Information-Theoretic Biodescriptors for Proteomics Maps: Development and Applications in Predictive Toxicology

SUBHASH C. BASAK,[1] BRIAN D. GUTE[1] and FRANK WITZMANN[2]
[1]Center for Water and the Environment
Natural Resources Research Institute
5013 Miller Trunk Highway, Duluth, Minnesota, 55811
UNITED STATES of AMERICA
[2]Department of Cellular & Integrative Physiology
Indiana University School of Medicine
1345 West 16th Street L3, Rm 308, Indianapolis, Indiana 46202
UNITED STATES of AMERICA
http://www.nrri.umn.edu/

Abstract: This paper describes an approach using information theory to derive a new complexity measure for proteomics maps generated using 2-dimensional gel electrophoresis. The maps used in this study were partitioned into 5x5 grids and the total abundance of protein material in each grid was compared to the total abundance for the entire map. Next, Shannon's relation was applied to characterize the distribution of spots across the proteomics map. Details of the approach are discussed here, including an illustrative example and an example of the calculations for a proteomics map containing 200 spots. Finally, results for the Map Information Content index are presented for a set of five maps calculated using 200 spots, 500 spots, and 1,054 spots. It is hoped that the application of information-theoretic techniques to characterize the complexity of these maps, thus reducing the amount of information presented to the researcher, will help in the analysis and comparison of maps containing a great deal of information.

Key-Words: 2-DE gel; biodescriptor; complexity; information theory; map information content; proteomics; proteomics maps

1 Introduction

In the aftermath of the Human Genome Project the emerging technologies of genomics, proteomics, and metabolomics are taking an increasingly important role in the prediction of biomedicinal activity and chemical toxicity. Whereas microarray studies provide an assessment of cellular transcriptional processes, proteomics provides a better understanding of the cell function because proteins are the workhorses of the living systems. The field of proteomics includes technologies such as two-dimensional gel electrophoresis (2-DE), matrix-assisted laser desorption-ionization (MALDI), surface-enhanced laser desorption ionization (SELDI), and isotope-coded affinity tagging (ICAT).

Many authors have used 2-DE gel technology in understanding the molecular basis of chemical toxicity [1, 2]. In this method the tissue or cell exposed to the toxicant is homogenized and the proteins are separated by charge and mass through two-dimensional electrophoresis. The abundance of each spot after separation gives the magnitude of a particular type of protein or a closely related set of proteins, which comprise an individual spot. A typical 2-DE get can have 1,500–2,000 identified protein spots.

An important goal of toxicoproteomics is to study the perturbation of protein expression in tissues under the influence of toxicants. However, characterizing patterns consisting of 1,500 or more objects is a daunting task and cannot be accomplished

simply through visual inspection. It requires rigorous mathematical/statistical methods for a thorough and objective analysis of such patterns. Our group has been involved in the characterization of toxicoproteomic patterns using four different techniques: a) invariants of graphs associated with proteomics maps [3], b) spectrum-like representations of proteomics maps based on projections of the 3-D space (mass, charge, and abundance) onto three (*xy*, *yz*, and *xz*) planes [4], c) selection of toxicologically-relevant spots based on robust statistical methods, and d) information-theoretic characterization of the pattern of protein spots of the 2-DE gel.

2 Information-Theoretic Formalism

A proteomic map can be looked upon as the two-dimensional distribution pattern of the total cellular mass of identifiable proteins based on charge (*x*) and mass (*y*). When a cell is exposed to a toxicant, its transcriptional and translational processes are perturbed, potentially resulting in a redistribution of protein spots and the appearance or disappearance of some proteins. Information theory is a suitable mathematical tool for characterizing such complex patterns. Previously, we have applied information theory to characterize the neighborhood complexity of atomic bonding patterns within molecules [5–9]. Here we report the application of information theory in characterizing the proteomic patterns of cells exposed to four peroxisome proliferators, *viz.*, perfluoro-octanoic acid (PFOA), perfluorodecanoic acid (PFDA), clofibrate, and diethylhexyl-phthalate (DEHP).

