Generalized lattice graphs for 2D-visualization of biological information

H. González-Díaz, L.G. Pérez-Montoto, A. Duardo-Sanchez, E. Paniagua, S. Vázquez-Prieto, R. Vilas, M.A. Dea-Ayuela, F. Bolas-Fernández, C.R. Munteanu, J. Dorado, et al.

www.elsevier.com/locate/yjtbi

Generalized lattice graphs for 2D-visualization of biological information

H. González-Díaz, L.G. Pérez-Montoto, A. Duardo-Sanchez, E. Paniagua, S. Vázquez-Prieto, R. Vilas, M.A. Dea-Ayuela, F. Bolas-Fernández, C.R. Munteanu, J. Dorado, J. Costas, F.M. Ubeira

PII: S0022-5193(09)00338-5
DOI: doi:10.1016/j.jtbi.2009.07.029
Reference: YJTBI 5643

To appear in: *Journal of Theoretical Biology*

Received date: 30 April 2009
Revised date: 18 July 2009
Accepted date: 20 July 2009

Cite this article as: H. González-Díaz, L.G. Pérez-Montoto, A. Duardo-Sanchez, E. Paniagua, S. Vázquez-Prieto, R. Vilas, M.A. Dea-Ayuela, F. Bolas-Fernández, C.R. Munteanu, J. Dorado, J. Costas and F.M. Ubeira, Generalized lattice graphs for 2D-visualization of biological information, *Journal of Theoretical Biology*, doi:10.1016/j.jtbi.2009.07.029

Generalized Lattice Graphs for 2D-Visualization of Biological Information

H. González-Díaz[1,*], L.G. Pérez-Montoto[1], A. Duardo-Sanchez[1], E. Paniagua[1], S. Vázquez-Prieto[1], R. Vilas[2], M.A. Dea-Ayuela[3], F. Bolas-Fernández[4], C.R. Munteanu[5], J. Dorado[5], J. Costas[6] and F.M. Ubeira[1]

[1] *Department of Microbiology & Parasitology, and Department of Organic Chemistry, Faculty of Pharmacy and Department of Special Public Law, Financial and Tributary Law Area, Faculty of Law, University of Santiago de Compostela (USC), 15782, Santiago de Compostela, Spain*
[2] *Departament of Genetics, Faculty of Veterinary, USC, 27002, Lugo, Spain.*
[3] *Department of Chemistry, Biochemistry and Molecular Biology, Faculty of Experimental & Health Sciences, University Cardenal Herrera, 46113, Moncada, Valencia, Spain.*
[4] *Department of Parasitology, Faculty of Pharmacy, Complutense University, 28040, Madrid, Spain.*
[5] *Department of Information and Communication Technologies, Computer Science Faculty, University of A Coruña, 15071, A Coruña, Spain.*
[6] *Fundación Pública Galega de Medicina Xenómica, Hospital Clínico Universitario de Santiago, E-15706 Santiago de Compostela, Spain.*

*** Corresponding author:** Humberto González-Díaz, Faculty of Pharmacy, University of Santiago de Compostela 15782, Spain. Fax: +34-981 594912. Email: gonzalezdiazh@yahoo.es or humberto.gonzalez@usc.es

Abstract

Several graph representations have been introduced for different data in theoretical biology. For instance, Complex Networks based on Graph theory are used to represent the structure and/or dynamics of different large biological systems such as protein-protein interaction networks. In addition, Randic, Liao, Nandy, Basak, and many others developed some special types of graph-based representations. This special type of graph includes geometrical constrains to node positioning in space and adopts final geometrical shapes that resemble lattice-like patterns. Lattice networks have been used to visually depict DNA and protein sequences but they are very flexible. However, despite the proved efficacy of new Lattice-like graph/networks to represent diverse systems, most works focus on only one specific type of biological data. This work proposes a generalized type of lattice and illustrates how to use it in order to represent and compare biological data from different sources. We exemplify the following cases: Protein sequence; Mass Spectra (MS) of protein Peptide Mass Fingerprints (PMF); Molecular Dynamic Trajectory (MDTs) from structural studies; mRNA Microarray data; Single Nucleotide Polymorphisms (SNPs); 1D or 2D-Electrophoresis study of protein Polymorphisms and Protein-research patent and/or copyright information. We used data available from public sources for some examples but for other, we used experimental results reported herein for the first time. This work may break new ground for the application of graph theory in theoretical biology and other areas of biomedical sciences.

Keywords: Graph theory; Complex Networks; Proteomics; Mass Spectrometry; Leishmaniosis; 2D Electrophoresis; Parasite population Polymorphism; Single Nucleotide Polymorphism; Schizophrenia; Microarray; Cancer; Patents & Copyright studies.

1. Introduction

Several graph representations have been introduced for different data in theoretical biology. For instance, Complex Networks based on Graph theory are used to represent the structure and/or dynamics of different large biological systems such as protein-protein interaction networks. Complex networks are made up of nodes and edges/arcs (node-node connections or links). Drugs, genes, RNAs, proteins, organisms, brain cortex regions, diseases, patients or environmental systems may play the role of nodes. In general, the edges represent similarity/dissimilarity relationships between the nodes. In Complex Networks, both nodes and edges are placed generally in space without any geometrical constrains; nodes do not need spatial coordinates and edges have not a specific length or shape (Barabasi and Oltvai, 2004; Boccaletti et al., 2006; Estrada, 2006). In addition, Randic, Nandy, Basak, Liao, and many others developed some special types of graph-based representations. This special type of graph includes geometrical constrains to node positioning in space and sometimes adopts final geometrical shapes that resemble lattice-like patterns (Chen et al., 2009; Huang et al., 2009; Liao, 2005; Liao and Wang, 2004; Liao and Ding, 2005; Liao et al., 2005; Liao et al., 2006; Liao et al., 2006 ; Liao et al., 2009; Novic and Randic, 2008; Randic, 2006; Randic and Balaban, 2003; Randic et al., 2007; Randic et al., 2008; Randic et al., 2009; Randič, 2002 ; Randic´ et al., 2005; Zhang et al., 2009).

Using graphical approaches to study biological problems can provide an intuitive picture or useful insights in order to support the analysis of complicated relations within these systems, as demonstrated by many previous studies on a series of important biological topics, such as enzyme-catalyzed reactions (Andraos, 2008; Cornish-Bowden, 1979; Chou, 1980; Chou, 1981; Chou, 1989; Chou and Forsen, 1980; Chou and Liu, 1981; Chou et al., 1979; King and Altman, 1956; Kuzmic et al., 1992; Myers and Palmer, 1985; Zhou and Deng, 1984), protein folding kinetics and folding rates (Chou, 1990; Chou and Shen, 2009; Shen et al., 2009), inhibition kinetics of processive nucleic acid polymerases and nucleases (Althaus et al., 1993a; Althaus et al., 1993b; Althaus et al., 1993c; Althaus et al., 1996; Althaus et al., 1994a; Althaus et al., 1994b; Chou et al., 1994), analysis of codon usage (Chou and Zhang, 1992; Zhang and Chou, 1993; Zhang and Chou, 1994), base frequencies in the anti-sense strands (Chou et al., 1996), analysis of DNA sequence (Qi et al., 2007). Moreover, graphical methods have been introduced for a QSAR study (González-Díaz et al., 2006; González-Díaz et al., 2007b; Prado-Prado et al., 2008) and they have also been used to deal with complicated network systems (Diao et al., 2007; Gonzalez-Diaz et al., 2008b; González-Díaz et al., 2007a). Recently, the "cellular automaton image" (Wolfram, 1984; Wolfram, 2002) has also been applied to study hepatitis B viral infections (Xiao et al., 2006a), HBV virus gene missense mutation (Xiao et al., 2005b), and visual analysis of SARS-CoV (Gao et al., 2006; Wang et al., 2005), as well as in representing complicated biological sequences (Xiao et al., 2005a) and helping to identify various protein attributes (Xiao and Chou, 2007; Xiao et al., 2009; Xiao et al., 2006b). In this study, we attempted to propose a different 2D graphical representation for some relevant areas.

In recent reviews, we have discussed the applications of these ones and other graphs in Proteomics and other Biomedical Sciences (Gonzalez-Diaz, 2008; Gonzalez-Diaz et al., 2008a; González-Díaz et al., 2008). However, despite the proved efficacy of new lattice-like graph/networks to represent diverse systems, most works focus on only one specific type of biological data. This work proposes a generalized type of lattice and illustrates how to use it in order to represent and compare biological data from different sources. Specifically, we extend the method from Protein sequence to Mass Spectra (MS) of Peptide Mass Fingerprints (PMF), Molecular Dynamic (MD) results from protein structural studies, mRNA Microarray data, Single Nucleotide Polymorphisms (SNPs), 1D or 2D-Electrophoresis (2DE) study of protein Polymorphisms and Protein-research patent and/or copyright information.

2. Methods

2.1. Generalized Lattice graphs

Let there be a set of n elements or signals of a biological system, each one identified by a label s_j, and arranged in the form of a sequence of objects or numeric series. For instance, the one-letter code for all bases in a DNA sequence, amino acids in a protein sequence, gene in a chromosome, signals in a Mass Spectrum, values from the microarray data results, etc. First, we arrange all these elements n_j as a vector $\mathbf{s} = [s_1, s_2, s_3, s_j, \ldots s_n]$. Next, we assign to each element a_j one or more up to m properties or weights ($^k w_j$) arranged also as vectors: $^1\mathbf{w} = [^1w_1, {}^1w_2, {}^1w_3, \ldots {}^1w_j, \ldots {}^1w_n]$; $^2\mathbf{w} = [^2w_1, {}^2w_2, {}^2w_3, \ldots {}^2w_j, \ldots {}^2w_n]$; … $^k\mathbf{w} = [^kw_1, {}^kw_2, {}^kw_3, \ldots {}^kw_j \ldots {}^kw_n]$;… and $^m\mathbf{w} = [^mw_1, {}^mw_2, {}^mw_3, \ldots {}^mw_j \ldots {}^mw_n]$. For instance; given an MS we can consider as elements a_j the n signals in the MS and we

can assign at least two weights to each signal a_j: 1) the mass/charge ratio $(m/z)_j$ of s_j and 2) the intensity I_j of s_j. Consequently, we have two weight vectors: ${}^{1}\mathbf{w} = [(m/z)_1, (m/z)_2, (m/z)_3,\ldots (m/z)_j,\ldots (m/z)_n]$ and ${}^{2}\mathbf{w} = [I_1, I_2, I_3,\ldots I_j,\ldots I_n]$. In addition, we can regroup all the elements or signals of the biological system (s_j) into one or more classes (q) if they obey certain sets of conditions $\mathbf{C}_q$. These are usually simple or even composed logical conditions and we assign one letter symbol to all the elements of the same class. For instance, we can label as A, T, G, or C each nucleotide in a sequence if it belongs to the class of Adenine, Thymine, Guanine, or Cytosine. Another example is that we can label as H or L each signal s_j in an MS if the respective intensity value I_j is Higher (H) or Lower than the average of all intensities in the MS. Given all these starting facts, we deal here with the following question. How could we graphically visualize, in a simple way, all the information related to systems (sequences or numeric series), elements or signals, weights or properties, sets of conditions and classes if we have one or, even more complicated, up to i^{th} systems or sequences altogether? Our method assigns each element/signal of one sequence as a point with the Cartesian coordinates $\mathbf{r}_2 = (x, y)$ in a 2D Euclidean space. To this end, we star with the first node (it is not necessarily a data point) at the center of the system placed at $\mathbf{r}_2 = (0, 0)$ coordinate. The coordinates of the successive data points are calculated as follows in a similar manner to those for DNA spaces (Randic, 2004) but extended to multiple weight and condition sets for these weights or properties ${}^{k}w_j$ of the elements or signals s_j.

a) Increases in +1 the y axe if ${}^{k}w_j$ obey the set of conditions $\mathbf{C}_1$ (upwards-step) or:
b) Increases in +1 the x axe if ${}^{k}w_j$ obey the set of conditions $\mathbf{C}_2$ but not $\mathbf{C}_1$ (leftwards-step) or:
c) Decreases in -1 the y axe if ${}^{k}w_j$ obey the set of conditions $\mathbf{C}_3$ but not $\mathbf{C}_1$ nor $\mathbf{C}_2$ (rightwards-step) or:
d) Decreases in -1 the x axe otherwise (downwards-step).

