Graph Potentials Method and Its Application for Chemical Information Processing

V. E. GOLENDER,* V. V. DRBOGLAV, and A. B. ROSENBLIT

Institute of Organic Synthesis, Latvian Academy of Sciences, Riga 226006, USSR

Received June 16, 1981

A graph potentials method for the solution of isomorphism and automorphism partitioning problems is proposed. The method is based on an analogy between graphs and electrical networks and serves to compute graph invariants, including vertex potentials and other characteristics, and to provide an efficient solution to the problems. The relationships between potentials and some other graph invariants, including the extended connectivity and the characteristic polynomial, have been established. Some computational results demonstrating sensitivity of the proposed method are presented.

I. INTRODUCTION

Efficient isomorphism detection and topological symmetry perception algorithms are of great importance for chemical information processing. Reduction of the computational complexity of algorithms is usually achieved by use of graph invariants (properties that do not depend on the numbering of vertexes).

Graph invariants can be divided into two groups. Invariants of the first group are vectors with components corresponding to vertexes of a graph and therefore can be used both for the perception of topologically nonequivalent vertexes and for the testing of graph nonisomorphism. Invariants of the second group contain integral information on all vertexes of a graph and can be used as preliminary screens for isomorphism testing but cannot be used for topological symmetry perception.

The first group contains such invariants as extended connectivity, used in Morgan's canonical numbering algorithm[1] and in procedures for the determination of the symmetry group of graphs.[2] Shelly and Munk[3] have described the extension of Morgan's scheme which is based on a vector representation of the extended connectivity. This modification is close to the Corneil and Gotlieb algorithm I.[4] An analysis of the limitations of these schemes for symmetry perception has been given by Carhart.[5]

Randić[6] described the application of the eigenvector associated with the highest eigenvalue (maximal eigenvector) for the perception of topological symmetry.

A number of papers have been devoted to the application of metric invariants, i.e., invariants based on the paths and distances between vertexes of a graph. Thus, Jochum and Gastaiger[7] used a simple metrical invariant—the number of the outermost occupied neighbor sphere (NOON). Randić and Wilkins[8] proposed to describe each atom by the number of paths of different length starting with this atom (atomic path code). Iterative procedures for automorphism partitioning based on distance matrices have been proposed by Weisfeiler and Lehman,[9] Schubert and Ugi,[10] and Uchino.[11] As a single graph invariant is not sufficient for the perception of topological symmetry, Shelly and Munk proposed an automorphism partitioning algorithm using a set of invariants.[12]

The most widespread invariant of the second group is the characteristic polynomial of the adjacency matrix of a graph. The controversy regarding its application for chemical information processing had been the subject of several papers.[13-17] The use of topological indexes as invariants of the second kind has been described.[18] Metrical properties have been used to obtain such invariants as the number of paths of different lengths in a molecule (molecular path code) and the total number of paths in a molecule.[8]

In this paper a family of graph invariants based on the calculation of vertex potentials is proposed. Their properties and relations to other invariants are being investigated both mathematically and experimentally. Some problems concerning their application for chemical information processing are discussed here too.

Just like the application of the characteristic polynomial as a graph invariant was to a considerable degree determined by the quantum chemical model, the calculation of vertex potentials is based on a certain physical model. In this case it is the model of an electrical network. It can be assumed that any graph depicts the topology of a certain electrical network. The flow of the electric current in the network causes certain nodal potentials, the values of these potentials being a function of the topology of the network. Physical considerations have led to the conclusion that topologically nonequivalent vertexes should have for a certain degree different potentials. A more detailed description of the model is given below.

II. THE PHYSICAL MODEL

Graphical representation of electrical networks is widely used in electrical engineering.[19] For the aims of the present discussion it is necessary to reverse the analogy between electrical networks and graphs. It can be suggested that every graph represents a resistive electrical network. Resistors of the network correspond to edges of the graph and the network nodes (junction points of resistors) correspond to graph vertexes. To cause the flow of electrical current in a network one must connect it with current sources. For this purpose an additional node is created in the network. This node is connected by branches containing resistors and current sources with all other nodes of the network. An example illustrating the transition from the styrene graph to the corresponding network is presented in Figure 1.

The flow of electrical current in the network causes certain potential differences on its nodes. Taking the potential of the additional node to be zero, one can calculate the potentials of all other nodes according to the method of nodal analysis from the following system of linear equations:

$$\begin{aligned} g_{11}u_1 - \ldots - g_{1k}u_k - \ldots - g_{1n}u_n &= C_1 \\ &\vdots \\ -g_{k1}u_1 - \ldots + g_{kk}u_k - \ldots - g_{kn}u_n &= C_k \\ &\vdots \\ -g_{n1}u_1 - \ldots - g_{nk}u_k - \ldots + g_{nn}u_n &= C_n \end{aligned}$$

where u_k is the potential of the kth node, g_{ij}, $i \neq j$ is the conductance (quantity reciprocal to resistance) of the resistor connecting the ith and the jth nodes, $g_{kk} = g_{k1} + g_{k2} + \ldots + g_{kn} + g_{k,n+1}$ is the sum of the conductances of all branches connected with the kth node, and C_k is the current value of the current source connected with the kth node.

One can set branch conductances g_{ij}, $i, j = 1, \ldots, n$ equal to the multiplicity of the corresponding graph edge and con-

0095-2338/81/1621-0196$01.25/0 © 1981 American Chemical Society

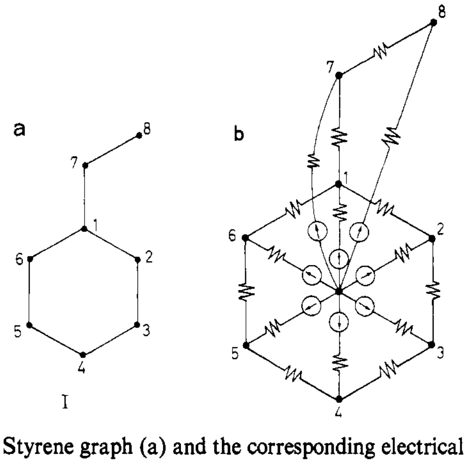


Figure 1. Styrene graph (a) and the corresponding electrical network (b) used by the potentials method. Zigzag lines denote resistors and encircled arrows represent current sources.

Table I. Selected Invariants for the Styrene Graph I

vertex	deg	potentials (first kind)	extended connectivity	maximal eigenvector	NOON	atomic path codes
1	3	2.28	14	.51	3	3, 3, 2, 2
2	2	2.11	10	.39	3	2, 3, 3, 2
3	2	2.05	9	.33	4	2, 2, 3, 3
4	2	2.03	8	.31	5	2, 2, 2, 4
5	2	2.05	9	.33	4	2, 2, 3, 3
6	2	2.11	10	.39	3	2, 3, 3, 2
7	2	1.91	8	.31	4	2, 2, 2, 2
8	1	1.46	4	.14	5	1, 1, 2, 2

ductances of additional branches $g_{i,n+1}$ dependent upon the ith vertex type.

Thus, matrix elements g_{ij} are completely determined by the structure of a graph. At the same time, there is a certain freedom of choice of the right-hand side elements C_k (currents). Our experimental findings and the results of mathematical analysis show that the sensitivity of obtained potentials depends on the procedure used for the selection of the right-hand side elements. In the simplest case, one can choose the value C_k equal to the degree of the kth vertex of a graph.

