

alternating, block, graft, random, etc. End groups should also be reported. If the polymer has been chemically modified, the type of modification and its extent is needed.

**Physical Properties.** Physical properties such as transition temperatures, swell index, and viscosity give information about the nature of the polymer. The polymer should also be characterized as to whether it is thermoplastic or a network-type polymer.

**Preparation.** Further information that is relevant to the characterization of a polymer is the method used for its preparation. An indication as to whether it was synthesized by bulk, solution, or emulsion polymerization is important. Also whether it is anionic or cationic, etc. Catalysts and solvents used should also be indicated.

**Postprocessing.** Processing of a polymer after its preparation can affect its properties. Information as to such processing also needs to be included.

By enumerating all these different types of information I have tried to make the point that we have to view polymers as significantly different from "simple" organic chemical substances. We must provide for a means of representing polymers and their mixtures in such a way that all this information can be associated together and searched in conjunction with one another.

Such a change will require a major shift in the way we index

substances and in the software that will be needed to retrieve them. However, unless we do so we will continue to struggle with large numbers of unwanted hits and missing important information.

The papers in the Symposium show that of the major vendors, only one is really starting to think about approaching polymer information in a manner somewhat different from "simple" organic substances. Others have not yet made the shift to chemical structures, much less to integrated information. Coding schemes are not made better by increasing the number of codes. A different approach is needed.

The work reported by Molecular Design Limited shows promise as a start toward integrating information in such a way that accurate searches can be made. Unfortunately, such an approach is not being pursued by any of the major vendors. Although individual companies and scientists can use this approach for their own files, the methodology must be applied to the open literature and eventually to back files. The prospect for this is not very promising.

We must shift out of our present thinking and approach polymers as complex collections of information. If we adjust ourselves to this concept, innovative solutions to the problem will emerge and in years to come we will be able to retrieve information with the precision that we can now obtain for "simple" substances.

## Random Walks: Computations and Applications to Chemistry

A. S. SHALABI

Department of Chemistry, Faculty of Sciences, United Arab Emirates University, P.O. Box 17551, Al-Ain, United Arab Emirates

Received December 18, 1989

The concept of random walks (self-returning and self-avoiding walks) in molecular graphs is reconsidered. An algorithm that allows the manual calculation of self-avoiding walks is proposed. Neither self-avoiding walks or self-returning walks provide a totally reliable basis for property prediction, but self-returning walks are generally superior. A program which calculates a similarity matrix and which is dimensioned for up to  $50 \times 200$  entries has been written in Basic for use on IBM PCs, and its use in the determination of similarities of different structures is described. A method is suggested for the calculation of the diagnostic power of a graph theoretical invariant in characterizing structures. Ring closure effects in relation to graph coloring problems are investigated, and applications to total  $\pi$  and  $\omega$ - $\pi$  electron energies and physicochemical properties are considered.

### INTRODUCTION

Characterization of structures is a central problem in chemistry and mathematics.<sup>1-5</sup> As a basis for characterization, some different graph theoretical invariants have been examined. The concept of random walks (RW) developed by Randić<sup>6</sup> as a graph theoretical invariant deserves special attention because it provides a unique characterization for atom environments, is easily applied, and allows the verification by means of reconstruction algorithms. Definition of the concept of similarity, which may apply to a selected property, a dominant property, or the overall features of the system under investigation<sup>7-9</sup> requires the selection of a graph theoretical invariant upon which comparisons among structures can be based.

In this paper, some conceptual definitions for random walks are proposed. These suggest a pencil-and-paper algorithm for the calculation of self-avoiding walks (SAW) in simple molecular graph structures that have real structural correspondences.<sup>10</sup> The potentials in characterization of molecular graphs of self-avoiding walks and self-returning walks (SRW)

are compared. A similarity matrix program has been written to facilitate such comparisons. These comparisons have allowed the measurement of the relative diagnostic power of SAWs and SRWs, and it has been observed that, as predicted by Randić,<sup>6</sup> the SRWs give a superior performance. The calculations of SAWs and SRWs were accompanied by the observation of some subtle relationships between the ring closure process and the coloring system of the graph structures investigated.

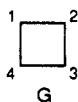
These calculations and similarity comparisons have been compared to the similarities derived from the measured physical and chemical properties of two groups of structurally related chemicals. The first group consists of the isomers of hexane, and the second is a group of alkylbenzenes. This study can be considered an application of one of the Randić similarity measures to two sets of chemical compounds.

### METHODS

There is no general agreement upon the terminology that is used for many graph theoretical concepts, and subsequently,

the definitions used here will be those previously adopted by Randić.<sup>6</sup> Thus a **random walk** (RW) is a sequence of edges that can be continuously traversed, starting at any vertex and ending at any vertex. Repetitive use of the same edge or edges is permitted. A **self-returning walk** (SRW) is a random walk starting and ending at the same vertex. A **self-avoiding walk** (SAW) is a sequence of edges that can be continuously traversed starting at any vertex and ending at any vertex and in which no vertex is visited more than twice.

**Calculation of SRWs.** A four-membered ring can be represented by the graph (G)



and this graph in turn can be represented by the adjacency matrix  $A(G)$

$$A(G) = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$$

which is defined as a square symmetric matrix whose entries can assume a value of either 1 or 0 depending upon whether two vertices in a graph are adjacent or not. An entry of 1 indicates two adjacent vertices, while 0 indicates two nonadjacent vertices. A considerable amount of information can be extracted from an adjacency matrix, because it reflects the topology of a graph. With appropriate modifications, it is possible to calculate random walks; the roots of the corresponding characteristic polynomials (eigenvalues or graph spectra); the count of Kekulé structures and algebraic structures; the sequences of 1s and 0s, i.e., the binary codes; the sum of the coefficients of the inverse polynomial; and so on.

The square of this adjacency matrix,  $A(G)^2$ , is

$$A(G)^2 = \begin{bmatrix} 2 & 0 & 2 & 0 \\ 0 & 2 & 0 & 2 \\ 2 & 0 & 2 & 0 \\ 0 & 2 & 0 & 2 \end{bmatrix}$$

The **trace** of the square of the adjacency matrix,  $\text{Tr}[A(G)^2]$ , is the sum of the diagonal elements of  $A(G)^2$ . Thus

$$\text{Tr} = \sum_{jj}^k [A(G)^2]_{jj} = [A(G)^2]_{11} + [A(G)^2]_{22} + [A(G)^2]_{33} + [A(G)^2]_{44}$$

and so, in this case

$$\text{Tr} = 2 + 2 + 2 + 2 = 8$$

The trace corresponds to the total number of SRWs of length 2, and consequently, in this particular graph, there are eight such SRWs, as can be confirmed by examination of the graph. If the fourth power of  $A(G)$  is considered, its trace, the sum of the diagonal elements of  $A(G)^4$ , will correspond to the number of SRWs of length 4. This process can be repeated for any higher power of the adjacency matrix  $A(G)^m$  where  $m$  is any even number.

