# A Comparative Study of Proteomics Maps Using Graph Theoretical Biodescriptors

Milan Randić[†] and Subhash C. Basak*[,‡]

National Institute of Chemistry, Ljubljana, Slovenia, Department of Mathematics and Computer Science, Drake University, Des Moines, Iowa 50311, and Natural Resources Research Institute, University of Minnesota at Duluth, 5013 Miller Trunk Highway, Duluth, Minnesota 55811

Received August 10, 2001

This paper reports the development of new methods for mathematical characterization of effects of different toxic agents on the cellular proteome. We describe numerical characterization of proteomics maps based on mathematical invariants. A graph is first associated with a proteomics map by considering partial ordering of spots on 2-D gels by ordering proteins with respect to the mass and the charge, the two properties by which proteins are separated. The graph is then embedded over the map, and several graph theoretical invariants have been constructed. In particular we consider invariants that can be extracted from the Euclidean distance-adjacency matrix of the embedded graph, in which only Euclidean distances between adjacent vertices of a graph are considered. The approach is illustrated using proteomics patterns of normal liver cells of rats and those derived from liver cells of animals exposed to four peroxisome proliferators. In contrast to direct comparison of spot abundance our approach incorporates information on spots locations. The difference between the two approaches is that in the first case only changes in abundances are considered as a measure of perturbation of the proteome map, but in the second case not only the charge but also the mass of proteins are used for ordering protein spots.

## INTRODUCTION

In this paper we have attempted the development of novel numerical descriptors for characterization of proteomics maps, that could facilitate comparison of proteomics data, whether from a single laboratory or from different laboratories. A typical proteomics map may have 2000 and more separated proteins that appear as spots on a 2-D gel in which they are separated by charge (*x*-axis) and mass (*y*-axis). When an experimental animal is exposed to drugs or toxins, the protein content of cells usually change. As a consequence the relative abundance of various proteins that characterize a normal cell is perturbed resulting in an altered proteomics map. To facilitate comparison of such changes we want to characterize individual proteomics maps by a set of map descriptors, which are mathematical invariants, derived from proteomics patterns.

Information contained in a map can be (and often is) represented as a list of coordinates and abundance of individual spots, as illustrated in Table 1 for 20 most intensive spots from the liver of normal rats and those exposed to four toxic chemicals. Figure 1 shows the positions of the 20 most abundant protein spots labeled A−T based on information in the columns listing coordinates *x, y* of Table 1. We consider the information in Table 1 as *input* information, and our goal is to arrive at some numerical descriptors that characterize a map as a whole so that they can be used for quantitative comparison of different proteomics maps. What makes our approach different from a direct comparison of spot abundance, as shown in different columns of Table 1, is that the approach incorporates besides the information on the abundance also the information on spot locations. The difference between the two approaches is that in the first case only changes in abundances are considered as a measure of perturbation of a proteomics map, but in the second case also locations of protein spots, i.e., the charge and mass of proteins, make an influence.

**Table 1.** Gel Coordinates and Abundance for the 20 Most Intensive Protein Spots of Rat Liver Cells of the Normal Cells and the Corresponding Abundance for Rat Liver Cells Exposed to Four Chemicals

| # | *x* | *y* | control | PFOA | PFDA | clofibrate | DEHP |
|---|---|---|---|---|---|---|---|
| A | 2111.7 | 2278.6 | 144357 | 108713 | 95028 | 147081 | 165886 |
| B | 2804.3 | 903.6 | 143630 | 155565 | 188582 | 159898 | 155055 |
| C | 1183.9 | 959.6 | 136653 | 113859 | 150253 | 163645 | 8111 |
| D | 2182.2 | 928.8 | 127195 | 99160 | 73071 | 76642 | 112096 |
| E | 2685.6 | 1196.1 | 118581 | 112790 | 49769 | 109856 | 138795 |
| F | 1527.9 | 825.5 | 114929 | 192437 | 221567 | 166080 | 180590 |
| G | 1346.0 | 1352.5 | 112251 | 58669 | 38915 | 73159 | 77075 |
| H | 2868.5 | 778.0 | 108883 | 26105 | 50735 | 45923 | 116849 |
| I | 1406.3 | 1118.1 | 98224 | 91147 | 82963 | 84196 | 92942 |
| J | 2450.2 | 409.2 | 93601 | 83172 | 62934 | 79870 | 109381 |
| K | 1474.0 | 665.1 | 90004 | 129340 | 112361 | 112655 | 119402 |
| L | 2974.9 | 772.8 | 86730 | 70746 | 78691 | 105760 | 116281 |
| M | 2068.4 | 823.1 | 84842 | 73814 | 45482 | 71911 | 97444 |
| N | 642.2 | 669.8 | 82492 | 73974 | 74466 | 84703 | 88545 |
| O | 2860.7 | 1649.9 | 81965 | 16137 | 16501 | 60077 | 148992 |
| P | 2032.7 | 902.8 | 80015 | 77314 | 80072 | 76027 | 100836 |
| Q | 2752.7 | 765.6 | 79847 | 20782 | 13103 | 38816 | 53830 |
| R | 2334.2 | 982.2 | 72791 | 76369 | 52749 | 55599 | 77432 |
| S | 1053.6 | 864.3 | 72173 | 77982 | 60376 | 46808 | 78121 |
| T | 2519.5 | 1365.9 | 69452 | 37838 | 16129 | 57167 | 71274 |

Before continuing with details we need to respond to the question: "Why the 20 most abundant proteins are chosen for analysis?" The answer is that for the development of a mathematical approach neither the number of selected points

* Corresponding author phone: (218)720-4230; fax: (218)720-4328; e-mail: sbasak@nrri.umn.edu.
[†] National Institute of Chemistry and Drake University.
[‡] University of Minnesota at Duluth.

10.1021/ci0100797 CCC: $22.00 
Published on Web 08/09/2002

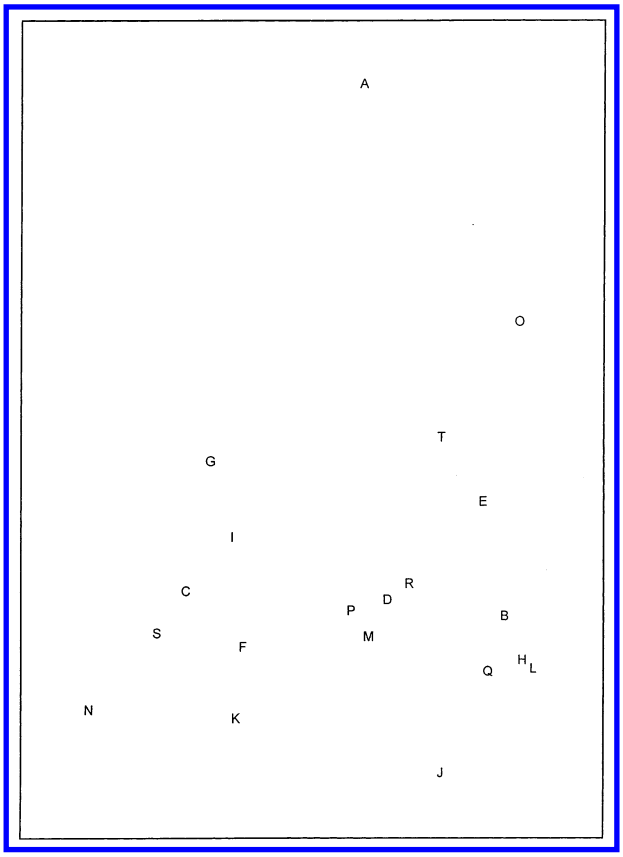


**Figure 1.** The position of 20 most abundant protein spots of rat liver cells labeled alphabetically.

nor the criteria of selection is essential. Because of the wide variability of experimental data on proteomics in general we focused on the most abundant proteins. Whether 20 most abundant proteins suffice to represent a map as a whole or one should select 50, or 200, remains yet to be better explored. Equally it remains to be seen if the selection of a subset of N proteins rather than all of them could be justified. If N is sufficiently large (and that may be 20, 50, or 200) one hopes that such subsets may capture the most features of a map that may suffice for comparison of different proteomics maps.

A proteomics map invariant is a quantity that is independent of the coordinates and the labels of the proteins in proteomics maps or 2-D gels. Why is it so important for mathematical descriptors to be independent of the adopted protein labels? A quantity independent of the arbitrary labeling of proteins characterizes a system as a *whole* and does not pertain to any of its local features or any particular protein. To be more correct we should say that such characterization pertains to a set of protein spots selected as the basis for numerical characterization of a map. Hence, map invariants are properties of the cell/tissue considered as represented by a group of proteins as a whole and not of any particular protein. This, however, in no way prevents one to compare protein expressions across different samples taken from the same species.

