Novel Characterization of Proteomics Maps by Sequential Neighborhoods of Protein Spots

Milan Randić,* Marjana Novič, and Marjan Vračko

National Institute of Chemistry, Ljubljana, Hajdrihova 19, Slovenia

Received July 31, 2004

We consider a characterization of proteomics maps based on an alternative kind of neighborhood graphs for the protein spots on 2-D gel. The novel approach considers for every protein spot only the nearest neighborhood consisting of protein spots of higher abundance. The approach has the simplicity and advantages of the recently introduced characterization of proteome maps based on considering the nearest neighborhoods of protein spots, but it also has important additional desirable computational features. The characterization of the nearest neighborhood graphs of 2-D gel proteomics maps is sensitive to the number of spots considered and may lead to changes in the degree of similarity of different maps when the number of points has been changed, thus imposing restrictions on the protocol used for comparison of maps. The novel approach presented in this work is less sensitive to the number of points used in the analysis because graphs are constructed in a stepwise process in which the role of more distant neighbors has been diminished by linking a new spot to the nearest spot that has been already part of the neighborhood graph. In this way a graph with $N + 1$ spots is obtained from the graph on N spots by adding a single new link, while in the case of the nearest neighborhood graphs adding a new spot introduces novel neighborhoods and generally results in a graph that may differ significantly from the neighborhood graph on N points.

INTRODUCTION

It has been generally recognized that “innovative computational tools and methods to process, analyze and interpret prodigious amount of data” emerging in the world of proteomics is essential for further advances in this field.[1] For a brief review on future directions in proteomics, one should consult the article “Proteomics in Genomeland” by Stanley Fields, from which we will cite but a few sections that clearly point to and justify the work that we continue to develop in the present paper:

“So far most proteomic measurements have been performed in a cataloguing mode, but the future will see more studies that address the dynamics of cellular processes. The protein composition of a cell is not static, therefore, it is crucial to obtain quantitative comparisons after a cell’s environment changes. ... Increasingly, proteins will undergo wholesale analyses to probe for their various modifications.”

“Protein databases will need to become much more sophisticated if they are to help a scientist make sense of the staggering number of experimental measurements that will soon emerge. ... As proteomic data accumulates, we will become better at triangulating from multiple disparate bits of information to obtain bearing on what a protein does in a cell.”

“An interdisciplinary spirit will come to guide those excited by the global analysis of protein function. Geneticists need to talk to chemists, physiologists to physicists, cell biologist to computer science.”

Let us end these introductory quotations by augmenting the last quote with “and all of them need to talk to mathematical chemists”. The reason for this is that mathematical chemistry offers concepts and language that is very appropriate for characterization of complex systems and is not necessarily so widely known. The first step in trying to follow the above outlined task of helping scientists to “make sense of the staggering number of experimental measurements” is to develop better systems of recording and processing raw data given by proteomics maps that is a straightforward list of spots coordinates and their abundances. This has been our goal, and in this paper we continue with developing further such an approach.

QUANTITATIVE CHARACTERIZATION OF PROTEOMICS MAPS

In a series of publications, we outlined several alternative ways to arrive at numerical characterization of proteomics map. We consider as “characterization” or “description” of a map any useful set of map invariants, that is, numerical descriptors that neither depend on assumed numbering of protein spots nor depend on the orientation of the map. Our strategy is to associate with a map suitable mathematical objects, which are using matrix algebra transformed into matrices, from which a set of invariants are extracted.[2−13] In Table 1, we have summarized several approaches that have shown promise for quantitative characterization of 2-D proteomics maps. To these we would like to add a novel scheme to be introduced in this work, which is closely related to the recently outlined approach based on construction on the concept of the nearest neighbor graphs.

On one hand, the four approaches listed in Table 1 may be viewed to represent an evolutionary process in arriving at numerical characterization of proteomics maps that culminates with the approach to be outlined in this work. However, on the other hand, because each of the four

10.1021/ci0497612 CCC: $30.25
Published on Web 07/27/2005

Table 1. Survey of Evolutionary Methods for Numerical Characterization of Proteomics Maps Based on Mathematical Invariants of a Map

mathematical object		ref
2-D and 3-D zigzag line	spots are numbered by relative abundance (unless desired differently) and sequentially connected; distance matrix for vertices of zigzag line set of invariants	2−4
graph of partial order	spots are ordered with respect to dominance by mass and charge (pH); invariants are obtained by considering the resulting partial order graph of fixed geometry (superimposed over the map)	5, 6
cluster graph	spots are connected in a graph if they are at selected critical distance or closer; invariants are obtained from distance matrix of embedded cluster graph (superimposed over the map)	8
nearest neighbor graph	each protein spot is connected to its NN[a]; matrix giving distance between adjacent spots in graph (embedded graph over the 2-D map) offers suitable invariants	11, 12
sequential neighbor-hood	spots are numbered by relative abundance (unless desired differently); each spot is connected only to its NN[a] of lower order; associated distance matrices offer novel map invariants	this work

[a] NN = nearest neighbors.

approaches involves different structural elements of proteomics maps, they are not necessarily competitive and may find different use in different applications. Let us briefly outline their advantages and limitations to better appreciate features of the here introduced novel approach. The approach based on construction of a zigzag line by connecting spots of similar abundance[2−4] despite being rather simple and straightforward nevertheless offers useful characterization of complex proteomics maps. It is unlikely that different maps will match in the relative abundance of two dozen or more proteins, thus one may expect unique characteristics for individual proteomics maps determined by as few as $N - 1$ line segments connecting proteins spots of decreasing abundance. Partial order on proteins (with respect to mass and charge) introduces for set of N spots a graph with some N^2 lines, and thus captures more features of a map as a whole.[5,6] An advantage of the graph of partial order is that its construction does not require a prior ordering of spots, as was the case with the zigzag line. In contrast to the graph of partial order, the cluster graph[8] in which all spots within certain distance are connected can be dense, that is, having a large number of edges, thus incorporating a higher amount of information on distances of spots in 2-D map. The question that remains to be investigated is whether such additional information is essential for characterization of maps or is perhaps to some extent redundant. This approach has the advantage that it allows the user to select the critical distance for construction of cluster and thus determine how many lines will have the cluster graph. A disadvantage of the approach is that finding all the shortest path between pair of vertices in a graph needed for the construction of the distance matrix that accompany the approach is known to be associated with the NP (non-polynomial) algorithm.

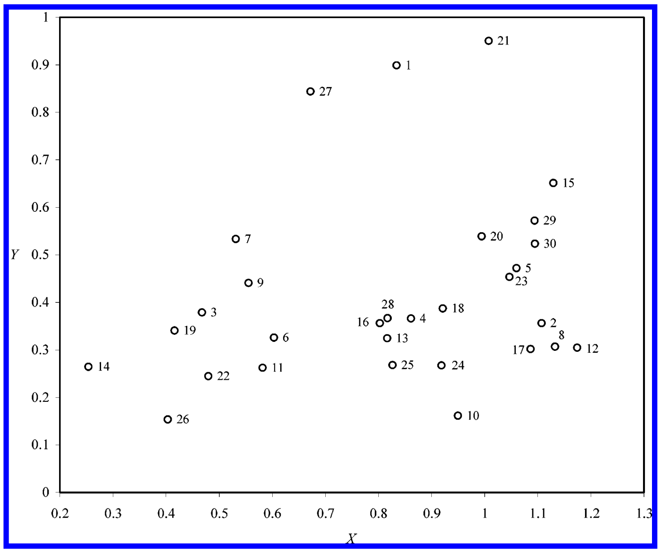

Figure 1. 30 most abundant proteins of a simplified proteome map of liver cells of rats listed in Table 2.