In the information theoretic formalism, a set *A* of *N* objects is partitioned into subsets A_i with cardinalities N_i; $\Sigma\ N_i = N$. A probability scheme is then associated to the distribution:

$$A_1, A_2, \ldots, A_h$$
$$p_1, p_2, \ldots, p_h$$

Where $p_i = N_i / N$

The complexity of the system consisting of *N* objects is computed by Shannon's formula:

$$Complexity = -\sum_{i=1}^{h} p_i \log_2 p_i$$

3 Information Content of Proteomics Maps

Two-dimensional gel electrophoresis data on the effects of the four peroxisome proliferators; PFOA, PFDA, clofibrate, and DEHP; were developed in the laboratory of Frank Witzmann using liver tissue from treated male Fisher-344 rats [10]. A total of 1,054 distinct protein spots were quantified. Each spot has associated with it magnitudes of charge, mass, and abundance.

In applying information-theory to the study of proteomics maps, it was of interest to calculate the complexity (information content) of the distribution of proteins (abundance) over the mass-charge plane. To this end, the maps were divided on the *xy* plane into *n* x *n* cells, where *n* = 1, 2, …, 5. It may be noted that as the number of cells (n^2) increases, the protein spots are distributed across an increasingly larger and larger number of cells. With high enough values of *n*, it is conceivable that each spot might occur in its own cell, which would be much too reductionistic to yield interesting results. For this communication, the charge-mass proteomics maps for the most abundant 200, most abundant 500, and the complete maps of 1,054 spots have been divided into 5x5 grids for the calculation of information content indices.

4 Calculation Procedure

Since proteomics maps for Fisher-344 rat liver cells range in charge roughly from 0 to 3,100, mass from 0 to 2,500, and abundance from 0 to 163,000, let us first consider a simpler example of map characterization.

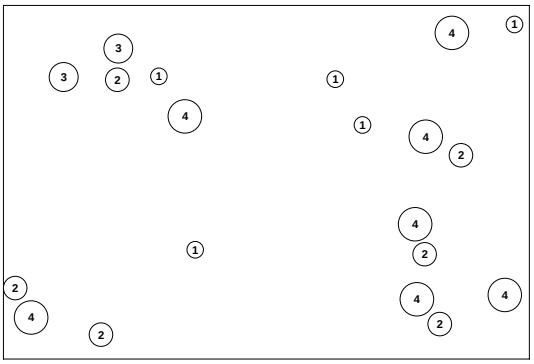

Figure 1. Simulated 20-spot proteomics map.

Fig. 1 presents a simplified sample map of 20 spots. As can be seen from Table 1, the values for *x*, *y*, and *z* all range from zero to four.

Table 1. Simulated *x*, *y*, and *z* coordinates for a 20-spot map.

Spot ID	x	y	z
1	2.73	1.35	1
2	1.18	0.80	1
3	0.87	0.49	3
4	3.81	3.28	4
5	3.14	3.32	4
6	3.41	0.31	4
7	3.13	2.48	4
8	3.21	1.49	4
9	0.21	3.53	4
10	3.32	3.60	2
11	0.74	3.72	2
12	3.48	1.70	2
13	3.89	0.22	1
14	2.52	0.83	1
15	1.46	2.76	1
16	0.46	0.81	3
17	0.09	3.20	2
18	3.20	2.82	2
19	0.87	0.84	2
20	1.38	1.26	4

Fig. 2 presents the same sample map divided into a 4x4 grid. Once the divisions have been made, it is a simple matter to sum the values of *z* (or N_i) for each sector, as shown in Fig. 3.

