On Graphical and Numerical Characterization of Proteomics Maps

Milan Randić[†]

National Institute of Chemistry, Ljubljana, Slovenia

Received January 3, 2001

We outlined a mathematical approach suitable for characterization of experimental data given by 2-D densitograms. In particular we consider numerical characterization of proteomics maps. The basis of our approach is to order “spots” of a 2-D map and assign them unique labels (that in general will depend on the criteria used for ordering). In this way a map is “translated” into a sequence. In the next step one associates with the generated sequence a geometrical path and views such a path as a mathematical object that needs characterization. We have ordered spots representing proteins in 2-D gel plates according to their relative intensities which results in a zigzag path that produces a complicated “fingerprint” pattern. Mathematical characterization of zigzag pattern follows similar mathematical characterizations of embedded patterns based on matrices, the elements of which are given as quotients of Euclidean distance between spots and the distance along the zigzag path. The leading eigenvalue of constructed matrices is taken to represent characterization of the original 2-D map. Comparison of different 2-D maps (simulated by using random generator) allows one to construct partial order, which although qualitative in nature gives some insight into perturbation induced by foreign agents to the proteome of the control cell.

When you can measure what you are speaking about, and express it in numbers, you know something about it. But when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager unsatisfactory kind...

Lord Kelvin[1]

INTRODUCTION

Combinatorial chemistry combined with throughput methodology and automated DNA sequencing has resulted in an overwhelming abundance of raw data that appears to characterize this period of transition of science to a new millennium. To take advantage of the situation a need for novel and imaginative approaches for data processing and data reduction is apparent. Any novel path to data digesting deserves attention, particularly if it is not computer intensive. Dirac’s dictum[2] that with the rise of quantum mechanics most of physics and all of chemistry is understood, the difficulties remaining in solving corresponding equations is of little comfort when considering biological complexities. That indeed novel approach to a complex and combinatorially rich problem may have promise which is well illustrated in a recent study of Lahana and co-workers[3] on a rational synthesis of biologically active compounds. By using a dozen molecular descriptors, in a way these descriptors have not been used before, these authors succeeded to screen a combinatorial library of 280 000 virtual dipeptides and come up with a short list of a dozen potentially interesting compounds. Additional more thorough screening of so selected compounds resulted in preparation of a compound with almost a 100 times higher immunosuppressive activity than the leading peptide on which the search was based.

In this article we will address a problem of a considerable complexity—a quantitative characterization of experimental data given as 2-D assays, such as proteomics maps or 2-D spectroscopic essays. In particular we will focus on proteomics maps. A need for quantitative analysis of proteomics maps is apparent—as only a quantitative study may offer, if not a better understanding, at least better characterization and possibility of predictability of complex phenomena that typify biological systems. Proteomics data are reported as tables of x, y coordinates of protein spots and their abundance on 2-D gels in which proteins are separated by mass and charge (Table 1). Such tabular data can be transformed into computer generated proteomics “bubble” map (Figure 1) which offers a better visual impression than original gel photographs in which the abundance of individual spots are shown as “bubbles” of different radius. This may facilitate visual inspection but essentially duplicate the information rather than analyze it. In contrast we are interested in analysis of data given by proteomics maps that will result in data reduction. Our goal is to arrive at a set of biodescriptors for proteomics maps, which can be used for comparison of different maps and even for relating changes in cell proteome induced by specific agents and toxic substances.

A GRAPHICAL REPRESENTATION OF PROTEOMICS MAPS

Proteomics map can be viewed as a summary of graphical output of experiments using electrophoretic and chromatographic techniques. The map shown in Figure 1, which came from the laboratory of Frank Witzmann,[4] is a typical illustration of data that we want to analyze and reduce to a mathematical object. We are interested in alternative graphical representations and suitable numerical characterizations of such maps. A map that is represented by an alternative mathematical format may facilitate comparison with other

[†] Corresponding author fax: (515)292-8629; e-mail: milan.randic@ki.si. Current address: 3225 Kingman Rd., Ames, IA 50311.

10.1021/ci000167b CCC: $20.00
Published on Web 07/14/2001

Table 1. Coordinates and the Relative Abundance of Protein Spots for a Portion of the Proteomics Map of Figure 1

n	*x*	*y*	*A*	*n*	*x*	*y*	*A*
1	21	23	144.4	11	25	4	93.6
2	28	9	143.6	12	16	6	90.0
3	12	10	136.7	13	30	8	86.7
4	22	9	127.2	14	21	8	84.8
5	27	12	118.6	15	6	7	82.5
6	15	8	114.9	16	29	19	82.0
7	13	14	112.3	17	20	9	80.0
8	29	8	108.9	18	28	8	79.8
9	14	11	98.2	19	23	10	72.8
10	26	13	94.1	20	11	9	72.2

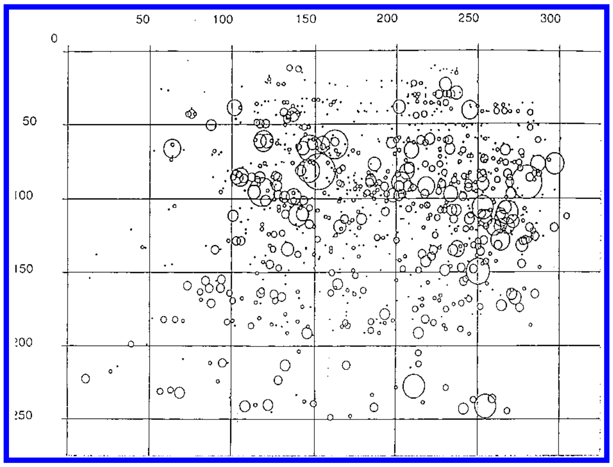

Figure 1. An illustration of a proteomics map with spots represented as "bubbles".

maps similarly represented. The question is how can one arrive at an alternative representation of data that is suitable for mathematical characterization.