Table 2. Position Coordinates and Relative Abundance Values for Proteomics Map of Figure 1

	x	*y*	*z*		*x*	*y*	*z*
1	2117.7	2278.6	**1443.57**	16	2032.7	902.8	**800.15**
2	2804.3	903.6	**1436.30**	17	2752.7	765.6	**798.7**
3	1183.9	959.6	**1366.63**	18	2334.2	982.2	**727.91**
4	2182.2	928.8	**1272.95**	19	1053.6	864.3	**721.73**
5	2685.6	1196.1	**1185.81**	20	2519.5	1365.9	**694.52**
6	1527.9	825.5	**1149.29**	21	2552.5	2409.4	**677.72**
7	1346.0	1352.5	**1122.51**	22	1214.3	620.0	**646.84**
8	2868.5	778.0	**1088.93**	23	2651.1	1149.6	**610.74**
9	1406.3	1118.1	**982.24**	24	2327.9	677.3	**592.94**
10	2450.2	409.2	**936.01**	25	2094.5	680.5	**589.77**
11	1474.0	665.1	**900.04**	26	1021.7	390.2	**580.01**
12	2974.9	772.8	**867.30**	27	1702.7	2138.3	**574.00**
13	2068.4	823.1	**848.42**	28	2070.4	929.6	**554.02**
14	642.2	669.8	**824.92**	29	2771.7	1451.0	**538.96**
15	2860.7	1649.9	**819.65**	30	2772.8	1326.9	**513.47**

The latest addition to the collection of numerical approaches for characterization of proteomics map is based on construction of graphs of the nearest neighborhoods for selected spots of proteome map.[11,12] In this approach, again the user selects the number of nearest neighbors to be considered, which results in graphs with variable numbers of lines. An important advantage of this approach is conceptual and computational straightforwardness, which makes it easy to use and suitable to persons in an experimental laboratory to apply as it does not require familiarity with the notion of partial order and Hasse diagrams[14−18] or familiarity with the Dijsktra algorithm[19] for searching the shortest paths in a graph.

GRAPH OF THE NEAREST NEIGHBORHOODS

In Figure 1, we illustrate a simplified proteomics map with 30 protein spots based on data from the laboratory of F. Witzmann,[20] which we used in previous studies.[4,10,11] The protein spots are numbered from 1 to 30 in decreasing values of the abundance (the z-coordinate of Table 2). In Figure 2a−c for the same map, we show the graphs in which we have connected each spot with its nearest neighbor spot. In Figure 2a, we consider $N = 20$ protein spots of the highest abundance; in Figure 2b, we consider $N = 25$ spots; in Figure

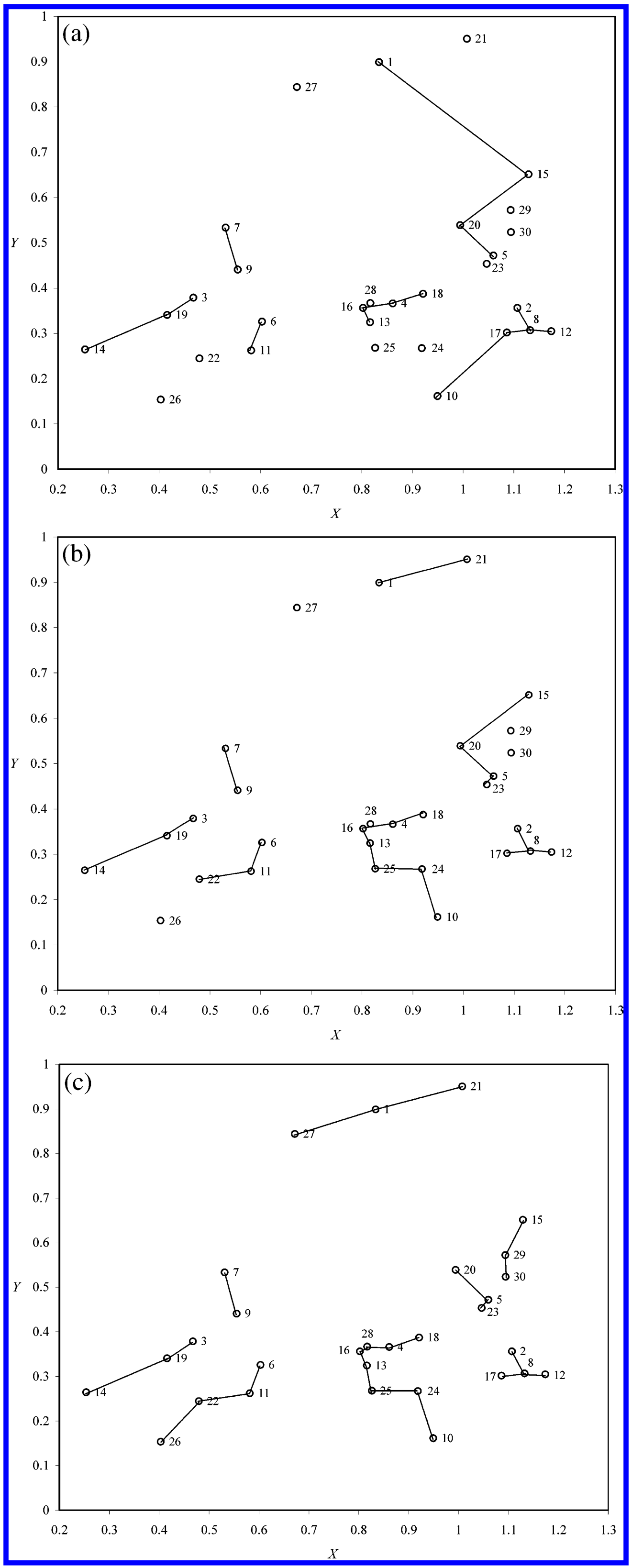

Figure 2. (a) Nearest neighborhood graph $N = 20$, NN = 1. (b) Nearest neighborhood graph $N = 25$, NN = 1. (c) Nearest neighborhood graph $N = 30$, NN = 1.