Once we have placed the first sequence or system using the following rules we can superpose over it the remnant q sequences. It allows us the display of large databases in a simple 2D picture. We can use colour-scales highlighting systems or sequences with a given property. For instance, use different colours for enzymes of different classes or for MS signals of the blood samples of healthy *vs.* cancer patients. This type of visual graphs may be interpreted as 2D overlapping or alignment maps. As follows, we give here below some examples to illustrate the high versatility of this approach.

3. Results and Discussion

3.1. Classic lattices for protein and peptide sequences

Several authors have used pseudo-folding lattice Hydrophobicity-Polarity (HP) models to simulate polymer folding by optimizing the lattice structure and resembling the real folding (Berger and Leighton, 1998). However, we can choose notably simpler polymer chain pseudo-folding rules to avoid optimization procedures and speed up notably the construction of the lattice. In this sense, useful graph representations of DNA, RNA and/or protein sequences have been introduced by Gates (Gates, 1986), Nandy (Nandy, 1996a), Leong (Leong and Morgenthaler, 1995), Randic, Balaban, Guo and Basak (Randic et al., 2001) based on 2D coordinate systems. We call these graph representations as polymer sequence pseudo-folding lattice networks because they look like lattice structures and in fact, we force a sequence to fold in a way that does not necessarily occur in nature. In this regard, a novel 2D-lattice representation for protein sequence similar to the one proposed by Nandy for DNA sequences was introduced by our group in the study of protein sequences (Nandy, 1996b; Nandy, 2003; Roy et al., 1998). In this 2D graph, each of the four amino acid groups is assigned to each axis direction according to the physicochemical nature of the amino acids (non-polar and non-charged, polar but non-charged, positively charged, or negatively charged) (Aguero-Chapin et al., 2006). These four classes characterize the physicochemical nature of the amino acids as: polar, non-polar, acid, or basic. Classification as positively or negatively charged prevails over polar/non-polar classification in such a way that the four classes do not overlap each other. In mathematical terms, it means that in this example, we used the vector $\mathbf{s} = [s_0, s_1, s_2,.. s_j,\ldots s_n]$ to list the labels for the *n* amino acids s_j, which are the elements of the system (protein sequence). Here we also used two vectors of weights to characterize numerically the s_j. The first ${}^{1}\mathbf{w} = [q_0, q_1, q_2,\ldots q_j,\ldots q_n]$ lists the electrostatic charge of each one of the *n* amino acids in the sequence of the protein or peptide. The second vector ${}^{2}\mathbf{w} = [\mu_1, \mu_2, \mu_3,\ldots \mu_i, \ldots \mu_n]$ lists the dipolar moments of each amino acid. We also used herein the sets of conditions $\mathbf{C}_1$, $\mathbf{C}_2$, and $\mathbf{C}_3$ that consist of two logic order operations. First, we place the node of the initial amino acid s_0 at the coordinates (0, 0) in a Cartesian 2D space. The coordinates of the successive amino acids are calculated as follows in a similar manner to that for DNA spaces:

$\mathbf{C}_1$: Increases in +1 the y axe if $q_j > 0$ (upwards-step) or:
$\mathbf{C}_2$: Increases in +1 the x axe if $q_j = 0$ and $\mu_j \neq 0$ (rightwards-step) or:
$\mathbf{C}_3$: Decreases in -1 the y axe if $q_j < 0$ (downwards-step) or:
$\mathbf{C}_4$: Decreases in -1 the x axe otherwise (leftwards-step).

The reader must note that the new representation is very similar to the ones previously reported for DNA but it contains a protein sequence of 20 amino acid types instead a DNA sequence of 4 base types. The key of the method we propose to overcome the above-mentioned 10D-space bottleneck is the previous grouping of the twenty natural amino acids into only four groups. As an illustrative example, we have under study a protein that belongs to the family of dyneins. This protein has the accession number LmjF25.0980 in the public database GenDB related to the Sanger institute (http://www.genedb.org/genedb/) and it is expressed by the parasite *Leishmania major*. *Leishmania spp.* is required intracellular protozoa that exist in two forms, a promastigote form (elongated cells with a long flagellum) and an amastigote one (ovoid cells that have a very short flagellum). The flagellum is responsible for the motility of trypanosomatids and for their early interaction with the hosts, either by adhering to the insect digestive tract, or by initiating the contact with mammalian cells. Trypanosomatids depend on this adhesion to survive and differentiate. This surface organelle plays a key role in *Leishmania* motility and sensory reception, and it is essential for parasite migration, invasion and persistence in host tissues. In this regard, some authors have applied lattice representations to study dyneins (Dea-Ayuela et al., 2008). Due to both the high interest of dyneins for the mechanism involved in protein-protein interaction or binding process, some authors have proposed experimental studies of peptide sets found in these proteins (Lajoix et al., 2004). **Figure 1 (A)** illustrates the isolated and overlapped lattice graphs only for the first two peptides found in this protein presented in **Table 1**. In this table we give the sequence of these peptides and other relevant information (see also next section). We used the first peptide (P01) with the sequences "vlmntlrdir" as example, where the vectors are $\mathbf{s} = [v_0,l_1,m_2,n_3,t_4,l_5,r_6,d_7,i_8,r_9]$, ${}^1\mathbf{w} = [0_0, 0_1, 0_2, 0_3, 0_4, 0_5, 1_6, -1_7, 0_8, 1_9]$ and ${}^2\mathbf{w} = [0_0, 0_1, 0_2, 1_3, 0_4, 0_5, 1_6, 1_7, 0_8, 1_9]$. The vector ${}^1\mathbf{w}$ is based on amino acid net charges and the vector ${}^2\mathbf{w}$ is based on the discrete dipole moments (see **Table S1** from the Supplementary material for more details).

Table 1 comes about here

3.2. Lattices for MD outcomes

The 3D structure of the *L. major* dynein protein sequence represented in the previous example is unknown; which is the case for many other proteins nowadays. In this sense, and taking into consideration the issues discussed in the previous section, the study of the 3D structure of its component peptides is of major interest. Since the advent of MD in bioscience with the study carried out by McCammon *et al.* on the dynamics of the bovine pancreatic trypsin inhibitor, MD has become the by the foremost a well-established computational technique to investigate the 3D structure and function of peptides and proteins (Karplus and McCammon, 2002; McCammon et al., 1977). Consequently, MD studies of peptides of the template protein used in the previous example are also interesting. In general, the analysis of the MD-Trajectories (MDTs) resulting from the integration of the motion equations in MD remains, however, the greatest challenge and requires a great deal of insight, experience, and effort. In a recent and very important work, Hamacher (Hamacher, 2007) has proposed a new, theoretical sound, and versatile analysis procedure that provides scientists with a semi-quantitative tool to compare various scenarios of their respective simulations. In this regard, we extended the lattice representations of proteins to allow easy comparison of MDTs. In **Figure 2** we illustrate an example that consists of the superposition or 2D-Alignment of 18 lattices derived from 100-steps MDTs results. Each MDT was obtained after a Monte Carlo study of 18 peptides found on the PMF of a very important parasite protein. In **Table 1** we summarize some details on the MD study used here as example. The key of the method we propose is the regrouping into four classes the Energy values E_j obtained from different steps (s) of one MD trajectory obtained in peptide structure study with the Monte Carlo method. These four classes characterize the deviation of the energy value E_j from the average energy of the same MDT at different steps (MD-average); or the deviation from average energy values in the same step for other MDTs (Step-average).

First, we place the values of energy for a MDT in a Cartesian 2D space starting with s_0 at the coordinates (0, 0). In this example, we used the vector $\mathbf{s} = [s_0, s_1, s_2,.. s_j,\ldots s_{100}]$ to list the labels for optimization steps s_j in the MDT, which are the signals or elements of the system. Herein, we also used three vectors of weights to numerically characterize s_j. The first ${}^1\mathbf{w} = [E_0, E_1, E_2,\ldots E_j,\ldots E_{100}]_{n=101}$ lists the energy values for each s_j in the MDT numeric

sequence (of one peptide). The second vector $^{2}\mathbf{w} = [^{avg}E_1, {}^{avg}E_2, {}^{avg}E_3,\ldots {}^{avg}E_i, \ldots {}^{avg}E_9]$ lists the average of E_j for each one of the i^{th} MDT for the i^{th} peptides. The third vector $^{3}\mathbf{w} = [^{avg}E_0, {}^{avg}E_1, {}^{avg}E_2,\ldots {}^{avg}E_j, \ldots {}^{avg}E_{100}]$ lists the average of E_j for each one of the j^{th} steps of all MDT for all peptides. We also used the sets of conditions $\mathbf{C}_1$, $\mathbf{C}_2$, and $\mathbf{C}_3$ that consist of two logical order operations. These operations perform the comparison (">" or "<") with respect to the average values MDT-average and Step-average:

$\mathbf{C}_1$: Increases in +1 the y axe; if $E_j > {}^{avg}E_j$ and $E_j > {}^{avg}E_i$ (upwards-step) or:

$\mathbf{C}_2$: Increases in +1 the x axe; if $E_j > {}^{avg}E_j$ and $E_j < {}^{avg}E_i$ (rightwards-step) or:

$\mathbf{C}_3$: Decreases in -1 the y axe; if $E_j < {}^{avg}E_j$ and $E_j < {}^{avg}E_i$ (downwards-step) or:

$\mathbf{C}_4$: Decreases in -1 the x axe; otherwise (leftwards-step).

In **Figure 1 (B),** we depict the 2D alignment for MD results obtained after the optimization of the structure of 18 peptides and successive Monte Carlo search of different conformations. Remarkably, the MDT lattice obtained for peptide P03 notably deviates from the other peptides while the graph for P15 lies in the middle of the rest of peptides. It may indicate that this type of lattice is useful to differentiate visually peptides with high initial energy after MD geometry optimization (E_1) and not optimal MDT from the rest of peptides (see **Table 1**). The reader may note the differences between sequence and MD lattice graphs; which indicates that both types of graphs may be used as complementary information visualization techniques.

3.3. Lattices for MS of Peptide Mass Fingerprints

The study of peptides found on the PMFs of new proteins may become an interesting source to discover new peptides with potential use as drug, in vaccine design, or as disease biomarkers. In particular, toxicity and inefficacy of actual organic drugs against Leishmaniosis justify research projects to find new drugs or drug molecular targets in *Leishmania* species including *L. infantum* and *L. major*, both important pathogens (Chenik et al., 2006; Dea-Ayuela et al., 2008; Roldos et al., 2008; Sarciron et al., 2005). In the two previous examples, we used lattices to study the sequences and MD results of peptides found in a dynein of *L. major*. In this example, we propose to construct PMF lattices, in analogy to sequences and MDTs lattices. To this end, we use a real experiment as example. We isolate all the peptides found on the PMF of a protein expressed on the parasite *L. infantum* with 2DE and characterize them with MALDI-TOF MS. After a MASCOT search of similar PMF-MS, we found that this new protein is similar to the protein of *L. major* studied in the previous examples. In **Figure 2 (A),** we illustrate the 2DE map experimentally obtained and highlight the position of the spot for the new protein. In **Table 1** we give details on the $(m/z)_j$ values for the peptides found on the PMF of the new protein. We report the experimental study of this protein for the first time but the method used is essentially the same we had used before for other dynein protein. That is why we omit the experimental details in this work and refer to the previous work (Dea-Ayuela et al., 2008).

Figure 2 comes about here

Next, we report the generation of the 2D lattice graphs for large MS data generated in PMF experiments. The idea of using the graph to study MS is a promising field of research. Bartels proposed for the first time the application of graph theory to MS for peptide sequencing (Bartels, 1990). The fundamental idea consists in transforming an MS into a graph called the *spectrum graph*, each peak in the experimental spectrum being represented as a graph node (or several nodes). Directed edges (or arc) connect between two vertices if the mass difference of the two vertices equals the mass of one or several amino acids. "SeqMS" (Fernandez-de-Cossio et al., 1995), "Lutefisk" (Taylor and Johnson, 1997), "Sherenga" (Dancík et al., 1999) and "PepNovo" (Frank and Pevzner, 2005) are the most popular algorithms that make use of spectrum graphs based on the basic idea proposed by Bartels.