A parallel with representation of data on primary DNA sequences is here of some interest. DNA data are typically given by long list of four letters A, C, G, T representing the four nucleic acid bases (adenine, cytosine, guanine, and thymine, respectively). To facilitate comparison of DNA sequences Nandy[5] and others[6−8] have developed virtual graphical (geometrical) representation of DNA sequences, which allow one to visualize differences among long sequences of nucleic acid bases. Recently a scheme was outlined which allows these graphical representations of DNA primary sequences to be transformed into numerical data. In the first step one constructs numerical matrices to be associated with DNA sequence, and then one selects a set of suitable matrix invariants as a characterization of DNA sequence.[9−13] In this way a tedious and computer intensive direct comparison of DNA sequences is replaced by comparison of a relatively short list of mathematical invariants associated with sequences. Can a proteomics map be associated with a numerical matrix and then reduced to a similar characterization by a set of invariants?

We will outline an approach that accomplished just that. As we will see we will first from a map construct a sequential list of proteins, and then we will construct a matrix associated with the sequence. From the matrix we will extract a list of invariants that will allow for fast comparison of different proteomics maps. To illustrate the approach we have selected a smaller section of Figure 1, the region with coordinates $500 < x < 1000$ and $1500 < y < 2000$, schematically illustrated in Figure 2. We have reduced the scale to arbitrary units so that the fragment of a map corresponds to a 35 × 30 grid. The input data for 20 spots with the highest abundance are summarized in Table 1. The problem that we wish to consider is how to arrive at a quantitative characterization of Figure 2. There are several ways as to how a map (set of points in a 2-D plane) can be quantitatively described. One can divide the plane into regions using the notion of Voronei polyhedra and then connect those "vertices" (spots) which correspond to regions that have a common edge. In this way one arrives at a graph (embedded in the plane), the adjacency and the distance matrix of which offer numerical manipulations of experimental data. Alternatively one can consider various triangulation of the map which also leads to graphs embedded in a planar. In subsequent analysis one considers various invariants of so embedded graph. One can also consider superimposition of an n × n coordinate grid over a map and view the experimental abundance values as the entries of so designed n × n matrix. Each such approach will be accompanied with some loss of information, and it remains to be seen whether such schemes will offer useful characterization of experimental data given by 2-D maps.

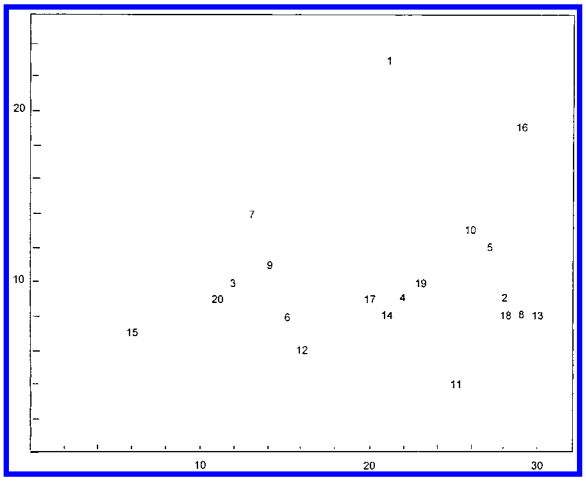

Figure 2. Schematic representation of a portion of the "bubble" map of Figure 1 showing the location of the 20 most intensive spots.

Instead of these geometrically well defined approaches we will consider a somewhat "unusual" approach by focusing attention on the path between suitably selected spots of a map. First we will order spots by assigning to them labels that will rank spots relative to their abundance giving to the most abundant protein spot label 1. In Figure 2 we show locations belonging to each of the 20 most intense protein spots. In the next step we connect points having adjacent numerical labels which results in a somewhat complicated path which overlaps itself many times (Figure 3). We will refer to the zigzag path as the graphical "fingerprint" pattern of the proteomics map. Clearly different proteomics maps will have distinct "fingerprints" because an abundance of proteins and proteins themselves that occur will vary from case to case. The zigzag line embedded in a plane in which spot labels are replaced by abundance fully represents a proteomics map, that is, there is a 1:1 correspondence between a map and its graphical "fingerprint". The problem of the characterization of proteomics map becomes the problem of the characterization of associated graphical "fingerprint" patterns.

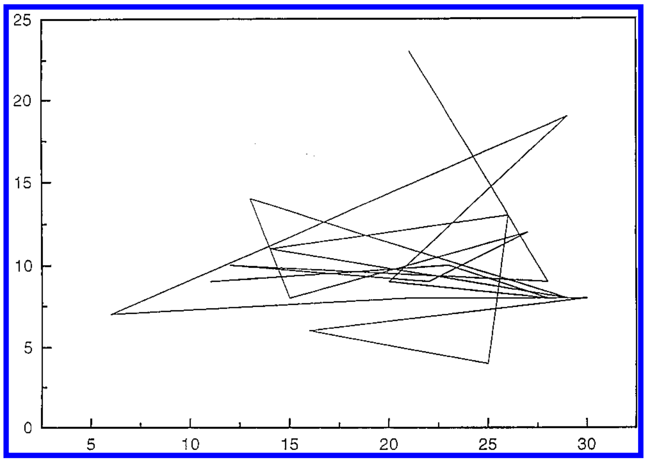

Figure 3. Zigzag line connecting the first 20 most intensive spots of Figure 2.

ON CHARACTERIZATION OF 2-D MAP "FINGERPRINT" PATTERN

The "fingerprint" zigzag path offers a novel graphical representation of proteomics map. Numerical characterization of embedded zigzag path representing the map "fingerprint" pattern is mathematically related to characterization of mathematical curves embedded in 2-D and 3-D space,[14-16] molecular skeletons embedded on a graphite,[17-20] diamond lattice, or 3-D space,[20] and extension of such computations for model proteins embedded in 3-D plane.[21] Selected points of such geometrically rigid structures can be considered as vertices of an embedded graph. For such structures one constructs the so-called D/D matrix[17] in which matrix elements are given as quotient of distances between a pair of vertices measured through the space and along the bonds. Because we consider quotients of distances one can measure distance in arbitrary units. There is a difference between the D/D matrices of embedded graphs and the D/D matrices here considered embedded zigzag paths. Now the line sections are of unequal length, while in the case of graphs all line segments (edges) had the same length. Unequal line segments, however, neither introduce conceptual nor computational difficulties, as will be seen from the outline of such characterization that follows.