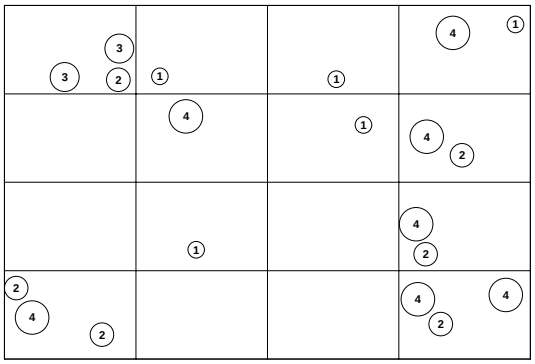

Figure 2. The simulated proteomics map evenly divided into a 4x4 grid.

8	1	1	5
0	4	1	6
0	1	0	6
8	0	0	10

Figure 3. Values of N_i for each of the 16 sectors on the 4x4 grid.

The value of p_i is then calculated for each sector. For this map, the total sum of the values of *z* for the entire map (*N*) is equal to 51. Therefore, p_i for the sector in the uppermost lefthand corner of the map is equal to N_i / *N*, where N_i equals 8 and *N* equals 51. Thus the value of p_i for this uppermost lefthand cell is approximately 0.1569

0.1569	0.0196	0.0196	0.0980
0	0.0784	0.0196	0.1176
0	0.0196	0	0.1176
0.1569	0	0	0.1961

Figure 4. Values of p_i for each of the 16 sectors on the 4x4 grid.

Once all of the values of p_i have been calculated, the resulting map complexity

index (MIC) can be calculated. However, seeing that several of the cells have zero values, an approximation will have to be made since it is meaningless to take the logarithm of zero. This is a very realistic problem when dealing with proteomics maps, since the proteins are not evenly distributed across the gel. As the values of p_i approach zero, the value of $p_i \log_2 p_i$ also approaches zero. Therefore, if the value of ($p_i \log_2 p_i$) for zero is approximated as zero, the value of MIC can be calculated for map grids containing empty cells. Fig. 5 presents the values of $p_i \log_2 p_i$ for each of the cells which are then summed to determine the value of MIC. In the case of our 20-spot map, MIC = 3.0872.

Before extrapolating this simple example to a proteomics map, there are several additional issues that must be addressed.

-0.4192	-0.1112	-0.1112	-0.3285
0	-0.2880	-0.1112	-0.3632
0	-0.1112	0	-0.3632
-0.4192	0	0	-0.4609

Figure 5. Values of $p_i log_2 p_i$ for each of the 16 sectors on the 4x4 grid.

First, the exact boundaries of a proteomics map vary depending on the technique being used and the cells being studied. For the Fisher-344 rat liver cells, theoretically the maps range from a charge of 0 to 3,100, but the lowest value for charge that appears in the data set is 104, while the highest value is 3,050. However, if we normalize the range based on the minima and maxima of the dataset, we will not retain the overall dimensionality of the map, which could create problems later on if other maps have proteins that fall outside of those normalized ranges. So, it is important to consider the true dimensions of the gel, rather than simply the apparent range of the data.

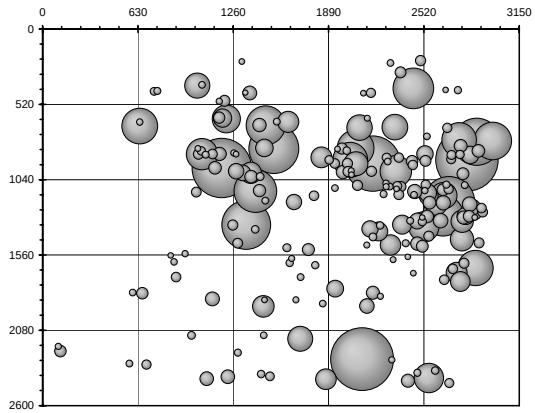

Figure 6. Most abundant 200 spots for the control proteomics map divided into a 5x5 grid.