**Atom Indexing Problem.** For random walks to be graph theoretical invariants, the indexing of atoms in different ways in different isomers must not alter the number of calculated walks. The indexing is arbitrary (noncanonical) and does not follow rules. Canonical numbering is only necessary in certain special cases where a specific need exists, for example, to reduce the number of steps in the Hessenberg reduction process.<sup>5</sup> Such canonical numbering does not affect the invariance of the topological index, as the name suggests.

The **characteristic polynomial**,  $P(\lambda)$ , of the matrix  $A(G)$  is calculated by expanding the determinant  $|A(G) - \lambda I|$ . In this

way, a polynomial equation  $|A(G) - \lambda I| = 0$  in  $\lambda$  of degree  $p$  is obtained. The  $p$  roots of this equation are called eigenvalues of  $A(G)$ . For

$$A(G) = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$$

the characteristic equation is

$$|A(G) - \lambda I| = \begin{vmatrix} -\lambda & 1 & 0 & 1 \\ 1 & -\lambda & 1 & 0 \\ 0 & 1 & -\lambda & 1 \\ 1 & 0 & 1 & -\lambda \end{vmatrix} = 0$$

and the characteristic polynomial is

$$P(\lambda) = \lambda^4 - 4\lambda^2 = 0$$

The resulting eigenvalues are

$$\lambda_1 = \lambda_2 = 0$$

$$\lambda_3 = 2$$

$$\lambda_4 = -2$$

and the sum of the coefficients of the characteristic polynomial ( $1-4 = -3$ ) is a **graph invariant**.

The spectra of the graph consist of eigenvalues of  $A(G)$ , which is a symmetric matrix, and consequently, all the eigenvalues are real and can be calculated by means of the Jacobi matrix diagonalization subroutine.

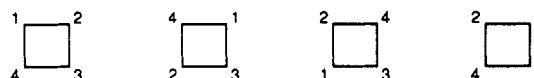
The **inverse polynomial** is calculated from  $A(G)^{-1}$ , the inverse of the adjacency matrix  $A(G)$ :

$$A(G) \cdot A(G)^{-1} = 1$$

The calculation follows that in which the characteristic polynomial is derived from  $A(G)^{-1}$ .

**Comparison of Nonisomeric Structures.** Comparison of different, nonisomeric structures can be accomplished by assignment of a single variable (e.g., the sum of the coefficients of the characteristic polynomial of a graph structure) or a set of variables, such as SAWs and SRWs, to the complete structure, followed by comparison of the variables with a specific property. When a set of variables is used, an equal number of variables must be considered for each pair of nonisomeric structures before the similarity test can be applied.

Considering the four numbering systems below:

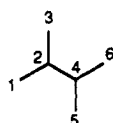


The characteristic polynomial, in every case is

$$P(\lambda) = \lambda^4 - 4\lambda^2$$

and the sums of coefficients in every case are  $-3$ . That is, no matter whether the atom indexing is arbitrary or canonical, the sum of coefficients is the same, and it is therefore a **graph theoretical invariant**.

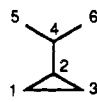
**Manual Construction of Self-Avoiding Walks.** A simple, manual, noncomputational approach to the calculation of the SAW parts of random walks is as follows. The lower triangular part of the distance matrix  $D(G)$  for a molecular graph (G) is constructed. The distance matrix  $D(G)$  is defined as the matrix consisting of the elements  $D_{ij}(G)$  that represent the shortest path from vertex  $i$  to vertex  $j$ . The second step is to count the number ( $x$ ) of repetitions of the number ( $y$ ) in the lower or upper triangular part of the distance matrix  $D(G)$ . As an illustration of this process, consider the structures below: For structure 1 there are five SAWs of length 1, six of length 2, and four of length 3. A similar analysis applies to structure 2. From the previous considerations, the relationship between



structure 1

1 0  
2 1 0  
3 2 1 0  
4 2 1 2 0  
5 3 2 3 1 0  
6 3 2 3 1 2 0

SAWs: 5(1), 6(2), 4(3)



structure 2

1 0  
2 1,2 0  
3 1,2 1,2 0  
4 2,3 1 2,3 0  
5 3,4 2 3,4 1 0  
6 3,4 2 3,4 1 2 0

SAWs: 6(1), 8(2), 6(3), 4(4)

SAWs and  $D(G)$  is confirmed and may be used for the manual construction of SAWs for complex molecular graphs representing real molecules of chemical interest. It should be noted that only half of the SAWs are counted, i.e., each one is counted in one direction only. This operates as a scaling factor, but otherwise does not appear to be significant.

In order to test the similarity between these two structures, it is necessary to consider an equal number of variables for each structure. This may be done either by eliminating the variable 4(4) from the SAWs of structure 2 or by adding a variable 0(4) to the SAWs of structure 1. Thus, given the SAWs derived from both structures:

5(1), 6(2), 4(3) for structure 1

6(1), 8(2), 6(3), 4(4) for structure 2

one arrives at either 5(1), 6(2), 4(3) and 6(1), 8(2), 6(3) or 5(1), 6(2), 4(3), 0(4) and 6(1), 8(2), 6(3), 4(4). Either of these pairs can be used as described in the next section to obtain a measure of the similarity between structures 1 and 2.

**Construction of Self-Returning Walks.** The SRWs of a graph  $G$  are calculated from the sum of the diagonal elements of the adjacency matrix of the corresponding graph  $A(G)$  raised to the appropriate power,  $m$ :

$$n(m) = \text{Tr}[A(G)^m] = \sum_{j=1}^k [A(G)^m]_{jj}$$

where  $n(m)$  is the number,  $n$ , of SRWs of length  $m$ ;  $m$  is an even number; and  $k$  is the order of the square symmetric matrix  $A(G)$ , which is equal to the total number of vertices in the graph ( $G$ ). The value  $m$  should be set to the appropriate SRW length. The larger the number of terms in the computation of the SRW (the value  $m$ ), the better the characterization of the structure will be.

**Computation of the Similarity Matrix.** The similarity matrix (SM) is defined as a square symmetric matrix each of whose off-diagonal elements represent the degree of similarity ( $S$ ) between a pair of structures as

$$(S) = 100 - D(AB)$$

where  $D(AB)$  is the Euclidean distance between the similarity "vectors" or basis sets (e.g., SRWs) of the pair of structures under investigation.  $D(AB)$  is given by

$$D(AB) = [\sum_i (A_i - B_i)^2]^{1/2}$$

The quantity  $S$  in Table II is negative because the corresponding value for  $D(AB)$  is greater than 100. When large numbers of graphs or graph structures are considered, it becomes necessary to use a computer, in spite of the straightforward nature of the calculation of  $(S)$ .