Our goal is to associate with the embedded zigzag path of Figure 3 a D/D matrix constructed by using the Euclidean distance (measured through the 2-D space) and the path distance (measured along the path). A quick way to construct a D/D matrix is to measure all the distances directly from the map, such as is Figure 3. For example, the Euclidean distances between vertices shown in the numerators of the quotients shown in Table 2, listed only for the 10 leading vertices, were obtained by measuring distances of Figure 3 in millimeters. From a table of Euclidean distances it is straightforward to construct a table of distances as measured along the zigzag path for any pair of points along the path. These are obtained by the successive adding of distances of adjacent points. The distances measured along the path are shown in denominators of the quotients shown in Table 2. The elements of the D/D matrix, shown in Table 2, are obtained as quotients of the corresponding distances.

Observe that in the construction of the D/D matrix we have not used directly the numerical information on the abundance, but have used this information only indirectly when we assigned the labels to individual spots. At the first it may look as though we have not taken sufficient information from a map, but in the view that there are so many spots that almost continually vary in their abundance the information on the relative abundance (which prescribe numerical labels for protein spots ordering of spots) appears to hold enough information of interest. Besides, such a choice offers some robustness to the approach, which is not necessarily so sensitive to minor variations in the reported abundance. As is known proteomics maps from different laboratories vary somewhat due to different experimental conditions and a robust approach is almost a necessity in such circumstances.

D/D MATRIX INVARIANTS

Now the task is to extract form the D/D matrix a set of invariants that characterize such matrix. This problem has been considered in chemical graph theory[22] where D/D matrices served as a source for novel molecular descriptors. We will apply here essentially the same methodology in order to arrive at a set of structural invariants for proteomics maps. In the case of linear chains embedded in the 2-D or 3-D space the leading eigenvalue (the largest positive eigenvalue of a matrix) of the D/D matrix has been interpreted as a measure of the degree of folding or bending of such a chain in a plane or the space.[17,21] This interpretation is empirical in origin and followed from comparisons of the leading eigenvalues of D/D matrices of smaller carbon atom chains embedded on graphite lattice. Extension of such studies to a selection of mathematical curves[14,15] showed that indeed the interpretation is essentially correct. This can be understood intuitively by recognizing that each "sharp" bend of the zigzag path considerably reduces the Euclidean distance between points, while leaving the distances measured along the path intact. As a result the corresponding matrix elements decrease in magnitude, and consequently the corresponding row sums of the D/D matrix become smaller. According to the Frobenius-Perron theorem,[23] which states that the smallest row (or column) sum and the largest row (or column) sum of a matrix represent the lower and the upper bound on the leading eigenvalue, this has as a consequence a decrease of the magnitude of the leading eigenvalue for a chain that has a large number of locally bent sites.

We have selected the leading eigenvalue of D/D matrix as the invariant to characterize the map. Use of the leading eigenvalue for characterization of smaller molecules, including "toy" proteins of Tang and co-workers,[24] has shown that this invariant is sufficiently sensitive to minor changes in molecular structure. But clearly complex and complicated proteomics maps need more than a single invariant for their characterization. Rather than resorting to other possible map invariants we will construct additional descriptors of a map by considering the leading eigenvalues the "higher" order D/D matrices.

THE HIGHER ORDER D/D MATRICES

The "higher" order D/D matrices are derived from D/D matrix by considering higher powers of the individual matrix elements. We designated the higher order D/D matrices as ${}^{m}D/{}^{m}D$ where m is positive integer different from 1. In the past applications typically m has taken values from 2 to 20. Matrix element (i, j) of ${}^{m}D/{}^{m}D$ is obtained from the

Table 2. Part of the D/D Matrix for the Ten Most Intense Spots of the Proteomics Map of Figure 2

	1	2	3	4	5	6	7	8	9	10
1	0	84/84	87/180	78/240	68/274	86/349	68/383	92/483	76/574	60/647
2		0	96/96	35/156	16/190	77/265	93/299	8/399	84/490	23/563
3			0	60/60	91/94	21/169	22/203	102/303	18/394	86/467
4				0	34/34	42/109	60/143	42/243	49/334	32/407
5					0	75/75	85/109	24/209	78/300	8/373
6						0	34/34	83/134	17/225	71/298
7							0	100/100	16/191	78/264
8								0	91/91	31/164
9									0	73/73
10										0

Table 3. Leading Eigenvalue of the D/D Matrix and the Higher Order ${}^mD/{}^mD$ Matrices for Portions of the Map of Figure 2 for Cases When an Increasing Number of Spots Is Considered

	10 × 10	15 × 15	20 × 20	20 × 20[a]
1	3.71221	4.37802	4.82401	4.83627
2	2.81558	2.91190	3.00714	3.01532
3	2.55679	2.57280	2.60174	2.60380
4	2.44338	2.44866	2.47019	2.47133
5	2.37923	2.38232	2.42433	2.42348
6	2.33645	2.33892	2.40803	2.40464
7	2.30460	2.30700	2.40075	2.39548
8	2.27909	2.28185	2.39697	2.39047
9	2.25766	2.26158	2.39478	2.38755
10	2.23906	2.24699	2.39336	2.38579
11	2.22255	2.24065	2.39237	2.38469
12	2.20764	2.23859	2.39162	2.38400
13	2.19402	2.23768	2.39102	2.38356
14	2.18145	2.23719	2.39052	2.38328
15	2.16977	2.23688	2.39009	2.38309
16	2.15886	2.23669	2.38971	2.38298
17	2.14861	2.23655	2.38936	2.38290
18	2.13895	2.23646	2.38905	2.38285
19	2.12983	2.23639	2.38876	2.38905
20	2.12118	2.23634	2.38849	2.38279

[a] The last column belongs to a modified matrix (see text).

corresponding matrix element of D/D matrix elements such as $(i, j)^m$. Such a matrix is a special case of the procedure in matrix algebra known as Kronecker's product $A \wedge B$ of two matrices, which when $A = B = D/D$ gives ${}^2D/{}^2D$. Users of MATLAB[25] will find this procedure available among possible matrix manipulations.