2c, we consider all $N = 30$ protein spots of Figure 1. As we see from Figure 2a−c, the pattern of connections in the neighborhood graphs is sensitive to the number of spots selected. When considering additional spots we have to redraw the neighborhood graphs because novel spots will alter the neighborhood relations. Spots that were the nearest after inclusion of additional spots need no longer be the closest neighbors, and thus the graph depicting the nearest neighbors depends on the number of spots used. Thus in Figure 2a when considering only $N = 20$ spots, the nearest neighbor to spot 1 is spot 15; in Figure 2b with $N = 25$, the nearest spot to spot 1 is spot 21; and finally when $N = 30$, the nearest spot to spot 1 is spot 27. Observe that some vertices in the nearest neighbor graphs may have two or more neighbors. For instance, spots 21 and 27 at the top part of Figure 2c are both connected to spot 1, their nearest neighbor, which therefore is connected to two spots. While the characterization of proteomics maps based on the neighborhood graphs have important computational advantages, its main disadvantage is that the connectivity of neighborhood graphs depends critically on the number of spots used and can change drastically when additional spots are included in the analysis, as is visible from a close comparison of Figure 2a with Figure 2, panels b and c.

GRAPH OF THE SEQUENTIAL NEAREST NEIGHBORHOODS

In Figure 3a−c, we show also graphs depicting neighborhoods for $N = 20$, $N = 25$, and $N = 30$ spots, but now we have connected each spot only to its nearest spot if the spot has a smaller label, that is, if it belongs to more abundant protein. Obviously for protein spot 2, spot 1 is the only nearest spot with a smaller label. The next spot, spot 3 is connected to spot 1 because it is closer to spot 3 than to spot 2. The process continues with spot 4, which among spots 1−3 is the closest to 2, and similarly among spots 1−4, spot 5 is closest to protein 2. As we continue, we obtain for the first $N = 20$ spots Figure 3a. By inclusion of additional five spots, we obtain Figure 3b, which when compared to Figure 3a shows additional lines for spots with labels 21−25, but inclusion of new spots does not change the already constructed part of the graph involving spots 1−20. By continuing to include the first 30 most abundant spots, we obtain the graph of Figure 3c, which incorporates Figure 3a,b as its parts.

The modified rule for connecting N spots only to those of higher abundance always produces a connected graph having $N - 1$ edges. Thus, Figure 3c has 29 edges (lines) in contrast to Figure 2c with only 22 lines. We will refer to both graphs as graphs of protein map nearest neighborhood; specifically graphs of Figures 2 and 3 are graphs for NN = 1 (NN = nearest neighbor), and when it is necessary to differentiate between the two, we will refer to sequential neighborhood for the later (Figure 3). As we can see from Figure 3c, again spots 21 and 27 are connected to spot 1, but now in addition spot 1 is also connected to protein spots 2 and 3. In contrast to graphs of Figure 2a−c, graphs of Figure 3a−c representing the sequential neighborhood will critically depend on assumed ordering of spots. Unless specified differently, we will continue to assume that spots have been labeled according to their relative abundance of the proteomics map belonging to the control group. The advantage of the ordering of spots relative to their abundance is that errors in experimental measurements of abundance make smaller influence on characterization of a map than would be otherwise the case.

In analogy with Figure 3 (NN = 1), graphs can be drawn for the sequential nearest neighborhoods for NN = 2, NN

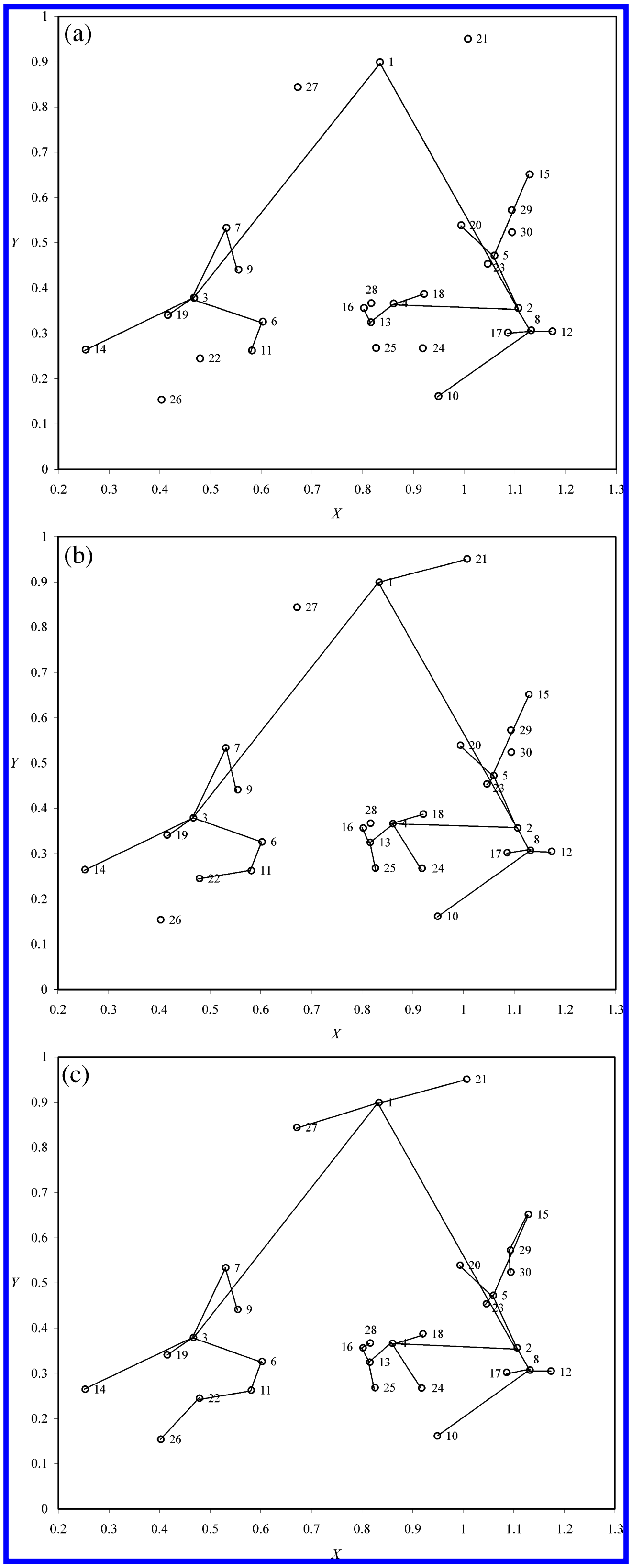

Figure 3. (a) Nearest sequential neighbors $N = 20$, NN = 1. (b) Nearest sequential neighbors $N = 25$, NN = 1. (c) Nearest sequential neighborhood graph $N = 30$, NN = 1.

= 3, and so on. A comparison of the graphs of the NN = 1 for $N = 20$, $N = 25$, and $N = 30$ case shows that the larger graph contains the smaller graph as a subgraph. Thus the graph for $N = 30$ points contains the graph for $N = 25$ spots as a subgraph, which in turn contains graph $N = 20$ as its subgraph. In general, a sequential neighborhood graph on $N + 1$ spots contains a sequential neighborhood graph on N spots. Hence, one can obtain the neighborhood graph for $N = 30$ from the graph of $N = 25$ by adding the five new vertices and connecting them sequentially to the nearest vertices of the subgraph $N = 25$. The advantage of this novel approach is that it has robustness and that the neighborhood graph for $N + 1$ spots does not change the already constructed subgraph for N spots as has been the case with the nearest neighbor neighborhood graphs; moreover, the distances involved in adding successive spots, as N increases, are associated with shorter and shorter magnitudes. This is a consequence of the fact that the density of spots in the map continually increases, and thus new spots have neighbors at smaller separations.

