On the Dependence of a Characterization of Proteomics Maps on the Number of Protein Spots Considered

Milan Randić,*,† Frank A. Witzmann,‡ Varshna Kodali,§ and Subhash C. Basak§

National Institute of Chemistry, Ljubljana, Slovenia, Department of Cellular and Integrative Physiology, Indiana University School of Medicine, Indianapolis, Indiana 46202, and Natural Resources Research Institute, University of Minnesota, Duluth, Minnesota 55811

Received April 12, 2005

We have reexamined the numerical characterization of proteomics maps based on the construction of novel distance matrices associated with the nearest neighbor graph for the protein spots. In particular we consider dependence of a characterization of proteomics map on the number of proteins considered in the analysis. We examined a collection of proteomics maps in which we approximately doubled the number of spots to be used for quantitative analysis, considering cases of maps having 30, 50, 100, 250, 500, and 1054 protein spots. For each case we have compared the similarity−dissimilarity results for five proteomics maps of rat liver cells associated with the control group and four proliferators administrated by intraperitoneal injection. We found that proteins maps based on a set of about the 250 most abundant proteins spots suffice for a satisfactory numerical characterization of such maps.

INTRODUCTION

Today research in many areas of chemistry concerns highly complex systems and results in highly complex experimental and computational outputs which immediately call for further processing in order to allow an interpretation of such results. To make things worse, continuing advances in instrumentation (automatization and computerization) produces by far too fast of an abundance of data that waits to be digested. It is obvious that unless we develop novel efficient approaches for data processing for some time much of the useful information may remain hidden in the plentitude of raw data. In contrast to obtaining experimental physicochemical data with standard instruments of the past, which typically presented results either as a list of numbers or spectral charts, experiments on biological systems often produce very large and complex outputs, which require further processing. Clearly a different kind of experimental data will require different considerations. In our previous work on quantitative characterization of DNA[1−15] we have shown how one can associate with lengthy DNA sequences a set of numerical descriptors that allow quantitative characterization and quantitative comparison of such sequences. More recently this work has been extended to graphical and nongraphical representation and accompanying numerical characterization of proteins, again based on mathematical invariants extracted for such sequences.[16−20] On the other hand in the case of 2-D proteomics maps,[21−35] which besides visual format are also reported by lengthy lists of (x, y, z) coordinates, where (x, y) give the location and (z) the abundance of proteins spots on 2-D gel, similarly it has been possible to construct a set of mathematical invariants, or map descriptors, which characterize 2-D gel maps and allow their quantitative comparative study. However, it is interesting that even though information on DNA represents one-dimensional sequential data, while a proteomics map constitutes two-dimensional data, mathematical characterizations of both of them have some common and overlapping features. Thus not long ago it was shown that lengthy strings of four letter sequences, which represent the experimental output on DNA, can be also represented as 2-D maps. This then allows the use of approaches that were initially developed for characterization of maps to be used for characterization of DNA sequences. For example, by using a spiral representation of a DNA sequence one obtains a rather condensed 2-D binary map of DNA in which distances between bases are given by integers.[8,36] Alternatively, again using the spiral format for writing a DNA sequence one can obtain "four color" maps of DNA, in which the drawing plane is divided into local regions belonging to the four bases.[37,38] One can also arrive at a 2-D graphical representation of codons by applying Jeffrey's ingenious scheme for inscribing lengthy DNA sequences within a square to triplets of bases. [39] This leads to a 8×8 Table of Codons,[11] which offers an extension of the graphical representation of DNA to proteins and thus produces quantitative characterization of proteins.[18]

In this article we will focus our attention on a particular computational aspect of the quantitative characterization of proteomics maps. We will consider the role of the number of protein (spots) considered in an analysis on characterization of maps as a whole, that is, the role of the truncation of data in the construction of reduced maps that include a limited number of protein spots. As has been outlined in several of our publications a way to arrive at the quantitative characterization of proteomics maps is first to associate with a proteomic map suitable 2-D geometrical object, which will allow an adequate mathematical characterization of the map. Such are various geometrical objects of fixed geometry. The

* Corresponding author e-mail: mrandic@msn.com. Permanent address: 3225 Kingman Road, Ames, IA 50014.
† National Institute of Chemistry.
‡ Indiana University School of Medicine.
§ University of Minnesota.

10.1021/ci050132h CCC: $33.50
Published on Web 12/27/2005

first numerical characterization of a proteomics map was based on the construction of a zigzag curve connecting protein spots in order of their decreasing abundance.[21,22] Later additional geometrical objects for the characterization of proteomics maps were considered such as graphs of partial ordering,[24] graphs depicting the clustering of spots,[29] and graphs depicting various neighborhoods of protein spots.[30,31,34] In all these cases one has to select a single dominant parameter: the number N of protein spots of a map that one includes in the analysis.