The basic principles of potentials calculation can be illustrated by the following example. The system of nodal equations for styrene graph I (Figure 1) is given by (all conductances g_{ij} are equal to 1)

$$4u_1 - u_2 - u_6 - u_7 = 3$$

$$-u_1 + 3u_2 - u_3 = 2$$

$$-u_2 + 3u_3 - u_4 = 2$$

$$-u_3 + 3u_4 - u_5 = 2$$

$$-u_4 + 3u_5 - u_6 = 2$$

$$-u_5 + 3u_6 - u_1 = 2$$

$$-u_1 + 3u_7 - u_8 = 2$$

$$-u_7 + 2u_8 = 1$$

By solving the system one can obtain the potential values presented in Table I. These potentials along with truncated atomic path codes differentiate all nonequivalent vertexes of the styrene graph. At the same time, as can be seen from Table I, the extended connectivity values and components of the maximal eigenvector fail to establish the nonequivalence of nodes 4 and 7; invariant NOON has equal values for nonequivalent pairs of vertexes 8 and 4, 1 and 2, 1 and 6, 3 and 7, 5 and 7. Thus, this simple example shows that the potentials method allows to calculate comparatively sensitive graph invariants. The calculations are based on the use of standard programs for the solution of linear simultaneous equations.

A more rigorous treatment of the potentials method is given below.

III. MATHEMATICAL FORMULATION AND SOME PROPERTIES OF THE METHOD

This section contains mathematical formulation of the method and analysis of some of its properties and relations to other methods. Several special concepts and theorems of graph theory and linear algebra are used for this purpose.

Let a graph $\Gamma = (V, E)$, where $V = \{v_i\}$, $i = 1, \ldots, n$ is the set of vertexes and $E = \{(v_{i_1}, v_{i_2}), \ldots, (v_{i_k}, v_{i_l})\}$ is the set of edges, be represented by the adjacency matrix $\mathbf{A} = [a_{ij}]$, whose entries are given by

$$a_{ij} = 1 \text{ if vertexes } v_i \text{ and } v_j \text{ are adjacent}$$
$$0 \text{ otherwise}$$

A matrix $\mathbf{G} = [g_{ij}]$ is a real symmetric $n \times n$ matrix defined as

$$g_{ij} = -a_{ij} \text{ if } i = \mathrm{j}$$
$$\sum_{j=1}^{n} a_{ij} + 1 \text{ if } i = \mathrm{j} \qquad (1)$$

Thus, the off-diagonal entries of the G matrix are equal to −1 for adjacent vertexes and to zero for nonadjacent vertexes. The diagonal entries are equal to the degree of the corresponding vertex plus one: $g_{ii} = \deg(v_i) + 1$. One can rewrite eq 1 using matrix notation as

$$\mathbf{G} = \mathbf{D} + \mathbf{I} - \mathbf{A} \qquad (1')$$

where **D** is the diagonal $n \times n$ matrix with diagonal entries equal to the degree of the corresponding vertex and off-diagonal entries equal to zero, **I** is the identity matrix, and **A** is the adjacency matrix.

Structural formulas of chemical compounds can be represented by graphs with multiple edges and loops, the multiplicity of the edge corresponding to the type of the chemical bond and the multiplicity of the loop corresponding to the type of atom. Entries of the adjacency matrix of such graphs are equal to the multiplicity of the corresponding edge or loop. The **G** matrices in this case are also calculated from eq 1.

Vertex potentials are calculated as the roots of the following linear simultaneous equations

$$\mathbf{GU} = \mathbf{C} \qquad (2)$$

where **U** is the column vector of potentials and **C** is the right-hand-side column vector.

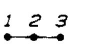
Together with potentials, the **G** matrix determinant and the inverse matrix $\mathbf{H} = \mathbf{G}^{-1}$ are important topological characteristics of a graph.

A. G-Matrix Determinant. The determinant of the **G** matrix is an interesting graph invariant with clear topological meaning. It can be considered as a principal cofactor of the matrix $\mathbf{G}^+$. $\mathbf{G}^+$ is a **G** matrix for the graph formed by the addition to the graph Γ of a new vertex v_{n+1} sharing edges with all vertexes v_i, $i = 1, \ldots, n$. According to the matrix-tree theorem of graph theory,[20] the principal cofactor of the matrix $\mathbf{G}^+$ is equal to the number of spanning trees of graph Γ^+. This relationship is illustrated by Table II.

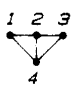
The determinant det(G) can be used as a topological index. This property is illustrated by the case of 6-vertex trees in Figure 2. The det(G) value is maximal for the chain and minimal for the star. Determinants of other graphs are placed within the limits of determinants for the extreme graphs in the sequence corresponding to the degree of their branching.

Table II. Topological Interpretation of the G-matrix Determinant and the Inverse Matrix $\mathbf{G}^{-1}$ Entries

Graph Γ



Graph Γ^+



G-matrix

2	-1	0
-1	3	-1
0	-1	2

G^+matrix

2	-1	0	-1
-1	3	-1	-1
0	-1	-1	-1
-1	-1	-1	-1

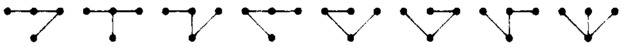
Inverse matrix G^{-1}

.625	.250	.125
.250	.500	.250
.125	.250	.625

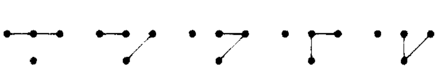
Matrix of cofactors

5	2	1
2	4	2
1	2	5

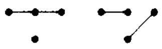
Spanning trees of graph Γ^+ corresponding to the det(G) = 8



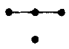
The 2-trees $T_{2_{1,4}}$ corresponding to the cofactor F_{11} = 5



The 2-trees $T_{2_{12,4}}$ corresponding to the cofactor F_{12} = 2



The 2-tree $T_{2_{13,4}}$ corresponding to the cofactor F_{13} = 1



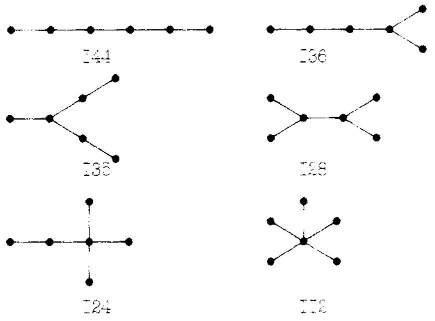
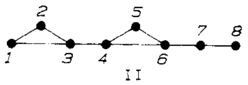


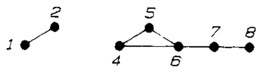
Figure 2. The 6-vertex trees ordered according to the G-matrix determinant values.

Table III. Perception of the Topological Nonequivalence of Vertexes Using the Diagonal Entries of the H Matrix

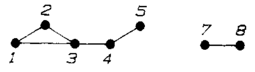
Harary-Palmer graph



Residual graph for vertex 3



Residual graph for vertex 6



	G-matrix								H-matrix							
	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
1	3	-1	-1	0	0	0	0	0	.468	.218	.187	.060	.028	.024	.010	.005
2	-1	3	-1	0	0	0	0	0	.218	.468	.187	.060	.028	.024	.010	.005
3	-1	-1	4	-1	0	0	0	0	.187	.187	.373	.120	.056	.049	.020	.010
4	0	0	-1	4	-1	-1	0	0	.060	.060	.120	.359	.168	.146	.059	.029
5	0	0	0	-1	3	-1	0	0	.028	.028	.056	.168	.446	.171	.068	.034
6	0	0	0	-1	-1	4	-1	0	.024	.024	.049	.146	.171	.366	.146	.073
7	0	0	0	0	0	-1	3	-1	.010	.010	.020	.059	.068	.146	.459	.229
8	0	0	0	0	0	0	-1	2	.005	.005	.010	.029	.034	.073	.229	.615

h_{33} = .373, $h_{6,6}$ = .366

Thus, det(G) can be used either as a second-kind invariant for testing of graph isomorphism or as a branching index for structure–property correlations.

