J Math Chem (2011) 49:1759–1768
DOI 10.1007/s10910-011-9856-x

ORIGINAL PAPER

On numerical characterization of proteomics maps based on partitioning of 2-D maps into Voronoi regions

Milan Randić · Rok Orel

Received: 5 September 2010 / Accepted: 23 May 2011 / Published online: 8 June 2011
© Springer Science+Business Media, LLC 2011

Abstract We consider the problem of partitioning of the 2-D proteomics maps into regions associated with set of selected protein spots so that one can construct the adjacency matrix for proteomics maps, which will be a source of 2-D map descriptors. We selected a set of N most abundant spots for which we construct the Voronoi regions. Once Voronoi regions have been obtained one can construct the accompanying adjacency matrix A, and adjacency-distance matrix AD. The binary elements A(i, j) of the adjacency matrix are determined by the adjacency of corresponding Voronoi regions, and the elements AD(i, j) of the adjacency-distance matrix are determined by the distances of corresponding centers of Voronoi regions. The approach is illustrated on a smaller map for 20 most abundant proteins from a selection of available data in the literature.

Keywords Proteomics maps · Voronoi regions · Peroxisome proliferators

1 Introduction

Proteins have constant charge and constant mass. In 2-D proteomics maps they are separated by charge electrophoresis and by mass using chromatography. Consequently proteins will appear with the same (x, y) coordinates in all proteomics maps while their

M. Randić · R. Orel
National Institute of Chemistry, Hajdrihova 19, Ljubljana, Slovenia

R. Orel
e-mail: rok.orel@gmail.com

M. Randić (✉)
Department of Mathematics and Computer Science, Drake University, Des Moines, IA, USA
e-mail: mrandic@msn.com

abundance varies from case to case. In order to make quantitative comparisons of different maps one has first numerically to characterize individual maps M(x_i, y_i, z_i), where x_i and y_i represent mass and charge, and z_i the abundance of ith protein (i = 1, …, N), N being the number of proteins considered. A way to do this is to associate with the map M(x_i, y_i, z_i) a suitable geometrical object, numerical characterization of which will lead to mathematical representation of the proteomic map. Numerical characterization of proteomics maps, initiated in 2003 [1], considered the following geometrical objects for numerical analysis of proteomics maps:

(1) Zigzag curve connecting *N* most abundant points [1–4];
(2) Graph of partial ordering of spots with respect to mass and charge [5,6];
(3) Cluster graph connecting spots with distances smaller of an assumed threshold value [7];
(4) Graph connecting the nearest pairs of spots among the set of N spots selected [8];
(5) Graph connecting the nearest pairs of spots sequentially, restricting attention only to spots having larger abundance [9].

Each of the above schemes leads to sets of invariants, which could be used as map descriptors [10]. It is desirable to have alternative approaches for characterization of maps, which offer different sets of descriptors. In this way in comparative studies that follow numerical characterization of proteomics maps one can at eliminate false positives, that is, one can identify as dissimilar maps that would accidentally be characterized by similar numerical descriptors. As is well known any characterization of mathematical objects by set of invariants is accompanied with loss of information, however, when descriptors based on different representations are combined part of lost information may be recovered. This reason alone suffices to encourage search for additional alternative graphical and numerical representations of proteomics maps. In addition, some of mentioned approaches are computationally more involved or use *dense* matrices (matrices with few zeros [11]). Such are the distance/distance [12] matrices for zigzag curves in which one computes all N(N-1)/2 distances between *N* points. The same is also with use of the shortest paths in cluster graphs [13]. In contrast graphs resulting from partial ordering and graphs depicting the nearest neighbors lead to *sparse* matrices, which are computationally more suitable for use with large maps. Clearly, representation of 2-D proteomics maps by *sparse* matrices is desirable.

In this contribution we will outline an approach to numerical characterization of 2-D proteomics maps by sparse matrices obtained by partitioning of 2-D maps into Voronoi regions [14,15].