In most previous studies on the characterization of proteomics maps N was selected, somewhat arbitrary, to be in the range 20−30, up to 100. The essential difference between approaches based on the zigzag curve, the graph of partial ordering, the cluster graph, and the graphs representing neighborhoods of protein spots using the same number of spots N is in their "density", that is, in the number of "lines" that such geometrical objects involve. Preliminary results appear to indicate that the number of "lines" used for given N is not a critical factor for the characterization of proteomics maps. In particular the use of the neighborhood graphs allowed one to vary for a given N the number of lines used for the characterization of a proteomics map. Unexpectedly and somewhat surprisingly it turned out that variation in the number of lines in the neighborhood graphs makes predictable changes for the map invariants considered,[30] which were the average row sum of accompanying distance/distance matrices. We decided therefore in this study to direct attention to the role of the variable N and to find out how the selection of N influences the relative magnitude of selected map invariants (descriptors).

We selected to consider five proteome maps associated with the liver cell of mice. The control group was represented by healthy mice, and the other four groups of mice were subjected to certain doses of four proliferators: PFOA, PFDA, clofibrate, and DEHP. The five proteomics maps used in this work involving the four proliferators have been previously used in several of our publications.

A comment concerning "small spots" (that is the protein spots of rather low abundance) that have been neglected in the previous studies as well as in this approach is in place. Could these neglected spots nevertheless involve some essential hidden information? Well, the "small spots" can be included and treated as any other protein spots, provided they are detected and quantified in all samples considered with sufficient accuracy. In fact, as will be seen in this contribution, which for the first time in quantitative characterization of proteomics maps includes over 1000 protein spots, there is some uncertainty in the precision of experimental data reported for spots of low abundance. Thus the inclusion of "small spots" may be questioned from the point of view of reliability of conclusions based on their incorporation in the analysis. The distinction between "small spots" versus "large spots" is a function of the resolution of spots and their sensitivity to staining. The small spots of *very* low abundance challenge the linearity of the staining procedure, which in this case Coomassie blue stain is linear approximately only for 2 orders of magnitude. The spots of *very* low abundance being at the threshold of detection by the stain may not accurately reflect their *actual* abundances and thus inaccurately predict chemical effects. As is true in any quantitative proteomics technique, very low abundance proteins are not as reliable as those quantified with moderate to high abundance, no matter how the data are to be used subsequently.

EXPERIMENTAL SECTION

Animal Care and Intoxication. Male Fisher-344 rats (225−250 g) were obtained from Charles River Breeding Labs, individually housed, and maintained on rat chow and water ad libitum. Rats were injected intraperitoneally with either of the following:

(1) 2 mg, 20 mg, or 50 mg PFDA/kg body weight; single injection; animals sacrificed on day 8 of exposure;

(2) 150 mg PFOA/kg (perfluorooctanoate, PFDA's eight carbon analog); single injection; animals sacrificed on day 3 of exposure;

(3) 250 mg clofibrate/kg; single injection on each of 3 successive days; animals sacrificed on day 5 of exposure;

(4) 1200 mg DEHP/kg (di[2-ethylhexyl]phthalate); single injection on each of 3 successive days; animals sacrificed on day 5 of exposure.

Sample Preparation. After the appropriate exposure period, livers were surgically removed from the ketamine/xylazine anesthetized rats and manually perfused with ice-cold saline to remove excess blood. One 0.5 g piece was removed, minced, and homogenized in 8 volumes (4 mL) of a lysis buffer containing 9 M urea, 4% NP-40, 2% DTE (dithioerythritol), and 2% ampholytes (Serva pH 9−11) pH 9.5 for 2-DE.

Two-Dimensional Electrophoresis. The strength of the 2-D electrophoretic technique lies in its ability to resolve literally thousands of cellular proteins first based on their content of acidic and basic amino acids (isoelectric focusing) and second by molecular weight (SDS electrophoresis). In combination, these two separation techniques produce a two-dimensional protein pattern unique for a specific group of cells/tissues. Individual proteins within the pattern can be analyzed for alterations in volume (density), charge, and molecular weight. Changes in volume or spot density reflect alterations in a protein's abundance and suggest up- or down-regulation of the genome or altered protein turnover rates. Charge modifications suggest either post-translational modification alterations or point mutations in the genome.

Ten microliters of the solubilized protein sample was placed on each of 20 first dimension gels (25 cm × 1.5 mm) containing 4% acrylamide, 2% NP-40, and 2% ampholyte (BDH pH 4−8) and electrophoresed for 32 000 VHr at room temperature. Each first dimension gel was then placed on a second-dimension DALT slab gel (20 cm × 25 cm × 1.5 mm) containing a linear 9−17% acrylamide gradient. Gradient slab gels were poured reproducibly using the ANGELIQUE computer-controlled gradient maker (Large Scale Biology Corp.). This system enables one to reduce run-to-run variability in the polyacrylamide gel concentration, an essential characteristic for protein pattern image analysis. Slab gels were run for 18 h at 4 °C and later stained with Coomassie brilliant blue G-250.