Consider three hypothetical structures, A, B, and C, whose similarity matrix is

$$\begin{vmatrix} 100 & 99 & 60 \\ 99 & 100 & 86 \\ 60 & 86 & 100 \end{vmatrix}$$

The element in row 1, column 1 is 100, signifying that structure A is identical to itself. The element in row 3, column 1 (60)

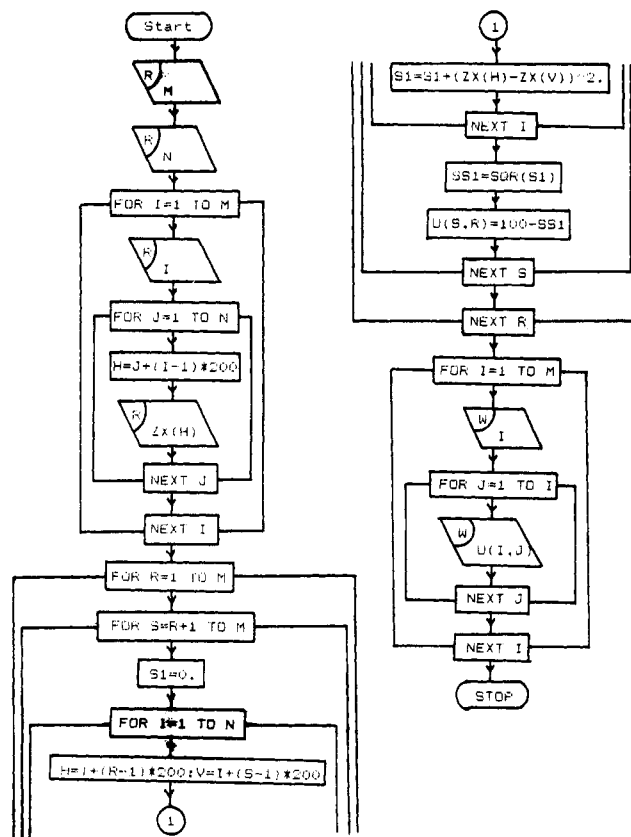
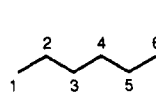


Figure 1. Flow chart of the SM program segment.

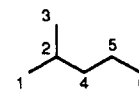
indicates that structures C and A enjoy only 60% similarity. This element (60) is derived from

$$(S) = 100 - D(AB)$$

The distance  $D(AB)$  is calculated as follows. Consider two structures A and B:



structure A



structure B

SRWs: 10, 26, 76, 234, 740, 2372 SRWs: 10, 30, 100, 350, 1250, 4500

For these two graphs, the value of  $D(AB)$  is given by

$$D(AB) = [(10 - 10)^2 + (26 - 30)^2 + (76 - 100)^2 + (234 - 350)^2 + (740 - 1250)^2 + (2372 - 4500)^2]^{1/2} = 2191.468$$

If  $D(AB) > 100$ , then  $S$  will be negative. If  $(S)_1$  is the degree of similarity between A and B and  $(S)_2$  is the degree of similarity between B and C, then in the case where  $D(AB) = 140$ , for example, and  $D(BC) = 160$ :

$$(S)_1 = 100 - (140) = -40$$

$$(S)_2 = 100 - (160) = -60$$

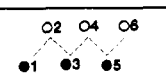
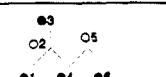

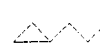



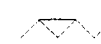

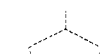


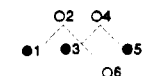



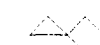
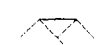
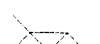

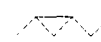


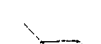



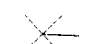
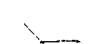


which means that A and B are more similar to one another than are B and C.

The basic program is dimensioned for up to  $50 \times 200$  entries, and its output is designed as the lower triangular part of each similarity matrix. The flow chart shown in Figure 1 and the program segment given in Figure 2 can support an interactive session which will run on an IBM PC or compatible under MS-DOS.

#### APPLICATIONS OF THE SIMILARITY MATRIX METHOD

**1. Diagnostic Powers of SAWs and SRWs.** In Table I, five isomeric forms of a graph ( $G$ ) are given. These were derived

**Table I.** Graphs (G1-G6), Graph Structures (GS), Ring-Closure Vertices (RCV), Self-Returning Walks (SRW), and Self-Avoiding Walks (SAW)

|   |   |  |   |   |   |
|---|---|--|---|---|---|
| <b>Graph 1</b><br>   |   | <b>Graph 2</b><br>   |   | <b>GS 2.4</b><br>  |   |
| SRWs  | 10 26 76 234 740 2372   | SRWs   | 10 30 100 350 1250 4500   | RCVs  | 1,6 ●●  |
| SAWs  | 5 4 3 2 1 0   | SAWs   | 5 5 3 2 0 0   | SRWs  | 12 40 150 592 2412 10054  |
| GS 1.1  |    | GS 1.4   |    | GS 2.1  |     |
| RCVs  | 1,3 ●●  | RCVs   | 1,6 ●○  | RCVs  | 1,3 ●●  |
| SRWs  | 12 40 162 720 3352 16006  | SRWs   | 12 36 132 516 2052 8196   | SRWs  | 12 40 162 720 3352 16006  |
| SAWs  | 6 7 5 4 2 0   | SAWs   | 6 6 6 6 6 0   | SAWs  | 6 7 5 4 2 0   |
| GS 1.2  |    | GS 1.5   |    | GS 2.2  |     |
| RCVs  | 1,4 ●○  | RCVs   | 2,4 ○○  | RCVs  | 1,4 ○○  |
| SRWs  | 12 48 216 1008 4752 22464   | SRWs   | 12 44 198 980 5082 26978  | SRWs  | 12 44 198 980 5082 26978  |
| SAWs  | 6 7 8 4 2 0   | SAWs   | 6 8 7 4 1 0   | SAWs  | 6 7 8 4 1 0   |
| GS 1.3  |    | GS 1.6   |    | GS 2.3  |     |
| RCVs  | 1,5 ●●  | RCVs   | 2,5 ●○  | RCVs  | 1,5 ●○  |
| SRWs  | 12 40 150 592 2412 10054  | SRWs   | 12 52 252 1252 6252 31252   | SRWs  | 12 52 252 1252 6252 31252   |
| SAWs  | 6 7 7 7 2 0   | SAWs   | 6 8 8 6 0 0   | SAWs  | 6 8 8 6 0 0   |
| <b>Graph 3</b><br>   |   | <b>Graph 4</b><br>  |   | <b>GS 3.4</b><br>   |   |
| SRWs  | 10 30 106 390 1450 5406   | SRWs   | 10 34 130 514 2050 8194   | SAWs  | 5 6 4 0 0 0   |
| SAWs  | 5 5 4 1 0   | SAWs   | 5 6 4 0 0 0   | GS 4.1  |  |
| GS 3.1  |   | GS 3.5   |  | GS 4.2  |  |
| RCVs  | 1,3 ●●  | RCVs   | 1,6 ●○  | RCVs  | 1,3 ○○  |
| SRWs  | 10 30 109 422 1680 6801   | SRWs   | 12 48 216 1008 4752 22464   | SRWs  | 12 44 192 900 4372 21692  |
| SAWs  | 6 9 7 2 1 1   | SAWs   | 6 7 8 4 2 0   | SAWs  | 6 8 6 4 0 0   |
| GS 3.2  |  | GS 3.6   |  | GS 4.3  |  |
| RCVs  | 1,4 ●○  | RCVs   | 2,4 ○○  | RCVs  | 1,4 ○○  |
| SRWs  | 12 52 258 1300 6562 33130   | SRWs   | 12 48 234 1248 6972 39846   | SRWs  | 10 34 133 554 2385 10477  |
| SAWs  | 6 8 9 4 1 0   | SAWs   | 6 9 9 3 0 0   | SAWs  | 6 10 8 2 1 0  |
| GS 3.3  |  | GS 5.1   |  | GS 5.2  |    |
| RCVs  | 1,5 ●●  | RCVs   | 2,6 ○○  | RCVs  | 1,3 ●●  |
| SRWs  | 12 40 150 592 2412 10054  | SRWs   | 12 44 198 980 5082 26978  | SRWs  | 10 34 133 554 2385 10477  |
| SAWs  | 6 7 7 7 2 0   | SAWs   | 6 8 7 4 1 0   | SAWs  | 6 10 8 2 1 0  |
| <b>Graph 5</b><br> |   | <b>Graph 6</b><br> |   | <b>GS 5.4</b><br> |   |
| SRWs  | 10 38 160 686 2950 12692  | SRWs   | 10 30 100 350 1250 4500   | RCVs  | 2,4 ○○  |
| SAWs  | 5 7 3 0 0 0   | SAWs   | 5 5 3 2 0 0   | SAWs  | 12 60 336 1956 11652 70572  |
| GS 5.1  |  | GS 5.2   |  | GS 5.3  |    |
| RCVs  | 1,3 ●●  | RCVs   | 1,4 ●○  | RCVs  | 1,5 ●●  |
| SRWs  | 10 34 133 554 2385 10477  | SRWs   | 12 56 288 1508 7872 41216   | SRWs  | 10 30 109 422 1680 6801   |
| SAWs  | 6 10 8 2 1 0  | SAWs   | 6 9 8 4 0 1   | SAWs  | 6 9 7 2 1 1   |