In our lattice graph, MS signals are placed in a Cartesian 2D space starting with the first data point at the coordinate $\mathbf{r}_2 = (0, 0)$. The coordinates of the successive data points are calculated as follows. In this example, we used the vector $\mathbf{s} = [s_1, s_2, s_3,.. s_j,\ldots s_{68}]$ to list the labels for the 68 MS signals s_j of the new protein. We also used four vectors of weights; the first $^{1}\mathbf{w} = [^{1}(m/z)_j] = [^{1}(m/z)_1, {}^{1}(m/z)_2,\ldots {}^{1}(m/z)_j,\ldots {}^{1}(m/z)_{68}]$ lists the mass/charge ratio values for each s_j. The other three vectors are: $^{2}\mathbf{w} = [^{2}(m/z)_1, {}^{2}(m/z)_2,\ldots {}^{2}(m/z)_{j \le 68}]$; $^{3}\mathbf{w} = [^{3}(m/z)_1, {}^{3}(m/z)_2,\ldots {}^{3}(m/z)_j,\ldots {}^{3}(m/z)_{j \le 68}]$; and $^{4}\mathbf{w} = [^{4}(m/z)_1, {}^{4}(m/z)_2,\ldots {}^{4}(m/z)_j,\ldots {}^{4}(m/z)_{j \le 68}]$. These vectors list, in an increasing order, the $(m/z)_j$ values for s_j also present in the MS of the three most similar template proteins found after the MASCOT search. In order to generalize the procedure, we can refer to the vectors: $^{k+1}\mathbf{w} = [^{k+1}(m/z)_j]$, $^{k+2}\mathbf{w} = [^{k+2}(m/z)_j]$, and $^{k+3}\mathbf{w} = [^{k+3}(m/z)_j]$. These vectors list the $(m/z)_j$ values of the three proteins (k^{th}-triad)

placed at positions k, k + 1, and k +2 in the list of template proteins, found after the MASCOT search, ordered from higher to lower similarity to the query protein . Next, we can use a set of conditions $\mathbf{C}_1$, $\mathbf{C}_2$, and $\mathbf{C}_3$ to align many triads and detect the similarity patterns.

$\mathbf{C}_1$: Increases in +1 the y axe if $^{1}(m/z)_j \in {}^{k}\mathbf{w}$ and $\notin {}^{k+1}\mathbf{w}$ and $\notin {}^{k+2}\mathbf{w}$ (upwards-step) or:

$\mathbf{C}_2$: Increases in +1 the x axe if $^{1}(m/z)_j \notin {}^{k}\mathbf{w}$ and $\in {}^{k+1}\mathbf{w}$ and $\notin {}^{k+2}\mathbf{w}$ (rightwards-step) or:

$\mathbf{C}_3$: Decreases in -1 the y axe if $^{1}(m/z)_j \notin {}^{k}\mathbf{w}$ and $\notin {}^{k+1}\mathbf{w}$ nor $\in {}^{k+2}\mathbf{w}$ (downwards-step).

$\mathbf{C}_4$: Decreases in -1 the x axe if otherwise (leftwards-step).

In **Figure 1 (C),** we depict the alignment of these types of lattice graphs for the query protein *vs.* 17 triads found with MASCOT in the template database. It is relevant that the method perfectly discriminates the alignment (black colour) with the best triad (more similar proteins), with respect to triads formed by other less similar or dissimilar proteins (gray colour). Last, we can apply alternatively and somehow complementary operations ${}^{*}\mathbf{C}_1$, ${}^{*}\mathbf{C}_2$, ${}^{*}\mathbf{C}_3$ and ${}^{*}\mathbf{C}_4$ if our aim is the study of the s_j in the query protein that does not match up with any template protein.

${}^{*}\mathbf{C}_1$: Increases in +1 the y axe if $^{1}(m/z)_j \notin {}^{2}\mathbf{w}$ (upwards-step) or:

${}^{*}\mathbf{C}_2$: Increases in +1 the x axe if $^{1}(m/z)_j \notin {}^{2}\mathbf{w}$ and $\notin {}^{3}\mathbf{w}$ neither (rightwards-step) or:

${}^{*}\mathbf{C}_3$: Decreases in -1 the y axe if $^{1}(m/z)_j \notin {}^{2}\mathbf{w}$ and $\notin {}^{3}\mathbf{w}$ and $\notin {}^{4}\mathbf{w}$ neither (downwards-step).

${}^{*}\mathbf{C}_4$: Decreases in -1 the x axe if otherwise (leftwards-step).

Note that the first lattice graph based on $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_3$ and $\mathbf{C}_4$ plots the MS signals present in both the query protein and at least one of the triad of template proteins selected. Consequently, this graph gives us a visual idea on how similar our query protein is with respect to the known template proteins (like in BLAST). Conversely, the second type of graph based on ${}^{*}\mathbf{C}_1$, ${}^{*}\mathbf{C}_2$, ${}^{*}\mathbf{C}_3$ and ${}^{*}\mathbf{C}_4$ plots precisely those MS signals that do not match up with MS signals found on the triad of template proteins. In consonance, this graph may give an idea on how dissimilar this protein is and then how useful it may be to decide an investigation of unknown peptides.

3.4. Lattices for Mass Spectra of Proteins Serum Profiles (PSP-MS)

In the previous paragraphs, we have introduced 2D lattice graph representations for DNA/protein sequences, MDT results, and PMF-MS experiments. Now, we report the generation of the 2D lattice graphs for large MS data generated in PSP-MS experiments with blood samples. Blood proteome is continuously changing due to the effect of the drug-induced damage in the affected organ. After the separation of the small peptide fragments from the actual insult, the remaining mixture of peptides retains the specificity of the disease due to the specific biomarker amplification process in a unique tissue microenvironment in the organ where the toxicity occurs (Hu et al., 2006). Therefore, we can use the serum, the saliva, or the urine because they are protein-rich information reservoirs containing blood traces (Hu et al., 2006). In addition, it is well-known the optimal performance in the low mass range demonstrated by the mass spectroscopy (Kantor, 2002; McDonald and Yates, 2002) applied to proteomics by offering the great chance of discovering these early stage composition changes. The main problems in the identification of a single disease-related protein are the following: there are thousands of intact and cleaved proteins in the PSP that require the separation and identification of each protein biomarker and most toxicity biomarkers appear only when significant organ damage has occurred. Thus, the pattern identification in PSP-MS becomes a realistic complementary approach compared with the direct identification of a single marker candidate. Consequently, we can state that PSP-MS may allow detecting disease biomarkers at the first stages. In this regard, the development of new graph representations becomes significant to visually depict interesting similarity/dissimilarity patterns between PSP-MS of different groups of patients. In a recent work, we have introduced novel Randic's Spiral network representation of PSP-MS (Cruz-Monteagudo et al., 2008a). Other example is the previous theoretical study of Human Prostate Cancer with new graph representations (Ferino et al., 2008). In previous works, our group has extended for the first time the Spiral, Star (Cruz-Monteagudo et al., 2008b), and Lattice (Petricoin et al., 2004) graphs to represent the PSP-MS with a very high number of intensity (I_j) signals and wide $(m/z)_j$ bandwidth. As these types of graphs had been studied before, we do not depict a PSP-MS lattice here, by reasons of space. However, we give as follows a mathematical formalization of this type of graph in order to generalize them and show more possibilities to codify information in PSP-MS experiments. For it, each signal in the MS is placed in a Cartesian 2D space starting with the first data point at the coordinate $\mathbf{r}_2$ = (0, 0). The coordinates of the successive data points are calculated by using the following mathematical formalism. Although the binned process reduces efficiently the number of data points, it is still unmanageable for

graph generation. Hence, the number of data points in the binned data files was condensed by taking the averaged ${}^{*}(m/z)_j$ and ${}^{*}I_j$ values for consecutive regions containing a fixed number n^* of $(m/z)_j$ and I_j data points. The value n^* may be changed according to the interest of the research; in particular when we keep $n^* = 1$ the number of averaged regions ${}^{*}s_j$ is equal to the number of original signals s_j. In this example, we used the vector $\mathbf{s} = [s_1, s_2, s_3,.. s_j,\ldots s_{n*}]$ to list the labels for the MS signals s_j of the new protein. We also used two vectors of weights; the first ${}^{1}\mathbf{w} = [{}^{*}(m/z)_j] = [{}^{*}(m/z)_1, {}^{*}(m/z)_2,\ldots {}^{*}(m/z)_j,\ldots {}^{*}(m/z)_{*n}]$ list the average mass/charge ratio values for each region ${}^{*}s_j$ out of ${}^{*}n$ altogether. The other vector lists in a similar way the average intensity values ${}^{2}\mathbf{w} = [{}^{*}I_1, {}^{*}I_2,\ldots {}^{*}I_j,\ldots {}^{*}I_{*n}]$. We also used the sets of conditions $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_3$ and $\mathbf{C}_4$:

$\mathbf{C}_1$: Increases in +1 the y axe if ${}^{*}(m/z)_j > 0.5$ and ${}^{*}I_j > 0.5$ for ${}^{*}s_j$ (upwards-step) or:

$\mathbf{C}_2$: Increases in +1 the x axe if ${}^{*}(m/z)_j > 0.5$ and ${}^{*}I_j < 0.5$ for ${}^{*}s_j$ (rightwards-step) or:

$\mathbf{C}_3$: Decreases in -1 the y axe if ${}^{*}(m/z)_j < 0.5$ and ${}^{*}I_j < 0.5$ for ${}^{*}s_j$ (downwards-step).

$\mathbf{C}_4$: Decreases in -1 the x axe if otherwise (leftwards-step):

3.5. Lattices of Protein Polymorphisms Determined by Electrophoresis

Different electrophoresis such as: immunofixation electrophoresis, capillary electrophoresis, 2D-gel electrophoresis or 2DE, are used to characterize protein polymorphism in populations (Alper and Johnson, 1969; Hadi et al., 1998; Kanamori-Kataoka and Seto, 2009; Lopez-Galvez et al., 1995). In fact, the amount of protein variation undetected by electrophoresis may be reasonably small and at the protein level, a typical sexually-reproducing organism may be heterozygous at 20 or more percent of the gene loci. Although the evidence is limited, it seems that at the level of the DNA nucleotide sequence every individual is heterozygous at every locus - if introns as well as exons are taken into account (Ayala, 1983). In the present example, we characterized experimentally for the first time the polymorphism for 17 enzymes in three populations of *Fasciola hepatica* (*F. hepatica*). The parasite *F. hepatica* is the causal agent of fasciolosis infection, an important cause of lost productivity in livestock worldwide. Effective control of fasciolosis is difficult, especially in milking cows, which can only be treated during dry periods, a control strategy that has not been evaluated yet. Recently, our group has studied the effect of the type of flukicide treatment on the prevalence and intensity of infection in dairy cattle from Galicia, an area where fasciolosis is endemic and which is also the main milk-producing region in Spain (Mezo et al., 2008). In the present preliminary study, we found that 8 loci out of 17 studied presented polymorphisms expressing up to 3 different isoforms of the enzyme. The polymorphic enzymes were: Aconitate Hydratase or Aconitase (ACO), Adenilate Kynase (AK), Glutamate Oxaloacetate Transaminase (GOT), Hexokinase (HK), Isocitrate Dehydrogenase (IDH), Phosphogluconate Dehydrogenase (PGD),: Phosphoglucomutase 1 (PGM1), and Phosphoglucomutase 2 (PGM2). Considering that: 1) heterozygous organisms are common, 2) each protein has one or even two locus, and 3) each locus may present one out of two or more alleles; we can construct large databases with the information obtained by electrophoresis for individuals in different populations. This situation determines the necessity of the use of computational techniques. Actually, the necessity of the use of computational techniques for phenotypic analysis in adults and eggs of *F. hepatica* has been recently proposed by Valero and Panova *et al* (Valero et al., 2005).