Image Analysis. Stained gels were digitized at 125 micron resolution using an Ektron 1412 CCD scanner that produces 8 bit images in the optical density domain with up to 2048 × 2048 pixels although most images were 1800 × 2000 pixels on a side. The gel images were processed on a DEC

Table 1. *x, y* Coordinates for the Ten Most Abundant Proteins Spots of the Control Group, the Abundance, and Abundance of the Same Spots for Four Proliferators

x	*y*	control	PFOA	PFDA	clofibrate	DEHP
2111.7	2278.6	144357	108713	95028	147081	165886
2804.3	903.6	143630	155565	188582	159898	155055
1183.9	959.6	136653	113859	150253	163645	8111
2182.2	928.8	127195	99160	73071	76642	112096
2685.6	1196.1	118581	112790	49769	109856	138795
1527.9	825.5	114929	192437	221567	166080	180590
1346.0	1352.5	112251	58669	38915	73159	77075
2868.5	778.0	108883	26105	50735	45923	116849
1406.3	1118.1	98224	91147	82963	84196	92942
2646.6	1288.0	94128	100030	84933	100303	143490

VAXStation 3100 M76 workstation using the KEPLER software system (Large Scale Biology Corp.) with procedure PROC008 which includes background and streak subtraction, erosion/dilation spot cutout, and 2-D Gaussian fitting to generate a spotlist giving *x, y* position, shape, and density information for each detected spot. Groups of numerous sample gels corresponding to all the animal treatment groups were assembled and matched to a standard master pattern for that particular experiment and a set of running conditions. This master pattern was initially constructed from a representative clone, and additional spots were added that were detected in other gel image clones to develop a master pattern containing all detectable protein spots. Individual gel patterns were scaled together using a linear fit to the abundances of selected matched spots to compensate for variations in sample protein loading. Raw abundance data for each matched protein spot on the various maps was exported to Excel for further mathematical analysis.

Outline of the Calculations. In Table 1 in the first three columns we have listed for the control group the 10 most abundant points of the proteome map having 1054 listed spots with their *x, y* coordinates and the relative abundance, all based on arbitrary units for distances measured over the 2-D gel and the density of the spots. The remaining four columns of Table 1 show the corresponding abundance values for animals exposed to four different proliferators: PFOA, PFDA, clofibrate, and DEHP. Because the units used for measuring the abundance values produce numbers by 2 orders of magnitude greater than the range of the *x* and *y* coordinate, following the recommendation of Bender and Kowalski[40] we normalized the data of Table 1. However, instead of rescaling the values of all three variables in the interval of $(-1, +1)$, which is the procedure advocated by Bender and Kowalski, which gives the three variables *x, y, z* approximately the same relative weight, we have scaled the coordinates by dividing the original coordinates of the spots by the maximal Euclidean distance between two spots on the map.

The first step in preparation for the numerical characterization of the proteomics map is construction of the graph of the nearest neighbors. This is easy to construct once one calculates the distance matrix for the protein spots. For example, when $N = 30$ as we can see from Figure 1 the nearest neighbor of spot 1 is spot 27; of spot 2 is spot 8; of spot 3 is spot 19; and so on. In this way by connecting the nearest neighbors we obtain the graph of the nearest neighbors for any value of *N*. In the next step we constructed a Weighted Adjacency Matrix (WA matrix) for the graph of the nearest neighbors, which is defined as follows:

$$\mathrm{WA}(i, j) = \text{Euclidean distance between } i \text{ and } j, \quad \text{if adjacent;}$$

$$\mathrm{WA}(i, j) = 0 \text{ otherwise}$$

In other words the WA matrix is a hybrid between the graph adjacency matrix and the distance matrix, where the distances are given by the length of the connections between adjacent