2-D GEL MAP DESCRIPTORS

We will illustrate our novel approach on the proteomics map shown in Figure 1 that has $N = 30$ spots. In Table 2, we have listed the x, y, and z coordinates of the selected protein spots. The (x, y) coordinates in a 2-D gel are associated with location of spots on the gel (x relates to the relative mass and y relates to the relative charge), while the z coordinate reflects the relative abundance of a protein spot. As already mentioned, the first step in analysis involves ordering the spots, which we will assume to be ordered relative to their abundance values. To construct neighborhood graphs we need to examine all distances, which requires the construction of the spot distance matrix. Once we have the distance matrix, it is not difficult to sort out spots and to find easily for each spot its NN sequential nearest neighbors. In Table 3, we have listed for the protein map of Figure 1 and Table 2 the six sequential nearest neighbors. In the first column of Table 3 where the spots run from $N = 2$ to $N = 30$, we have listed for each new spot in each row the nearest spots among those already listed above. Thus the entries of the first column of Table 3 give the graphs depicted in Figure 3a–c showing the neighboring graphs for NN = 1. Observe that in the first column, as well as in all other columns, the second index, which indicates the nearest sequential neighbor for each row, is always smaller than the first (running) index.

The remaining columns of Table 3 list from the second to the sixth sequential nearest neighbor. By including third and higher order nearest neighbors, one would obtain graphs with increasing number of lines offering limited visual advantage because they would be fairly dense.[21] Therefore, we will limit attention only to visual inspection of Figure 3a–c (portraying graphs for $N = 20$, $N = 25$, and $N = 30$), which depict the sequential neighborhoods NN = 1. If one is to compare the sequential neighborhood graphs with the corresponding nearest neighborhood graphs, one would find that the graphs of Figure 3a–c are more dense than the corresponding graphs depicting the nearest neighborhoods (depicted in Figure 2 for NN = 1). This is clearly seen from Table 4 in which the numbers of vertices and edges in the two types of neighborhood graphs for NN = 1 and NN = 2 are listed as the number of included vertices increases. In Table 4 we also included the average valence (degree) of vertices given by $2E/V$, where E is the number of edges and V is the number of vertices. It is interesting to see that as V increases, in the case of the nearest neighborhood graphs for NN = 2, the average vertex

Table 3. First Six Sequential Nearest Neighbors for Proteomics Map Spots of Figure 1

first	second	third	fourth	fifth	sixth
0					
(2,1) 1536.89	0				
(3,1) 1616.01	(3,2) 1621.37	0			
(4,2) 622.61	(4,3) 998.77	(4,1) 1351.34	0		
(5,2) 315.67	(5,4) 569.97	(5,1) 1222.42	(5,3) 1520.21	0	
(6,3) −369.21	(6,4) 662.4	(6,5) 1215.57	(6,2) 1278.79	(6,1) 1568.24	0
(7,3) 425.03	(7,6) 557.51	(7,4) 937.42	(7,1) 1205.48	(7,5) 1348.70	(7,2) 1525.83
(8,2) 141.06	(8,5) 456.36	(8,4) 702.67	(8,6) 1341.44	(8,7) 1627.29	(8,1) 1677.95
(9,7) 242.03	(9,3) 273.10	(9,6) 316.86	(9,4) 798.66	(9,5) 1281.68	(9,1) 1361.19
(10,8) 557.66	(10,4) 584.64	(10,2) 608.13	(10,5) 821.36	(10,6) 1011.90	(10,9) 1261.85
(11,6) 169.21	(11,3) 413.39	(11,9) 458.03	(11,7) 699.22	(11,4) 755.70	(11,10) 1009.18
(12,8) 106.5	(12,2) 215.0	(12,5) 512.7	(12,10) 638.4	(12,4) 807.9	(12,6) 1448.0
(13,4) 155.32	(13,6) 540.51	(13,10) 563.10	(13,11) 615.04	(13,5) 721.16	(13,9) 724.85
(14,3) 614.35	(14,11) 831.81	(14,9) 885.90	(14,6) 899.28	(14,7) 980.52	(14,13) 1434.42
(15,5) 486.41	(15,2) 748.43	(15,8) 871.93	(15,12) 884.50	(15,1) 973.30	(15,4) 990.13
(16,13) 87.33	(16,4) 151.74	(16,6) 510.68	(16,11) 607.16	(16,10) 646.49	(16,9) 662.37
(17,8) 116.46	(17,2) 147.33	(17,12) 222.32	(17,5) 435.7	(17,10) 467.47	(17,4) 593.38
(18,4) 161.11	(18,13) 309.78	(18,16) 311.78	(18,5) 411.38	(18,17) 471.23	(18,2) 476.63
(19,3) 161.43	(19,9) 434.52	(19,14) 455.06	(19,11) 465.21	(19,6) 475.88	(19,7) 569.07
(20,5) 237.53	(20,18) 426.10	(20,15) 443.93	(20,2) 542.98	(20,4) 552.11	(20,17) 644.00
(21,1) 454.05	(21,15) 819.65	(21,20) 1044.02	(21,5) 1220.58	(21,18) 1443.80	(21,4) 1526.20
(22,11) 263.59	(22,19) 292.42	(22,3) 340.96	(22,6) 374.93	(22,9) 533.82	(22,14) 574.26
(23,5) 57.90	(23,20) 253.19	(23,2) 289.80	(23,18) 253.19	(23,17) 397.21	(23,8) 430.52
(24,4) 290.66	(24,10) 294.68	(24,13) 297.65	(24,18) 304.97	(24,16) 371.47	(24,17) 433.88
(25,13) 144.97	(25,16) 230.73	(25,24) 233.42	(25,4) 263.33	(25,18) 385.33	(25,10) 447.35
(26,22) 299.84	(26,14) 471.38	(26,19) 475.17	(26,11) 529.29	(26,3) 592.05	(26,6) 667.63
(27,1) 438.07	(27,7) 862.97	(27,21) 892.00	(27,9) 1062.38	(27,20) 1124.17	(27,15) 1256.78
(28,16) 46.26	(28,13) 106.52	(28,4) 111.80	(28,25) 250.26	(28,18) 268.99	(28,24) 360.50
(29,15) 217.90	(29,20) 266.17	(29,5) 269.05	(29,23) 324.63	(29,2) 548.37	(29,18) 641.23
(30,29) 124.10	(30,5) 157.20	(30,23) 215.05	(30,20) 256.28	(30,15) 334.75	(30,2) 424.47

Table 4. Increase in the Number of Edges as the Number of Vertices Grow and the Average Vertex Valence for Two Kind of Nearest Neighborhood Graphs

no. of vertices	nearest neighbors (Figure 2a−c)		sequential nearest neighbors (Figure 3a−c)	
	NN = 1			
20	14 (14/20 = 0.7000)	1.40	18 (18/20 = 0.9000)	1.80
25	18 (18/25 = 0.7200)	1.44	23 (23/25 = 0.9200)	1.84
30	22 (22/30 = 0.7333)	1.47	28 (28/30 = 0.9333)	1.87
	NN = 2			
20	28 (28/20 = 1.4000)	2.80	35 (35/20 = 1.7500)	3.50
25	34 (34/25 = 1.3600)	2.72	45 (45/25 = 1.8000)	3.60
30	40 (40/30 = 1.3333)	2.67	55 (55/30 = 1.8333)	3.67

degree decreases, while in the case of the graphs of sequential neighborhoods the average vertex degree increases as *V* increases. Clearly the later graphs contain more information on the connectivity among the proteomics map spots.