We have constructed D/D matrices for the first 10, then for the first 15, and finally for the first 20 most abundant proteins of Figure 2. In Table 3 we have listed the corresponding leading eigenvalues for $m = 1-20$. The leading eigenvalues decrease with increasing m, as expected, showing a slow but definite convergence. To indicate the convergence of the leading eigenvalues as m tends to infinity we show in Table 4 the leading eigenvalue for a selection of large values of m. The convergence is to be expected because as m increases the individual elements of ${}^mD/{}^mD$ matrices continually decrease and in the limit assume the value zero. In the limit the ${}^mD/{}^mD$ matrix becomes a binary matrix representing the adjacency matrix of a chain of length n. The leading eigenvalue of the adjacency matrix of a chain of length n is given as $2\cos[\pi/(n+1)]$, which we approach already for $m = 1000$ to 12 digits.

SENSITIVITY ANALYSIS

It is desirable that a matrix associated with a map is not only sensitive to minor variations in a map but also that different maps do not produce the same numerical values for matrix invariants. It is well-known in chemical graph theory[22] that many topological indices show degeneracy (i.e., duplicate values) already for relatively small graphs. To explore the sensitivity of the proposed "fingerprint" patterns and associated D/D matrices we have reexamined the input data of Table 1 and have permuted the labels for the spots 17 and 18, which show a minor difference in their relative abundance. Hence, we assumed that spot 18 (with abundance 79.8) is to be assumed to have a larger abundance. If experimental error in the estimate of the abundance would be ±0.2 this could easily be the case. Equally, somewhat different experimental conditions in different laboratories can cause the same situation. As a result of the permutation of the relative order of two spots we obtain a zigzag pattern shown in Figure 4, which differs but slightly from the corresponding zigzag pattern of Figure 3. The difference is contained only within the triangle with vertices at the points 16, 17, and 18. In Figure 3 we used the sides 16−17 and 17−18 of this triangle, while in Figure 4 we used the sides 16−18 and 17−18.

The permutation of labels does not change the Euclidean distances between spots at all, but it does alter the distances *along* the edges of the "fingerprint" pattern for the row and the column 17−20. In Table 5 we have listed the columns 17−20 for unperturbed and the perturbed matrices that give distances along the zigzag path (that enter as denominators in the quotients of distance in D/D matrices) in order to illustrate the magnitudes and the changes involved, which are minor. The columns 17 and 18 are identical in both cases till row 18 when columns 19 and 20 have changed for rows 1−18. As we see from Table 5 the differences between the unperturbed and the perturbed entries of D/D matrices are not large. Nevertheless, the perturbed "profile" of the corresponding "fingerprints" differs sufficiently, as illustrated in Table 5 where the last column belongs to the leading eigenvalues of a 20 × 20 matrix of the perturbed ${}^mD/{}^mD$ matrix. The results for the 10 × 10 and the 15 × 15 matrices remain the same, because the perturbation considered occurred at higher spot labels. A comparison of the last two columns in Table 3 shows that the unperturbed 20 × 20 and the perturbed 20 × 20* matrices have the same limiting leading eigenvalue. This is expected because the limiting binary matrix in both cases is an adjacency matrix of a linear graph having 20 vertices.

ON COMPARISON OF PERTURBED PROTEOMICS MAPS

If a drug or a toxin is administered to an animal for which cell proteome is analyzed the protein abundance will be

Table 4. Illustration of the Convergence of the $^{m}D/^{m}D$ Matrices as the Exponent *m* Tends to Infinity

	10 × 10	15 × 15	20 × 20	20 × 20[a]
25	2.08383085	2.23621521	2.38738446	2.38274709
50	1.98213908	2.23609904	2.38468716	2.38273384
100	1.92998853	2.23607137	2.38325678	2.38273164
250	1.91906781	2.23606675	2.38275465	2.38273124318428
500	1.91898597	2.23606671	2.28273142	2.38273124032431
1000	1.91898594722900	2.23606671884894	2.28273124033440	2.38273124032345
2500	1.91898594722899	2.23606671884893	2.38273124032344	2.38273124032344

[a] The last column belongs to a modified matrix (see text).

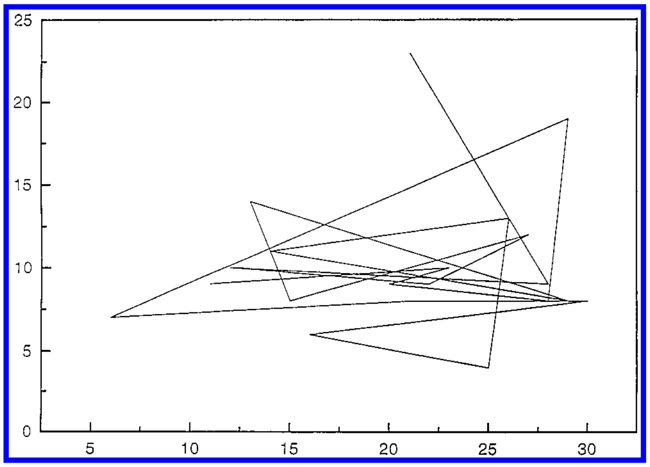

Figure 4. Zigzag line connecting the first 20 most intensive spots of Figure 2 with an exchange of relative intensities for spots having labels 17 and 18.