**Table II.** Lower Triangular Parts (LTP) of the Similarity Matrices (SM) for G1–G5 Based on SRWs and SAWs as well as the LTPs of the SMs for Each Graph Together with Its Graph Structures (GS) Based on SRWs<sup>a</sup>

| LTP of SM for Graphs (G) 1–5 Based on the SRW Basis Set                      |           |           |           |           |           |           |           |       |
|--|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-------|
|  | G1        | G2        | G3        | G4        | G5        |           |           |       |
| G1   | 100       |           |           |           |           |           |           |       |
| G2   | -2091.496 | 100       |           |           |           |           |           |       |
| G3   | -828.6938 | -828.6938 | 100       |           |           |           |           |       |
| G4   | -5874.376 | -3683.312 | -2754.63  | 100       |           |           |           |       |
| G5   | -10464    | -8273.495 | -7344.89  | -4490.48  | 100       |           |           |       |
| LTP of SM for Graphs (G) 1–5 Based on the SAW Basis Set                      |           |           |           |           |           |           |           |       |
|  | G1        | G2        | G3        | G4        | G5        |           |           |       |
| G1   | 100       |           |           |           |           |           |           |       |
| G2   | 98.58579  | 100       |           |           |           |           |           |       |
| G3   | 98        | 98.58579  | 100       |           |           |           |           |       |
| G4   | 96.83772  | 97.55051  | 98.58579  | 100       |           |           |           |       |
| G5   | 96.25834  | 97.17157  | 97.55051  | 98.58579  | 100       |           |           |       |
| LTP of SM for G1 and Its Graph Structures (GS1.n) Based on the SRW Basis Set |           |           |           |           |           |           |           |       |
|  | G1        | GS1.1     | GS1.2     | GS1.3     | GS1.4     | GS1.5     | GS1.6     |       |
| G1   | 100       |           |           |           |           |           |           |       |
| GS1.1  | -13790.73 | 100       |           |           |           |           |           |       |
| GS1.2  | -20403.75 | -6514.506 | 100       |           |           |           |           |       |
| GS1.3  | -7770.359 | -5927.141 | -12535.71 | 100       |           |           |           |       |
| GS1.4  | -5876.879 | -7820.141 | -14429.80 | -1794.17  | 100       |           |           |       |
| GS1.5  | -24897.60 | -11010.65 | -4426.71  | -17037.78 | -18930.61 | 100       |           |       |
| GS1.6  | -29319.46 | -15428.74 | -8818.509 | -21453.35 | -23347.29 | -4339.926 | 100       |       |
| LTP of SM for G2 and Its Graph Structures (GS2.n) Based on the SRW Basis Set |           |           |           |           |           |           |           |       |
|  | G2        | GS2.1     | GS2.2     | GS2.3     | GS2.4     | GS2.5     | GS2.6     | GS2.7 |
| G2   | 100       |           |           |           |           |           |           |       |
| GS2.1  | -11602.45 | 100       |           |           |           |           |           |       |
| GS2.2  | -22711.21 | -11010.65 | 100       |           |           |           |           |       |
| GS2.3  | -27130.99 | -15428.74 | -4339.926 | 100       |           |           |           |       |
| GS2.4  | -5579.642 | -5927.141 | -17037.78 | -21453.35 | 100       |           |           |       |
| GS2.5  | -6987.321 | -5515.458 | -16625.51 | -21043.70 | -325.9472 | 100       |           |       |
| GS2.6  | -37226.71 | -25524.31 | -14418.53 | -998.045  | -31550.17 | -31139.72 | 100       |       |
| GS2.7  | -17382.07 | -5679.647 | -5234.073 | -9649.644 | -11705.99 | -11295.07 | -19744.67 | 100   |
| LTP of SM for G3 and Its Graph Structures (GS3.n) Based on the SRW Basis Set |           |           |           |           |           |           |           |       |
|  | G3        | GS3.1     | GS3.2     | GS3.3     | GS3.4     | GS3.5     | GS3.6     |       |
| G3   | 100       |           |           |           |           |           |           |       |
| GS3.1  | 1314.199  | 100       |           |           |           |           |           |       |
| GS3.2  | -28016.46 | -26692.61 | 100       |           |           |           |           |       |
| GS3.3  | -4651.02  | -3238.94  | -23357.14 | 100       |           |           |           |       |
| GS3.4  | -17286    | -15872.54 | -10722.51 | -12535.71 | 100       |           |           |       |
| GS3.5  | -34790.67 | -33376.49 | -6628.748 | -30046.22 | -17424.85 | 100       |           |       |
| GS3.6  | -21783.77 | -20369.60 | -6235.895 | -17037.78 | -4426.171 | -12908.87 | 100       |       |
| LTP of SM for G4 and Its Graph Structures (GS4.n) Based on the SRW Basis Set |           |           |           |           |           |           |           |       |
|  | G4        | GS4.1     | GS4.2     | GS4.3     |           |           |           |       |
| G4   | 100       |           |           |           |           |           |           |       |
| GS4.1  | -13601.85 | 100       |           |           |           |           |           |       |
| GS4.2  | -2207.796 | -11295.07 | 100       |           |           |           |           |       |
| GS4.3  | -25253.44 | -11552.83 | -22947.30 | 100       |           |           |           |       |
| LTP of SM for G5 and Its Graph Structures (GS5.n) Based on the SRW Basis Set |           |           |           |           |           |           |           |       |
|  | G5        | GS5.1     | GS5.2     | GS5.3     | GS5.4     |           |           |       |
| G5   | 100       |           |           |           |           |           |           |       |
| GS5.1  | -2189.895 | 100       |           |           |           |           |           |       |
| GS5.2  | -28857.51 | -31139.85 | 100       |           |           |           |           |       |
| GS5.3  | -5932.342 | -3645.40  | -34884.93 | 100       |           |           |           |       |
| GS5.4  | -58444.55 | -60721.82 | -29501.79 | -64464.60 | 100       |           |           |       |