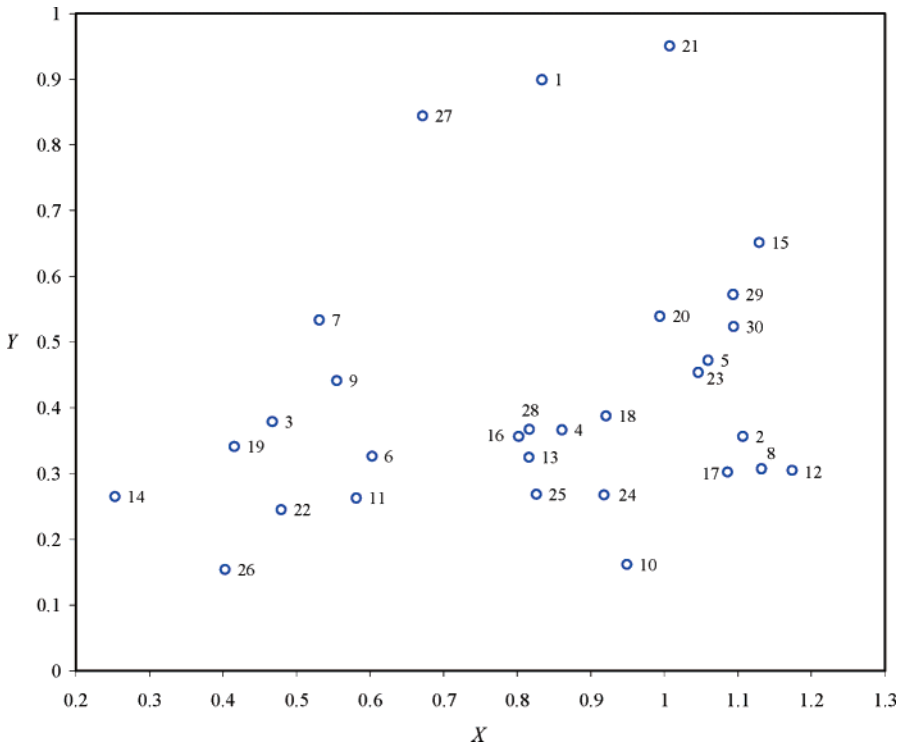

Figure 1. Simplified proteomics map with $N = 30$ the most abundant protein spots.

spots. Observe that the graph of the nearest neighbors, when one considers for each spot only one of its nearest neighbors, is as a rule disconnected. The consequence of this is that matrices associated with such graphs can be factored in smaller matrices, which represent individual components (fragments of the map). This means that the eigenvalues of such a matrix, including also the leading eigenvalue, will characterize individual fragments of the map, rather than the map as a whole. We, of course, are interested in map descriptors for the map as a whole. Because of this we have selected the average row sum of the matrix as the map descriptors of choice. As is known from algebra, the largest row sum and the smallest row sum (or the column sums, which for symmetric matrices are the same) of a matrix represent the upper and the lower bound on the leading eigenvalue of the matrix. In view of this the average row sum will have the magnitude in the same range as the leading eigenvalue, would the matrix belong to a connected graph (what happens when one considers several nearest neighbors of each spot).

As a map descriptor we will take the magnitude of a 2-component vector constructed as the average value of contributions of individual spots, the contributions of which are obtained by constructing for each spot a vector with two components: One component of the vector is given by the row sum belonging to the spot and the second component is given by the abundance of the spot, that is, as

$$(1/N)\Sigma \sqrt{\{[\text{row } \Sigma(i)]^2 + [\text{abundance}(i)]^2\}}$$

In this way we obtain in the case of $N = 30$ the following map descriptors:

control group	0.623987
PFOA	0.567060
PFDA	0.498219
clofibrate	0.598282
DEHP	0.712407

Just a glance on these numbers suggests, if we assume that the descriptor considered is a measure of the similarity between maps, that the clofibrate proteomics map is the most similar to the original (control group) map, while the maps of the remaining three peroxisome proliferators show a larger departure from the map of the control group. In general a single map invariant contains rather limited information, and if we want to make more conclusive statements we have to use additional map invariants. However, as we will see in our case already a single invariant can give useful information on the degree of similarity of proteomics maps. Our focus in this study is to see how map invariants may depend on the amount of information carried by a different number of spots, that is, the sensitivity of the numerical characterization of proteomics maps on N, the number of protein spots selected for the numerical analysis.

Sensitivity of Map Invariants on the Number of Spots. To give a visual impression on calculations made in Figure 2 we illustrate simplified forms of three proteomics maps having $N = 50$, $N = 100$, and $N = 250$ belonging to the same proteomics map of the control group having $N = 1054$ spots. The question that we want to address is as follows: Is there a number of spots of a proteomics map that suffices to characterize proteomics maps adequately. In other words

