# Milestones in Graphical Bioinformatics

Milan Randić,*[a] Marjana Novič,[a] and Dejan Plavšić[b]

After reviewing the field of graphical bioinformatics, we have selected two dozen of the most significant publications that represent milestones of graphical bioinformatics. These publications can be viewed as forming the backbone of graphical bioinformatics, the branch of bioinformatics that initiates analysis of DNA, RNA, and proteins by considering various graphical representations of these sequences. Graphical bioinformatics, a division of bioinformatics that analyzes sequences of DNA, RNA, proteins, and proteomics maps by developing and using tools of discrete mathematics and graph theory in particular, has expanded since the year 2000, although pioneering contributions date back to Hamory (1983) and Jeffrey (1990). We chronologically follow the development of graphical bioinformatics, without assuming that readers are familiar with discrete mathematics or graph theory. Readers unfamiliar with graph theory may even have some advantage over those who have been only superficially exposed to graph theory, in view of wide misconceptions and misinformation about chemical graph theory among quantum chemists, physical chemists, and medicinal chemists in past decades. © 2013 Wiley Periodicals, Inc.

**DOI: 10.1002/qua.24479**

## Introduction

We introduce the term "graphical bioinformatics" to emphasize the distinction between the part of bioinformatics concerned with comparative studies of biosequences based on direct computer-driven comparisons of primary DNA and protein sequences, and the part of bioinformatics dealing with graphical representations of DNA and proteins and their numerical characterization based on mathematical invariants extracted from graphical representations. As an important distinction between the two branches of bioinformatics, the former always simultaneously considers at least two sequences, while in graphical bioinformatics one can focus attention and characterize a single DNA, RNA, protein, or proteome.

Because a comprehensive review on graphical bioinformatics was recently published in the journal *Chemical Reviews*,[1] we will not dwell on details described therein. We will focus on our selection of the most significant results of graphical bioinformatics, to which we refer as the "milestones" of graphical bioinformatics. They are listed in Table 1. We will elaborate on a few recent results in graphical bioinformatics, reported during the last two years, and appearing after the publication of the above-mentioned review on the graphical representation of proteins.

We use the word "milestone" to signify "an important event in the advancement of knowledge in a field." The word "bioinformatics" does not have a uniform definition, and may be put in parallel with the widely used chemical concept "aromaticity," which most people know about, yet at the same time have difficulty in defining. The same can be said of bioinformatics; most people know what it is, yet at the same time have difficulty in formally defining it—but here the parallelism ends. In the case of aromaticity, chemists try to capture diverse aspects of aromatic molecules under an evasive unified theoretical model, yet have difficulty in accomplishing such a problematic task. Numerous developments in bioinformatics have introduced often unexpected novel directions that broaden previously established frontiers of this discipline. One such novel direction is graphical bioinformatics, a term recently coined. The origin of this discipline can be traced to the 1983 paper by Hamory and Ruskin,[2,27,28] who depicted DNA as a path in three-dimensional (3D) space, making it possible to visually compare different DNAs. Another outstanding early contribution of graphical bioinformatics was by Jeffrey,[29,30] who in 1990 modified the chaos game, (a mathematical construction for graphical representations of lengthy sequences of digits) for the graphical representation of DNA. The mathematician M. F. Barnsley, who developed an algorithm for graphical representations of lengthy mathematical sequences (often including random sequences of digits), named his algorithm "chaos game."[3,31,32] The chaos game graphical representations of DNA and other bio-sequences have been subsequently used for qualitative and visual inspections and comparisons of different DNAs. The introductory section of the above-mentioned review[1] illustrates these early graphical representations of DNA.

A. Nandy,[4,5,33,34] one of the early contributors to graphical bioinformatics, advocated a two-dimensional (2D) graphical representation of DNA, which has good visual qualities, despite a loss of information caused by the overlap of

[a] *M. Randić and M. Novič*
*National Institute of Chemistry, Ljubljana, Hajdrihova 19, Slovenia*
*E-mail: mrandic@msn.com*
[b] *D. Plavšić*
*NMR Center, Institute Rudjer Bošković, Zagreb, Bijenička cesta 54, Croatia*

Contract grant sponsor: The Ministry of Science and Higher Education of the Republic of Slovenia; contract grant number: P1-0017.
Contract grant sponsor: The Ministry of Science, Education and Sports of the Republic of Croatia; contract grant number: 098-0982929-2917.

opposite steps in plotting DNA as paths in 2D over the Cartesian grid.

One of the first important breakthroughs of graphical bioinformatics is the visual recognition of relative abundances and the distribution of bases in DNA, which can be used to determine potential protein coding regions, demonstrating the use of a 2D graphical representation of DNA sequences for intron-exon discrimination in intron-rich sequences.[6,7,35,36] For early developments of graphical bioinformatics, see the review article by A. Roy, C. Raychaudhury, and A. Nandy.[8]

In the year 2000, graphical bioinformatics saw an important novelty that resulted in the expansion from this so-far essentially qualitative graphical bioinformatics, a visual discipline, into a quantitative discipline of graphical bioinformatics, defined by the numerical characterization of DNA.[9,37,38] Soon followed extensions of this numerical characterization to RNA and the introduction of the first graphical representations of proteins accompanied by the numerical characterization of proteins. In 2001, the numerical characterization was extended to analyses of experimental data on proteomics maps, thus extending graphical bioinformatics to the quantitative (numerical) analysis of proteomics maps.[10,11] For early developments of the quantitative study of proteomics maps, readers may consult a review article on numerical characterization of proteomics maps by matrix invariants,[12] which appeared a year or two after the publication of the first article in this area, indicating the significance of the emergence of the quantitative study of proteomics maps.

This review cites the most significant publications in graphical bioinformatics, and readers can conclude which publications deserve recognition as milestones of graphical bioinformatics, and which elaborate on already introduced results.

## Milestones in Graphical Bioinformatics

Table 1 lists our view of the milestones of graphical bioinformatics by year, informative titles, and references. Table 1 covers 30 years of the initially very slow growth of graphical informatics, which was reborn in the year 2000 with a publication dealing with the numerical characterization of the graphical representation of the first exon of the human $\beta$-globin gene, as proposed by Nandy[33] and illustrated in Figure 1.

Nandy, from Calcutta, India, was visiting S. C. Basak at the Natural Resources Research Institute in Duluth, MN (associated with the University of Minnesota in Duluth) and presented a seminar on the graphical representation of DNA. At that time one of the authors (MR) too was visiting Basak and attending the seminar where Nandy also presented the DNA plot of the complete human $\beta$-globin gene, part of which is illustrated in Figure 1. The distance/distance (D/D) matrices[44,45], which were introduced into chemical graph theory half a dozen years ago to characterize the degree of bending of chain-like molecules, can be used for the numerical characterization of graphical representations of DNA, even though the DNA graphical representation is not a path graph, but a path over the Cartesian coordinate system. By numerical characterization, the construction of a set of invariants of graphical objects is understood, not a single number or a pair of numbers. This can be used for indexing DNA sequences instead of allowing numerical comparative studies of such diagrams.

**Table 1.** Milestones in graphical bioinformatics.

| | Year | | Ref. |
|---|---|---|---|
| 1 | 1983 | 3-D graphical representation of DNA | [27] |
| 2 | 1990 | Chaos Game representation of DNA | [3] |
| 3 | 1995 | Simplified graphical 2-D representation of DNA | [5] |
| 4 | 1996 | Recognition of potential coding regions in DNA | [6,7] |
| 5 | 1999 | Indexing macromolecular sequences | [8] |
| 6 | 2000 | Numerical characterization of 2-D DNA plots | [37] |
| 7 | 2001 | Numerical characterization of proteomics maps | [9,38] |
| 8 | 2003 | Spectral representation of DNA | [13] |
| 9 | 2004 | Graphical representation of DNA as a map | [14] |
| 10 | 2004 | Virtual genetic code | [39] |
| 11 | 2005 | Sequential neighbor labels for vertices of maps | [15] |
| 12 | 2005 | Hormesis at the proteome level | [16] |
| 13 | 2005 | Viral targeted applications | [17] |
| 14 | 2006 | Graphical alignment of DNA | [18] |
| 15 | 2006 | Alignment-free approach to phylogenetic analysis | [19,20] |
| 16 | 2007 | Graphical representation of proteins by graphs | [21,22] |
| 17 | 2008 | Graphical alignment of proteins | [109] |
| 18 | 2008 | Amino acid adjacency matrix | [23] |
| 19 | 2008 | Representation of RNA without loss of information | [25] |
| 20 | 2009 | Prediction of protein functional regions | [49] |
| 21 | 2012 | Novel 2D Representation of proteomics maps | [41] |
| 22 | 2012 | Exact solution to protein alignments | [42] |
| 23 | 2013 | Exact solution to nucleotide alignments | [43] |
| 24 | 2013 | Canonical labels for maps | |

Soon after the seminar with Nandy, we constructed the 92 $\times$ 92 size D/D for the first exon of the human $\beta$-globin gene. The DNA is shown in Figure 1, and a small portion is shown in Table 2. The location of the initial 12 nucleotides is shown in Figure 2. The significance of this work, which is outlined in Ref. [37], was that this step upgraded graphical bioinformatics into a quantitative theoretical discipline. Until that time graphical bioinformatics was a qualitative discipline, in which comparisons between graphical representations of different DNA were performed visually. As seen from Refs. [1] and [37] , the construction of the D/D matrices allows one to recover the lost information of the 2D graphical representation of DNA, making such graphical representations more useful than previously. The nature of the D/D matrix and some of its invariants, which can serve as DNA descriptors, are outlined here.

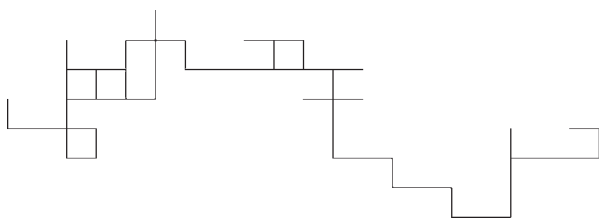

**Figure 1.** Graphical representation of the first exon of human $\beta$-globin gene according to the approach of A. Nandy. Reproduced with permission from Ref. [1].

**Table 2.** A small portion of the D/D matrix of the first exon of human $\beta$-globin gene. ATG GTG CAC CTG ACT CCT GAG GAG AAG TCT GCC GTT ACT GCC CTG TGG GGC AAG GTG AAC GTG GAT GAA GTT GGT GGT GAG GCC CTG GGC AG.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 1/1 | $\sqrt{2}/2$ | $\sqrt{5}/3$ | $2\sqrt{2}/4$ | $\sqrt{13}/5$ | $\sqrt{10}/6$ | $\sqrt{5}/7$ | 2/8 | $\sqrt{5}/9$ | 2/10 | $\sqrt{3}/11$ |
| 2 | | 0 | 1/1 | 2/2 | $\sqrt{5}/3$ | $\sqrt{10}/4$ | 3/5 | $\sqrt{2}/6$ | $\sqrt{2}/7$ | $\sqrt{8}/8$ | $\sqrt{2}/9$ | $\sqrt{10}/10$ |
| 3 | | | 0 | 1/1 | $\sqrt{2}/2$ | $\sqrt{5}/3$ | 2/4 | 1/5 | $\sqrt{2}/6$ | $\sqrt{5}/7$ | $\sqrt{2}/8$ | $\sqrt{5}/9$ |
| 4 | | | | 0 | 1/1 | $\sqrt{2}/2$ | 1/3 | 0 | 1/5 | 2/6 | 1/7 | $\sqrt{2}/8$ |
| 5 | | | | | 0 | 1/1 | $\sqrt{2}/2$ | 1/3 | 2/4 | 3/5 | 2/6 | $\sqrt{5}/7$ |
| 6 | | | | | | 0 | 1/1 | $\sqrt{2}/2$ | $\sqrt{5}/3$ | $\sqrt{10}/4$ | $\sqrt{5}/5$ | 2/6 |
| 7 | | | | | | | 0 | 1/1 | $\sqrt{2}/2$ | $\sqrt{5}/3$ | $\sqrt{2}/4$ | 1/5 |
| 8 | | | | | | | | 0 | 1/1 | 2/2 | 1/3 | $\sqrt{2}/4$ |
| 9 | | | | | | | | | 0 | 1/1 | 0 | 1/3 |
| 10 | | | | | | | | | | 0 | 1/1 | $\sqrt{2}/2$ |
| 11 | | | | | | | | | | | 0 | 1/1 |
| 12 | | | | | | | | | | | | 0 |

## D/D Matrix

The D/D matrix, or DD matrix, was initially constructed for the characterization of chain-like structures of fixed geometry with bonds of the same length but embedded in space with edges oriented in different directions. For example, the D/D matrix has been used for the characterization of graphs of Figure 3, which illustrates short paths that can be obtained by walking over the graphite network. The matrix elements (*i, j*) of the D/D matrix are given by the quotient of the Euclidean distance between vertices (*i, j*) and the length of the distance between vertices (*i, j*) along the path connecting them. From this definition, it is clear that if two structures have the same overall length (the same number of vertices), then the one that is more bent will have smaller D/D matrix elements, and consequently, smaller matrix row sums. Following Perron's theorem,[46–48] which states that (in the symmetrical matrices) the largest and the smallest row sums give the upper and the lower bounds on the leading eigenvalue, one expects that more bent structures will also have smaller leading eigenvalues. Hence, the leading eigenvalue is a good index measuring the degree of bending or folding of such structures.

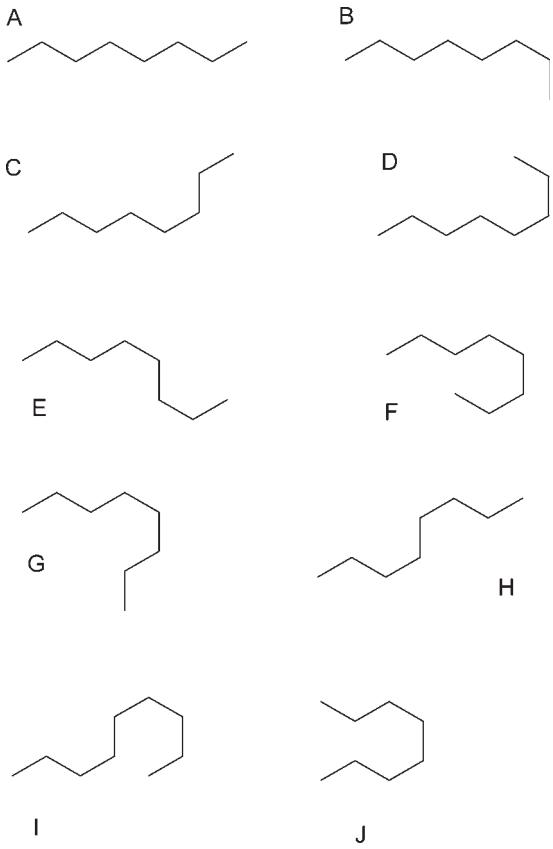


**Figure 3**. Graphs representing path of length 7 over graphite lattice.

The D/D matrix was later generalized to embedded paths in 2D or 3D having links of different lengths, which is useful for the characterization of proteomics maps.[12] Recently, the use of D/D matrices has been extended to acyclic graphs, that is, graphs having branching vertices and branches (Randić and Plavšić, in preparation). The leading eigenvalue also continues to be a useful structure descriptor for acyclic graphs. Beside the leading eigenvalue, the set of all eigenvalues of D/D matrices is of interest, as is the set of row sums, which must be first-ordered to qualify as a set of invariants. Recently, the coefficients of the leading eigenvector were found to parallel the abundances (the relative magnitudes of spots) in proteomics maps,[49] and thus are useful structure descriptors.

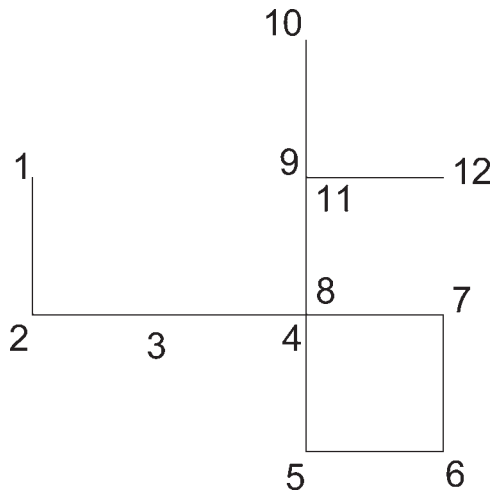


**Figure 2**. Graphical representations of the initial 12 nucleotides of the first exon of human $\beta$-globin gene of Figure 1. Reproduced with permission from Ref. [1].

## Lattice Representations of DNA without Loss of Information

The graphical representation of DNA by Hamory and by Nandy can be considered as 3D and 2D lattice representations of DNA such that all nucleotides have integer coordinates, ($x_i$, $y_i$,

$z_i$) and ($x_i$, $y_i$), respectively. The 2D graphical representations of DNA by Nandy, by Gates,[50,51] and by Leong and Morgenthaler[52] are accompanied by loss of information, because walking in opposite directions over the Cartesian coordinate grid introduces cancellations of random walk steps. Thus in the graphical representations of DNA by Nandy, each adenine (A) followed by guanine (G) and vice versa, and each thymine (T) followed by cytosine (C) and vice versa, retraces a previous step in the DNA sequence, and thus introduces loss of information in graphical representation. The resulting graphical representation is not unique and may stand for several different DNA sequences.

This serious limitation of 2D lattice representations of DNA can be lifted when such graphical representations are analyzed numerically by using D/D matrices, because in constructing the D/D matrix one follows the path and knows the exact coordinates of each nucleotide as construction proceeds. As shown first by Gou et al.[53] and later by others,[54–63] it is also possible to modify the graphical representation of DNA by Nandy,[33] and arrive at somewhat modified 2D graphical representations of DNA that are not accompanied by loss of information. The same applies to graphical representations of DNA by Gates,[50,51] and Leong and Morgenthaler.[52] Finally, one can design alternative 2D graphical representations of DNA that from the start are not accompanied by loss of information on DNA in the input information. Such graphical representations allow the reconstruction of the DNA sequence, as was the case with Hamory's 3D representations of DNA and Jeffrey's 2D chaos game representations of DNA. The next section outlines the four-line DNA representation, which depicts DNA by plotting successive nucleotides over four horizontal lines, each associated with a single nucleotide. Such 2D representations of DNA are referred to as "spectral representations of DNA" because they visually resemble molecular spectra.

## Spectral Representation of DNA

Spectral representations of DNA, proteins, and RNA have an advantage over many other 2D graphical representations of biological sequences in that the horizontal lines (4 lines in the case of DNA, 8 or 12 lines in the case of RNA, and 20 lines in the case of proteins) can be associated with numerical magnitudes and can be manipulated arithmetically. This allows cancellations of values when the differences in graphical representations are considered, if two different graphical representations are superimposed. Every cancellation identifies the same nucleotides or amino acids in different sequences, which facilitates the arrival at DNA, RNA, or protein alignments graphically.

The top of Figure 4 illustrates the spectral representation of the first exon of the human $\beta$-globin gene, and immediately below shows the spectral representation of the first exon of the opossum $\beta$-globin gene. The spots on the first horizontal line are assigned the numerical value of $+1$ and indicate nucleotide adenine; the spots on the second horizontal line are assigned numerical value of $+2$ and indicate cytosine; the spots on the third horizontal line are assigned numerical value of $+3$ and indicate guanine; while the spots on the fourth horizontal line are assigned numerical value of $+4$ and indicate thymine. Visual comparison of the two spectra shows that the degree of variations in the $\beta$-globin gene of humans and opossums are considerable. In contrast, Figure 5 shows spectral representations of the first exon of the goat and bovine $\beta$-globin genes, which are fairly similar. Figure 5 shows that graphical representations based on four horizontal lines allow one to identify that some spectra are more different than others, and also exactly where they are different. For example, Figure 5 shows that goat and bovine first exons of the $\beta$-globin gene differ around the site 39 and in the region 58–61. The detection of these minor differences is not as easy in many other 2D DNA graphical representations, as has been the case with four-line spectral representations of DNA.

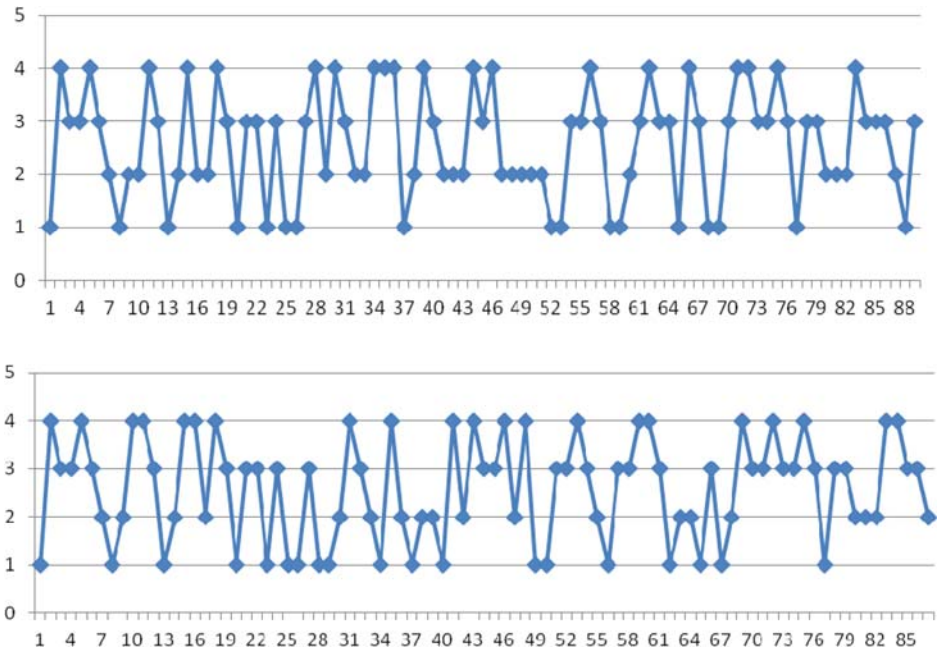


**Figure 4**. Spectral representation of the first exon of human (top) opossum (bottom) $\beta$-globin gene.

A criticism has been raised that spectral representations have limited visual qualities, which we dispute. After plotting the complete

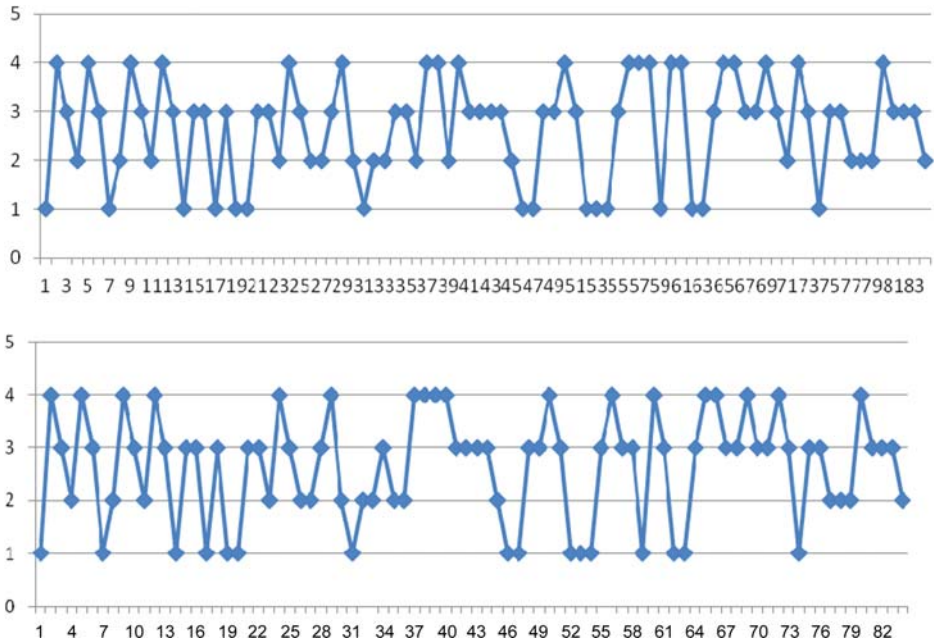

**Figure 5**. The first exon of goat (top) and bovine (bottom) $\beta$-globin gene.

$\beta$-globin genes (all three exons) of human and opossum, which have over 1400 nucleotides, Z.-J. Zhang[62] commented, "It is difficult to identify that the sequences [of human and opossum] are different. Because the visualization of this method become difficult when the DNA sequence is >300 bp" (sic).[62] This statement is subjective because, even though the two spectra in reference [63] have been reduced to 30 $cm^2$, they are different upon close examination. It is difficult to see quantitatively how different the spectra are, which is also true for other graphical representations, including the dual-vector curves of Z.-J. Zhang. Nandy's representations, which are already 2D, allowing the visual identification of different and similar DNA sequences despite loss of information. The same is true of graphical representations of DNA or proteins, which use lattice coordinates. The difference between spectral and lattice representations of DNA is that in spectral representations one assigns a single coordinate (value) to each nucleotide, but in lattice representations one assigns a pair of coordinates to each nucleotide. Figure 6 shows lattice representations for the first exons of the $\beta$-globin genes of human and opossum, and Figure 7 shows lattice representations the first exons of the $\beta$-globin genes of goat and bovine.

