Journal of Theoretical Biology 256 (2009) 458–466

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Alignment-free prediction of mycobacterial DNA promoters based on pseudo-folding lattice network or star-graph topological indices

Alcides Perez-Bello [a,b,c], Cristian Robert Munteanu [d], Florencio M. Ubeira [a], Alexandre Lopes De Magalhães [d], Eugenio Uriarte [c], Humberto González-Díaz [a,*]

[a] Department of Microbiology and Parasitology, University of Santiago de Compostela, Santiago de Compostela 15782, Spain
[b] Department of Veterinary Medicine, UCLV, Santa Clara 54830, Cuba
[c] Department of Organic Chemistry, Faculty of Pharmacy, University of Santiago de Compostela, Santiago de Compostela 15782, Spain
[d] REQUIMTE/University of Porto, Faculty of Science, Chemistry Department, Porto 4169-007, Portugal

ARTICLE INFO

Article history:
Received 6 August 2008
Received in revised form
23 September 2008
Accepted 25 September 2008
Available online 17 October 2008

Keywords:
QSAR
Markov model
Mycobacterial promoters
Star-graph
Lattice network
Topological indices

ABSTRACT

The importance of the promoter sequences in the function regulation of several important mycobacterial pathogens creates the necessity to design simple and fast theoretical models that can predict them. This work proposes two DNA promoter QSAR models based on pseudo-folding lattice network (LN) and star-graphs (SG) topological indices. In addition, a comparative study with the previous RNA electrostatic parameters of thermodynamically-driven secondary structure folding representations has been carried out. The best model of this work was obtained with only two LN stochastic electrostatic potentials and it is characterized by accuracy, selectivity and specificity of 90.87%, 82.96% and 92.95%, respectively. In addition, we pointed out the SG result dependence on the DNA sequence codification and we proposed a QSAR model based on codons and only three SG spectral moments.

1. Introduction

Protein synthesis promoter sequences play an important role in the function regulation of several important mycobacterial pathogens (Levine and Tjian, 2003; Wyrick and Young, 2002). In this sense, the prediction of the mycobacterial promoter sequences (Mps) could be interesting for the future discovery of new anti-mycobacterial drug targets or for the study of protein metabolism. Mycobacteria have a low transcription rate and a low RNA content per DNA unit. Thus, the transcription and translation signals in Mycobacteria may be different from those in other bacteria such as Esccherichia coli. The large variations among the characterized mycobacterial promoters suggest that the consensus sequences are not representative of these promoters. Consequently, a number of conflicting opinions regarding the presence and characteristics of consensus promoter sequences in the Mycobacteria have been presented in the literature (Mulder et al., 1997). Therefore, understanding the factors responsible for the low level of transcription and the possible mechanisms of regulation of gene expression in Mycobacteria, involves the examination of the mycobacterial promoter structure and the promoter transcription machinery, including chemical information about the involved RNA molecules (Arnvig et al., 2005; Harshey and Ramakrishnan, 1977). Efforts have been made to develop statistical algorithms for the sequence analysis and motif prediction by searching for homologous regions or by comparing the sequence information with a consensus sequence (O'Neill and Chiafari, 1989). Wide variations existing within individual promoter sequences are primarily responsible for the unsatisfactory results yielded by the promoter-site-searching algorithms that in essence perform statistical analysis (Mulligan and McClure, 1986; Mulligan et al., 1984). Therefore, it can be inferred that the recognition of Mps requires a powerful technique capable of unravelling those hidden patterns in the promoter regions, which are difficult to identify directly by sequence alignment.

The bioinformatics methods, based on sequence alignment, may fail in general in cases of low sequence homology between the database query and the template sequences. The lack of function annotation (defined biological function) of the sequences used as template for function prediction constitutes another weakness of the alignment approaches (Dobson and Doig, 2005;

* Corresponding author at: Faculty of Pharmacy, University of Santiago de Compostela. Tel.: +34 981 563100; fax: +34 981 594912.
E-mail addresses: alcidopb@gmail.com (A. Perez-Bello), muntisa@gmail.com (C.R. Munteanu), mpubeira@usc.es (F.M. Ubeira), almagalh@fc.up.pt (A. Lopes De Magalhães), eugenio.uriarte@usc.es (E. Uriarte), humberto.gonzalez@usc.es (H. González-Díaz).

0022-5193/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2008.09.035

Dobson et al., 2004, 2005). In addition, Chou demonstrated that the 3D structures developed, based on homology modelling are very sensitive to the sequence alignment of the query protein with the structure-known protein (Chou, 2004). A group of researchers shows the growing importance of machine learning methods for predicting protein functional class, independently of sequence similarity (Han et al., 2006). These methods often use the input the 1D sequence numerical parameters, specifically defined to seek sequence–function relationships. For instance, the so-called pseudo amino acid composition approach (Chou, 2001a, 2005), based on 1D sequence coupling numbers, has been widely used to predict subcellular localization, enzyme family class, structural class, as well as other attributes of proteins based on their sequence similarity (Caballero et al., 2006; Chou and Shen, 2006; Du et al., 2008). Alternatively, the molecular indices that are classically used for small molecules (Aguero-Chapin et al., 2006; Liao and Wang, 2004a, b, c; Liao and Ding, 2005; Liao et al., 2005, 2006; Liu et al., 2002; Nandy, 1994, 1996; Nandy and Basak, 2000; Randic and Vracko, 2000; Randic and Balaban, 2003; Randic and Zupan, 2004; Randic et al., 2000; Song and Tang, 2005; Woodcock et al., 1992; Zupan and Randic, 2005) have been adapted to describe the protein sequences. On the other hand, many authors have introduced 2D or higher dimension representations of sequences prior to the calculation of numerical parameters. This is an important step in order to uncover useful higher-order information not encoded by 1D sequence parameters (Randic, 2004). An example of the 2D representations is the graphs used for proteins and DNA sequences. For instance, the spectral-like and zigzag representations have been used in order to suggest an algorithm to enclose long strings of building blocks (like four DNA bases, 20 natural amino acids, or all 64 possible base triplets) (Aguero-Chapin et al., 2006). The use of the graphic approaches to study biological systems can provide useful insights, as indicated by many previous studies on a series of important biological topics, such as enzyme-catalyzed reactions (Andraos, 2008; Chou, 1981, 1989; Chou and Forsen, 1980; Chou and Liu, 1981; Chou et al., 1979; Cornish-Bowden, 1979; King and Altman, 1956; Kuzmic et al., 1992; Myers and Palmer, 1985; Zhou and Deng, 1984), protein folding kinetics (Chou, 1990), inhibition kinetics of processive nucleic acid polymerases and nucleases (Althaus et al., 1993a, b, c; Chou et al., 1994), analysis of codon usage (Chou and Zhang, 1992; Zhang and Chou, 1993, 1994), analysis of DNA sequence (Qi et al., 2007). Moreover, graphical methods have been introduced for the quantitative structure-activity relationships (QSAR) study (González-Díaz et al., 2006c, 2007b; Prado-Prado et al., 2008) as well as used to deal with complicated network systems (Diao et al., 2007; González-Díaz et al., 2007a, 2008). Recently, the "cellular automaton image" (Wolfram, 1984, 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 to represent complicated biological sequences (Xiao et al., 2005a) and help to identify protein attributes (Xiao and Chou, 2007; Xiao et al., 2006b).

In this work, we are proposing a comparative study of the Mycobacterial DNA promoter prediction using pseudo-folding lattice network (LN) and star-graph (SG) topological indices. The first group of indices contains the mean stochastic electrostatic potential (${}^{\mathrm{LN}}\xi_k$), Markov spectral moments (${}^{\mathrm{LN}}\pi_k$) and Markov entropies (${}^{\mathrm{LN}}\theta_k$) of a Markov model (MM) associated to a 2D network that numerically characterize DNA sequences and build a QSAR model to predict Mps. The lattice-like representations (also called maps or graphs) for Mps and control group sequences (Cgs) were derived (González-Díaz et al., 2003, 2006a, 2005c; González-Díaz, 2007d). The ξ_k, π_k and θ_k values of several types of graphs/networks have been the base for different QSAR studies of DNA/RNA and protein sequences (Du et al., 2007a, b; Garcia-Garcia et al., 2004; Marrero-Ponce et al., 2004a, b, 2005b; Meneses-Marcel et al., 2005; Santana et al., 2006). The second group of TIs is derived from the SG representations (Harary, 1969). We subsequently developed a classifier to connect Mps information (represented by the ξ_k, π_k, θ_k and SG TIs values) with the prediction of Cgs as Mps. The Linear Discriminant Analysis (LDA) was selected as a simple but powerful technique (González-Díaz et al., 2006b; González-Díaz, 2003a).