The fact that det(G) is equal to the number of spanning trees in graph Γ^+ indicates that det(G) is always greater than zero. Therefore, the inverse matrix $\mathbf{G}^{-1}$ always exists. Some properties of this matrix are examined below.

B. The Inverse Matrix $\mathbf{G}^{-1}$. According to the general properties of matrices the ijth entry of the inverse matrix $\mathbf{H} = \mathbf{G}^{-1}$ is equal to the cofactor F_{ij} divided by the G-matrix determinant:

$$h_{ij} = F_{ij}/\det(\mathbf{G})$$

The cofactor F_{ij} is the determinant of matrix $\mathbf{G}_{ij}$ which is obtained from matrix $\mathbf{G}$ by deleting the ith row and the jth column.

Note that matrices $\mathbf{G}_{ii}$ do not coincide with G matrices of residual graphs Γ_i formed from Γ by the deletion of the ith vertex with its edges. The difference is caused by the property of diagonal entries $g_{kk} = \deg(v_k) + 1$ corresponding to vertexes connected with the deleted vertex v_i to retain information on edges deleted with this vertex. The Harary–Palmer graph[20] can serve to illustrate the importance of this difference (see Table III). This graph is a counterexample to the conjecture stating that the vertexes of a graph having isomorphic residual subgraphs are similar. In the graph theory, vertexes v_i and v_j are called similar or A equivalent if there is an adjacency preserving permutation (automorphism) mapping v_i into v_j. In chemical informatics such vertexes are usually called topologically equivalent. As one can see from Table III the vertexes v_3 and v_6, being nonequivalent, have isomorphic residual graphs. The nonequivalence of these vertexes is recognized by the inequality of corresponding matrix elements of the **H** matrix, $h_{33} = 0.373$ and $h_{6,6} = 0.365$.

Now we consider the topological meaning of **H**-matrix entries. In view of the fact that det(G) is equal to the number of spanning trees in the graph Γ^+, it seems likely that the values of cofactors F_{ij} are connected with the number of certain subgraphs of that graph. Actually, from the works on the topological analysis of electrical networks[19] it follows that the diagonal cofactors F_{ij} are equal to the number of 2-trees $T_{2_{i,n+1}}$ in the graph Γ^+, and the values of the off-diagonal cofactors F_{ij} are equal to the number of 2-trees $T_{2_{ij,n+1}}$. The 2-tree of a graph is formed from two connected subgraphs without cycles separated between themselves. Both subgraphs cover all vertexes of the graph Γ^+. Vertexes v_i and v_{n+1} of the 2-tree $T_{2_{i,n+1}}$ belong to different parts of that tree. Vertexes v_i and v_j are in one part of the 2-tree $T_{2_{ij,n+1}}$, but vertex v_{n+1} is in the other part of that tree. The 2-trees $T_{2_{i,n+1}}$ can be obtained from spanning trees of graph Γ^+ containing the edge (v_i, v_{n+1}) by deletion of that edge. An example illustrating the relation of **H**-matrix entries to the number of 2-trees is presented in Table II.

This topological interpretation of the **H**-matrix entries indicates that the value of the diagonal **H**-matrix entry is inversely correlated with the degree of the corresponding graph vertex. Actually, every spanning tree of a graph must contain at least one edge connecting each vertex v_i with other vertexes. The 2-trees $T_{2_{i,n+1}}$ are formed from spanning trees containing the edge (v_i, v_{n+1}). It is clear therefore that the smaller number of edges incident to the vertex v_i the greater is the part of spanning trees covered by the 2-trees $T_{2_{i,n+1}}$.

The corollary of the Cayley–Hamilton theorem[21] connecting the inverse matrix with powers of the inverted matrix and coefficients of the characteristic polynomial provides additional insight into the topological meaning of the **H**-matrix entries. This connection is given by

$$\mathbf{G}^{-1} = -1/p_n(\mathbf{G}^{n-1} + p_1\mathbf{G}^{n-2} + \ldots + p_{n-1}\mathbf{I}) \quad (3)$$

where $\mathbf{G}^i$ is the ith power of the **G** matrix and p_i is the ith coefficient of the characteristic polynomial of the **G** matrix.

The entries of the matrix $\mathbf{G}^i$ contain information on the number of walks of length i between all pairs of vertexes. Thus the entries of the inverse matrix $\mathbf{G}^{-1}$ contain in a contracted form information on the number of walks of different length between all pairs of vertexes.

The diagonal entries of the **H** matrix reflect to a certain degree the structural nonequivalence of vertexes. Similarly, the off-diagonal entries of the ith row reflect the topological nonequivalence of other vertexes with respect to the ith vertex. In terms of the group theory the diagonal entries of the **H** matrix contain information about the automorphism group orbits, but the off-diagonal entries of the ith row contain information about the orbits of the stabilizers of the ith vertex. The stabilizer of the ith vertex is a subgroup of the automorphism group and contains permutations that fix that vertex.[20] The off-diagonal entries of the **H** matrix also bear information on the topological nonequivalence of edges. Thus, the **H** matrix can be usefully employed for the reduction of trials and errors in the graph symmetry group determination algorithms.

At the same time, the **H** matrix can be used for the detection of graph isomorphism, because it defines a graph up to the isomorphism. It can be easily shown that if two graphs have equal **H** matrices, then these graphs are isomorphic. Actually, these graphs have equal **G** matrices because $\mathbf{G} = \mathbf{H}^{-1}$ and consequently they are isomorphic. Two rigorous graph isomorphism procedures based on the use of **H** matrices can be proposed:

(1) search for the renumeration of graph vertexes producing equal **H** matrices

(2) comparison of the canonical representations of **H** matrices.

A more detailed description of these procedures is given in section VI. It must be pointed out here that due to the above described properties of **H** matrices these algorithms require less computational resources than analogical algorithms based on the use of the adjacency matrix. Nevertheless, for some graphs the required number of steps can be quite extensive. For the reduction of computational complexity it is expedient to derive graph invariants from the **H** matrix. Vertex potentials represent such kind of invariants. Note that the transition from the **H** matrix to the vertex potentials can result in the loss of significant information on the structure of graphs. This loss of information represents the cost that must be paid for the reduction of the number of steps required for graph analysis.

C. First-Kind Potentials. Vertex potentials are computed from eq 2. For the appropriate selection of the vector **C** the following property of the **H** matrix is important: the sum of its entries belonging to one row or column is equal to 1. This property follows from the same property of the **G** matrix and the general property of matrices—$\mathbf{G}^{-1}\mathbf{G} = \mathbf{I}$. Therefore if all components of the **C** vector are equal, potentials of all vertexes are also equal, and $\mathbf{U} = \mathbf{C}$.

In order to obtain potentials reflecting the topology of a graph one must provide the initial differentiation of vertexes in vector **C**.