2 Voronoi regions

Voronoi regions result from partitioning of a plane into cells (or regions) the boundaries of which separate regions so that all points in each region are closer to spot N_i than any other spot. The number of the regions in which (x, y) plane is partitioned is given by *N* (the number of selected spots or points). Partitioning of proteomics maps in Voronoi regions leaves close proteins having similar both mass and charge. In contrast graph of partial ordering relates proteins having similar mass and charge only

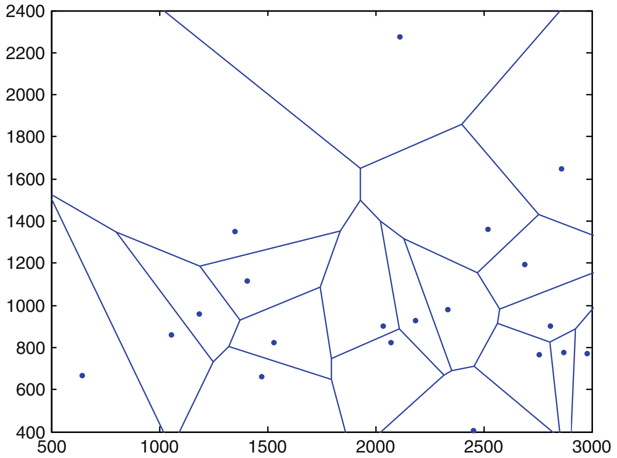

Fig. 1 Voronoi regions for the proteomic maps considered in this contribution

Table 1 The (x, y) coordinates of twenty most abundant protein spots in the control group and the corresponding abundances when four different proliferators were added to feed of rats (from work of Anderson et al. [18])

	x	y	Control	PFOA	PFDA	Clofibrate	DEHP
1	2111.7	2278.6	144357	108713	95028	147081	165886
2	2804.3	903.6	143630	155565	188582	159898	155055
3	1183.9	959.6	136653	113859	150253	163645	8111
4	2182.2	928.8	127195	99160	73071	76642	112096
5	2685.6	1196.1	118581	112790	49769	109856	138795
6	1527.9	825.5	114929	192437	221567	166080	180590
7	1346.0	1352.5	112251	58669	38915	73159	77075
8	2868.5	778.0	108883	26105	50735	45923	116849
9	1406.3	1118.1	98224	91147	82963	84196	92942
10	2450.2	409.2	93601	83172	62934	79870	109381
11	1474.0	665.1	90004	129340	112361	112655	119402
12	2974.9	772.8	86730	70746	78691	105760	116281
13	2068.4	823.1	84842	73814	45482	71911	97444
14	642.2	669.8	82492	73974	74466	84703	88545
15	2860.7	1649.9	81965	16137	16501	60077	148992
16	2032.7	902.8	80015	77314	80072	76027	100836
17	2752.7	765.6	79847	20782	13103	38816	53830
18	2334.2	982.2	72791	76369	52749	55599	77432
19	1053.6	864.3	72173	77982	60376	46808	78121
20	2519.5	1365.9	69452	37838	16129	57167	71274

if one protein dominates the other (or is dominated by the other) in both properties. In Fig. 1 we illustrate one simplified proteome map showing twenty Voronoi regions, determined by the twenty points with the coordinates (x, y) listed in Table 1. Observe, however that the adjacency relationships for some regions at the periphery need not be obvious, but the adjacencies can be identified by construction of convex hull. In the case of Fig. 1 convex hull is pentagon obtained by connecting points 1, 15, 2, 10

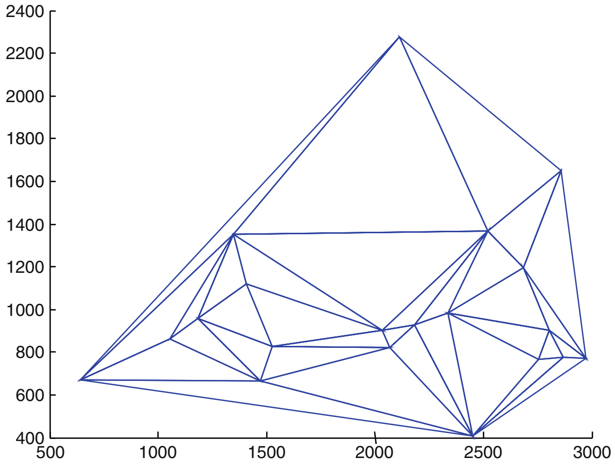

Fig. 2 The accompanying Delaunay triangulation of the Voronoi regions of Fig. 1

and 14. Adjacent points of the pentagon signify Voronoi regions that have common boundary if the diagram would be extended indefinitely.