In this regard, the present type of data is another interesting candidate to be studied with lattice graphs. Consequently, in this example we need 1 lattice graph for each parasite individual with 8 enzymes. Altogether, each of the 6 enzymes is encoded by 1 gene; which presents 2 loci that may express 1 out of 3 possible isoforms of the enzyme. In addition, one of the enzymes PGM is codified by two different gene producing two different proteins, PGM1 and PGM2. We used the vectors ${}^{a}\mathbf{s} = [{}^{a}s_1, {}^{a}s_2, {}^{a}s_3,.. {}^{a}s_j,\ldots {}^{a}s_8]$ and ${}^{b}\mathbf{s} = [{}^{b}s_1, {}^{b}s_2, {}^{b}s_3,.. {}^{b}s_j,\ldots {}^{b}s_8]$ to list the labels for the two possible alleles ${}^{a}s_j$ and ${}^{b}s_j$ for two enzyme isoforms A and B codified by a gen. Commonly, we use the magnitude called Retention factor (Rf_j) to characterize each signal in 1D electrophoresis; which measures the chromatographic displacement distance (electrophoresis in this case) of the band from the point of application. In **Figure 2 (B),** we illustrate the 1D electrophoresis bands for one individual parasite experimentally characterized in this work. Herein, we used an integer-value scale for $Rf_j = 1$, 2, or 3 for the bands with lower absolute displacement, second higher displacement, etc. Then, we also used two vectors of weights; the first ${}^{a}\mathbf{w} = [{}^{a}Rf_j] = [{}^{a}Rf_1, {}^{a}Rf_2,\ldots {}^{a}Rf_j,\ldots {}^{a}Rf_8]$ and the second ${}^{b}\mathbf{w} = [{}^{b}Rf_j] = [{}^{b}Rf_1, {}^{b}Rf_2,\ldots {}^{b}Rf_j,\ldots {}^{b}Rf_8]$. These vectors list the first ${}^{1}Rf_j$ and the second ${}^{2}Rf_j$ values of Rf for two isoenzymes codified by the same gene. In general, when ${}^{1}Rf_j \neq {}^{2}Rf_j$ the organism is heterozygote for this character (enzyme) it means that the two loci of the gene codify different alleles. In this case we assign the lower absolute value of retention factor (${}^{1}Rf_j > {}^{2}Rf_j$) to the first vector

$^{1}\mathbf{s}$ the enzyme. Otherwise, $^{1}Rf_j > {}^{2}Rf_j$ the two alleles of the gen produce the same enzyme and the organism is homozygote for this character. We can use the following sets of conditions $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_3$, and $\mathbf{C}_4$ to obtain the lattice graph for $^{1}\mathbf{s}$ or $^{2}\mathbf{s}$ separately. In particular, $\mathbf{C}_4$ refers to cases when the Rf_j could not be accurately determined and the genotypic information is not clear; the condition sets are as follows:

$\mathbf{C}_1$: Increases in +1 the y axe if $^{k}Rf_j = 1$ (upwards-step) or:

$\mathbf{C}_2$: Increases in +1 the x axe if $^{k}Rf_j = 2$ (rightwards-step) or:

$\mathbf{C}_3$: Decreases in -1 the y axe if $^{k}Rf_j = 3$ (downwards-step).

$\mathbf{C}_4$: Decreases in -1 the x axe otherwise (leftwards-step).

However, both alleles are determinant in the polymorphism. Therefore, it is more interesting to generate graphical plots for one individual containing both alleles at the same time. In this regard, we extended our mathematical formalism as follows (see the previous example on SNPs). Let there be, $^{c}\mathbf{s} = {}^{a}\mathbf{s}\cup{}^{b}\mathbf{s} = [{}^{a}s_1, {}^{a}s_2, {}^{a}s_3,..{}^{a}s_j,\ldots{}^{b}s_n, {}^{b}s_1, {}^{b}s_2, {}^{b}s_3,..{}^{b}s_j,\ldots{}^{b}s_n]$ the vector that list the labels of two possible alleles of n genes and $^{c}\mathbf{w} = {}^{a}\mathbf{w}\cup{}^{b}\mathbf{w} = [{}^{1}Rf_1, {}^{1}Rf_2,\ldots {}^{1}Rf_j,\ldots{}^{1}Rf_8, {}^{2}Rf_1, {}^{2}Rf_2,\ldots{}^{2}Rf_j,\ldots{}^{2}Rf_8]$; we can use it to characterize the polymorphism of one individual in the following manner. We can refer to **s** and **w** as genotypic-polymorphism vectors and apply the same rules outlined above. **Figure 3** depicts the $^{c}\mathbf{s}$-alignment of all individuals belonging to different populations using the same above-mentioned rules $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_3$ and $\mathbf{C}_4$.

Figure 3 comes about here

In any case, as the elements $^{a}s_i$ and $^{b}s_j$ for i = j are haplotypes of the same loci that codify the same enzyme, it is easier to list them successively. In this regard, it is probably easier to use the vectors $\mathbf{s}_c = {}^{a}\mathbf{s}\cap{}^{b}\mathbf{s} = [{}^{a}s_1, {}^{b}s_1, {}^{a}s_2, {}^{b}s_2, {}^{a}s_3, {}^{b}s_3, {}^{a}s_j, {}^{b}s_j,\ldots{}^{a}s_n, {}^{b}s_n]$ and $\mathbf{w}_c = {}^{a}\mathbf{w}\cap{}^{b}\mathbf{w} = [{}^{1}Rf_1, {}^{2}Rf_1, {}^{1}Rf_2, {}^{2}Rf_2,\ldots{}^{1}Rf_j, {}^{2}Rf_j,\ldots{}^{1}Rf_8, {}^{2}Rf_8]$ to list the labels and weights of two possible alleles of n genes. In terms of computational cost, both procedures are equivalent but, with respect to facilitating data input, the $\mathbf{s}_c$ vectors are more user-friendly. In any case, it is important to note that, in general, the lattice graph is different for $^{c}\mathbf{s}$ and $\mathbf{s}_c$ schemes; which may offer alternative solutions to the same problem. In **Figure 2 (B),** we also give examples of lattice graphs for one individual parasite using $^{c}\mathbf{s} = {}^{a}\mathbf{s}\cup{}^{b}\mathbf{s}$ or $^{c}\mathbf{s} = {}^{a}\mathbf{s}\cap{}^{b}\mathbf{s}$ as alternative schemes. On the other hand, if you are interested not in the characterization of the polymorphism of individuals within a population but in specific enzymes in different individuals you have to invert the previous approach using one vector of labels to list individuals and vectors of weights to characterize a haplotypes of the specific a enzyme in different individuals.

Last, in other type of electrophoresis methods such as 2D electrophoresis (see **Figure 2** for instance) the different proteins are characterized by Mass (Mj) and isoelectric point (pI_j) instead of only one Rf_j value (Dea-Ayuela and Bolás-Fernández, 2005). In these cases, we may use one label vector **s**, two weight vectors $^{1}\mathbf{w} = [M_j]$ and $^{2}\mathbf{w} = [pI_j]$, and cut-off values δ_1 and δ_2, to define condition sets $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_3$ and $\mathbf{C}_4$ similar to those used for PSP-MS above:

$\mathbf{C}_1$: Increases in +1 the y axe if $M_j > \delta_1$ and $pI_j > \delta_2$ (upwards-step) or:

$\mathbf{C}_2$: Increases in +1 the x axe if $M_j > \delta_1$ and $pI_j < \delta_2$ (rightwards-step) or:

$\mathbf{C}_3$: Decreases in -1 the y axe if $M_j < \delta_1$ and $pI_j < \delta_2$ (downwards-step).

$\mathbf{C}_4$: Decreases in -1 the x axe otherwise (leftwards-step).

3.6. Lattices of Single Nucleotide Polymorphisms (SNPs)

In the previous paragraphs, we have introduced 2D lattice graph representations for DNA/protein sequences, MDT results, and MS outcomes. Now, we report the generation of the 2D lattice graphs for large SNP of schizophrenia patients. The HTR2A and DRD3 genes codify protein receptors for the biogenic amine serotonin (5-HT) and dopamine (DA) neurotransmitter; which are the primary targets of the antipsychotic drugs in the schizophrenia treatment (Meltzer et al., 1989). The silent SNP T102C (rs6313) at HTR2A as well as the non-synonymous SNP Ser9Gly (rs6280) at DRD3 have been extensively analysed in schizophrenia case-control studies (Abdolmaleky et al., 2004; Jonsson et al., 2003). In this example we aim to study a SNPs database based on the 17bp-long and 32bp-long SNPs of the DRD3 and HTR2A gene respectively, from 260 schizophrenic patients and 354 control subjects (Dominguez et al., 2007). These SNPs are codified with the following haplotypes: 0 if the first allele is homozygous, 1 if heterozygous, 2 if the second allele is homozygous and 3 if it is unknown. As a result, we have a large amount of information contained in raw data of 614 patients with 17 inputs each one making 17 x 614 = 10,438 input data points altogether for DRD3 gene or 32 x 614 = 19,648 input

data points altogether for HTR2A gene. In this type of cases, the use of simple graph methods may be very interesting in order to perform a fast visualization of the large database. In addition, the alignment of superposition in some way of all inputs may unravel hidden patterns of similarity/dissimilarity between all patients. In any case, to the best of our knowledge, no lattice graph has been reported to represent and perform 2D alignment of SNPs in schizophrenia patients.

In this example, the sequences of the SNPs genotype information are transformed into lattice graphs using the following mathematical formalism. To this end, each nucleotide in the 32-bp SNPs sequence of one patient was placed as a node in a Cartesian 2D space starting with the first data point at the coordinate $\mathbf{r}_2 = (0, 0)$. The coordinates of the successive data points are calculated as follows. In this example, we used the vector $^1\mathbf{s} = [^1s_1, {}^1s_2, {}^1s_3,.. {}^1s_j,\ldots {}^1s_{17}]$ and $^2\mathbf{s} = [^2s_1, {}^2s_2, {}^2s_3,..{}^2s_j,\ldots {}^2s_{32}]$ to list the labels for each of the 17 signals s_j for gene DRD3 or 32 signals for gene HTR2A of one patient. We also used two vectors of weights; the first $^1\mathbf{w} = [^1h_j] = [^1h_1, {}^1h_2,\ldots {}^1h_j,\ldots {}^1h_{17}]$ and the second $^2\mathbf{w} = [^2h_j] = [^2h_1, {}^2h_2,\ldots {}^2h_j,\ldots {}^2h_{32}]$ to list the haplotype types ^1h_j or $^2h_j = 0, 1, 2,$ or 3 for each s_j of DRD3 or HTR2A gene respectively. We also used the following sets of conditions $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_3$, and $\mathbf{C}_4$:

$\mathbf{C}_1$: Increases in +1 the y axe if $^1h_j = 0$ (upwards-step) or:

$\mathbf{C}_2$: Increases in +1 the x axe if $^1h_j = 1$ (rightwards-step) or:

$\mathbf{C}_3$: Decreases in -1 the y axe if $^1h_j = 2$ (downwards-step).

$\mathbf{C}_4$: Decreases in -1 the x axe if otherwise (leftwards-step).

We may apply these conditions to the vectors $^1\mathbf{s}$ and $^1\mathbf{w}$ in order to obtain a 2D alignment of all SNPs for all patients by using gene DRD3 or HTR2A separately. However, it has been admitted that both genes were involved in schizophrenia so the generation of graphical plots for both sets of SNPs is interesting. In this regard, we extended our mathematical formalism as follows. Let there be, $^n\mathbf{s} = {}^1\mathbf{s} \cup {}^2\mathbf{s}\ldots\cup {}^n\mathbf{s} = [^2s_1, {}^2s_2, {}^2s_3,..{}^2s_j,\ldots {}^2s_{n1}, {}^2s_1, {}^2s_2, {}^2s_3,..{}^2s_j,\ldots {}^2s_{n2}, \ldots {}^2s_1, {}^2s_2, {}^2s_3,..{}^2s_j,\ldots {}^2s_{nn}]$ the vector that lists the labels of all SNPs for n genes with n1, n2, …nn SNPs and $^n\mathbf{w} = {}^1\mathbf{w} \cup {}^2\mathbf{w}\ldots\cup {}^n\mathbf{w} = [^2h_1, {}^2h_2, {}^2h_3,..{}^2h_j,\ldots {}^2h_{n1}, {}^2h_1, {}^2h_2, {}^2h_3,..{}^2h_j,\ldots {}^2h_{n2}, \ldots {}^2h_1, {}^2h_2, {}^2h_3,..{}^2h_j,\ldots {}^2h_{nn}]$; we can consider it as a single list of SNPs for n genes instead of only one. We can refer to $^n\mathbf{s}$ and $^n\mathbf{w}$ as partial or total chromosome vectors if they incorporate all gene in the same chromosome or only some of them. We can refer to $^n\mathbf{s}$ and $^n\mathbf{w}$ as ordered if the order of union of vectors is the same as in the original chromosome. Last, $^n\mathbf{s}$ and $^n\mathbf{w}$ are mixed and/or disordered if they assemble vectors coming from different chromosomes and/or in another order, different from the one specific to natural chromosome order. Last, the vectors $^n\mathbf{s}$ and $^n\mathbf{w}$ generated with all the genes of an organism may be classified as hole-SNPs genome vectors. In this example we can construct the vectors: $^2\mathbf{s} = {}^1\mathbf{s} \cup {}^2\mathbf{s} = [^2s_1, {}^2s_2, {}^2s_3,..{}^2s_j,\ldots {}^2s_{17}, {}^2s_1, {}^2s_2, {}^2s_3,..{}^2s_j,\ldots {}^2s_{32}]$ and $^2\mathbf{w} = {}^1\mathbf{w} \cup {}^2\mathbf{w} = [^2h_1, {}^2h_2, {}^2h_3,..{}^2h_j,\ldots {}^2h_{17}, {}^2h_1, {}^2h_2, {}^2h_3,..{}^2h_j,\ldots {}^2h_{n2}]$ that list the labels and weights for SNPs of a patient and the two genes DRD3 and HTR2A at the same time. **Figure 4, A** depicts the 2D-alignment of all these vectors for all patients using the same above-mentioned rules $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_3$ and $\mathbf{C}_4$. We can note that both groups overlap notably in the lattices of schizophrenia patients (in grey), by expanding leftwards to areas not covered by healthy patients, but there are no significant results because this region is the consequence of the unknown allele-type component ($^1h_j = 3$). The area with interesting overlap differences that can be used to perform further research to find SNPs biomarkers for schizophrenia is the upper part of the lattice that is generated by the homozygous allele ($^1h_j = 0$). The graph allowed us to depict visually 17 x 614 + 32 x 614 = 10,438 + 19,648 = 30,086 SNPs points for 614 healthy *vs.* schizophrenia patients in a single 2D graph.