It is obvious that we need many invariants if we are to capture the majority of salient features of information-rich proteomics maps. Thus this paper represents a continuation of our recent efforts[1−6] to develop mathematical characterization of proteomics maps. In particular, in this article we will consider sets of invariants that can be constructed by a procedure outlined in ref 4 which can be viewed as part I of this paper, in which an approach is outlined in which one associates an *embedded graph* to a proteomics map through *partial ordering* of spots with respect to their charge and mass.

Because the intensity of a spot in the 2D gel will depend not only on the concentration of proteins but also on the total amount of protein used in running the gel, reliability of experimental data should be known. We used experimental data from a laboratory known for following a rigorous protocol,[7] and thus we have some confidence that the overall effects of various peroxisome proliferators resulted in reduced overall production or proteins.

## PARTIAL ORDERING BASED ON CHARGE AND MASS

Partial ordering has not been widely used in chemistry despite the fact that it offers a useful tool for comparison of objects characterized by sequences. For example, partial ordering allows one to compare properties of isomers,[8,9] and it can be used to search for pharmacophore in QSAR studies.[10,11] For a recent review of use of partial orderings in chemistry see the special issue of MATCH.[12] In this paper we will use partial order to generate a *graph* to be associated with a map. Once we have a graph, we can construct various graph matrices, and from such matrices we can construct graph invariants to serve as map descriptors.

To obtain a graph associated with protein spots listed in Table 1, we will *order* the 20 spots of Table 1 with respect to the *charge* and the *mass*. By ordering protein spots with respect to mass we obtain the sequence A, O, T, G, E, I . . . which gives an ordering in which spots appear when we list them as they appear in going from the top of the map to the bottom, beginning with the protein having the greatest mass and ending with the protein having the smallest mass. As we see at the top of Figure 1, the protein spot A has the largest mass and is followed by spot O, the protein with the second largest mass, and so on. When we order the 20 protein spots with respect to the charge we obtain the sequence L, H, O, B, Q, E . . . which reflects the order in which protein spots appear if we go from the right (the largest charge) to the left (the smallest charge). As we see from Figure 1 protein L has the largest charge, followed by protein H, and so on.

*Partial ordering* is defined as a set of *all* sequential orderings given by subsets of the two sequences, the elements of which maintain the ordering for both sequences. To obtain all sequential orderings preserved in both sequences it is convenient to construct a diagram, as shown in Figure 2, in which the two sequences are listed, one above the other and lines are drawn that connect the same element in both sequences. Any set of noncrossing lines constitutes a subset of partial ordering, the subsequence the relative order of which is present in both sequences. For example, the lines O − O, E − E, R − R, D − D, P − P, F − F, and N − N do not cross giving the subsequence O > E > R > D > P > F > N. In Table 2 we have listed all such fragmentary orderings that start with protein spot A. The totality of all fragmentary orderings makes the partial ordering for the selected 20 proteins of the map of Figure 1.

A complete list of all sequences of partial ordering is neither a necessary nor particularly useful representation of the partial ordering sought. It is customary to represent partial

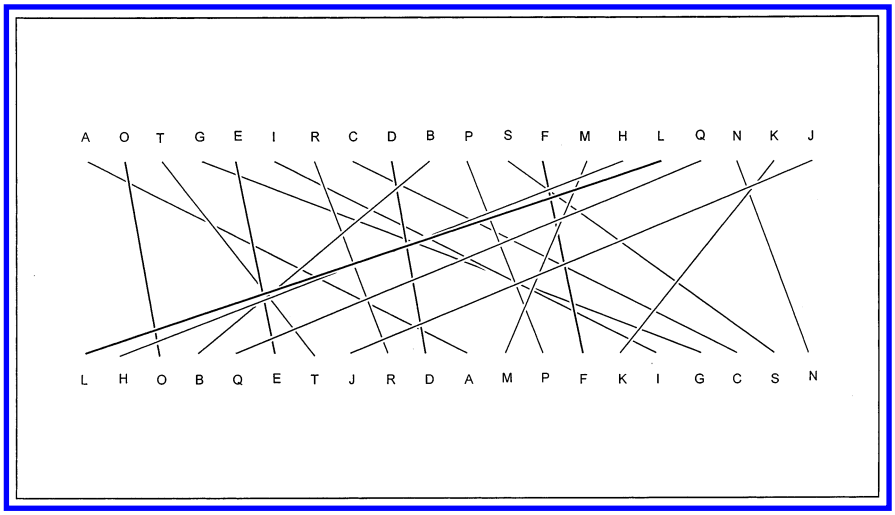


**Figure 2.** Ordering of proteins of Figure 1 relative to their mass (the top sequence) and their charge (the bottom sequence). The same labels in both sequences are connected by line to help to identify partial orderings.

**Table 2.** Fragmentary Orderings of Protein Spots Holding with Respect to Charge and Mass Respectively that Start with Protein A

| | | | | | |
|---|---|---|---|---|---|
| 1 | A ⟹ | G ⟹ | C ⟹ | S ⟹ | N |
| 2 | A ⟹ | I ⟹ | C ⟹ | S ⟹ | N |
| 3 | A ⟹ | M ⟹ | K | | |
| 4 | A ⟹ | M ⟹ | N | | |
| 5 | A ⟹ | P ⟹ | F ⟹ | K | |
| 6 | A ⟹ | P ⟹ | F ⟹ | N | |
| 7 | A ⟹ | P ⟹ | S ⟹ | N | |

ordering by a directed graph or a hierarchical diagram shown in Figure 3 in which points at the left *dominate* points to the right. *Dominance* here means that an element at the *left* in Figure 3 is also at the *left* in both sequences of Figure 2. Figure 3 facilitates visual inspection of the partial ordering and immediately shows which protein dominates others by mass and charge or is dominated by others. In making Figure 3 one can place vertices (labels) in arbitrary positions as long as the dominance relation is preserved. Nevertheless, one tries to construct partial ordering diagrams that will have, for better visibility, as few crossings of lines as possible.

The next important step in analysis of proteomics map based on partial ordering is to *embed* the partial ordering diagram of Figure 3 directly over the map. This is a critical step that has been for the first time considered in relationship to partial ordering in general and proteomics maps in particular in ref 4. *Embedding* of graphs in 2-D space (plane) or 3-D space in Graph Theory implies that a fixed geometry is assigned to the graph that can no longer be viewed as only giving *adjacency* relationships between vertices but also giving the *distances* between linked elements. While Graph Theory[13] is usually concerned with *topological* and *combinatorial* properties of mathematical objects, in the case of embedded graphs it is concerned solely with the geometry and the combinatorial properties of the same objects, while topology of objects is no longer of concern. In Figure 4 we have imposed the adjacency relations of Figure 3 directly over the map of Figure 1, and as a result we obtain a partial ordering graph *embedded* on the map. Because the protein spots have definite (*x*, *y*) coordinates now proteins at the top dominate those below if connected by lines and proteins at the right dominate those at the left if connected by lines.

Observe an important and interesting feature of the embedded graph of Figure 4 in that all the lines (links) in the graph have positive slope. This is a consequence of the dominance relation and is a property which can be exploited for a direct construction of embedded graph for a given map, *without* a need for construction of Figure 2, listing the various fragmentary orderings, and depicting the hierarchical diagram of the partial ordering, such as Figure 3. To obtain Figure 4 directly from Figure 1 one can start with the top vertex (spot A in Figure 4) and connect it to the next lower spot to the left of A which is spot G. One can continue the same with vertex G and connect it to the next lower vertex below it and to the left, which is spot C. We continue to connect C to S and finally S to N. By exhausting this particular trail we return to vertex A and repeat the process: we connect A to I and then I to C. In the next step, we connect A to P and finally P to S. By backtracking we connect P to F and F both to K and N, since both K and N are below P and at the left of P. Finally we backtrack again to protein A and we connect it to M, which is connected to K and N. This has exhausted all the fragmentary orders starting with protein spot A, that have been listed in Table 2. The process continues with spot O, then L and H, which completes construction of the embedded graph of partial order for the map considered.