For the computation of first-kind potentials the components of the vector **C** are selected to be equal to the degree of the corresponding vertex. An example of the calculation of first-kind potentials for the styrene graph I is presented in Table I. Analysis of the potential values shows the redistribution of potentials among vertexes of the graph. Thus, the low-degree vertexes receive a potential exceeding their degree ($u_8 = 1.46$, $\deg(v_8) = 1$), and the high-degree vertexes receive a potential smaller than their degree ($u_1 = 2.28$, $\deg(v_1) = 3$). The redistribution of potentials takes place according to the following law: the sum of potentials is equal to the sum of vertex degrees. In the general case, the method of potentials has the property

$$\sum_i u_i = \sum_i c_i$$

In spite of the fact that these potentials precisely differentiate nonequivalent vertexes in various kinds of graphs, they fail to do so for the vertexes of regular graphs. Vertexes of these graphs have the same degree, and as a result they obtain equal potentials. Note that vertexes of regular graphs are not differentiated by the extended connectivity[1] including its vector representation[3] too.

In order to increase the sensitivity of the potentials it is necessary to employ properties providing initial differentiation of the regular graph vertexes as components of **C** vectors. Within the framework of the potential method, it appears possible to obtain such initial differentiation without any increase in computational complexity.

D. Second-Kind Potentials. This version of the method is based on the application of the diagonal entries of the **H** matrix for the initial differentiation of vertexes. Actually, as shown in section IIIB, the diagonal entries of the **H** matrix can serve to recognize topologically nonequivalent vertexes. Hence, it is possible to use as components of the **C** vector

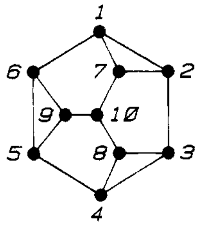
(1) diagonal entries of the **H** matrix

(2) values reciprocal to the diagonal entries of the **H** matrix

(3) ranks of the diagonal entries of the **H**-matrix, sorted in the descending or ascending order.

Each of the mentioned methods has its own advantages and drawbacks, and none of them can guarantee that second-kind potentials can differentiate vertexes that are differentiated by entries of the **H** matrix. Calculations of second-kind potentials presented in this paper have been performed by using values reciprocal to the diagonal entries of the **H** matrix as components of the **C** vector.

The regular 10-vertex cubic graph III proposed[5] as a counterexample to the Shelley–Munk method[3] can be used to illustrate the sensitivity of second-kind potentials.



III

The diagonal entries of the **H** matrix recognize two classes of the graph vertexes

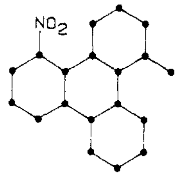
$$h_{11} = h_{22} = h_{33} = h_{44} = h_{55} = h_{66} = h_{77} = h_{88} = 0.344,\ h_{1010} = 0.333$$

and second-kind potentials recognize all three groups of topologically nonequivalent vertexes.

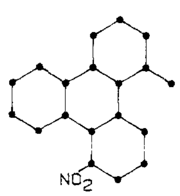
$$u_{11} = u_{22} = u_{33} = u_{44} = u_{55} = u_{66} = 2.910;\ u_{77} = u_{88} = u_{99} = 2.916,\ u_{1010} = 2.93$$

Another illustrative example is presented by the three Herndon[16] graphs IV, V, and VI.

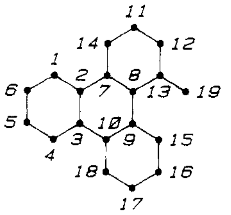
The graphs IV and V are obtained from the graph VI by substitution in vertexes 1 and 18, respectively. These vertexes being topologically nonequivalent are cospectral and, as a consequence, graphs IV and V are cospectral.[16,17] Second-kind potentials for vertexes 1 and 18 are different, $u_1 = 2.46966$,



IV



V



VI

u_{18} = 2.46957. The difference is reliable for computations with the single precision arithmetic. Graphs IV and V have different sets of potentials (see Table IV). These potentials were calculated by using an additional weight to the diagonal entry of the **G** matrix corresponding to the NO_2 group equal to 5 $- g_{2020}$ = 7. The **G** matrix determinants for these graphs are equal, and their common value is 5101296224.

A more detailed account on the investigation of second-kind potentials is presented in section V. In the present section it is necessary to point out that in some classes of graphs second-kind potentials possess low sensitivity. Thus, all nodes of strongly regular graphs have equal potentials. A regular graph is called strongly regular if every pair of connected vertexes has n_1 common neighbors and every pair of disconnected vertexes has n_2 common neighbors.[22] Among chemical graphs there are no strongly regular graphs with topologically nonequivalent vertexes, but this class of graphs represents a good test set for graph analysis algorithms. Graph VII[12] (see Figure 3) represents another counterexample to second-kind potentials method. All the nodes of this graph are topologically nonequivalent, but second-kind potentials along with the invariants described by Shelley and Munk[12] recognize only three classes of vertexes. Third-kind potentials can be used for recognition of nonequivalent vertexes in such difficult cases.

E. Third-Kind Potentials. The algorithm for the calculation of third-kind potentials contains the following steps:

1. Residual graphs Γ_i are formed by deleting every vertex v_i, i = 1, . . ., n, along with its neighbors.
2. Determinants D_i = det($\mathbf{G}_i$) and second-kind potentials are calculated for all residual graphs Γ_i.
3. The residual graphs Γ_i are sorted according to the value of D_i. Graphs with equal determinants are sorted according to the values of potential vectors. Ranks presenting the position of each graph in the sorted lists are determined and placed into the array **C**.
4. Third-kind potentials are calculated—$\mathbf{U} = \mathbf{G}^{-1}\mathbf{C}$.

In the case of graphs with a small vertex degree it is expedient sometimes to use neighbor graphs[22] instead of residual graphs. The neighbor graph for a vertex contains that vertex, its immediate neighbors, and the edges connecting them. However, the use of neighbor graphs instead of residual graphs can change the sensitivity of the method. For instance, for graph VII the neighbor graphs recognize only three classes of nonequivalent nodes. At the same time third-kind potentials based on the residual graphs permit to establish nonequivalence of all vertexes.

F. Computer Implementation and Computational Complexity. The computer implementation of the method for calculating vertex potentials is extremely simple and is based on the use of standard subroutines for the matrix inversion and sorting. Many modern computers have microprograms and array processors for fast implementation of these common routines.

Experiments show that in the vast majority of cases the use of the common Gauss elimination procedure with the single precision arithmetics provide the necessary accuracy of calculations. But this method is not the best for the **G**-matrix inversion. It can be shown that this matrix is positively defined. Therefore for its inversion, the Cholesky's square root method can be used. This method has excellent properties of stability and economy.[23] Its application reduces the number of operations required for matrix inversion by twofold as compared with the Gauss elimination method.