The lattice representation of DNAs in Figures 6 and 7 are based on grouping pairs of nucleotides, to which the following coordinates are assigned:

$$\begin{array}{cccc} AA(8,-6) & CA(7,8) & GA(8,2) & TA(8,-8) \\ AC(8,-4) & CC(5,8) & GC(8,4) & TC(6,-8) \\ AG(8,-2) & CG(3,8) & GG(8,6) & TG(4,-8) \\ AT(8,0) & CT(1,8) & GT(8,8) & TT(2,-8) \end{array}$$

Other choices of coordinates are possible and will show similar results. Because Nandy's 2D representation of DNA

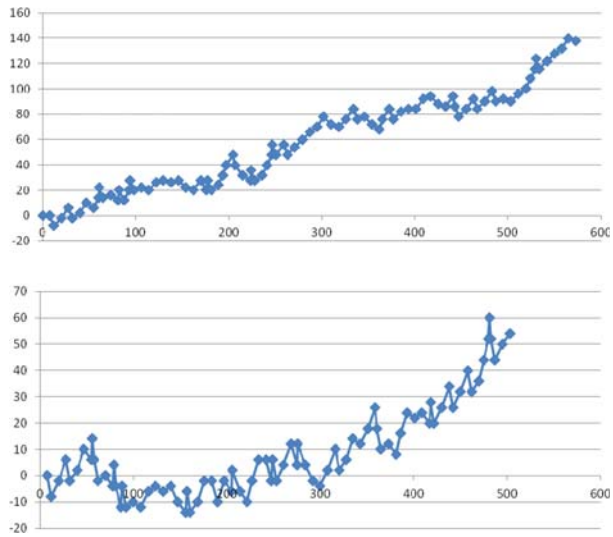

**Figure 6**. Lattice representation of the first exon of human (top) and opossum (bottom) $\beta$-globin gene.

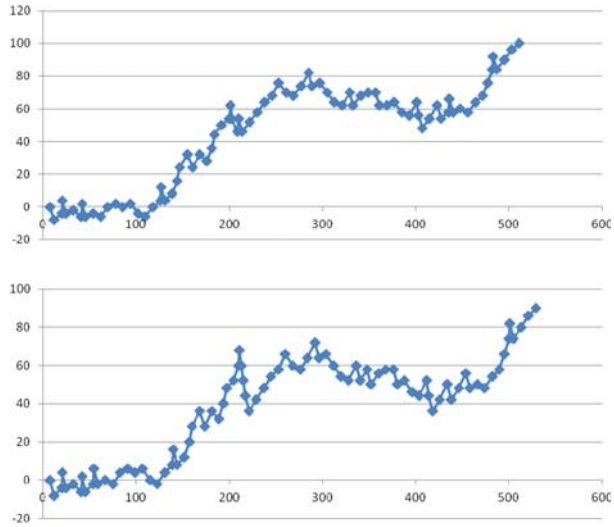

**Figure 7**. Lattice representation of the first exon of goat (top) and bovine (bottom) $\beta$-globin gene.

nucleotides A and G move along the *x*-coordinate in opposite directions, the coordinates for pairs starting with A and G are chosen to move forward along the *x*-coordinate, while nucleotides C and T, which move along the *y*-coordinate in opposite directions in Nandy's 2D representation of DNA, are chosen to move in opposite directions along the *y*-coordinate. We have selected the coordinates of C and T in opposite (but nonoverlapping) directions so that the length of the overall spectra is somewhat reduced. DNA can be plotted as a lattice graph by assigning vectors to single nucleotides A, C, G, and T directed to the set of coordinates

$$\mathrm{A}(1,-2);\mathrm{C}(2,-1);\mathrm{G}(2,1);\ \text{and}\ \mathrm{T}(1,2).$$

Figure 8 illustrates the lattice graph for the human $\beta$-globin gene based on the above coordinates. This graph is similar to one obtained using the set of coordinates considered by Yau et al.:[54]

$$\mathrm{A}(1/2,-\sqrt{3}/2);\mathrm{C}(\sqrt{3}/2,-1/2);\mathrm{G}(\sqrt{3}/2,1/2);\ \text{and}\ \mathrm{T}(1/2,\sqrt{3}/2),$$

except that now the (*x*, *y*) coordinates are not lattice points (integers).

To avoid information loss by accidental cancellations of opposite movements, the directions for up and down movements are shifted by changing the respective *x*-coordinates by one unit. The lattice representation in Figure 6 shows that the first exon in the human and opossum $\beta$-globin gene are fairly different, while Figure 7 shows that the first exon in goat and bovine are fairly similar.

The 2D ladder-like graphical representation of DNA by Li and Hu,[63] which follows a binary code for a 3-component vector, is an illustration of lattice representation of DNA. This originates from the pairwise partitions of A, C, G, and T as purine and pyrimidine, as amino and keto groups, and as weak and strong hydrogen bonds. For example, when the first exon of human $\beta$-globin gene is coded based on purine and pyrimidine classification of nucleotides, according to Li and Hu the following binary sequence is obtained:

10110101000110000001111111110001001001001000010111101111101110101110111100110110111100000111011

If one starts at the origin (0, 0) and moves along the *x* coordinate for each nucleotide shown as "one" and along *y*-coordinates for each nucleotide shown as "zero," one obtains one component of the 2D ladder-like graphical representation of DNA shown in Figure 9.

In our view, whether graphical representations of biosequences appear pleasing to the eye is less important than the numerical characterizations that they carry, which allow quantitative estimates of the degree of similarity or dissimilarity between different DNA, RNA, or proteins. To find how quantitatively different two or more DNAs, RNAs, or proteins are, constructed graphical curves should be analyzed numerically.

Figures 10 and 11 show spectral representations of the first exons of human and opossum, and goat and bovine, but instead of plotting the sites of nucleotides individually, we have grouped nucleotides into codons and assigned them to triplets of nucleotides, making a codon the average value of the numerical values of the three nucleotides forming the codon. For example, for the first codon of the human $\beta$-globin gene ATG, we assigned the value 2.6667: the average of 1, 4, and 3, which correspond to A, T, and G, respectively. The spectral representations of the first exon of the human, opossum, goat, and bovine $\beta$-globin gene based on codons show even more clearly that the first exons of human and opossum are very different, but that of goat and bovine differ little.

## Graphical Approach to the Alignment of DNA

To Find alignments of two DNA sequences, one can take advantage of numerical values associated with the four horizontal lines that represent A, C, G, and T and subtract the spectra of two DNAs to be aligned. This identifies sites where nucleotides in two sequences are equal. Then, with shifting, the two DNA sequences relative to one another are followed by one or more steps, and again their spectra are subtracted. Each time a coincidence in nucleotides is present, it will show as zero in the difference spectra. This is illustrated in Figure 12

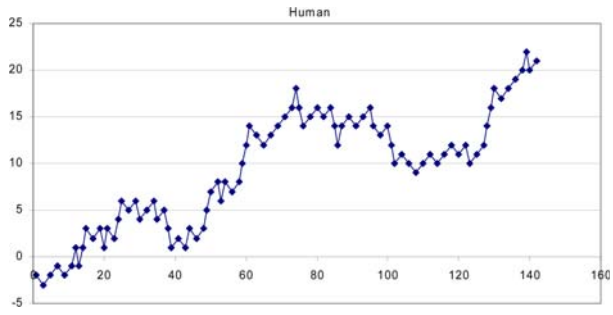


**Figure 8**. Novel lattice representation of DNA with no loss of information: The first exon of human $\beta$-globin gene using vectors: A →(+ 1, – 2,); T →(+ 2, –1); G →(+2, + 1); C →(+1, +2). Reproduced with permission from Ref. [1].

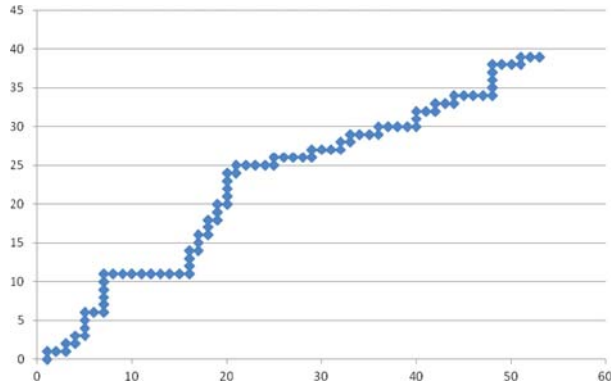


**Figure 9**. The 2-D ladder-like graphical representation of one of the component of the first exon of the human $\beta$-globin gene.

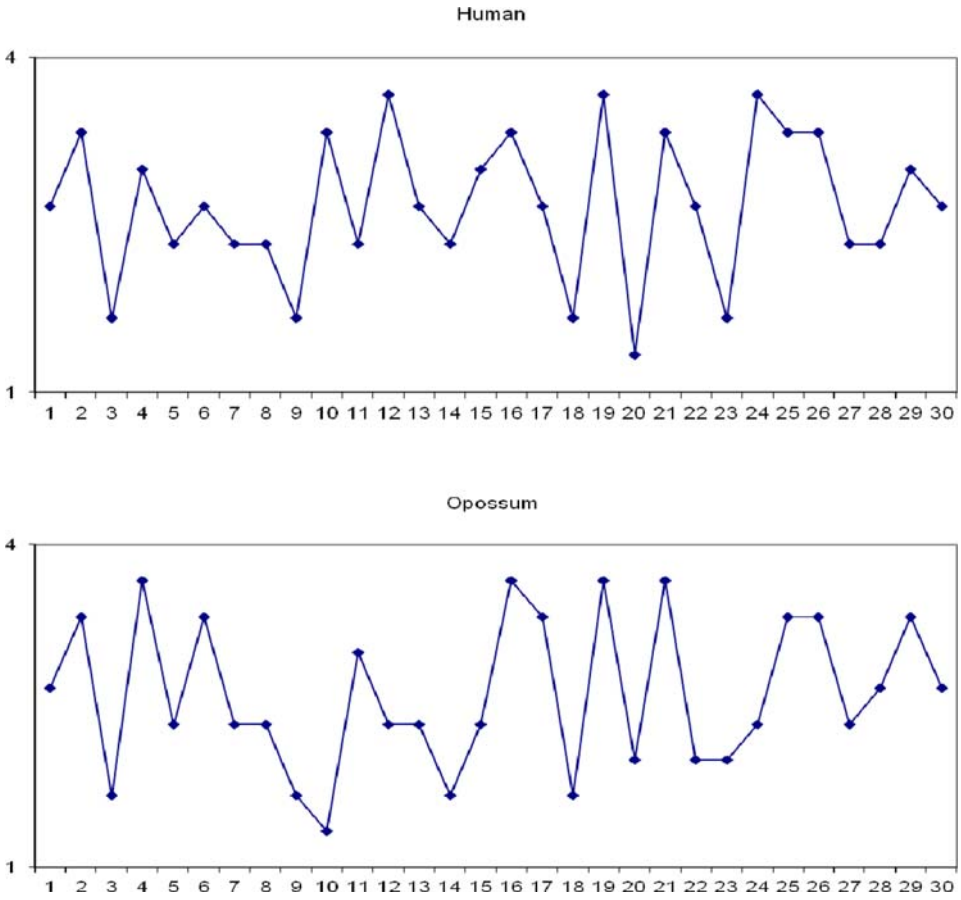


**Figure 10**. Graphical representation of codons of the first exon of human and opossum $\beta$-globin gene.

in the search for the alignment of the first exons of the $\beta$-globin genes of goat and bovine.

Figure 12 shows four different spectra more closely. The top shows the difference in the spectra of the first exons of the $\beta$-globin genes of goat and bovine. There is a full cancellation of spectral amplitudes only at the leftmost part of the spectra, signifying that the initial seven nucleotide doublets of the two DNA are identical. Figure 4 shows that the initial eight nucleotides are the same, but upon consideration of pairs of adjacent nucleotides, this gives seven doublets. The second picture of Figure 12 shows the spectral difference when the two DNA sequences have been shifted by a single place. Suddenly, a long segment of zeros signifies identical fragments of about 40 nucleotides in both sequences (with the exception of a few nucleotides in the middle of this fragment). When the two DNA sequences have been shifted by two steps, as shown in the third picture, the tail part of the two DNAs is practically identical (with a single nucleotide pair exception). This almost accounts for all differences between the two DNA, except for a short section involving a half dozen pairs of the last nucleotide in the central part of the two DNA. Continuing from the last spectral difference obtained by shifting the two sequences for an additional step, an additional half dozen nucleotides are fully aligned.

Figure 12 demonstrates the graphical approach to search for the alignment of DNA based on spectral representations of DNA, which was first demonstrated in 2006.[18] A graphical approach to searching for alignment of proteins based on spectral representations of proteins was developed the following year.[22] Both these publications introduced the use of this novel tool of graphical alignment to solve problems in biology. Based on the limited number of citations that these publications received, it appears that interest is low, even though both papers on the graphical alignment (of DNA and proteins) were published in respectable journals. The initial paper on DNA alignment was based on graphical representations of individual nucleotides A, C, G, T,[22] while here (Figures 4–9 and 12) the spectral representations are based on pairs of adjacent nucleotides.

## Graphical Approach to the Alignment of Proteins

In order to obtain a one-dimensional (1D) spectral graphical representation of proteins analogous to the four-line DNA representation, each of the 20 amino acids can be assigned a numerical value, such as entries from 1 to 20. Another possibility is the use of angular polar coordinates of amino acids, which are uniformly arranged on the circumference of the unit circle. Similarly, the 64 codons can be arranged uniformly on the periphery of the unit circle and assigned polar angles (multiples of $2\pi/64$ radians), leading to a 1D "spectrum-like" representation of DNA based on codons. In these 1D representations of DNA (based on the four nucleotides or codons), or of protein sequences, alphabetic sequences of nucleotides or amino acids are transformed into numerical sequences. Numerical sequences allow simple numerical operations to be performed on the elements of the sequence, such as subtracting the corresponding members of two sequences, or subtracting sequences that have been shifted one relative to one another. The next section shows that this is the essential step for graphical solutions to the problem of DNA and protein alignment.

Until now there was no rigorous solution to the problem of protein–protein alignments. The existing algorithms for protein alignments[64–69] involve dynamic programming, probabilistic approaches, genetic algorithms, graph-theoretical approaches, and empirical parameters. Some computer-based approaches consider penalties for the deletion, substitution, and permutation of sequence labels (i.e., amino acids), which are associated

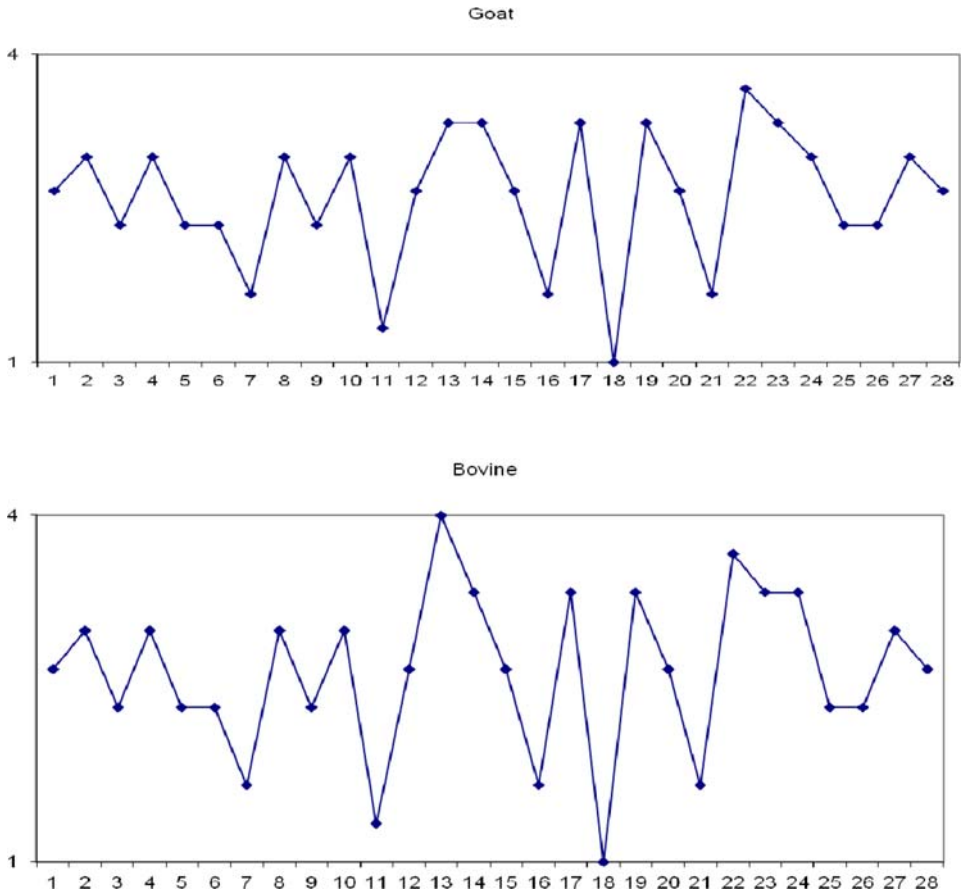


**Figure 11.** Graphical representation of codons of the first exon of goat and bovine $\beta$-globin gene.

with the metrics of Levenshtein,[70] also known as "edit distance." In contrast to these computer-based programs for protein alignment, which search for an optimal alignment of proteins when various penalties for deletions, substitutions, and gaps are assumed, graphical approaches consider direct comparisons of two protein sequences after numerical values have been assigned to different amino acids. The recently outlined graphical approach to protein alignment identifies the same amino acids in two protein sequences by locating the zeros on the plot of the difference between two numerical representations of two proteins. To arrive at a complete analysis, however, the differences between sequences of proteins when shifted by one or more positions relative to the other, both to the left and to the right, must be considered.

The graphical alignments of the two proteins of Table 3, which have almost 170 amino acids,[71] are illustrated. The first protein pertains to carboxypeptidase Y from *Saccharomyces cerevisiae* (baker's yeast), and the second belongs to the mature putative serine carboxypeptidase in ESR1-IRA1 intergenic region, also from *Saccharomyces cerevisiae*. Figure 13 illustrates 1D graphical representation of the two proteins.

Figure 13 reveals 20 different spectral amplitudes. The values having the same "height" in the spectrum correspond to the same amino acid. Thus, the spots at the top line in Figure 13 and the bottom line of the spectra corresponding to valine and alanine immediately indicate that protein 1 has 12 valine ($y = 6$) and seven alanine AAs ($y = 0$), whereas protein 2 has five valine and nine alanine AAs. The count of the number of spots on the same horizontal line gives the abundance count for amino acids in proteins. In the case of protein 1 and protein 2 of Figure 13, for the 20 amino acids, ordered (A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V), which is alphabetical according to their three-letter codes, one obtains

$(7, 2, 13, 11, 1, 4, 7, 18, 10, 4, 15, 2, 0, 15, 11, 16, 8, 3, 8, 12)$ for protein 1, and
$(9, 3, 11, 9, 1, 3, 9, 17, 13, 6, 13, 9, 4, 14, 11, 15, 7, 4, 8, 5)$ for protein 2.

To identify repeating adjacent occurrences of the same amino acid in a sequence, the spectra are searched for locations of adjacent "spots" on the same horizontal line. In protein 1 there are AA, GG, FFF, FF twice, and SS thrice; while in protein 2 there are NNN, GG, HH, LL, FF thrice, and SS twice.

The 20-component "abundance" vectors allow a fast preliminary screening of proteins for their similarity or lack thereof. The similarity of the 20-component vectors is a necessary, but not sufficient, condition for similarity among proteins. Abundance vectors tell nothing about distributions of amino acids, but a glance at such vectors for two proteins gives insight on their degree of similarity of two proteins. A comparison of the above two 20-component abundance vectors suggests that protein 1 and protein 2 have an appreciable degree of similarity. A plot of the two 20-component vectors representing the abundance of the two proteins against one another is shown in Figure 14, which shows a fair correlation with three outliers: alanine (A), methionine (M), and leucine (L).

The plot of the difference of the spectral representations of protein 1 and protein 2 oscillates above and below the *x*-axis (Fig. 15). This is because there are no significant segments of amino acids in two proteins that overlap, which would result in differences equal to zero, except for a few accidental cases. But when the two sequences are shifted by one or two steps, the diagrams in Figure 16 show alignments for amino acids in a significant portion of two proteins. The shift of two sequences by one step gives alignments of amino acid in the region 22–99; the shift of the two protein sequences by two steps

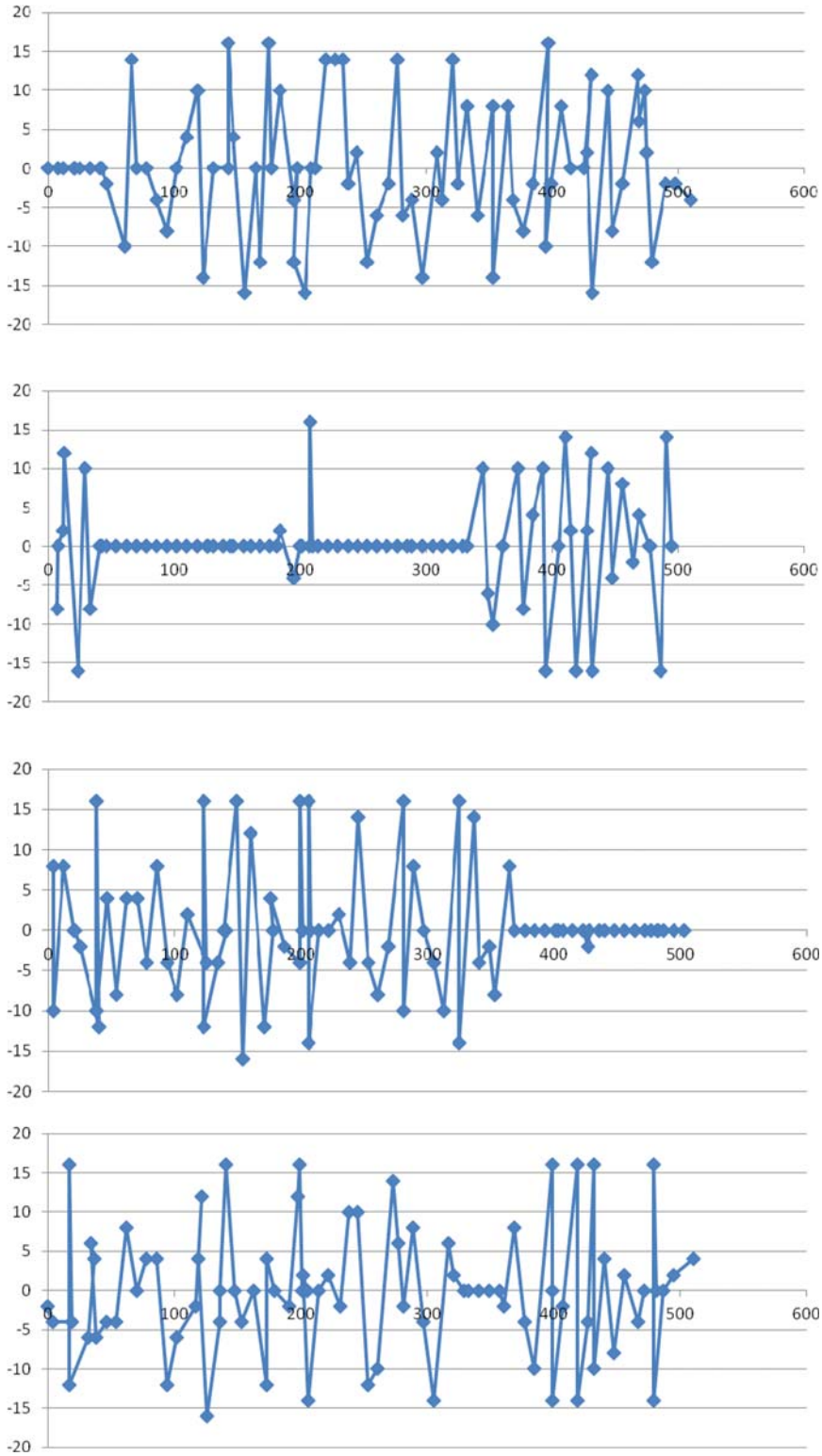

**Figure 12.** The difference between goat and bovine specteral representations of $\beta$-globin gene (top) and difference when the two sequences are shifted by one step (next), two steps (next), and three steps (bottom).

shows alignment between the two proteins in the region 108–120. When the shift of the two sequences continues farther, by three and four steps, additional local alignments of amino acids are in the regions 159–169 and 130–145, respectively.