2. Materials and methods

2.1. Pseudo-folding LN

The first MM, also called MARCH-INSIDE, was used to codify the information of 135 Mps (González-Díaz et al., 2005a, 2006a, 2007d) and 511 random Cgs (see Table S.1 in the Supplementary material). Our methodology considers as states of the Markov Chain (MC) any atom, nucleotide or amino acid depending on the class of molecule to be described (González-Díaz et al., 2005e, 2003b). Therefore, MM deals with the calculation of the probabilities (${}^{k}p_{ij}$) where the charge distribution of nucleotide moves from any nucleotide in the vicinity i at time t_0 to another nucleotide j along the protein backbone in discrete time periods, until a stationary state is achieved (Yuan, 1999). As seen from the discussion above, we selected ${}^{\mathrm{LN}}\xi_{k,}$ ${}^{\mathrm{LN}}\pi_k$ and ${}^{\mathrm{LN}}\theta_k$ based on the utility of non-stochastic (González-Díaz and Uriarte, 2005; González-Díaz et al., 2005d; Ramos de Armas et al., 2004) and stochastic parameters (Randic and Vracko, 2000). Many researchers have demonstrated the possibility of predicting RNA from sequences (Aguero-Chapin et al., 2006) and we used 2D graphs to encode information about Mps sequences (Estrada, 2000, 2002; Estrada and González-Díaz, 2003; González-Díaz et al., 2005b; González and Moldes del Carmen Teran, 2004; González et al., 2005, 2006; Vilar et al., 2005, 2006). This RNA 2D graphical representation is similar to those previously reported for DNA (Jacchieri, 2000; Nandy, 1994, 1996) using four different nucleotides. The construction of the 2D lattice graph corresponding to the Mps of the gene Alpha in *Mycobacterum bovis* (BCG) is shown in Table 1 and Fig. 1. Each nucleotide in the sequence is placed in a Cartesian 2D space starting with the first monomer at the (0, 0) coordinates. The coordinates of the successive nucleotide are calculated according to with the following rules:

(a) Increase by +1 the abscissa axis coordinate for thymine (rightwards-step) or
(b) Decrease by −1 the abscissa axis coordinate for cytosine (leftwards-step) or
(c) Increase by +1 the ordinate axis coordinate for adenine (upwards-step) or
(d) Decrease by −1 the ordinate axis coordinate for guanine (downwards-step).

In the next step, we assigned a stochastic matrix ${}^{1}\mathbf{\Pi}$ to each graph. The elements of ${}^{1}\mathbf{\Pi}$ are the probabilities ${}^{1}p_{ij}$ of reaching a node n_i with the charge Q_i moving through a walk of length of $k = 1$ from another node n_j with charge Q_j (Aguero-Chapin et al., 2006):

$$p_{ij} = \frac{\dfrac{Q_j}{d_{j0}}}{\sum_{m=l}^{n} \alpha_{il} \cdot \dfrac{Q_j}{d_{l0}}} = \frac{\varphi_j}{\sum_{m=l}^{n} \alpha_{il} \cdot \varphi_l} \tag{1}$$

Table 1
LN construction rules for the Mps of the gene Alpha in *Mycobacterum bovis* (BCG)

DNA LN
$c_1g_2a_3c_4t_5t_6t_7c_8g_9c_{10}c_{11}c_{12}g_{13}a_{14}a_{15}t_{16}c_{17}g_{18}a_{19}c_{20}$
$a_{21}t_{22}t_{23}t_{24}g_{25}g_{26}c_{27}c_{28}t_{29}c_{30}c_{31}a_{32}c_{33}a_{34}c_{35}a_{36}c_{37}g_{38}g_{39}t_{40}$
$a_{41}t_{42}g_{43}t_{44}t_{45}c_{46}t_{47}g_{48}g_{49}c_{50}c_{51}c_{52}g_{53}a_{54}g_{55}c_{56}a_{57}c_{58}a_{59}c_{60}$
$g_{61}a_{62}c_{63}g_{64}a_{65}$

n	Nucleotide	*x*	*y*
1	$c_1a_3t_5g_{25}$	0	0
2	$g_2c_{10}g_{26}$	0	−1
3	c_4t_{16}	−1	0
4	t_6c_8	1	0
5	t_7	2	0
6	g_9	1	−1
7	$c_{11}c_{27}t_{29}$	−1	-1
8	$c_{12}a_{14}g_{18}c_{28}c_{30}g_{48}$	−2	−1
9	$g_{13}g_{49}$	−2	−2
10	$a_{15}c_{17}a_{19}t_{45}t_{47}$	−2	0
11	$c_{20}a_{32}t_{44}c_{46}$	−3	0
12	a_{21}	−3	1
13	t_{22}	−2	1
14	t_{23}	−1	1
15	t_{24}	0	1
16	c_{31}	−3	−1
17	$c_{33}g_{43}$	−4	0
18	$a_{34}t_{42}$	−4	1
19	$c_{35}a_{41}$	−5	1
20	a_{36}	−5	2
21	c_{37}	−6	2
22	g_{38}	−6	1
23	g_{39}	−6	0
24	t_{40}	−5	0
25	c_{50}	−3	−2
26	c_{51}	−4	−2
27	$c_{52}a_{54}$	−5	−2
28	$g_{53}g_{55}$	−5	−3
29	c_{56}	−6	−3
30	a_{57}	−6	−2
31	c_{58}	−7	−2
32	a_{59}	−7	−1
33	$c_{60}a_{62}$	−8	−1
34	g_{61}	−8	−2
35	$c_{63}a_{65}$	−9	−1
36	g_{64}	−9	−2

$$p_j = \frac{\dfrac{Q_j}{d_{j0}}}{\sum_{m=l}^{n}\dfrac{Q_j}{d_{l0}}} = \frac{\varphi_j}{\sum_{m=l}^{n}\varphi_l} \qquad (2)$$

where α_{ij} equals to 1 if the nodes n_i and n_j are adjacent in the graph or equal to 0 otherwise; Q_j is equal to the sum of the electrostatic charges of all nucleotides placed at this node. Note that the number of nodes (n) in the graph is equal to the number of rows and columns in ${}^1\mathbf{\Pi}$ but it may be equal or even smaller than the number of DNA bases in the sequence. It then becomes straightforward to calculate different types of invariant parameters for ${}^1\mathbf{\Pi}$ in order to numerically characterize the DNA sequence. In this work we calculated the following invariants:

$${}^{\mathrm{LN}}\pi_k = \sum_{i=j}^{n} {}^k p_{ij} \qquad (3)$$

$${}^{\mathrm{LN}}\xi_k = \sum_{i=j}^{n} {}^k p_j \cdot \varphi_j \qquad (4)$$

$${}^{\mathrm{LN}}\theta_k = -\sum_{i=j}^{n} {}^k p_j \cdot \log({}^k p_j) \qquad (5)$$