Secondly, many of the spots on a 2-DE gel do fit neatly into the arbitrarily defined sectors. As such, the centroid of each spot is used to determine the sector to which the protein belongs and all of its abundance is attributed to that sector.

In this analysis, a control sample and four treatments were considered, generating a total of 15 MIC values, five for the most abundant 200 spots, five for the 500 most abundant spots, and five for the complete maps.

Upon examining the five proteomics maps, it was found that empty bands appeared along two edges of the maps, while the other two edges were somewhat crowded. No proteins appeared at relatively low values for charge (x) or mass (y). The lowest value for charge was 104 and the lowest value for mass was 110. As such, it was decided that the upper bounds of the maps should be extended as well, based on the assumption that higher values for charge and mass than those observed could also occur. So, when dividing the maps into sectors it was decided that the x-axis would range between 0 and 3,150 (3,050 being the highest observed value of charge), and the y-axis would range from 0 to 2600 (2,486 being the highest observed value of charge). Fig. 6 presents a graphical representation of the proteomics map showing only the 200 most abundant spots for the control with a 5x5 grid superimposed onto the map. Figs. 7 & 8 show the control values for p_i and

$p_i \log_2 p_i$, respectively, on a 5x5 grid for the most abundant 200 spots.

0	0.0220	0.0086	0.0287	0.0044
0	0.0882	0.0869	0.1315	0.1049
0	0.0091	0.0592	0.1012	0.1580
0.0023	0.0163	0.0235	0.0233	0.0361
0.0085	0.0152	0.0260	0.0305	0.0157

Figure 7. Values of p_i for each of the 25 sectors on the 5x5 grid for the most abundant 200 spots of the control map.

0	-0.1211	-0.0591	-0.1470	-0.0345
0	-0.3090	-0.3062	-0.3849	-0.3413
0	-0.0615	-0.2415	-0.3343	-0.4206
-0.0202	-0.0966	-0.1271	-0.1263	-0.1728
-0.0587	-0.0919	-0.1369	-0.1534	-0.0940

Figure 8. Values of $p_i log_2 p_i$ for each of the 25 sectors on the 5x5 grid for the most abundant 200 spots of the control map.

5 Results and Discussion

The magnitudes of the Map Information Content (MIC) indices for the most abundant 200 spots, the most abundant 500 spots, and for the entire set of 1,054 proteins spots reported by Witzmann for peroxisome proliferators [10] are presented in Table 2 and Fig. 9. It may be noted that the magnitude of MIC decreases for the proteomics maps of cells exposed to toxicants.

A toxicologically interesting fact is that the two structurally related peroxisome proliferators PFOA and PFDA, along with clofibrate, have relatively similar MIC values as compared to the control and DEHP. This indicates that the MIC biodescriptor reported here may be capable of characterizing toxicity and toxic modes of action of toxicants.

Table 2. Calculated values of MIC for the five treatments using the most abundant 200, most abundant 500, and entire set of 1,054 proteins.

	200	**500**	**1054**
Control	3.8390	3.9486	3.9892
PFOA	3.7865	3.9112	3.9519
PFDA	3.7702	3.8861	3.9289
Clofibrate	3.7769	3.9051	3.9584
DEHP	3.7033	3.8387	3.8923

One desirable property of chemo-descriptors is the ability to discriminate among closely related chemical structures. Analogously, we would expect that biodescriptors derived from biological systems would be able to discriminate among closely related biochemical processes. The MIC index not only discriminates among maps derived for different structural classes of peroxisome proliferators, it also discriminates between closely related compounds, *e.g.*, PFOA and PFDA. It is expected that the MIC index will find applications in pattern recognition for proteomics maps pertinent to biomedicinal chemistry, pharmacology, pathology, and toxicology.