Figure 16 represents the essence of the novel graphical approach to the protein alignment problem. By combining the information obtained by considering the difference in spectra for the four shifts of protein 1 and protein 2, the alignment pattern for the two proteins can be constructed. The search for additional local alignments can be continued, but this is not essential for the outline of the novel graphical approach for protein alignment. The four shifts of the spectrum-like (20 lines) representations of proteins achieved an overall matching in 117 sites out of 169.

The resulting graphical alignment was obtained without considering penalties for various gaps. The graphical alignment approach represents an alternative searching route for protein alignments, which is conceptually and computationally simple. But even at this early stage of its development, it is possible to conceive further improvements. All graphical displays and all computations discussed in this review can be easily performed in Excel, which is particularly suitable for such work.

Some readers may view this route to protein alignment as having limited potential, not competitive with currently available computer packages such as FASTA[66,67] and BLAST.[68] This may be true now and in the immediate future, but the graphical approach to protein alignment has just emerged, while many computer-graphic packages have been available for longer (25 and 15 years, respectively). Novel aspects of the graphical alignment of proteins may be seen in the future. In the case of a DNA alignment, which is described in reference[31] and follows the route outlined here for graphical alignments of proteins[22] (even though the publication on DNA appeared earlier), graphical alignment can successfully reproduce the computer-based result, and has also shown that there are better solutions not detected by the particular computer program.

## Hormesis at the Proteome Level

This section briefly outlines the route used for the numerical characterization of proteomics maps, which has an outstanding result: The recognition of the presence of hormesis at the

**Table 3.** Two proteins of *Saccharomyces cerevisiae* selected for outline of the VESPA algorithm. Amino acids are listed in groups of ten for easier reading.

| |
|---|
| Protein 1 |
| KILGIDPNVT QYTGYLDVED EDKHFFFWTF ESRNDPAKDP VILWLNGGPG |
| CSSLTGLFFE LGPSSIGPDL KPIGNPYSWN SNATVIFLDQ PVNVGFSYSG |
| SSGVSNTVAA GKDVYNFLEL FFDQFPEYVN KGQDFHIAGE SYAGHYIPVF |
| ASEILSHKDR NFNLTSVLIG NGLT |
| Protein 2 |
| PSKLGIDTVK QWSGYMDYKD SKHFFYWFFE SRNDPANDPI ILWLNGGPGC |
| SSFTGLLFEL GPSSIGADMK PIHNPYSWNN NASMIFLEQP LGVGFSYGDE |
| KVSSTKLAGK DAYIFLELEF EAFPHLRSND FHIAGESYAG HYIPQIAHEI |
| VVKNPERTFN LTSVMIGNGI T |

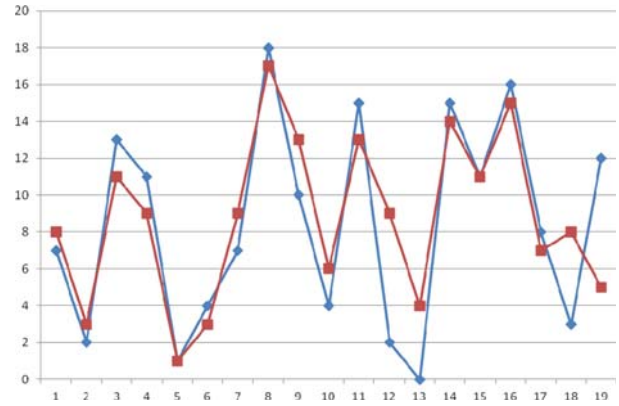

**Figure 14.** The relative abundances of the 20 amino acids in the two proteins.

proteome level. Up to that time hormesis, which advocated a J-shaped response curve rather than a simple linear dose-response, had been recognized for many years by a number of research circles as a possible dose-response of the whole organism. An early illustration, for example, is the effect of a lethal dose of radiation on rats never exposed to radiation and rats previously exposed to small doses of radiation.[72] Despite available evidence for the J-shaped response curve for some time, hormesis has not been accepted or acknowledged by several leading authorities.[73,74] By reinvestigating the available proteomics data of Andersen et al.,[75] it was demonstrated for the first time in 2005 that a J-shaped dose-response is also characteristic of the proteome variations in individual cells of an organism, even though the variation of individual protein abundance appears chaotic.[12] When this article was reviewed, an anonymous referee sent the single sentence report, "This paper will be highly cited."

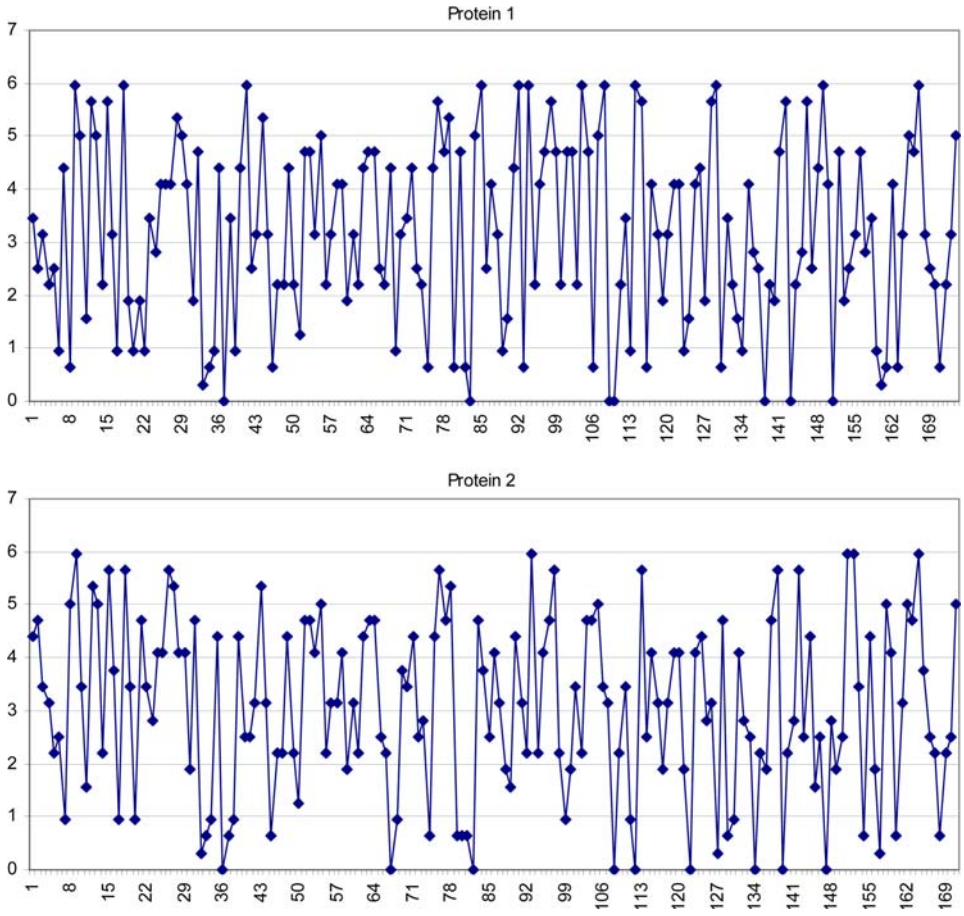


**Figure 13.** The radian coordinates of the corresponding amino acids of the two proteins of Table 3. Reproduced with permission from Ref. [22].

About seven years have passed since the publication of this work, but as of September 2012, the total number of citations is only 35. This is about five citations per year (which includes self-citations), allowing three conclusions:

1. One of the most difficult jobs is to predict the future
2. There are too few researchers who can recognize and appreciate the significance of novelty in research and the significance of results that are outside their narrow field of interest.
3. There is at least a single authority (the anonymous referee) in the field who recognized an important discovery at its early stage.

It could have happened, although it did not in this case, that not a single supporting scientist would appreciate the novelty of this work. This is not unknown in science when true

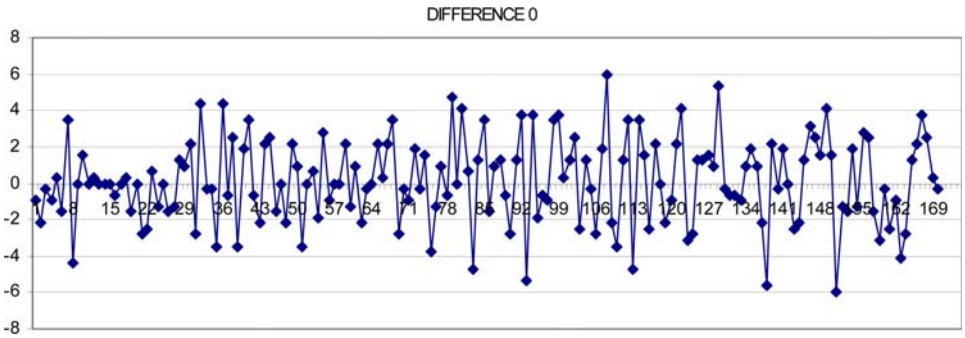


**Figure 15.** The difference in the radian coordinates of the corresponding amino acids of the two proteins of Figure 13. Reproduced with permission from Ref. [35].

novelty has been discovered. This continual overlook by authorities of the novelty of some scientific contributions is discouraging and inspired the quote, "It is more important to have a view of a single scientist who understands what one is doing than worry about 100 that do not understand what one is doing."[76,77]

For example, in theoretical chemistry this was the case with the emergence of the density functional theory (DFT), when very few quantum chemists recognized the significance of the work of Kohn and were hostile to DFT (exceptions were R. G. Parr and J. A. Pople, the two leading theoretical chemists in the world). This is how Walter Kohn describes reception of his work:[78]

In those early years of DFT, the community of theoretical chemists felt, almost without exception, that this approach had nothing useful to offer to them. Occasionally, I was invited to give a paper on their meetings, but I had the feeling that most of the audience expected to confirm their conviction that it was full of irremediable defects, in particular, insufficient accuracy and the absence of guaranteed, systematic procedure to improve it. The most notable exception was Bob Parr.

The situation changed dramatically in 1998 when Walter Kohn received the Nobel Prize in Chemistry for his work! He shared the Nobel Prize with John A. Pople, another nonhostile exception among quantum chemists toward DFT.

Sooner or later graphical bioinformatics will gain recognition, due to an undeniable continuation of growth that will eventually make its presence obvious. Perhaps the disappointing citation results should have been expected, because most chemists, including theoretical chemists and particularly quantum chemists, are unfamiliar with discrete mathematics and graph theory (which can be viewed as a part of discrete mathematics), as were their professors and will be their students. However, this is not the case with computer scientists, the "tool" makers in chemistry and bioinformatics, although it may continue for a while with tool-users until some spectacular new result emerges. We believe that the situation with graphical bioinformatics will soon change, possibly dramatically and at least in bioinformatics circles, when most users learn of the latest results in graphical bioinformatics that cannot be overlooked: the exact solution to protein and DNA sequence alignments, to be outlined in the final sections of this review.

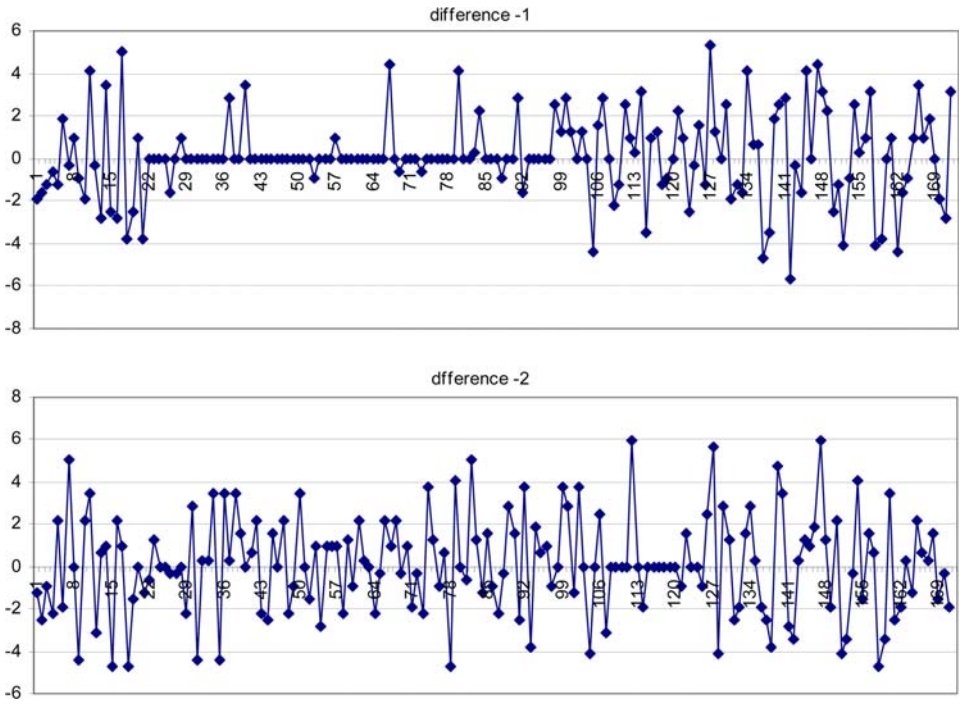


**Figure 16.** The difference in spectral coordinates of the corresponding amino acids of carboxypeptidase Y from Saccharomyces cerevisiae (top) and amino acids of mature putative serine carboxypeptidase in ESR1-IRA1 intergenic region also from saccharomyces cerevisi shifted to the left for one to four places and to the opposite direction by one step. Reproduced with permission from Ref. [22].

## Proteomics Map and Their Numerical Characterization

The proteomics maps data of Anderson et al.,[75] a leading authority of this experimentally difficult area of reporting high-quality, reproducible data, is considered here. Table 4 lists scaled abundance values for the control group (based on the 20 most abundant proteins of liver cells of mice) and four additional cases of mice after the

**Table 4.** Scaled abundance values.

| Spot | Control | PFOA | PFDA | Clofibrate | DEHP |
|---|---|---|---|---|---|
| 1 | 5.200 | 3.916 | 3.423 | 5.299 | 5.976 |
| 2 | 5.174 | 5.604 | 6.794 | 5.760 | 5.586 |
| 3 | 4.923 | 4.102 | 5.413 | 5.895 | 0.292 |
| 4 | 4.582 | 3.572 | 2.632 | 2.761 | 4.038 |
| 5 | 4.272 | 4.063 | 1.793 | 3.958 | 5.000 |
| 6 | 4.140 | 6.933 | 7.982 | 5.983 | 6.506 |
| 7 | 4.044 | 2.114 | 1.402 | 2.636 | 2.777 |
| 8 | 3.923 | 0.940 | 1.828 | 1.654 | 4.209 |
| 9 | 3.539 | 3.284 | 2.989 | 3.033 | 3.348 |
| 10 | 3.372 | 2.996 | 2.267 | 2.877 | 3.940 |
| 11 | 3.242 | 4.659 | 4.048 | 4.058 | 4.301 |
| 12 | 3.124 | 2.549 | 2.835 | 3.810 | 4.189 |
| 13 | 3.056 | 2.659 | 1.638 | 2.590 | 3.510 |
| 14 | 2.972 | 2.665 | 2.683 | 3.051 | 3.190 |
| 15 | 2.953 | 0.581 | 0.594 | 2.164 | 5.367 |
| 16 | 2.883 | 2.785 | 2.885 | 2.739 | 3.633 |
| 17 | 2.876 | 0.749 | 0.472 | 1.398 | 1.939 |
| 18 | 2.622 | 2.751 | 1.900 | 2.003 | 2.789 |
| 19 | 2.600 | 2.809 | 2.175 | 1.686 | 2.814 |
| 20 | 2.502 | 1.363 | 0.581 | 2.059 | 2.568 |
| Sum | 72 | 61.095 | 56.334 | 65.417 | 75.974 |

ingestion of four different peroxisome proliferators. The scaling is based on experimental data taken from the work of Anderson et al.[75] Figure 17 shows the positions of the 20 most abundant protein spots labeled 1–20 in this proteomics map. Only 20 protein spots are chosen for analysis because neither the number of selected points nor the criteria of selection is essential for the development of a mathematical approach. The variability of experimental data in proteomics could be significant, so it is best to focus on the most abundant proteins, the experimental errors for which are expected to be the least. The number of protein spots sufficient to represent a map or cellular proteome as a whole was considered[79,80] and appears to be one order of magnitude larger, not two or three orders of magnitude.

It is obvious that many invariants are needed to capture salient features of information-rich proteomics maps. The use of partial ordering is one of several routes to the numerical characterization of proteomics maps, and represents a continuation of our efforts[81–87] to develop the mathematical characterization of DNA, proteins, and proteome. The partial ordering diagram shown in Figure 18 is based on protein spots ordered with respect to their charge and mass, and connecting lines are embedded over the proteomics map. An important feature of the embedded graph of Figure 18 is that all lines connecting spots in the graph have positive slope. This is a consequence of partial ordering and the underlying dominance relation, and it is the property that can be used for a direct construction of the partial ordering diagram for a given map without the need to search for components of partial ordering. Partial ordering means ordering items (here, points having two coordinates) so that if one follows the diagram along the connecting lines from top to bottom, both components ($x$, $y$ coordinates) always dominate (are bigger than) those that follow.

To obtain Figure 18 directly from Figure 17, one can start with the top vertex (spot 1 in Figure 17) and connect it to the most left lower spot of 1, which is spot 7. Continue the same with vertex 7 and connect it to the next lower vertex below it and to the left, which is spot 3. Continue to connect 3 to 19, and finally 19 to 14. By exhausting this particular trail, return to vertex 1 and repeat the process: connect 1 to the next most left lower spot still unconnected, spot 9, and then 9 to 3. In the next step connect 1 again to the next most left lower point still unconnected, which is spot 6, and finally connect 6 to 14. By backtracking, connect 6 to 11. Finally, connect 1 to 16 and 1 to 13, which are connected to 6 and 11, respectively.

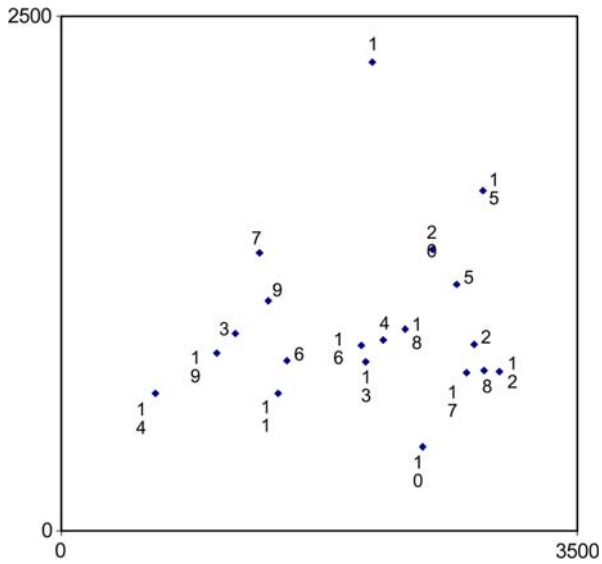


**Figure 17.** Location of 20 most abundant proteins for proteomics maps of the control group. Reproduced with permission from Ref. [1].

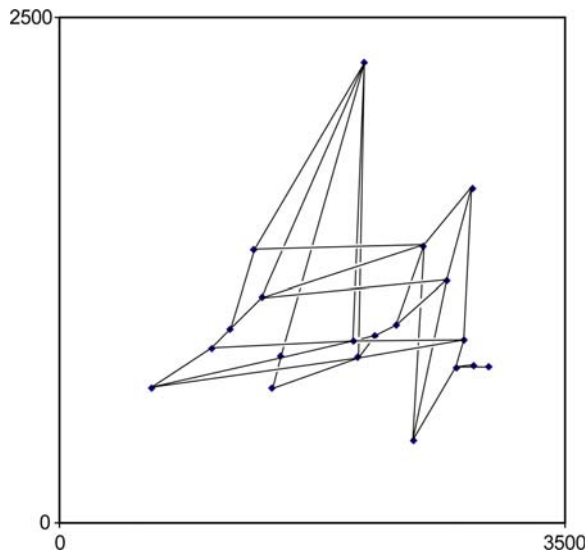


**Figure 18.** Partial ordering diagram for 20 protein spots of Figure 17. Reproduced with permission from Ref. [1].

This exhausts all the fragmentary orders starting with protein spot 1. The process continues with spot 15, then 8 and 12, which completes the construction of the embedded graph of partial order for the map considered.

The adjacency matrix of the partial ordering diagram can now be constructed, the matrix elements of which are defined as

$A_{ij}=1$ if vertices $i$ and $j$ are adjacent; and

$A_{ij}=0$ otherwise.

For the graph of partial ordering of the proteomics map illustrated in Figure 18, the adjacency matrix is shown in Table 5. The numerical characterization of the five proteomics maps of Table 4 is based on the augmented adjacency matrix, which is obtained by replacing zeros on the main diagonal of the matrix by the relative abundances of individual protein spots in the corresponding proteomics maps.

When experimental quantities are measured in different units, such entries must be suitably normalized so that neither of the properties ($x$, $y$) numerically dominates the other. In such situations, Kowalski and Bender[88] recommended that one rescale the units used to the same numerical interval, such as ($-1$, $+1$). There are three quantities that are combined into our analysis: protein charge (coordinate $x$), protein mass (coordinate $y$), and protein abundance (coordinate $z$). The $x$, $y$ coordinates do not enter directly into analysis, but determine the adjacency of the spots, while the abundance of the 20 proteins are incorporated by augmenting the adjacency matrix by the introduction of 20 nonzero diagonal elements.

Our problem has an additional complication because we use matrices (mathematical objects), not just a list of tabular data. In such situations, it is important that scaling is size-consistent, so that if the matrix is enlarged with new data, its elements are renormalized. Reference [89] suggests that a way to arrive at a matrix in which both the off-diagonal entries and the diagonal entries have balanced roles is by scaling both such that their sum is equal. The normalized abundances for the five proteomics maps have been listed in Table 4. The last row in Table 4 gives the abundance sums for the five maps, which immediately show the overall decrease of protein total for the most abundant 20 proteins for three peroxisome proliferators—perfluorooctanic acid (PFOA), perfluorodecanic acid (PFDA), and clofibrate—and an increase for the peroxisome proliferator di(2-ethylhexyl)phthalate (DEHP). The five matrices for the five proteomics maps differ in diagonal entries, which reflect on the role of drugs inducing changes in the proteomics maps. The constructed augmented matrices are analogous to similar matrices that differentiate heteroatoms in molecules in the construction of the variable connectivity indices.[90–103] A similar approach of differentiation among proteomics maps associated with different drugs and other xenobiotic agents was used earlier in the literature on the mathematical characterization of proteomics maps using zigzag lines.[10,104] However, the normalizations used there were not adjusted to incorporate the dependence of matrix elements on the matrix size.

Table 4 compares variations in abundances of individual proteins, when different drugs have been tested. In many cases, there are considerable changes in abundances in comparison with that of the control group. Protein 15, in the case of PFOA and PFDA, has considerably decreased its abundance, but in the case of DEHP it has increased its abundance. Assuming that the changes are statistically significant, abundances of proteins increased slightly after exposure to the four peroxisome proliferators (like proteins 2 and 11). Similarly, some proteins diminished their abundance, although often not evenly (like proteins 7, 9, and 17). Protein 14 appears to be among the least affected by any of the four agents considered. Quantitative characterizations of such changes in the relative abundance of proteins in cells exposed to different agents may facilitate a better understanding of the possible

**Table 5.** Adjacency matrix for the partial ordering graph of Figure 3.

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

**Table 6.** Similarity/dissimilarity among perturbations of abundances of proteome rat liver cells for the control and the four peroxisome proliferators based on 20-components leading eigenvectors shown in Table 7.

| | Control | PFOA | PFDA | Clofibrate | DEHP |
|---|---|---|---|---|---|
| Control | 0 | 2.279 | 2.758 | 1.182 | 1.264 |
| PFOA | | 0 | 0.651 | 1.097 | 1.102 |
| PFDA | | | 0 | 1.600 | 1.540 |
| Clofibrate | | | | 0 | 0.493 |
| DEHP | | | | | 0 |

existence of "stationary" states of cell proteomes and their diversity.