## THE EUCLIDEAN−ADJACENCY MATRIX *E* OF A MAP

The adjacency matrix is a well-known binary matrix of Graph Theory. Its matrix elements are defined as

$$A_{ij} = 1 \quad \text{if vertices } i \text{ and } j \text{ are adjacent}$$

$$A_{ij} = 0 \quad \text{otherwise}$$

The Euclidean-adjacency matrix E is a novel matrix only recently suggested,[13] that is defined only for *embedded* graphs (because only embedded graph in 2-D or 3-D space have fixed geometry). Its matrix elements are defined by

$$E_{ij} = \text{Euclidean distance} \quad \text{if vertices } i \text{ and } j \text{ are adjacent}$$

$$E_{ij} = 0 \quad \text{otherwise}$$

Hence, the difference between A and E matrices is only in the magnitudes of the nonzero entries, which in the former

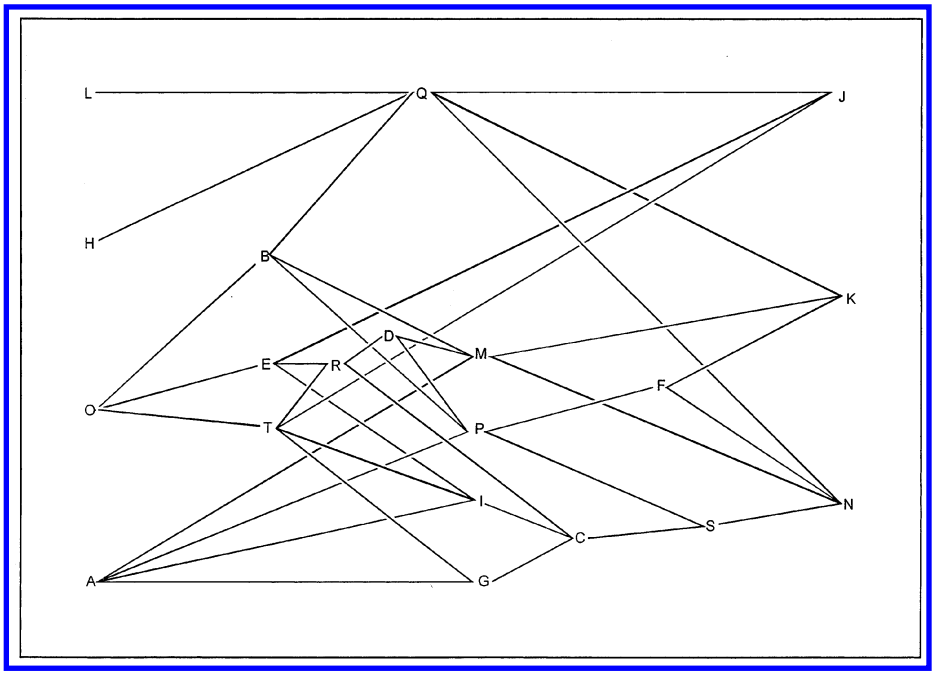


**Figure 3.** Graphical representation of the partial ordering of Figure 1.

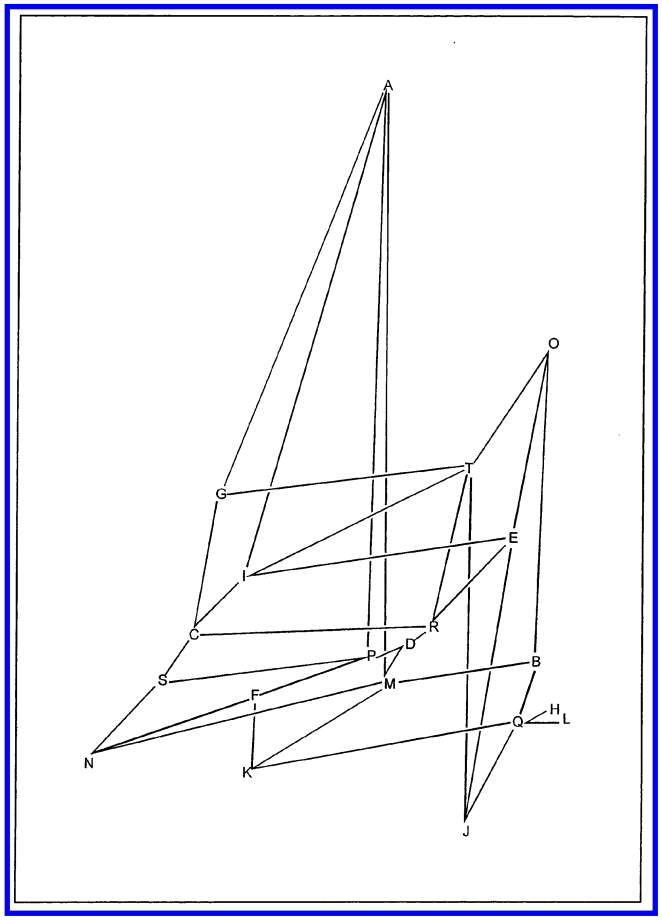


**Figure 4.** The adjacency relationships of Figure 3 embedded upon the proteomics map of Figure 1 that results in embedded graph of the partial order of proteins with respect to mass and charge.

take the value 1 and in the latter assume numerical values determined by the distance between points considered.

The embedded graph of Figure 4 is the starting point for our mathematical characterization of proteomics maps. First, in view of the fact that that $x$, $y$ coordinates represent physically distinct properties and are therefore measured in different units, we will follow a recommendation of Kowalski and Bender[14] and will rescale the coordinate units. However, rather than using the scaling that reduces both coordinate values to an interval $(-1, +1)$, that Kowalski and Bender suggested, we will scale the distances between adjacent spots of Figure 4 so that the average distance becomes 1. The reason for this is that then not only the entries in the Euclidean-adjacency matrix E and the adjacency matrix A are of comparable magnitude but also because their averages are the same, for both matrices the sum of all entries above the main diagonal is the same.[20,21] In Table 3 in the column "norm" we show the rescaled Euclidean distances between the adjacent vertices of the embedded graph of Figure 4.

Adjacency and the Euclidean-adjacency matrices will generate a number of different topological or topographic indices.[15] As we can see from Table 4, where we have listed the row sums for the adjacency matrix and the Euclidean-

**Table 3.** Reported Distances, Normalized Distances between Adjacent Spots, as Defined by the Partial Order for the 20 Intensive Protein Spots of Normal Rat Liver Cells, as the Corresponding Matrix Elements of Euclidean-Adjacency Matrix

| $E_{ij}$ | line | distance | norm | $E_{ij}$ | line | distance | norm |
|---|---|---|---|---|---|---|---|
| 1, 7 | AG | 1201.7 | 1.6207 | 5, 18 | ER | 411.4 | 0.5549 |
| 1, 9 | AI | 1358.1 | 1.8319 | 6, 11 | FK | 169.2 | 0.2282 |
| 1, 13 | AM | 1456.1 | 1.9641 | 6, 14 | FN | 899.3 | 1.2130 |
| 1, 16 | AP | 1378.1 | 1.8588 | 6, 16 | FP | 510.7 | 0.6888 |
| 2, 13 | BM | 740.3 | 0.9986 | 7, 20 | GT | 1173.6 | 1.5830 |
| 2, 15 | BO | 748.4 | 1.0095 | 8, 18 | HQ | 116.5 | 0.1571 |
| 2, 16 | BP | 771.6 | 1.0408 | 9, 20 | IT | 1140.4 | 1.5383 |
| 2, 17 | BQ | 147.3 | 0.1987 | 10, 17 | JQ | 467.5 | 0.6306 |
| 3, 7 | CG | 425.0 | 0.5733 | 10, 20 | JT | 959.2 | 1.2938 |
| 3, 9 | CI | 273.1 | 0.3684 | 11, 13 | KM | 615.0 | 0.8296 |
| 3, 18 | CR | 1150.5 | 1.5519 | 11, 17 | KQ | 1282.6 | 1.7301 |
| 3, 19 | CS | 161.4 | 0.2177 | 12, 17 | LQ | 222.3 | 0.2999 |
| 4, 13 | DM | 155.3 | 0.2095 | 13, 14 | MN | 1434.4 | 1.9348 |
| 4, 16 | DP | 151.7 | 0.2047 | 14, 17 | NQ | 2112.7 | 2.8497 |
| 4, 18 | DR | 161.1 | 0.2173 | 14, 19 | NS | 455.1 | 0.6138 |
| 5, 9 | EI | 1281.7 | 1.7288 | 15, 20 | OT | 443.9 | 0.5988 |
| 5, 10 | EJ | 821.4 | 1.1079 | 16, 19 | PS | 979.9 | 1.3217 |
| 5, 15 | EO | 486.4 | 0.6561 | 18, 20 | RT | 426.1 | 0.5748 |

**Table 4.** Row Sums for the Adjacency Matrix (A) and the Euclidean-Adjacency Matrix (E) of Graph of Figure 4

| spot | spot | row sum for A | row sum for E | new label |
|---|---|---|---|---|
| 1 | A | 4 | 7.2756 | 1 |
| 2 | B | 4 | 3.2476 | 10 |
| 3 | C | 4 | 2.7113 | 14 |
| 4 | D | 3 | 0.6315 | 18 |
| 5 | E | 4 | 4.0477 | 8 |
| 6 | F | 3 | 2.1302 | 17 |
| 7 | G | 3 | 3.7771 | 9 |
| 8 | H | 1 | 0.1571 | 20 |
| 9 | I | 4 | 5.4673 | 6 |
| 10 | J | 3 | 3.0323 | 11 |
| 11 | K | 3 | 2.7880 | 13 |
| 12 | L | 1 | 0.2999 | 29 |
| 13 | M | 5 | 5.9476 | 3 |
| 14 | N | 4 | 6.6114 | 2 |
| 15 | O | 3 | 2.2544 | 25 |
| 16 | P | 5 | 5.1148 | 7 |
| 17 | Q | 6 | 5.8661 | 5 |
| 18 | R | 4 | 2.8989 | 12 |
| 19 | S | 3 | 2.1533 | 16 |
| 20 | T | 5 | 5.887 | 4 |

adjacency matrix of the graph of Figure 4, already the two matrices have different row sums and therefore matrix invariants based on them will be distinct. In the case of the adjacency matrix of a graph row sum represents the valence (or the degree) of the vertex, hence, we can interpret the row sums of the Euclidean-adjacency matrix as E as the "distance valences" of vertices of the embedded graph. Vertices that have as neighbors more vertices at larger distances will have a larger value for their "distance valence," while vertices having fewer neighbors and at shorter distances will have lower values of their "distance" valence.