As the first map descriptor we have selected the average row sum of distance weighted adjacency matrices of the sequential neighborhood graphs. These in fact are the weighted adjacency matrices of the neighborhood graphs portion of which is illustrated in Table 5. The weights used are the Euclidean distances between the connected vertices. In Table 5 we show only a 10 × 10 portion of the 30 × 30 matrix associated with the sequential neighborhood graph for $N = 30$, NN = 2. The average row sums can be obtained directly from information listed in Table 5 without actually constructing the accompanying neighborhood graph by summing all the entries of Table 5 (and extending summation to columns 11−30 not listed in Table 5). This sum is proportional to the generalized Wiener number of a graph based on the actual lengths of connecting lines, except that it includes only distances between adjacent vertices, while the Wiener number,[22] as shown by Hosoya,[23] can be obtained by summing all entries above the main diagonal of the distance matrix of a graph. In Table 6 we have listed the increments to the row sums when adding additional nearest neighbors as well as the overall average row sums, the later being illustrated in Figure 4.

Observe a rather regular quadratic dependence of the average row sums on the variable *N*, the number of spots. This regularity is somewhat surprising because no a priori reason would require that apparently random-like proteomics map should show a degree of organization. The same regularity has also been observed for the nearest neighbor graphs (see refs 11 and 12) and is of similar unexpected occurrence. The significance of this finding is that one can conclude that distribution of proteins in proteomic map, which appears random, is not random. This follows from a comparison of Figure 5 with the corresponding figure of the distribution of the average row distances for randomly generated maps—the comparison that has shown marked differences—the random map figures grouping well below those of genuine proteomics maps.[25] It is interesting to mention in this respect that graphical illustrations of 10 and 100 thousands of nucleic bases of DNA that Jeffrey[26] considered in his “Chaos game” graphical representation of DNA also visually appear as scatter of random points but, as Jeffrey has shown, are not representing random distribution of points in 2-D plane.

In Table 7, we have listed the magnitudes of the row sums for the case NN = 1−6. Observe that some row sums in adjacent columns of Table 7 remain constant when NN changes, others continue to increase, due to presence of additional connections. As we move down the table on average the magnitude of row sum decrease, though not uniformly. This is even better seen from Table 8 in which in the upper part we show the average row sums for every

Table 5. Portion of Distance-Weighted Adjacency Matrix for Sequential Neighborhood (NN = 2)

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	0	1536.9	1616.1	0	0	0	0	0	0	0
2	1536.9	0	0	622.6	315.7	0	0	141.1	0	0
3	1616.1	0	0	988.8	0	369.2	425.0	0	273.1	0
4	0	622.6	998.8	0	570.0	662.4	0	0	0	584.6
5	0	315.7	0	570.0	0	0	0	456.4	0	0
6	0	0	369.2	662.4	0	0	557.5	0	0	0
7	0	0	425.0	0	0	557.5	0	0	242.0	
8	0	141.1	0	0	456.4	0	0	0	0	557.7
9	0	0	273.1	0	0	0	242.0	0	0	0
10	0	0	0	584.6	0	0	0	557.7	0	0

Table 6. Increment in the Average Row Sum by Adding the Next Nearest Neighbor and the Overall Average Row Sums for NN = 2 for Graphs Having from 1 to 6 Nearest Neighbors

	increment	average row sum
1st	697.3	697.3
2nd	913.2	1610.5
3rd	1050.6	2661.0
4th	1207.3	3868.4
5th	1312.6	5181.0
6th	1409.4	6590.4

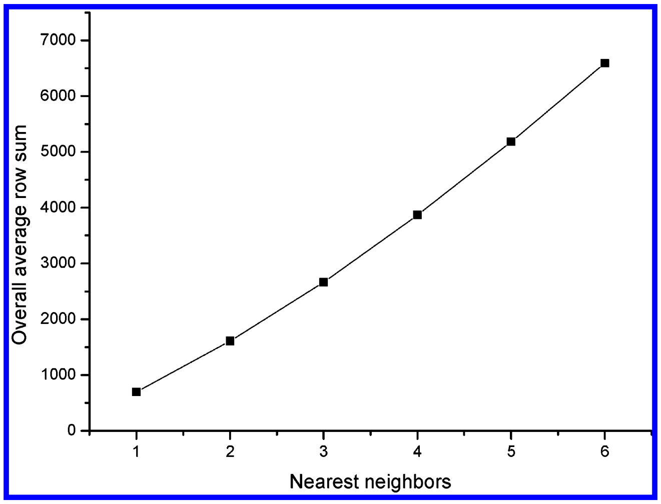

Figure 4. Plot of the average row sums for NN = 1−6. R^2 = 0.99394; SD = 0.16278; F = 656.45.

five rows, and in the lower part we show the differences of successive row sums. A close look at Table 8 shows irregularities in the increments of the average row sums associated with low N values and low NN values. Ideally one would see regular increments in the average row sums after adding a group of five spots and after adding additional neighbors. We see that the sought regularities occur in the lower right-hand side of Table 8, for rows 20−30 and columns NN = 5 and 6, which suggests that maps with less than N = 20 spots and NN = 5 are not likely to offer a sufficiently reliable information for characterization of proteomics maps. Table 8 also suggests that by increasing the number N of spots, if we reduce the number of the nearest neighbors NN, we still obtain a robust stable characterization of the proteomics map.

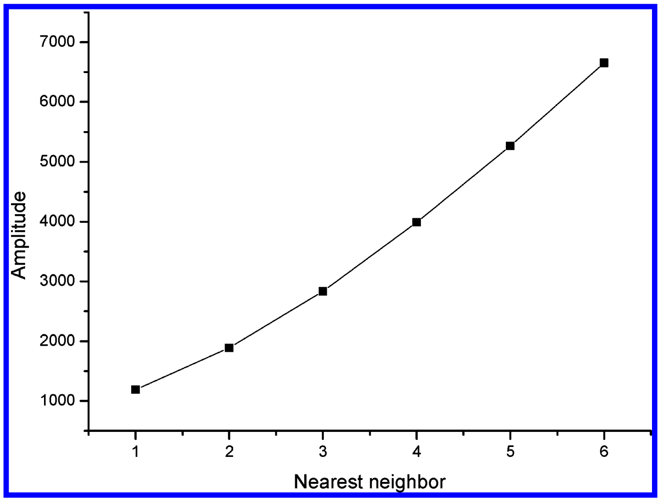

Figure 5. Plot of the average magnitude of two-component vectors for NN = 1−6. R^2 = 0.98724; SD = 0.23626; F = 309.50.