Table 6 shows pair-wise similarity/dissimilarity comparisons of the five proteomics maps based on the degree of similarity/dissimilarity for the corresponding leading eigenvectors, which are listed in Table 7. The values in the table were computed by viewing each column in Table 4 as a 20-component vector. The Euclidean distance (in 20-dimensional vector space) gives the distance between the corresponding endpoints of vectors. The smaller are the distance, and the more similar are the vectors (or alternatively the more similar are the corresponding proteomics maps). The first row in Table 6 gives the similarity of the four perturbed maps, with the map of the control group based on the 20 most abundant spots. As shown between the four peroxisome proliferators, clofibrate and DEHP cause the least perturbation of liver cell proteome, while the most similar proteomics maps are those of PFOA and PFDA. However, such comparisons may obscure details of how each agent affects individual protein types, and overall similarity does not imply that the two chemicals have necessarily similar effects on all proteins. PFOA makes little change on the abundance of protein 5, while PFDA drastically reduces the abundance of protein 5 in liver cells.

For the dose-response curves for LY1711883 peroxisome proliferator, Anderson et al.[75] reported proteomics maps for six different concentrations. Using their data, Randić and Estrada[16] selected 99 protein spots for which they measured the difference of the abundance of individual proteins from the abundance in the control group. Figure 19 shows calculated differences for the six concentrations reported, which include the values

$$c = 0.003;\ c = 0.01;\ c = 0.03;\ c = 0.1;\ c = 0.3;\ \text{and } c = 0.6.$$

In this analysis no information on $x$, $y$ coordinates were used, and thus the analysis pertains to cell proteome, and not to proteomics maps. Figure 19 shows that variations of protein abundance for the 99 protein spots vary chaotically, but going from the smallest concentration ($c = 0.003$) toward higher concentrations, initially the perturbations decrease, and only at higher concentrations ($c = 0.3$ and $c = 0.6$) do they start to increase significantly. This qualitative observation can easily be characterized numerically by calculating the total degree of dispersion with respect to the unperturbed proteome of the control group, which gives six concentrations, respectively:

$$s = 51.997;\ s = 40.481;\ s = 28.952;\ s = 41.152;\ s = 75.408;\ \text{and } s = 94.177.$$

When $s$ is plotted against the concentration ($c$), a J-shape curve is obtained, typical of hormesis, illustrated in Figure 20. We would appreciate feedback from readers on the significance of observing hormesis at the cellular level.

**Table 7.** Leading eigenvector for the 20 most intensive protein spots of rat liver cells of the normal cells and cells exposed to four chemicals.

| Protein | Control | PFOA | PFDA | Clofibrate | DEHP |
|---|---|---|---|---|---|
| 1 | 0.5314 | 0.3195 | 0.2374 | 0.5065 | 0.5287 |
| 2 | 0.3059 | 0.1973 | 0.1853 | 0.2823 | 0.2471 |
| 3 | 0.2157 | 0.0608 | 0.0447 | 0.1918 | 0.0417 |
| 4 | 0.2134 | 0.1052 | 0.0573 | 0.1073 | 0.1358 |
| 5 | 0.1629 | 0.0431 | 0.0138 | 0.0800 | 0.1067 |
| 6 | 0.3039 | 0.7737 | 0.8699 | 0.5646 | 0.5716 |
| 7 | 0.2404 | 0.0675 | 0.0396 | 0.1429 | 0.1119 |
| 8 | 0.0231 | 0.0041 | 0.0033 | 0.0080 | 0.0109 |
| 9 | 0.2501 | 0.0919 | 0.0524 | 0.1701 | 0.1443 |
| 10 | 0.0919 | 0.0207 | 0.0081 | 0.0404 | 0.0507 |
| 11 | 0.1284 | 0.2663 | 0.2012 | 0.1987 | 0.1904 |
| 12 | 0.0192 | 0.0053 | 0.0039 | 0.0121 | 0.0108 |
| 13 | 0.2802 | 0.1915 | 0.1162 | 0.2326 | 0.2490 |
| 14 | 0.1478 | 0.1887 | 0.1679 | 0.1796 | 0.1681 |
| 15 | 0.1287 | 0.0359 | 0.0258 | 0.0746 | 0.1344 |
| 16 | 0.3021 | 0.2711 | 0.2335 | 0.2935 | 0.3162 |
| 17 | 0.0896 | 0.0302 | 0.0237 | 0.0514 | 0.0479 |
| 18 | 0.1026 | 0.0333 | 0.0124 | 0.0438 | 0.0558 |
| 19 | 0.1282 | 0.0951 | 0.0659 | 0.1042 | 0.0907 |
| 20 | 0.5314 | 0.0360 | 0.0165 | 0.0785 | 0.0823 |

## Canonical Labels for Maps

Ending this section on proteome and proteomics maps is a brief outline of the most recent work in this area, which considers the search for canonical labels for proteomics maps, and in general any "spot-like" 2D maps. In the case of graphs, canonical labels are important for at least two reasons:

1. They can solve the problem of graph isomorphism, that is, facilitate the recognition of identical graphs that may be presented in different geometrical forms or with matrices with different labels for vertices; and
2. They can facilitate finding the automorphism of a graph (that is, finding the symmetry property of a graph).

By analogy, canonical labels of proteomics maps (and maps in general) will similarly help in checking if two maps are identical, which may then facilitate the construction of catalogues of maps.

For a number of maps, there may be a "natural" way to assign unique labels to spots in a map. For example, with the chaos game representation of DNA, spots can simply assume their sequential position in the DNA sequence as their label, like the map shown in Figure 21, which shows the chaos game representation of the first exon of the human $\beta$-globin gene, according to the algorithm proposed by Jeffrey.[29] Table 8

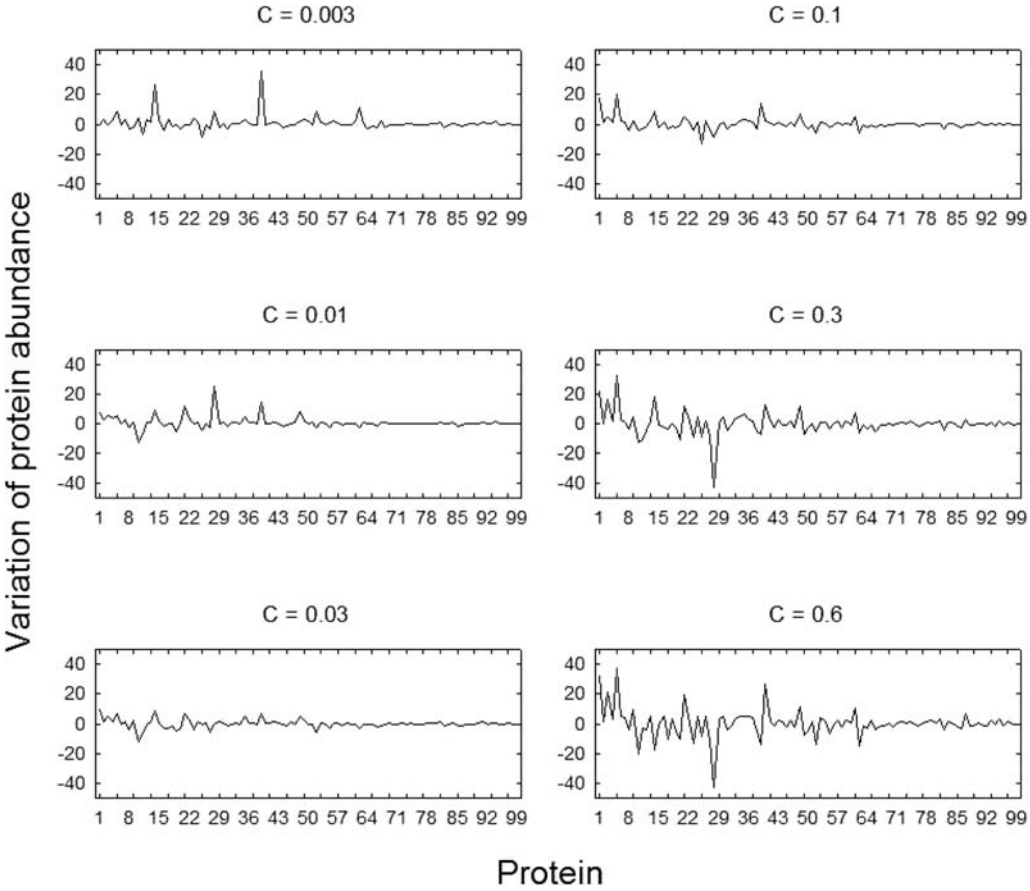


**Figure 19.** Variations in abundance of 99 protein spots with variations in dose concentration of LY171883. Reproduced with permission from Ref. [1].

shows the coordinates of the first dozen nucleotides, which are cumulative coordinates based on:

$$A=(-1,-1);\ C=(-1,1);\quad G=(1,1);\ \text{and}\ T=(1,-1).$$

The same algorithm does not apply to maps that have spots in general positions, like the map shown in Figure 22, which is based on 20 points that have random coordinates. However, the "spirit" of this algorithm can be applied to assign labels to random points by searching for spots closest to the position where the random game assigns locations for spots.

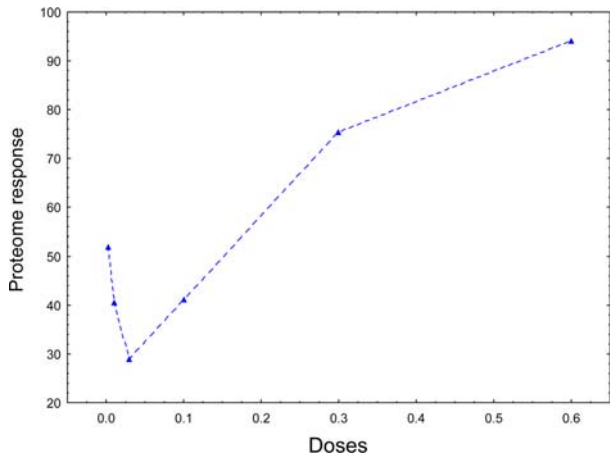


**Figure 20.** The *J*-shaped dose response showing hormesis at the cellular level. Reproduced with permission from Ref. [1].

## Modified Labeling Algorithm for General Maps

A modified approach of chaos game labeling DNA maps for general maps is illustrated on an arbitrary one of 20 vertices in Figure 22. The map of Figure 22 is obtained by selecting coordinates (*x*, *y*) at random (using a random number generator) in the domain (1, 100) and excluding repetitive numbers. Excluding the repetition of random numbers is not essential unless they produce (*x*, *y*) coordinates that have already been selected. Equally, it is not essential that coordinates be integers, but using integers between 1 and 100 makes the illustration simpler. Table 9 shows the selected random (*x*, *y*) coordinates for the 20 unlabeled spots of Figure 2. Let us label the four corners of the 100 × 100 units square by labels A, B, C, and D (which have the coordinates: (0, 0); (0, 100); (100, 100); and (0, 100), respectively). The following canonical rule is adopted for labeling map spots:

Label 1 is assigned to the vertex nearest to the center of one of the four rays from the center of the square to the four corners A, C, G, and T. Let us assume that there is only one such point, which is given label 1. The next point, given label 2, is the point nearest to the center of one of the four rays from the point 1 to the four corners A, B, C, and D. Let us again assume that there is only one such point. The process continues. The next point, given label 3, is the point nearest to the center of one of the four rays from the point 2 to the four corners A, B, C, and D, and so on.

In the illustration that introduces the canonical labels, it is assumed that there is no case of two spots at the same distance from the centers of one of the rays in any step of this process. Should more than one point occur at the same distance from the centers of one of rays, the point having the smaller *x* coordinate is selected. If two (or more) points have

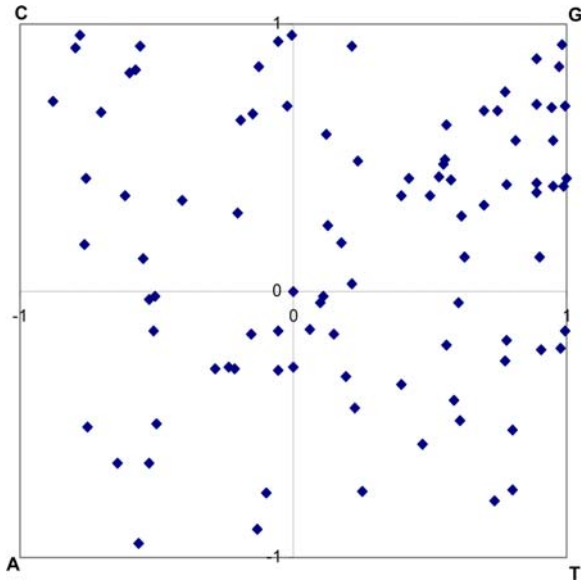


**Figure 21**. The chaos game representation of the first exon of human $\beta$-globgin gene (92 nucloeotides). This rerpesentation allows one to recover the DNA sequence A T G G T G C A C C T … by reversing the constructioin and thus assigns labels 1-92 to all nucleotides of the Figure 4 (top). Reproduced with permission from Ref. [1].

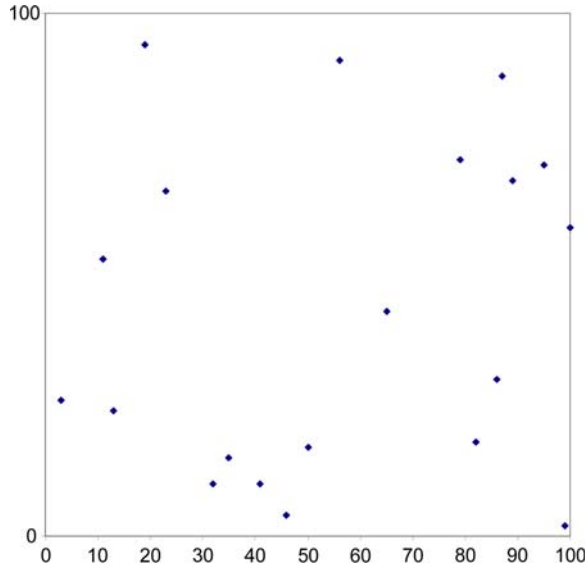


**Figure 22**. Map having 20 unlabelled vertices at random positions.

the same *x* coordinate, the point having the smaller *y* coordinate is selected.

Table 10 illustrates the search for the spot of Figure 2 to be given the canonical label 1. The entries in Table 10 are the distances of all 20 spots of Figure 2 from the centers of the four rays from the origin to the four corners A, B, C, and D, respectively. The first point in Table 10 with coordinates (3, 26) is at distance 22.02 from the point (25, 25), which is the center of the ray from the center of the square to the corner A. The next entry in the first row of Table 10 is 72.01, which is the distance of the point (3, 26) from the point (75, 25), which is the center of the ray from the center of the square to the corner B; the next entry in the first row of Table 10 is 87.09, which is the distance of the point (3, 26) from the point (75, 75), which is the center of the ray from the center of the square to the corner C; and the last entry in the first row of Table 10 is 53.71, which is the distance of the point (3, 26) from the point (25, 75), which is the center of the ray from the center of the square to the corner D. The point (3, 26) in the first quadrant (A) is clearly closer to the center of ray A than any other point from the centers of the remaining three rays. However, we are interested in all 20 points and want to find the point nearest to any four centers of available rays. The smallest entry in Table 10, shown in bold, is in row 14 and column C, signifying that the spot having coordinates (79, 72) and currently having (an arbitrary) label 14 should have the canonical label 1.

In the next step, distances of the remaining 19 points are calculated from the mid points of the four rays from the point having coordinates (79, 72) and the four corners A, B, C, or D. Table 11 shows the critical distances to the four corners of the

**Table 8.** Chaos Game soordinates of the first dozen nucleotides of the first exon of the human $\beta$-globine gene.

| | 0 | 0 |
|---|---|---|
| A | −0.5 | −0.5 |
| T | 0.25 | −0.75 |
| G | 0.625 | 0.125 |
| G | 0.8125 | 0.5625 |
| T | 0.90625 | −0.21875 |
| G | 0.953125 | 0.390625 |
| C | −0.02344 | 0.695313 |
| A | −0.51172 | −0.15234 |
| C | −0.75586 | 0.423828 |
| C | −0.87793 | 0.711914 |
| T | 0.061035 | −0.14404 |
| G | 0.530518 | 0.427979 |
| … | … | … |

**Table 9.** The random (*x*, *y*) coordinates for 20 points of Figure 18.

| | *x* | *y* |
|---|---|---|
| 1 | 3 | 26 |
| 2 | 19 | 94 |
| 3 | 23 | 66 |
| 4 | 86 | 30 |
| 5 | 13 | 24 |
| 6 | 100 | 59 |
| 7 | 95 | 71 |
| 8 | 65 | 43 |
| 9 | 89 | 68 |
| 10 | 46 | 4 |
| 11 | 50 | 17 |
| 12 | 35 | 15 |
| 13 | 56 | 91 |
| 14 | 79 | 72 |
| 15 | 82 | 18 |
| 16 | 41 | 1 |
| 17 | 11 | 53 |
| 18 | 99 | 2 |
| 19 | 32 | 10 |
| 20 | 87 | 88 |

**Table 10.** Distances of the 20 spots of Figure 18 from the centers of the rays between the center of the square and the four corners.

| | A | B | C | D |
|---|---|---|---|---|
| 1 | 22.02 | 72.01 | 87.09 | 53.71 |
| 2 | 69.26 | 88.87 | 59.14 | 19.92 |
| 3 | 41.05 | 66.22 | 52.77 | 9.22 |
| 4 | 61.20 | 12.08 | 46.32 | 75.80 |
| 5 | 12.04 | 62.01 | 80.28 | 52.39 |
| 6 | 82.35 | 42.20 | 29.68 | 76.69 |
| 7 | 83.76 | 50.16 | 20.40 | 70.11 |
| 8 | 43.86 | 20.59 | 33.52 | 51.22 |
| 9 | 77.10 | 45.22 | 15.65 | 64.38 |
| 10 | 29.70 | 35.81 | 76.69 | 74.04 |
| 11 | 26.25 | 26.25 | 63.16 | 63.16 |
| 12 | 14.14 | 41.23 | 72.11 | 60.83 |
| 13 | 72.92 | 68.68 | 24.84 | 34.89 |
| 14 | 71.59 | 47.17 | **5.00** | 54.08 |
| 15 | 57.43 | 9.90 | 57.43 | 80.61 |
| 16 | 28.84 | 41.62 | 81.44 | 75.71 |
| 17 | 31.30 | 69.86 | 67.68 | 26.08 |
| 18 | 77.49 | 33.24 | 76.84 | 103.95 |
| 19 | 16.55 | 45.54 | 77.94 | 65.38 |
| 20 | 88.39 | 64.13 | 17.69 | 63.35 |

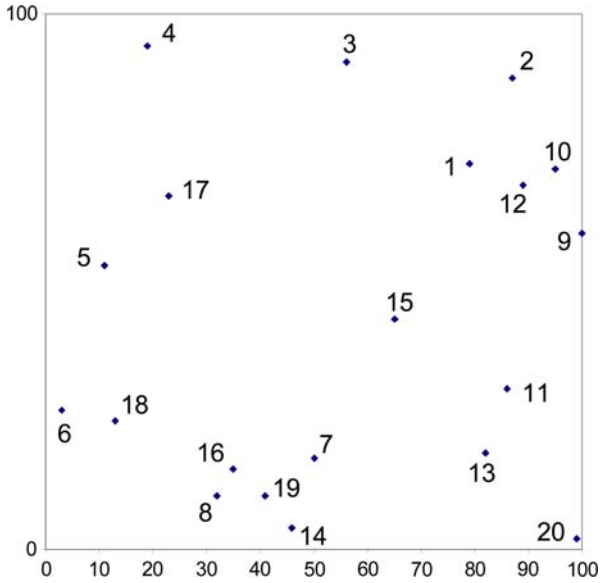


**Figure 23**. Canonical labels for vertices of the map of Figure 22.

square at each step in the search; the shortest distance always determines the spot to which the next canonical label is assigned. Table 11 also lists the canonical labels, the quadrants, the $(x, y)$ coordinates for spots of the map, and initial labels. The new, canonical labels for the map of Figure 22 are also illustrated in Figure 23, which gives the solution to the problem of unique canonical labeling of unlabeled quadratic maps.

## Characterization of Maps Based on Canonical Labels

Once the canonical labels for vertices of a map are found, the characterization of the map can be considered by invariants of sparse matrices to be associated with the map. A way to arrive at a sparse matrix for a map is to consider geometrical objects that overlap the map. We illustrate (1) the construction of the partial ordering graph of the map vertices[105,106]; and (2) the construction of the graph of sequential nearest neighbors for the vertices of the map.[107] In both cases, maps can be represented by a sparse binary matrix.

**Table 11.** The canonical labels, the quadrants, the minimal distances, the $(x, y)$ coordinates, the old labels and for spots of the map.

| Canonical labels | Quadrant | Critical distance | Coordinates | Old label |
|---|---|---|---|---|
| 1 | C | 5.00 | (79, 72) | 14 |
| 2 | C | 3.20 | (87, 88) | 20 |
| 3 | D | 12.85 | (56, 91) | 13 |
| 4 | D | 9.12 | (19, 94) | 2 |
| 5 | A | 6.18 | (11, 53) | 17 |
| 6 | A | 2.55 | (3, 26) | 1 |
| 7 | B | 4.27 | (50, 17) | 11 |
| 8 | A | 7.16 | (32, 10) | 19 |
| 9 | C | 11.70 | (100, 59) | 6 |
| 10 | C | 9.86 | (95, 71) | 7 |
| 11 | B | 12.75 | (86, 30) | 4 |
| 12 | C | 5.00 | (89, 68) | 9 |
| 13 | B | 20.30 | (82, 18) | 15 |
| 14 | A | 7.07 | (46, 4) | 10 |
| 15 | C | 12.04 | (65, 43) | 8 |
| 16 | A | 6.98 | (35, 15) | 12 |
| 17 | D | 10.12 | (23, 66) | 3 |
| 18 | A | 9.12 | (13, 24) | 5 |
| 19 | B | 19.01 | (41, 10) | 16 |
| 20 | C | 21.27 | (99, 2) | 18 |

### Graph of partial ordering of vertices of a map

Figure 24 shows the diagram of partial ordering of the 20 vertices of Figure 23, based on domination of coordinates $(x, y)$ for all pairs of vertices. Let vertex $i$ have coordinates $(x_i, y_i)$ and vertex $j$ have coordinates $(x_j, y_j)$. If $x_i \geq x_j$ and $y_i \geq y_j$ then vertex $i$ is said to dominate vertex $j$. If the two vertices are connected by a line, then the line has a positive slope, because both $x_i = x_j$ and $y_i = y_j$ cannot occur simultaneously. If vertex $j$ dominates vertex $k$, then vertices $j$ and $k$ are similarly connected with a line, but vertex $i$ is not connected to vertex $k$ because dominance is implied by $i \rightarrow j \rightarrow k$ dominance. If the inequalities $x_i \geq x_j$ and $y_i \geq y_j$ are not satisfied, the corresponding vertices are referred to as noncomparable and are left unconnected. Once the partial ordering diagram is constructed, its adjacency matrix can be constructed and used to generate a set of graph invariants.

### Graph of sequential nearest neighbors for the vertices of the map

Figure 25 graphs sequential nearest neighbors for the 20 vertices of the map of Figure 23. This graph is constructed by first connecting vertices 1 and 2. Vertex 3 is then connected to either 1 or 2, depending on which of the two already connected

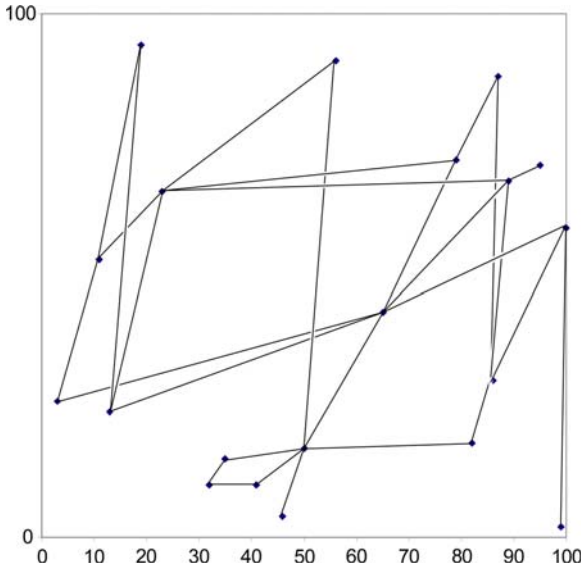


**Figure 24**. The graph of partial ordering of vertices of the map of Figure 22. Observe that slopes of all connecting lines are positive (as they should be).

vertices are closer to vertex 3. If both vertices are at the same distance, then vertex 3 is connected, by convention, to the vertex having the smaller label. Connection of vertices is continued by connecting vertex 4 to the nearest vertex of those already considered. When all vertices are connected, the process ends and the result is an acyclic graph superimposed over the map of Figure 23, as illustrated in Figure 25. For the map of Figure 23, only vertex 19 is at the same distance from the vertices 14 and 16. By following our convention, we connected vertex 19 to vertex 14. Again, the adjacency matrix of this graph can be constructed, and from it sets of invariants can be constructed to serve as map descriptors.