## ADDITIONAL MAP MATRICES

An advantage of having a graph (and embedded graph) associated with a map as compared to the case of associating a map with a zigzag curve is that for such a graph we can construct a number of additional matrices, which subsequently will lead to additional alternative characterizations of the same map. We will briefly outline several of such additional matrices associated with the embedded graph of Figure 4 that may be of interest as a source of novel map invariants.

**The Geometry Matrix G.** In contrast to the adjacency and the Euclidean−adjacency matrices which are sparse matrices (that is matrices having many zero elements)[16] geometrical matrix G is a dense matrix having zero entries only on the main diagonal. Its matrix element (i, j) represents the Euclidean distance between vertex i and vertex j regardless whether i and j are adjacent or not.

**D/D Matrix.** As is known, the elements of the Graph Theoretical Distance Matrix D express *graph theoretical distances* between a pair of vertices, which are defined by the smallest number of edges that separate two vertices.[17] Once D is constructed, it can be combined with the G matrix for construction of the so-called D/D matrix,[18] the elements of which are given as the quotient of the corresponding elements of D and G matrix.

**The Map Connectivity Matrices $D_{k\chi}$.** These new matrices are reported for the first time in this article. They are based on partitioning of the connectivity index[19] and the higher order connectivity indices[20] into contributions arising from paths of length k.

**Table 5.** Nonzero Matrix Elements of the Map Connectivity Matrix $D_{1\chi}$

| element | | element | | element | |
|---|---|---|---|---|---|
| 1, 7 | 0.19076 | 4, 13 | 0.51599 | 9, 20 | 0.17627 |
| 1, 9 | 0.15855 | 4, 16 | 0.55642 | 10, 17 | 0.23710 |
| 1, 13 | 0.15201 | 4, 18 | 0.73909 | 10, 20 | 0.23668 |
| 1, 16 | 0.16393 | 5, 9 | 0.21257 | 11, 13 | 0.24557 |
| 2, 13 | 0.22753 | 5, 10 | 0.28544 | 11, 17 | 0.24727 |
| 2, 15 | 0.36958 | 5, 15 | 0.33104 | 12, 17 | 0.75395 |
| 2, 16 | 0.24536 | 5, 18 | 0.29193 | 13, 14 | 0.15947 |
| 2, 17 | 0.22911 | 6, 11 | 0.41034 | 14, 17 | 0.16057 |
| 3, 7 | 0.31249 | 6, 14 | 0.26647 | 14, 19 | 0.26503 |
| 3, 9 | 0.25973 | 6, 16 | 0.30295 | 15, 20 | 0.27450 |
| 3, 18 | 0.35669 | 7, 20 | 0.21207 | 16, 19 | 0.30132 |
| 3, 19 | 0.41387 | 8, 17 | 1.04169 | 18, 20 | 0.24207 |

**Table 6.** Nonzero Matrix Elements of the Map Connectivity Matrix $D_{2\chi}$

| element | | element | | element | |
|---|---|---|---|---|---|
| 1, 2 | 0.06575 | 3, 4 | 0.48660 | 7, 18 | 0.13858 |
| 1, 3 | 0.22998 | 3, 5 | 0.33215 | 8, 10 | 0.64853 |
| 1, 4 | 0.45814 | 3, 14 | 0.12049 | 8, 12 | 2.06227 |
| 1, 5 | 0.08544 | 3, 16 | 0.15350 | 8, 14 | 0.43921 |
| 1, 6 | 0.10142 | 3, 20 | 0.42577 | 9, 10 | 0.11265 |
| 1, 11 | 0.09879 | 4, 5 | 0.39825 | 9, 15 | 0.13033 |
| 1, 14 | 0.06416 | 4, 6 | 0.34423 | 9, 18 | 0.11519 |
| 1, 19 | 0.00332 | 4, 11 | 0.33532 | 10, 12 | 0.46939 |
| 1, 20 | 0.16018 | 4, 14 | 0.21775 | 10, 14 | 0.09997 |
| 2, 3 | 0.12350 | 4, 19 | 0.31807 | 10, 15 | 0.17500 |
| 2, 4 | 0.22318 | 4, 20 | 0.33893 | 10, 18 | 0.15467 |
| 2, 5 | 0.14273 | 5, 17 | 0.12776 | 11, 12 | 0.48952 |
| 2, 8 | 0.45016 | 5, 20 | 0.51371 | 11, 14 | 0.20789 |
| 2, 10 | 0.10246 | 6, 13 | 0.13572 | 12, 14 | 0.31789 |
| 2, 11 | 0.23803 | 6, 17 | 0.22519 | 13, 17 | 0.07145 |
| 2, 12 | 0.32581 | 6, 19 | 0.22979 | 13, 19 | 0.08143 |
| 2, 14 | 0.30305 | 7, 9 | 0.10091 | 14, 16 | 0.09830 |
| 2, 16 | 0.20979 | 7, 10 | 0.13550 | 15, 18 | 0.17898 |
| 2, 20 | 0.12147 | 7, 15 | 0.15680 | | |

**Table 7.** Row Sums of the Map Connectivity Matrix $D_{1\chi}$ and the Row Sums of the Map Connectivity Matrix $D_{2\chi}$

| row | sum | row | sum | row | sum | row | sum |
|---|---|---|---|---|---|---|---|
| 1 | 0.7230 | 11 | 0.8422 | 1 | 1.2672 | 11 | 1.3696 |
| 2 | 1.0381 | 12 | 0.7957 | 2 | 2.3059 | 12 | 3.6645 |
| 3 | 1.2819 | 13 | 1.3061 | 3 | 1.8720 | 13 | 0.2886 |
| 4 | 1.9829 | 14 | 0.7401 | 4 | 3.1205 | 14 | 1.6608 |
| 5 | 1.1822 | 15 | 0.9248 | 5 | 1.6000 | 15 | 0.4661 |
| 6 | 1.0115 | 16 | 1.3641 | 6 | 1.0363 | 16 | 0.4616 |
| 7 | 0.7608 | 17 | 1.8465 | 7 | 0.5318 | 17 | 0.4244 |
| 8 | 1.0993 | 18 | 1.7266 | 8 | 3.6002 | 18 | 0.5874 |
| 9 | 0.8567 | 19 | 0.9388 | 9 | 0.4591 | 19 | 0.6312 |
| 10 | 0.8078 | 20 | 1.2358 | 10 | 1.7855 | 20 | 1.5601 |

In the case of the first order connectivity index (referred to as the Randić index of order 1 in CODESSA, the computer program of Katritzky, Lobanov, and Kerelson[21] for computation of molecular descriptors) matrix elements represent contributions of paths of length one (bonds) to the $^1\chi$. In Table 5 we give the nonzero matrix elements for the Map Connectivity Matrix $D_{1\chi}$. Similarly, in Table 6 we give the nonzero matrix elements for the Map Connectivity Matrix $D_{2\chi}$, which is based on contributions from paths of length two. Finally in Table 7 we have listed the row sums for these two matrices. Tables 4−7 offer new possibilities for construction of additional numerical map descriptors. The situation is similar to that of the construction of novel topological indices from matrices associated with graphs, which produce novel chemodescriptors for QSAR. The difference is that descriptors obtained from matrices associ-