Figure 4 comes about here

3.7. Lattices for mRNA Microarrays

In the previous sections, we have introduced 2D lattice graph representations for DNA/protein sequences, MDT results, MS outcomes, and SNPs. In this example, we introduce 2D lattice graphs for the results obtained in mRNA microarray experiments. Microarrays have been used to find gene expression patterns with special relevance as molecular biomarkers for different diseases including cancer. Specifically, Human Breast Cancer (HBCa) is the most common neoplasia in women since approximately 211,000 women are diagnosed with it annually in the United States. In spite of earlier detection and improved treatment, it remains the second leading cause of cancer-related death in the United States and in other developed countries. The genetic background of patients and the tumor's genetic and epigenetic anomalies create, in combination, molecularly distinct subtypes arising from distinct cell types within the ductal epithelium. This genetic complexity underlies the clinical

heterogeneity of HBCa limiting a rational selection of treatment tailored to individual patient/tumor characteristics. In this regard, Modlich and Prisack *et al.* (Modlich et al., 2005) published a very interesting study whose declared goal was to identify gene signatures predictive of response to preoperative systemic chemotherapy (PST) with epirubicin/cyclophosphamide in patients with primary HBCa. The authors obtained pre-treatment needle biopsies from 83 patients with breast cancer and profiled mRNA on Affymetrix HG-U133A arrays. Response ranged from pathologically confirmed Complete Remission (pCR), to partial remission (PR), to stable or progressive disease, "No Change" (NC). A primary analysis was performed in breast tissue samples from 56 patients and 5 normal healthy individuals as a training cohort for predictive marker identification. The high complexity of this dataset makes these results another interesting candidate to be visually depicted with lattice graphs. In addition, the 2D alignment or superposition of all inputs may unravel hidden patterns of similarity/dissimilarity between all patients. In any case, we have not found a previous report using lattice graphs to represent and/or perform 2D alignment of mRNA microarray results in cancer patients or another disease.

In this example, values of mRNA levels for each patient obtained with Affymetrix HG-U133A arrays are directly transformed into one lattice graph using the following mathematical formalism. To this end, each value for one specific mRNA for one patient is placed as point (node) in a Cartesian 2D space starting with the first data point at the coordinate $\mathbf{r}_2 = (0, 0)$. We calculated the coordinates of the successive data points as follows. In this example, we used the vector ${}^1\mathbf{s} = [{}^1s_1, {}^1s_2, {}^1s_3,.. {}^1s_j,\ldots {}^1s_n]$ to list the labels s_j for the different mRNA profiled with the Affymetrix kit. We also used the vector of weights: ${}^1\mathbf{w} = [{}^1c_j] = [{}^1c_1, {}^1c_2,\ldots {}^1c_j,\ldots {}^1c_n]$ to list the numeric value of the level of the mRNA. Last, we used the following sets of conditions $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_3$, and $\mathbf{C}_4$:

$\mathbf{C}_1$: Increases in +1 the y axe; if ${}^1c_j > {}^{avg}c_j$ and ${}^1c_j > {}^{avg}c_i$ (upwards-step) or:

$\mathbf{C}_2$: Increases in +1 the x axe; if ${}^1c_j > {}^{avg}c_j$ and ${}^1c_j < {}^{avg}c_i$ (rightwards-step) or:

$\mathbf{C}_3$: Decreases in -1 the y axe; if ${}^1c_j < {}^{avg}c_j$ and ${}^1c_j < {}^{avg}c_i$ (downwards-step) or:

$\mathbf{C}_4$: Decreases in -1 the x axe; otherwise (leftwards-step),

where ${}^{avg}c_j$ is the average of ${}^1c_j$ of the same mRNA value for all patients whereas ${}^{avg}c_i$ is the average of ${}^1c_j$ for mRNA value of all gen in a given patient. **Figure 4, B** depicts the 2D-alignment of all these vectors for a sub-set of patients using the same above-mentioned rules $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_3$ and $\mathbf{C}_4$. In this graph, both NC and PR patients are displayed in black whereas pCR patients are coloured in grey. The lattice shows that in fact, both populations share common areas but NC and PR patients with no positive answer to drug treatment distribute downwards to regions not covered by lattices of healthy patients. In any case, this is only a technical-note illustrative example on how to carry out the construction of mRNA Microarrays lattices and we need to perform further research with larger databases in order to draw more convincing conclusions.

3.8. Lattices for Research Trends, Copyright & Patent protection in biological research

In the previous sections, we have introduced 2D lattice graphs for different molecular experiments. However, the applications in proteome research of these lattices may have further implications. For instance, we can use these graphs to analyse the scientific production and copyright or patent protection of this scientific production. It may help proteome research scientists, development managers, and/or politicians to decide which directions on proteome I + D are promising for further investment in order to introduce final protected products in the market. It may help also to detect relevant communities, groups, and/or research networks in their respective areas of interest. The use of graph theory to analyze scientific production trends is not new (Malin and Carley, 2007; Rosvall and Bergstrom, 2008). Thus, in this example we report for the first time the construction of 2D lattice graphs with this aim.

In this example, we get outputs of patent search including the last 500 inputs containing the word protein in the field title from the European Patent Office (EPO) web (http://ep.espacenet.com/). Now, we report the generation of the 2D lattice graphs for this dataset as a sort of illustrative example, more detailed research is expected to be used in other fields. The starting point has coordinates $\mathbf{r}_2 = (0,0)$ placed at the center of a Cartesian 2D space. The coordinates of the successive data points were calculated as follows. First, we assign to each patent a vector $\mathbf{s} = [s_0, s_1, s_2,.. s_j,\ldots s_{20}]$ that lists the labels for different search terms s_j (s_0 is the word method + 20 additional terms). We also used two weighting vectors; the first ${}^1\mathbf{w} = [{}^1f_j] = [{}^1f_0, {}^1f_2,\ldots {}^1f_j,\ldots {}^1f_{20}]$ lists the frequency of each term s_j in the 500 patents studied. The other vector: ${}^1\mathbf{w} = [{}^1\delta_j] = [{}^1\delta_0, {}^1\delta_2,\ldots {}^1\delta_j,\ldots {}^1\delta_{20}]$ lists the values ${}^1\delta_j = 1$ if the term s_j is present in the field of the patent; ${}^1\delta_j = 0$ otherwise. We also used the sets of conditions $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_3$ and $\mathbf{C}_4$.

$\mathbf{C}_1$: Increases in +1 the y axe if ${}^1\delta_0 = 1$ and ${}^1\delta_j = 1$ and ${}^1f_j > {}^*f_j$ (upwards-step) or:

$\mathbf{C}_2$: Increases in +1 the x axe if ${}^1\delta_0 = 0$ and ${}^1\delta_j = 1$ and ${}^1f_j > {}^*f_j$ (rightwards-step) or:
$\mathbf{C}_3$: Decreases in -1 the y axe if ${}^1\delta_0 = 1$ and ${}^1\delta_j = 1$ and ${}^1f_j < {}^*f_j$ (downwards-step).
$\mathbf{C}_4$: Decreases in -1 the x axe if otherwise (leftwards-step).

In **Figure 4, C** we depict the alignment of the 500 protein-research related patents studied here using these rules. For instance, in this case, we can note which areas are common to US, WO, or other patents and which are not, in order to profile patenting strategies in US.

4. Conclusions

The construction of 2D-Generalized Lattice graphs constrained into a Cartesian coordinate system is a useful technique for biological data visualization not necessarily limited to DNA sequences. For instance, we demonstrated how to extend it in order to depict protein sequences, SNPs, parasite enzyme poly-morfirms, peptide MD, protein MS, PSP-MS, mRNA microarray outcomes and protein-research patent information. The present results break new ground in applying the graph theory for knowledge discovery in proteome research as well as other areas of biological sciences.

Acknowledgments

We acknowledge the kind attention and useful comments of the editor and the referees. González-Díaz H., Vilas R. and Munteanu C. R. acknowledge the funding for a research position by Programme Isidro Parga Pondal, Xunta de Galicia. S. Vázquez-Prieto is grateful for the scholarship support from Maria Barbeito Programme, Xunta de Galicia. The authors thank for the partial financial support from project (AGL2006-13936-C01/C02) Ministry of Education and Science, Spain, which is co-financed with European Union funds (FEDER) and for the grants 2007/127 and 2007/144 from the General Directorate of Scientific and Technologic Promotion of the Galician University System of the Xunta de Galicia.

References

Abdolmaleky, H.M., Faraone, S.V., Glatt, S.J., and Tsuang, M.T., 2004. Meta-analysis of association between the T102C polymorphism of the 5HT2a receptor gene and schizophrenia. Schizophr Res 67, 53-62.

Aguero-Chapin, G., Gonzalez-Diaz, H., Molina, R., Varona-Santos, J., Uriarte, E., and Gonzalez-Diaz, Y., 2006. Novel 2D maps and coupling numbers for protein sequences. The first QSAR study of polygalacturonases; isolation and prediction of a novel sequence from Psidium guajava L. FEBS Lett 580, 723-30.

Alper, C.A., and Johnson, A.M., 1969. Immunofixation electrophoresis: a technique for the study of protein polymorphism. Vox Sang 17, 445-52.

Althaus, I.W., Chou, J.J., Gonzales, A.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Aristoff, P.A., Tarpley, W.G., and Reusser, F., 1993a. Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E. Biochemistry 32, 6548-6554.

Althaus, I.W., Chou, J.J., Gonzales, A.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Aristoff, P.A., Tarpley, W.G., and Reusser, F., 1993b. Steady-state kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-87201E. J Biol Chem 268, 6119-6124.

Althaus, I.W., Gonzales, A.J., Chou, J.J., Diebel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Aristoff, P.A., Tarpley, W.G., and Reusser, F., 1993c. The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. J Biol Chem 268, 14875-14880.

Althaus, I.W., Chou, K.C., Franks, K.M., Diebel, M.R., Kezdy, F.J., Romero, D.L., Thomas, R.C., Aristoff, P.A., Tarpley, W.G., and Reusser, F., 1996. The benzylthio-pyrididine U-31,355 is a potent inhibitor of HIV-1 reverse transcriptase. Biochemical Pharmacology 51, 743-750.

Althaus, I.W., Chou, J.J., Gonzales, A.J., Deibel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Thomas, R.C., Aristoff, P.A., Tarpley, W.G., and et al., 1994a. Kinetic studies with the non-nucleoside human immunodeficiency virus type-1 reverse transcriptase inhibitor U-90152E. Biochem Pharmacol 47, 2017-28.