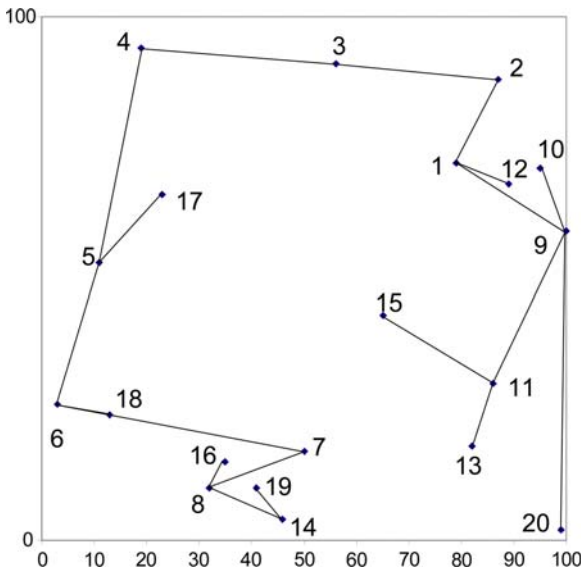


**Figure 25**. The graph of sequential nearest neighbors for the map of Figure 22.

Having an acyclic graph superimposed over a map allows an elementary binary code for a map to be constructed. Such code need not be unique to a map, because two maps may produce the same acyclic graph; but it is unique to the graph and appears to have high discriminatory power. The code to be presented is based on a significant modification the »walk around« code for graphs introduced in graph theory by R. C. Read.[107] To arrive at the »walk around« binary code for trees, each edge of a graph is assigned labels 0 or 1 as follows: One draws a graph on a paper, starts a »walk around« the graph (a tree) at any vertex, and moves clockwise (or counter-clockwise) around the graph assigning label 0 to any edge that is passed for the first time. When the same edge is viewed again from the other side, that is, passed for the second time, it is assigned label 1. Because one can start at an arbitrary edge and circle in either direction around the graph, the resulting code is not unique.

In this case, graph vertices already have labels, which allow one to start at vertex 1. We will not »walk around« the graph, but »walk above« the graph. As we arrive at any branching vertex, by convention we select to move in the direction that leads to the next vertex having the smallest available label. The resulting binary code is unique. Now the binary code for the graph of Figure 5 can be constructed. We start with vertex 1; move toward vertices 2, 3, and 4; and arrive at vertex 5, which is a branching vertex. To the four edges (1, 2); (2, 3); (3, 4); and (4, 5) are assigned labels 0; thus the code starts with 0 0 0 0. At the branching vertex 5, according to our rule, we move toward the vertex having the smaller label, which is vertex 6, which is also a branching vertex. Following our rule, we continue to vertex 7, and follow to vertex 8. Here again is branching, and we move to vertex 14 (having the smaller label) and end with vertex 19. In this way, we passed above the additional five edges: (5, 6); (6, 7); (7, 8); (8, 14); and (14, 19), adding five more zeros to our code, the beginning of which is now 0 0 0 0 0 0 0 0 0. The vertex 19 is the end of travel so far, and we have to go back toward vertex 14 and 8. We assign to edges (19, 14) and (14, 8) labels 1, because we passed these edges before. This continues the code: 0 0 0 0 0 0 0 0 0 1 1. Returning to vertex 8, we first go to vertex 16, because the edge (8, 16) has not yet been visited, rather than returning to vertex 7, because edge (8, 7) has already been visited. Edges that have not yet obtained label 0 have precedent over edges that already have binary assignment 0. Our code thus continues with 0 for edge (8, 16), then 1 for edge (16, 8) and 1 for edge (8, 7), giving this point: 0 0 0 0 0 0 0 0 0 1 1 0 1 1. With this introductory information, one can complete the code which in its entirety, which is:

0 0 0 0 0 0 0 0 0 1 1 0 1 1 1 0 1 1 0 1 1 1 1 1 1 0 0 1 0 0 0 1 1 0 1 1 1 0 1

The code has 38 binary characters, twice the number of edges of the graph. To an edge ($i$, $j$) is assigned the label 0 if $i < j$ and the label 1 if $i > j$.

## Dual of the Map

The »walk above the graph« code (just as the »walk around the graph« code of R. C. Read) allows a dual of the graph embedded on the map to be constructed, referred to as the dual of the map. To obtain this dual, one follows the algorithm outlined in the book of Rouse Ball,[108] starting by replacing the binary entries 0 and 1 of the code by the left and the right brackets. For the above »walk above the graph« code, this gives:

( ( ( ( ( ( ( ( ) ( ) ) ) ( ) ) ( ) ) )
) ) ( ( ) ( ( ( ) ) ( ) ) ) ( )

In the following step, the adjacent left and right brackets are connected, forming circles:

( ( ( ( ( ( ( ( ○ ) ○ ) ) ○ ) ○ ) ) ) )
( ○ ( ( ○ ) ○ ) ) ○

There are eight such instances, which correspond to the eight terminal vertices of the graph (19, 16, 18, 17, 10, 20, 15, and 12). In continuation, the constructed circles are ignored (it is pretended that they have been erased), and newly formed adjacent left and right brackets are connected. The process continues till all left brackets are connected to corresponding right brackets, which results in the map dual shown in Figure 26.

From the dual of Figure 26, the acyclic graph on which it has been based and the labels of the vertices can be reconstructed. To obtain the graph, a vertex is inserted inside each of the circles or ellipses of the dual graph, adding a vertex to the outside area. By connecting vertices in adjacent areas, the original acyclic graph is obtained. The map dual in Figure 26 consists of three disjointed segments, which correspond to three branches of the starting vertex 1 of the map. Map duals offer a qualitative representation of a map, which may have visual advantages in comparative studies of maps and map classifications.

## Route to the Exact Solution to the Alignment of DNA and Proteins

This discussion of DNA ends with an outline of very recent, probably the most outstanding, accomplishment of graphical bioinformatics, and of bioinformatics: Finding the exact solution to the alignment of DNA. The problem of finding the exact solution to the alignment of proteins was solved first,[41] before finding the exact solution to the alignment of DNA. This solution for proteins was immediately, by suitable modifications, extended to DNA.[42] The two approaches have been named Very Efficient Search for Protein Alignment (VESPA) and Very Efficient Search for Nucleotides Alignment (VESNA), respectively. VESPA is the Italian word for wasp and the name of popular and elegant scooter, suggesting elegance in the searching algorithm for protein alignment; VESNA is a common name for ladies and girls in several countries, and it is the Russian word for spring, suggesting the coming of good weather after a long winter (time without having an exact algorithm for DNA alignments).

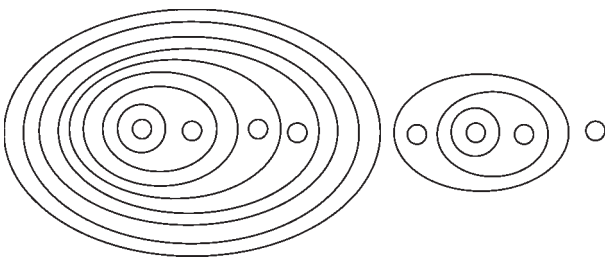

**Figure 26.** Dual of the map based on the sequential nearest neighbor graph of Figure 25.

It is common in science for an unsolved problem to be solved by attempting to solve something else. The problem that led to the exact solution of sequence alignment was the search for the construction of an alternative matrix representation of proteins by sparse matrices, without loss of information. This unique and compact representation can be accomplished by the modification of the amino acid adjacency matrix (AAA matrix), such that instead of using numbers as matrix elements, sets (collections of numbers) are used as matrix entries. After the construction of such a matrix, such matrices also offer an exact solution to the problem of protein alignment.

The exact solution to the problem of alignment of proteins, its development, and its unexpected result are described by exact methods. The perception that this was an unsolvable problem may have been a reason that an exact solution to this problem was not sought, which includes the present author who reported the exact solution of protein alignment and was seeking sparse matrices that can represent proteins. Sparse matrices have many off-diagonal zeros, making computations with them less intensive, which is an important advantage when dealing with a large number of proteins and large proteins. However, the exact alignment of the proteins could have been solved about 5 years earlier because the basic tool, the AAA matrix, was available in 2008.[109]

## AAA Matrix

The AAA matrix is a 20 × 20 matrix with rows and columns belonging to 20 natural amino acids. It has been introduced in search for a uniform representation of proteins. The matrix elements of the AAA matrix count the frequency of occurrence of pairs of amino acids in a protein primary sequence. This is illustrated by the sequence of amino acids of two proteins of *Saccharomyces cerevisiae* selected for outline by the VESPA algorithm: Protein 1 and protein 2 of Table 3, which start with amino acids K I L G I D P N V T Q Y…and P S K L G I D T V K Q W…, respectively. Tables 12 and 13 show the AAA matrices for the two proteins. For better visibility of the nonzero entries in the matrix, the zero entries are not shown, except on the main diagonal. Matrix elements (*i*, *j*) have entries 0, 1, 2, 3, or 4 (in the cases considered), which indicate that the amino acid $A_i$ is followed by amino acid A zero, one, two, three, or four

**Table 12.** Amino acid adjacency matrix for Protein 1.

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 | | | | | | | 3 | | | | 1 | | | | 1 | 1 | | | |
| R | | | 2 | | | | | | | | | | | | | 1 | 1 | | | |
| N | 1 | | | 1 | | | | 2 | | | 1 | 1 | | 2 | 1 | 1 | 1 | | | 2 |
| D | | 1 | | | | 2 | 1 | | | | 1 | 1 | | 1 | 3 | | | | | 2 |
| C | | | | | | | | | | | | | | | | 1 | | | | |
| Q | | | | 1 | | | | | | | | | | 1 | 1 | | | | 1 | |
| E | | | | 2 | | | | | | 1 | 2 | | | | | 2 | | | 1 | |
| G | | | 2 | | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | | 1 | 3 | 1 | | | 1 | 1 |
| H | | | | | | | | | | 1 | | 1 | | 1 | | | | | 1 | |
| I | 1 | | | 1 | | | | 3 | | | 3 | | | 1 | 1 | | | | | |
| L | | | 1 | 2 | | | 1 | 2 | | 1 | | 1 | | 2 | | 1 | 3 | 1 | | |
| K | | | | 3 | | | | 1 | 1 | 1 | | | | | 1 | | | | | |
| M | | | | | | | | | | | | | | | | | | | | |
| F | 1 | | 1 | 1 | | | 2 | | 1 | | 2 | | | 4 | 1 | 1 | | 1 | | |
| P | 1 | | 1 | 1 | | | 1 | 1 | | 1 | | | | | | 1 | | | 1 | 3 |
| S | | 1 | 2 | | | | 1 | 2 | 1 | 1 | 1 | | | | | 3 | | 1 | 2 | 1 |
| T | | | | | | 1 | | 2 | | | | | | 1 | | 1 | 1 | | | 2 |
| W | | | 1 | | | | | | | | 1 | | | | | | 1 | | | |
| Y | 1 | | 1 | | | | | | | 1 | 1 | | | | | 2 | 1 | | | 1 |
| V | 1 | | 2 | | | | 1 | 1 | | 2 | 1 | | | 1 | | 1 | 1 | | 1 | |

times, respectively. In the case of protein 1, the first entry (1, 1) in the first row means that in the protein 1 sequence there is a succession of two alanine. The entry 3 in the same row and column G means that adjacent AG (alanine, guanine) occur three times. Clearly, the matrix is nonsymmetrical, while both the row sums and the column sums of the corresponding amino acids are the same (giving the abundance of individual amino acids), except for the first and last amino acids, which are counted only once.

Obviously, there is loss of information in using the AAA matrix when representing proteins, because locations of individual amino acids in sequences are not known. Nevertheless, AAA matrices have been found useful in the comparative study of proteins and in a study of individual proteins, as has been demonstrated by Roy Choudhury and coworkers,[25,26] who used artificial neural networks and properties of AAA matrices to identify fragments of membrane proteins that are inside a membrane, which we consider as one of the milestones of graphical bioinformatics. Membrane proteins are vital to the survival of organisms because they are involved in a variety of biochemical processes and functions. The active transport of molecules or signals through the biological membranes is one of the most important functions of membrane proteins. An estimated up to 30% of all genes in most genomes encode membrane proteins.

Information about a 3D structure of a membrane transporter is required in the study of the protein and small molecule transport mechanisms, which is important for drug design because membrane proteins are targets of over 50% of modern medicinal drugs. Due to difficulties encountered in

**Table 13.** Amino acid adjacency matrix for Protein 2.

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | | | 1 | 1 | | | | 3 | 1 | | | | | 1 | | 1 | | | 1 | |
| R | | | 1 | | | | | | | | | | | | | 1 | 1 | | | |
| N | 1 | | 2 | 3 | | | | 2 | | | 1 | | | | 2 | | | | | |
| D | 1 | | | | | | 1 | | | | | | 1 | 1 | 2 | 1 | 1 | | 1 | |
| C | | | | | | | | | | | | | | | | 1 | | | | |
| Q | | | | | | | | | | 1 | | | | | 1 | | | 1 | | |
| E | 1 | 1 | | | | 1 | | | | 1 | 2 | 1 | | 1 | | 2 | | | | |
| G | 1 | | 1 | 1 | 1 | | 1 | 1 | 1 | 2 | 1 | 1 | | 1 | 2 | | | | 1 | 1 |
| H | | | 1 | | | | 1 | | | 1 | 1 | | | 1 | | | | | 1 | |
| I | 2 | | | 1 | | | | 2 | 1 | 1 | 1 | | | 2 | 1 | | 1 | | | 1 |
| L | 1 | 1 | 1 | | | | 3 | 3 | | | 1 | | | 1 | | | 1 | 1 | | |
| K | | | 1 | 2 | | 1 | | | 1 | | 2 | | | | 1 | | | | | 1 |
| M | | | | 1 | | | | | | 2 | | 1 | | | | | | | | |
| F | | | 1 | | | | 3 | | 1 | | 2 | | | 2 | 1 | 1 | 1 | | 1 | |
| P | 1 | | | | | 1 | 1 | 1 | 1 | 2 | 1 | | | | | 2 | | | 1 | |
| S | | 1 | 1 | | | | | 1 | | 1 | | 2 | 1 | | 1 | 3 | 1 | 1 | 2 | 1 |
| T | | | | | | | | 1 | | | | 1 | | 1 | | 1 | | | | 1 |
| W | | | 1 | | | | | | | | 1 | | | | | 1 | | | | |
| Y | 1 | | | | | | | 1 | | 2 | | 1 | 1 | | | 1 | | 1 | | |
| V | | | | | | | | 1 | | | | 2 | 1 | | | 1 | | | | 1 |

biomembrane research, only a limited number of membrane transport proteins have been solved experimentally for their 3D structure. The tertiary structures of a large number of membrane proteins are still unresolved; however, *in silico* methods may fill the information gap and offer a possibility to hypothesize the transport mechanisms. The topology of an integral membrane protein, which describes the number of transmembrane segments and the orientation in the membrane, may be predicted using computation methods based solely on the AA sequence information, as demonstrated in [25] using AAA matrices.

## Sequential AAA Matrix

Although AAA matrices are accompanied by a loss of information on the location of individual pairs of amino acid within the primary sequence, they carry significant information on individual protein sequences. For example, Table 12, which is AAA of protein 1, shows the occurrence of repeated entries. Thus AA, GG, and TT appear once, while SS appears three times, and FF appears four times. Without further inspection of the primary sequence, it is uncertain if SS appears twice or SSS appears once, but it is easy to see that the former is the case. In the case of four FF occurrences, there could be four FF, two FF and a single FFF, two FFF, or a single FFFFF, and it is easy to see that there are two FF and a single FFF. If row sums and column sums are computed, an abundance of individual amino acids are obtained, which in the case of protein 1 give the following:

$$\text{Row sum}: \quad (7,\ 2,\ 13,\ 12,\ 1,\ 4,\ 8,\ 18,\ 4,\ 10,\ 15,\ 7,\ 0,\ 15,\ 11,\ 16,\ 8,\ 3,\ 8,\ 12)$$

$$\text{Column sum}: \quad (7,\ 2,\ 13,\ 12,\ 1,\ 4,\ 8,\ 18,\ 4,\ 10,\ 15,\ 6,\ 0,\ 15,\ 11,\ 16,\ 9,\ 3,\ 8,\ 12)$$

A comparison of the row sum and the column sum shows that this protein starts with K and ends with T. The row sums are 7 and 8 for K and T, and column sums are 6 and 9 for K and T, respectively.

An overlay of Tables 12 and 13 shows about 50 pairs of adjacent amino acids, which appear in protein 1 and do not appear in protein 2, and vice versa. This observation significantly increases efficiency in the search for protein alignment, because such pairs of amino acids can be ignored in a search. For example, the pairs AA, AK, and AT appear in protein 1 but do not occur in protein 2, and pairs AN, AD, AH, AF, and AY appear in protein 2 but do not appear in protein 1. This leaves only three pairs of AG and a single occurrence of AS as possible pair components in an aligned fragment of the two proteins. Table 14 shows the AAA matrix obtained by superposition of AAA matrices of protein 1 and protein 2 after eliminating the pairs, which are unique for either protein. The symbol *x* indicates that those matrix elements in two matrices do not match in number. For example, the *x* for the RN element in the AAA superposition matrix arises because in protein 1 there are two adjacent RN pairs, but in protein 2 there is only one, and without further examination it is unknown which pair is matched, if any. Similarly, the *x* for the ND element in the AAA superposition matrix arises because in protein 1 there is one RN pair, but in protein 2 there are three, and again without further examination it is unknown which of the three pairs or RN in protein 2 is matched RN of protein 1, if any.

Finding the exact solution to the problem of protein alignment is just one step away, which consists of inserting the sequential numbers of amino acids as matrix elements instead of just recording their frequency of occurrence. All that needs to be done is to construct the sequential AA matrices for proteins and combine them to extract common neighborhoods. In the sequential AA adjacency matrix, the matrix elements do not

**Table 14.** Superposition of amino acid adjacency matrices after unique pairs of adjacent amino acids have been eliminated. Symbol x indicates that those matrix elements in two matrices do not match in number.

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | | | | | | | | 3 | | | | | | | | 1 | | | | |
| R | | | x | | | | | | | | | | | | | | | | | |
| N | 1 | | | x | | | | 2 | | | 1 | | | | x | | | | | |
| D | | | | | | | 1 | | | | | | | 1 | x | | | | | |
| C | | | | | | | | | | | | | | | | 1 | | | | |
| Q | | | | | | | | | | | | | | | 1 | | | | | |
| E | | | | | | | | | | 1 | 2 | | | | | 2 | | | | |
| G | | | x | | 1 | | 1 | 1 | 1 | x | x | 1 | | 1 | x | | | | 1 | 1 |
| H | | | | | | | | | | 1 | | | | 1 | | | | | 1 | |
| I | x | | | 1 | | | | x | | | x | | | x | 1 | | | | | |
| L | | | 1 | | | | x | x | | | | | | x | | | x | 1 | | |
| K | | | | x | | | | | 1 | | | | | | 1 | | | | | |
| M | | | | | | | | | | | | | | | | | | | | |
| F | | | | | | | x | | 1 | | 2 | | | x | 1 | 1 | | | | |
| P | 1 | | | | | | 1 | 1 | | x | | | | | | x | | | 1 | |
| S | | 1 | x | | | | | x | | 1 | | | | | | 3 | | 1 | 2 | 1 |
| T | | | | | | | | x | | | | | | 1 | | 1 | | | | x |
| W | | | 1 | | | | | | | | I | | | | | | | | | |
| Y | 1 | | | | | | | | | x | | | | | | x | | | | |
| V | | | | | | | | 1 | | | | | | | | 1 | | | | |

**Table 15.** The initial 20 entries of the Sequential AA adjacency matrix of protein 1.

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | | | | | | | | | | | | | | | | | | | | |
| R | | | | | | | | | | | | | | | | | | | | |
| N | | | | | | | | | | | | | | | | | | | | 8 |
| D | | | | | | | 20 | | | | | | | | 6 | | | | | 17 |
| C | | | | | | | | | | | | | | | | | | | | |
| Q | | | | | | | | | | | | | | | | | | | 11 | |
| E | | | | 19 | | | | | | | | | | | | | | | | |
| G | | | | | | | | | | 4 | | | | | | | | | 14 | |
| H | | | | | | | | | | | | | | | | | | | | |
| I | | | | 5 | | | | | | | 2 | | | | | | | | | |
| L | | | | 16 | | | | 3 | | | | | | | | | | | | |
| K | | | | | | | | | | 1 | | | | | | | | | | |
| M | | | | | | | | | | | | | | | | | | | | |
| F | | | | | | | | | | | | | | | | | | | | |
| P | | | 7 | | | | | | | | | | | | | | | | | |
| S | | | | | | | | | | | | | | | | | | | | |
| T | | | | | | 10 | | 13 | | | | | | | | | | | | |
| W | | | | | | | | | | | | | | | | | | | | |
| Y | | | | | | | | | | | 15 | | | | | | 12 | | | |
| V | | | | | | | 18 | | | | | | | | | | 9 | | | |

count the occurrence of individual pairs of amino acids, but indicate their locations in the primary sequence. Table 15 shows the initial 20 steps in construction of the sequential AA matrix for protein 1. Proceeding to the next entry, the 21st pair of adjacent amino acids, which is again ED, just as was the 19th pair, is added to the present entry of 19. Clearly, the entries of the sequential amino acid matrix besides numbers (individual amino acid sequential labels) can also be sets of numbers. For clarity, instead of writing in the standard matrix form, it may be better to simply list nonzero matrix elements, as shown in Table 16.

## Exact Solution to the Protein Alignment Problem

The exact solution of the protein alignment problems for a pair of proteins has no approximation of any kind. The two proteins of Table 3 are selected for illustration, and their sequential AAA matrices are shown in Tables 16 and 17, respectively. To combine these two sequential AA adjacency matrices, only their common elements are listed. The first column of Tables 16 and 17 shows that, of adjacent pairs starting with A (alanine), only AG and AS are common to both proteins; hence AG and AS are starting amino acid pairs in our list of common AA pairs for two proteins. Table 18 reproduces in its first two entries, one above the other (using the color blue for protein 1 and the color red for protein 2), the sequential labels of alanine–glycine pairs, followed by sequential labels for adjacent alanine–serine amino acids. A continuation with the rest of amino acid pairs present in both proteins results in Table 18.