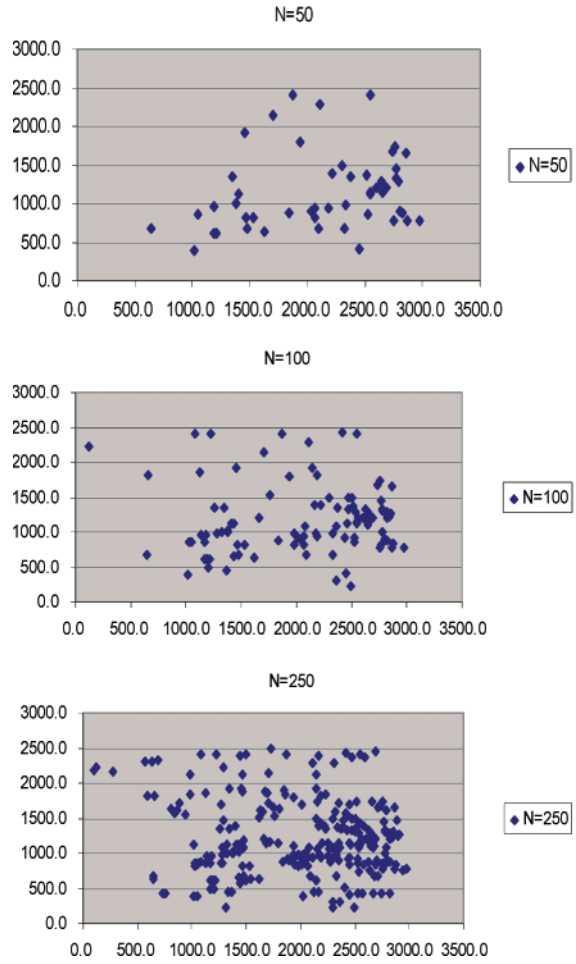

Figure 2. The same protein map depicted for a variable number of protein spots for $N = 50$, 100, and 250.

do we need all 1054 listed spots (and their abundances) for a useful comparative study of proteomics maps? Or could 50 or 100 suffice? We will see that we can obtain a definite answer to the question. On one side the answer to the question can indicate whether one can base calculations on a few data. If that would be possible this would result in a considerable increase in the efficiency of analysis and comparisons of proteomics maps when one considers a large data set. On the other side as the number of spots to be included increases their relative abundance decreases, and thus their inclusion introduces data associated with a greater relative experimental error in measurements of the density of spots, which makes analysis less reliable. Hence, one has to find a balance between using limited but more reliable data and using most of the data available regardless their relative quality.

To get some insight into the sensitivity of the numerical characterizations of proteomics maps on N we constructed for each of the five proteomics maps the WA matrices associated with $N = 30$, $N = 50$, $N = 100$, $N = 250$, $N = 500$, $N = 750$, and $N = 1054$. For each of them we constructed two component vectors, the magnitude of which are listed in Table 2.

The question to consider now is do the vectors of Table 2 for different N's have similar information content. To answer this question we will construct the similarity−dissimilarity table based on the magnitudes of two-component vectors. In Table 3 we have listed entries of the symmetrical

Table 2. Magnitudes of Two-Component Vectors Representing Individual Proteomics Maps

	$N = 30$	$N = 50$	$N = 100$	$N = 250$	$N = 500$	$N = 750$	$N = 1054$
control	0.6240	0.5043	0.3578	0.2076	0.1291	0.0952	0.0717
PFOA	0.5671	0.4588	0.3218	0.1909	0.1205	0.0907	0.0695
PFDA	0.4982	0.4248	0.3006	0.1772	0.1132	0.0862	0.0676
clofibrate	0.5983	0.4975	0.3543	0.2150	0.1333	0.0994	0.0751
DEHP	0.7124	0.5734	0.4159	0.2437	0.1494	0.1114	0.0841

Table 3. Symmetrical Similarity−Dissimilarity Matrix Based on the Magnitudes of the Vectors of Table 2 for $N = 30$[a]

		control A	PFOA B	PFDA C	clofibrate D	DEHP E
control	A	0	0.0569	0.1258	0.0257	0.0884
PGOA	B	S_{AB}	0	0.0688	0.3122	0.1453
PFDA	C	S_{AC}	S_{BC}	0	0.1001	0.2142
clofibrate	D	S_{AD}	S_{BD}	S_{CD}	0	0.1141
DEHP	E	S_{AE}	S_{BE}	S_{CE}	S_{DE}	0

[a] Numerical values are shown above the diagonal and the corresponding labels below the diagonal.

Table 4. Nonzero Entries of the Similarity−Dissimilarity Matrix Based on the Magnitudes of the Vectors of Table 2 for $N = 30$ to $N = 1054$

	30	50	100	250	500	750	1054
S_{AB}	0.0569	0.0455	0.0361	0.0167	0.0086	0.0044	0.0022
S_{AC}	0.1258	0.0795	0.0573	0.0305	0.0158	0.0089	0.0041
S_{AD}	0.0257	0.0067	0.0035	0.0074	0.0042	0.0042	0.0035
S_{AE}	0.0884	0.0692	0.0581	0.0360	0.0203	0.0163	0.0124
S_{BC}	0.0688	0.0340	0.0212	0.0137	0.0072	0.0045	0.0019
S_{BD}	0.0312	0.0387	0.0325	0.0241	0.0128	0.0087	0.0056
S_{BE}	0.1453	0.1146	0.0942	0.0528	0.0289	0.0207	0.0146
S_{CD}	0.1001	0.0727	0.0537	0.0379	0.0200	0.0132	0.0076
S_{CE}	0.2142	0.1486	0.1154	0.0665	0.0362	0.0252	0.0165
S_{DE}	0.1141	0.0759	0.0616	0.0287	0.0161	0.0120	0.0090