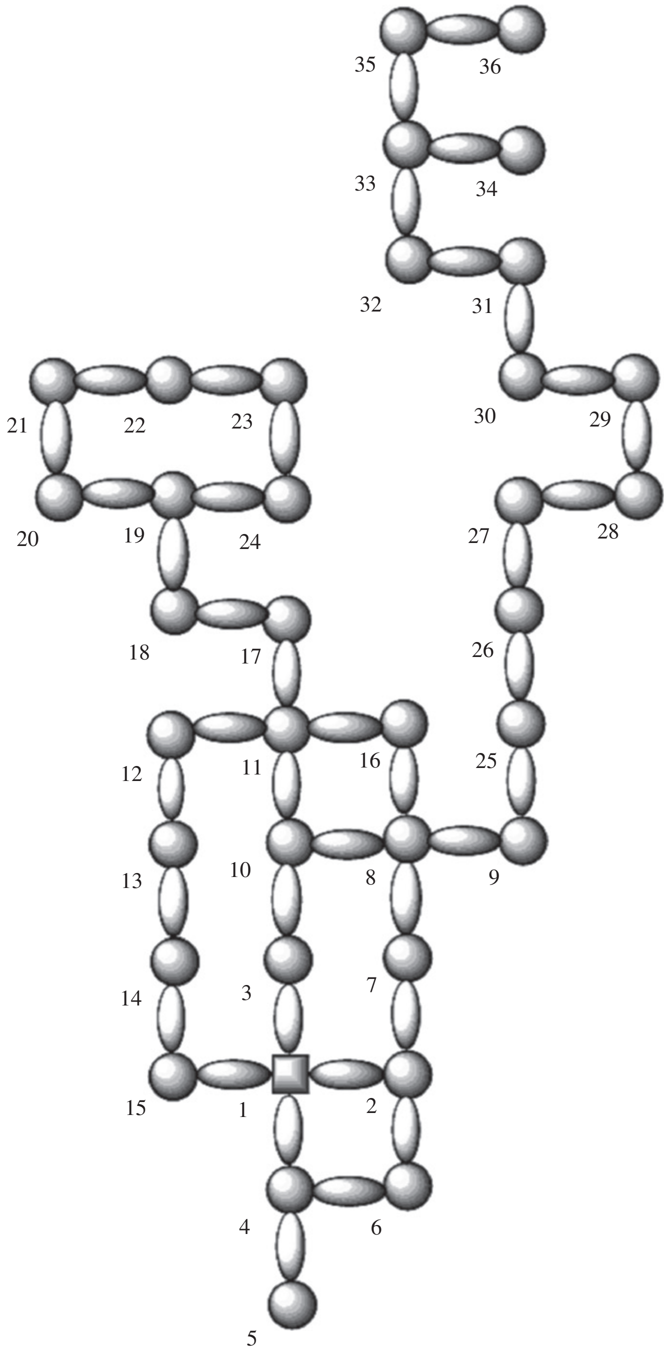

Fig. 1. LN for the Mps of the gene Alpha in *Mycobacterum bovis* (BCG).

where ${}^{\mathrm{LN}}\pi_k$ are the Markov spectral moments and indicate that we sum all the values in the main diagonal of the matrices ${}^{\mathrm{LN}}\pi_k = \mathrm{tr}({}^k\mathbf{\Pi}) = \mathrm{tr}[({}^1\mathbf{\Pi})^k]$ (tr is the trace operator), ${}^{\mathrm{LN}}\xi_k$ are the mean values of electrostatic potentials and ${}^{\mathrm{LN}}\theta_k$ are the Markov entropies (González-Díaz et al., 2007e). All calculations of the ${}^{\mathrm{LN}}\xi_k$, ${}^{\mathrm{LN}}\pi_k$ and ${}^{\mathrm{LN}}\theta_k$ values for the DNA sequences of both groups (Mps and Cgs) were carried out with our in-house software MARCH-INSIDE, *version* 2.0® (González-Díaz et al., 2007e), including sequence representation.

2.2. SG topological indices

Each DNA sequence is a real network where the nucleotides are the vertices/nodes, connected in a specific sequence by the phosphodiester bonds. SG is an abstract representation of the real network that has a dummy non-nucleotide centre and a number of "rays" equal with the nucleotide types. In the case of DNA, we can consider two codifications: the nucleotide code (as in the case of the amino acid protein sequences) and the DNA codons (the final incomplete codons are ignored). In the first codification, there are only four branches ("rays") of the star corresponding to the four types of nucleotides: adenosine (a), thymidine (t), cytidine (c) and guanosine (g). Using the codons, the DNA sequences are virtually translated into amino acid sequences that generate 21 branches, 20 standard amino acids and an extra *X* non-amino acid corresponding to the STOP DNA codons (Griffiths et al., 1999). Even if the promoters are not naturally translated in proteins, the second codification is useful for a comparison with the protein SG calculations. The same DNA/protein can be represented by different forms associated to distinct distance matrices (Randic et al., 2007). Standard SG were constructed for each DNA promoter: each nucleotide/vertex holds the position in the original sequence and the branches are labelled by the standard letters of the nucleotides (a, t, c and g). If the initial connectivity in the DNA sequence is included, the graph is embedded. In order to qualitatively evaluate the graphs, it is necessary to transform the graphical representation into correspondent connectivity matrix, distance matrix and degree matrix. In the case of embedded graph, the matrices of the connectivity in the sequence and in the SG are combined. These matrices and the normalized ones are the base of the calculation of the topological indices.

For a visual comparison of the lattice and SG representations, the same promoter sequence from Table 1 was used to generate a standard SG based on codons that are virtually translated to amino acids (see Table 2 and Fig. 2).

The SG topological indices are obtained with the in-house sequence to star networks (S2SNet) python application. This tool can transform any character string in SG topological indices. Our recent works (Munteanu et al., 2008a, b) proved the potential of S2SNet in protein QSAR models. The calculations presented in this work are characterized by embedded (E) and non-embedded (nE) TIs, non-weights, Markov normalization and power of matrices/indices (*n*) up to 5. The result file contains the following embedded (super index "*e*") or non-embedded TIs (Todeschini and Consonni, 2002):

Shannon Entropy of the *n* powered Markov matrices (${}^{SG}\theta_n$):

$$ {}^{SG}\theta_n^{(e)} = -\sum_i p_i * \log(p_i) \tag{6} $$

where p_i are the n_i elements of the p vector, resulted from the matrix multiplication of the powered Markov normalized matrix $(n_i \times n_i)$ and a vector $(n_i \times 1)$ with each element equal to $1/n_i$;

The trace of the *n* connectivity matrices (${}^{SG}\pi_n$):

$$ {}^{SG}\pi_n^{(e)} = -\sum_i (M^n)_{ii} \tag{7} $$

where $n = 0$–power limit, ${}^{SG}M = \text{SG}$ connectivity matrix ($i*i$ dimension); $ii = i$th diagonal element;

Harary number (H):

$$ H^{(e)} = \sum_{i<j} (m_{ij}/d_{ij}) \tag{8} $$

where d_{ij} are the elements of the distance matrix and m_{ij} are the elements of the M connectivity matrix;

Wiener index (W):

$$ W^{(e)} = \sum_{i<j} d_{ij} \tag{9} $$

Gutman topological index (S_6):

$$ S_6^{(e)} = \sum_{ij} deg_i * deg_j / deg_{ij} \tag{10} $$

where deg_i are the elements of the degree matrix;

Schultz topological index (non-trivial part) (S):

$$ S^{(e)} = \sum_{i<j} (deg_i + deg_j) * d \tag{11} $$

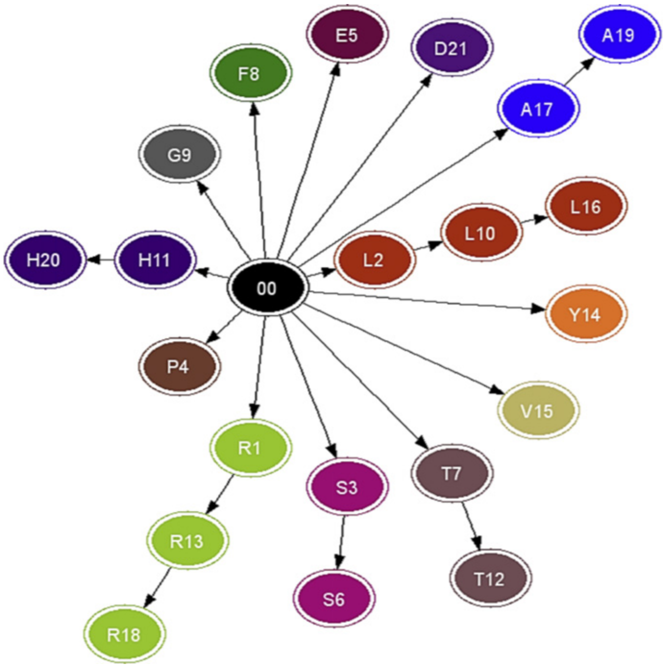

Fig. 2. SG for the Mps of the gene Alpha in *Mycobacterum bovis* (BCG).

