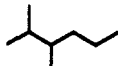
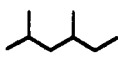
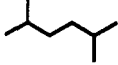
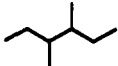
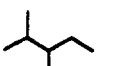
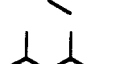


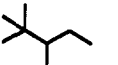
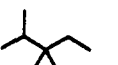
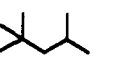
For our current application we will use the path 3 count, 3p , as a filter to examine its usefulness in the filtering step of the design process for the same example as was used in part 1. Table III lists the eleven candidate molecules along with their 1p , 2p , and 3p values as well their edge type counts.

The solution of the QSAR equations given in Figure 1a in part 1 led to values for the χ indexes as follows: $^1\chi$, from 3.498 to 3.616; $^2\chi$, from 3.275 to 3.413. The results for $^3\chi_p$ were not reported in part 1; they are for $^3\chi_p$, 1.718–1.798. These index values correspond to the following path count values: $^1p = 7$; $^2p = 8, 9, 10$; $^3p = 6, 7, 8$. These path counts were obtained in the manner stated in part 1, from the plot of each χ index versus the corresponding path count.¹⁰ In the case of 3p , we obtained from our database of hydrocarbons the list of graphs with $^3\chi_p$ values in the range of 1.718 to 1.798. This list confirmed the 3p values obtained from the plot.

The data given in Table III indicate that candidates 3, 7, 10, and 11 do not have 3p values which correspond to the target values for 3p : 6, 7, 8. These four molecules, for which $6 > ^3p > 8$, are filtered out. The target set obtained in this fashion contains one more member than the target set obtained in part 1. The additional member is candidate 2 in Table III: 2,4-dimethylhexane. Its experimental molar volume is 163.1 mL as reported in part 1.

Examination of the edge type counts in Table III reveals that they also correspond to the same candidate graphs as were indicated by the 3p values. The set of path counts given above do not generate edge type counts which correspond to graphs 3, 7, 10, and 11. If the candidate set had been generated from edge type counts, graphs numbered 3, 7, 10, and 11 would not have been generated.

Table III. Path Counts and Edge Type Counts for the Candidate Set of Molecules Produced from the Inversing Process for Acyclic Alkanes with Molar Volume in the Range 158–162 mL

graph	path counts				edge type counts								
	0p	1p	2p	3p	12	13	14	22	23	24	33	34	44
	8	7	8	7	1	3	0	1	1	0	1	0	0
	8	7	8	6	1	3	0	0	3	0	0	0	0
	8	7	8	5 ^a	<i>b</i>								
	8	7	8	8	2	2	0	0	2	0	0	1	0
	8	7	8	8	2	2	0	0	2	0	0	1	0
	8	7	9	8	0	5	0	0	0	0	0	2	0
	8	7	9	5 ^a	<i>b</i>								
	8	7	9	7	2	0	2	1	0	2	0	0	0
	8	7	10	8	1	1	3	0	1	0	0	1	0
	8	7	10	9 ^a	<i>b</i>								
	8	7	10	5 ^a	<i>b</i>								

^a Outside the target range based on $^3p = 6, 7$, and 8 . See text. ^b No edge set was generated from the target set of path counts, filtering out this candidate. See text.

DISCUSSION

The focus on path, vertex degree, and edge type counts provides a basis for molecular structure information. Graphs of different molecules possess different profiles of these counts. In this sense, these counts constitute, at least in part, a representation of molecular structure. This representation is used in these three papers as a basis for the construction of molecules with desired property values.

General relations have been developed between vertex degree counts, edge type counts, and the path counts. Together with the restricting relations which were also developed, these equations provide a basis for describing molecular structure in considerable detail. Further, the equations relating path counts to edge type counts are the basis for an algorithm which generates edge type counts. These edge type counts may be used in the filtering step as part of the inversing process.

The restricting relations developed in this paper place precise limitations upon the sets of vertices and edges which are valid, that is, those which can be used to construct a graph which represents an actual organic molecule. Further, for a given valid set of vertex degree counts and/or edge type counts, the relations place limitations upon which molecules can actually be constructed. In this manner, the number of actual graphs which can be constructed from a set of connectivity index values is limited so that a combinatoric explosion of possibilities is avoided.