similarity−dissimilarity table for the case $N = 30$. As we see there are 10 nonzero entries which fully describe the relative degree of similarity−dissimilarities among the five proteomics maps. The smaller entries in the table point to similarity among the corresponding proteomics maps, while the large entries point to dissimilar maps. As we see from Table 3 and as it was already mentioned in the case $N = 30$ the most similar to the control map is the clofibrate map and the least similar to the control map is the map obtained for PFDA. If we are comparing all five maps among themselves, then we find that PFOA and clofibrate maps are the most similar, while PFOA and DEHP are among the least similar. Will the same conclusions hold if we consider cases $N = 50$, $N = 100$, and others?

To answer these questions we will examine similar similarity−dissimilarity data for the cases of maps for $N = 50$ to $N = 1054$. In Table 4 we have collected the similarity−dissimilarity values for $N = 50$ to $N = 1054$ for the 10 nonzero entries of the corresponding similarity−dissimilarity matrices. We have labeled the control group as A and the four proliferators with B, C, D, and E, respectively. The first numerical column in Table 4 reproduces the nonzero

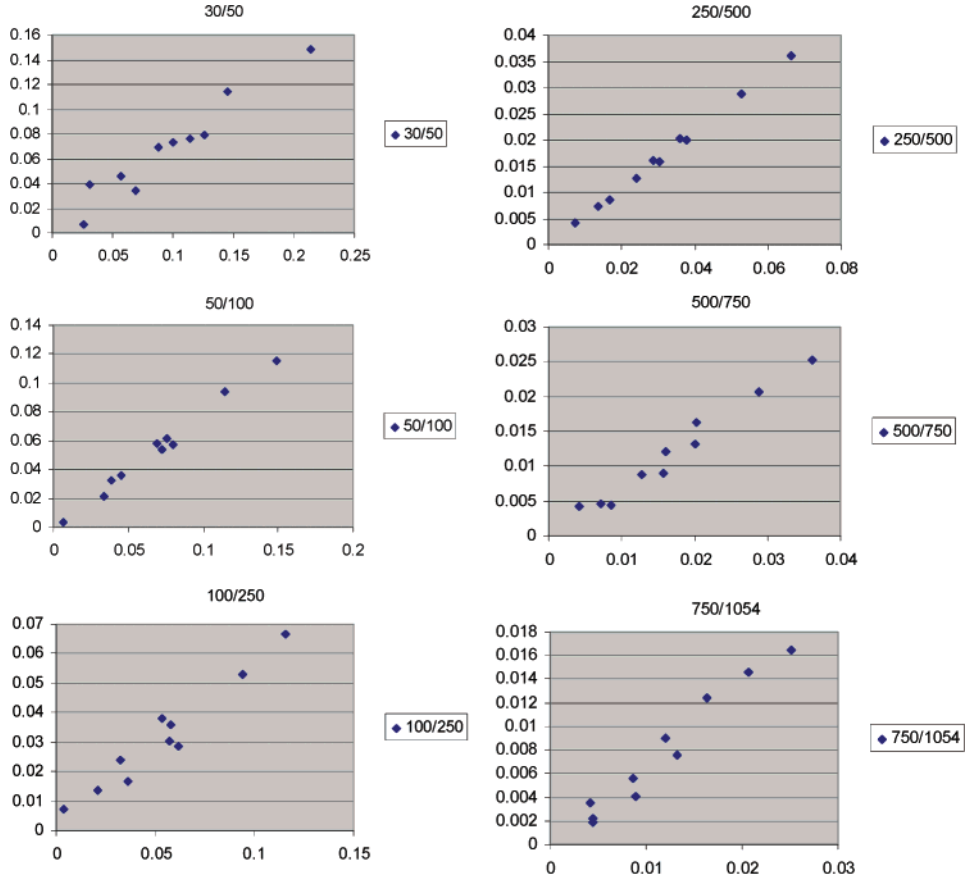

Figure 3. Correlations of the adjacent columns of Table 4 showing the degree of parallelism between maps of the variable number of protein spots.

elements of the similarity−dissimilarity matrix of Table 3, the first element in this column S_{AB} being the similarity−dissimilarity between the control group and PFOA, and the last element in the column S_{DE} being the similarity−dissimilarity between clofibrate and DEHP. The values in each column correspond to the increasing values of *N*. The similarity−dissimilarity values decrease with *N* because as the number of spots increases the distances between the neighboring protein in the 2-D gel maps decrease. To see if the increase of *N* is accompanied with an increase in the information content of the WA matrices we have to examine the entries of Table 4 more closely. If all the numbers in successive columns are proportional, then *N* is not a very critical parameter for the considered numerical characterization of the proteomics map, while if numbers in adjacent columns show considerable scatter the opposite will be the case.