An alternative way to establish adjacencies for regions at the periphery of a Voronoi diagram is to construct the corresponding Delaunay triangulation [16,17]. This triangulation has a property that for given set of points it leads to a collection of edges with a property that for each edge one can find a circle containing the edge's endpoints but not containing any other points of the given set. The Delaunay triangulation is the *dual* of the Voronoi diagram, which is constructed by connecting the points that define Voronoi regions if they have a common boundary. For the Voronoi diagram of Fig. 1 we illustrate in Fig. 2 the accompanying Delaunay triangulation and in Table 2 show the corresponding adjacency matrix.

3 Characterization of proteomics map using the first eigenvector

Before continuing we should draw attention of readers to the fact that the three coordinates (x, y, z), where z gives the relative abundance of proteins, each measure different quality: the mass, the charge, the abundance, each on its own scale. Because we will combine the x, y, z triplets into a matrix, which is dimensionless, question of different scales has to be considered. Kowalski and Bender [19] were addressing some aspects of this problem and they recommended re-scaling initial x, y, z quantities to a [0, 1] interval. Because the intervals of the x and y coordinates for the 20 proteins considered are of the same magnitude in this particular study there is no need to rescale them. In our approach we want to insert the information on the relative abundance of proteins as diagonal entries into adjacency matrices. Therefore we have to consider adequate scaling of the relative abundance of proteins. This poses a problem that the recommendation of Bender and Kowalski cannot resolve.

Table 2 The adjacency matrix for Voronoi regions of Fig. 1

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	1
2	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	1	1	0	0
3	0	0	0	0	0	1	1	0	1	0	1	0	0	0	0	0	0	0	1	0
4	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	1	0	1
5	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	1
6	0	0	1	0	0	0	0	0	1	0	1	0	1	0	0	1	0	0	0	0
7	1	0	1	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0	1	1
8	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0
9	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0
10	0	0	0	1	0	0	0	1	0	0	1	1	1	1	0	0	1	1	0	0
11	0	0	1	0	0	1	0	0	0	1	0	0	1	1	0	0	0	0	1	0
12	0	1	0	0	1	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0
13	0	0	0	1	0	1	0	0	0	1	1	0	0	0	0	1	0	0	0	0
14	1	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	1	0
15	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
16	0	0	0	1	0	1	1	0	1	0	0	0	1	0	0	0	0	0	0	1
17	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0
18	0	1	0	1	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1
19	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0
20	1	0	0	1	1	0	1	0	0	0	0	0	0	0	1	1	0	1	0	0

That the simple rescaling of Bender and Kowalski will not be suitable is easy to see because such scaling does not take into account the *size* of adjacency matrix into consideration. For example, scaling of diagonal entries of the 20 × 20 adjacency matrix of Table 1 to balance the influence of non-diagonal entries will not hold if one wishes to extend the approach by considering 40 proteins, because the in the 40 × 40 adjacency matrix the off diagonal matrix elements would outweigh the role of the diagonal element of a 20 × 20 matrix. A way out of this difficulty is to normalize diagonal entries of a matrix so that the matrix trace (the sum of diagonal entries) equals the sum of off-diagonal entries in the matrix (the Wiener number of the matrix, or 2W) [19]. This procedure allows extension of analysis by inclusion of larger number of protein spots in proteomic maps. It is of some interest to mention that Laplace matrix [20], which is of interest also in Chemical Graph Theory [21,22], has inherently built in this "variable" aspect of scaling the trace of a matrix, which is determined by the number of edges in a graph.

We have adopted this notion of the "variable" scaling for abundances, which for the entries of Table 1 give the relative abundances listed in Table 3. The scaling factor used is: 102/1998615 = 1/1959.26, where 102 is the sum of off-diagonal entries in the adjacency matrix of Table 2 and 1998615 is the sum of abundances of all 20 most abundant protein for the control group (listed in the "control" column in Table 1). Observe that we use the *same* normalization factor for all five proteomic maps of Table 1. This immediately indicates the overall increase or decrease of proteins in cells exposed to different conditions. For the first 20 proteins of the control group this is visible from the last row of Table 3, which shows that for PFOA, PFDA and Clofibrate the total abundance of proteins has decreased, while for DEHP it has somewhat increased.