Table 18 contains an exact solution for alignment of the two proteins to be extracted out, which is shown in Table 19. Table 18 compares sequential labels for the two proteins. If the labels are in the same vicinity/neighborhood, the difference of the corresponding labels can be 0, or $\pm$ a few steps. Table 19 shows the differences $+2$, $+1$, 0, $-1$, $-2$, $-3$, and $-4$, which are the actual differences found in Table 18. The first entry of Table 18 for AG shows the difference of $-2$ [for (110, 108)], and the difference $-4$ [for (138, 134) and (143, 139)]. The sequential neighbors for AS are not in the

**Table 16.** The non-zero matrix elements of the sequential AA matrix for protein 1.

| | | | |
|---|---|---|---|
| AA 109 | ES 31, 140 | LF 57, 120 | SH 156 |
| AG 110, 138, 143 | EY 127 | LS 155 | SI 65 |
| AK 37 | GN 74, 170 | LT 54, 164, 173 | SL 53 |
| AS 151 | GC 50 | LW 43 | SS 52, 64, 101 |
| AT 83 | GQ 132 | KD 38, 112, 158 | SW 78 |
| RN 33, 160 | GE 139 | KG 131 | SY 97, 141 |
| NA 82 | GG 47 | KH 23 | SV 166 |
| ND 34 | GH 144 | KI 1 | TQ 10 |
| NG 46, 171 | GI 4 | KP 71 | TG 13, 55 |
| NL 163 | GL 56, 172 | FA 150 | TF 29 |
| NK 130 | GK 111, 132 | FN 162 | TS 165 |
| NF 116, 161 | GF 95 | FD 122 | TV 84, 107 |
| NP 75 | GP 48, 62, 67 | FE 30, 59 | WN 79 |
| NS 80 | GS 100 | FH 135 | WL 44 |
| NT 106 | GY 14 | FL 87, 117 | WT 28 |
| NV 8, 93 | GV 103 | FF 25, 26, 58, 121 | YA 142 |
| DR 159 | HI 136 | FP 125 | YN 115 |
| DQ 89, 123 | HK 157 | FS 96 | YI 146 |
| DE 20 | HF 24 | FW 27 | YL 15 |
| DL 69 | HY 145 | PA 36 | YS 77, 98 |
| DK 22 | IA 137 | PN 7 | YT 12 |
| DF 134 | ID 5 | PD 68 | YV 128 |
| DP 6, 35, 39 | IG 66, 73, 169 | PE 126 | VA 108 |
| DV 17, 113 | IL 2, 42, 154 | PG 49 | VN 92, 129 |
| CS 51 | IF 86 | PI 72 | VE 18 |
| QD 124, 133 | IP 147 | PS 63 | VG 94 |
| QF | LN 45 | PY 76 | VI 41, 85 |
| QP 90 | LD 16, 88 | PV 40, 81, 148 | VL 167 |
| QY 11 | LE 118 | SR 32 | VF 149 |
| ED 19, 21 | LG 3, 61 | SN 81, 105 | VS 104 |
| EI 153 | LI 168 | SE 152 | VT 9 |
| EL 60, 119 | LK 70 | SG 99, 102 | VY 114 |

**Table 17.** The non-zero matrix elements of the sequential AA matrix for protein 2.

| | | | |
|---|---|---|---|
| AN 36 | ES 30, 136 | LG 4, 60, 91 | SN 128 |
| AD 67 | GA 66 | LL 56 | SG 13 |
| AG 108, 134, 139 | GN 167 | LF 57 | SI 64 |
| AH 147 | GD 98 | LT 161 | SK 2, 21 |
| AF 122 | GC 49 | LW 42 | SM 83 |
| AS 82 | GE 135 | KN 153 | SF 52 |
| AY 112 | GG 46 | KD 19 | SP |
| RN 32 | GH 140 | KQ 10 | SS 51, 63, 103 |
| RS 127 | GI 5, 169 | KG 110 | ST 104 |
| RT 157 | GL 55 | KH 22 | SW 77 |
| NA 81 | GK 109 | KL 106 | SY 96, 137 |
| NN 79, 80 | GF 94 | KP 3, 70 | SV 163 |
| ND 33, 37, 129 | GP 47, 61 | KV 101 | TG 54 |
| NG 45, 168 | GY 14 | MD 16 | TK 105 |
| NL 160 | GV 92 | MI 84, 165 | TF 158 |
| NP 74, 154 | HN 73 | MK 69 | TS 162 |
| DA 111 | HE 148 | FN 160 | TV 8 |
| DE 99 | HI 132 | FE 29, 58, 120 | WN 78 |
| DM 68 | HL 125 | FH 131 | WL 43 |
| DF 130 | HF 23 | FL 86, 115 | WF 27 |
| DP 34, 38 | HY 141 | FF 24, 28 | WS 12 |
| DS 20 | IA 133, 146 | FP 123 | YA 138 |
| DT 7 | ID 6 | FS 95 | YG 97 |
| DY 17 | IG 65, 166 | FT 53 | YI 113, 142 |
| CS 50 | IH 72 | FY 25 | YK 18 |
| QI 145 | II 40 | PA 35 | YM 15 |
| QP 89 | IL 41 | PQ 144 | YS 76 |
| QW 11 | IF 85, 114 | PE 155 | YW 26 |
| EA 121 | IP 143 | PG 48 | VG 93 |
| ER 156 | IT 160 | PH 124 | VK 9, 152 |
| EQ 88 | IV 150 | PI 39, 71 | VM 164 |
| EI 149 | LA 107 | PL 90 | VS 102 |
| EL 59, 117 | LR 126 | PS 1, 62 | VV 151 |
| EK 100 | LN 44 | PY 75 | |
| EF 119 | LE 87, 116, 118 | SR 31 | |

**Table 18.** Common pairs of AA in protein 1 (top) and protein 2 (bottom).

| | | | |
|---|---|---|---|
| AG 110, 138, 143 | GN 74, 170 | ID 5 | FF 25, 26, 58, 121 |
| AG 108, 134, 139 | GN 167 | ID 6 | FF 24, 28 |
| AS 151 | GC 50 | IG 66, 73, 169 | FP 125 |
| AS 82 | GC 49 | IG 65, 166 | FP 123 |
| RN 33, 160 | GE 139 | IL 2, 42, 154 | FS 96 |
| RN 32 | GE 135 | IL 41 | FS 95 |
| NA 82 | GG 47 | IF 86 | PA 36 |
| NA 81 | GG 46 | IF 85, 114 | PA 35 |
| ND 34 | GH 144 | IP 147 | PE 126 |
| ND 33, 37, 129 | GH 140 | IP 143 | PE 155 |
| NG 46, 171 | GI 4 | LN 45 | PG 49 |
| NG 45, 168 | GI 5, 169 | LN 44 | PG 48 |
| NL 163 | GL 56, 172 | LE 118 | PI 72 |
| NL 160 | GL 55 | LE 86, 116, 118 | PI 39, 71 |
| NP 75 | GK 111, 132 | LG 3, 61 | PS 63 |
| NP 74, 154 | GK 109 | LG 4, 60, 91 | PS 1, 62 |
| DE 20 | GF 95 | LF 57, 120 | SR 32 |
| DE 99 | GF 94 | LF 57 | SR 31 |
| DF 134 | GP 48, 62, 67 | LT 54, 164, 173 | SN 81, 105 |
| DF 130 | GP 47, 61 | LT 161 | SN 128 |
| DP 6, 35, 39 | GY 14 | KD 38, 112, 158 | SG 99, 102 |
| DP 34, 38 | GY 14 | KD 19 | SG 13 |
| CS 51 | GV 103 | KH 23 | SI 65 |
| CS 50 | GV 92 | KH 22 | SI 64 |
| QP 90 | HI 136 | KP 71 | SS 52, 64, 101 |
| QP 89 | HI 132 | KP 3, 70 | SS 51, 63, 103 |
| EI 153 | HF 24 | FE 30, 59 | SW 78 |
| EI 149 | HF 23 | FE 29, 58, 120 | SW 77 |
| EL 60, 119 | HY 145 | FH 135 | SY 97, 141 |
| EL 59, 117 | HY 141 | FH 131 | SY 96, 137 |
| ES 31, 140 | IA 137 | FL 87, 117 | SV 166 |
| ES 30, 136 | IA 133, 146 | FL 86, 115 | SV 163 |
| TG 13, 55 | TV 84, 107 | YA 142 | VG 94 |
| TG 54 | TV 8 | YA 138 | VG 93 |
| TF 29 | WN 79 | YI 146 | VS 104 |
| TF 158 | WN 78 | YI 113, 142 | VS 102 |
| TS 165 | WL 44 | YS 77, 98 | |
| TS 162 | WL 43 | YS 76 | |

neighborhood (151 and 82) and are ignored. The next cell is (33, 32) for RN with the difference of −1, while RN at position 160 in protein 1 is ignored with nothing to match. Continuing this process ends with Table 19, which leads to the solution of the alignment of protein 1 and protein 2.

Table 19 suggests that amino acid pairs having differences of +2 and 0 can be ignored, as they represent individual (chance) alignments at great separations. Thus there are a short segment with the difference of −1, a sizable segment

**Table 19.** Aligned segments of protein 1 and protein 2.

Difference +2
(26, 28), (101, 103)
Difference +1
(3, 4), (4, 5), (5, 6)
Difference 0
(14, 14), (57, 57)
Difference −1
(23, 22), (24, 23), (25, 24), (30, 29), (31, 30), (32, 31), (33, 32), (34, 33), (35, 34), (36, 35), (39, 38), (42, 41), (44, 43), (45, 44), (46, 45), (47, 46), (48, 47), (49, 48), (50, 49), (51, 50), (52, 51), (55, 54), (56, 55), (59, 58), (60, 59), (61, 60), (62, 61), (63, 62), (64, 63), (65, 64), (66, 65), (71, 70), (72, 71), (75, 74), (77, 76), (78, 77), (79, 78), (82, 81), (86, 85), (87, 86), (90, 89), (94, 93), (95, 94), (96, 95), (97, 96)
Difference −2
(104, 102), (110, 108), (111, 109), (117, 115), (118, 116), (119, 117), (125, 123)
Difference −3
(163, 160), (164, 161), (165, 162), (166, 163), (169, 166), (170, 167), (171, 168)
Difference −4
(134, 130), (135, 131), (136, 132), (137, 133), (138, 134), (139, 135), (140, 136), (141, 137), (142, 138), (143, 139), (144, 140), (145, 141), (146, 142), (147, 143), (153, 149)

around (23, 96), two shorter segments with differences −2 and −3 around (110, 125) and (163, 171), respectively, and an additional intermediate length segment with a difference of −4 around (134, 153).

## Comment on the Exact Solution of the Protein Alignment Problem

There are a number of famous problems in mathematics, described informally as problems that everyone (even nonprofessional mathematicians) can understand but apparently nobody can solve. Many of these problems remain unsolved for a long time. They include the problems listed in Table 20, but the complete list is longer. The history of solving some of these problems can be followed in the literature (e.g., Ref. [110]). The problem of the exact solution to the protein alignment problem is not as famous, but it shares some common features with famous problems in mathematics. For example, it has not been known whether a rigorous solution exists at all for the problem. The problem can be understood by all, or at least by undergraduate students of chemistry and biology. The problem has also existed for more than 40 years. Similarly, there may be additional famous problems in chemistry that remain unsolved for a long time, even though they do not receive as much publicity in chemistry as do famous problems of mathematics among mathematicians and the general public. For example, the problem of the four center molecular integrals over Slater-type functions may be one such famous problem of chemistry, because it has existed for well over 50 years, it is well defined, and there is no proof that it cannot be solved. The current mathematical tool for solving such problems may not be adequate. In situations when the current tool shows limitations, development of new tool, if possible, could help solve the problem.

This is precisely what happened with finding the "Exact Solution to the Protein Alignment Problem," which was found not because of a search for the rigorous solution to the problem of protein alignment, but because of a search for a novel tool for the characterization of proteins. The answer was the use of AA adjacency matrices, but instead of counting the frequency of adjacent pairs of amino acids, such information is replaced with sequential labels of corresponding adjacent amino acids. The solution can be obtained by overlapping two such matrices for the two proteins of interest and simply extracting pairs of AA that are in the same neighborhood, as illustrated in analyzing Table 18 and constructing Table 19. Table 19, ignoring entries for differences of +2 and 0, gives the exact solution to the problem.

The following two comments qualify the use of the terms "rigorous" and "exact," and the nature and simplicity of the solution. The manuscript on the rigorous solution of the protein alignment problem was sent to the *Journal of Computational Chemistry*, where it was immediately accepted. However, one of the reviewers failed to recognize that the article reports on an exact solution of the problem. This may be in part because the words "exact solution" were not used in the title, and the title of the article may not have been the best choice for the message. A better title would be "Rigorous Solution to the Protein Alignment Problem" or, even better, "Exact Solution to the Protein Alignment Problem." This became clear when the same referee requested more details on the approximations used; but exact solutions have no approximations. The referee also objected that the word "rigorous" was used for this approach, as if the available computer programs are not rigorous; but computer programs for protein alignment are not rigorous in a strict mathematical sense.

In summary, whatever is known and understood today in bioinformatics and related biology—and that is an amazing amount of novelty and insight with a plethora of highly significant results—is due to the existing available computer-based programs and packages. But technically, particularly with mathematical terms as used by mathematicians and not as used by laypersons, "rigorous" implies a solution that does not use approximations, empirical parameters, statistical methods, and so on. On such grounds, the current existing available computer-based programs and packages do not qualify as rigorous, though they are mathematically well-defined. According to Wikipedia, such programs have been described as "rigorous."

The simplicity of the solution, which can informally be qualified as a solution that everyone (at least undergraduate students) can understand, is interesting. The problem of protein alignment differs visibly from famous problems of mathematics, which are generally easy to understand while the details of their solutions are difficult to understand. In contrast, the solution to the problem of protein alignment is as easy to understand as the problem. However, this does not reflect on those who tried to solve this problem and did not find a solution, but it reflects on the novelty of the tool used (starting with the AA adjacency matrix), which has not been available in the past. Some may refer to the exact solution of the protein alignment problem as so simple that anyone could have found it. That may be true, but it has not been done before! If such

**Table 20.** A selection of famous mathematical problems.

| Problem | Informative description |
|---|---|
| Four Color Conjecture | Any map drawn in a plane can be colored with at most four colors |
| Traveling Salesman Problem | Find the shortest route for a person to travel over given network visiting each place just once |
| Fermat' Last Theorem | Show that equation: $A^n + B^n = C^n$ has no solution in integer A, B and n, except for $n = 2$ |
| Graph Reconstruction | Prove that set of subgraphs in which each vertex is removed separately allows reconstruction of the initial graph |
| Goldbach's Conjecture | Prove that any integer bigger/equal 4 can be expressed as the sum of two prime numbers |
| Trisection of an angle | Design geometrical construction that allows any given angle to be divided in three equal sections |

comments appear, they will be a reminder of the story of an egg of Columbus, which "refers to a brilliant idea or discovery that seems simple or easy after the fact" (It is difficult to find original reference to the well-known story of *Columbus Breaking the Egg*, but in 1752 an engraving is already made by the English artist William Hogarth entitled and depicting *Columbus Breaking the Egg*). Current titles of papers on a rigorous approach to the alignment of proteins and DNA are "Very efficient search for protein alignment" and "Very efficient search for nucleotide alignment." If the word "search" were replaced with "solution," the title would be less confusing for readers who may not recognize the exact solution as the solution of a problem that was unsolved for about half a century.

## Very Efficient Search for Nucleotide Alignment (VESNA)

The novel approach to DNA alignment, VESNA, parallels the approach to the exact solution to protein alignment, VESPA, after an important modification at the start. It is based on the following steps:

1. The construction of 4 × 4 nucleotide adjacency table (Table 21), in which sequential positions of all adjacent pairs of nucleotides in DNA sequential labels for adjacent nucleotides are listed in the corresponding matrix elements.
2. For very long DNA sequences, instead of considering a 4 × 4 matrix (which has only 16 distinct matrix elements), a 16 × 16 matrix can be considered, the matrix elements of which are the 16 pairs of nucleotides of the 4 × 4 matrix (Table 21). This leads to 256 distinct matrix elements, which is comparable to 400 distinct matrix elements of the AAA matrix used in the search for protein alignment.
3. The resulting matrices have set of numbers as elements. Instead of constructing the nucleotide adjacency matrices, the cardinality of the sets forming their elements that may be large, it is often more convenient to just construct the list of matrix elements, even for shorter DNA sequences.
4. Superposition of such matrices, or a list of matrix elements, for two DNA sequences allows the immediate identification of nucleotides in two sequences that differ in sequence locations by the same amount.
5. Grouping of matrix elements that have the same difference in their sequential labels resolves the problem of DNA alignment.

**Table 21.** The 4×4 non-symmetrical nucleotide adjacency matrix.

| | A | C | G | T |
|---|---|---|---|---|
| A | AA | AC | AG | AT |
| C | CA | CC | CG | CT |
| G | GA | GC | GG | GT |
| T | TA | TC | TG | TT |

The last step immediately reveals all segments in two proteins that have the same relative shift, and the differences indicate the number of steps that such segments are shifted. In general, the 4 × 4 nucleotide adjacency tables (or 16 × 16 tables) are nonsymmetrical, except in the special case of palindromic DNA sequences.

The exact solution to the alignment of DNA is illustrated on the α-globin genes (GenBank sequence CHPHBA and RABHBA belonging to the chimpanzee and rabbit, respectively, having just over 110 nucleotides). Their initial 20 nucleotides are shown below.

CHPHBA : GACTCAGAAACCCACCATG . . .

RABHBA : GACTCAGAACCCACCATGG . . .

These are the two proteins considered by Pearson and Lipman in their article on the construction of improved tools for biological sequence comparison.[67] Table 22 alphabetically lists the 16 matrix elements of the 4 × 4 nucleotide adjacency matrix for both proteins, one above the other (blue for protein 1 and red for protein 2). For better visibility of pairs of nucleotides that are aligned, blank spaces are added in-between.

Table 22 shows that the first nucleotide pair GA appears at position 1 and at sites 7, 11, 38, 43, 89, 101, 110, and 112 in protein 1, and at the locations 1, 5, 7, 11, 37, 42, 54, 64, 69, 88, 100, 109, and 113 in the second DNA. The two GA sets of (ordered) labels show that GA appears at the same locations

**Table 22.** The nucleotide sequential adjacency matrices for the two DNA sequences CHPHBA (upper in blue) and RABHBA (lower in red).

| | |
|---|---|
| AA | 8, 9, 12, 41, 47, 53, 68, 90, 102, 111, 113 |
| | 8, 12, 40, 46, 52, 65, 66, 67 |
| AC | 2, 13, 17, 39, 44, 48, 81 |
| | 2, 13, 16, 38, 43, 47, 55, 80 |
| AG | 6, 10, 42, 54, 69 |
| | 6, 9, 41, 53, 68, 76, 89, 101, 110, 112 |
| AT | 20, 93 |
| | 19, 49, 70, 92 |
| CA | 5, 16, 19, 40, 46, 52, 80 |
| | 15, 18, 39, 45, 48, 51, 75, 79, |
| CC | 14, 15, 18, 32, 36, 45, 57, 60, 105, 106 |
| | 14, 17, 31, 32, 44, 59, 78, 98, 104 |
| CG | 37, 49, 58, 73, 76, 78, 82, 88, 99 |
| | 33, 72, 81, 87, 96, 99, 105 |
| CT | 3, 26, 30, 33, 61, 84, 107 |
| | 3, 25, 29, 35, 56, 60 |
| GA | 1, 7, 11, 38, 43, 89, 101, 110, 112 |
| | 1, 5, 7, 11, 37, 42, 54, 64, 69, 88, 100, 109, 111 |
| GC | 25, 35, 56, 59, 75, 77, 79, 83, 87, 98, 104 |
| | 24, 34, 58, 74, 77, 86, 95, 97, 103 |
| GG | 22, 55, 63, 64, 65, 70, 74, 86, 95, 100, 103, 109, 114 |
| | 10, 21, 62, 63, 73, 82, 85, 94, 102, 108, 113 |
| GT | 23, 28, 50, 66, 71, 91, 96 |
| | 22, 27, 83, 90, 106 |
| TA | 11, 25, 34, 67, 92 |
| | 91 |
| TC | 4, 29, 31, 51, 72 |
| | 28, 30, 50, 71 |
| TG | 21, 24, 27, 34, 62, 85, 94, 97, 108 |
| | 20, 23, 26, 36, 57, 61, 84, 93, 107 |
| TT | 33, 37, 38, |

**Table 23.** List of matching of nucleotides in DNA sequence 1 and 2.

Difference = 0
(1,1) (2,2) (3,3) (4,4) (5,5) (6,6) (7,7) (8,8) (9,9) (10, 10) (12,12) (13, 13) (77, 77) (99, 99)
Difference = −1
(10, 9) (16, 15) (17, 16) (18, 17) (19, 18) (20, 19) (21, 20) (22, 21) (23, 22) (24, 23) (25, 24) (26, 25) (27, 26) (28, 27) (29, 28) (30, 29) (31, 30) (33, 31) (35, 34) (38, 37) (39, 38) (40, 39) (42, 41) (43, 42) (44, 43) (45, 44) (46, 45) (48, 47) (52, 51) (53, 52) (54, 53) (59, 58) (60, 59) (61, 60) (62, 61) (63, 62) (64, 63) (68, 67) (69, 68) (72, 71) (73, 72) (74, 73) (75, 74) (80, 79) (81, 80) (82, 81) (84, 84) (86, 85) (87, 86) (88, 87) (89, 88) (91, 90) (92, 91) (93, 92) (94, 93) (95, 94) (98, 97) (101, 100) (103, 102) (104, 103) (105, 104) (108, 107) (109, 108) (110, 109) (112, 111) (114, 113)

in both sequences only at sites 1, 7, and 11, while the same pair of nucleotides is moved by one position at the locations 38, 43, 89, 101, 110, and 112. The nucleotide pairs GA that appear in RABHBA at locations 5, 54, 64, and 69 have no corresponding nucleotides in CHPHBA and can be ignored in further analysis.

Table 22 contains information on the alignment of all 16 pairs of nucleotides. Table 22 shows nucleotides that are at the same sequential sites in both sequences, and nucleotides that are shifted by the same amount, to the left or right. In the case of the two DNA sequences selected for illustration of this search for DNA alignment, nucleotide pairs are either at the same site, or shifted by one place. Table 23 shows, extracted from Table 22, nucleotide pairs that are nonshifted or shifted by one place, which ends the search for DNA alignments of the sequences considered.

A convenient way to view Table 23 is to construct a spectral representation of the CHPHBA and RABHBA DNA sequences, shown in Figure 27, and to consider the difference of the spectral representation of the CHPHBA and RABHBA DNA sequences, which are illustrated in Figure 28. In contrast to the use of spectral representations of DNA and various differences of these spectral representations in search for the graphical alignment of DNA (as described in Ref. [18] when DNA sequences have been systematically shifted, both to the left and to the right relative to one another, here it is known exactly how much, and to which side, two sequences need to be shifted to visually illustrate the DNA alignment, as shown in Figure 27 for the shifts of zero or one sites. Figure 27 shows segments of DNA that are fully aligned, where adjacent nucleotides are at the *x*-axis, and illustrates occasional sites, such as (32, 33); (63, 64); and (77, 78) as aligned. These sites are of no consequence, as they illustrate "accidental" alignments of isolated pairs of nucleotides. Locally aligned segments of DNA are characterized by additional matching nucleotides that follow.

The graphical display of the aligned spectral differences of DNA sequences shows the Crick–Watson pairing of C–G and A–T when nucleotides are not the same. Because A, C, G, and T are assigned the numerical values of 1, 2, 3, and 4, respectively, the difference in pairing of C–G is ±1, and the difference for pairing of A–T is ±3. Hence, spots in Figure 27 that are on the horizontal lines ±1 and ±3 show the sites of Crick–Watson pairing. Similarly, the spots in Figure 27 that are on the horizontal lines ±2 correspond to the non-Crick–Watson pairing of A–G and C–T. This, of course, holds only when attention is restricted to the aligned segments (segment 1–15 for the spectral difference of 0 and the segment 9–114 for the spectral difference of 1). The small overlap of the above two intervals points to the possibility of locally alternative assignments for nucleotides in the overlapping regions.