Table 5. Cumulative Distances for the Perturbed Proteomics Map for the Four Last Spots Affected by Perturbation

	17	18	19	20	17	18	19	20
1	1199	1247	1279	1351	1199	1247	1265	1337
2	1115	1163	1195	1267	1115	1163	1181	1253
3	1019	1067	1099	1171	1019	1067	1085	1157
4	959	1007	1039	1111	959	1007	1025	1097
5	925	973	1005	1077	925	973	991	1063
6	850	898	930	1002	850	898	916	988
7	816	864	896	968	816	864	882	954
8	716	764	796	868	716	764	782	854
9	625	673	705	777	625	673	691	763
10	552	600	632	704	552	600	618	690
11	503	551	583	655	503	551	569	641
12	442	490	522	594	442	490	508	580
13	353	401	433	505	353	401	419	491
14	300	348	380	452	300	348	366	438
15	210	258	290	362	210	258	276	348
16	65	113	145	217	65	113	131	203
17	0	48	80	152	0	48	66	138
18	48	0	32	104	113	0	18	90
19	80	32	0	72	131	18	0	72
20	152	104	72	0	203	90	72	0

affected and may show dramatic changes. A change in the relative abundance of spots will induce an assignment of novel labels to spots on the map. The zigzag graphical “fingerprints” pattern of modified proteomics maps will be considerably different from the zigzag pattern of the original (control) sample. Visual inspection of graphical “fingerprints” is somewhat tedious because of a dense overlapping of the zigzag curve. A more convenient approach to comparison of such proteomics maps can be accomplished: (1) using a set of invariants, such as the outlined map “profiles”, and (2) constructing the partial order based on comparison of old and new labels for spots.

A way to relate changes of individual protein spots when comparing different maps is to compare the original list of proteins ordered according to their abundance to the new list in which ordering is based on perturbed abundance. In Table 6 we show five perturbed orderings of spots that were generated at random to simulate changes in abundance when five different substances are experimentally used on animals of which tissue cells were collected. The newly obtained ordering of protein spots leads to zigzag patterns illustrated in Figure 5a−e. Already a visual inspection of Figure 5a−e shows that the outlined approach of construction of the “fingerprint” patterns represents a very sensitive scheme for discrimination among 2-D maps. It is just difficult to imagine that different proteomics maps will generate similar “fingerprint” patterns, when already only slightly perturbed data show variations in their characterization as has been previously illustrated in Figure 4 when we permuted two spots having close values of abundance, while keeping all other abundance values constant. Hence, the novel numerical characterization of proteomics maps can be expected to yield unique characterizations of proteomics maps.

PARTIAL ORDERING OF PROTEIN SPOTS

The original ordering of spots and the new ordering allow one to consider partial ordering of proteins that can characterize each of the five situations (toxins or drugs). A simple representation of induced permutation of labels is given as an ordered list of permutations of labels already illustrated in Table 6. The same information can be presented graphically as a partial order diagram illustrated in Figure 5 for the first permutation pattern of Table 6.

The randomly generated proteomics maps of Figure 5a−e simulate effects of a set of agents (drugs, toxic substances, or changes induced by). When attention is confined to protein spots that have changed their abundance considerably, the zigzag patterns do not offer insights as to what happened to individual protein spots. Nevertheless, qualitative information presented by partial order may serve to characterize pictorial maps, assist in classification of maps, facilitate comparisons of data from different sources (animals, organs, or laboratories), and search large data banks. If one wishes to analyze changes that occurred for a particular proteomics map induced by different agents the mathematical technique known as the partial order could be of use. Despite the fact that this approach is still qualitative in nature, it offers a different viewing of data which may be quite instructive. To illustrate the partial order as an approach we will consider in some details the case of the first randomly generated perturbed simulation listed at the top of Table 6 and illustrated in Figure 5a. In this case the most abundant is

Table 6. Five Randomly Generated Proteomics Maps for $n = 20$ Spots Obtained by Permutation of Labels 1−20 and the Corresponding Decomposition in Cycles

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
9	3	14	13	7	18	15	2	12	5	4	6	17	20	8	1	11	16	10	19
(1, 9, 12, 6, 18, 16) (2, 3, 14, 20, 19, 10, 5, 7, 15, 8) (4, 13, 17, 11)																			
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
6	12	2	8	11	15	13	3	18	9	14	17	5	7	1	20	10	4	19	16
(1, 6, 15) (2, 12, 17, 10, 9, 18, 4, 8, 3) (5, 11, 14, 7, 13) (16, 20) (19)																			
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
19	10	9	5	1	11	14	15	17	6	18	7	2	8	4	3	12	20	16	13
(1, 19, 16, 3, 9, 17, 12, 7, 14, 8, 15, 4, 5) (2, 10, 6, 11, 18, 20, 13)																			
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
16	13	6	17	5	12	10	19	1	11	20	9	15	14	8	2	18	4	7	3
(1, 16, 2, 13, 15, 8, 19, 7, 10, 20, 3, 6, 12, 9) (4, 17, 18) (5) (14)																			
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	16	17	4	14	12	11	15	9	5	3	7	20	18	19	10	8	6	13	2
(1) (2, 16 10, 5, 14, 18, 6, 12, 7, 11, 3, 17, 8, 15, 19, 13, 20)																			

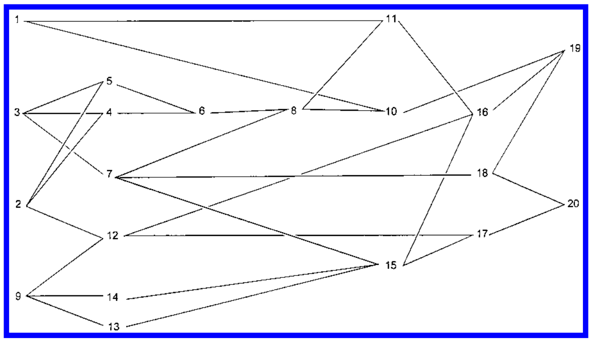

Figure 5. The partial order depicting permuted labels of the first randomly generated proteomics map of Table 6.

protein 9, the next most abundant is protein 3, then protein 14, etc.