Table 3 Scaled abundances (diagonal elements) of augmented adjacency matrix

Control	PFOA	PFDA	Clofribrate	DEHP
7.3673	5.5482	4.8498	7.5063	8.4661
7.3302	7.9393	9.6243	8.1605	7.9133
6.9741	5.8108	7.6682	8.3517	0.4140
6.4914	5.0607	3.7292	3.9115	5.7209
6.0518	5.7563	2.5400	5.6065	7.0835
5.8654	9.8211	11.3078	8.4760	9.2165
5.7288	2.9942	1.9860	3.7337	3.9336
5.5569	1.3323	2.5893	2.3437	5.9634
5.0129	4.6517	4.2340	4.2970	4.7433
4.7770	4.2447	3.2119	4.0762	5.5823
4.5934	6.6009	5.7344	5.7494	6.0937
4.4263	3.6105	4.0160	5.3975	5.9344
4.3299	3.7671	2.3212	3.6700	4.9731
4.2100	3.7753	3.8004	4.3228	4.5189
4.1831	0.8236	0.8421	3.0661	7.6039
4.0836	3.9457	4.0865	3.8801	5.1462
4.0750	1.0606	0.6687	1.9810	2.7472
3.7149	3.8975	2.6921	2.8375	3.9518
3.6834	3.9798	3.0813	2.3889	3.9869
3.5445	1.9311	0.8231	2.9175	3.6375
102	86.55	79.81	92.67	107.62

4 Novel map invariants

The second novelty of this article is use of novel map descriptors. In the past as map descriptors often were used the leading eigenvalues of the distance/distance D/D matrices, and the "higher order" matrices obtained from D/D matrices by using Kronecker product of matrix by itself. This amounts to raising the individual matrix elements a_{ij} to higher powers. Rather than using the leading eigenvalues, which for the five cases of Table 3 are: 10.5037; 11.3171; 12.4136, 10.9696 and 10.9276, for the control, PFOA, PFDA, Clofibrate and DEHP groups, respectively, we decided to consider the corresponding *first eigenvectors*. As is known, it is not uncommon that the leading eigenvalues of matrices have some structural interpretation [12,23,24], which make them suitable as descriptors in structure-property studies. On the other hand the accompanying leading eigenvectors have no nodes (that is all coefficients are of the same sign), which makes them of interest for interpretation. In the case of MO calculations in Quantum Chemistry the eigenvectors are of interest for construction of atom charges and bond orders, however, in contrast here we will continue to confine attention only to the first eigenvector of matrices considered.

5 Illustration

To illustrate the novel approach we have selected data from the work of Anderson *et al* for two reasons: (1) The data from this laboratory appear to be of very high quality, which reduces chance for experimental variations know to plague preparations

Table 4 The first eigenvector of augmented adjacency matrices for the five proteomics maps considered

ψ_1 (A*)	ψ_1 (B*)	ψ_1 (C*)	ψ_1 (D*)	ψ_1 (E*)
0.2676	0.0506	$0.021\underline{5}$	0.1379	0.3646
0.2717	0.0489	0.0196	0.0971	0.3790
$0.304\underline{5}$	0.2790	0.2973	0.5412	0.0444
0.2670	0.1004	0.0453	0.0937	0.1684
0.2209	0.0376	0.0084	0.0694	0.3721
0.2365	0.7827	0.8768	0.5679	0.0469
0.3162	0.1196	0.0766	0.2013	0.1502
0.1664	$0.022\underline{5}$	0.0095	0.0389	0.2136
$0.191\underline{5}$	0.2051	0.1721	0.2227	0.0602
0.2639	0.1211	0.0541	0.1354	0.2550
0.2210	0.3337	0.2214	0.3217	0.1417
0.1710	0.2018	0.0116	0.0688	0.3174
0.1916	0.0318	0.1343	0.1777	0.1242
0.1895	0.0970	0.0513	0.1394	0.1519
0.1379	0.0171	0.0061	0.0471	0.3813
0.2188	0.1979	0.1598	0.1904	0.1342
0.1378	0.0237	0.0085	0.0371	0.1199
0.2011	0.0252	0.0169	0.0649	0.1391
0.1484	0.1121	0.0692	0.1392	0.0698
0.2299	0.0610	0.0288	0.0991	0.2329

of proteomic maps; (2) The same data have been previously analyzed by different approaches, which makes possible to compare results from different studies. In Table 3 we have already listed the diagonal entries for the corresponding adjacency matrices and in Table 4 we have listed the first eigenvector for the five proteomic maps. They have been labeled as ψ_1, in parallel with labels ψ_i used in quantum chemistry for molecular orbitals (electron wavefunctions). All the calculations, construction of Voronoi regions, Delaunay triangulation and computations of the leading eigevector were made using MATLAB [25]. We now view the entries in each column of Table 4 as 20-component vector to be used and proteomic map descriptors.