In the inversing process, the sequential use of the path count information from zero order up to order three and higher provides a focusing effect by reducing the number of candidate

graphs as each successive order is used. If the inversing process is attempted only with low level structure information, a very large number of graphs is generated. An example here may be useful to illustrate this point; the numbers used are not drawn from a specific QSAR model but are typical of inverse cases.

Consider an inversing process in which the target number of atoms is 12: $A = 12$. There are 355 acyclic dodecanes, and, in addition, the number of cyclic dodecanes is larger. Even the number of monocyclic graphs is larger than the number of acyclic graphs. However, when we include the additional information from a 1p count, the number of candidate graphs produced is significantly reduced. For example, if $A = 12$ and $^1p = 11$, then only the acyclic graphs remain: $n = 355$. When we supply further information for the count of paths of length 2 the number of candidates is reduced again. Let us take $A = 12$, $^1p = 11$, $^2p = 15, 16$, and 17 ; then, $n = 144$, a reduction to less than half. When we proceed further to specify the path 3 count, for example $^3p = 17, 18$, and 19 , the number is greatly reduced again: $n = 30$. If higher order path count information is available in a particular QSAR case, the number of candidate graphs can be further reduced. For example, if $^4p = 13, 14$, and 15 , then $n = 12$. In this specific example, the number of candidates resulting from the inversing process is reduced from several thousand, when only A is specified, to 30 when 3p is specified and even further for 4p specification.

The example inverse imaging process given in part 1, and elaborated here, is based on the molar volume of hydrocarbons.

In this example, the path 3 count was used as a filter to remove undesired members from the candidate set of graphs. Four of the eleven members were filtered out, leaving seven members in the target set. This target set contains one more member than was obtained in part 1 using the best QSAR equation as the filter. The extra member, no. 2 in Table III, does have a molar volume outside the target range: 163.1 compared to the range of 160–162. This "error" is not considered large, although it is real. In fact, in this case the use of the best QSAR equation leads to better results because no. 2 is filtered out.

The use of the generated edge type counts leads to the same results as use of the path 3 count, as is expected since 3p is included in the algorithm for generated edge type counts. If the candidates had been constructed directly from edge type counts, rather than vertex degree counts, then the candidate set would have contained only seven members. Future work will explore the relative merits of construction from A , 1p , and 2p with subsequent 3p filtering to that of construction directly from edge type counts which includes 3p information from the start.

CONCLUSIONS

In this paper we discuss one aspect of the graph reconstruction problem, namely, a method for obtaining graph primitive information (vertex degrees and edge types) from QSAR equations. The problem of graph reconstruction has been receiving much attention.³⁻⁷ This recent activity indicates the importance attached to the problem. As usually stated (the construction of graphs from the set of vertex degrees and/or edge types), the graph reconstruction problem is considered difficult.¹² The method must generate all possible graphs. Such exhaustive generation has been demonstrated with computer methods.^{13,14} The more challenging aspect of the problem is generation of a nonredundant list of graphs. This problem, known as the graph isomorphism problem, means avoiding generation of graphs which differ only in the numbering of the vertices.¹⁴⁻¹⁸

The development of a canonical indexing method by Faradzhev,⁵ using a backtracking algorithm, has advanced the state of the art significantly. Further, the work of Pospíchal et al.⁶ and Contreras et al.⁷ have contributed to the practical aspects of the problem.

In our work, presented in these first three papers, we have sought to present the foundations necessary for development of graph primitives from QSAR equations based on graph theoretic indexes. These papers include examples with relatively simple structures. Current work is exploring the practical application of the newer algorithms⁵⁻⁷ to more complex structures.

In this third paper in the series on design of molecules from QSAR equations, we have developed relations among path counts, degree counts, and edge counts. This set of equations includes direct relations as well as restricting relations, in addition to inequalities which place limitations upon the generated edge type counts. These relations are general relations for molecular graphs. The set of relations provides insight into the various levels of structure information for molecular graphs and contributes to a representation of molecular structure.