<sup>a</sup>The SM elements which correspond to the *most* and *least* similar pairs are underlined.

by assuming different connectivities between the same number of vertices. Each vertex in these graphs is colored. The first vertex is colored "black" (●), the second "white" (○), and so on. A graph in which no two neighboring nodes (adjacent vertices) have the same color is termed "bipartite"; otherwise it is non-bipartite. Under each of the graphs 1–5, two lines of entries are given. The first of these designates the SRWs and the second the SAWs, both calculated as described above. For each of the graph structures considered, the ring closure vertices (the RCVs, the pair of vertices involved in a ring closure) as well as the color of each particular pair are shown. The SRWs and SAWs for each graph structure are shown in the second and third rows, respectively.

The lower triangular parts of the similarity matrices from the graphs 1–5 based on SRWs and SAWs were calculated, as described above and are given in Table II, together with the lower triangular parts of the similarity matrices for each graph (G) and the graph structures (GS) based on SRWs. In each set, the extremes (-828.6938 and -10464 in the first set) are underlined and represent the smallest and largest negative values, respectively. The value of -828.6938 means that the two graphs G3 and G2 are the most similar, and likewise, the -10464 means the G5 and G1 are the least similar.

In a model application of the similarity matrix method, the most similar pair of graphs 1–5 are G3 and G2, while the most dissimilar pair are G5 and G1 (based on the basis set of SRWs

```

10 DIM ZX(10000),U(50,50)
11 PRINT "SIMILARITY MATRIX PROGRAM (SM)":PRINT
12 PRINT "note:"
13 PRINT
14 PRINT "The max no. of sets is 50 and the max. no. of elements is 200"
15 PRINT:PRINT
20 INPUT "enter no. of sets " :M
30 INPUT "enter no. of elements in each set " :N
40 FOR I=1 TO M
44 PRINT
45 PRINT "enter elements of set " :I
46 PRINT
50 FOR J=1 TO N
59 H=J*(I-1)*200
60 PRINT "enter element no. "J" : " :INPUT ZX(H)
70 NEXT J
80 NEXT I
90 FOR R=1 TO M
91 FOR S=R+1 TO M
100 S1=0
101 FOR I=1 TO N
102 H=I*(R-1)*200;V=I*(S-1)*200
110 S1=S1+(ZX(H)-ZX(V))^2
115 NEXT I
120 SS1=SQR(S1)
130 U(S,R)=100-SS1
140 NEXT S
150 NEXT R
151 PRINT
152 PRINT
156 PRINT
160 FOR I=1 TO M
165 PRINT I " "
170 FOR J=1 TO I
180 PRINT U(I,J):
190 NEXT J
200 PRINT
210 NEXT I
OK

```

Figure 2. The SM program segment.

given in Table II). The most similar and least similar pairs of graphs (G) 1–5 and their graph structures (GS) are given below:

| most similar pairs          | least similar pairs |
|-----------------------------|---------------------|
| GS 1.4, GS 1.3              | GS 1.6, G 1         |
| GS 2.5, GS 2.4              | GS 2.6, G 2         |
| GS 3.1, G 3                 | GS 3.5, G 3         |
| consequently<br>GS 5.1, G 5 | GS 5.4, GS 5.3      |

The degrees of similarity can easily be calculated for such a small collection of graphs and graph structures.

Turning to the diagnostic powers of SRWs and SAWs, inspection of the top of Table II shows that the similarity matrix elements that are based on the basis set of SRWs are completely different, but this is not true for those based on the basis set of SAWs. Those elements that are based on the basis set of SAWs are grouped in descending order, as shown below:

| graph pair | degree of similarity |
|------------|----------------------|
| G2, G1     | 98.58579             |
| G3, G2     |                      |
| G4, G3     |                      |
| G5, G4     |                      |
| G3, G1     | 98.00000             |
| G4, G2     | 97.55051             |
| G5, G3     |                      |
| G5, G2     | 97.17157             |
| G4, G1     | 96.83772             |
| G5, G1     | 96.25834             |

This shows that there are four pairs of graphs (G2,G1; G3,G2; G4,G3; and G5,G4) which have the same degree of similarity ( $S = 98.58579$ ) and two pairs (G4,G2; and G5,G3) with the same degree of similarity ( $S = 97.55051$ ). For the elements that are based on the SRW basis set, no identical degrees of similarity are found, i.e., each graph is dissimilar to its partner—this is the case under investigation with graphs 1–5. It is concluded from this that the diagnostic power of the SRW basis set is superior to that of the SAW basis set.

**2. Ring Closure Effects and Graphic Coloring.** A bipartite graph is defined as a graph with no adjacent vertices having the same color. Using this definition, graphs 1–5 in Table I are bipartite. Inspection of Table I shows that ring closure between two differently colored vertices leads to a larger increase in the SRWs than does a ring closure between two similarly colored vertices. This can be seen by comparing the SRWs of graph 1 (G1) with those of graph structures GS1.1 and GS1.2; the ring closure vertices (RCVs 1 and 3) of GS1.1

Table III. Total  $\pi$ -Energy ( $E_A$ ), Total  $\omega$ - $\pi$ -Energy ( $E_B$ ) at  $\omega = 1.0$ ,  $E_B^*$  at  $\omega = 1.0$ , and SRWs for Graphs 1–5

|  |   |
|--|---|
| <p>Graph 1.</p> <p> <math>E_A = 6.988 \beta</math><br/> <math>E_B = 6.988 \beta</math><br/> <math>E_B^* = 6.988 \beta</math> </p> <p>SRWs: 10,26,76,234,740,2372</p>   | <p>Graph 4.</p> <p> <math>E_A = 6.000 \beta</math><br/> <math>E_B = 7.123 \beta</math><br/> <math>E_B^* = 8.228 \beta</math> </p> <p>SRWs: 10,34,130,514,2050,8194</p>  |
| <p>Graph 2.</p> <p> <math>E_A = 6.155 \beta</math><br/> <math>E_B = 6.888 \beta</math><br/> <math>E_B^* = 7.657 \beta</math> </p> <p>SRWs: 10,30,100,330,1250,4500</p> | <p>Graph 5.</p> <p> <math>E_A = 5.819 \beta</math><br/> <math>E_B = 6.611 \beta</math><br/> <math>E_B^* = 7.451 \beta</math> </p> <p>SRWs: 10,38,160,686,2950,12692</p> |
| <p>Graph 3.</p> <p> <math>E_A = 6.899 \beta</math><br/> <math>E_B = 6.899 \beta</math><br/> <math>E_B^* = 6.899 \beta</math> </p> <p>SRWs: 10,30,106,390,1450,5406</p> |   |