Althaus, I.W., Chou, J.J., Gonzales, A.J., LeMay, R.J., Deibel, M.R., Chou, K.C., Kezdy, F.J., Romero, D.L., Thomas, R.C., Aristoff, P.A., and et al., 1994b. Steady-state kinetic studies with the polysulfonate U-9843, an HIV reverse transcriptase inhibitor. Experientia 50, 23-8.

Andraos, J., 2008. Kinetic plasticity and the determination of product ratios for kinetic schemes leading to multiple products without rate laws: new methods based on directed graphs. Canadian Journal of Chemistry 86, 342-357.

Ayala, F.J., 1983. Genetic polymorphism: from electrophoresis to DNA sequences. Experientia 39, 813-23.

Barabasi, A.L., and Oltvai, Z.N., 2004. Network biology: understanding the cell's functional organization. Nat Rev Genet 5, 101-13.

Bartels, C., 1990. Fast algorithm for peptide sequencing by mass spectroscopy. Biomed. Environ. Mass Spectrom. 19, 363–368.

Berger, B., and Leighton, T., 1998. Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. J Comput Biol 5, 27-40.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D.U., 2006. Complex networks: Structure and dynamics. Physics Reports 424, 175-308.

Cornish-Bowden, A., Chapter 4. Butterworths, Fundamentals of Enzyme Kinetics, London 1979.

Cruz-Monteagudo, M., Munteanu, C.R., Borges, F., Cordeiro, M.N., Uriarte, E., and Gonzalez-Diaz, H., 2008a. Quantitative Proteome-Property Relationships (QPPRs). Part 1: finding biomarkers of organic drugs with mean Markov connectivity indices of spiral networks of blood mass spectra. Bioorg Med Chem 16, 9684-93.

Cruz-Monteagudo, M., Munteanu, C.R., Borges, F., Cordeiro, M.N., Uriarte, E., Chou, K.C., and González-Díaz, H., 2008b. Stochastic molecular descriptors for polymers. 4. Study of complex mixtures with topological indices of mass spectra spiral and star networks: The blood proteome case. Polymer 49, 5575–5587.

Chen, W., Liao, B., Zhu, W., and Xiang, X., 2009. Multiple sequence alignment algorithm based on a dispersion graph and ant colony algorithm. J Comput Chem.

Chenik, M., Chaabouni, N., Achour-Chenik, Y.B., Ouakad, M., Lakhal-Naouar, I., Louzir, H., and Dellagi, K., 2006. Identification of a new developmentally regulated Leishmania major large RAB GTPase. BBRC 341, 541-548.

Chou, K.C., 1980. A new schematic method in enzyme kinetics. European Journal of Biochemistry 113, 195-8.

Chou, K.C., 1981. Two new schematic rules for rate laws of enzyme-catalyzed reactions. Journal of Theoretical Biology 89, 581-592.

Chou, K.C., 1989. Graphical rules in steady and non-steady enzyme kinetics J Biol Chem 264, 12074-12079.

Chou, K.C., 1990. Review: Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems Biophys Chem 35, 1-24.

Chou, K.C., and Forsen, S., 1980. Graphical rules for enzyme-catalyzed rate laws. Biochem J 187, 829-835.

Chou, K.C., and Liu, W.M., 1981. Graphical rules for non-steady state enzyme kinetics. J Theor Biol 91, 637-54.

Chou, K.C., and Zhang, C.T., 1992. Diagrammatization of codon usage in 339 HIV proteins and its biological implication. AIDS Research and Human Retroviruses 8, 1967-1976.

Chou, K.C., and Shen, H.B., 2009. FoldRate: A web-server for predicting protein folding rates from primary sequence. The Open Bioinformatics Journal 3, 31-50.

Chou, K.C., Kezdy, F.J., and Reusser, F., 1994. Review: Steady-state inhibition kinetics of processive nucleic acid polymerases and nucleases. Anal Biochem 221, 217-230.

Chou, K.C., Zhang, C.T., and Elrod, D.W., 1996. Do antisense proteins exist? Journal of Protein Chemistry 15, 59-61.

Chou, K.C., Jiang, S.P., Liu, W.M., and Fee, C.H., 1979. Graph theory of enzyme kinetics: 1. Steady-state reaction system. Scientia Sinica 22, 341-358.

Dancík, V., Addona, T.A., Clauser, K.R., Vath, J.E., and Pevzner, P.A., 1999. De Novo Peptide Sequencing via Tandem Mass Spectrometry. J Comput Biol 6, 327–342.

Dea-Ayuela, M.A., and Bolás-Fernández, F., 2005. Two-dimensional electrophoresis and mass spectrometry for the identification of species-specific Trichinella antigens. Vet Parasitol 132, 43-49.

Dea-Ayuela, M.A., Perez-Castillo, Y., Meneses-Marcel, A., Ubeira, F.M., Bolas-Fernandez, F., Chou, K.C., and Gonzalez-Diaz, H., 2008. HP-Lattice QSAR for dynein proteins: experimental proteomics (2D-electrophoresis, mass spectrometry) and theoretic study of a Leishmania infantum sequence. Bioorg Med Chem 16, 7770-6.

Diao, Y., Li, M., Feng, Z., Yin, J., and Pan, Y., 2007. The community structure of human cellular signaling network. J Theor Biol 247, 608-15.

Dominguez, E., Loza, M.I., Padin, F., Gesteira, A., Paz, E., Paramo, M., Brenlla, J., Pumar, E., Iglesias, F., Cibeira, A., Castro, M., Caruncho, H., Carracedo, A., and Costas, J., 2007. Extensive linkage disequilibrium

mapping at HTR2A and DRD3 for schizophrenia susceptibility genes in the Galician population. Schizophr Res 90, 123-9.

Estrada, E., 2006. Protein bipartivity and essentiality in the yeast protein-protein interaction network. J Proteome Res 5, 2177-84.

Ferino, G., Gonzalez-Diaz, H., Delogu, G., Podda, G., and Uriarte, E., 2008. Using spectral moments of spiral networks based on PSA/mass spectra outcomes to derive quantitative proteome-disease relationships (QPDRs) and predicting prostate cancer. Biochem Biophys Res Commun 372, 320-5.

Fernandez-de-Cossio, J., Gonzalez, J., and Besada, V., 1995. A computer program to aid the sequencing of peptides in collision-activated decomposition experiments. Comput Appl Biosci 11, 427-34.

Frank, A., and Pevzner, P., 2005. P. PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling. Anal. Chem. 77, 964-973.

Gao, L., Ding, Y.S., Dai, H., Shao, S.H., Huang, Z.D., and Chou, K.C., 2006. A novel fingerprint map for detecting SARS-CoV. Journal of Pharmaceutical and Biomedical Analysis 41, 246-50.

Gates, M.A., 1986. A simple way to look at DNA. J Theor Biol 119, 319-328.

Gonzalez-Diaz, H., 2008. Quantitative studies on Structure-Activity and Structure-Property Relationships (QSAR/QSPR). Curr Top Med Chem 8, 1554.

Gonzalez-Diaz, H., Prado-Prado, F., and Ubeira, F.M., 2008a. Predicting antimicrobial drugs and targets with the MARCH-INSIDE approach. Curr Top Med Chem 8, 1676-90.

Gonzalez-Diaz, H., Gonzalez-Diaz, Y., Santana, L., Ubeira, F.M., and Uriarte, E., 2008b. Proteomics, networks and connectivity indices. Proteomics 8, 750-78.

González-Díaz, H., Sanchez-Gonzalez, A., and Gonzalez-Diaz, Y., 2006. 3D-QSAR study for DNA cleavage proteins with a potential anti-tumor ATCUN-like motif. J Inorg Biochem 100, 1290-7.

González-Díaz, H., Vilar, S., Santana, L., and Uriarte, E., 2007a. Medicinal Chemistry and Bioinformatics – Current Trends in Drugs Discovery with Networks Topological Indices. Curr Top Med Chem 7, 1025-39.

González-Díaz, H., González-Díaz, Y., Santana, L., Ubeira, F.M., and Uriarte, E., 2008. Proteomics, networks and connectivity indices. 1615-9853 8, 750-778.

González-Díaz, H., Bonet, I., Terán, C., de Clercq, E., Bello, R., García, M., Santana, L., and Uriarte, E., 2007b. ANN-QSAR model for selection of anticancer leads from structurally heterogeneous series of compounds. European Journal of Medicinal Chemistry 42, 580-585.

Hadi, H.A., Mohran, Z.S., Hakam, A.A., Mourad, A., and Oyofo, B.A., 1998. Characterization of Campylobacter spp. using restriction fragment length polymorphism and SDS-polyacrylamide gel electrophoresis. J Egypt Public Health Assoc 73, 1-10.

Hamacher, K., 2007. Information theoretical measures to analyze trajectories in rational molecular design. J Comput Chem 28, 2576-80.

Hu, S., Loo, J.A., and Wong, D.T., 2006. Human body fluid proteome analysis. Proteomics 6, 6326-53.

Huang, G., Liao, B., Li, Y., and Yu, Y., 2009. Similarity studies of DNA sequences based on a new 2D graphical representation. Biophys Chem 143, 55-9.

Jonsson, E.G., Flyckt, L., Burgert, E., Crocq, M.A., Forslund, K., Mattila-Evenden, M., Rylander, G., Asberg, M., Nimgaonkar, V.L., Edman, G., Bjerkenstedt, L., Wiesel, F.A., and Sedvall, G.C., 2003. Dopamine D3 receptor gene Ser9Gly variant and schizophrenia: association study and meta-analysis. Psychiatr Genet 13, 1-12.

Kanamori-Kataoka, M., and Seto, Y., 2009. Paraoxonase activity against nerve gases measured by capillary electrophoresis and characterization of human serum paraoxonase (PON1) polymorphism in the coding region (Q192R). Anal Biochem 385, 94-100.

Kantor, A.B., 2002. Comprehensive phenotyping and biological marker discovery. Dis Markers 18, 91-7.

Karplus, M., and McCammon, J.A., 2002. Molecular dynamics simulations of biomolecules. Nat Struct Biol 9, 646-52.

King, E.L., and Altman, C., 1956. A schematic method of deriving the rate laws for enzyme-catalyzed reactions. Journal of Physical Chemistry 60, 1375-1378.

Kuzmic, P., Ng, K.Y., and Heath, T.D., 1992. Mixtures of tight-binding enzyme inhibitors. Kinetic analysis by a recursive rate equation Anal Biochem 200 68-73.

Lajoix, A.D., Gross, R., Aknin, C., Dietz, S., Granier, C., and Laune, D., 2004. Cellulose membrane supported peptide arrays for deciphering protein-protein interaction sites: the case of PIN, a protein with multiple natural partners. Mol Divers 8, 281-90.

Leong, P.M., and Morgenthaler, S., 1995. Random walk and gap plots of DNA sequences. Comput Applic Biosci 11, 503-507.

Liao, B., 2005. A 2D graphical representation of DNA sequence. Chem Phys Lett 401, 196-199.

Liao, B., and Wang, T.M., 2004. New 2D graphical representation of DNA sequences. J Comput Chem 25, 1364-8.

Liao, B., and Ding, K., 2005. Graphical approach to analyzing DNA sequences. J Comput Chem 26, 1519-23.

Liao, B., Tan, M., and Ding, K., 2005. Application of 2-D graphical representation of DNA sequence. Chem Phys Lett 414, 296-300.

Liao, B., Xiang, X., and Zhu, W., 2006. Coronavirus phylogeny based on 2D graphical representation of DNA sequence. J Comput Chem 27, 1196-1202.

Liao, B., Luo, J., Li, R., and Zhu, W., 2006 RNA Secondary structure 2D graphical representation without degeneracy. International Journal of Quantum Chemistry 106 1749-1755.

Liao, B., Chen, W., Sun, X., and Zhu, W., 2009. A binary coding method of RNA secondary structure and its application. J Comput Chem.

Lopez-Galvez, G., Juarez, M., and Ramos, M., 1995. Two dimensional electrophoresis and immunoblotting for the study of ovine whey protein polymorphism. J Dairy Res 62, 311-20.

Malin, B., and Carley, K., 2007. A longitudinal social network analysis of the editorial boards of medical informatics and bioinformatics journals. J Am Med Inform Assoc 14, 340-8.

McCammon, J.A., Gelin, B.R., and Karplus, M., 1977. Dynamics of folded proteins. Nature 267, 585-90.