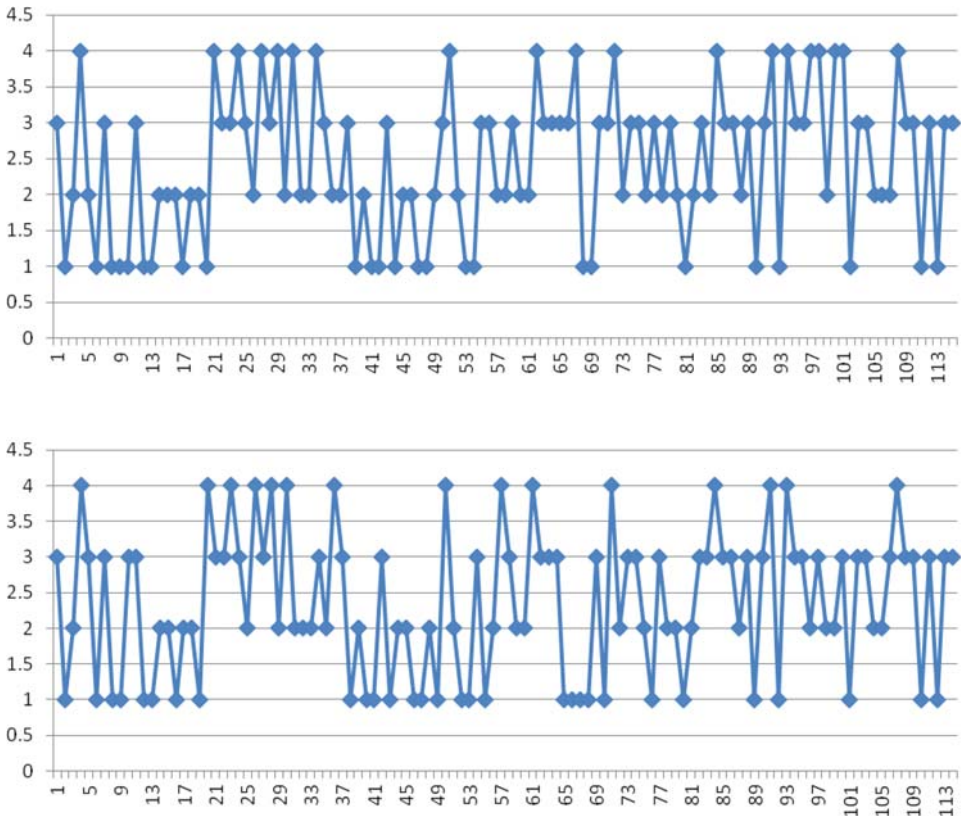


**Figure 27**. Spectral representation of the DNA sequences CHPHBA (top) and RABHBA (bottom).

## Milestones and Beyond

A closer look at the collections of seminal contributions to graphical bioinformatics listed in Table 1, selected as the milestones of graphical bioinformatics, shows that most of the selected articles have introduced novel methodologies or novel routes to the comparative

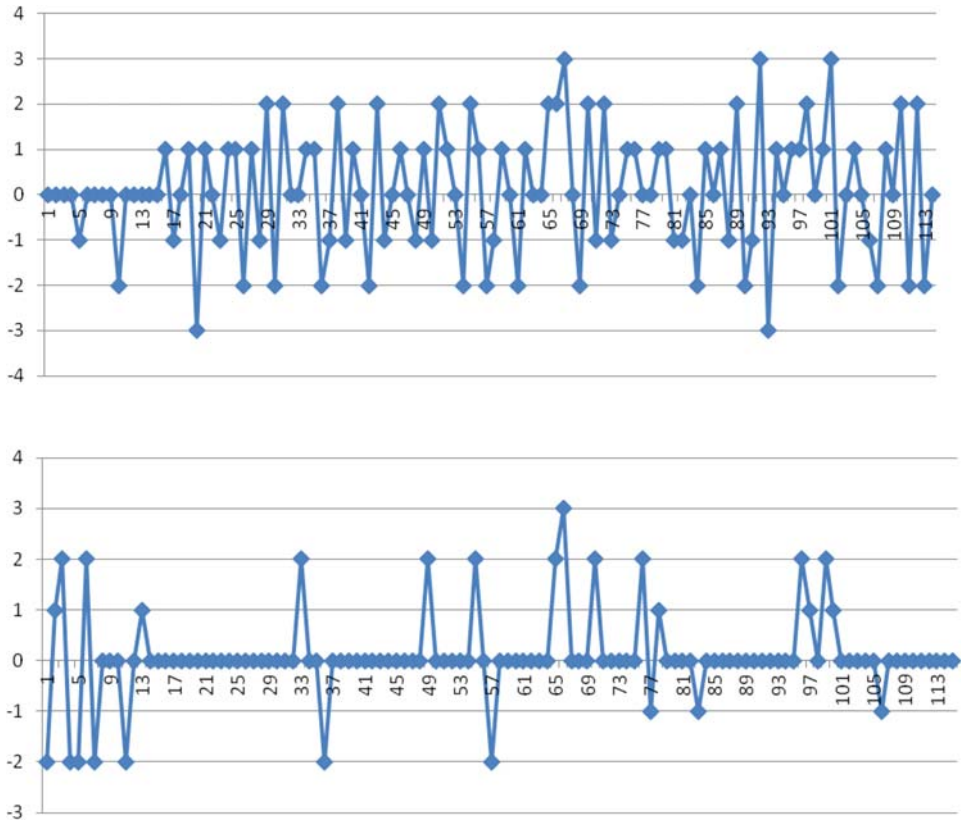

**Figure 28.** The difference of spectral representations of the DNA sequence.

study of the proteins, DNA and RNA. Their major contributions were not in solving problems of interest to biology; instead they were concerned with developing novel tools for solving important problems of biology. It may take some time to test the new tool, modify it if necessary to improve its performance, find the optimal one for specific tasks, and select it among competing variations. While attempts to perfect the existing approaches may be seen, the general conclusion is that novel tools have novel potential applications and may be important for solving both new problems and old unsolved problems.

Unsolved problems, not only in biology but also in chemistry, physics, and even mathematics, raise the question of why many of them have been so elusive. In some cases, including the central problem of protein alignment in bioinformatics, it appears that the reason for the delay was not due to the lack of imagination of scientists, but the lack of an adequate tool. The new tool for solving exactly the problem of protein alignment in bioinformatics is the modification of the AAA matrix, so that its elements are sets (collections of numbers), instead of numbers. The article describing VESPA may have been the first article to consider sets as matrix elements, not only in mathematical chemistry, but also in mathematics. In mathematics and mathematical chemistry, besides standard numerical matrices, more general matrices with subgraphs as matrix elements have also been used,[111,112] and alphanumeric matrices have been used in chemical documentation for some time.[113,114] Sets as matrix elements appear to be a novelty, which lead to solving the protein alignment problem. Matrices with sets as matrix elements were introduced not in an attempt to solve the protein alignment problem but in an attempt to recover the lost information that accompanies the construction of the AAA matrices. When this problem was solved, it immediately became clear that the use of the novel matrices solves one of the central problems of bioinformatics, the protein alignment problem.

Beside significant novel developments in methodologies to analyze proteins, in more recent years DNA and RNA biosequences have been seen, several of which are included in Table 1, as well as significant novel developments in applications of graphical methodologies to analyze proteins DNA and RNA biosequences. For example, simple numerical descriptors for quantifying effects of toxic substances on DNA[115] have been used to index single-nucleotide polymorphism (SNP) related gene sequences,[116] and to analyze the spread of avian flu and the numerical characterization of the H5N1 avian flu neuraminidase gene sequence,[117] including the study of dispersion and the extent of mutated and duplicated sequences of H5N1 influenza neuraminidase over twelve years (1997–2008).[118,119] More recently, the work on the viral-targeted applications of graphical bioinformatics was continued by A. Nandy and colleagues to identify targets for developing vaccines for flu and rotavirus varieties that should be immune to several cycles of mutations.[120,121] The same methodology was extended the numerical characterization of proteins.

Some of this work qualifies as milestones of graphical bioinformatics, but as the number of applications of graphical bioinformatics grows, the border between bioinformatics and biology is becoming less clear in the sense that some contributions involving elements of graphical bioinformatics also involve elements of biology and relate to problems of biology. The situation is similar to that of mathematical chemistry, and mathematics and chemistry, which sometimes have overlapping borders. It has been said semiseriously that mathematicians know how to solve problems of chemistry, but do not know what problems to consider; while chemists know which problems to consider, but do not know how to solve them. Together, mathematicians and chemists will form strong teams that may solve important problems. If scientists in bioinformatics and mathematical chemistry develop the tools, and scientists in chemistry offer problems, then their combined talents may lead to solutions to important problems of biology.

Graphical bioinformatics may build novel bridges between mathematics and computer science on one side; and chemistry, biochemistry, and biology on the other side, which are scientific disciplines that have their own languages, sometimes hampering communications. It is therefore important to encourage scientists on both sides of the gap to prepare general reviews of their research that may facilitate and strengthen further collaboration between the two sides. A few reviews on graphical bioinformatics are the previously mentioned "Graphical Representation of Proteins"[1] and "Novel Techniques of Graphical Representation and Analysis of DNA Sequences."[8]

We highly recommend the following recent reviews: "Proteomics, Networks, and Connectivity Indices,"[122] "Mathematical Descriptor of DNA Sequences: Development and Application,"[123] and "New Approaches to Drug–DNA Interactions Based on Graphical Representation and Numerical Characterization of DNA Sequences."[124]

## Other Voices

Even though the number of researchers in graphical bioinformatics is not large, except for China, there have been contributions from other research centers of graphical bioinformatics. Table 24 collects research groups worldwide in Europe, Asia, Africa, and the America; and Table 25 lists a fraction of contributions coming from China.

Table 24 shows that the chaos game representation of DNA and proteins has received attention. The research group of S. C. Basak at the Natural Resources Research Institute of the University of Minnesota at Duluth is the most active in graphical bioinformatics in the U. S. The same research group has also been very active in structure–activity relationship and quantitative structure–activity relationship, with particular interest in toxicity, including toxicoproteomics. Another visible group in graphical bioinformatics is led by one of early pioneers of graphical representations of DNA, Nandy in Calcutta, India. Table 1 includes several contributions by this group as important steps in the evolution of graphical bioinformatics since its beginning in 1983.

Finally, another visible and active group is the research group of Humberto Gonzáles-Díaz in Santiago de Compostela, Galicia, Spain. Their interest is in describing the connectivity of chemical and biological systems using networks, including very large networks, and developing tools for the study and characterization of proteomics maps, and also for describing protein interaction networks, which tend to be very complex. Those interested in complex networks, including protein interaction networks, and their analysis should consult several papers of Estrada and colleagues as a good introduction in this topic.[125,153–167]

Table 25 shows that research in China in graphical bioinformatics has deep roots, and it appears that China will soon, if not already, be the leading country in the development of graphical bioinformatics. In China, the dominant groups of researchers come from mathematical institutions, and they are interested in discrete mathematics and graph theory.

## Other Directions

Graphical bioinformatics as reviewed here was mostly confined to the application of discrete mathematics (in particular graph theory and partial ordering) and other methods of

**Table 24.** Selection of publications on discrete mathematics and graphical bioinformatics from different research centers worldwide.

| Authors | Topic | Country | Ref. |
|---|---|---|---|
| H. Gonzales-Díaz Y. Gonzales-Díaz L. Santana F. M. Ubeira E. Uriarte | Proteomics, networks and connectivity | Santiago de Compostela, Spain | [122] |
| E. Estrada | Protein interaction networks | Strachlide, Scotland (U. K.) | [125] |
| R. Todeschini V. Consonni A. Mauri D. Ballabio | Use of partial ordering for characterization of DNA | Milano, Italy | [126] |
| C. Lee C. Grasso M. F. Sharlow | Use of partial order graph for multiple sequence alignment | Los Angeles, California | [127] |
| N. Goldman | Chaos game representation of DNA and proteins | London, England (U. K.) | [128] |
| P. J. Deschavanne A. Giron J. Vilain G. Fagot B. Fertu | Characterization and Classification of spicies by chaos game representation of DNA sequences | Paris, France | [129] |
| A. Fiser G. E. Tusnady I. Simon | Chaos game representation of protein structures | Budapest, Hungary | [130] |
| S. Basu, A. Pan, C. Dutta J. Das. | Chaos game representation of protein structures | Calcutta, India | [131] |
| P. D. Cristea | DNA genomic signals | Bucharest, Rumania | [132,133] |
| A.Verma R. K. Singh | Ladder like structure for DNA | | [134] |
| A. Nandy P. Nandy | On uniqueness of DNA descriptors | Calcutta, India | [135] |
| D. Bielisnska-Waz T. Clark P. Waz W. Nowak A. Nandy | 2D dynamic representation DNA | Warszawa, Poland | [136] |
| S. Larionov A. Loskutov E.Ryadcheno | Palindromic context of life | | [137] |
| A. Perdih, A. Roy Choudhury Š. ŽuperlE. Sikorska I. Zhukov T. Šolmajer M. Novič | Sructural analysis of peptide fragment of transmembrane transporter protein bilitranslocase | National Institute of Chemistry, Ljubljana, Slovenia | [138] |
| A. T. Balaban M. Randić | 8×8 tabular representation of the genetic code | Texas A&M University, atGalveston TX | [139] |
| M. Randić A. T. Balaban T. Pisanski M. Novič | Novel graphical representation of proteins | National Institute of Chemistry, Ljubljana, Slovenia | [140] |

**Table 25.** Small fraction of publications in graphical bioinformatics from different research centers in China.

| Authors | Topic | City/University | Ref. |
|---|---|---|---|
| F. Bai T. Wang | 2D graphical representation of proteins based on codons | Dalian Univ. Techn., Dalian | [141] |
| J. Song | Similarity of DNA based on 3-DGraphical representation | Shaoguan Univ., Shaoguan | [142] |
| P-A. He Y-P. Zhang Y-H. Yao Y-F. Tang X-Y. Nan | Graphical representation of proteins based on their physic-chemical properties | Zhejiang Univ., Hangzhou and Chinese Academy of Sciences, Beijing | [143] |
| W. Wang B. Liao T. Wang W. Zhu | Graphical method for construction of phylogenetic tree | Dalian Univ., Dalian and Hunan Univ. Changsha | [144] |
| R. Wu R. Li H. Yan M. Yang | DNA sequence visualization | Hunan Univ., Changsha and Hunan Jaixing Univ., Jaixing Zhejiang | [145] |
| Y. Guo T.-m. Wang | Graphical method to analyze similarity of DNA | Dalian Univ. of Technology, Dalian | [146] |
| Y-H. Yao X-Y. Nan T.-m. Wang | Classification and similarity/dissimilarity of DNA | Zhejiang Univ., Hangzhou and Hainan Normal Univ., Haikou | [147] |
| C. Yu Q. Liang C. Yin R. L. He S. S.-T. Yau | Novel construction of genome space | Chinese Univ. of Hong Kong, Hong Kong | [148] |
| B. Liao Y. Zhang K. Ding T. Wang | Similarity/dissimilarity of DNA | Dalian Univ., Dalian and Hunan Univ. Changsha | [149] |
| F. Bai D. Li T. Wang | Mapping of RNA secondary structure | | [150] |
| B. Liao X. Shan W. Zhu R. Li | Phylogenetic tree construction based on 2D graphical representation | Dalian Univ., Dalian and Hunan Univ. Changsha | [151] |
| B. Liao | 2D graphical representation of DNA | Dalian Univ., Dalian and Hunan Univ. Changsha | [152] |
| See also Ref.: | [17,19,20,53–63] | | |

mathematical chemistry to problems considered in bioinformatics, and the use of such approaches to transform qualitative results into quantitative results. However, this is not the only possible route to transforming qualitative results into quantitative results, and to considering the visual representation of quantitative results once they are obtained. One such approach is based on lattice models, which have been used in polymer physics[168] and in biopolymers in chemistry.[169–173] Another approach to the study of DNA and proteins is the recurrence quantification analysis (RQA), a nonlinear technique initially developed as a purely graphical method[174] and soon upgraded to a quantitative method.[175,176] These "computational biochemistry" approaches focus on the relationship between sequence embedded information and protein folding, the sequence–structure puzzle,[177,178] which is one of the central concerns in theoretical and applied biochemical research.

### Lattice model

The lattice model starts by embedding biosequences (proteins, DNA) on a square grid, limiting interactions to residues with "topological" neighbors. Interactions are based on potential functions constructed on selected physicochemical properties, such as hydrophobicity. A brief discussion of lattice models can be found in the introduction of the review article "Nonlinear signal analysis methods in the elucidation of protein sequence–structure relationship" by A. Giuliani et al.[179]

Figure 29 illustrates a small 3 × 3 lattice and one conformation of a polymer having nine monomer units embedded in this lattice, taken from the paper by Chan and Dill.[170] Table 26 shows a 9 × 9 matrix of the 3 × 3 lattice "contact map", which has nonzero entries for the topological contacts (1, 4); (3, 6); (3, 8); and (2, 9). The order $k$ of a contact is the chain length between the two monomers in contact.

Figure 30 shows the 27 amino acids protein embedded in a 3 × 3 × 3 lattice by Šali et al.,[173] which has 28 topological contacts. The corresponding 28 × 28 adjacency matrix corresponds to the graph illustrated in Figure 31. All topological contacts of the conformation illustrated on the 3 × 3 × 3 lattice are identified on the graph in Figure 31, by assuming consecutive numbering of vertices along the protein. Figure 30 identifies the first and the last vertex of the embedded protein.

For lattice proteins, as outlined by Šali et al., one can calculate the total energy of the conformation ($E$), which is given as the sum of the contact energies $B_{ij}$ between nonbonded adjacent amino acids on the lattice:

$$E = \sum_{i<j} \Delta(r_i, r_j) B_{ij}$$

The $\Delta(r_i, r_j)$ equals 1 if amino acids are in contact (nonbonded but adjacent), and 0 otherwise. In this model, two amino acids are in contact if they are not adjacent in the protein sequence and are at the unit distance from each other in

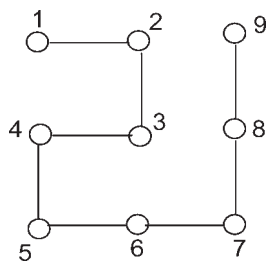


**Figure 29**. A conformation of a nine monomer polymer embedded on a 3 × 3 lattice with topological contacts (1,4) (2, 9) (3, 6) (3, 8).

**Table 26.** The contact map matrix for 3×3 contact map of Figure 28.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 3 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| 4 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

the lattice. Assuming that all $B_{ij} = 1$, the evaluation of the total energy of conformation is reduced to the "hard ball" potential model of Bloch[180] used for the calculation of electron mobility in metals, which Erich Hückel adopted for his molecular orbital calculations of benzene and other $\pi$-electron systems.[181–183] The secular equation for both is represented by a binary matrix.

### Signal analysis methods

Signal analysis methods, developed in physics and engineering, typically apply to very long signal inputs. In biology applications, amino acids of protein sequences are viewed as a string of signals, which are relatively short (most are fewer than several hundred amino acids, and as few as a dozen or two dozen), limiting the use of some techniques of signal analysis. Proteins are reduced to 1D numerical sequences, which resemble "spectra" when represented graphically. One important advantage of such representations of proteins is that they allow the analysis of individual (single) proteins, rather than considering pairwise alignment, which is typical of computer-based bioinformatics analyses.

Figure 32 shows the hydrophobic profile of the protein 1 of *Saccharomyces cerevisiae*, the amino acids of which are listed in Table 3. The hydrophobicity scale by Schneider and Wrede[184,185] is used:

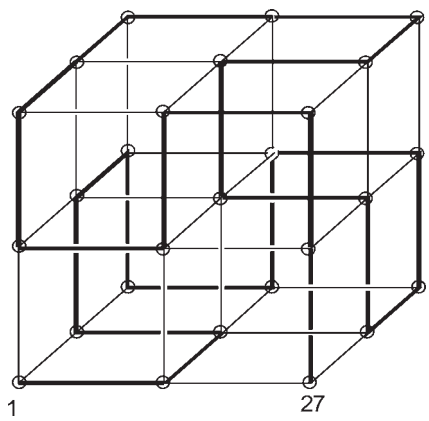


**Figure 30**. A conformation of a 27 monomer polymer embedded on a 3 × 3 × 3 lattice with 28 topological contacts. Reproduced with permission from Ref. [1].

| A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| W | Y | V | | | | | | | | | | | | | | |
| 1.6 | −12.3 | −4.8 | −9.2 | 2 | −4.1 | −8.2 | 1 | −3 | 3.1 | | | | | | | |
| 2.8 | −8.8 | 3.4 | 3.7 | −0.2 | 0.6 | 1.2 | 1.9 | −0.7 | 2.6 | | | | | | | |

High values (positive values) correspond to hydrophobic amino acids (A, C, G, I, L, M, F, S, T, W, and V), while negative values correspond to hydrophilic amino acids (R, N, D, Q, E, H, K, P, and Y). According to A. Giuliani (personal communication), Palliser and Parry[184] are quoted in Giuliani's review[179]- because their article is a great general summary of hydrophobic scales. Schenider and Wrede used the Engelmann scale,[185] but this scale is normally referred to as "Schneider and Wrede."

While Figure 32 is similar to Figure 13, which shows the spectral representation of the same protein, Figure 32 has an advantage in that amplitudes of spectral peaks have physicochemical interpretation, that of hydrophobicity, while spectral amplitudes in Figure 13 are arbitrary.

### RQA

RQA was originally developed by Eckmann in 1987,[174] about 25 years ago, as a purely qualitative approach. Several years later, Webber and Zbilut[175] ungraded the RQA by developing quantitative methods for the analysis of qualitative recurrence plots, which are essentially an adjacency matrix. The concept of recurrence is simple: recurrence in a protein (or DNA) sequence is the element that repeats itself. The concept of

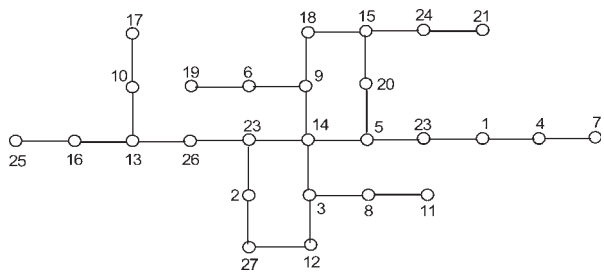


**Figure 31**. Graph corresponding to the topological contacts of the conformation of polymer of Figure 30 embedded in a 3 × 3 × 3 lattice.

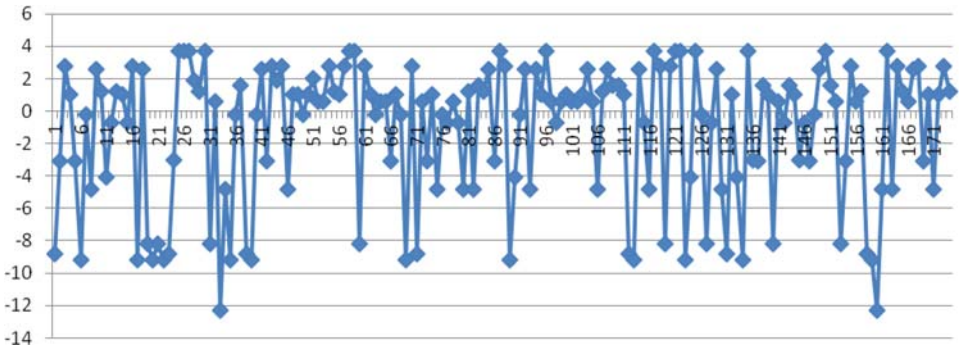


**Figure 32.** Hydrphobicity profile of the protein 1 of *Saccharomyces cerevisiae* (amino acids of which have been listed in Table 3).

recurrence in 3D is formally expressed as follows: given an element $X_0$ and a sphere of radius $R$, a point $X$ is said to recur with respect to $X_0$ if

$$B_R(X_0)=\{X : \| X-X_0 \ \| \leq R\}$$

In the case of protein sequences, the recurrence corresponds to segments of amino acids of considered length, associated with their hydrophobicity profile, shared with other segments along the sequence having the same hydrophobicity profiles. The recurrence plots represent a graphical record of the recurrences in the form of the symmetrical $N \times N$ matrix, in which an element $(i, j)$ is represented as a spot if the distance between $X_i$ and $X_j$ is smaller than the radius $R$. When spots are replaced by 1, and all blanks are assigned 0 values, the adjacency matrix is obtained for the recurrence plot. This matrix allows the construction of several matrix invariants to be used as recurrence plot descriptors. Webber and Zbilut[175] considered the following:

1. The percentage of plot filled with recurrent points;
2. The percentage of recurrent points forming line segments parallel to the main diagonal, with a minimum of line segments having two points;
3. The Shannon information entropy of the line length distribution;
4. The length of the longest line segment; and
5. The measure of the boundary of recurrent points away from the central diagonal.

These five statistical data allow the construction of five component vectors or a five-dimensional representation of autocorrelation structures of protein sequences, which parallels the visual impression of the plots by unbiased observers.[186] The above five (statistical) matrix invariants can be rephrased by replacing "spots" with 1s and "blanks" with 0s to make the adjacency matrix more apparent. The five descriptors of RQA recurrence plots are unknown in chemical graph theory and its extension to bioinformatics, in both of which the adjacency matrix plays a dominant role. The adjacency matrix of the RQA is an ordered adjacency matrix, but several of the above descriptors are matrix invariants (an ordered or not-ordered matrix is considered). In chemical graph theory and its extension to bioinformatics, commonly used matrix invariants are the leading eigenvalues of the matrix, the determinant of the matrix, the set of eigenvalues, the leading eigenvector, the coefficients of the characteristic polynomial, and the ordered row sums. Here is an opportunity for both groups to benefit by considering alternative sets of adjacency matrix invariants. On this helpful and hopeful note, we end this review article on graphical bioinformatics.

## Concluding Remarks

This review tries to outline major accomplishments of graphical bioinformatics, with which many in bioinformatics may not have been familiar. Because graphical bioinformatics may not have received sufficient attention in some circles interested in bioinformatics, the main purpose of this article is to draw attention of researchers in bioinformatics to graphical bioinformatics, which deserves attention for at least two reasons:

1. In graphical bioinformatics, in contrast to standard bioinformatics, a single DNA, a single RNA, and a single protein can be characterized numerically. This allows results to be compiled on a single DNA, a single RNA, and a single protein to eventually build up an atlas or a catalogue of DNA, RNA, and proteins, analogous to such catalogues or atlases of chemicals, fullerenes, and so on.
2. Graphical bioinformatics has led to significant novel insights and results in bioinformatics, which are listed in Table 1, and should not be overlooked. We encourage other researchers in graphical bioinformatics and in bioinformatics to supplement the material here with reports on work that we have not discussed. In particular, we invite leading authorities in bioinformatics to come forward with their own tables of "Milestones in Bioinformatics" and share the most significant results and directions of research in bioinformatics. It would be interesting to see how many of the topics listed in Table 1 would be included in more general tables on milestones in bioinformatics.