Table 2
SG codifications for the virtually translated Mps of the gene Alpha in *Mycobacterum bovis* (BCG)

DNA codon SG	
DNA nucleotide sequence	$c_1g_2a_3c_4t_5t_6t_7c_8g_9c_{10}c_{11}c_{12}g_{13}a_{14}a_{15}t_{16}c_{17}g_{18}a_{19}c_{20}a_{21}$ $t_{22}t_{23}t_{24}g_{25}g_{26}c_{27}c_{28}t_{29}c_{30}c_{31}a_{32}c_{33}a_{34}c_{35}a_{36}c_{37}g_{38}g_{39}$ $t_{40}a_{41}t_{42}g_{43}t_{44}t_{45}c_{46}t_{47}g_{48}g_{49}c_{50}c_{51}c_{52}g_{53}a_{54}g_{55}c_{56}a_{57}c_{58}a_{59}c_{60}g_{61}a_{62}c_{63}$
DNA codons sequence	$cga_1ctt_2tcg_3ccc_4gaa_5tcg_6aca_7$ $ttt_8ggc_9ctc_{10}cac_{11}aca_{12}cgg_{13}$ $tat_{14}gtt_{15}ctg_{16}gcc_{17}cga_{18}gca_{19}cac_{20}gac_{21}$
Virtually translated amino acid sequence	$R_1L_2S_3P_4E_5S_6T_7F_8G_9L_{10}H_{11}T_{12}R_{13}Y_{14}V_{15}L_{16}A_{17}R_{18}A_{19}H_{20}D_{21}$

Balaban distance connectivity index (J):

$$J^{(e)} = (edges - nodes + 2) * \sum_{i<j} m_{ij} * \mathrm{sqrt}\left(\sum_k d_{ik} * \sum_k d_{kj}\right) \tag{12}$$

where $nodes$+1 = AA numbers/node number in the SG+origin, $\sum_k d_{ik}$ is the node distance degree;

Kier–Hall connectivity indices (^nX):

$$^0X^{(e)} = \sum_i 1/\mathrm{sqrt}(deg_i) \tag{13}$$

$$^0X^{(e)} = \sum_{i<j<k} m_{ij} * m_{jk}/\mathrm{sqrt}(deg_i * deg_j * deg_k) \tag{14}$$

$$^3X^{(e)} = \sum_{i<j<k<m} m_{ij} * m_{jk} * m_{km}/\mathrm{sqrt}(deg_i * deg_j * deg_k * deg_m) \tag{15}$$

$$^4X^{(e)} = \sum_{i<j<k<m<0} m_{ij} * m_{jk} * m_{km} * m_{mo}/\mathrm{sqrt}(deg_i * deg_j * deg_k * deg_m * deg_o) \tag{16}$$

$$^5X^{(e)} = \sum_{i<j<k<m<o<q} m_{ij} * m_{jk} * m_{km} * m_{mo} * m_{oq}/\mathrm{sqrt}(deg_i * deg_j * deg_k * deg_m * deg_o * deg_q) \tag{17}$$

Randic connectivity index (^1XR):

$$^1XR^{(e)} = \sum_{i<j} m_{ij}/\mathrm{sqrt}(deg_i * deg_j) \tag{18}$$

The embedded and non-embedded SG TIs are used to construct a DNA promoter classification model using the LDA statistical methods.

2.3. Linear discriminant analysis

LDA forward stepwise analysis from STATISTICA (StatSoft.Inc., 2002) was carried out for a variable selection to build up the model (Garcia-Garcia et al., 2004; Kutner et al., 2005; Marrero-Ponce et al., 2004a, b, 2005b; Meneses-Marcel et al., 2005; Santana et al., 2006). In order to decide whether a DNA sequence is classified as a mycobacterial promoter (Prom) or not (nProm), we added an extra dummy variable named Prom/nProm (binary values of 1/−1 for LN and 1/0 for SG) and a cross-validation variable (CV). The best cross-validation methods in practice are the independent dataset test, the subsampling test and the jackknife test (Chou and Zhang, 1995). The jackknife test has been increasingly used by investigators to examine the accuracy of various predictors (Chen and Li, 2007; Chou and Shen, 2007a, 2008; Diao et al., 2007; Ding et al., 2007; Lin, 2008; Xiao and Chou, 2007). In the actual work, the independent data test is used by splitting the data at random in a training series (train, 75%) used for model construction and a prediction one (val, 25%) for model validation (the CV column is filled by repeating 3 train and 1 val). All the variables included in the models were standardized in order to bring them onto the same scale. Subsequently, standardized linear discriminant equations that allow comparison of their coefficients were obtained (Chiti et al., 2003; Pawar et al., 2005).

In the case of LN, the general QSAR formula is the following:

$$^{\mathrm{LN}}\mathrm{Mps\text{-}score} = a_0 + \sum_{k=0}^{5} b_k \times {}^{\mathrm{LN}}\pi_k + \sum_{k=0}^{5} c_k \times {}^{\mathrm{LN}}\theta_k + \sum_{k=0}^{5} d_k \times {}^{\mathrm{LN}}\xi_k \tag{19}$$

where $^{\mathrm{LN}}$Mps-score is the continue score value for the DNA mycobacterial promoter classification corresponding to the lattice representation, $^{\mathrm{LN}}\pi_k$ are Markov spectral moments (traces), $^{\mathrm{LN}}\theta_k$ are the Markov entropies, $^{\mathrm{LN}}\xi_k$ the mean stochastic electrostatic potential, b_k, c_k, d_k are the coefficients of the previous indices and a_0 is the independent term. A similar formula defines the SG QSAR model in Eq. (20).

$$^{\mathrm{SG}}\mathrm{Mps\text{-}score} = e_0 + \sum_{k=0}^{5} f_k \times {}^{\mathrm{SG}}\pi_k + \sum_{k=0}^{5} f_k^e \times {}^{\mathrm{SG}}\pi_k^e + \sum_{k=0}^{5} g_k \times {}^{\mathrm{SG}}\theta_k + \sum_{k=0}^{5} g_k^e \times {}^{\mathrm{SG}}\theta_k^e + \sum_{k=0}^{10} e_k \times \mathrm{TI}_k + \sum_{k=0}^{10} e_k \times \mathrm{TI}_k^e \tag{20}$$

where $^{\mathrm{SG}}$Mps-score is the continue score value for the DNA mycobacterial promoter classification corresponding to the SG representation, $^{\mathrm{SG}}\pi_k{}^e/^{\mathrm{SG}}\pi_k$ and $^{\mathrm{SG}}\theta_k{}^e/^{\mathrm{SG}}\theta_k$ are embedded/non-embedded traces (Markov spectral moments) and the Shannon entropies, $\mathrm{TI}_k{}^e/\mathrm{TI}_k$ are the other 22 standard SG embedded and non-embedded TIs (H, W, S_6, S, J, ^{0}X, $^{2-5}$X, ^1XR, H^e, W^e, S_6^e, S^e, J^e, $^0X^e$, $^{2-5}X^e$, $^1XR^e$), $f_k{}^e/f_k$, $g_k{}^e/g_k$ and $e_k{}^e/e_k$ are the TIs coefficients and e_0 is the independent term. Accuracy, specificity, sensitivity, F, Wilk's (λ) statistic ($\lambda = 0$ perfect discrimination, being $0<\lambda<1$) were examined in order to assess the discriminatory power of the model.

3. Results and discussion

Many different parameters can be used to encode RNA sequence information and further assign or predict the function or physical properties (González-Díaz and Uriarte, 2005). The present approach involves the calculation of different sequence parameters, which can be applied to different types of molecular graphs (Aguero-Chapin et al., 2006), including DNA, RNA and proteins (Di Francesco, 1999; González-Díaz et al., 2005c). MM has been applied successfully to Genomics and Proteomics and represents an important tool to analyse biological sequence data. In particular, MM has been used for protein folding recognition (Chou, 2001b) and for prediction of protein signal sequences (Chou and Shen, 2007b; Van Waterbeemd, 1995). This work compared two models based on different TIs including π_k and θ_k values of the stochastic matrices $^1\mathbf{\Pi}$(LN) and $^1\mathbf{\Pi}$(SG) ($^{\mathrm{SG}}$M) associated with LN and SG, $^{\mathrm{LN}}\xi_k$ parameters of $^1\mathbf{\Pi}$(LN) as well as classic TIs for $^1\mathbf{\Pi}$(SG). These parameters describe the distribution of the nucleotides of the DNA sequence in the above graphs/networks. This calculation was carried out for two groups of DNA sequences, one made up of Mps and the other formed by Cgs. In addition, previous results of the RNA secondary structure (2S) QSAR are compared.