**Table 8.** Normalized Abundance for the 20 Most Intensive Protein Spots of Rat Liver Cells of the Normal Cells and the Corresponding Abundance for Rat Liver Cells Exposed to Four Chemicals

| | control | PFOA | PFDA | clofibrate | DEHP |
|---|---|---|---|---|---|
| A | 1.44457 | 1.08788 | 0.95094 | 1.47183 | 1.66001 |
| B | 1.43730 | 1.55673 | 1.88713 | 1.60009 | 1.55162 |
| C | 1.36748 | 1.13938 | 1.50357 | 1.63758 | **0.08116** |
| D | 1.27283 | 0.99229 | 0.73122 | 0.76695 | 1.12174 |
| E | 1.18663 | 1.12868 | **0.49803** | 1.09932 | 1.38891 |
| F | 1.15009 | 1.62550 | **2.21721** | 1.66195 | 1.80715 |
| G | 1.12329 | 0.58710 | **0.38942** | 0.73210 | 0.77128 |
| H | 1.08958 | **0.26123** | **0.50770** | **0.45923** | 1.16930 |
| I | 0.98292 | 0.91210 | 0.83020 | **0.19614** | 0.93006 |
| J | 0.93666 | 0.83230 | 0.62978 | 0.79925 | 1.09457 |
| K | 0.90066 | 1.29430 | 1.12439 | 1.12733 | 1.19485 |
| L | 0.86790 | 0.70795 | 0.78746 | 1.05833 | 1.16362 |
| M | 0.84901 | 0.73865 | 0.45514 | 0.71961 | 0.97512 |
| N | 0.82549 | 0.74025 | 0.74518 | 0.84762 | 0.88606 |
| O | 0.82022 | **0.16148** | **0.16512** | 0.60119 | **1.49095** |
| P | 0.80070 | 0.77368 | 0.80127 | 0.76080 | 1.00906 |
| Q | 0.79902 | **0.20796** | **0.13112** | **0.38843** | 0.53867 |
| R | 0.72841 | 0.76422 | 0.52786 | 0.55638 | 0.77486 |
| S | 0.72223 | 0.78036 | 0.60418 | 0.46840 | 0.78175 |
| T | 0.69500 | 0.37864 | **0.16140** | 0.57207 | 0.71323 |

ated with proteomics maps serve as potential biodescriptors in an integrated QSAR,[22] in which chemodescriptors (topological indices, 3-D descriptors, and quantum chemical indices) are used in combination with invariants of proteomics maps (graph theoretical indices derived from embedded graphs) to predict the bioactivity/toxicity of chemicals. Moreover, any of the new matrices can be algebraically manipulated to produce the so-called "higher order" matrices either by using the Kronecker multiplication of matrices or the standard matrix multiplication.

## AUGMENTED MAP MATRICES

So far we have not used information on abundance of proteins as measured by the intensity of gel spots. This information is critical when considering the role of various drugs and xenobiotic agents that perturb the proteome. The information on abundance can be combined in two ways with the information of locations of spots to produce invariants that will characterize altered proteomics maps. One approach would be to introduce the numerical values of experimental abundances as the diagonal entries of the matrix in analogy with differentiation of heteroatoms in construction of variable connectivity indices.[23−27] Another way to incorporate the abundance is to view it as the third coordinate to the 2-D map based on the charge and the mass of proteins as separated by electrophoresis and chromatography.[2] This approach has been outlined in the introductory papers on the mathematical characterization of proteomics maps based on 2-D zigzag pattern of proteomics map and its generalization to 3-D zigzag pattern.[2] In both cases we will obtain novel matrices for proteomics maps of control cells as well as those derived from cells exposed to different chemicals.

Before one augments a map matrix by either including abundance as the diagonal entries or considers them as the third coordinate, one has to rescale the abundance values so that they are of the same order of magnitude as are the other two coordinates used for construction of the matrix. We have rescaled abundance values so that the average abundance equals one. In Table 8 we listed the rescaled abundance values for the 20 most intense protein spots of normal rat hepatocytes and the corresponding abundance values for rat hepatocytes exposed to peroxisome proliferators: perfluorooctanoic acid (PFOA), perfluorodecanoic acid (PFDA), clofibrate, and diethylhexyl phthalate (DEHP). As is known, the sites of the protein spots for the same protein always occur in the same locations on the gel (the same $x$, $y$ coordinates), so correspondence can be established without identifying proteins. The 20 most abundant spots of the control group determine the selection of the spots for the remaining four proteomics maps corresponding to PFOA, PFDA, clofibrate, and DEHP. As can be seen in the first numerical column of Table 8, belonging to the control cells, the relative abundance values steadily decrease because spots were ordered by abundance. This, however, need not and is not the case with the remaining four columns, which show considerable oscillations in the relative magnitudes of the abundances when compared to the regular decrease of spots for the control group.

The most abundant spot appears to be the protein having label F for liver cells exposed to PFDA (2.217), and the least abundant is protein C for liver cells exposed to DEHP (0.081). From looking at Table 8, the question to consider is the following: how significant are the changes in abundance and how reproducible would the values in Table 8 be if the experiments were repeated. These are questions directed to *experimentalists* to which we cannot answer, except that we may add that the reproducibility of our numerical analysis is guaranteed once the reliability of the experimental data is established. Our contribution is toward *developing* theoretical methodology for quantitative comparison of proteomics maps, and the questions of reproducibility/error of protein data, which may be a major unanswered question in experimental proteomics, will not affect our *methodology*. In other words, as experimental techniques improve, the proposed numerical analysis will gain more and more reliability, which currently depends solely on the reliability of the experimental data.

In Table 8 we have indicated in bold type abundances of proteins, which show considerable change in rats, treated with peroxisome proliferators. As a threshold we have taken $\pm 50\%$ change in the abundance as compared with that of the control group. It is interesting to observe protein O, which in the case of PFOA and PFDA has *decreased* in abundance considerably; but in the case of DEHP, its abundance has *increased*. In many cases there appear to be considerable change in abundances in comparison with that of the control group even if the change is within the set bounds of $\pm 50\%$. In some cases abundances of proteins increased slightly after exposure to chemicals. Similarly, for some proteins the abundance decreased, although often not evenly (G, I, and Q). Protein N appears to have been the least affected by any of the four chemical agents.

Table 9 gives the degree of similarity/dissimilarity for pairwise comparisons of proteomics maps. The values in the table were computed by viewing each column in Table 8 as a 20-component vector. The Euclidean distance (in 20-dimensional vector space) gives the distance between the corresponding end-points of vectors. The smaller is the distance the more similar are vectors (or alternatively the more similar are the corresponding proteomics maps). The first row gives the similarity with the map of the control group, which suggests that among the four peroxisome

**Table 9.** Similarity/Dissimilarity among Perturbations of Abundances of Proteome Rat Liver Cells for the Control and the Four Peroxisome Proliferators Based on 20-Component Vectors of Table 8

| | control | PFOA | PFDA | clofibrate | DEHP |
|---|---|---|---|---|---|
| control | 0 | 1.606 | 2.190 | 1.499 | 1.769 |
| PFOA | | 0 | 1.217 | 1.261 | 2.198 |
| PFDA | | | 0 | 1.481 | 2.714 |
| Clofibrate | | | | 0 | 2.224 |
| DEHP | | | | | 0 |

proliferators clofibrate is causing the least perturbation in the liver cell proteome. However, such comparisons may obscure important details on how each agent affects individual protein types. Thus, clofibrate may give an overall (average) least perturbation of the map but it introduces the largest change in abundance of protein I, while the other three chemicals hardly change the abundance of the protein I in proteome of liver rat cells exposed to them.

Clearly a single type of descriptor, such as are the components of vectors in 20-D space already mentioned, cannot capture the complexity of variations occurring even for relatively small maps such as a map considered here that is based on information on only 20 proteins. In the next section we will consider 3-D representation of the proteomics maps for control groups, PFOA, PFDA, clofibrate, and DEHP from which we will construct additional map descriptors.

## 3-D MAP MATRICES

By viewing normalized abundances shown in Table 8 as the third coordinate of a 3-D space, one can combine this information with the data on ($x$, $y$) coordinates of Table 3 and compute Euclidean distances for a graph embedded in 3-D space, of which Figure 4 now represents projection on the ($x$, $y$) plane. The nonzero matrix elements of the 3-D distance-adjacency graphs are listed in Table 10, where the alphabetic labels indicate line segments of the graph and the numerical labels indicate the corresponding matrix elements for graph considered. A comparison of Tables 3 and 10 for the control group shows some (but not large) changes for corresponding matrix elements. As expected, the new ($x$, $y$, $z$) distances between adjacent spots in 3-D map representation have somewhat increased compared to ($x$, $y$) distances of 2-D map when the third dimension is added to the graph. The changes in the abundance between the proteins extracted from control cells and cells treated with the four peroxisome proliferators introduced uneven changes in the matrix elements for the four corresponding matrices.