Table IV. Second-Kind Potentials for Graphs IV and V

	graph IV		graph V	
rank	potential	vertex	potential	vertex
1	1.2992	20	1.2992	20
2	2.1118	19	2.1119	19
3	2.3567	17	2.3564	6
4	2.3574	16	2.3566	5
5	2.3649	6	2.3654	17
6	2.3660	5	2.3676	16
7	2.3792	11	2.3785	11
8	2.3959	1	2.3959	18
9	2.4018	12	2.4016	12
10	2.4708	18	2.4707	1
11	2.4735	15	2.4712	4
12	2.4764	4	2.4806	15
13	2.4843	14	2.4824	14
14	2.5253	13	2.2554	13
15	2.7406	10	2.7407	2
16	2.7437	18	2.7415	3
17	2.7475	2	2.7468	10
18	2.7477	9	2.7506	7
19	2.7547	7	2.7540	9
20	3.7725	8	2.7731	8

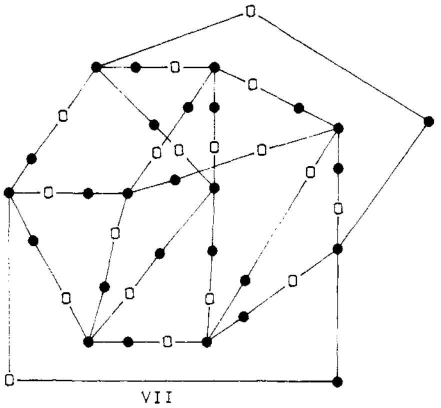

VII

Figure 3. Graph proposed as a counterexample to the automorphism partitioning algorithm.[12] Second-kind potentials recognize three classes of nonequivalent vertexes and third-kind potentials serve to establish nonequivalence of all vertexes.

The matrix inversion procedures require that the number of steps be approximately proportional to the third power of the matrix size—n^3. Thus, the computational (time) complexity[24] of the algorithms for the computation of first-kind and second-kind potentials is $O(n^3)$. Computation of third-kind potentials requires, in the worst case, n times more operations; hence its computational complexity is $O(n^4)$.

Graph potentials presented in this paper have been calculated by using a Wang 2200 VP and a HP-1000 minicomputer. The Wang programs are written in Basic, and they employ microcoded instructions for the matrix inversion and sorting. The HP-1000 program is in Fortran. Computation of second-kind potentials on a HP-1000 computer requires 0.01 s for a 10-vertex graph, 0.05 s for a 20-vertex graph, 1 s for a 30-vertex graph, and 3 s for a 40-vertex graph.

IV. RELATION OF POTENTIALS TO SOME OTHER GRAPH INVARIANTS

A. Relation to the Extended Connectivity. Morgan's algorithm for the calculation of the extended connectivity (EC)[1,25] consists of the following steps:

1. Let the initial value of the EC for each vertex be equal to the degree of that vertex—s_i^1 = deg(v_i). Compute the number of different values of the EC ranks.

2. Calculate the new EC values for each vertex as the sum of the previous EC values of its neighbors

$$s_i^{k+1} = \sum_j s_j^k$$

where the summation is over all vertexes v_j connected with vertex v_i.

3. Calculate the new number of EC ranks. If it is greater than the previous one go to step 2, otherwise go to step 4.

4. Choose values s_i^k, $i = 1, \ldots, n$, corresponding to the iteration k with the maximal number of ranks as the final values of the EC.

It can be shown that Morgan's algorithm performs iterative solution of linear simultaneous equations similar to those of the potential method (eq 2), namely, of the system

$$\mathbf{MS} = 0 \qquad (4)$$

where $\mathbf{S}$ is the column matrix of the EC values and the matrix $\mathbf{M} = [m_{ij}]$ is given by

$$m_{ij} = \begin{cases} -1 & \text{if } v_i \text{ is connected with } v_j \\ 0 & \text{if } v_i \text{ is not connected with } v_j \\ 1 & \text{if } i = j \end{cases}$$

in the matrix notation $\mathbf{M} = \mathbf{I} - \mathbf{A}$.

Actually, the recurrent procedure for the solution of the system (eq 4) by the method of iterations[21] is given by

$$s_i^{k+1} = \sum_{j \neq i} m_{ij} s_j^k$$

Let the starting values of $s_i - s_i^0$ be equal to 1 for all i. Then the first approximation of $s_i = s_i^1$ is equal to the degree of vertex v_i. The following iterations will give the s_i^k values coinciding with that obtained by Morgan's algorithm. It can be shown that the EC values do not converge to any final value. Therefore Morgan's algorithm contains special exit criteria (step 3). The fact that Morgan's algorithm is based on the method of iterations explains the nature of EC values oscillations discovered by Randić.[6] Such oscillations are typical for the method of iterations.

First-kind and second-kind potentials are closely related to the extended connectivity. All of them are calculated from similar systems, the difference being in the diagonal entries of $\mathbf{G}$ and $\mathbf{M}$ matrices and in the right-hand side vectors. It appears quite plausible that Morgan's algorithm fails to provide better partitioning of vertexes than first-kind potentials. At the same time there exist a number of graphs for which first and especially second-kind potentials are more sensitive invariants than the EC values.

B. Relation to the Characteristic Polynomial. The characteristic polynomial of the adjacency matrix of a graph det($\mathbf{A} - \lambda\mathbf{I}$) is frequently employed in chemical informatics and graph theory. Eigenvalues λ_i, $i = 1, \ldots, n$, are calculated as the roots of the so-called secular equation

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0$$

Eigenvector $\mathbf{X}$ corresponding to the eigenvalue λ_i is calculated from the system

$$(\mathbf{A} - \lambda_i\mathbf{I})\mathbf{X} = 0 \qquad (7)$$

Equation 7 possesses considerable similarity with eq 2 of the potentials method and especially with eq 4 of Morgan's algorithm. Note that if $\lambda = 1$ eq 4 and 7 coincide up to the sign preceding the $\mathbf{M}$ matrix. This similarity explains the coincidence in vertex numberings obtained by Morgan's algorithm and the maximal eigenvector method.[6]

It is a plausible conjecture that second-kind potentials are more sensitive graph invariants than the characteristic polynomial or the maximal eigenvector. This conjecture is true for several graph classes. For a number of graphs isospectral vertexes are correctly differentiated by second-kind or even first-kind potentials. Such cases include graphs I–IV presented earlier. The conjecture is also true for regular graphs. For such graphs maximal eigenvector is equal to the degree of a graph, and the corresponding eigenvector contains all components equal to one.[26] Thus, the maximal eigenvector, like the EC and first-kind potentials, do not recognize nonequivalent vertexes of regular graphs. At the same time other eigenvectors differentiate to a certain extent vertexes of the regular graphs.

A number of isospectral graphs have been found to have different second-kind potentials, including the already presented graphs IV and V; other examples are mentioned in section VI.

Besides the adjacency matrix characteristic polynomial, characteristic polynomials of other matrices are also known from the mathematical literature. Thus, Kelmans[27] proposed the $\mathbf{L}$-matrix characteristic polynomial. Diagonal entries of the $\mathbf{L}$ matrix l_{ii} are equal to the degree of the ith vertex, and the off-diagonal entries l_{ij} are equal to the negative value of the multiplicity of the edge connecting vertexes v_i and v_j. The $\mathbf{L}$-matrix polynomial is equivalent to the adjacency matrix characteristic polynomial for regular graphs, i.e., they have the same isospectral graphs.[27] For nonregular graphs there are cases differentiated by both polynomials, by one or neither of them.

The relation between the $\mathbf{L}$-matrix characteristic polynomial

$$\det(\mathbf{L} - \lambda\mathbf{I}) = \sum_{i=0}^{n} (-1)^{n-i} p_i^{n-i}$$

and the $\mathbf{G}$-matrix determinant is given by

$$\det(\mathbf{G}) = \sum_{i=0}^{n} p_i$$

Thus, the determinant of the $\mathbf{G}$ matrix is equal to the sum of the $\mathbf{L}$-matrix characteristic polynomial coefficients. This proposition follows from the theorem on the sum of two matrices.

The computational complexity of the eigenvalue problem is $O(n^3)$.