Table IV. Lower Triangular Parts (LTP) of Similarity Matrices (SM) Based on Total  $\pi$ -Energies and  $\omega$ - $\pi$ -Energies for Graphs 1–5<sup>a</sup>

|  | 1              | 2             | 3      | 4      | 5   |
|--|----------------|---------------|--------|--------|-----|
| LTP of SM Based on Total $\pi$ -Electron Energy                      |                |               |        |        |     |
| 1  | 100            |               |        |        |     |
| 2  | 99.167         | 100           |        |        |     |
| 3  | <u>99.9911</u> | 99.25599      | 100    |        |     |
| 4  | 99.012         | 99.845        | 99.101 | 100    |     |
| 5  | 98.831         | 99.664        | 98.92  | 99.814 | 100 |
| LTP of SM Based on Total $\omega$ - $\pi$ -Energy ( $\omega = 1.0$ ) |                |               |        |        |     |
| 1  | 100            |               |        |        |     |
| 2  | 99.9           | 100           |        |        |     |
| 3  | 99.911         | <u>99.989</u> | 100    |        |     |
| 4  | 99.865         | <u>99.765</u> | 99.776 | 100    |     |
| 5  | 99.623         | 99.723        | 99.712 | 99.488 | 100 |
| LTP of SM Based on Total $\omega$ - $\pi$ -Energy ( $\omega = 1.4$ ) |                |               |        |        |     |
| 1  | 100            |               |        |        |     |
| 2  | 99.331         | 100           |        |        |     |
| 3  | <u>99.911</u>  | 99.24199      | 100    |        |     |
| 4  | 98.76          |               | 99.429 | 98.671 | 100 |
| 5  | 99.537         | 99.794        | 99.448 | 99.223 | 100 |

<sup>a</sup>Underlined elements of the similarity matrices identify the most similar pairs of structures.

have the same color (OO), while those of GS1.2 have different colors (O●).

However, exceptions are noted for SRWs with lengths  $m = 2$  and  $m = 4$  as well as for SAWs with lengths  $m = 1$  and  $m = 2$ . Clearly, the trend applies to the graph structures, and these exceptions indicate that only SRWs whose length exceeds 4 can discriminate between structures obtained by ring closure. The order was also distributed in the graph structures GS1.4 and GS5.4, i.e., for cyclic structures as well as those containing vertices with valencies  $\geq 5$ .

**3. Chemical Applications.** The application of the similarity matrix method based upon SRWs is discussed in this section. Two types of properties are investigated, namely, molecular properties such as total  $\pi$ -electron energy and physicochemical properties such as boiling point (BP), condensation point (CP), freezing point (FP), melting point (MP), density (D), and refractive index ( $n_D$ ).

*i.  $\Pi$ -Electron Energies.* Applications of the similarity matrix method to the estimation of total  $\pi$  and  $\Omega$ - $\pi$  energies in terms of  $\beta$  units were carried out for graphs 1–5, and the

Table V. C<sub>9</sub>H<sub>12</sub> Physicochemical Properties and SRWs

|  |   |
|--|---|
|  |   |
| <b>1,2,3-Trimethylbenzene</b><br>BP/CP = 176.2, FP/MP < -15, D = 0.8944, $n_D$ = 1.5139<br>SRWs: 18, 66, 294, 1426, 7158, 36402, 186022                            | <b>1-Methyl-3-ethylbenzene</b><br>BP/CP = 161.3, FP/MP -, D = 0.8645, $n_D$ = 1.4966<br>SRWs: 18, 62, 258, 1166, 5468, 26078, 125346                      |
|  |   |
| <b>1,2,4-Trimethylbenzene</b><br>BP/CP = 169.2, FP/MP = -57.4, D = 0.9758, $n_D$ = 1.5049<br>SRWs: 18, 66, 288, 1362, 6678, 33264, 166842                          | <b>1-Methyl-4-ethylbenzene</b><br>BP/CP = 162.1, FP/MP = -20, D = 0.8612, $n_D$ = 1.4950<br>SRWs: 18, 62, 258, 1158, 5378, 25406, 121090                  |
|  |   |
| <b>1,3,5-Trimethylbenzene</b><br>BP/CP = 169.2, FP/MP = -57.4, D = 0.9758, $n_D$ = 1.5049<br>SRWs: 18, 66, 288, 1362, 6678, 33264, 166842                          | <b>n-Propylbenzene (<i>p</i>-Ethyltoluene)</b><br>BP/CP = 159.2, FP/MP = -101.6, D = 0.8620, $n_D$ = 1.4920<br>SRWs: 18, 58, 228, 978, 4358, 19768, 90458 |
|  |   |
| <b>1-Methyl-2-Ethylbenzene (<i>o</i>-ethyltoluene)</b><br>BP/CP = 169.2, FP/MP = -57.4, D = 0.9758, $n_D$ = 1.5049<br>SRWs: 18, 66, 288, 1362, 6678, 33264, 166842 | <b>Isopropylbenzene (Cumene)</b><br>BP/CP = 152.4, FP/MP = -96.9, D = 0.8618, $n_D$ = 1.4915<br>SRWs: 18, 62, 258, 1158, 5378, 25406, 121090              |

results are given in Table IV. Graphs 1–4 in Table V represent real molecular structures after adding suppressed hydrogen atoms to give the corresponding hexane isomers. The  $\pi$ -electron energies were calculated from Hückel HMO theory with the  $\Omega$  technique.<sup>11</sup> The first step is to calculate the Hückel matrix (which is equivalent to the adjacency matrix) for the hydrocarbon with hydrogens suppressed. This is a molecular graph whose vertices represent the carbon atoms and whose edges represent the bonds. The resulting energies are usually recorded in  $\beta$  units. Calculations of  $\Omega$ - $\pi$  energies for  $\Omega = 1.0$  and 1.4 were carried out, and the similarity matrices based on the three  $\pi$  energies are given, together with the SRWs, in Table III.