## THE LEADING EIGENVALUE OF MAP MATRICES

In the following we will consider the leading eigenvalues of the 3-D matrices as map descriptors. In Table 11 in the first numerical row we have listed the leading eigenvalue of the 3-D distance-adjacency matrices of proteomics maps for the normal (control) rat liver cells and those treated with the four peroxisome proliferators. To obtain additional map invariants we constructed the "higher order" map matrices, as outlined in ref 4 by considering the Kronecker products of each of the five matrices with itself. The leading eigenvalues of the matrices thus obtained were normalized using the reciprocal factorial ($1/n!$) as a normalization factor,

**Table 10.** Nonzero Matrix Elements of the Euclidean Distance-Adjacency Matrix for the 20 Most Intensive Protein Spots of Rat Liver Cells of the Normal Cells (Control) and the Corresponding Abundance for Rat Liver Cells Exposed to Four Chemicals

| line | | control | PFOA | PFDA | clofibrate | DEHP |
|---|---|---|---|---|---|---|
| AG | 1, 7 | 1.6516 | 1.6963 | 1.7152 | 1.7815 | 1.8484 |
| AI | 1, 9 | 1.8892 | 1.8403 | 1.8359 | 2.2323 | 1.9720 |
| AM | 1, 13 | 2.0524 | 1.9949 | 2.0257 | 2.1032 | 2.0801 |
| AP | 1, 16 | 1.9672 | 1.8852 | 1.8648 | 1.9902 | 1.9687 |
| BM | 2, 13 | 1.1590 | 1.2909 | 1.7458 | 1.3313 | 1.1531 |
| BO | 2, 15 | 1.1832 | 1.7222 | 1.9961 | 1.4202 | 1.0113 |
| BP | 2, 16 | 1.2201 | 1.3025 | 1.5041 | 1.3370 | 1.1737 |
| BQ | 2, 17 | 0.6685 | 1.3633 | 1.7672 | 1.2278 | 1.0324 |
| CG | 3, 7 | 0.6231 | 0.7960 | 1.2530 | 1.0717 | 0.8972 |
| CI | 3, 9 | 0.5325 | 0.4329 | 0.7676 | 1.4878 | 0.9254 |
| CR | 3, 18 | 1.6783 | 1.5966 | 1.8331 | 1.8914 | 1.0347 |
| CS | 3, 19 | 0.6810 | 0.4199 | 0.9254 | 1.1893 | 0.7336 |
| DM | 4, 13 | 0.4728 | 0.3290 | 0.3466 | 0.2148 | 0.2557 |
| DP | 4, 16 | 0.5146 | 0.2995 | 0.2189 | 0.2048 | 0.2337 |
| DR | 4, 18 | 0.5862 | 0.3150 | 0.2976 | 0.3026 | 0.4093 |
| EI | 5, 9 | 1.7408 | 1.7423 | 1.7604 | 1.9505 | 1.7887 |
| EJ | 5, 10 | 1.1357 | 1.1469 | 1.1157 | 1.1478 | 1.1463 |
| EO | 5, 15 | 0.7515 | 1.1687 | 0.7357 | 0.8238 | 0.6640 |
| ER | 5, 18 | 0.7196 | 0.6639 | 0.5557 | 0.7763 | 0.8276 |
| FK | 6, 11 | 0.3381 | 0.4022 | 1.1164 | 0.5813 | 0.6534 |
| FN | 6, 14 | 1.2557 | 1.5017 | 1.9074 | 1.4610 | 1.5231 |
| FP | 6, 16 | 0.7723 | 1.0955 | 1.5746 | 1.1342 | 1.0542 |
| GT | 7, 20 | 1.6399 | 1.5967 | 1.5993 | 1.5911 | 1.5841 |
| HQ | 8, 17 | 0.3303 | 0.1659 | 0.4080 | 0.1723 | 0.6499 |
| IT | 9, 20 | 1.5650 | 1.6282 | 1.6774 | 1.5836 | 1.5535 |
| JQ | 10, 17 | 0.6454 | 0.8874 | 0.8039 | 0.7526 | 0.8406 |
| JT | 10, 20 | 1.3162 | 1.3710 | 1.3760 | 1.3136 | 1.3488 |
| KM | 11, 13 | 0.8312 | 0.9985 | 1.0659 | 0.9334 | 0.8582 |
| KQ | 11, 17 | 1.7331 | 2.0429 | 1.9950 | 1.8813 | 1.8504 |
| LQ | 12, 17 | 0.3077 | 0.5830 | 0.7216 | 0.7340 | 0.6932 |
| MN | 13, 14 | 1.9349 | 1.9348 | 1.9564 | 1.9390 | 1.9368 |
| NQ | 14, 17 | 2.8498 | 2.8990 | 2.9151 | 2.8865 | 2.8708 |
| NS | 14, 19 | 0.6224 | 0.6151 | 0.6298 | 0.7215 | 0.6226 |
| OT | 15, 20 | 0.6118 | 0.6370 | 0.5988 | 0.5995 | 0.9815 |
| PS | 16, 19 | 1.3240 | 1.3217 | 1.3363 | 1.3537 | 1.3411 |
| RT | 18, 20 | 0.5758 | 0.6921 | 0.6817 | 0.5750 | 0.5781 |

**Table 11.** Leading Eigenvalue of the Euclidean Distance-Adjacency Matrix E and the Higher Order Matrices Obtained by Kronecker Product of a Matrix E by Itself

| power | control | PFOA | PFDA | clofibrate | DEHP |
|---|---|---|---|---|---|
| 1 | 5.1421 | 5.4972 | 6.0381 | 5.6998 | 5.4203 |
| 2 | 5.0141 | 5.4730 | 6.0599 | 5.4770 | 5.2738 |
| 3 | 4.1788 | 4.5492 | 4.8440 | 4.4267 | 4.3376 |
| 4 | 2.8409 | 3.0998 | 3.2291 | 3.0112 | 2.9445 |
| 5 | 1.5887 | 1.7483 | 1.8098 | 1.6985 | 1.6526 |
| 6 | 0.7486 | 0.8340 | 0.8642 | 0.8093 | 0.7832 |
| 7 | 0.3037 | 0.3433 | 0.3571 | 0.3323 | 0.3199 |
| 8 | 0.1080 | 0.1241 | 0.1297 | 0.1197 | 0.1146 |
| 9 | 0.0342 | 0.0399 | 0.0419 | 0.0384 | 0.0365 |
| 10 | 0.0097 | 0.0116 | 0.0122 | 0.0111 | 0.0105 |

where $n$ is the power to which matrices were raised. The normalized leading eigenvalues listed in Table 11 show a fast convergence. Already when $n = 10$ they practically approach zero. We can view columns in Table 11 as vectors in 10-dimensional vector space. In Table 12 entries above the main diagonal represent the similarity/dissimilarity matrix between the five vectors corresponding to the control and the treated animals. Some similarities and differences can be observed when Table 12 is compared to Table 9 in which the similarity/dissimilarity between the five proteomics maps was based solely on the relative abundances of the leading 20 spots of the control map.

**Table 12.** Similarity/Dissimilarity Table between the Control Data and Data for the Four Chemicals[a]

| | control | PFOA | PFDA | clofibrate | DEHP |
|---|---|---|---|---|---|
| control | 0 | 0.7586 | 0.8635 | 0.8318 | 0.3671 |
| PFOA | 25.88 | 0 | 1.5985 | 0.2591 | 0.3562 |
| PFDA | 85.69 | 52.86 | 0 | 0.8302 | 1.1705 |
| Clofibrate | 44.63 | 18.83 | 41.23 | 0 | 0.3671 |
| DEHP | 19.34 | 6.56 | 66.51 | 25.33 | 0 |

[a] Entries above the main diagonal: based on the vectors using the leading eigenvalue of the Euclidean distance-adjacency matrix and the higher order matrices obtained by Kronecker product of a matrix. Entries below the main diagonal: based on the vectors using the leading eigenvalue of the distance-adjacency matrix and the higher order matrices obtained by use of the standard product of a matrices.