McDonald, W.H., and Yates, J.R., 3rd, 2002. Shotgun proteomics and biomarker discovery. Dis Markers 18, 99-105.

Meltzer, H.Y., Matsubara, S., and Lee, J.C., 1989. Classification of typical and atypical antipsychotic drugs on the basis of dopamine D-1, D-2 and serotonin2 pKi values. J Pharmacol Exp Ther 251, 238-46.

Mezo, M., Gonzalez-Warleta, M., Castro-Hermida, J.A., and Ubeira, F.M., 2008. Evaluation of the flukicide treatment policy for dairy cattle in Galicia (NW Spain). Vet Parasitol 157, 235-43.

Modlich, O., Prisack, H.B., Munnes, M., Audretsch, W., and Bojar, H., 2005. Predictors of primary breast cancers responsiveness to preoperative epirubicin/cyclophosphamide-based chemotherapy: translation of microarray data into clinically useful predictive signatures. J Transl Med 3, 32.

Myers, D., and Palmer, G., 1985. Microcomputer tools for steady-state enzyme kinetics. Bioinformatics (original: Computer Applied Bioscience) 1, 105-110.

Nandy, A., 1996a. Two-dimensional graphical representation of DNA sequences and intron-exon discrimination in intron-rich sequences. Comput Appl Biosci 12, 55-62.

Nandy, A., 1996b. Two-dimensional graphical representation of DNA sequences and intron-exon discrimination in intron-rich sequences. CABIOS (Comput-Appl-Biosci.) 12, 55-62.

Nandy, A., 2003. Novel Method for Discrimination of Conserved Genes through Numerical Characterization of DNA Sequences. Int E J Mol Design 2, 000–000.

Novic, M., and Randic, M., 2008. Representation of proteins as walks in 20-D space. SAR QSAR Environ Res 19, 317-37.

Petricoin, E.F., Rajapaske, V., Herman, E.H., Arekani, A.M., Ross, S., Johann, D., Knapton, A., Zhang, J., Hitt, B.A., Conrads, T.P., Veenstra, T.D., Liotta, L.A., and Sistare, F.D., 2004. Toxicoproteomics: serum proteomic pattern diagnostics for early detection of drug induced cardiac toxicities and cardioprotection. Toxicol Pathol 32 Suppl 1, 122-30.

Prado-Prado, F.J., González-Díaz, H., Martinez de la Vega, O., Ubeira, F.M., and Chou, K.C., 2008. Unified QSAR approach to antimicrobials. Part 3: First multi-tasking QSAR model for Input-Coded prediction, structural back-projection, and complex networks clustering of antiprotozoal compounds. Bioorganic and Medicinal Chemistry 16, 5871–5880.

Qi, X.Q., Wen, J., and Qi, Z.H., 2007. New 3D graphical representation of DNA sequence based on dual nucleotides. J Theor Biol 249, 681-90.

Randic, M., 2004. Graphical representations of DNA as 2-D map. Chemical Physics Letters 386, 468-471.

Randic, M., 2006. Quantitative characterizations of proteome: dependence on the number of proteins considered. J Proteome Res 5, 1575-9.
Randic, M., and Balaban, A.T., 2003. On a four-dimensional representation of DNA primary sequences. J Chem Inf Comput Sci 43, 532-9.
Randic, M., Guo, X., and Basak, S.C., 2001. On the characterization of DNA primary sequences by triplet of nucleic acid bases. J Chem Inf Comput Sci 41, 619-26.
Randic, M., Zupan, J., and Vikic-Topic, D., 2007. On representation of proteins by star-like graphs. J Mol Graph Model 26, 290-305.
Randic, M., Novic, M., and Vracko, M., 2008. On novel representation of proteins based on amino acid adjacency matrix. SAR QSAR Environ Res 19, 339-49.
Randic, M., Mehulic, K., Vukicevic, D., Pisanski, T., Vikic-Topic, D., and Plavsic, D., 2009. Graphical representation of proteins as four-color maps and their numerical characterization. J Mol Graph Model 27, 637-41.
Randič, M., 2002 A Graph Theoretical Characterization of Proteomics Maps. Int J Quant Chem 90, 848–858.
Randic´, M., Lers, N., Plavsic´, D., Basak, S., and Balaban, A.T., 2005. Four-color map representation of DNA or RNA sequences and their numerical characterization. Chem Phys Lett 407, 205-208.
Roldos, V., Nakayama, H., Rolon, M., Montero-Torres, A., Trucco, F., Torres, S., Vega, C., Marrero-Ponce, Y., Heguaburu, V., Yaluff, G., Gomez-Barrio, A., Sanabria, L., Ferreira, M.E., Rojas de Arias, A., and Pandolfi, E., 2008. Activity of a hydroxybibenzyl bryophyte constituent against Leishmania spp. and Trypanosoma cruzi: In silico, in vitro and in vivo activity studies. Eur J Med Chem 43, 1797-807.
Rosvall, M., and Bergstrom, C.T., 2008. Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci U S A 105, 1118-23.
Roy, A., Raychaudhur, C., and Nandy, A., 1998. Novel techniques of graphical representation and analysis of DNA sequences-A review. J. Biosci. 23, 55-71.
Sarciron, M.E., Terreux, R., Prieto, Y., Cortes, M., Cuellar, M.A., Tapia, R.A., Domard, M., Walchshofer, N., and Petavy, A.F., 2005. Antileishmanial activity of polycyclic derivatives. Parasite 12, 251-8.
Shen, H.B., Song, J.N., and Chou, K.C., 2009. Prediction of protein folding rates from primary sequence by fusing multiple sequential features. Journal of Biomedical Science and Engineering (JBiSE) 2, 136-143 (open accessible at http://www.srpublishing.org/journal/jbise/).
Taylor, J.A., and Johnson, R.S., 1997. Sequence Database Searches via de Novo Peptide Sequencing by Tandem Mass Spectrometry. Rapid Communications In Mass Spectrometry 11, 1067–1075.
Valero, M.A., Panova, M., and Mas-Coma, S., 2005. Phenotypic analysis of adults and eggs of Fasciola hepatica by computer image analysis system. J Helminthol 79, 217-25.
Wang, M., Yao, J.S., Huang, Z.D., Xu, Z.J., Liu, G.P., Zhao, H.Y., Wang, X.Y., Yang, J., Zhu, Y.S., and Chou, K.C., 2005. A new nucleotide-composition based fingerprint of SARS-CoV with visualization analysis. Med Chem 1, 39-47.
Wolfram, S., 1984. Cellular automation as models of complexity. Nature 311, 419-424.
Wolfram, S., 2002. A New Kind of Science. Wolfram Media Inc., Champaign, IL.
Xiao, X., and Chou, K.C., 2007. Digital coding of amino acids based on hydrophobic index. Protein Pept Lett 14, 871-5.
Xiao, X., Shao, S.H., and Chou, K.C., 2006a. A probability cellular automaton model for hepatitis B viral infections. Biochemical and Biophysical Research Communications 342, 605-10.
Xiao, X., Wang, P., and Chou, K.C., 2009. GPCR-CA: A cellular automaton image approach for predicting G-protein-coupled receptor functional classes. J Comput Chem 30, 1414-23.
Xiao, X., Shao, S., Ding, Y., Huang, Z., and Chou, K.C., 2006b. Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. Amino Acids 30, 49-54.
Xiao, X., Shao, S., Ding, Y., Huang, Z., Chen, X., and Chou, K.C., 2005a. Using cellular automata to generate image representation for biological sequences. Amino Acids 28, 29-35.
Xiao, X., Shao, S., Ding, Y., Huang, Z., Chen, X., and Chou, K.C., 2005b. An application of gene comparative image for predicting the effect on replication ratio by HBV virus gene missense mutation. J Theor Biol 235, 555-65.

Zhang, C.T., and Chou, K.C., 1993. Graphic analysis of codon usage strategy in 1490 human proteins. J Protein Chem 12, 329-35.

Zhang, C.T., and Chou, K.C., 1994. Analysis of codon usage in 1562 E. Coli protein coding sequences. J Mol Biol 238, 1-8.

Zhang, L., Liao, B., Li, D., and Zhu, W., 2009. A novel representation for apoptosis protein subcellular localization prediction using support vector machine. J Theor Biol 259, 361-5.

Zhou, G.P., and Deng, M.H., 1984. An extension of Chou's graphical rules for deriving enzyme kinetic equations to system involving parallel reaction pathways. Biochem J 222, 169-176.

FIGURE LEGENDS

Figure 1. Sequence, *vs.* MDT and MS lattice graphs for peptides found on PMF of proteins

Figure 2. 2D/1D Electrophoresis experiments reported in this work and examples of lattices

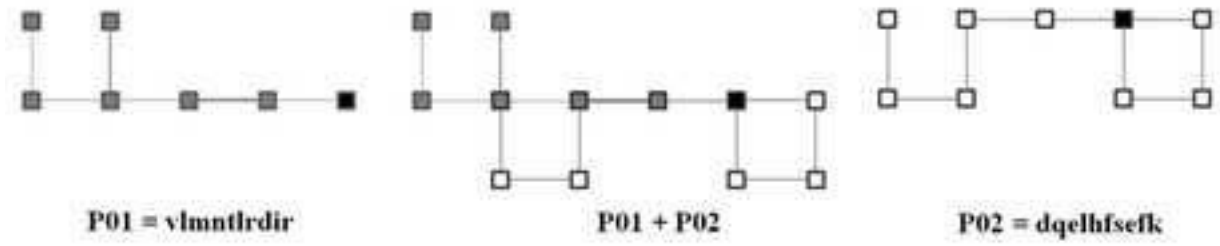
Figure 3. Parasite Polymorphism lattices for different populations derived with 1D Electrophoresis results

Figure 4. Examples of lattices for: **(A)** SNPs of Schizophrenia patients, **(B)** Microarray for Cancer patients and **(C)** Patents related to protein-research methods

Table 1. Some information for peptides used to construct Sequence, MDT, or PMF-MS lattices.

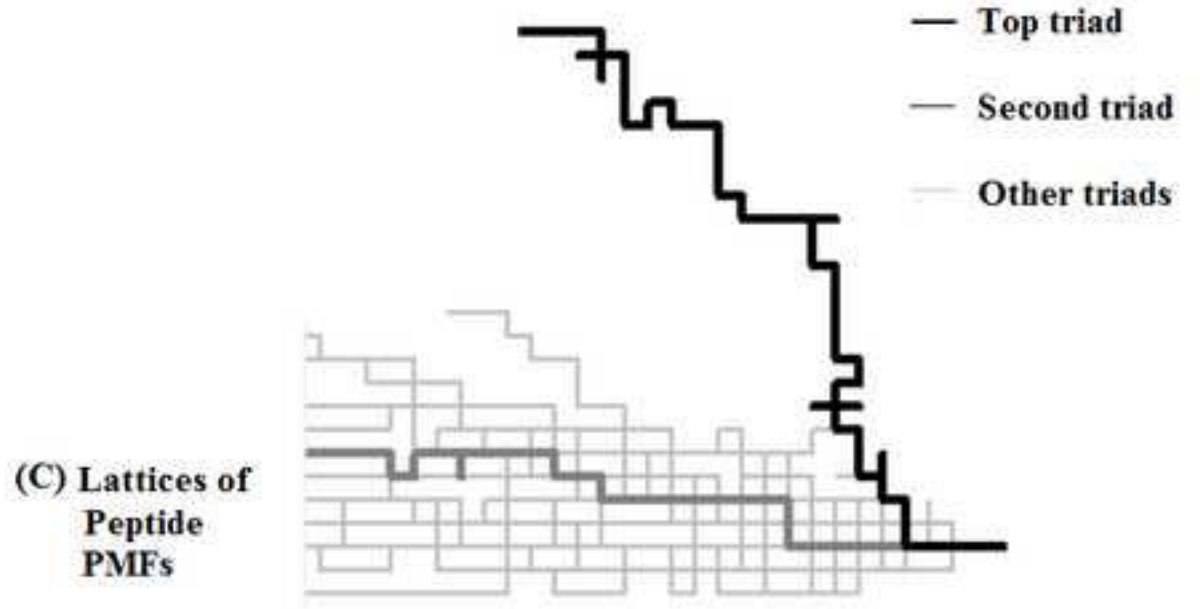
Peptide	Sequence	AAs	E_0	E_1	E_{100}	ACCR	$(m/z)_j$
P01	vlmntlrdir	10	635.4	-175.49	-22.38	0.48	1246.67
P02	dqelhfsefk	10	264.1	-15.71	107.65	0.46	1279.71
P03	hgimvvgpamcgk	13	17719.8	67.89	211.26	0.47	1356.70
P04	hwqeimkvsgr	11	10779.3	-34.40	133.33	0.47	1370.71
P05	qvmeylchfr	10	456.1	-75.29	75.14	0.47	1382.71
P06	mdsanglidalsger	15	714.9	-84.51	65.21	0.48	1564.81
P07	mnpkaitapqmfgr	14	15383.8	-42.80	137.60	0.47	1593.84
P08	mmytiaryyptr	12	16704.0	-116.06	62.84	0.47	1597.85
P09	lratmnadgqmlpr	14	14499.2	-145.61	26.08	0.48	1605.85
P10	ldfsslfiptadsvr	15	1325865.3	-80.48	98.25	0.47	1667.86
P11	lvrhgimvvgpamcgk	16	18520.6	18.45	197.03	0.48	1740.96
P12	eavahdaaivahgeaeakk	19	1343.8	13.43	222.83	0.47	1917.03
P13	qvvemsqvydlskpgvr	17	15611.8	-124.14	84.37	0.47	1935.04
P14	qvvemsqvydlskpgvrr	18	15565.5	-184.71	56.38	0.48	2091.14
P15	ylqsldtyfdvlyssnlqr	19	1532.4	-184.84	73.61	0.47	2325.15
P16	aqskpwetitdavtllrvwk	20	43367.0	-104.21	167.46	0.47	2342.16
P17	ldfsslfiptadsvrlhylak	21	$1.4\cdot10^{-7}$	-62.93	193.21	0.47	2393.28
P18	iwvtsephnsvpigllqmsikltneppqgik	31	$1.5\cdot10^{-7}$	-66.01	298.05	0.47	3442.90