It is further shown that the path 3 count, 3p , can be used in the filtering step in the inverse imaging process which was outlined in part 1. The path counts, from zero up to third order, form the foundation for generating sets of edge type counts which may also be used in the filtering step. Alter-

natively the set of path counts, 0p , 1p , 2p , and 3p , can be used directly to construct candidate graphs via degree counts and edge counts. Such an approach greatly limits the number of candidates generated as compared to use of only atom count and path 1 information. It remains to be seen just how these approaches will work out in future examples. We will explore various avenues to investigate the behavior of several practical strategies and determine which will actually produce the best results.

These first three papers are intended to present the basic elements of the inverting process. The inverse imaging example presented in this series so far is rather elementary, just for acyclic alkanes. Further work will explore practical aspects of inverting more complex data sets. We will also examine more direct methods for obtaining path counts from χ and κ indexes, for obtaining the target vertex sets, and edge sets. In addition, more development is required for explicit inclusion of heteroatoms.

ACKNOWLEDGMENT

We wish to express our appreciation for support of this work from Sterling-Winthrop Pharmaceutical Research Division, Malvern, PA.

REFERENCES AND NOTES

- (1) Kier, L. B.; Hall, L. H.; Frazer, J. W. Design of Molecules from Quantitative Structure-Activity Relationship Models. 1. Information Transfer between Path and Vertex Degree Counts. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 143–147.
- (2) Hall, L. H.; Kier, L. B.; Frazer, J. W. Design of Molecules from Quantitative Structure-Activity Relationship Models. 2. Derivation and Proof of Information Transfer Relating Equations. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 148–152.
- (3) Baskin, I. I.; Gordeeva, E. V.; Devdariani, R. O.; Zefirov, N. S.; Palyulin, V. A.; Stankevitch, M. I. Solving the Inverse Problem of Structure-Property Relations for the Case of Topological Indexes. *Dokl. Akad. Nauk. USSR*, **1989**, *307*, 613–617.
- (4) 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.*, in press.
- (5) Faradzhev, I. Generation of Nonisomorphic Graphs with Given Partition of Vertex Degrees. In *Algorithmic Investigations in Combinatorics* (Russian); Faradzhev, I. A., Ed.; Nauka: Moscow, 1978; pp 11–19.
- (6) Kvasnička, V.; Pospíchal, J. Canonical Indexing and Constructive Enumeration of Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 105.
- (7) Contreras, M. L.; Valdivia, R.; Rozas, R. Exhaustive Generation of Organic Isomers 1. Acyclic Structures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 323–330.
- (8) Wiener, H. Structure Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17–69, 2636–2638.
- (9) Wiener, H. Vapor Pressure-Temperature Relationships among the Branched Paraffin Hydrocarbons. *J. Phys. Chem.* **1948**, *52*, 425–430.
- (10) The computation of molecular connectivity χ indexes, path counts, and degree counts for molecular graphs was carried out using the software package Molconn-X, version 2.0, from Hall Associates Consulting; for information contact author L. H. Hall.
- (11) POWER MATH-II; Charles E. Roth and James H. Davenport, Industrial Computations, Inc.: 40 Washington St., Wellesley, MA 02181.
- (12) Harary, F.; Palmer, E. M. *Graphical Enumeration*; Academic Press: New York, 1973.
- (13) Trinastić, N.; Jericevic, Z.; Knop, J. V.; Muller, W. R.; Szymanski, K. Computer Generation of Isomeric Structures. *Pure Appl. Chem.* **1983**, *55*, 379–390.
- (14) Randić, M.; Brissey, G. M.; Wilkins, C. L. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 52.
- (15) Shelly, C. A.; Munk, M. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 110.
- (16) Carhart, R. E. *J. Chem. Inf. Comput. Sci.* **1987**, *18*, 108.
- (17) Smith, D. H.; Carhart, R. E. *Tetrahedron* **1976**, *32*, 2513.
- (18) Read, R. C.; Corneil, D. G. *J. Graph Theory* **1977**, *1*, 339.