A close look at Table 4 is instructive. First observe that all entries in the first column (the control group) are of similar magnitude, between 0.3262 and 0.1378, which is also the case with the entries in the first column of Table 3, though in Table 4 the values do not decrease monotonically. The other columns of Table 4, corresponding to PFOA, PFDA, Clofibrate and DEHP, show considerable variations in the magnitudes of the coefficients that define the first eigenvector. Thus in the case of PFOA the extreme values are (0.0171 and 0.7827); the extreme values are for PFDA (0.0061 and 0.8768); and the extreme values for Clofibrate and DEHP are (0.0371 and 0.5679) and (0.0603 and 0.379), respectively. Very interesting, however, is the parallelism between the columns of Tables 3 and 4, which we have illustrated for PFOA columns in Fig. 3, which suggests that the first eigenvectors can serve as proteomic map descriptors and may allow one to estimate numerically the degree of similarity between different proteomic maps.

To illustrate this point in Fig. 4 we have plotted the relative (experimental) abundances for the five proteomic maps considered, which have been tabulated in Table 3.

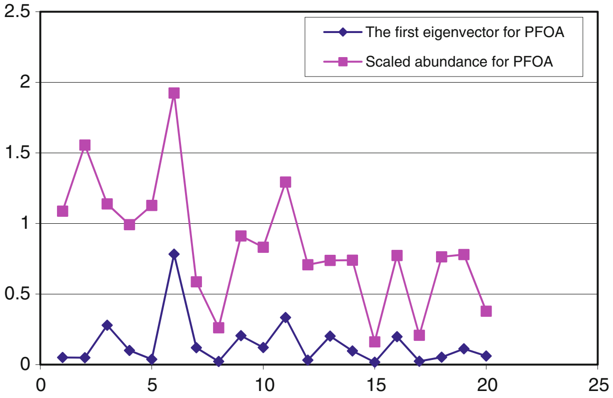

Fig. 3 Plot of the first eigenvector for PFOA (*bottom*) and the experimental scaled abundances for 20 most abundant proteins of the control group

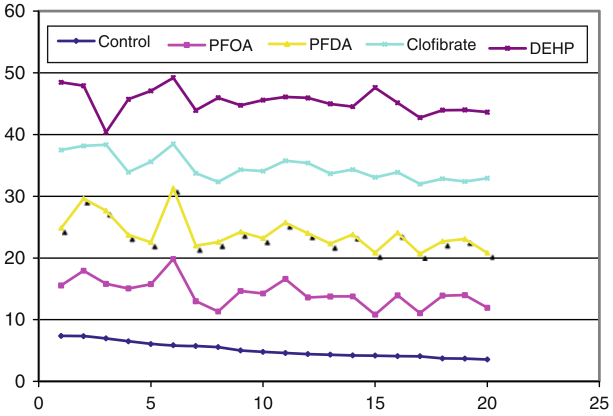

Fig. 4 Experimentally reported relative abundance for 20 proteins of the five proteomic maps considered for illustration of novel scheme for numerical characterization of proteomic maps

From Fig. 4 visually it appears that the effect of PFOA and PFDA on cell proteome are comparable, the effects of Clofibrate following and the effect of DEHP appear to be most dissimilar. In order to check this qualitative observation we computed the degree of similarity among the five proteomic maps using the 20-component vectors based on the first eigenvector. The results are collected in Table 5, which confirms the expectations that indeed the most similar proteomic maps are those belonging to PFOA and PFDA, with some similarity between these two proteomic maps and those of Clofibrate—all of course based on considerations of the twenty most abundant protein spots. It is not, of course, topic of this contribution to argue on limitations of characterizations of proteomics maps on 20 proteins, but to outline an approach, which may eventually address even that question, which received some attention in the literature [26,27].