3.1. Results for DNA LN indices

In the first study of the DNA LN representations, the best QSAR equation that classifies a novel sequence as Mps or not is the following (Table 3):

$$^{\mathrm{LN}}\mathrm{Mps\text{-}score} = -1.2 - 4.1 \times {}^{\mathrm{LN}}\xi_1 + 2.1 \times {}^{\mathrm{LN}}\xi_5 \tag{21}$$

The statistical parameters of this equation were Wilk's statistic ($\lambda = 0.95$) and the error level (p-level<0.001). This discriminant function misclassified only 36 cases out of 511 Cgs used, reaching a high level of accuracy (90.87%). More specifically, the model classified correctly 112/135 (82.9%) of Mps and 475/511 (92.9%) of the control group. Conversely, the remaining four descriptors $^{\mathrm{LN}}\xi_{0,}$ $^{\mathrm{LN}}\xi_{2,}$ $^{\mathrm{LN}}\xi_3$ and $^{\mathrm{LN}}\xi_4$ do not have a significant relationship with the Mps characteristic. The use of only six molecular descriptors to

Table 3
Summary of the LDA results for DNA LN and SG models vs. RNA 2S folding representations

TI	Ac (%)	Se (%)	Sp (%)	Final TIs	Vars.	λ	F	p	Ref.
Primary structure of DNA nucleotide & LN									
${}^{LN}\theta_k$	78.33	72.59	79.84	${}^{LN}\theta_0$	1	0.74	230.5	0.0001	a
${}^{LN}\pi_k$	81.73	78.52	82.58	${}^{LN}\pi_0$, ${}^{LN}\pi_1$, ${}^{LN}\pi_5$,	3	0.89	76.3	0.0001	a
${}^{LN}\xi_k$	90.87	82.96	92.95	${}^{LN}\xi_1$, ${}^{LN}\xi_5$	2	0.82	142.1	0.0001	a
Pool	92.88	75.56	97.46	${}^{LN}\theta_0$, ${}^{LN}\pi_0$, ${}^{LN}\xi_1$, ${}^{LN}\xi_5$	4	0.83	130.8	0.0001	a
Primary structure of DNA nucleotide sequences & SG									
${}^{SG}\theta_k$	66.25	81.48	62.23	${}^{SG}\theta_1^e$, ${}^{SG}\theta_4^e$	2	0.78	69.62	0.001	a
${}^{SG}\pi_k$	71.21	85.19	67.51	${}^{SG}\pi_0^e$, ${}^{SG}\pi_2^e$, ${}^{SG}\pi_5^e$	3	0.76	49.54	0.001	a
TI$_k$	75.39	68.15	77.30	W, J^e, ${}^0X^e$	3	0.73	58.19	0.001	a
Pool	81.58	68.15	85.13	${}^{SG}\pi_5^e$, H, ${}^1XR^e$	3	0.67	79.94	0.001	a
Primary structure of DNA *codon* sequences & SG									
${}^{SG}\theta_k$	70.43	76.30	68.88	${}^{SG}\theta_0$, ${}^{SG}\theta_1$, ${}^{SG}\theta_4^e$	3	0.75	52.31	0.001	a
${}^{SG}\pi_k$	74.77	82.96	72.60	${}^{SG}\pi_4$, ${}^{SG}\pi_4^e$, ${}^{SG}\pi_5^e$	3	0.74	56.37	0.001	a
TI$_k$	76.16	59.26	80.63	S, ${}^0X$, ${}^1XR^e$	3	0.72	60.98	0.001	a
Pool	80.80	74.81	82.39	${}^{SG}\theta_0$, ${}^{SG}\theta_4^e$, ${}^{SG}\pi_4^e$, ${}^{SG}\pi_5^e$, W	5	0.67	47.04	0.001	a
RNA electrostatic parameter of thermodynamically-driven 2S folding									
${}^{2S}\theta_k$	97.60	93.30	100.00	${}^{2S}\theta_0$	1	0.34	724.47	0.001	b
${}^{2S}\pi_k$	93.83	83.70	98.89	${}^{2S}\pi_0$, ${}^{2S}\pi_2$	2	0.44	515.03	0.05	c
${}^{2S}\xi_k$	96.58	85.19	100.00	${}^{2S}\xi_0$, ${}^{2S}\xi_1$	2	0.41	38.8	0.001	d

Note: the terms Ac, Se, and Sp mean accuracy, sensitivity and specificity, and measure the ratio of the number of total, Mps, or Cgs sequences correctly classified by the model with respect to the real classification; Vars. = no of variables in the QSAR equations; SG = star-graph; LN = lattice network; 2S = secondary structure; super index "*e*" represents the embedded calculations; references (Ref.) are a: this work, b: (González-Díaz et al., 2007c), c: (González-Díaz et al., 2005a) and d: (González-Díaz et al., 2006a).

model a data set of 585 sequences prevents us by large from chance correlation. In physical terms, the above results confirm other studies about the relationship between the electrostatic potential of the DNA molecule and its biological activity. However, in this case, not all the electrostatic interactions affect the activity in the same way. Finally, long-term electrostatic interaction potentials (${}^{LN}\xi_{0,}$ ${}^{LN}\xi_{2,}$ ${}^{LN}\xi_3$ and ${}^{LN}\xi_4$) do not correlate with the Mps activity. The detailed results of the forward stepwise analysis are given in Table 3.

Analyzing the above equations, it is important to highlight that, the combination of a negative contribution of ${}^{LN}\xi_1$ and a positive contribution of ${}^{LN}\xi_5$ in Eq. (21) point to a pseudo-folding rule for the biological activity. A validation procedure was subsequently performed in order to assess the model predictability. This validation was carried out with an external series of Mps and randomized Cgs. The present model showed an accuracy of 90.87%, which is similar in comparison to results obtained by other researchers when using the LDA method in QSAR studies (González-Díaz et al., 2007f). These results are also consistent with many others that we have recently reviewed in-depth and published as a review article where we used different network-like indices in small-sized, nucleic acid, and protein QSAR (González-Díaz et al., 2005d, 2007d, f; Marrero-Ponce et al., 2005a, b; Van Waterbeemd, 1995).

3.2. Results for DNA SG indices

The second study used the SG–QSAR models in order to evaluate the same mycobacterial DNA promoter property (see Table 3). The grouping of the embedded and non-embedded TIs was done similar to the lattice models: the traces (${}^{SG}\pi_k{}^e/{}^{SG}\pi_k$), the Shannon entropies (${}^{SG}\theta_k{}^e/{}^{SG}\theta_k$), the rest of embedded and non-embedded TIs (H, W, S_6, S, J, ${}^0X$, ${}^{2-5}X$, ${}^1XR$, H^e, W^e, S_6^e, S^e, J^e, ${}^0X^e$, ${}^{2-5}X^e$, ${}^1XR^e$) and all SG TIs (pool). The *forward stepwise* selection variable method, conjugated with the nE & E TIs of the virtually translated DNA sequences, provides better results for the codon grouping of the nucleotides, with accuracy, sensitivity and specificity greater than 70% for the ${}^{SG}\pi_k{}^e/{}^{SG}\pi_k$ and for the pool (Table 3). Even if the accuracy of the simple nucleotide sequences are up to 81.58% (pool), the selectivity and the specificity have values lower than 70%. The best QSAR model using the SG based on the codon sequences is defined with the ${}^{SG}\pi_k{}^e/{}^{SG}\pi_k$ group of indices in Eq. (22) and is characterized by 74.77% accuracy, 82.96% sensitivity and 72.60% specificity.

$$ {}^{SG}\text{Mps-score} = -1.9 + 1.3 \times {}^{SG}\pi_4 - 1.9 \times {}^{SG}\pi_4^e - 1.2 \times {}^{SG}\pi_5^e \qquad (22) $$

Despite the good values of accuracy, sensitivity and specificity (80.80%, 74.81%, 82.39%) for the pool group of TIs (${}^{SG}\theta_0$, ${}^{SG}\theta_4^e$, ${}^{SG}\pi_4^e$, ${}^{SG}\pi_5^e$, W), the QSAR model cannot be considered due to the low sensitivity for the CV set (66.67%). Thus, the results based on the traces (spectral moments) are similar in the case of LN and SG representations, maintaining the ${}^{SG}\pi_5^e/{}^{LN}\pi_5$ in the equations.

3.3. Comparison with RNA 2S and other indices

In previous works, we have published QSAR models to predict Mps using RNA electrostatic parameters of thermodynamically-driven 2S folding representations. These models were based on the ${}^{2S}\theta_k$ (González-Díaz et al., 2007c), ${}^{2S}\pi_k$ (González-Díaz et al., 2005a) and ${}^{2S}\xi_k$ (González-Díaz et al., 2006a) values for the ${}^1\mathbf{\Pi}(2S)$ matrix associated to RNA 2S folding representations. In Table 3 we illustrate that the best values of accuracy, sensitivity and specificity of 97.60%, 93.30% and 100% were found for ${}^{2S}\theta_0$. This TI is present in the QSAR equations for DNA LN/SG and RNA 2S folding representations. All these observations pointed out the importance of the spectral moments, entropies and in the stochastic electrostatic potentials in the DNA/RNA QSAR models. In general, the results for RNA 2S folding representation are better, but they require additional calculations for the optimization of

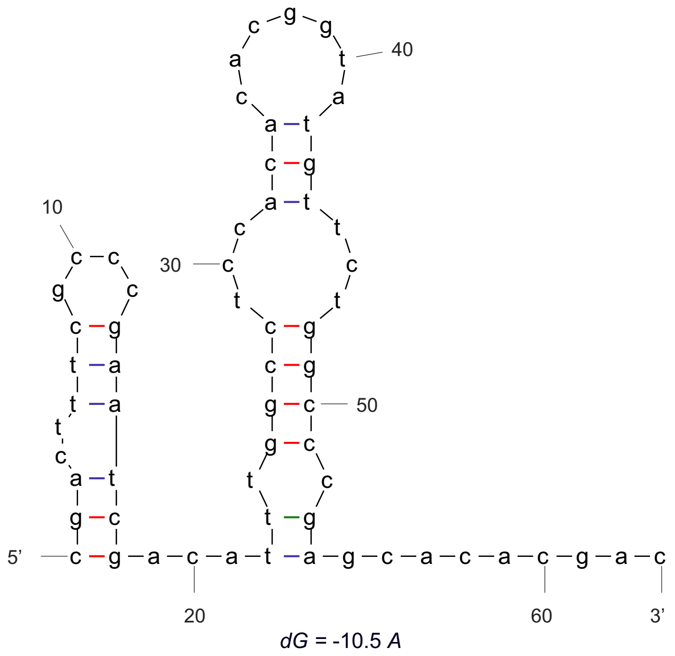

Fig. 3. RNA 2S for the Mps of the gene Alpha in *Mycobacterium bovis* (BCG).