The concept of partial order originates when objects having two (or more) attributes need to be ordered. Each property considered leads to a sequence in which objects are ranks differently. In our case, the "control" sequence is 1, 2, 3, 4, ..., while the perturbed sequence is as follows: 9, 3, 14, 13, ... Partial ordering from a mathematical point of view is concerned with identification of all ordered subsets for two sequences considered. Although each sequence represents a different ordering of elements of the set (protein spots), there may be shorter or longer fragments of the initial sequence that have preserved the relative order in both sequences. If we are comparing closely the first two rows of Table 6 we can find out, for example, that label 9 is in both lists ahead of label 10, which is again in both lists ahead of label 19. This observation can be formally represented by the following dominance relationship:

$$\mathrm{P}\ 9 > \mathrm{P}\ 10 > \mathrm{P}\ 19$$

Where P 9 stands for protein at spot 9 in the original proteomics map, P 10 stands for protein at spot 10 in the original proteomics map, etc. By further examination of the two lists of relative abundance we will see that the protein with label 9 is in both lists also ahead of the protein 12, which is again in both lists ahead of the protein 16 and the protein 16 is ahead of the protein 19. This yields to yet another fragmentary ordering

$$\mathrm{P}\ 9 > \mathrm{P}\ 12 > \mathrm{P}\ 16 > \mathrm{P}\ 19$$

The partial order consists of a complete list of all these fragmentary orders. However, rather than making a long list of all fragmentary orders it is customary to combine such results into a single diagram (already shown in Figure 5). The elements (protein spots) are drawn so that the entries at the left dominate (in the relative magnitudes of abundance) those at the right. The dominance is pictorially represented by connecting such points by a path. Thus in the lower portion of Figure 5 we see that P 9 is ahead of P 10, which is ahead of P 19. The relationship P 9 > P 10 > P 19 is graphically represented as a branch in the partial order of Figure 5. Another branch emanating from protein 9 is P 9 > P 12 > P 16 > P 19. While searching for fragmentary orders by straightforward comparison of two sequences is tedious, the graphical representation of partial order allows one to immediately view other fragmentary orders by tracing various paths from the dominant spots at the left side of the diagram. In Table 7 we have listed all fragmentary orderings of Figure 5 merely to indicate limitations of tabular presentation of data to diagrammatic presentation of Figure 5.

It is needless to say that Figure 5 can be drawn in many different ways. All that is important is to respect the relative order from the left to the right, but one can move points in the diagram (labels) up and down along vertical lines in an arbitrary manner, and also left and right as long as one does not alter implied hierarchy defined by the relative magnitudes of spots. Partial order implies directions along edges and hence is either represented graphically as an embedded (ordinary) graph as shown in Figure 5 or as a directed graph. The adjacency matrix of directed graphs has as entries besides zeros for nonadjacent vertices +1 or −1, depending on whether an edge is coming into a vertex or going out. Therefore information such as given by Figure 5 can be stored as a signed adjacency matrix.

When considering pictorial representations of partial order it is desirable for better visibility to have a diagram with as few crossing of lines as possible. Construction of diagram having the smallest number of crossing lines need not be always easy, thus diagrams with a relatively small number of crossings may suffice even if not optimal because pictorial form is here just an auxiliary format to summarize all partial orders for sequences considered. Often the time and the effort to get the optimal diagram may not be justified. For example, if in Figure 5 we place point 10 above point 11 we will reduce the number of crossings by one but visual inspection of both diagrams has hardly changed. Nevertheless, from a mathematical point of view finding an optimal diagram of partial order remains to be a problem of some interest.