To find out what is the case in Figure 3 we have plotted correlations for adjacent columns: 30/50; 50/100; 100/250; 250/500; 500/750; and 750/1054, the *x* coordinate is shown as the numerator and the *y* coordinate as the denominator. The plots displayed in Figure 3 are very instructive. The best correlation between the adjacent columns of Table 4 happens for the plot 250/500, which shows almost full proportionality between the values of Table 4 in column $N = 250$ and column $N = 500$. That means that after $N = 250$ we have not gained any significant information for characterization of the proteomics map as a whole, by adding another 250 protein spot in the analysis. In other words we could have stopped our calculations at $N = 250$ and obtained all information of interest without further increasing labor by considering a larger set of compounds. The plots corresponding to cases 30/50; 50/100; and 100/250 show a fair correlations associated, however, with some scatter of the points, which gradually decrease as we approach the 250/500 plot. On the other hand, when we go beyond the 250/500 spots we can observe again some scatter of points around the correlation line, which we believe are due to inclusion of an increasing number of spots of low abundance, the measurements of which are less accurate and thus their abundances are accompanied with higher noise. Nevertheless, the strong correlations of the entries of adjacent rows of Table 4 continue even for these cases, supporting the main conclusion that no significant additional information is obtained after we increase the number of spots in a map beyond 250.

CONCLUDING REMARKS

Most of the previous reports on the development of novel characterization of proteomics maps by a set of mathematical invariants were considering relatively small number of protein spots, even though these approaches could be applied to a larger set of data. The question of the role of the size of the proteins considered in the analysis was thus quite open. This question relates both to the characterization of proteomics maps as well as the characterization of cell proteome, a topic that only very recently received some attention,[41] in a study of the dose response for proliferators LY171883 (data from the work of Anderson et al[42]). In this case the presence of hormesis at the proteome level was for the first time demonstrated.[41] The issue on the role of the number of proteins considered for the characterization of proteome, which is analogous to the issue considered in this article with respect to proteomic maps, was considered recently.[43] In such a study one uses the information on the abundance of protein spots as detected in 2-D gels, but one is not using the information on the locations of protein spots in the 2-D gel. It was found in that study that some 200−300 most abundant proteins suffice for the characterization of proteome as a whole.[43] In this paper we have shown that for satisfactory numerical characterizations of proteomics maps, which typically have a few thousand proteins spots, it suffices to limit attention again to some 250 spots of the higher abundance, thus both studies arriving at the essentially the same conclusion.

ACKNOWLEDGMENT

This manuscript is contribution number 395 from the Center for Water and the Environment of the Natural Resources Research Institute. This material is based on research sponsored by the Air Force Research Laboratory, under agreement numbers F49620-02-1-0138 and FA9550-05-1-0456.

REFERENCES AND NOTES

(1) Randić, M.; Vračko, M.; Nandy, A.; Basak, S. C. On 3-D Graphical Representation of DNA Sequences and Their Numerical Characterization. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 1235−1244.
(2) Randić, M. On Characterization of DNA Primary Sequences by a Condensed Matrix. *Chem. Phys. Lett*. **2000**, *317*, 29−34.
(3) Randić, M.; Vračko, M. On The Similarity of DNA Primary Sequences. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 599−606.
(4) Randić, M. Condensed Representation of DNA Primary Sequences. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 50−56.
(5) Guo, X.; Randić, M.; Basak, S. C. A Novel 2-D Graphical Representation of DNA Sequences of Low Degeneracy. *Chem. Phys. Lett*. **2000**, *350*, 106−112.
(6) Randić, M.; Vračko, M.; Lerš, N.; Plavšić, D. Novel 2-D Graphical Representation of DNA Sequences And Their Numerical Characterization. *Chem. Phys. Lett*. **2003**, *368*, 1−6.
(7) Randić, M.; Vračko, M.; Lerš, N.; Plavšić, D. Analysis of Similarity/Dissimilarity of DNA Sequences Based on Novel 2-D Graphical Representation. *Chem. Phys. Lett*. **2003**, *371*, 202−207.
(8) Randić, M.; Vračko, M.; Zupan, J.; Novič, M. Compact 2-D Graphical Representation of DNA. *Chem. Phys. Lett*. **2003**, *373*, 558−562.
(9) Balaban, A. T.; Plavšić, D.; Randić, M. DNA Invariants Based on Nonoverlapping Triplets of Nucleotide Basis. *Chem. Phys. Lett*. **2003**, *379*, 147−154.
(10) Randić, M. Graphical Representation of DNA As a 2-D Map. *Chem. Phys. Lett*. **2004**, *386*, 468−4717.
(11) Randić, M.; Zupan, J.; Balaban, A. T. Unique Graphical Representation of Protein Sequences Based on Nucleotide Triplet Codons. *Chem. Phys. Lett.* **2004**, *397*, 247−252.
(12) Randić, M.; Zupan, J. Highly Compact 2-D Graphical Representation of DNA Sequences. *SAR QSAR Environ. Res.* **2004**, *15*, 191−205.
(13) Zupan, J.; Randić, M. Algorithm for Coding DNA Sequences into ≫Spectrum-Like≪ and ≫Zigzag≪ Representations. *J. Chem. Inf. Model*. **2005**, *45*, 309−000.
(14) Randić, M. Novel 1-Dimensional Representation of DNA. *Period. Biol*. **2005**, in press.
(15) Randić, M.; Vikić-Topić, D.; Graovac, A.; Lerš, N.; Plavšić, D. Novel Graphical And Numerical Characterization of DNA. *Period. Biol*. **2005**, in press.
(16) Randić, M. 2-D Graphical Representation of Proteins Based on Virtual Genetic Code. *SAR QSAR Environ. Res.* **2004**, *15*, 147−157.
(17) Randić, M.; Butina, D.; Zupan, J. Novel 2-D Graphical Representation of Proteins. *J. Chem. Phys. Lett.* **2005**, in press.
(18) Randić, M.; Balaban, A. T.; Novič, M.; Založnik, A.; Pisanski, T. A Novel Graphical Representation of Proteins. *Period. Biol*. **2005**, in press.
(19) Randić, M. Star Graphs. *J. Mol. Graphics Modell*. Submitted for publication.
(20) Novič, M.; Randić, M. Representation of Proteins as a Walk in 20-D Space. *Chem. Phys. Lett*. Manuscript to be submitted.