Table 7. Magnitude of Individual Row Sums for Different NN Values

	NN = 1	NN = 2	NN = 3	NN = 4	NN = 5	NN = 6
1	4045.1	4045.1	6618.9	7824.4	10365.9	13405.0
2	2616.2	5348.3	6246.3	8068.0	8616.4	11043.3
3	3186.1	6492.7	6833.7	8353.9	8946.0	8946.0
4	1229.7	4197.2	7300.5	8362.4	10478.2	13587.9
5	1097.5	2281.0	5500.8	9910.0	13261.6	13261.5
6	538.4	2298.8	4342.0	8236.4	11292.4	13408.0
7	667.1	2087.5	3025.0	4929.7	8886.2	10981.0
8	921.7	1378.1	2952.7	4294.1	5921.4	8029.9
9	242.0	949.7	2610.5	4471.5	6287.0	10297.3
10	557.7	1437.0	2608.2	4067.9	6193.8	8912.2
11	432.8	1678.0	2136.0	5051.9	5807.6	6816.8
12	106.5	321.5	1056.5	2579.4	3387.3	4835.3
13	387.6	1344.4	2205.2	2820.2	3541.4	5700.6
14	614.3	1917.5	3258.6	4157.8	5138.3	7147.0
15	704.3	2272.4	3588.3	4472.8	5780.8	8027.7
16	133.6	516.1	1338.5	1945.7	2963.7	3626.0
17	116.5	263.8	486.1	921.8	2257.7	3929.0
18	161.1	897.0	1208.8	2283.5	4852.9	5970.7
19	161.4	888.4	1818.6	2283.8	2759.7	3328.8
20	237.5	1183.0	2670.9	3470.2	5146.5	5790.5
21	454.0	1273.7	3209.7	4430.3	5874.1	7400.3
22	563.4	855.8	1196.8	1571.7	2105.6	2679.8
23	57.9	311.1	815.9	1499.0	1896.2	2326.7
24	290.7	585.3	1116.4	1421.4	1792.8	2587.2
25	145.0	375.7	609.1	1122.7	1508.0	1955.4
26	299.8	771.2	1246.4	1775.7	2367.7	3035.4
27	438.1	1301.0	2193.0	3255.4	4379.6	5636.4
28	46.3	152.8	264.6	514.8	783.8	1144.3
29	342.0	608.2	877.2	1201.9	1750.2	2391.5
30	124.1	281.3	496.4	752.6	1087.4	1511.9
average	**697.3**	**1610.5**	**2661.0**	**3868.4**	**5181.0**	**6590.4**

DESCRIPTORS FOR ABUNDANCE VARIATIONS

In this section we will describe construction of map invariants that will incorporate information on spot abundance. Recall that the numerical characterizations based on Figures 2 and 3 for both the nearest and the sequential neighborhood approaches used solely information on the x, y coordinates of a proteomics map. It is important to incorporate into the numerical characterization of proteome

Table 8. Average Row Sums for a Group of Five Spots from (1−5) to (25−30) as NN Increases from NN = 1 to NN = 6

rows	NN = 1	NN = 2	NN = 3	NN = 4	NN = 5	NN = 6
1−5	2434.9	4472.9	6500.0	8503.7	10333.6	12048.7
6−10	585.4	1630.2	3107.7	5199.9	7716.2	10325.7
11−15	449.1	1506.8	2448.9	3816.4	4731.1	6505.5
16−20	162.0	749.7	1504.6	2181.0	3596.1	4529.0
21−25	303.2	680.3	1389.6	2009.0	2536.3	3389.9
26−30	250.0	622.9	1015.5	1500.1	2073.7	2743.9

difference	NN = 1	NN = 2	NN = 3	NN = 4	NN = 5	NN = 6
(1−5)−(6−10)	1849.5	2842.7	3392.3	3303.8	2617.4	1723.0
(6−10)−(11−15)	136.3	123.4	658.8	1383.5	2985.1	3820.2
(11−15)−(16−20)	287.1	757.1	944.3	1635.4	1135.0	1976.5
(16−20)−(21−25)	−140.2	69.4	115.0	172.0	960.8	1139.1
(21−25)−(26−30)	52.2	57.4	374.1	508.9	561.9	645.0

maps also the information on the abundance of proteins in a proteome. We will do this by following the scheme outlined in ref 11 in which a two-component vector is constructed for each protein spot of the control proteomics map: one component is the row sum corresponding to individual spot and the other component is a *suitably* scaled abundance of the spot. We have emphasized “suitably” in order to draw attention of users to the fact that the coordinates *x*, *y*, and *z* each express a different protein property and are measured in different and mutually arbitrary units. In order not to give preference to any of the coordinates, Kowalski and Bender[25] recommended that all quantities be independently scaled on the interval (−1, +1). This introduces an effective way of assigning to all the three quantities similar weights. Because in our case already the *x* and *y* coordinates are of the same order of magnitude, we have adjusted the experimentally reported abundances by simply dividing the values by 100, which gave the abundances in the same range of values as those found for *x* and *y* in Table 2.

In Table 9, we show the magnitudes of 30 two-component vectors for cases NN = 1−6, and the last row of Table 9 now shows the average values for the $N = 30$ spots considered. The last row of Table 9 is illustrated in Figure 5 again displaying a high-quality quadratic dependence of the average on the number of the nearest neighbors, showing the same regularity that we have already seen in Figure 4. The entries in the last row of Table 9, the sequence 1188.0, 1889.4, 2836.6, 3987.2, 5267.2, 6656.4, ..., is one set of our map descriptors. If we divide each entry in the above sequence by the corresponding NN value, we obtain the sequence 1188.0, 944.7, 945.5, 996.8, 1053.4, 1109.4, ... by changing entries. In Table 10, we compare the above sequence with the corresponding sequences obtained by considering from $N = 21$ to $N = 30$ spots. As we see, the novel map invariants are rather stable and do not change dramatically by inclusion or exclusion of a few spots from the analysis, which is a very desirable property for map descriptors. The average values of the two-component vectors for cases NN = 1−6 gradually decrease because as *N* increases, the abundance of spots also decreases.