C. Relation to the Del-Re Method. From the quantum chemist's point of view eq 2 of the method of potentials looks quite similar to that of the Del-Re method for computing atomic charges[28]

$$\mathbf{R}\Delta = \Delta_0$$

where matrix $\mathbf{R} = [r_{ij}]$ is given by

$$r_{ij} = \begin{cases} -p_{ij} & \text{if atoms } i \text{ and } j \text{ are bonded} \\ 0 & \text{if atom } i \text{ and } j \text{ are not bonded} \\ 1 & \text{if } i = j \end{cases}$$

p_{ij} is an empirical parameter depending upon types of atoms i and j, Δ_0 is the column vector with the ith component being dependent on the ith atom kind, and Δ is the column matrix of unknowns used for the calculation of bond and atomic charges.

Thus, the Del-Re method is very similar to the method of the first-kind potentials. Several topological indexes based on the Del-Re method have been proposed by Cammarata.[29]

D. Relation to the Weisfeiler–Lehman method. Weisfeiler and Lehman[9,22] have developed one of the most sensitive approaches to the differentiation of topologically nonequivalent vertexes. The method is based on the utilization of information on paths and antipaths in a graph. This information is obtained by symbolic exponentiation of the modified adjacency matrix. According to eq 3, part of this information is reflected in the inverse matrix $\mathbf{G}^{-1}$. As a result, the sensitivity of the method of potentials is close to that of the Weisfeiler–Lehman method. Thus, for the differentiation of strongly regular graph vertexes it is necessary to apply the modification of the Weisfeiler–Lehman method analogical to that employed in third-kind potentials method.

The computational complexity of the basic version of the Weisfeiler–Lehman method is $O(n^4)$. Computer implementation of the algorithm requires special routines for symbolic manipulations; therefore the computer program for the Weisfeiler–Lehman method is considerably more complex than that for the method of potentials.

The recently proposed Uchino algorithm[11] is quite similar to the Weisfeiler–Lehman algorithm.

V. SOME EXPERIMENTAL RESULTS

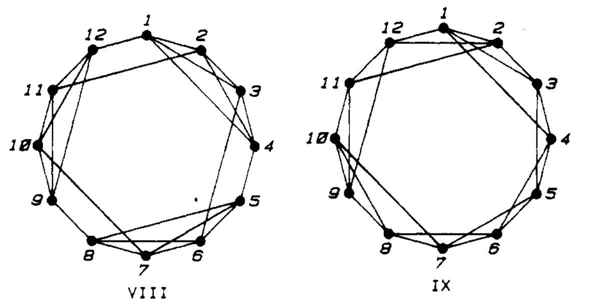
The sensitivity of the method has been investigated on a number of graph families. First of all, potentials were computed for graphs serving as counterexamples to other invariants. Then the method was investigated on several classes of regular and strongly regular graphs.

Second-kind potentials recognize topological nonequivalence of vertexes in many graphs that serve as counterexamples to Morgan's algorithm, maximal eigenvector, and other algorithms. Graphs I–VI presented earlier are examples of such graphs. Graph VII is an exception. Its vertexes are not differentiated to a full extent by second-kind potentials, but third-kind potentials solve this difficult problem.

Systematic investigation of the sensitivity of second-kind potentials was conducted on several families of regular graphs, including 5 8-vertex, 19 10-vertex, and 85 12-vertex graphs with degree 3 and 16 9-vertex and 59 10-vertex graphs with degree 4 from the compilation.[30] The sets of equipotential vertexes for all these graphs coincide with the orbits of the automorphism groups.[30] Furthermore, all nonisomorphic graphs from these families have different sets of potentials. Note that for 10-vertex regular graphs with degree 4 there are two pairs of nonisomorphic graphs with equal characteristic polynomials. These are graphs 11, 21 and 12, 22 from the above mentioned list.[30] The **G**-matrix determinants have equal values only for these pairs out of all 10-vertex graphs with degree 4—$\det(\mathbf{G}_{11}) = \det(\mathbf{G}_{21}) = 2\,695\,680$ and $\det(\mathbf{G}_{12}) = \det(\mathbf{G}_{22}) = 2\,736\,324$. Second-kind potentials for these graphs are different. The situation was analogous for 509 14-vertex cubic graphs from the compilation.[31] Three pairs of graphs—225 and 226, 336 and 337, and 384 and 385—are isospectral. The determinant of the **G** matrix has equal values only for these graphs, but second-kind potentials for them are different.

Additional families included regular graphs with the transitive automorphism group (all vertexes in such graphs are A equivalent). This set included 2 11-vertex, 10 12-vertex, 3 13-vertex, and 5 14-vertex regular graphs with degree 4 and 12 12-vertex graphs with degree 5.[30] A pair of nonisomorphic isopotential graphs VIII and IX has been found. This pair of graphs together with the determinants, second-kind potentials, and the first rows of the inverse matrix $\mathbf{G}^{-1}$ is presented in Table V. It follows from Table V that inverse matrices have nonequal entries. Thus, **H** matrices bear information necessary for the recognition of graphs VIII and IX nonisomorphism, but the employed version of second-kind potentials fails to extract it.

Table V. Isopotential Graphs with Different H-Matrix Entries



det(G) = 5072128, u_i = 3.52713, i = 1,...,12

The first rows of the inverse matrices

	1	2	3	4	5	6	7	8	9	10	11	12
graph VIII	.284	.112	.110	.110	.044	.044	.033	.033	.044	.044	.057	.086
graph IX	.284	.110	.110	.099	.057	.044	.033	.031	.044	.033	.057	.099

Third-kind potentials were investigated on a family of strongly regular graphs including 2 16-vertex, 15 25-vertex, 10 26-vertex, and 4 28-vertex graphs.[22] Sets of equipotential vertexes coincided with the lists of A-equivalent vertexes.[22] Note that the neighbor graphs of strongly regular graphs produced an additional testing set for investigating sensitivity of second-kind potentials. In all these cases but one, det(**G**) and potentials correctly recognized the number of different neighbor graphs. The exception was represented by two isopotential graphs isomorphic to the earlier mentioned graphs VIII and IX.

VI. APPLICATIONS OF THE METHOD

Owing to the high sensitivity, simplicity of programming, and small computational complexity, second-kind potentials can be employed in different existing graph coding and symmetry perception algorithms as substitutes for the extended connectivity,[1,2] its modifications[3,4] and other invariants.[12] But the more rational way of the application of the potential method is to develop special algorithms, taking into account properties of the method. Such algorithms are briefly described in this section.

A. Chemical Structure Searching. The following procedure can be employed for the chemical structure searching:

1. Determinants det(G) and second-kind potentials are computed for all compounds of a file.

2. Compounds are sorted according to the values of det(**G**). Compounds with equal determinants are further sorted according to the values of the potentials.

3. The search for isomorphic graphs is performed first by searching in the sorted numerical files of determinants and potentials.

4. For graphs with equal determinants and potentials rigorous isomorphism testing is performed. As pointed out in section IIIB, two different algorithms based on the analysis of the **H** matrix can be employed for this purpose.

(i) The **H** matrices of each pair of compared graphs are used for the construction of the compatibility graph. Existence of a clique of the compatibility graph of size n, where n is the number of vertexes in each graph, serves as a criterion of graph isomorphism.[32,33]

(ii) Canonical codes of tested graphs are compared. In the case of their equality, graphs are isomorphic.