4. Figure



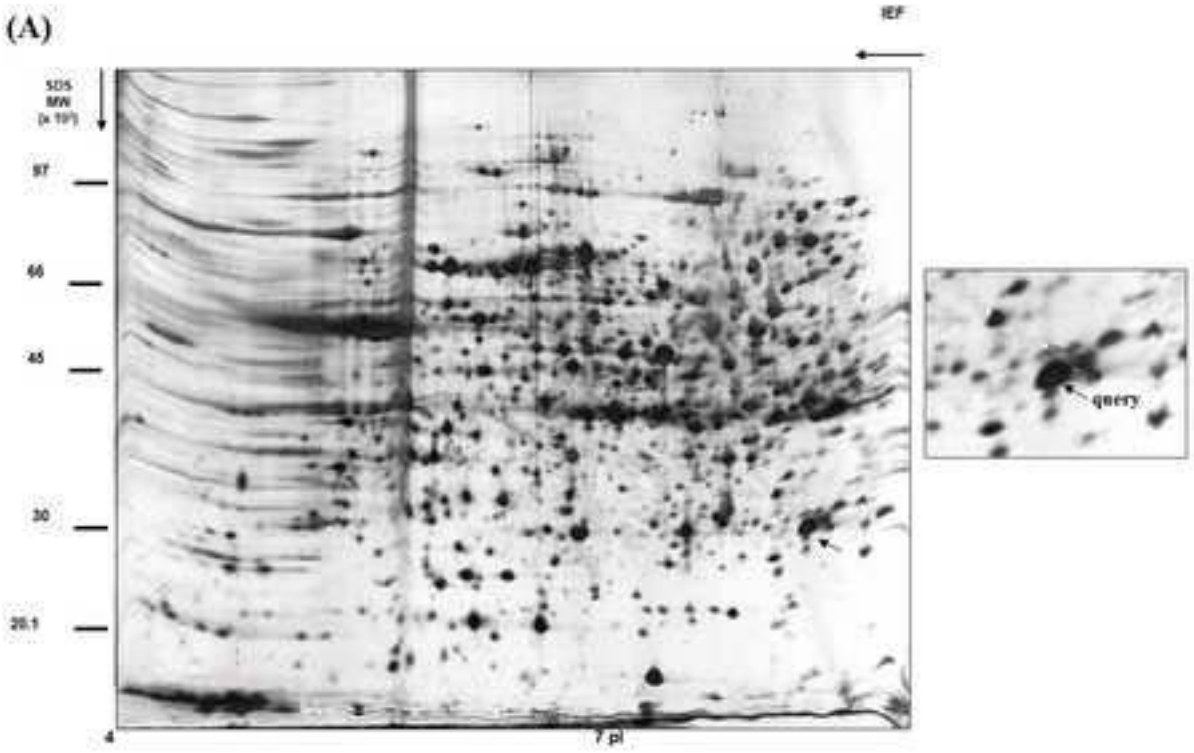
(A) Lattices of Peptide Sequences

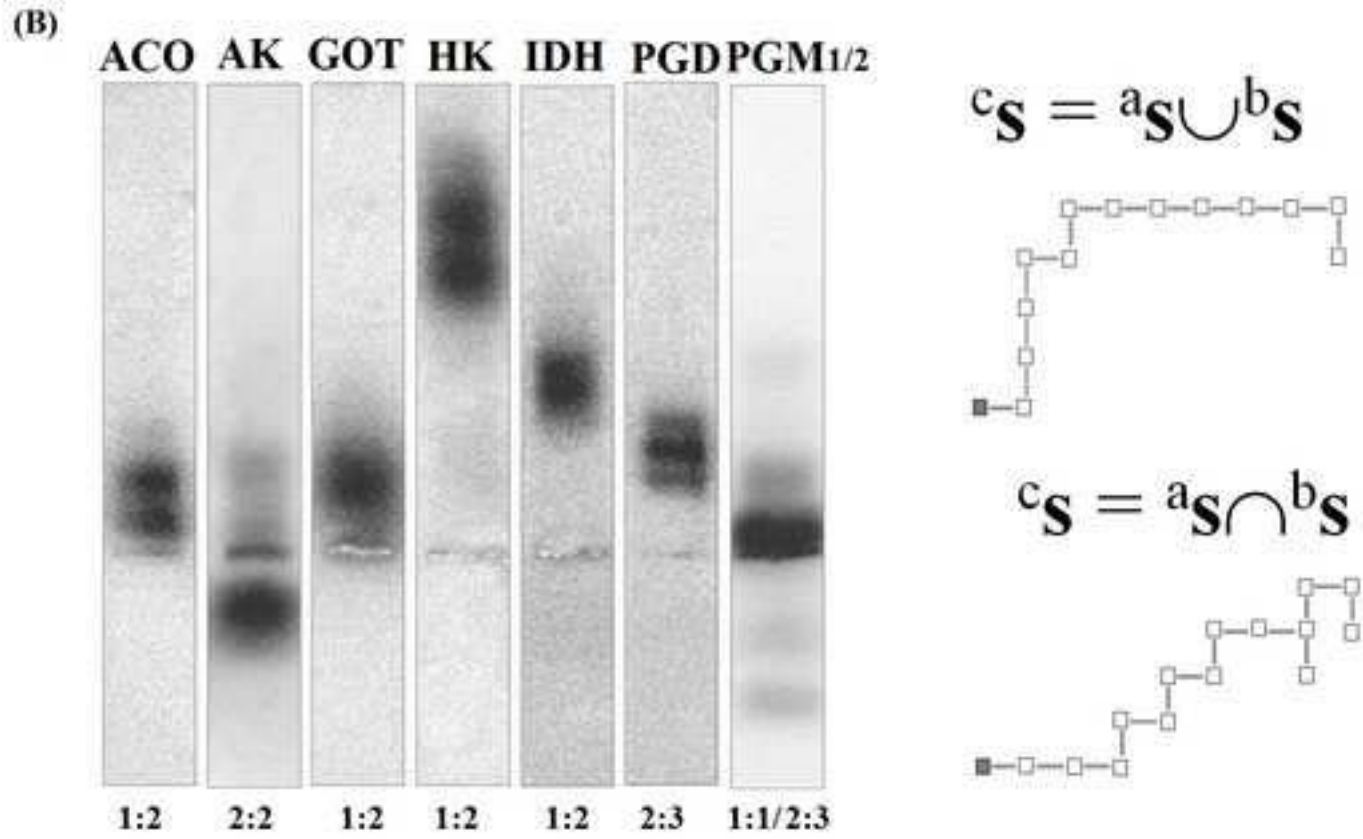
(B) Lattices of Peptide MDTs



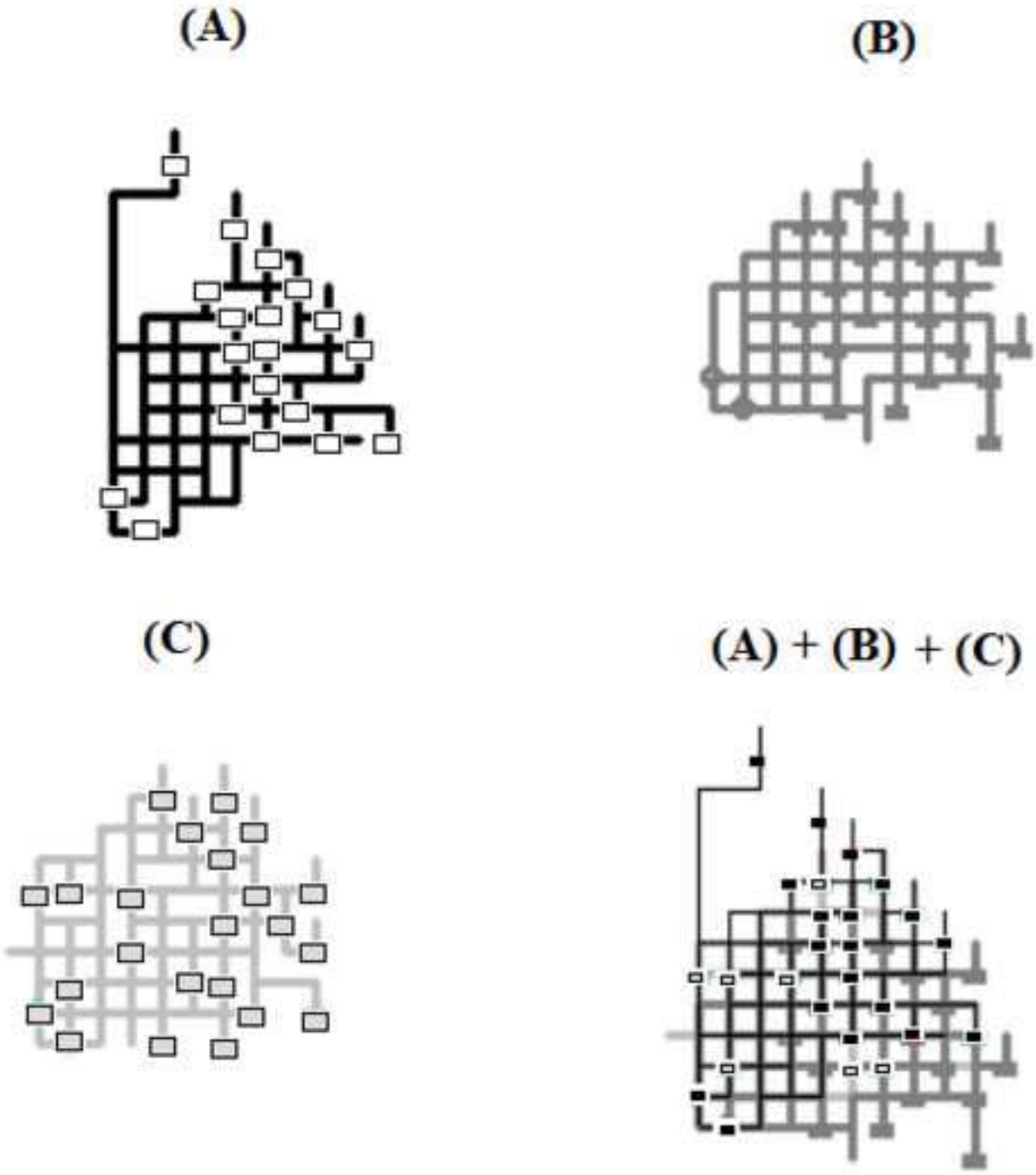
(C) Lattices of Peptide PMFs

4. Figure





4. Figure



4. Figure