Table 9. Magnitude of Two-Component Vectors Involving Abundance Values

	NN = 1	NN = 2	NN = 3	NN = 4	NN = 5	NN = 6
1	4295.0	4295.0	6774.5	7956.4	10465.9	13482.5
2	2984.6	5537.8	6409.3	8194.9	8735.3	11136.3
3	3466.8	6635.0	6969.0	8465.0	9049.7	9049.7
4	1769.9	4386.0	7410.6	8458.8	10555.2	13647.4
5	1615.8	2570.9	5627.2	9980.7	13314.5	13314.5
6	1269.2	2570.1	4491.5	8316.2	11350.8	13457.2
7	1305.8	2370.2	3226.5	5055.8	8956.8	11038.3
8	1426.7	1756.4	3147.1	4430.0	6020.7	8103.4
9	1011.6	1366.3	2789.1	4578.1	6363.3	10344.0
10	1089.5	1715.0	2771.1	4174.2	6264.1	8961.2
11	998.7	1904.1	2317.9	5131.5	5877.0	6876.0
12	873.8	925.0	1366.9	2721.3	3496.6	4912.4
13	932.8	1589.7	2362.8	2945.1	3641.6	5763.4
14	1028.6	2087.5	3361.3	4238.8	5204.1	7194.4
15	1080.7	2415.7	3680.7	4547.2	5838.6	8069.5
16	811.2	952.1	1559.5	2103.8	3069.8	3713.3
17	807.2	841.1	935.0	1219.7	2394.8	4009.4
18	745.5	1155.2	1411.0	2396.7	4907.2	6014.9
19	739.6	1144.6	1956.6	2395.1	2852.5	3406.1
20	734.0	1371.8	2759.8	3539.0	5193.2	5832.0
21	815.8	1442.8	3280.5	4481.8	5913.1	7431.3
22	857.8	1072.8	1360.4	1699.6	2202.7	2756.8
23	613.5	685.4	1019.2	1618.6	1992.1	2405.5
24	660.4	833.2	1264.1	1540.1	1888.4	2654.3
25	607.3	699.3	847.9	1268.2	1619.3	2042.4
26	652.9	965.0	1374.7	1868.0	2437.7	3090.3
27	722.1	1422.0	2266.9	3305.6	4417.1	5665.5
28	556.0	574.7	614.0	756.3	959.9	1271.4
29	638.3	812.6	1029.6	1317.2	1831.3	2451.4
30	528.3	585.5	714.2	911.1	1202.5	1596.7
average	**1188.0**	**1889.4**	**2836.6**	**3987.2**	**5267.2**	**6656.4**

Table 10. Gradual Decrease of the Average Row Sums as an Additional Spot Is Added to Map

	NN = 1	NN = 2	NN = 3	NN = 4	NN = 5	NN = 6
21	1449.3	2379.5	3566.4	5042.4	6677.6	8416.3
22	1419.2	2334.9	3552.7	5015.7	6641.2	8369.4
23	1393.7	2277.5	3453.1	4865.0	6439.4	8114.3
24	1359.7	2208.3	3347.3	4723.9	6246.1	7866.1
25	1330.6	2151.0	3260.5	4591.2	6064.5	7648.9
26	1276.7	2049.5	3095.2	4358.6	5754.0	7257.9
27	1256.2	2026.3	3064.5	4319.6	5704.5	7199.0
28	1231.2	1974.4	2977.0	4192.4	5535.5	6987.3
29	1210.7	1934.4	2909.8	4093.2	5407.3	6830.9
30	1188.0	1889.4	2836.6	3987.2	5267.2	6656.4

DISCUSSION

Currently experimental data are represented as lengthy tables of protein coordinates and abundances (such as shown in our Table 2). Although such tables can be screened by computers, the information in such tables is simply a collection of numerous data on individual proteins. In contrast, our approach focuses on data that represents global information on individual maps, not individual proteins. The underlying concept here is, one may say, modification of old paradigm that can be traced to Paul Erlich in that similar drugs will produce similar effects. In our context, similar proteomics maps will be described by a set of similar descriptors. That being accepted as a working hypothesis, what our algorithms offer is the possibility to catalog and search large files of proteomics maps for finding maps that are more similar to a target map. To facilitate such developments (in which clearly many laboratory ought to participate), we have outlined in this paper (in the next section) a computer program that will compute set of map descriptors used for comparisons of different maps.

We do not expect that any of the five available approaches (of Table 1) for characterization of proteomics maps will lead to significantly different conclusions when applied to the same data. The advantage, however, of the present approach is that if one completes an analysis by setting let us say $N = 30$ and then later decides to reconsider analysis and choose $N = 40$, the derived numerical descriptors for the two cases will not change as much as they would with other methods

Any of the available numerical characterizations of proteomics map listed in Table 1 can be applied to include large number of spots. It is true that in most past illustrations we have used from 20 to 30 spots, merely to outline the methodology. In some work we used up to 100 spots, but there are no inherent restrictions on N.

THE PROMISE PROGRAM

All calculations of map invariants were obtained by using an in-house program PROMISE (PROteomics Map Invariants SEarch). We feel that the acronym name has been well selected in view that from the five evolutionary methodologies so far developed for construction of map invariants (listed in Table 1), the methodology outlined here, the last on the list of Table 1, appears the most promising. In Figure 6, we show a chart that gives an overview of the program PROMISE. As we can see from the chart, the input data for quantitative characterization of proteomics maps are the experimentally reported x, y, and z coordinates. Before proceeding with the analysis, one should select a scaling procedure to ensure that all measured quantities are expressed by numbers of the same order of magnitude. Besides the scaling advocated by Kowalski and Bender,[25] one can consider dividing all quantities by the maximal entry (maximal x for variable x, maximal y for variable y, and maximal z for variable z), or by the average entry. After spots have been ordered according to their relative abundance, one constructs the (x, y) distance matrix. The prime reason for not constructing the (x, y, z) distance matrix is that in this way all proteomics map originating from the same tissue, organ, and species but associated with different experimental regimes (toxins or other causes for change of proteome abundance) will have the same "parent" proteome map, that of the control group.

From the distance matrix, one forms a list of NN nearest neighbors. Once the domain of the NN has been selected (in this paper, we used NN = 1−6), the program (written in Excel) constructs the corresponding distance-adjacency matrix (D/A). The D/A matrix, in contrast to the distance matrix, keeps information only on distances of vertices in a graph, which are adjacent. An advantage of the D/A matrix is that it is sparse matrix (that is a matrix having numerous zero elements) and when manipulated (used for construction of "higher order" matrices or solving for eigenvalues) it does not introduces an unusual amount of computations. In the final stages of computations, from the average row sums of the D/A matrix, two-component vectors are constructed, which will differentiate among proteomics maps associated with different experimental conditions.

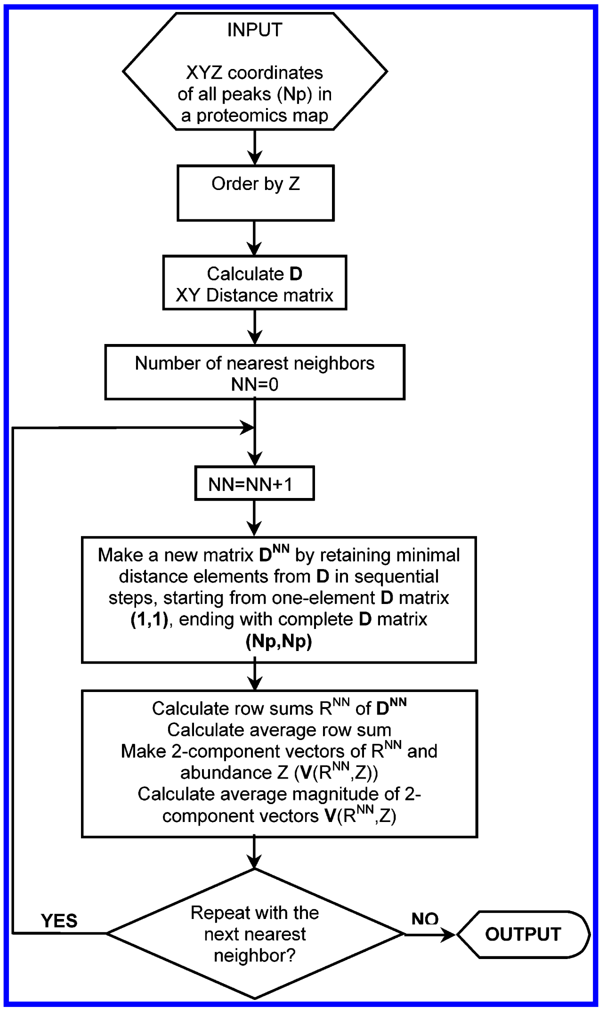
INPUT
XYZ coordinates of all peaks (Np) in a proteomics map
Order by Z
Calculate **D** XY Distance matrix
Number of nearest neighbors NN=0
NN=NN+1
Make a new matrix $\mathbf{D}^{NN}$ by retaining minimal distance elements from **D** in sequential steps, starting from one-element **D** matrix **(1,1)**, ending with complete **D** matrix **(Np,Np)**
Calculate row sums R^{NN} of $\mathbf{D}^{NN}$
Calculate average row sum
Make 2-component vectors of R^{NN} and abundance Z ($\mathbf{V}(R^{NN},Z)$)
Calculate average magnitude of 2-component vectors $\mathbf{V}(R^{NN},Z)$
Repeat with the next nearest neighbor?
YES
NO
OUTPUT