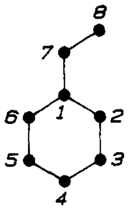
Procedure for the construction of the compatibility graph on the basis of the **H** matrix is an obvious extension of the procedure based on the adjacency matrix, and therefore it is not considered here. On the other hand, the graph coding algorithm based on the method of potentials has several peculiarities described below.

B. Canonical Numbering Algorithm. The proposed canonical numbering algorithm is based on the reduction of **H** matrix to its lexicographically maximal form (LMF). LMF is such a form when the maximal entries have the minimal possible indexes. Matrix indexes are assumed to have the following order: 11, 12, . . ., 1n, 21, . . ., 2n, . . ., nn.

An algorithm has been developed for the reduction of the **H** matrix to the LMF. It is based on the depth-first search

Table VI. Canonical Numbering of the Styrene Graph

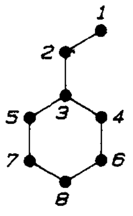
Starting numbering



Starting H-matrix

	1	2	3	4	5	6	7	8
1	.354	.138	.059	.039	.059	.138	.142	.071
2	.138	.436	.169	.071	.044	.061	.055	.028
3	.059	.169	.447	.173	.072	.044	.024	.012
4	.039	.071	.173	.449	.173	.071	.016	.008
5	.059	.044	.072	.173	.447	.169	.024	.012
6	.138	.061	.044	.071	.169	.436	.055	.028
7	.142	.055	.024	.016	.024	.055	.457	.228
8	.071	.028	.012	.008	.012	.028	.228	.614

Canonical numbering



Canonical H-matrix

	1	2	3	4	5	6	7	8
1	.614	.228	.071	.028	.028	.012	.012	.008
2	.228	.457	.142	.055	.055	.024	.024	.016
3	.071	.142	.354	.138	.138	.059	.059	.039
4	.028	.055	.138	.436	.061	.169	.044	.071
5	.028	.055	.138	.061	.436	.044	.169	.071
6	.012	.024	.059	.169	.044	.447	.072	.173
7	.012	.024	.059	.044	.169	.072	.447	.173
8	.008	.016	.039	.071	.071	.173	.173	.449

and employs sorting procedures. During the search, recognition of equivalent graph vertexes is performed. As the **H**-matrix entries are, as a rule, more different than the entries of the adjacency matrix, the proposed algorithm usually requires less search steps than algorithms using the adjacency matrix.[34-36] Since the **H** matrix defines graphs up to isomorphism, it is not necessary to store canonical **H** matrix. One can use canonical numbering of the **H** matrix for the canonicalization of the adjacency matrix.

Table VI presents an illustrative example of the styrene graph canonicalization. One can note that the canonical vertex numbers propagate from the peripheral vertex to the vertex that is maximally remote from it.

A detailed description of the canonicalization procedure and the corresponding Fortran program will be published elsewhere. Other publications will be devoted to the application of the topological indexes derived from the potential method for structure–property studies and to the use of the method for designing a structure–activity data base.[37]

VII. CONCLUSIONS

The potentials method is based on the analogy between graphs and electrical networks. This method relates every n-vertex graph with a $n \times n$ matrix **G**. Diagonal entries of the **G** matrix are equal to the degree of the corresponding vertex plus one, and off-diagonal entries g_{ij} are equal to the negative value of the multiplicity of the edge connecting vertexes v_i and v_j. Several additional graph characteristics are related to the **G** matrix, including its determinant det(**G**), the inverse matrix $\mathbf{H} = \mathbf{G}^{-1}$, and vertex potentials $\mathbf{U} = \mathbf{G}^{-1}\mathbf{C}$.

The **G**-matrix determinant is a graph invariant and can be used for the testing of nonisomorphism and as a topological index. Ample information on the structure of a graph is present in the inverse matrix $\mathbf{G}^{-1}$. It can be used for the design of efficient algorithms for the isomorphism testing, automorphism partitioning, computation of the symmetry group, and canonicalization.

Vertex potentials represent in a contracted form the **H**-matrix entries and can be used for a fast topological nonequivalence and nonisomorphism testing. Sensitivity of potentials depends on the procedure used for the selection of the **C** vector. Second-kind potentials are computed by using components of the **C** vector equal to the reciprocal values of the diagonal entries of the **G** matrix. These potentials can be computed for n^3 steps and have high sensitivity for chemical graphs.

The method of potentials is closely related to Morgan's algorithm, Del-Re, characteristic polynomial, and distance matrix methods. Experiments and the analysis indicate that the potentials method is superior to the above-mentioned methods with regards to its sensitivity and/or computational complexity.

REFERENCES AND NOTES

(1) Morgan, H. L. "The Generation of a Unique Machine Description for Chemical Structures—a Technique Developed at Chemical Abstracts Service". *J. Chem. Doc.* **1965**, *5*, 107–113.

(2) Brown, H. "Molecular Structure Elucidation, III". *SIAM J. Appl. Math.* **1977**, *32*, 534–551.

(3) Shelley, C. A.; Munk, M. E. "Computer Perception of Topological Symmetry". *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 110–113.

(4) Corneil, D. E.; Gotlieb, G. O. "An Efficient Algorithm for Graph Isomorphism". *J. Assoc. Comput. Mach.* **1970**, *17*, 51–64.

(5) Carhart, R. E. "Erroneous Claims Concerning the Perception of Topological Symmetry". *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 108–110.

(6) Randić, M. "On Unique Numbering of Atoms and Unique Codes for Molecular Graphs". *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 105–112.

(7) Jochum, C.; Gastaiger, J. "Canonical Numbering and Constitutional Symmetry". *J. Chem. Inf. Comput. Sci.* **1977**, *1*, 113–117.

(8) Randić, M.; Wilkins, C. L. "Graph Theoretical Approach to Recognition Structural Similarity in Molecules". *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 31–37.

(9) Weisfeiler, B.; Lehman, A. "A Reduction of a Graph to a Canonical Form and an Algebra Arising During this Reduction". *Nauchno-Tekh. Inf., Ser. 2* **1968**, *9*, 12–16 (in Russian).

(10) Shubert, W.; Ugi, S. "Constitutional Symmetry and Unique Descriptors of Molecular Structure". *J. Am. Chem. Soc.* **1978**, *100*, 37–41.

(11) Uchino, M. "Algorithms for Unique and Unambiguous Coding and Symmetry Perception of Molecular Structure Diagram. I. Vector Functions for Automorphism Partitioning". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 116–120.

(12) Shelley, C. A.; Munk, M. E. "An Approach to the Assignment of Canonical Connection Tables and Topological Symmetry Perception". *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 247–250.

(13) Spialter, A. "The Atom Connectivity Matrix (ACM) and its Characteristic Polynomial (ACMCP)". *J. Chem. Doc.* **1964**, *4*, 261–269.

(14) Balaban, A. T.; Harary, F. "The Characteristic Polynomial Does Not Uniquely Determine the Topology of a Molecule". *J. Chem. Doc.* **1971**, *11*, 258–259.

(15) Cudo, J.; Yamasaki, T.; Sasaki, S. "The Characteristic Polynomial Uniquely Represents the Topology of Molecule". *J. Chem. Doc.* **1973**, *13*, 225–227.

(16) Herndon, W. C. "The Characteristic Polynomial Does Not Uniquely Determine Molecular Topology". *J. Chem. Doc.* **1974**, *14*, 150–154.

(17) Herndon, W. C.; Ellzey, M. L. "Isospectral Graphs and Molecules". *Tetrahedron* **1975**, *31*, 99–107.