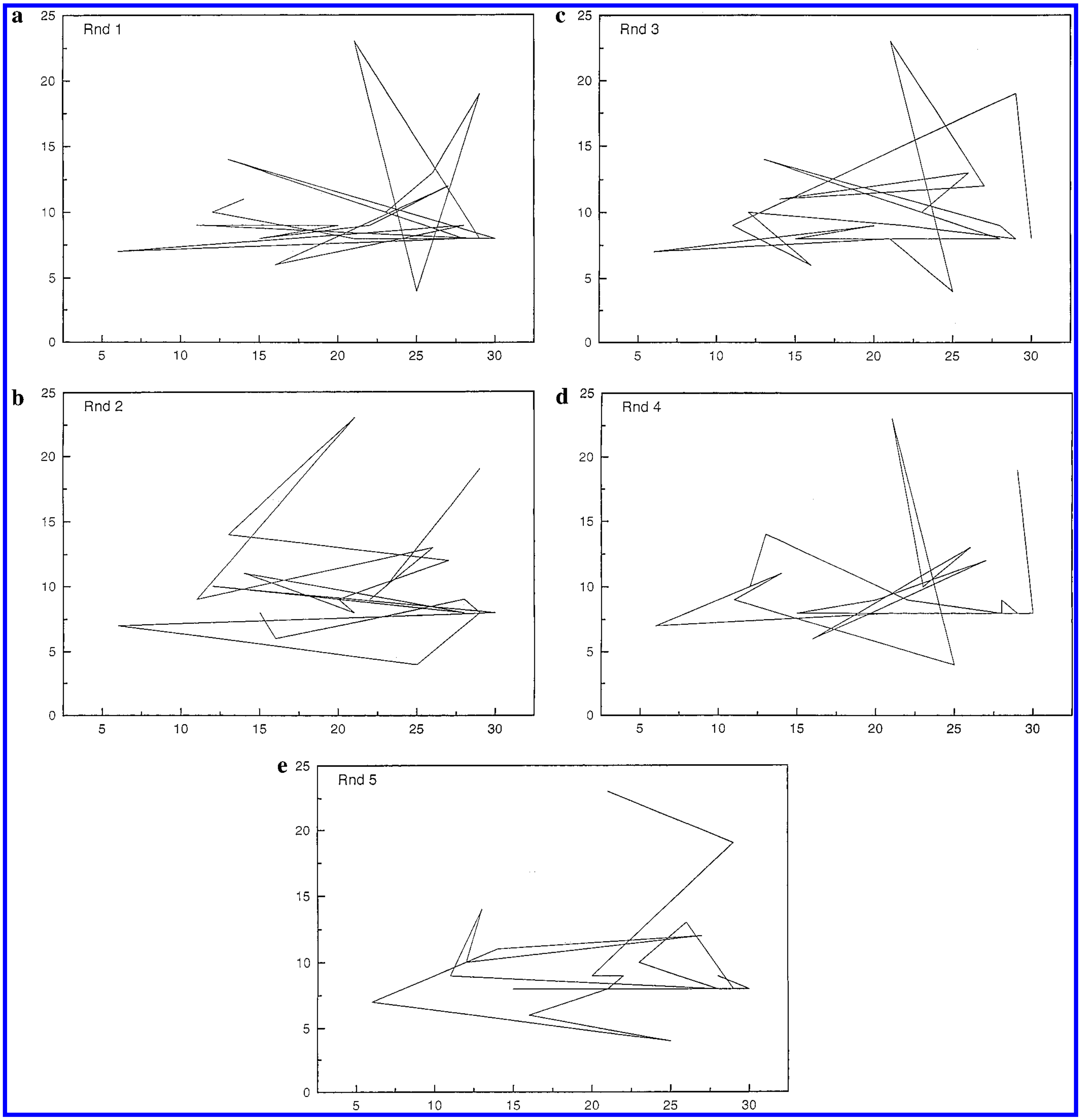

Figure 6. The “fingerprint” representations for five randomly generated proteomics maps of Table 6.

Table 7. List of Partial Orders of Figure 7

	path		path
1	1−11−16−19	11	3−7−18−20
2	2−4−6−8−10−19	12	9−10−19
3	2−5−6−8−10−19	13	9−12−17−19
4	2−12−16−19	14	9−12−17−19
5	2−12−16−20	15	9−13−15−16−19
6	3−4−6−8−10−19	16	9−14−15−16−19
7	3−5−6−8−10−19	17	9−14−17−19
8	3−7−8−10−19	18	9−14−18−19
9	3−7−15−16−19	19	9−14−17−20
10	3−7−18−19	20	9−14−18−20

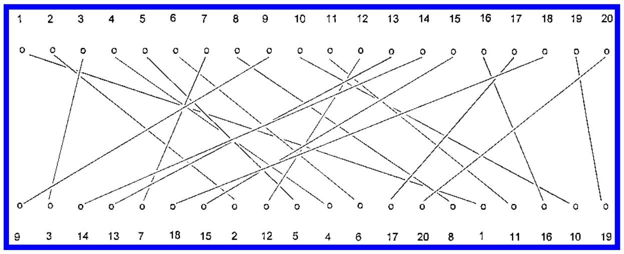

Figure 7. Construction of the partial order: The same labels in two sequences are connected by line. Any subset of noncrossing lines gives a branch in the partial order diagram.

A convenient way to a construct a partial order when both lists are given is first to list both sequences one above the other at some separation as illustrated in Figure 7. Then one connects the points having the same label in the upper and the lower sequence by a line, which results in a diagram having multiple crossings. Partial orders can now be extracted by listing an entry on the left side of the diagram that is followed by a sequence of labels at the right only if the lines that connect the same labels do not cross each other. For example, consider the label 9. As we see the line 9−9 does not cross the line 14−14, which does not cross the line 17−17 and this line does not cross the line 20−20. Hence, we obtain the partial (fragmentary) order:

$$P\ 9 > P\ 14 > P\ 17 > P\ 20$$

This is one of the paths depicted in Figure 5. Other similar fragmentary orders can be extracted in a same fashion. For example, following the same path at the last step we could have taken the line 19−19, instead of 20−20, because also the line 19−19 does not cross lines that are already listed before. This then gives

$$P\ 9 > P\ 14 > P\ 17 > P\ 19$$

This is another path that starts at label 9 but ends at label 19, not 20. In construction of the partial order diagram one does not depict lines between labels that can be reached by stepwise progression. Thus, for instance there is no need to draw a line between protein 12 and protein 19, or P 12 and 20, because one can reach label 19 and label 20 from site 12 by first getting to point 17, which is connected to 19 and 20.

Partial orders, like the one presented in Figure 5 are quite informative when one considers a comparison of an original control proteomics map with maps obtained after cells of test animals were perturbed by an agent administered to the animal. For example, we can see that the perturbation on the proteins 2 and 3 were such that their relative abundances have changed, but nevertheless both these proteins continue to dominate protein 4. This suggest that relative changes to proteins 2 and proteins 3 may have been drastic, but those on the protein 4 were in relation smaller and could not upset the relative abundance of 2 to 4 and 3 to 4. This in turn may suggest that protein 4 has some similarity (at least in the behavior) to both protein 2 and protein 3, but proteins 2 and 3 do not necessarily have common similarity (with respect to the effects of the agent administrated). The above discussion may appear somewhat speculative and this cannot be denied. What is important, however, is that we have now available a novel tool for comparison of proteomics maps. Future applications will show the extent to which the novel numerical characterization of proteomics will be useful.