RNA 2S. Therefore, several RNA 2S are possible for the same DNA sequence (theoretically because the promoters have no correspondent RNA) by introducing an indeterminacy in the final model prediction. In Fig. 3 we depict a possible 2S for the RNA sequence corresponding to the DNA sequences used in Figs. 1 and 2 (dG is the free energy). This RNA 2S was obtained with the online DINAMelt server (Markham and Zuker, 2005). The SG TIs that show to not be important for the DNA/RNA models (*H*, *W*, *S*, *J*) can successfully describe protein QSAR models (Munteanu et al., 2008b). This work pointed out the conclusion that the models based on SG, LN, as well as 2S, which are linear and have few variables, can be compared very favourably in terms of complexity with other models previously reported by Kalate et al. (2003). These authors used a non-linear artificial neural network and a large parameter space.

5. Conclusions

The work presents a comparative study of the parameters associated with LN and SG representations in order to predict the mycobacterial DNA promoters. LN QSAR classifier successfully discriminates between Mps and a control group, with values significantly better than the SG-QSAR results based on the DNA codon sequences. In addition, the DNA nucleotide sequences (used for LN) were not able to create a good model based on SG representations. The work promotes the use of the experience accumulated in small-molecules QSAR with spectral moments and other kind of indices (entropies and spectral moments) in new types of DNA QSAR studies, today in the focus of attention of many researchers worldwide.

Acknowledgements

Cristian R. Munteanu thanks the FCT (Portugal) for support from grant SFRH/BPD/24997/2005. González-Díaz H. acknowledges the program Isidro Parga Pondal of Xunta de Galicia for the financial support of a tenure-eligible research position at the Faculty of Pharmacy, University of Santiago de Compostela (Spain). The authors thank financial support from grant INCITE07PXI203141ES, Conselleria de Industria, Xunta de Galicia, Spain.

Appendix A. Supporting Information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jtbi.2008.09.035.

References

Aguero-Chapin, G., González-Díaz, H., Molina, R., Varona-Santos, J., Uriarte, E., González-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, 723–730.

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., Reusser, F., 1993a. 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., Reusser, F., 1993b. The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. J. Biol. Chem. 268, 14875–14880.

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., Reusser, F., 1993c. Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E. Biochemistry 32, 6548–6554.

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. Can. J. Chem. 86, 342–357.

Arnvig, K.B., Gopal, B., Papavinasasundaram, K.G., Cox, R.A., Colston, M.J., 2005. The mechanism of upstream activation in the rrnB operon of Mycobacterium smegmatis is different from the Escherichia coli paradigm. Microbiology 151, 467–473.

Caballero, J., Fernandez, L., Abreu, J.I., Fernandez, M., 2006. Amino acid sequence autocorrelation vectors and ensembles of Bayesian-regularized genetic neural networks for prediction of conformational stability of human lysozyme mutants. J. Chem. Inf. Model. 46, 1255–1268.

Chen, Y.L., Li, Q.Z., 2007. Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition. J. Theor. Biol. 248, 377–381.

Chiti, F., Stefani, M., Taddei, N., Ramponi, G., Dobson, C.M., 2003. Rationalization of the effects of mutations on peptide and protein aggregation rates. Nature 424, 805–808.

Chou, K.C., 1981. Two new schematic rules for rate laws of enzyme-catalyzed reactions. J. Theor. Biol. 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., 2001a. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 43, 246–255.

Chou, K.C., 2001b. Prediction of signal peptides using scaled window. Peptides 22, 1973–1979.

Chou, K.C., 2004. Review: structural bioinformatics and its impact to biomedical science. Curr. Med. Chem. 11, 2105–2134.

Chou, K.C., 2005. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21, 10–19.

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

Chou, K.C., Liu, W.M., 1981. Graphical rules for non-steady state enzyme kinetics. J. Theor. Biol. 91, 637–654.

Chou, K.C., Shen, H.B., 2006. Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochem. Biophys. Res. Commun. 347, 150–157.

Chou, K.C., Shen, H.B., 2007a. Recent progress in protein subcellular location prediction. Anal. Biochem. 370, 1–16.

Chou, K.C., Shen, H.B., 2007b. Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochem. Biophys. Res. Commun. 357, 633–640.

Chou, K.C., Shen, H.B., 2008. Cell-PLoc: a package of web-servers for predicting subcellular localization of proteins in various organisms. Nat. Protocols 3, 153–162.

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

Chou, K.C., Zhang, C.T., 1995. Prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 30, 275–349.

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

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

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

Di Francesco, V., Munson, P.J., Garnier, J., 1999. FORESST: fold recognition from secondary structure predictions of proteins. Bioinformatics 15, 131–140.

Diao, Y., Li, M., Feng, Z., Yin, J., Pan, Y., 2007. The community structure of human cellular signaling network. J. Theor. Biol. 247, 608–615.

Ding, Y.S., Zhang, T.L., Chou, K.C., 2007. Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. Protein. Pept. Lett. 14, 811–815.

Dobson, P.D., Doig, A.J., 2005. Predicting enzyme class from protein structure without alignments. J. Mol. Biol. 345, 187–199.

Dobson, P.M., Boyle, M., Loewenthal, M., 2004. Home intravenous antibiotic therapy and allergic drug reactions: is there a case for routine supply of anaphylaxis kits? J. Infus. Nurs. 27, 425–430.

Dobson, P.S., Weaver, J.M., Holder, M.N., Unwin, P.R., Macpherson, J.V., 2005. Characterization of batch-microfabricated scanning electrochemical-atomic force microscopy probes. Anal. Chem. 77, 424–434.

Du, Q.S., Wei, Y.T., Pang, Z.W., Chou, K.C., Huang, R.B., 2007a. Predicting the affinity of epitope-peptides with class I MHC molecule HLA-A*0201: an application of amino acid-based peptide prediction. Protein Eng. Des. Sel. 20, 417–423.

Du, Q.S., Huang, R.B., Wei, Y.T., Wang, C.H., Chou, K.C., 2007b. Peptide reagent design based on physical and chemical properties of amino acid residues. J. Comput. Chem. 28, 2043–2050.

Du, Q.S., Huang, R.B., Wei, Y.T., Du, L.Q., Chou, K.C., 2008. Multiple field three dimensional quantitative structure–activity relationship (MF-3D-QSAR). J. Comput. Chem. 29, 211–219.

Estrada, E., 2000. On the topological sub-structural molecular design (TOSS-MODE) in QSPR/QSAR and drug design research. SAR QSAR Environ. Res. 11, 55–73.

Estrada, E., 2002. Characterization of the folding degree of proteins. Bioinformatics 18, 697–704.

Estrada, E., González-Díaz, H., 2003. What are the limits of applicability for graph theoretic descriptors in QSPR/QSAR? Modeling dipole moments of aromatic compounds with TOPS-MODE descriptors. J. Chem. Inf. Comput. Sci. 43, 75–84.

Gao, L., Ding, Y.S., Dai, H., Shao, S.H., Huang, Z.D., Chou, K.C., 2006. A novel fingerprint map for detecting SARS-CoV. J. Pharm. Biomed. Anal. 41, 246–250.

Garcia-Garcia, A., Galvez, J., de Julian-Ortiz, J.V., Garcia-Domenech, R., Munoz, C., Guna, R., Borras, R., 2004. New agents active against Mycobacterium avium complex selected by molecular topology: a virtual screening method. J. Antimicrob. Chemother. 53, 65–73.