(21) Randić, M. On Graphical and Numerical Characterization of Proteomics Maps. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 1330−1338.
(22) 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.
(23) 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.
(24) Randić, M. A Graph Theoretical Characterization of Proteomics Maps. *Int. J. Quantum Chem.* **2002**, *90*, 848−858.
(25) Randić, M.; Basak, S. C. A Comparative Study of Proteomics Maps Using Graph Theoretical Biodescriptors. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 983−992.
(26) Randić, M.; Vračko, M.; Novič, M. On Characterization of Dose Variations of 2-D Proteomics Maps by Matrix Invariants. *J. Proteome Res.* **2002**, *1*, 217−226.
(27) 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.
(28) 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.
(29) 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.
(30) 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.
(31) 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.
(32) 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: in press.
(33) Randić, M.; Lerš, N.; Vukičević, D.; Plavšić, D.; Gute, B. D.; Basak, S. C. Canonical Labeling for Protein Spots and Proteomics Maps. *J. Proteome Res*. **2005**, *4*, 1347−1352.
(34) Randić, M.; Nović, M.; Vračko, M. Novel Characterization of Proteomics Maps by Sequential Neighborhoods of Protein Spots. *J. Chem. Inf. Model*. **2005**, *45*, 1205−1213.
(35) Vračko, M.; Basak, S. C. Similarity Study of Proteomic maps. *Chemom. Intell. Lab. Syst*. **2004**, *70*, 33−38.
(36) Randić, M. Graphical Representation of DNA As 2-D Map. *Chem. Phys. Lett.* **2004**, *36*, 468−471.
(37) Randić, M.; Lerš, N.; Plavšić, D.; Basak, S. C.; Balaban, A. T. Four-Color Map Representation of DNA or RNA Sequences and Their Numerical Characterization. *Chem. Phys. Lett.* **2005**, *407*, 205−208.
(38) Randić, M.; Balaban, A. T.; Pisanski, T.; Vikić-Topić, D.; Vukičević, D.; Plavšić, D. Graphical Representation of Proteins as Four-Color Maps And Their Numerical Characterization. *J. Proteome Res.* Manuscript submitted for publication.
(39) Jeffrey, H. I. Chaos Game Representation of Gene Structure. *Nucleic Acid Res*. **1990**, *18*, 2163-2170.
(40) Kowalski, B. R.; Bender, C. F. A Powerful Approach to Interpreting Chemical Data. *J. Am. Chem. Soc*. **1972**, *94*, 5632−5639.
(41) Randić, M.; Estrada, E. Order From Chaos: Observing Hormesis at the Proteome Level. *J. Protome Res*. **2005**, *4*, 2133-2136.
(42) Anderson, N. L.; Esquer-Blasco, R.; Richardson, F.; Foxworthy, P.; Eacho, P. The Effect of Peroxisome Proliferators on Protein Abundances in Mouse Liver. *Toxicol. Appl. Pharmacol*. **1996**, *137*, 75−89.
(43) Randić, M. Quantitative Characterizations of Proteome: Dependence on the Number of Proteins Considered. *J. Protome Res*. Manuscript submitted for publication.

CI050132H