Figure 6. Flow chart of the program PROMISE.

CONCLUDING REMARKS

We have outlined a novel approach for quantitative numerical characterizations of proteome maps based on the idea of neighborhood graphs, but differing from the earlier outlined scheme in an important feature ensuring that the present approach is numerically robust and not highly sensitive to the number of spots selected for analysis. The present approach incorporates some of the features of other schemes, like the zigzag approach that is based on ordering of spots according to their abundance and like the cluster method that results in graphs that are connected even for small number of the nearest neighbor considered. An important finding is that the approach apparently is not so sensitive to the number of the nearest neighbors considered, because, as can be seen from Figures 4 and 5, the presence of additional neighbors does not necessarily introduce novel information: the row sums for higher NN values can be predicted form the known values of average row sums for smaller NN. In view of the robustness of the approach, which was not the case with the scheme based on the nearest neighborhood graphs, we believe that the present scheme has all the advantages of the former but in addition has important advantages and thus appears promising.

ACKNOWLEDGMENT

M.R. thanks the National Institute of Chemistry of Slovenia and Professor Jure Zupan of Chemometrics Laboratory for cordial hospitality. This work has been supported by Ministry of Higher Education, Science and Technology of Republic of Slovenia under Research Grant P1-017. We also thank Professor A. T. Balaban (Galveston, TX) for kindly reviewing the manuscript and suggesting improvements in the presentation of our results.

REFERENCES AND NOTES

(1) Fields, S. Proteomics in genomeland. *Science* **2001**, *291*, 1221−1224.

(2) Randić, M. On graphical and numerical characterization of proteomics maps. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 1330−1338.

(3) Randić, M.; Zupan, J.; Novič, M. On 3-D graphical representation of proteomics maps and their numerical characterization. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 1339−1334.

(4) Randić, M.; Witzmann, F.; Vračko, M.; Basak, S. C. On characterization of proteomics maps and chemically induced changes in proteomics using matrix invariants: Application to peroxisome proliferators. *Med. Chem. Res.* **2001**, *10*, 456−479.

(5) Randić, M. A graph theoretical characterization of proteomics maps. *Int. J. Quantum Chem.* **2002**, *90*, 848−858.

(6) Randić, M.; Basak, S. C. A comparative study of proteomics maps using graph theoretical biodescriptors. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 983−992.

(7) Randić, M.; Lerš, N.; Vukičević, D.; Plavšić, D.; Gute, D. B.; Basak, S. C. Canonical labeling of Proteome Maps. *J. Proteome Res*. (in press).

(8) Bajzer, Ž.; Randić, M.; Plavšić, D.; Basak, S. C. Novel matrix invariants for characterization of toxic effects on proteomics maps. *J. Mol. Graphics Modell.* **2003**, *22*, 1−9.

(9) Randić, M. Quantitative characterization of proteomics maps by matrix invariants. In *Handbook of Proteomics Methods*; Conn, P. M., Ed.; Humana Press: Totowa, NJ, 2003; pp 429−450.

(10) Randić, M.; Zupan, J.; Novič, M.; Gute, B. D.; Basak, S. C. Novel matrix invariants for characterization of changes of proteomics maps. *SAR QSAR Environ. Res*. **2002**, *13*, 689−703.

(11) Randić, M.; Lerš, N.; Plavšić, D.; Basak, S. C. Characterization of 2-D proteome maps based on the nearest neighborhoods of spots. *Croat. Chem. Acta* **2004**, *77*, 345−351.

(12) Randić, M.; Lerš, N.; Plavšić, D.; Basak, S. C. On invariants of a 2-D proteome map derived from neighborhood graphs. *J. Proteome Res*. **2004**, *3*, 778−785.

(13) Bajzer, Ž.; Basak, S. C., Vračko Grobelšek, M.; Randić, M. Use of proteomics based biodescriptors in the characterization of chemical toxicity. In *Genomic and Proteomic Applications in Toxicity Testing*; Cunningham, M. J., Ed.; Humana Press: Totowa, NJ, 2005.

(14) According to G. Birkoff (as mentioned in ref 13), the idea of partial order can be traced to W. Leibniz. The set of inequalities that define dominance for sequences are due to Muirhead (ref 16). The diagram representation of partial ordering is known as Hasse diagram (ref 17).

(15) Klein, D. J. Prolegomenon on partial ordering in chemistry. *MATCH* **2000**, *42*, 7−21.

(16) Muirhead, R. F. *Proc. Edinburgh Math. Soc*. **1903**, *21*, 144.

(17) Hasse diagram, see: Weisstein, E. W. Hasse diagram. From *MathWorld*—A Wolfram web resource. http://mathworld.wolfram.com/HasseDiagram.html.

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

(19) Dijkstra, E. W. www.nist.gov/dads/HTLM/dijkstralgo.html (dads = Dictionary of Algorithms and Data Structures).

(20) Witzmann, F. Molecular Anatomy Laboratory, Department of Biology, Indiana University and Purdue University, Columbus, IN 47203.

(21) Randić, M.; DeAlba, L. M. Dense graphs and sparse matrices. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 1078−181.

(22) Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc*. **1947**, *69*, 17−20.

(23) Hosoya, H. Topological index. A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. *Bull. Chem. Soc. Jpn*. **1971**, *44*, 2332−2339.

(24) Randić, M.; Bajzer, Ž.; Vračko, M.; Novič, M.; Zupan, J. (manuscript in preparation).

(25) Kowalski, B. R.; Bender, C. F. A powerful approach to interpreting chemical data. *J. Am. Chem. Soc*. **1972**, *94*, 5632−5639.

(26) Randić, M.; Novič, M.; Vračko, M. On characterization of dose variations of 2-D proteomics maps by matrix invariants. *J. Proteome Res*. **2002**, *1*, 217−226.

CI0497612