CONCLUDING REMARKS

We proposed numerical characterization of proteomics maps that has several interesting aspects. First, a simple graphical representation in the form of a zigzag "fingerprint" pattern for proteomics maps was offered. To this graphical representation of proteomics map numerical characterization was prescribed based on the approach associated with D/D matrix in which the Euclidean distance and the distance measured along the zigzag path for any two points were used to construct the respective matrix elements. The leading eigenvalue of D/D matrix and the leading eigenvalues of similarly constructed higher order matrices ^{m}D/^{m}D are taken as a set of invariants that characterizes map. Other matrix invariants could also be considered. Comparison of proteomics maps has thus been transformed into a comparison of corresponding "profiles" of "fingerprints", which are represented by selected invariants, such as the leading eigenvalues of D/D matrices. We did not encounter situations that two or more spots would show the same abundance, which would cause ambiguity in assignment of labels to spots. If this would occur one could resolve such ambiguities by giving a preference, for example, to spots having a larger mass to those having a smaller mass (all having the same charge).

For a more complete comparison between different maps one can use also abundance (not directly used in this article) by constructing a vector for each proteomics map, the components of which are ordered relative magnitudes of abundance (listed in the last column of Table 1 for the case considered). Instead of a 2-D map on which zigzag curves were constructed one can also consider zigzag curves in 3-D. The 3-D space is constructed by using besides the x, y coordinates of the spots in the gel as the third coordinate the abundance.[26] Clearly the 2-D zigzag curve is a projection of the 3-D curve which has additional information on abundance (that has been here only indirectly taken into account through the assignment of labels to individual spots). The partial order that has been constructed when a proteomics map is analyzed for changes caused by an agent administrated to a test animal, however, will be the same whether we use 2-D or 3-D characterization of proteomics maps, because partial order does not use explicit values of the abundance but only the relative values which are implied in label assignments of spots. Partial orders, like one shown in Figure 5, relate to *two* proteomics maps, the "control" map and the "perturbed" map. Because of their inherent robustness partial orderings may prove valuable when comparing experimental data from different sources.

REFERENCES AND NOTES

(1) Lord Kelvin, *Popular Lectures & Addresses (1891−1894)*.

(2) Dirac, P. A. M. *Quantum Mechanics;* Oxford University Press: London 1935.

(3) Grassy, G.; Calas, B.; Yasry, A.; Lahana, R.; Woo, J.; Iyer, S.; Kaczorek, M.; Floc'h, R.; Buelow, R. Computer-assisted rational design of immunosuppressive compounds. *Nature Biotechn.* **1998**, *16,* 748−752.

(4) Witzmann, F. Indiana University Purdue University: Columbus, IN, private information.

(5) Nandy, A. New graphicalk representation and analysis of DNA sequence structure. I. Methodology and application to globin genes. *Curr. Sci.* **1994**, *66,* 309.

(6) Roy, A.; Raychaudhary, C.; Nandy, A. Novel techniques of graphical representation and analysis of DNA − A review. *J. Biosci.* **1998**, *23,* 55−71.

(7) Hamori, E. Graphical representation of long DNA sequences by methods of H curves, current results and future aspects. *BioTechniques* **1989**, *7,* 710−720.

(8) Leong, P. M.; Mogenthaler, S. Random walk and gap plots of DNA sequences. *Comput. Appl. Biosci.* **1995**, *12,* 503−511.

(9) Randić, M. On characterization of DNA sequences by condensed matrix. *Chem. Phys. Lett.* **2000**, *317,* 29−34.

(10) Randić, M. Condensed representation of DNA primary sequences. *J. Chem. Inf. Comput. Sci.* **2000**, *40,* 50−56.

(11) Randić, M.; Vračko, M. On similarity of DNA primary sequences. *J. Chem. Inf. Comput. Sci.* **2000**, *40,* 599−606.

(12) Randić, M.; Vračko, M.; Nandy, A.; Basak, S. C. On 3-D graphical representation of DNA primary sequences and their numerical characterization. *J. Chem. Inf. Comput. Sci.* **2000**, *40,* 1235−1244.

(13) Randić, M.; Basak, S. C. Characterization of DNA based on average distances between the nucleic acid bases. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 000−000.

(14) Randić, M.; Razinger, M. On characterization of three-dimensional molecular structure. In *From Chemical Topology to Three-Dimensional Geometry*; Balaban, A. T., Ed.; Plenum Press: New York, 1996; pp 159−236.

(15) Bytautas, L.; Klein, D. J.; Randic, M.; Pisanski, T. Foldedness in linear polymers: A difference between graphical and Euclidean distances. *DIMACS Ser. Discrete Mathematics Theor. Comput. Sci.* **2000**, *51*, 39−61.

(16) Randić, M.; Vračko, M.; Novic, M.; Basak, S. C. On ordering of folded structures. *MATCH* **2000**, *42*, 181−231.

(17) Randić, M.; Kleiner, A. F.; DeAlba, L. M. Distance/distance matrices. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 277−2866.

(18) Randić, M.; Razinger, M. On characterization of molecular shapes. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 594−606.

(19) Randić, M. Molecular profiles − Novel geometry-dependent molecular descriptors. *New J. Chem*. **1995**, *19*, 781−791.

(20) Randić, M. On characterization of the conformations of nine-membered rings. *Int. J. Quantum Chem: Quantum Biol. Symp.* **1995**, *22*, 61−73.

(21) Randić, M. Krilov, G. On characterization of the folding of proteins. *Int. J. Quantum Chem.* **1999**, *75*, 1017−1026.

(22) Trinajstić, N. *Chemical Graph Theory;* CRC Press: Boca Raton, Fl, 1992.

(23) Gantmacher, F. *Theory of matrices*; Chelsea Publishers: New York, 1959; Vol. II, Chapter 13.

(24) Li, H.; Helling, R.; Tang, C.; Wingreen, N. *Science* **1996**, *273*, 666−669.

(25) MATLAB; version Edu-MATLAB 87-91; The MathWorks, Inc.: Natick, MA.

(26) Randić, M.; Zupan, J.; Nović, M. On 3-D graphical representation of proteomics maps and their numerical characterization. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1339−1344.

CI000167B