González, M.P., Moldes del Carmen Teran, M., 2004. A TOPS-MODE approach to predict adenosine kinase inhibition. Bioorg. Med. Chem. Lett. 14, 3077–3079.

González, M.P., Helguera, A.M., Cabrera, M.A., 2005. Quantitative structure–activity relationship to predict toxicological properties of benzene derivative compounds. Bioorg. Med. Chem. 13, 1775–1781.

González, M.P., Teran, C., Teijeira, M., 2006. A topological function based on spectral moments for predicting affinity toward A3 adenosine receptors. Bioorg. Med. Chem. Lett. 16, 1291–1296.

González-Díaz, H., Uriarte, E., 2005. Biopolymer stochastic moments. I. Modeling human rhinovirus cellular recognition with protein surface electrostatic moments. Biopolymers 77, 296–303.

González-Díaz, H., de Armas, R.R., Molina, R., 2003a. Markovian negentropies in bioinformatics. 1. A picture of footprints after the interaction of the HIV-1 Psi-RNA packaging region with drugs. Bioinformatics 19, 2079–2087.

González-Díaz, H., Molina, R.R., Uriarte, E., 2003b. Stochastic molecular descriptors for polymers. 1. Modelling the properties of icosahedral viruses with 3D-Markovian negentropies. Polymer, 3845–3853.

González-Díaz, H., de Armas, R.R., Molina, R., 2003c. Vibrational Markovian modelling of footprints after the interaction of antibiotics with the packaging region of HIV type 1. Bull. Math. Biol. 65, 991–1002.

González-Díaz, H., Pérez-Bello, A., Uriarte, E., 2005a. Stochastic molecular descriptors for polymers. 3. Markov electrostatic moments as polymer 2D-folding descriptors: RNA–QSAR for mycobacterial promoters. Polymer 46, 6461–6473.

González-Díaz, H., Cruz-Monteagudo, M., Molina, R., Tenorio, E., Uriarte, E., 2005b. Predicting multiple drugs side effects with a general drug-target interaction thermodynamic Markov model. Bioorg. Med. Chem. 13, 1119–1129.

González-Díaz, H., Aguero-Chapin, G., Varona-Santos, J., Molina, R., de la Riva, G., Uriarte, E., 2005c. 2D RNA-QSAR: assigning ACC oxidase family membership with stochastic molecular descriptors; isolation and prediction of a sequence from Psidium guajava L. Bioorg. Med. Chem. Lett. 15, 2932–2937.

González-Díaz, H., Cruz-Monteagudo, M., Vina, D., Santana, L., Uriarte, E., De Clercq, E., 2005d. QSAR for anti-RNA-virus activity, synthesis, and assay of anti-RSV carbonucleosides given a unified representation of spectral moments, quadratic, and topologic indices. Bioorg. Med. Chem. Lett. 15, 1651–1657.

González-Díaz, H., Aguero, G., Cabrera, M.A., Molina, R., Santana, L., Uriarte, E., Delogu, G., Castanedo, N., 2005e. Unified Markov thermodynamics based on stochastic forms to classify drugs considering molecular structure, partition system, and biological species: distribution of the antimicrobial G1 on rat tissues. Bioorg. Med. Chem. Lett. 15, 551–557.

González-Díaz, H., Perez-Bello, A., Uriarte, E., González-Diaz, Y., 2006a. QSAR study for mycobacterial promoters with low sequence homology. Bioorg. Med. Chem. Lett. 16, 547–553.

González-Díaz, H., Vina, D., Santana, L., de Clercq, E., Uriarte, E., 2006b. Stochastic entropy QSAR for the in silico discovery of anticancer compounds: prediction, synthesis, and in vitro assay of new purine carbanucleosides. Bioorg. Med. Chem. 14, 1095–1107.

González-Diaz, H., Sanchez-González, A., González-Diaz, Y., 2006c. 3D-QSAR study for DNA cleavage proteins with a potential anti-tumor ATCUN-like motif. J. Inorg. Biochem. 100, 1290–1297.

González-Díaz, H., Molina-Ruiz, R., and Hernandez, I., 2007a. MARCH-INSIDE version 3.0 (MAR*kov* CH*ains* IN*variants for* SI*mulation* & DE*sign*); Windows supported version under request to the main author contact email: gonzalezdiazh@yahoo.es.

González-Díaz, H., Vilar, S., Santana, L., Uriarte, E., 2007b. Medicinal chemistry and bioinformatics—current trends in drugs discovery with networks topological indices. Curr. Top. Med. Chem. 7, 1025–1039.

González-Díaz, H., Pérez-Bello, A., Cruz-Monteagudo, M., González-Díaz, Y., Santana, L., Uriarte, E., 2007c. Chemometrics for QSAR with low sequence homology: Mycobacterial promoter sequences recognition with 2D-RNA entropies. Chemom. Intell. Lab. Syst. 85, 20–26.

González-Díaz, H., Agüero-Chapin, G., Varona, J., Molina, R., Delogu, G., Santana, L., Uriarte, E., Gianni, P., 2007d. 2D-RNA-coupling numbers: a new computational chemistry approach to link secondary structure topology with biological function. J. Comput. Chem. 28, 1049–1056.

González-Díaz, H., Vilar, S., Santana, L., Uriarte, E., 2007e. Medicinal chemistry and bioinformatics—current trends in drugs discovery with networks topological indices. Curr. Top. Med. Chem. 10, 1015–1029.

González-Díaz, H., Bonet, I., Teran, C., De Clercq, E., Bello, R., Garcia, M.M., Santana, L., Uriarte, E., 2007f. ANN-QSAR model for selection of anticancer leads from structurally heterogeneous series of compounds. Eur. J. Med. Chem. 42, 580–585.

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

Griffiths, A.J.F., Miller, J.H., Suzuki, D.T., Lewontin, R.C., Gelbart, W.M., 1999. Introduction to Genetic Analysis. Freeman, New York.

Han, L., Cui, J., Lin, H., Ji, Z., Cao, Z., Li, Y., Chen, Y., 2006. Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity. Proteomics 6, 4023–4037.

Harary, F., 1969. Graph theory, MA.

Harshey, R.M., Ramakrishnan, T., 1977. Rate of ribonucleic acid chain growth in Mycobacterium tuberculosis H37Rv. J. Bacteriol. 129, 616–622.

Jacchieri, S.G., 2000. Mining combinatorial data in protein sequences and structures. Mol. Divers. 5, 145–152.

Kalate, R.N., Tambe, S.S., Kulkarni, B.D., 2003. Artificial neural networks for prediction of mycobacterial promoter sequences. Comput. Biol. Chem. 27, 555–564.

King, E.L., Altman, C., 1956. A schematic method of deriving the rate laws for enzyme-catalyzed reactions. J. Phys. Chem. 60, 1375–1378.

Kutner, M.H., Nachtsheim, C.J., Neter, J., Li, W., 2005. Standardized Multiple Regression Model, Applied Linear Statistical Models. McGraw-Hill, New York, pp. 271-277.

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

Levine, M., Tjian, R., 2003. Transcription regulation and animal diversity. Nature 424, 147–151.

Liao, B., Ding, K., 2005. Graphical approach to analyzing DNA sequences. J. Comput. Chem. 26, 1519–1523.

Liao, B., Wang, T.M., 2004a. Analysis of similarity/dissimilarity of DNA sequences based on nonoverlapping triplets of nucleotide bases. J. Chem. Inf. Comput. Sci. 44, 1666–1670.

Liao, B., Wang, T.M., 2004b. New 2D graphical representation of DNA sequences. J. Comput. Chem. 25, 1364–1368.

Liao, B., Wang, T.M., 2004c. A 3D graphical representation of RNA secondary structures. J. Biomol. Struct. Dyn. 21, 827–832.

Liao, B., Ding, K., Wang, T.M., 2005. On a six-dimensional representation of RNA secondary structures. J. Biomol. Struct. Dyn. 22, 455–463.

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

Lin, H., 2008. The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. J. Theor. Biol. 252, 350–356.

Liu, Y., Guo, X., Xu, J., Pan, L., Wang, S., 2002. Some notes on 2-D graphical representation of DNA sequence. J. Chem. Inf. Comput. Sci. 42, 529–533.

Markham, N.R., Zuker, M., 2005. DINAMelt web server for nucleic acid melting prediction. Nucleic Acids Res. 33, W577–W581.

Marrero-Ponce, Y., Diaz, H.G., Zaldivar, V.R., Torrens, F., Castro, E.A., 2004a. 3D-chiral quadratic indices of the 'molecular pseudograph's atom adjacency matrix' and their application to central chirality codification: classification of ACE inhibitors and prediction of sigma-receptor antagonist activities. Bioorg. Med. Chem. 12, 5331–5342.