The similarity matrices based on SRWs (Table II) show that molecular graphs 3 and 2 are the most similar pair. This pair was reproduced for  $\Omega = 1.0$   $\pi$  energy (99.989%) for the element 3.2 in the similarity matrices based on that particular property (Table IV). It is however slightly shifted toward the element 3.1 in the similarity matrix based on the total  $\pi$  and  $\Omega = 1.4$   $\pi$  energy. This shift may be attributed to the crude nature of the HMO calculations and the graph theoretical invariant that was selected (SRWs).

ii. *Application to Physicochemical Properties.* Boiling point (BP), melting point (MP), density (D), and refractive index ( $n_D$ ) for eight structurally related isomers of C<sub>9</sub>H<sub>12</sub> were used to test the similarity matrix method, and the results are shown in Tables V and VI. These chemicals are all liquids, miscible

in ether and acetone, with a molecular weight of 120.19.<sup>12</sup> Isomers rather than compounds of different composition were selected since the validity of applying the method to the former ensures its validity with respect to the latter.

SRWs were calculated for the hydrogen-suppressed graphs in the usual way, and similarity matrices for both SRWs and the physicochemical properties are collected for comparison in Table V. The similarity matrix based on the MP was not constructed because no data could be found for 1-methyl-3-ethylbenzene (*m*-ethyltoluene) in addition to the rough estimates of that particular property that were necessitated by the fact that it is cited as a range.<sup>12</sup>

The similarity matrices based on SRWs, given in Table VI, show that the molecular structures of isopropylbenzene (cumene) and 1-methyl-4-ethylbenzene (*p*-ethyltoluene), i.e., structures 8 and 6, represent the most similar pair in the group, while the molecular structures of *n*-propylbenzene (1-phenylpropane) (structure 7) and 1,2,3-trimethylbenzene (structure 1) represents the most dissimilar pair.

Identical similarities, defined as the occurrence of identical similarity elements, were observed for the similarity matrix based on SRWs. Six pairs of identical similarities (marked by asterisks in Table VI) were observed. It may be worth noting that this was not the case when SAWs were compared with SRWs when testing the diagnostic efficiency. The elements of the similarity matrix based on SRWs at the top of Table II are different from one another. The top similarity

**Table VI.** Lower Triangular Parts (LTP) of Similarity Matrices (SM) for C<sub>9</sub>H<sub>12</sub> Based on SRWs and the Physicochemical Properties, BP/CP, *D*, and *n<sub>D</sub>*<sup>a</sup>

|   | 1             | 2          | 3          | 4          | 5          | 6             | 7          | 8   |
|---|---------------|------------|------------|------------|------------|---------------|------------|-----|
| LTP of SM Based on SRWs                 |               |            |            |            |            |               |            |     |
| 1                                       | 100           |            |            |            |            |               |            |     |
| 2                                       | -19341.04     | 100        |            |            |            |               |            |     |
| 3                                       | -29577.26     | -10136.67  | 100        |            |            |               |            |     |
| 4                                       | -43602.87     | -24162.83  | -13926.20  | 100        |            |               |            |     |
| 5                                       | -61471.80     | -42031.46  | -31794.81  | -17768.97  | 100        |               |            |     |
| 6                                       | -65781.10*    | -46340.57* | -36103.96* | -22078.37* | -4209.674  | 100           |            |     |
| 7                                       | -96942.33     | -77502.78  | -67266.15  | -53239.98  | -32371.92  | -31063.76*    | 100        |     |
| 8                                       | -65781.10*    | -46340.57* | -36103.96* | -22078.37* | -4209.674* | <u>100</u>    | -31063.76* | 100 |
| LTP of SM Based on BP/CP                |               |            |            |            |            |               |            |     |
| 1                                       | 100           |            |            |            |            |               |            |     |
| 2                                       | 93            | 100        |            |            |            |               |            |     |
| 3                                       | 88.5          | 95.5       | 100        |            |            |               |            |     |
| 4                                       | 89            | 96         | 99.5       | 100        |            |               |            |     |
| 5                                       | 85.1          | 92.1       | 96.6       | 96.6       | 100        |               |            |     |
| 6                                       | 85.9          | 92.9       | 97.4       | 96.9       | 99.2       | 100           |            |     |
| 7                                       | <u>83</u>     | 90         | 94.5       | 94         | 97.9       | 97.1          | 100        |     |
| 8                                       | 76.2          | 83.2       | 87.7       | 87.2       | 91.1       | <u>90.3</u>   | 93.2       | 100 |
| LTP of SM Based on <i>D</i>             |               |            |            |            |            |               |            |     |
| 1                                       | 100           |            |            |            |            |               |            |     |
| 2                                       | 99.981        | 100        |            |            |            |               |            |     |
| 3                                       | 99.971        | 99.989     | 100        |            |            |               |            |     |
| 4                                       | 99.986        | 99.995     | 99.984     | 100        |            |               |            |     |
| 5                                       | 99.970        | 99.989     | 99.999     | 99.984     | 100        |               |            |     |
| 6                                       | 99.967        | 99.985     | 99.996     | 99.980     | 99.997     | 100           |            |     |
| 7                                       | 99.969        | 99.986     | 99.997     | 99.981     | 99.997     | 99.999        | 100        |     |
| 8                                       | <u>99.967</u> | 99.986     | 99.997     | 99.981     | 99.997     | <u>99.999</u> | 99.999     | 100 |
| LTP of SM Based on <i>n<sub>D</sub></i> |               |            |            |            |            |               |            |     |
| 1                                       | 100           |            |            |            |            |               |            |     |
| 2                                       | 99.991        | 100        |            |            |            |               |            |     |
| 3                                       | 99.985        | 99.994     | 100        |            |            |               |            |     |
| 4                                       | 99.991        | 99.999     | 99.995     | 100        |            |               |            |     |
| 5                                       | 99.983        | 99.992     | 99.997     | 99.992     | 100        |               |            |     |
| 6                                       | 99.981        | 99.990     | 99.996     | 99.990     | 99.998     | 100           |            |     |
| 7                                       | <u>99.978</u> | 99.987     | 99.993     | 99.987     | 99.995     | 99.997        | 100        |     |
| 8                                       | 99.978        | 99.987     | 99.992     | 99.987     | 99.995     | <u>99.996</u> | 99.999     | 100 |

<sup>a</sup> Underlined elements in SMs identify the top similar and dissimilar pairs of structures. Identical similarities in the SM based on SRWs are asterisked.

between structures 8 and 6 is not well reflected by the BP data (Table VI), where the corresponding similarity matrix element (8,6) in the matrix based on the property BP (90.3%) was not the largest element. This top similarity, however, was in much better agreement with the properties *D* (99.999) and *n<sub>D</sub>* (99.996). The highest degree of dissimilarity, between structures 7 and 1, was also well reproduced for the full set of properties. In Table V, the *n<sub>D</sub>* of structure 1 is 1.5139, while that of structure 7 is 1.4920. These are the extreme *n<sub>D</sub>* values and suggest that structures 1 and 7 are least similar with respect to the refractive index at least. This result is also obtained from the lower triangular part of the similarity matrix in Table VI, where the value underlined (99.978) is the smallest value in the table.

For the melting points, whose similarity matrix was not constructed, the corresponding MP values for the previously determined most dissimilar pair (structures 7 and 1) are -101.6 and <-15. The difference between these two values represent the largest difference between any two values and is in good agreement with the previous result in which they were the most dissimilar on the SRW scale.