The selection of milestones in Table 1 is subjective. They are listed more-or-less chronologically, and only on few of them are elaborated upon. This was the case with VESPA and VESNA. We have also discussed the construction of sparse matrices, as they have computational advantages. Similarly, we have discussed partial ordering, as this concept is not well-known in chemistry. We have said nothing about the virtual genetic code or the representation of RNA without loss of information, and at best we have said very little about the

graphical representation of proteins by graphs. However, all these topics have been covered in the very recent review on graphical representation of proteins,[1] where interested readers can find more information. We could have said more about the graphical alignment of DNA and the graphical alignment of proteins, because both publications outlined a novel approach to the alignment of biosequences, which differ from the standard computer-based programs in that they do not involve empirical parameters, such as penalties for gaps and mismatches. We also have not elaborated on the pioneering work of Hamory, Jeffrey, and Nandy, but end by stating that we continue in the spirit of Hamori and Nandy by introducing additional graphical representations of DNA. We believe that their spectral representations are the most profound, because they lead to the graphical alignment approach for DNA and proteins. We have also adopted the chaos game representation of DNA introduced by Jeffrey, though with one significant distinction, in that we considered such representations only for relatively small $n$ (the number of nucleotides), including as the extreme the chaos game representations of codons (three nucleotide sequences), which present a new way to the graphical representation of proteins. In contrast, Jeffrey and those who followed considered very lengthy DNA sequences having 10,000 and more nucleotides.

In our opinion, the three most significant recent results of graphical bioinformatics are

1. The exact solution to protein and DNA alignment problems;
2. The numerical representation of proteomics maps and finding hormesis on cellular level; and
3. The spectral representation of DNA and proteins and their graphical alignment.

Of course, these significant results were not independent of most of the other topics discussed in this review.

## Acknowledgments

*M.R. wishes to thank the Chemometrics Laboratory of the National Institute of Chemistry of Slovenia for cordial hospitality. The authors would like to thank Professor A. T. Balaban (Texas A&M University at Galveston, Texas) for reading the manuscript and for his numerous suggestions that improved the presentation of the material. They also thank A. Nandy, one of early pioneers of Graphical Bioinformaics, for examining the manuscript and sending his comments, including a list of some overlooked publications.*

**Keywords:** graphical bioinformatics · chaos game · spectral representation of DNA · graphical alignment of DNA · graphical alignment of proteins · hormesis at cell level · exact solution of protein alignment · exact solution of DNA alignment

How to cite this article: M. Randić, M. Novič, D. Plavšić, *Int. J. Comput. Chem.* **2013**, DOI: 10.1002/qua.24479

[1] M. Randić, J. Zupan, A. T. Balaban, D. Vikić-Topić, D. Plavšić, *Chem. Rev.* **2011**, *111*, 790.
[2] E. Hamori, *BioTechniques* **1989**, *7*, 710.
[3] M. Barnsley, Fractals Everywhere; Morgan Kaufmann: San Francisco, **1996**.
[4] A. Nandy, *Indian J. Biochem. Biophys.* **1994**, *131*, 149.
[5] A. Nandy, P. Nandy, *Current Sci.* **1995**, *68*, 75.
[6] S. Gosh, A. Roy, S. Adhya, A. Nandy, *Current Sci.* **2003**, *84*, 1534.
[7] A. Nandy, *Comput. Appl. Biosci. (Cabios)* **1996**, *12*, 55.
[8] A. Roy, C. Raychaudhury, A. Nandy, *J. Biosci.* **1998**, *23*, 55.
[9] M. Randić, J. Zupan, M. Novič, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1339.
[10] M. Randić, *Int. J. Quantum Chem.* **2002**, *90*, 848.
[11] M. Randić, F. Witzmann, M. Vračko, S. C. Basak, *Med. Chem. Res.* **2001**, *10*, 456.
[12] M. Randić, In Handbook of Proteomics Methods; P. M. Conn, Ed.; Humana Press, Inc.: Totowa, NJ, **2003**; pp. 429–450.
[13] M. Randić, M. Vračko, N. Lerš, D. Plavšić, *Chem. Phys. Lett.* **2003**, *368*, 1.
[14] M. Randić, *Chem. Phys. Lett.* **2004**, *386*, 468.
[15] M. Randić, N. Novič, M. Vračko, *J. Chem. Inf. Model.* **2005**, *45*, 1205.
[16] M. Randić, E. Estrada, *J. Proteome Res.* **2005**, *4*, 2133.
[17] B. Liao, M. Tan, K. Ding, *Chem. Phys. Lett.* **2005**, *414*, 296.
[18] M. Randić, J. Zupan, D. Vikić-Topić, D. Plavšić, *Chem. Phys. Lett.* **2006**, *431*, 375.
[19] B. Liao, X. Xiang, W. Zhu, *J. Comput. Chem.* **2006**, *27*, 1196.
[20] B. Liao, M. Tan, K. Ding, *Chem. Phys. Lett.* **2005**, *414*, 296.
[21] M. Randić, J. Zupan, D. Vikić-Topić, *J. Mol. Graph. Model.* **2007**, *26*, 290.
[22] M. Randić, *J. Math. Chem.* **2008**, *43*, 756.
[23] M. Randić, M. Vračko, M. Novič, D. Plavšić, *Int. J. Quantum Chem.* **2009**, *109*, 2982.
[24] M. Randić, D. Plavšić, *Chem. Phys. Lett.* **2008**, *456*, 84.
[25] A. Roy Choudhury, M. Novič, *SAR QSAR Environ. Res.* **2009**, *20*, 741.
[26] A. Roy Choudhury, N. Zhukov, M. Novič, Sci. World J. (in press).
[27] E. Hamori, J. Ruskin, *J. Biol. Chem.* **1983**, *258*, 1318.
[28] E. Hamori, In Frontiers of Computing Science, Vol. *3*: Scientific Visualization; C. Pickover, S. K. Tewksbury, Eds.; Plenum Press: New York, **1994**; pp. 90–101.
[29] H. J. Jeffrey, *Nucleic Acids Res.* **1990**, *18*, 2163.
[30] H. J. Jeffrey, *Comput. Graphics* **1992**, *16*, 25.
[31] M. F. Barnsley, H. Rising, Fractals Everywhere, 2nd ed.; Academic Press: Boston, MA, **1993**.
[32] H.-O. Peitgen, H. Jürgens, D. Saupe, Chaos and Fractals: New Frontiers of Science; Springer-Verlag: Berlin, Germany, **1992**.
[33] A. Nandy, *Curr. Sci.* **1994**, *66*, 309.
[34] A. Nandy, *Curr. Sci.* **1994**, *66*, 821.
[35] A. Nandy, *Curr. Sci.* **1996**, *70*, 661.
[36] C. Raychaudhury, A. Nandy, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 243.
[37] M. Randić, M. Vračko, A. Nandy, S. C. Basak, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1325.
[38] M. Randić, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1330.
[39] M. Randić, *SAR QSAR Environ. Res.* **2004**, *15*, 147.
[40] R. Orel, M. Randić, *J. Math. Chem.* **2012**, *50*, 2689.
[41] M. Randić, *J. Comput. Chem.* **2012**, *33*, 702.
[42] M. Randić, *J. Comput. Chem.* **2013**, *34*, 77.
[43] M. Randić, R. Orel, *J. Math. Chem. (submitted).*
[44] M. Randić, A. F. Kleiner, L. M. DeAlba, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 277.
[45] M. Randić, M. Vračko, M. Novič, S. C. Basak, *MATCH Commun. Math. Comp. Chem.* **2000**, *42*, 181.
[46] O. Perron, *Math. Ann.* **1907**, *64*, 248.
[47] (a) G. Frobenius, *Sitzungberichte der königlichen preussischen Akademie der Wissenschaften zu Berlin* **1908**, 471; (b) G. Frobenius, *Sitzungberichte der königlichen preussischen Akademie der Wissenschaften zu Berlin* **1909**, 514.
[48] G. Frobenius, *Sitzungberichte der königlichen preussischen Akademie der Wissenschaften zu Berlin* **1912**, 456.
[49] M. Randić, R. Orel, *J. Math. Chem.* **2012**, *49*, 1759.
[50] M. A. Gates, *Nature*, **1985**, *316*, 219.
[51] M. A. Gates, *J. Theor. Biol.* **1986**, *119*, 319.
[52] P. M. Leong, S. Morgenthaler, *Comput. Appl. Biosci.* **1995**, *12*, 503.
[53] X. Guo, M. Randić, S. C. Basak, *Chem. Phys. Lett.* **2001**, *350*, 106.

[54] S. S.-T. Yau, J. Wang, A. Niknejad, C. Lu, N. Jin, Y.-K. Ho, *Nucleic Acid Res.* **2003**, *31*, 3078.
[55] Y. Wu, A. W.-C. Liew, H. Yan, M. Yang, *Chem. Phys. Lett.* **2003**, *367*, 170.
[56] X. Guo, A. Nandy, *Chem. Phys. Lett.* **2003**, *369*, 361.
[57] B. Liao, *Chem. Phys. Lett.* **2005**, *401*, 196.
[58] J. Wang, W. Wang, *Biophys. Rev. Lett.* **2006**, *1*, 133.
[59] S. S.-T. Yau, J. Wang, A. Niknejad, C. Lu, N. Jin, Y.-K. Ho, *Nucleic Acid Res.* **2013**, *31*, 3078.
[60] B. Liao, T.-M. Wang, *J. Comput. Chem.* **2004**, *25*, 1364.
[61] J. Luo, B. Liao, R. Li, W. Zhu, *J. Math. Chem.* **2006**, *39*, 629.
[62] Z.-J. Zhang, *Bioinformatics* **2009**, *25*, 1112.
[63] C. Li, J. Hu, *J. Biochem. Mol. Biol.* **2006**, *39*, 292.
[64] S. B. Needleman, C. D. Wunsch, *J. Mol. Biol.* **1970**, *48*, 443.
[65] T. F. Smith, M. S. Waterman, *J. Mol. Biol.* **1981**, *147*, 195.
[66] D. J. Lipman, W. R. Pearson, *Science* **1985**, *227*, 1435.
[67] W. R. Pearson, D. J. Lipman, *Proc. Natl. Acad. Sci. USA* **1988**, *85*, 2444.
[68] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, *Nucleic Acids Res.* **1997**, *25*, 3389.
[69] R. A. Lipert, H. Huang, M. S. Waterman, *Proc. Natl. Acad. Sci. USA* **2002**, *9*, 13980.
[70] V. I. Levenshtein. *Sov. Phys. Dokl.* **1966**, *10*, 707.
[71] J. Gasteiger, Ed., Handbook of Chemoinformatics—From Data to Knowledge in 4 Volumes, Vol. *3*; Wiley-VCH: Weinheim, **2003**.
[72] Z. Deanović, Z. Supek, M. Randić, *Nature*, **1964**, *201*, 411.
[73] J. Calabrese, Hormesis: a revolution in toxicology, risk assessment and medicine, *EMBO* **2004**, *5*, S37.
[74] J. Calabrese, L. A. Baldwin, *Toxicol. Sci.* **2001**, *62*, 330.
[75] N. L. Anderson, R. Esquer-Blasco, F. Richardson, P. Foxworthy, P. Eacho, *Toxicol. Appl. Pharmacol.* **1996**, *137*, 75.
[76] M. Randić, Selection of Quotes by Scientists or about Science, *Indian J. Mathematics Teaching*. **2000**, *26*, 11.
[77] M. Randić, *MATCH—Commun. Math. Comput. Chem.* **2008**, *59*, 5.
[78] W. Kohn, A celebration of the Contributions of Robert G. Parr, Vol. *1*; K. D. Sen, Ed.; World Scientific: Singapore, **2002**.
[79] M. Randić, *J. Proteome Res.* **2006**, *5*, 1575.
[80] M. Randić, F. A. Witzmann, V. Kodali, S. C. Basak, *J. Chem. Inf. Model.* **2006**, *46*, 116.
[81] M. Randić, C. L. Wilkins, *Chem. Phys. Lett.* **1979**, *63*, 332.
[82] M. Randić, C. L. Wilkins, *J. Phys. Chem.* **1979**, *83*, 1525.
[83] M. Randić, *MATCH Commun. Math. Comput. Chem.* **1979**, *5*, 3.
[84] M. Randić, *J. Chem. Educ.* **1992**, *69*, 713.
[85] M. Randić, M. Vračko, M. Novič, S. C. Basak, *MATCH Commun. Math. Comput. Chem.* **2000**, *42*, 181.
[86] M. Randić, S. C. Grossman, B. Jerman-Blažič, D. H. Rouvray, S. El-Basil, *Math. Comput. Model.* **1988**, *11*, 837.
[87] M. Randić, *Acta Chim. Slov.* **2000**, *47*, 143.
[88] B. R. Kowalski, C. F. Bender, *J. Am. Chem. Soc.* **1972**, *94*, 5632.
[89] M. Randić, M. Novič, M. Vračko, D. Plavšić, *J. Theor. Biol.* **2010**, *266*, 21.
[90] M. Pompe, M. Randić, *Acta Chim. Slov.* **2007**, *54*, 605.
[91] M. Pompe, M. Veber, M. Randić, A. T. Balaban, *Molecules* **2004**, *9*, 1160.
[92] M. Randić, M. Pompe, D. Mills, S. C. Basak, *Molecules* **2004**, *9*, 1177.
[93] M. Randić, S. C. Basak, M. Pompe, M. Novič, *Acta Chim. Slov.* **2001**, *48*, 169.
[94] M. Randić, *J. Mol. Graph. Model.* **2001**, *20*, 19.
[95] M. Randić, M. Pompe, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 575.
[96] M. Randić, S. C. Basak, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 614.
[97] M. Randić, D. Plavšić, N. Lerš, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 657.
[98] M. Randić, S. C. Basak, M. Pompe, M. Novič, *Acta Chim. Slov.* **2001**, *48*, 169.
[99] M. Randić, D. Mills, S. C. Basak, *Int. J. Quantum Chem. Quantum Biol. Symp.* **2000**, *80*, 1199.
[100] M. Randić, J. Cz. Dobrowolski, *Int. J. Quantum Chem.* **1998**, *70*, 1209.
[101] M. Randić, *J. Comput. Chem.* **1991**, *12*, 970.
[102] M. Randić, *Croat. Chem. Acta* **1991**, *64*, 43.
[103] M. Randić, *Intell. Lab. Syst.* **1991**, *10*, 213.
[104] M. Randić, J. Zupan, M. Novič, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1339.
[105] M. Randić, M. Novič, M. Vračko, *J. Chem. Inf. Comput. Sci.* **2005**, *45*, 1205.
[106] M. Randić, M. Novič, A. R. Choudhury, D. Plavšić, *SAR QSAR Environ Res.* **2012**, *23*, 327.
[107] R. C. Read, *Lect. Notes Math.* **1981**, *884*, 77.
[108] W. W. Rouse Ball, Mathematical Recreations and Essays; MacMillan and Co., Ltd: London, **1960**.
[109] M. Randić, M. Novič, M. Vračko, *SAR QSAR Environ. Res.* **2008**, *19*, 339.
[110] H. Tietze, Famous Problems of Mathematics: Solved and Unsolved Mathematical Problems from Antiquity to Modern Times; Greylock Press: New York, NY, **1965**.
[111] M. Randić, M. Plavšić, M. Razinger, *MATCH—Commun. Math. Comput. Chem.* **1997**, *35*, 243.
[112] M. Randić, N. Basak, D. Plavšić, *Croat. Chem. Acta* **2004**, *77*, 251.
[113] L. Spialter, *J. Am. Chem. Soc.* **1963**, *85*, 2012.
[114] L. Spialter, *J. Chem. Docum.* **1964**, *4*, 261.
[115] A. Nandy, S. C. Basak, *J Chem Inf. Comput. Sci.* **2000**, *40*, 915–919.
[116] A. Nandy, P. Nandy, S. C. Basak, *Internet. Electron. J. Mol. Design* **2002**, *1*, 367.
[117] A. Nandy, B. Gute, S. C. Basak, *J. Chem. Inf. Model.* **2007**, *47*, 945.
[118] A. Ghosh, A. Nandy, P. Nandy, B. D. Gute, S. C. Basak, *J. Chem. Inf. Model.* **2009**, *49*, 2627.
[119] A. Ghosh, A. Nandy, P. Nandy, *BMC Struct. Biol.* **2010**; doi:10.1186/1472-6807-10-6.
[120] A. Ghosh, S. Chattopadhyay, M. Chawla-Sarkar, P. Nandy, A. Nandy, In Silico Study of Rotavirus VP7 Surface Accessible Conserved Regions for Antiviral Drug/Vaccine Design. *PLoS One* **2012**; doi:10.1371/journal.pone.0040749.
[121] A. Nandy, A. Ghosh, P. Nandy, Numerical Characterization of Protein Sequences and Application to Voltage-Gated Sodium Channel Alpha Subunit Phylogeny, *In Silico Biol.* **2009**, *9*, 77.
[122] H. Gonzales-Díaz, Y. Gonzales-Díaz, L. Santana, F. M. Ubeira, E. Uriarte, *Proteomics* **2008**, *8*, 750.
[123] A. Nandy, M. Harle, S. C. Basak, Mathematical descriptors of DNA sequences: development and applications, *ARKIVOC* **2006**, *9*, 211.
[124] A. Nandy, S. C. Basak, *Curr. Comput. Aided Drug Des.* **2010**, *6*, 283.
[125] E. Estrada, *Proteomics* **2006**, *6*, 35.
[126] R. Todeschini, V. Consonni, A. Mauri, D. Ballabio, *J. Chem. Inf. Model.* **2006**, *46*, 1905.
[127] C. Lee, C. Grasso, M. F. Sharlow, *Bioinformatics* **2002**, *18*, 452.
[128] N. Goldman, *Nucleic Acid Res.* **1993**, *21*, 2487.
[129] P. J. Deschavanne, A. Giron, J. Vilain, G. Fagot, B. Fertu, *Mol. Biol. Evol.* **1999**, *16*, 1391.
[130] A. Fiser, G. E. Tusnády, I. Simon, *J. Mol. Graph.* **1994**, *12*, 302.
[131] S. Basu, A. Pan, C. Dutta, J. Das, *J. Mol. Graph. Model.* **1998**, *15*, 279.
[132] P. D. Cristea, In Genomic Signal Processing and Statistics, Vol. *2*; E. R. Dougherty, I. Shmulevich, J. Chen, Z. J. Wang, Eds.; Hindawi Publ. Corp: Cairo, Egypt, **2005**.
[133] P. D. Cristea, *J. Cell. Mol. Med.* **2002**, *6*, 279.
[134] A. Verma, P. K. Singh, *J. Comp. Sci. Inf. Techn.* **2012**, *3*, 4596.
[135] A. Nandy, P. Nandy, *Chem. Phys. Lett.* **2002**, *368*, 102.
[136] D. Bieliska-Waz, T. Clark, P. Waz, W. Nowak, A. Nandy, *Chem. Phys. Lett.* **2007**, *442*, 140.
[137] S. Larionov, A. Loskutov, E. Ryadcheno, *Chaos* **2008**, *18*, 013105.
[138] A. Perdih, A. Roy Choudhury, Š. Župerl, E. Sikorska, I. Zhukov, T. Šolmajer, M. Novič, *PLoS One* **2012**, *7*, e38967.
[139] A. T. Balaban and M. Randić, New Chessboard (8 × 8) Representation of the Standard Genetic Code, and its Application for Representing Primary Structures of Proteins, Biotechno. 2008 pp. 76–81.
[140] M. Randić, A. T. Balaban, T. Pisanski, M. Novič, A novel graphical representation of proteins, *Period. Biol.* **2005**, *197*, 404.
[141] F. Bai, T. Wang, *Chem. Phys. Lett.* **2005**, *413*, 458.
[142] J. Song, *Adv. Biomed. Eng.* **2012**, *9*, 29.
[143] P.-A. He, Y.-P. Zhang, Y.-H. Yao, Y. F. Tang, X.-Y. Nan, *J. Comput. Chem.* **2010**, *31*, 2136.
[144] W. Wang, B. Liao, T. Wang, W. Zhu, *Int. J. Quantum Chem.* **2006**, *106*, 1998.
[145] R. Wu, R. Li, B. Liao, G. Yue, *MATCH Commun. Math. Comp. Chem.* **2010**, *63*, 679.
[146] Y. Guo, T.-M. Wang, *J. Mol. Struct. (THEOCHEM)* **2008**, *853*, 62.
[147] Y. Yao, X.-Y. Nan, T.-M. Wang, *J. Mol. Struct. (THEOCHEM)* **2006**, *764*, 101.
[148] C. Yu, Q. Liang, C. Yin, R. L. He, S. S.-T. Yau, *DNA Res.* **2010**, *17*, 155.
[149] B. Liao, Y. Zhang, K. Ding, T. Wang, *J. Mol. Struct. (THEOCHEM)* **2005**, *717*, 199.

[150] F. Bai, D. Li, T. Wang, *J. Math. Chem.* **2008**, *43*, 932.
[151] B. Liao, X. Shan, W. Zhu, R. Li, *Chem. Phys. Lett.* **2006**, *422*, 282.
[152] B. Liao, *Chem. Phys. Lett.* **2005**, *401*, 196.
[153] E. Estrada, Chaos: Interdiscip. *J. Non Linear Sci.* **2011**, 21, 047101.
[154] E. Estrada, *Network heterogeneity, Phys. Rev. E* **2010**, *82*, 066102.
[155] E. Estrada, *Acta Chim. Slov.* **2010**, *57*, 597.
[156] E. Estrada, S. Gago, G. Caporossi, *Automatica* **2010**, *46*, 1835.
[157] J. J. Crofts, E. Estrada, D. H. Higham, A. Taylor, *Electron. Trans. Numer. Anal.* **2010**, *37*, 337.
[158] E. Estrada, N. Hatano, *Physica A* **2010**, *309*, 3648.
[159] E. Estrada, *J. Theor. Biol.* **2010**, *263*, 556.
[160] E. Estrada, D. J. Higham, *SIAM Rev.* **2010**, *52*, 696.
[161] E. Estrada, *Biophys. J.* **2010**, *98*, 890.
[162] E. Estrada, *Phys. Rev. E* **2009**, *90*, 0326104.
[163] E. Estrada, H. Natano, *Appl. Math. Comput.* **2009**, *214*, 500.
[164] E. Estrada, H. Natano, *Linear Algebra Appl.* **2009**, *490*, 1886.
[165] E. Estrada, D. J. Higham, H. Natano, *Physica A* **2009**, *388*, 764.
[166] E. Estrada, D. J. Higham, H. Natano, *Phys. Rev. E* **2008**, *78*, 026102.
[167] R. V. Solé, R. Pastor-Satorras, E. D. Smith, T. Kepler, *Adv. Complex Syst.* **2002**, *5*, 1.
[168] V. S. Pande, A. Y. Grosberg, T. Tanaka, *Rev. Mod. Phys.* **2000**, *71*, 259.
[169] H. Taketomi, Y. Ueda, N. Go, *Int. J. Pept. Protein Res.* **1975**, *7*, 445.
[170] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, H. S. Chan, *Protein Sci.* **1995**, *56*, 561.
[171] H. S. Chan, K. A. Dill, *J. Chem. Phys.* **1989**, *90*, 493.
[172] E. Shahknovich, A. Gutin, *J. Chem. Phys.* **1990**, *93*, 5967.
[173] A. Šali, E. Shahknovich, M. Karplus, *Nature* **1994**, *369*, 248.
[174] J. P. Eckmann, S. O. Kamphorst, D. Ruelle, *Europhys. Lett.* **1987**, *4*, 324.
[175] C. L. Weber, P. J. Zbilut, *J. Appl. Physiol.* **1994**, *76*, 965.
[176] A. Giuliani, G. Piccirillo, V. Marigliano, A. Colosimo, *Am. J. Physiol.* **1998**, *257*, H1455.
[177] C. M. Dobson, M. Karplus, *Curr. Opin. Struct. Biol.* **1999**, *9*, 92.
[178] K. T. Simons, C. Strauss, D. Baker, *J. Mol. Biol.* **2001**, *306*, 1191.
[179] A. Giuliani, R. Benigni, J. P. Zbilut, C. L. Webber, Jr., P. Sirabella, A. Colosimo, *Chem. Rev.* **2002**, *102*, 1471.
[180] F. Bloch, *Z. Phys.* **1928**, *52*, 555.
[181] E. Hückel, *Z. Phys.* **1930**, *60*, 423.
[182] E. Hückel, *Z. Phys.* **1931**, *71*, 204.
[183] E. Hückel, *Zeit. Phys.* **1932**, *76*, 628.
[184] C. C. Palliser, D. A. D. Parry, *Proteins: Struct. Funct. Genet.* **2001**, *42*, 243.
[185] D. M. Engelman, T. A. Seitz, *Ann. Rev. Biophys. Chem.* **1986**, *15*, 321.
[186] A. Giuliani, P. Sirabela, R. Benigni, A. Colosimo, *Protein Eng.* **2000**, *13*, 671.

Received: 11 December 2012
Revised: 2 April 2013
Accepted: 22 April 2013
Published online on