Marrero-Ponce, Y., Castillo-Garit, J.A., Olazabal, E., Serrano, H.S., Morales, A., Castanedo, N., Ibarra-Velarde, F., Huesca-Guillen, A., Jorge, E., del Valle, A., Torrens, F., Castro, E.A., 2004b. TOMOCOMD-CARDD, a novel approach for computer-aided 'rational' drug design: I. Theoretical and experimental assessment of a promising method for computational screening and in silico design of new anthelmintic compounds. J. Comput. Aided Mol. Des. 18, 615–634.

Marrero-Ponce, Y., Medina-Marrero, R., Castillo-Garit, J.A., Romero-Zaldivar, V., Torrens, F., Castro, E.A., 2005a. Protein linear indices of the 'macromolecular pseudograph alpha-carbon atom adjacency matrix' in bioinformatics. Part 1: prediction of protein stability effects of a complete set of alanine substitutions in arc repressor. Bioorg. Med. Chem. 13, 3003–3015.

Marrero-Ponce, Y., Castillo-Garit, J.A., Olazabal, E., Serrano, H.S., Morales, A., Castanedo, N., Ibarra-Velarde, F., Huesca-Guillen, A., Sanchez, A.M., Torrens, F., Castro, E.A., 2005b. Atom, atom-type and total molecular linear indices as a promising approach for bioorganic and medicinal chemistry: theoretical and experimental assessment of a novel method for virtual screening and rational design of new lead anthelmintic. Bioorg. Med. Chem. 13, 1005–1020.

Meneses-Marcel, A., Marrero-Ponce, Y., Machado-Tugores, Y., Montero-Torres, A., Pereira, D.M., Escario, J.A., Nogal-Ruiz, J.J., Ochoa, C., Aran, V.J., Martinez-Fernandez, A.R., Garcia Sanchez, R.N., 2005. A linear discrimination analysis based virtual screening of trichomonacidal lead-like compounds: outcomes of in silico studies supported by experimental results. Bioorg. Med. Chem. Lett. 15, 3838–3843.

Mulder, M.A., Zappe, H., Steyn, L.M., 1997. Mycobacterial promoters. Tuber. Lung Dis. 78, 211–223.

Mulligan, M.E., McClure, W.R., 1986. Analysis of the occurrence of promoter-sites in DNA. Nucleic Acids Res. 14, 109–126.

Mulligan, M.E., Hawley, D.K., Entriken, R., McClure, W.R., 1984. Escherichia coli promoter sequences predict in vitro RNA polymerase selectivity. Nucleic Acids Res. 12, 789–800.

Munteanu, C.R., González-Diaz, H., Magalhaes, A.L., 2008a. Enzymes/non-enzymes classification model complexity based on composition, sequence, 3D and topological indices. J. Theor. Biol. 254, 476–482.

Munteanu, C.R., González-Díaz, H., Borges, F., and Magalhães, A.L., 2008b. Natural/random protein classification models based on star network topological indices. J. Theor. Biol. 〈http://dx.doi.org/10.1016/j.jtbi.2008.07.018〉.

Myers, D., Palmer, G., 1985. Microcomputer tools for steady-state enzyme kinetics. Bioinformatics (original: Comput. Appl. Biosci.) 1, 105–110.

Nandy, A., 1994. Recent investigations into global characteristics of long DNA sequences. Indian J. Biochem. Biophys. 31, 149–155.

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

Nandy, A., Basak, S.C., 2000. Simple numerical descriptor for quantifying effect of toxic substances on DNA sequences. J. Chem. Inf. Comput. Sci. 40, 915–919.

O'Neill, M.C., Chiafari, F., 1989. Escherichia coli promoters. II. A spacing class-dependent promoter search protocol. J. Biol. Chem. 264, 5531–5534.

Pawar, A.P., Dubay, K.F., Zurdo, J., Chiti, F., Vendruscolo, M., Dobson, C.M., 2005. Prediction of "aggregation-prone" and "aggregation-susceptible" regions in proteins associated with neurodegenerative diseases. J. Mol. Biol. 350, 379–392.

Prado-Prado, F.J., González-Diaz, H., de la Vega, O.M., Ubeira, F.M., 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. Bioorg. Med. Chem. 16, 5871–5880.

Qi, X.Q., Wen, J., Qi, Z.H., 2007. New 3D graphical representation of DNA sequence based on dual nucleotides. J. Theor. Biol. 249, 681–690.

Ramos de Armas, R., González-Díaz, H., Molina, R., Perez Gonzalez, M., Uriarte, E., 2004. Stochastic-based descriptors studying peptides biological properties: modeling the bitter tasting threshold of dipeptides. Bioorg. Med. Chem. 12, 4815–4822.

Randic, M., 2004. 2-D graphical representation of proteins based on virtual genetic code. SAR QSAR Environ. Res. 15, 147–157.

Randic, M., Balaban, A.T., 2003. On a four-dimensional representation of DNA primary sequences. J. Chem. Inf. Comput. Sci. 43, 532–539.

Randic, M., Vracko, M., 2000. On the similarity of DNA primary sequences. J. Chem. Inf. Comput. Sci. 40, 599–606.

Randic, M., Zupan, J., 2004. Highly compact 2D graphical representation of DNA sequences. SAR QSAR Environ. Res. 15, 191–205.

Randic, M., Vracko, M., Nandy, A., Basak, S.C., 2000. On 3-D graphical representation of DNA primary sequences and their numerical characterization. J. Chem. Inf. Comput. Sci. 40, 1235–1244.

Randic, M., Zupan, J., Vikic-Topic, D., 2007. On representation of proteins by star-like graphs. J. Mol. Graph Model. 290-305.

Santana, L., Uriarte, E., González-Diaz, H., Zagotto, G., Soto-Otero, R., Mendez-Alvarez, E., 2006. A QSAR model for in silico screening of MAO-A inhibitors. Prediction, synthesis, and biological assay of novel coumarins. J. Med. Chem. 49, 1149–1156.

Song, J., Tang, H., 2005. A new 2-D graphical representation of DNA sequences and their numerical characterization. J. Biochem. Biophys. Meth. 63, 228–239.

StatSoft.Inc., STATISTICA (data analysis software system), version 6.0, 〈www.statsoft.com〉 Statsoft, 2002.

Todeschini, R., Consonni, V., 2002. Handbook of Molecular Descriptors. Wiley-VCH, New York.

Van Waterbeemd, H., 1995. Chemometric Methods in Molecular Design. Wiley-VCH, New York.

Vilar, S., Estrada, E., Uriarte, E., Santana, L., Gutierrez, Y., 2005. In silico studies toward the discovery of new anti-HIV nucleoside compounds through the use of TOPS-MODE and 2D/3D connectivity indices. 2. Purine derivatives. J. Chem. Inf. Model. 45, 502–514.

Vilar, S., Santana, L., Uriarte, E., 2006. Probabilistic neural network model for the in silico evaluation of anti-HIV activity and mechanism of action. J. Med. Chem. 49, 1118–1124.

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., 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.

Woodcock, S., Mornon, J.P., Henrissat, B., 1992. Detection of secondary structure elements in proteins by hydrophobic cluster analysis. Protein Eng. 5, 629–635.

Wyrick, J.J., Young, R.A., 2002. Deciphering gene expression regulatory networks. Curr. Opin. Genet. Dev. 12, 130–136.

Xiao, X., Chou, K.C., 2007. Digital coding of amino acids based on hydrophobic index. Protein Pept. Lett. 14, 871–875.

Xiao, X., Shao, S., Ding, Y., Huang, Z., Chen, X., 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., 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, 555–565.

Xiao, X., Shao, S.H., Chou, K.C., 2006a. A probability cellular automaton model for hepatitis B viral infections. Biochem. Biophys. Res. Comm. 342, 605–610.

Xiao, X., Shao, S.H., Ding, Y.S., Huang, Z.D., Chou, K.C., 2006b. Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. Amino Acids 30, 49–54.

Yuan, Z., 1999. Prediction of protein subcellular locations using Markov chain models. FEBS Lett. 451, 23–26.

Zhang, C.T., Chou, K.C., 1993. Graphic analysis of codon usage strategy in 1490 human proteins. J. Protein Chem. 12, 329–335.

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

Zhou, G.P., 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.

Zupan, J., Randic, M., 2005. Algorithm for coding DNA sequences into "spectrum-like" and "zigzag" representations. J. Chem. Inf. Model. 45, 309–313.