## CONCLUSIONS

It cannot be claimed that the basis set based on the SRWs is the best graph theoretical invariant for testing the degrees of similarity between the physicochemical properties under consideration. There are two reasons for this:

1. Occurrence of identical similarities in the similarity matrix based on SRWs.

2. The presence of elements smaller than the most dissimilar pair (structures 7 and 1) and greater than the most similar pair (structures 8 and 6) in the similarity matrices that are based on the selected physicochemical properties.

In our opinion, similarity measurements may in general be achieved through the following steps:

1. Select a readily constructed graph theoretical invariant that can be represented by sets of variables. This will be better than one represented by a single variable.
2. Test the diagnostic power of the invariant. The fewer the identical degrees of similarities that are reproduced, the more diagnostically powerful is the invariant.

The method is derived from the previously successful application of Euclidean distances. It is heuristic and reveals tendencies but does not produce results with a predictable reliability as far as similarity matrices based upon SRWs are concerned. The method appears to enjoy some success, but SRWs are not proposed as a basis for accurate description of the degree of similarities between structures. Rather, it allows an approach to the testing of other graph theoretical invariants and the development of a more rigorous basis for similarity measurements.

## ACKNOWLEDGMENT

The assistance of the Editor, Dr. G. W. A. Milne, in resolving language problems, improving the presentation, and



clarifying the results in this paper is greatly appreciated. The author also expresses his gratitude to the reviewers for their helpful comments.

## REFERENCES AND NOTES

- (1) Trinajstić, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1983.
- (2) Gati, G. *J. Graph Theory* 1979, 3, 95.
- (3) Read, R. C.; Corneil, D. G. *J. Graph Theory* 1977, 1, 339.
- (4) Weisfeiler, B. On Construction and Identification of Graphs. *Lectures in Mathematics*; Springer-Verlag: Berlin, 1976.
- (5) Shalabi, A. Hessenberg Matrix in Molecular Graph Theory. *Chem. Scr.* 1985, 25, 252-256.
- (6) Randić, M. Random Walks and their Diagnostic Value for Characterization of Atomic Environments. *J. Comput. Chem.* 1980, 1, 386-399 and references cited therein.
- (7) Randić, M.; Wilkins, C. L. Graph-Theoretical Approach to Recognition of Structural Similarity in Molecules. *J. Chem. Inf. Comput. Sci.* 1979, 19, 30-35 and references cited therein.
- (8) Bertz, S. H.; Herndon, W. C. The Similarity of Graphs and Molecules. In *Artificial Intelligence in Chemistry*; Pierce, T. H., Hohne, B. A., Eds.; ACS Symposium Series 306, American Chemical Society: Washington, DC, 1986.
- (9) Basak, S. C.; et al. *Discr. App. Math.* 1988, 19, 17.
- (10) See, for example, Randić, M.; et al. *J. Comput. Chem.* 1979, 3, 5.
- (11) Project SERAPHIM software, IB 1404, National Science Foundation Education, Department of Chemistry, University of Wisconsin, Madison, WI 53706. Details concerning the application of HMO theory can be found in: Greenwood, H. H. *Computing Methods in Quantum Organic Chemistry*; Wiley-Interscience: New York, 1972.
- (12) Helpin, W.; Burkart, A. *Tables for Laboratory and Industry*; Wiley Eastern Ltd.: New Delhi, 1979.

## The NEIC Organic Analysis Reporting System

JOHNNY LEE,\* K. ERIC NOTTINGHAM, and LAURENCE W. STRATTAN

Data Analysis Section, Chemistry Branch, Laboratory Services Division, National Enforcement Investigations Center, Environmental Protection Agency, Denver, Colorado 80225

Received July 2, 1990

The Organic Analysis Reporting System processes data from the Finnigan GC-MS Formaster Data System into a "matrix" report format. The standard report created by the GC-MS system is the "QUAN" report which, though very informative, does not present the data in a format that meets the Agency's needs in explaining analytical results to the lay public. Accordingly, the Organic Analysis Reporting System was created to convert the essential data from the GC-MS reports into a form more understandable by the nonscientist.

## INTRODUCTION

The National Enforcement Investigations Center (NEIC) is an investigative unit of the Environmental Protection Agency's (EPA) Office of Enforcement. The NEIC works on cases that involve potential civil and criminal violations of environmental laws. These cases result in litigation that is very often adversarial. We have found that the presentation of final analytical results in a matrix-like format best portrays our findings. The Organic Analysis Reporting System (OARS) is a PC-based system that extracts organic compound data from the Finnigan Formaster Data System and generates such matrix-type reports.

## METHODS

The goal of the Organic Analysis Reporting System (OARS) is to present organic compound results in a format that is definitive, complete, pertinent, and compact. The report from the Formaster Data System's QUAN Report, though informative and complete, contains instrumental and analytical conditions that are pertinent to the generation of quality data but superfluous and confusing when the results are presented in court. In an adversarial situation it is often prudent to present only the essential results of the analyses which show a violation of environmental law without including the intricacies of the required analytical chemistry.

A typical Finnigan Formaster Data System produces a report such as the one shown in Figure 1.

This report is important to the case development because it has information which demonstrates that each sample was analyzed by using appropriate analytical procedures. The disadvantage of this report is that it is cumbersome. Using this format, it is very difficult to cross reference the analytical results from one sample with results from other samples. The quantitative results are also reported to three significant fig-

ures, implying a precision that may not exist.

The NEIC Organic Analysis Reporting System reformats the information from the Formaster QUAN report into a matrix-type report, like that shown in Figure 2.

This NEIC matrix-type report is easier to read and more informative than the Formaster QUAN report. Seven sample results are presented on one page whereas the Formaster report gives the results of one sample per page. A particular advantage of OARS is the ease of comparison between samples. On a potential hazardous waste site, the extent and amount of contamination from different sampling stations can be easily ascertained. The spread of the contaminants could be of interest to the public as well as the courts.

An example of an actual case where the NEIC data led to a guilty plea by a polluter shows the advantage of matrix-type data presentation.

Investigators in Buffalo, NY, discovered that large amounts of solvents were accumulating in a storm drain under a baseball field. The investigators collected samples through a manhole cover in the playing field and from drums from a plant upstream of the storm drain. The NEIC analyses of the samples from both sources determined that they contained percent level solvents. All the samples exhibited the RCRA characteristic of ignitability when they flashed at less than 20 °C (68 °F) using Method 1020 of SW-846.

The EPA's definition of the characteristics of hazardous waste for a liquid is published in the Code of Federal Regulations (CFR), Title 40, Part 261, Subpart C, Section 261.21, paragraphs (a)(1).

### 261.21. Characteristics of Ignitability.

(a) A solid waste exhibits the characteristic of ignitability if a representative sample of the waste has any of the following properties:

(1) It is a liquid, other than an aqueous solution containing less than 24% alcohol by volume and has a