Table 5 Similarity/Dissimilarity matrix for the five proteomic maps considered

	Control	PFOA	PFDA	Clofibrate	DEHP
Control	0	2.822	3.345	2.362	2.077
PFOA		0	0.832	1.025	3.952
PFDA			0	1.720	4.198
Clofibrate				0	3.657
DEHP					0

Concluding Remarks

We outlined novel approach to numerical characterization of proteomic maps based on construction of Voronoi partition for proteomic maps, followed with construction of accompanying adjacency matrix that is augmented by diagonal elements, which are scaled to parallel contributions from the off-diagonal elements of binary adjacency matrix. As map descriptors we used the first eigenvector of the augmented matrices, which appear to reflect on the relative abundances of the proteins spots. It may be premature to speculate on the potential of this approach but for general case the impressive results here reported warrant further investigation along the lines outlined here.

Acknowledgments We would like to thank the Laboratory for Chemometric, National Institute of Chemistry, Ljubljana, Slovenia, for hospitality and partial financial support. This work has been supported in part by the Ministry of Science and Higher Education, of Republic of Slovenia under Research Grant P1017.

References

1. M. Randić, J. Chem. Inf. Comput. Sci. **41**, 1330 (2001)
2. M. Randić, J. Zupan, M. Novič, J. Chem. Inf. Comput. Sci. **41**, 1339 (2001)
3. M. Randić, F. Witzmann, M. Vračko, S.C. Basak, Med. Chem. Res. **10**, 456 (2001)
4. M. Randić, N. Novič, M. Vračko, J. Proteomic Res. **1**, 217 (2002)
5. M. Randić, Int. J. Quantum Chem. **90**, 848 (2002)
6. M. Randić, S.C. Basak, J. Chem. Inf. Comput. Sci. **42**, 983 (2002)
7. Ž. Bajzer, M. Randić, D. Plavšić, S.C. Basak, J. Mol. Graph. Modell. **22**, 1 (2003)
8. M. Randić, N. Lerš, D. Plavšić, S.C. Basak, Croat. Chem. Acta **77**, 345 (2004)
9. M. Randić, N. Novič, M. Vračko, J. Chem. Inf. Modell. **45**, 1205 (2005)
10. M. Randić, (ed) by P.M. Conn Handbook of Proteomics Methods (Human Press, Inc Totowa 2003) pp. 429–450
11. M. Randić, L.M. De Alba, J. Chem. Inf. Comput. Sci. **37**, 1078 (1997)
12. M. Randić, A.F. Kleiner, L.M. De Alba, J. Chem. Inf. Comput. Sci. **34**, 277 (1994)
13. E.W. Dijkstra, Numer. Math. **1**, 269 (1959)
14. G. Voronoi: Russian mathematician. His first paper (on factorization of polynomials) appeared while he was in high school (1868–1908)
15. G. Voronoi, J. Reine Angewandte Math. **133**, 97 (1907)
16. B. Delaunay: Russian mathematician, a student of G. Voronoii. The triangulation was invented in 1934
17. B. Delaunay, Izvestia Akad. Nauk SSSR, Otdel. Mat. Estest. Nauk **7**, 793 (1934)
18. N.L. Anderson, R. Esquer-Blasco, F. Richardson, P. Foxworthy, P. Eacho, Toxicol. Appl. Pharmacol. **137**, 75 (1996)
19. N. Trainjstić: Chemical Graph Theory, 2nd edn. (CRC Press, Boca Raton)

20. D. Cvetković, *A Combinatorial Approach to Matrix Theory and Its Applications* (CRC Press, Taylor and Francis Group, London, 2008)
21. D. Babić, D.J. Klein, I. Lukovits, Nikolić, N. Trinajstić, Int. J. Quant. Chem. **90**, 166 (2002)
22. O. Chan, T.K. Lam, R. Merris, J. Chem. Inf. Comput. Sci. **37**, 762 (1997)
23. L. Lovasz, J. Pelikan, Period Math. Hung. **3**, 175 (1973)
24. M. Randić, X.F. Guo, S. Bobst, DIMACS Ser. Discr. Math. Theor. Comput. Sci. **51**, 305 (2000)
25. MATLAB (abbreviation for Matrix Laboratory) is a product of The MathWorks, Inc., 3 Apple Hill Drive, Natick, MA 01760
26. M. Randić, F.A. Witzmann, V. Kodali, S.C. Basak, J. Chem. Inf. Model. **46**, 122 (2006)
27. M. Randić, J. Proteome Res. **5**, 1575 (2006)