(18) Evans, L. A.; Lynch, M. F.; Willet, P. "Structural Search Codes for On-Line Compound Registration". *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 146–149.

(19) Seshu, S.; Read, M. B. "Linear Graphs and Electrical Networks"; Addison-Wesley: Reading, MA, 1961.

(20) Harary, F. "Graph Theory"; Addison-Wesley: Reading, MA, 1969.

(21) Streng, G. "Linear Algebra and its Applications"; Academic Press: New York, 1978.

(22) Weisfeiler, B. "On Construction and Identification of Graphs"; Springer Verlag: Berlin, 1976; p 237.

(23) Wilkinson, J. H. "The Algebraic Eigenvalue Problem"; Claredon Press: Oxford, 1965.

(24) Aho, A. V.; Hopcroft, J. E.; Ullman, J. D. "The Design and Analysis of Computer Algorithms"; Addison-Wesley: Reading, MA, 1976.

(25) Wipke, W. T.; Dyott, T. M. "Stereochemically Unique Naming Algorithm". *J. Am. Chem. Soc.* **1974**, *96*, 4834–4842.

(26) Wilson, R. J. "On the Adjacency Matrix of a Graph". In "Combinatorics"; Welsh, D., Ed.; Southend-on-Sea, 1972; pp 295–321.

(27) Dinic, E. A.; Kelmans, A. K.; Zaitsev, M. A. "Nonisomorphic Trees with the Same T-Polynomial". *Inf. Process. Lett.* **1977**, *6*, 73–76.

(28) Del Re, G. "A Simple MO-LCAO Method for Calculation the Charge Distributions in Saturated Organic Molecules". *J. Chem. Soc.* **1958**, 4031–4040.

(29) Cammarata, A. "Molecular Topology and Aqueous Solubility of Aliphatic Alcohols". *J. Pharm. Sci.* **1979**, *68*, 839–842.

(30) Barayev, A. M.; Faradgev, L. A. "Computer Construction and Investigation of Regular and Bipartite Graphs". In "Algorithmic Investigations in Combinatorics"; Nauka: Moscow, 1978 (in Russian).

(31) Bussemaker, F. C.; Cobelic, S.; Cvetkovic, D. M.; Seidel, J. J. "Computer Investigation of Cubic Graphs"; Technological University: Eidhoven, 1976; T.H.-Report 76-WSK-01.

(32) Levi, G. "A Note on the Derivation of Maximal Common Subgraphs of the Two Directed or Undirected Graphs". *Calcolo* **1972**, *9*, 341–352.

(33) Barrow, H. G.; Burstall, R. M. "Subgraph Isomorphism, Matching Relational Structures and Maximal Cliques". *Inf. Process. Lett.* **1976**, *4*, 83–84.

(34) Raznikov, V. V.; Talroze, V. L. "Automatic Generation of Complete Set of Structural Isomers With a Given Molecular Composition and Molecular Weight". *Zh. Struct. Khim.* **1970**, *11*, 357–360 (in Russian).

(35) Arlazarov, V. L.; Zuev, I. I.; Uskov, A. V.; Faradgev, I. A. "An Algorithm for Reduction of Finite Undirected Graphs to a Canonical Form". *Zh. Vych. Math. y Math. Phys.* **1974**, *14*, 737–743.

(36) Randić, M. "On Canonical Numbering of Atoms in a Molecule and Graph Isomorphism". *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 171–180.

(37) Golender, V. E.; Rozenblit, A. B. "Logico-Structural Approach to Computer-Assisted Drug Design". *Med. Chem.* (*Academic*) **1980**, *11*, 300–337.

Comparison of Hierarchical Cluster Analysis Techniques for Automatic Classification of Chemical Structures

GEORGE W. ADAMSON*[†] and DAVID BAWDEN[‡]

Postgraduate School of Librarianship and Information Science, Sheffield University,
Sheffield, S10 2TN, England

Received April 13, 1981

Several hierarchical cluster analysis methods were applied to a set of benzenoid compounds by using structural features automatically derived from Wiswesser Line Notation. Comparisons of the differences in classification, due to choice of clustering algorithm and data standardization technique, were made.

INTRODUCTION

Cluster analysis, and similar techniques of numerical taxonomy, may be applied to descriptors of chemical structure to provide automatic classifications of sets of structures. Such classifications could be of value in information storage and retrieval, structure–property studies, and various areas of chemometrics.

Cluster analyses of sets of chemical substances of substituents have been shown to be of value in studying biological activity spectra[1,2] and selecting appropriate substituents for physicochemical property–biological activity studies.[3,4] These have, however, all used some molecular properties as variables in the clustering procedure.

Fewer examples have been reported of cluster analyses using variables directly representing chemical structure. Sneath described a classification of amino acids, based on both physical property and structural descriptor,[5] while Chu used augmented atom fragments in a structure–property study.[6] Adamson and Bush used various atom-and bond-centered fragments, automatically derived from connection tables, in clustering sets of amino acids[7] and diverse anaesthetic compounds.[8] This type of procedure, with automatic generation of variables from computer-readable representations of structure, could enable automatic classification to become a routine procedure within computerized chemical information systems. However, it is known that widely differing classifications, which may be equally valid representations of the data, are obtained by using different clustering algorithms.[9] The purpose of this work was to study this effect in the context of automatic classification of chemical structures and also to consider the related effect of using raw as against standardized data.

A small data set was constructed for this purpose, consisting of substituted benzene structures. This was chosen so that the effect of the choice of clustering methodology on the final result could be studied more easily than with "real", complicated data sets. Also it allowed classifications on the basis of "conventional chemical" structural features, in this case ring substituents, of the sort readily derived algorithmically from Wiswesser Line Notation (WLN).[10] These classifications could be compared with those obtained by using fixed-size atom- or bond-centered fragments as structural descriptors.[7,8]

EXPERIMENTAL SECTION

The structures were encoded in WLN[11] and descriptors derived algorithmically, as described previously.[10] The descriptors were counts of structural features (i.e., substituent type and relative position) of the kind readily derived from notation representation of structure.

Cluster analyses were carried out by using the CLUSTAN package.[12] Techniques of cluster analysis are fully described elsewhere,[9,13] and only an outline of significant points will be given here.

The methods used here fall into the category of hierarchical, agglomerative clustering techniques. By "hierarchical", it is meant that the classes, or clusters, are themselves classified into larger groups. Repetition of this process at different levels of similarity leads to the representation of the data set by a dendrogram or classification tree. By "agglomerative", it is meant that groups are formed, at each level of similarity, by fusions of existing groupings.

The first stage in such an analysis is the generation of a similarity or dissimilarity (or distance) matrix by computation of a similarity or dissimilarity coefficient between each pair of objects. A variety of such coefficients have been used.[9,13] However, earlier studies of chemical structure classification indicated that the choice of coefficient made little difference to the overall classification produced.[8] For the work reported here, therefore, it was decided to use a single coefficient. The Euclidean distance measure was chosen, because of its ready visualization, computational simplicity, and wide use in other areas.[13]

The Euclidean distance is defined as

$$d_{ij} = \sum_{k=1}^{n} [(X_{ik} - X_{jk})^2]^{1/2}$$

for the distance between objects i and j, where X_{ik} is the value of the kth variable for the ith object, and there are n variables

[†] ICI Pharmaceuticals Division, Alderley Park, Macclesfield, Cheshire, England.
[‡] Pfizer Central Research, Sandwich, Kent, England.

0095-2338/81/1621-0204$01.25/0 © 1981 American Chemical Society