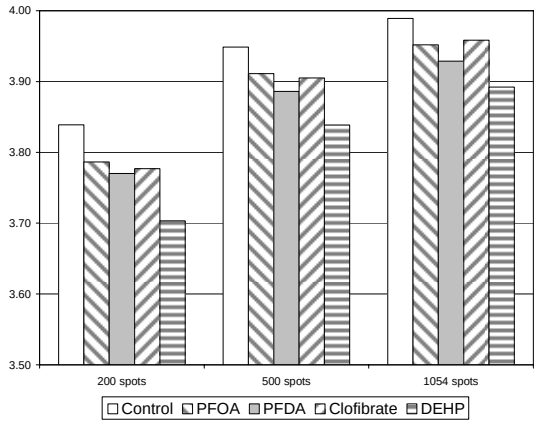

Figure 9. Bar chart comparing the values of MIC across the five treatments for the 200-spot, 500-spot, and 1,054-spot proteomics maps.

Acknowledgements

This manuscript is contribution number 379 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 number F49620-02-1-0138. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory or the U.S. Government.

References:

[1] Anderson, N.L., R. Esquer-Blasco, F. Richardson, P. Foxworthy and P. Eacho, The Effects of Peroxisome Proliferators on Protein Abundances in Mouse Liver. *Toxicology and Applied Pharmacology*, 137, 1996, pp. 75–89.

[2] Witzmann, F.A, Proteomic Applications in Toxicology, in *Comprehensive Toxicology, Vol. XIV: Cellular and Molecular Toxicology*, J.P. Vanden Heuvel, W.F. Greenlee, G.H. Perdew, W.B. Mattes (eds.), Elsevier, New York, NY, 2002, pp. 539–558.

[3] Randic, M., F. Witzmann, M. Vracko and S.C. Basak, On Characterization of Proteomics Maps and Chemically Induced Changes in Proteomes using Matrix Invariants: Application to Peroxisome Proliferators, *Medicinal Chemistry Research*, 10, 2001, 456–479.

[4] Vracko, M. and S.C. Basak, Similarity Study of Proteomic Maps, *Chemometr. Intell. Lab. Syst.*, 70, 2004, 33–38.

[5] Basak, S.C., A.B. Roy and J.J. Ghosh, Study of the Structure-Function Relationship of Pharmacological and Toxicological Agents using Information Theory, in *Proceedings of the Second International Conference on Mathematical Modelling*, X.J.R. Avula, R. Bellman, Y.L. Luke and A.K. Rigler (eds.), University of Missouri-Rolla, Rolla, Missouri, 1979, pp. 851–856.

[6] Magnuson, V.R., D.K. Harriss and S.C. Basak, Topological Indices Based on Neighborhood Symmetry: Chemical and Biological Applications, in *Chemical Applications of Topology and Graph Theory*, R.B. King (ed.), Elsevier, 1983, pp. 178–191.

[7] Roy, A.B., S.C. Basak, D.K. Harriss and V.R. Magnuson, Neighborhood Complexities and Symmetry of Chemical Graphs and Their Biological Applications, in *Mathl. Modelling Sci. Tech.*, X.J.R. Avula, R.E. Kalman, A.I. Liapis, and E.Y. Rodin (eds.), Pergamon Press, 1983, pp. 745–750.

[8] Basak, S.C., Use of Molecular Complexity Indices in Predictive Pharmacology and Toxicology: A QSAR Approach, *Medical Science Research*, 15, 1987, 605–609.

[9] Basak, S.C., Information Theoretic Indices of Neighborhood Complexity and Their Applications, in *Topological Indices and Related Descriptors in QSAR and QSPR*, J. Devillers and A.T. Balaban (eds.), Gordon and Breach Science Publishers, The Netherlands, 1999, pp. 563–593.

[10] Witzmann F.A., B.M. Jarnot, D.N. Parker, J.W. Clack, Modification of Hepatic Immunoglobulin Heavy Chain Binding Protein (BiP/Grp78) Following Exposure to Structurally Diverse Peroxisome Proliferators. *Fundam. Appl. Toxicol.*, 23, 1994, 1–8.