**Table 13.** Leading Eigenvalue of the Euclidean-Adjacency Matrix and the Higher Order Matrices Obtained by Use of the Standard Product of a Matrix by Itself

| power | control | PFOA | PFDA | clofibrate | DEHP |
|---|---|---|---|---|---|
| 1 | 5.1421 | 5.4972 | 6.0381 | 5.6998 | 5.4203 |
| 2 | 13.2204 | 15.1094 | 18.2290 | 16.2437 | 14.6897 |
| 3 | 22.6600 | 27.6864 | 36.6893 | 30.8618 | 26.5409 |
| 4 | 29.1298 | 38.7149 | 55.3830 | 43.9762 | 35.9647 |
| 5 | 29.9574 | 41.8326 | 66.8810 | 50.1309 | 38.9878 |
| 6 | 25.6738 | 38.3268 | 67.3052 | 47.6224 | 35.2209 |
| 7 | 18.8595 | 30.0984 | 58.0560 | 38.7767 | 27.2724 |
| 8 | 12.1221 | 20.6820 | 43.8182 | 27.6273 | 18.4780 |
| 9 | 6.9258 | 12.6325 | 29.3974 | 17.4966 | 11.1285 |
| 10 | 3.5613 | 6.9433 | 17.7503 | 9.9726 | 6.0319 |
| 11 | 1.6648 | 3.4704 | 9.7434 | 5.1674 | 2.9723 |
| 12 | 0.7134 | 1.5898 | 4.9026 | 2.4544 | 1.3425 |
| 13 | 0.2822 | 0.6722 | 2.2771 | 1.0761 | 0.5598 |
| 14 | 0.1036 | 0.2640 | 0.9821 | 0.4381 | 0.2167 |
| 15 | 0.0036 | 0.0967 | 0.3953 | 0.2665 | 0.0783 |

In addition to the Kronecker product of matrices (with themselves), we have also considered the standard product of matrices (with themselves). In Table 13 we show the leading eigenvalue of the corresponding "higher order" matrices. The first row in both Tables 11 and 13 is the same, but as we raise matrices to the higher powers using the standard matrix product we obtain larger matrix elements and consequently larger leading eigenvalues. We can view the columns of Table 13 as components of vectors in 15-dimensional vector space for which we can calculate the degree of mutual similarity/dissimilarity, which are shown as entries below the main diagonal in Table 12. The leading eigenvalue as well as other Kronecker and standard matrix manipulations were calculated using MATLAB software.[28]

## "FOX TRAIL" APPROACH

The zigzag curve introduced in the seminal work on quantification of proteomics maps[1−3] was based on selecting *n* most abundant spots and ordering them with respect to relative abundance. The resulting zigzag graph besides offering numerous invariants based on the leading eigenvalue of the D/D matrix as discussed in refs 1−3 allows one to construct additional, so-called, "band invariants" that were recently introduced by Randić and Balaban.[29] A possible disadvantage of the zigzag curve is that its form may be sensitive to fluctuations in the relative abundance as reported in different maps belonging to the same proteome. This shortcoming has been recently addressed by us[6] in which we introduced canonical labels for protein spots that are based on 2-D map and do not require information on

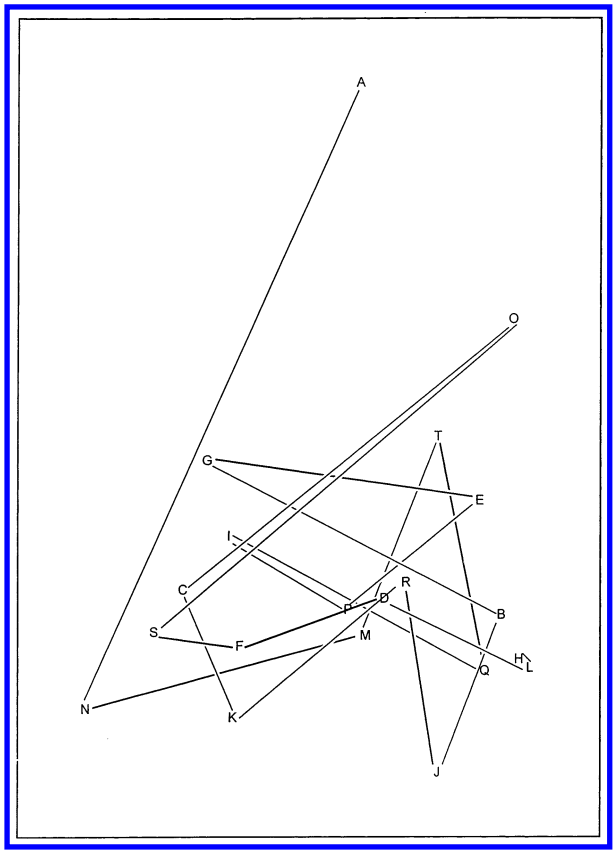


**Figure 5.** The "fox trail", the zigzag curve based on labels derived from two sums of the Euclidean-adjacency matrix E.

abundance. After arriving at canonical labels for protein spots, one simply connects spots using canonical labels as guidance. To differentiate such zigzag curves from the zigzag curve based on relative abundance we referred to the novel zigzag curve as "fox trail", in view of a great similarity of zigzag curves and the experimentally observed paths of a wandering fox recorded at constant time intervals.[30,31]

We will here outline an alternative "fox trail" associated with the proteomics map of Figure 1 and illustrated in Figure 5. Each time we assign numerical labels to a proteomics map, we can derive a zigzag curve by connecting spots having neighboring labels. The problem is in formulating useful rules that lead toward a *unique* labeling. Advantage of the canonical labels based on the smallest binary code for the embedded graph of the partial ordering is not only that it is unique, but once a set of spots is selected as a basis for quantitative characterization of a map, it offers canonical labels for a map as a whole and allows cataloguing and ordering such maps in a systematic way. A disadvantage of canonical labeling is that it is associated with an nonpolynomial (NP) algorithm, that is, finding canonical labels represent an NP problem,[32] which may become computationally intensive and even intractable as the number of spots grows larger.

We propose here to use row sums of the distance-adjacency matrix (listed in Table 4) as the criterion for labeling of vertices of Euclidean-adjacency graphs. In this way we bypassed the difficulties of NP associated with canonical labels and nevertheless obtain a "fox trail" graph that does not require relative abundance of proteomics spots for its construction, except for the *selection* of the 20 (or

whatever number is selected) spots used as a basis of mathematical analysis of proteomics map. In Figure 5 we illustrated the "fox trail" graph for graph of Figure 4.

## DISCUSSION

It is of interest to make a closer comparison between the three similarity/dissimilarity tables, Table 9, and both parts of Table 12. First, observe that there is more parallelism between the corresponding entries in Table 12 than those of Table 9, but even here there are a number of important differences. Both tables were based on the leading eigenvalues of the higher order map matrices, the difference being in the kind of matrix multiplication considered. Hence, it appears that computation of the Kronecker product of matrices and use of leading eigenvalues of the higher order Kronecker matrices as map descriptors, (which is simpler and faster), could not replace the standard matrix multiplication for construction of the higher order matrices, which offer somewhat different map descriptors.

The similarity/dissimilarity tables such as in Tables 9 and 12 do show some differences that illustrate the fact, that even though two approaches are based on similar methodologies, they capture somewhat different map information. For example, the smallest entry in Table 9 belongs to the pair (PFOA, PFDA), which is not the case with Table 12 where the smallest entry is associated with the pair (PFOA, clofibrate) and (PFOA, DEHP) for the 10-D and 15-D vectors, respectively. A relatively small value of the Euclidean distance in a similarity/dissimilarity table does not necessarily imply that the corresponding chemicals are the most similar, but it signifies that the effect of those chemicals on the 20 proteome spots of treated cells may be similar. The small amount of data presented may be merely anecdotal, and it would be premature to conclude that because of some parallelism Kronecker multiplication and standard matrix multiplication behave similarly in general. Future applications may clear this point; at this stage we are merely offering a tool for such applications.

Based on the entries in the first row of Table 9 one could conclude that clofibrate makes the least perturbation of the proteome of rat liver cells, but from the first row of Table 12 and the first column of Table 12 we see that DEHP apparently makes the least perturbation of the proteome of rat liver cells. The difference between these approaches is that in the first case, only changes in abundance are considered as a measure of perturbation; but in the other, the location of protein spots i.e., the charge and mass of proteins, makes an influence. Changes in abundance of protein spots which are in the vicinity (i.e., proteins having a similar mass and charge) will be more pronounced in the latter case.

It remains to be seen which graph invariants and what kind of weighting procedure may be more suitable when comparing different proteomics maps. At this stage it is more important to recognize the various possibilities, and it would be premature to suggest or to advocate one scheme over another. We should recollect that here we chose to consider a particular set of invariants, the leading eigenvalues of 3-D matrices as invariants, but other graph invariants ought to be examined. The main reason for the lack of definite conclusions with regard to one set of descriptors relative to another at this moment is the lack of suitable data for testing. The comparison of similarity/dissimilarity among proteomics maps with similarities/dissimilarities of chemical agents causing perturbations would be facilitated by having data on a set of chemical compounds showing closer structural features. If there is some parallelism between the biological descriptors of proteomics maps and chemodescriptors inducing changes in proteome, one could test the degree to which the paradigm that similar chemical structures induce similar properties, that holds in QSPR and QSAR, also applies to the proteome, even if proteomics data may point to drastic changes of the cell proteome under minor perturbations of biological systems. Another testing ground for the utility of biodescriptors derived from proteomics maps could be the study of perturbations of the cellular proteome induced by a gradual increase of concentrations of chemicals that produce visible perturbations of the proteome. Hopefully, as indicated in ref 22 for the first time, one may use a combination of chemodescriptors and biodescriptors derived from proteomics maps in an effort to arrive at an integrated QSAR.

## CONCLUDING REMARKS

Here we have developed two measures of "proteomics similarity" of the chemicals, one based solely on the abundances of the spots and the other based on the Euclidean distance in the eigenvalue space. These two measures give different pictures of the intermolecular similarity of the chemicals analyzed in this paper. This brings us to the well-known problem of similarity methods: Similarity methods are not uniquely defined, they are rather user-defined and context specific.[33,34] In the area of molecular similarity for relatively small molecules, Basak et al.[35−37] calculated and used various molecular similarity methods based on graph invariants as well as experimental properties to select analogues, cluster large libraries of chemicals,[38] and estimate properties of chemicals from their selected neighbors[39,40] in the selected spaces. It was concluded that no universal similarity method can handle all situations. It would be interesting to see how far map invariants are able to characterize various aspects of proteomics maps and their perturbation derived from the exposure of organisms/cells to drugs and xenobiotics.

## ACKNOWLEDGMENT

This is contribution number 315 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported in part by Grant F49620-02-1-013 from the United States Air Force Office of Scientific Research.

## REFERENCES AND NOTES

(1) Randić, M. On Graphical Representation of Proteomics and Their Numerical Characterization. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 1330−1338.

(2) Randić, M.; Zupan, J.; Nović, M. On 3-D Graphical Representation of Proteomics Maps. *Chem. Inf. Comput. Sci*. **2001**, *41*, 1339−1334.

(3) Randić, M.; Witzmann, F.; Vraćko, M.; Basak, S. C. On Characterization of Proteomics Maps and Chemically Induced changes in Proteome Using Matrix Invariants: Application to Peroxisome Proliferators. *Med. Chem. Rev.* **2001**, *10*, 456−479.

(4) Randić, M. A Graph Theoretical Characterization of Proteomics Maps. *Int. J. Quantum Chem.* In press.

(5) Randić, M.; Nović, M.; Vraćko, M. On Characterization of Dose Variations of 2-D Proteomics Maps by Matrix Invariants. *J. Proteome Res.* In press.
(6) Randić, M.; Basak, S. C. Canonical Labeling of Proteins and Proteomics Maps. *J. Chem. Inf. Comput. Sci.* Submitted for publication.
(7) Witzmann, F. Molecular Anatomy Laboratory, Department of Biology, Indiana University & Purdue University, Columbus, IN.
(8) Randić, M.; Wilkins, C. L. On a Graph Theoretical Basis for Ordering of Structures. *Chem. Phys. Lett.* **1979**, *63*, 332−336.
(9) Randić, M.; Wilkins, C. L. Graph Theoretical Ordering of Structures as a Basis for Systematic Searches for Regularities in Molecular Data. *J. Phys. Chem.* **1979**, *83*, 1525−1540.
(10) Randić, M.; Jerman-Blažic, B.; Rouvray, D. H.; Seybold, P. G.; Grossman, S. C. The Search for Active substructure in Structure−Activity Studies. *Int. J. Quantum Chem: Quantum Biol. Symp.* **1987**, *14*, 245−260.
(11) Randić, M. *Acta Chim. Slovenica* **2000**, *47*, 143.
(12) For a recent applications of partial order in chemistry, see: Partial Orderings in Chemistry. *MATCH* (*Mathematical Chemistry Communication*); Klein, D. J., Brickmann, J., guest Eds.; 2000; No. 42.
(13) Randić, M. Unpublished.
(14) Kowalski, B. R.; Bender, C. F. A Powerful Approach to Interpreting Chemical Data. *J. Am. Chem. Soc.* **1972**, *94*, 5632−5639.
(15) Randić, M. Topological Indices. In *The Encyclopedia of Computational Chemistry;* Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III., Schreiner, P. R., Eds.; John Wiley & Sons: Chichester, 1998; pp 3018−3032.
(16) Randić, M.; DeAlba, L. M. Dense Graphs and Sparse matrices. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 1078−1081.
(17) Harary, F. *Graph Theo*ry; Addison-Wesley: Reading, MA, 1969.
(18) Randić, M.; Kleiner, A. F.; DeAlba, L. M. Distance/Distance matrices. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 277−286.
(19) Randić, M. On Characterization of Molecular Branching *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.
(20) Kier, L. B.; Murray, W. J.; Randić, M.; Hall, L. H. Molecular Connectivity V. Connectivity Series Concept Applied to Density. *J. Pharm. Sci.* **1976**, *65*, 1226−1230.
(21) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. QSPR: The Correlation and Quantitative Prediction of Chemical and Physical Properties from Structure. *Chem. Soc. Rev.* **1995**, 279−287.
(22) Basak, S. C.; Mills, D.; Gute, B. D.; Grunwald, G. D.; Balaban, A. T. Applications of Topological Indices in the Property/Bioactivity/Toxicity Prediction of Chemicals. In *Topology in Chemistry: Discrete Mathematics of Molecules*; Rouvray, D. H., King, R. B., Eds.; Horwood Publishing Ltd.: Chichester, West Sussex, United Kingdom, 2001; pp 113−184.
(23) Randić, M. Novel Graph Theoretical Approach to Heteroatom in Quantitative Structure−Activity Relationships. *Chemometrics Intel. Lab. Systems* **1991**, *10*, 213−227.
(24) Randić, M. On Computation of Optimal Parameters for Multivariate Analysis of Structure−Property Relationship. *J. Comput. Chem.* **1991**, *12*, 970−980.
(25) Randić, M.; Dobrowolski, J. C. Optimal Molecular Connectivity Descriptors for Nitrogen-Containing Molecules. *Int. J. Quantum Chem.* **1998**, *70*, 1209−1215.
(26) Randić, M.; Mills, D.; Basak. S. C. On Use of Variable Connectivity Index for Characterization of Amino Acids. *Int. J. Quantum Chem.* **2000**, *80*, 1199.
(27) Randić, M.; Basak, S. C. In *Some Aspects of Mathematical Chemistry*; Sinha, D. K., Basak, S. C., Mohanty, R. K., Basumallick, I. N., Eds.; Visva-Bharati University Press: In press.
(28) MATLAB (abbreviation for Matrix Laboratory) is a product of The MathWorks, Inc., 3 Apple Hill Drive, Natick, MA 01760.
(29) Randić, M.; Balaban, A. T. 4-Dimensional DNA *Proc. Natl. Acad. Sci. U.S.A.* Submitted for publication.
(30) Siniff, D. H.; Jesson, C. R. Simulation Model of Animal Movement Patterns. *Adv. Ecological Res.* **1969**, *6*, 185−220.
(31) Hall, G. G. Modelling − A Philosophy for Applied Mathematicians. *Bull. Institute Mathematics Applications* **1972**, *8*, 226−228.
(32) Garey, M. R.; Johnson, D. S. *Computers and Intractability (A Guide to the Theory of NP-Completeness);* Freeman: San Francisco, 1979.
(33) Randić, M. Design of Molecules with Desired Properties. Molecular Similarity Approach to Property Optimization. In *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G., Eds.; Wiley: New York, 1990; pp 77−145.
(34) Randić, M. Similarity Methods of Interest in Chemistry. In *Mathematical Methods in Contemporary Chemistry*; Kuchanov, S., Ed.; Gordon & Breach Publ. Inc.: New York, 1995; pp 1−99.
(35) Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R. Determining Structural Similarity of Chemicals Using Graph-Theoretic Indices. *Discrete Appl. Math.* **1988**, *19*, 17−44.
(36) Basak, S. C.; Grunwald, G. D. Use of Topological Space and Property Space in Selecting Structural Analogues. *Math. Modelling Sci. Computing* **1994**, *4*, 464−469.
(37) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Development and Applications of Molecular Similarity Methods Using Nonempirical Parameters. *Math. Modelling Sci. Computing* In press.
(38) Basak, S. C.; Mills, D.; Gute, B. D.; Balaban, A. T.; Basak, K.; Grunwald, G. D. Use of Mathematical Structural Invariants in Analysis Combinatorial Libraries: A Case Study with Psoralen Derivatives. In *Some Aspects of Mathematical Chemistry*; Sinha, D. K., Basak, S. C., Mohanty, R. K., Basumallick, I. N., Eds.; Visva-Bharati University Press: In press.
(39) Gute, B, D.; Basak, S. C. Molecular Similarity-Based Estimation of Properties: A Comparison of Three Structure Spaces. *J. Mol. Graphics* **2001**, *20*, 95−109.
(40) Basak, S. C.; Grunwald, G. D.; Host, G. E.; Niemi, G. J.; Bradbury, S. P. A Comparative Study of Molecular Similarity, Statistical, and Neural Methods for Predicting Toxic Modes of Action. *Environ. Toxicol. Chem.* **1998**, *17*, 1056−1064.

CI0100797