Amino Acids (2011) 40:431–442
DOI 10.1007/s00726-010-0653-9

ORIGINAL ARTICLE

TI2BioP: Topological Indices to BioPolymers. Its practical use to unravel cryptic bacteriocin-like domains

Guillermín Agüero-Chapin · Gisselle Pérez-Machado · Reinaldo Molina-Ruiz · Yunierkis Pérez-Castillo · Aliuska Morales-Helguera · Vítor Vasconcelos · Agostinho Antunes

Received: 10 March 2010 / Accepted: 2 June 2010 / Published online: 19 June 2010

Abstract Bacteriocins are proteinaceous toxins produced and exported by both gram-negative and gram-positive bacteria as a defense mechanism. The bacteriocin protein family is highly diverse, which complicates the identification of bacteriocin-like sequences using alignment approaches. The use of topological indices (TIs) irrespective of sequence similarity can be a promising alternative to predict proteinaceous bacteriocins. Thus, we present Topological Indices to BioPolymers (TI2BioP) as an alignment-free approach inspired in both the Topological Substructural Molecular Design (TOPS-MODE) and Markov Chain Invariants for Network Selection and Design (MARCH-INSIDE) methodology. TI2BioP allows the calculation of the spectral moments as simple TIs to seek quantitative sequence-function relationships (QSFR) models. Since hydrophobicity and basicity are major criteria for the bactericide activity of bacteriocins, the spectral moments (${}^{HP}\mu_k$) were derived for the first time from protein artificial secondary structures based on amino acid clustering into a Cartesian system of hydrophobicity and polarity. Several orders of ${}^{HP}\mu_k$ characterized numerically 196 bacteriocin-like sequences and a control group made up of 200 representative CATH domains. Subsequently, they were used to develop an alignment-free QSFR model allowing a 76.92% discrimination of bacteriocin proteins from other domains, a relevant result considering the high sequence diversity among the members of both groups. The model showed a prediction overall performance of 72.16%, detecting specifically 66.7% of proteinaceous bacteriocins whereas the InterProScan retrieved just 60.2%. As a practical validation, the model also predicted successfully the cryptic bactericide function of the Cry 1Ab C-terminal domain from *Bacillus thuringiensis*'s endotoxin, which has not been detected by classical alignment methods.

Keywords Bacteriocin · Topological indices · Spectral moments · Alignment methods · Artificial secondary structure

Introduction

Bacteriocins are proteinaceous toxins produced and exported by both gram-negative and gram-positive bacteria

Electronic supplementary material The online version of this article (doi:10.1007/s00726-010-0653-9) contains supplementary material, which is available to authorized users.

G. Agüero-Chapin · V. Vasconcelos · A. Antunes (✉)
CIMAR/CIIMAR, Centro Interdisciplinar de Investigação
Marinha e Ambiental, Universidade do Porto,
Rua dos Bragas, 177, 4050-123 Porto, Portugal
e-mail: aantunes@ciimar.up.pt

G. Agüero-Chapin · G. Pérez-Machado · R. Molina-Ruiz ·
Y. Pérez-Castillo · A. Morales-Helguera
Molecular Simulation and Drug Design (CBQ),
Central University of Las Villas,
54830 Santa Clara, Cuba

Y. Pérez-Castillo
Department of Organic Chemistry, Vigo University,
36200 Vigo, Spain

A. Morales-Helguera
Department of Chemistry, Central University of Las Villas,
Santa Clara 54830, Villa Clara, Cuba

A. Morales-Helguera
REQUIMTE, Department of Chemistry, University of Porto,
4169-007 Porto, Portugal

G. Agüero-Chapin · V. Vasconcelos
Departamento de Biologia, Faculdade de Ciências,
Universidade do Porto, Porto, Portugal

to inhibit the growth of similar or more distant bacteria species (de Jong et al. 2006; Hammami et al. 2007). Bacteriocins can be applied as food preservatives (Cotter et al. 2005) and are of great interest for novel antibiotics development (Gillor et al. 2005) and as a diagnostic agents for some cancers (Cruz-Chamorro et al. 2006; Sand et al. 2007). The classical way to identify a bacteriocin includes the determination of its biological activity, which is accomplished by the extensive testing of the (putative) producer strain ability to inhibit the growth of other bacteria.

The bacteriocin family includes a diversity of proteins in terms of size, method of killing, method of production, genetics, microbial target, immunity mechanisms, and release. Given such high diversity, bacteriocin classification has been challenging (Cotter et al. 2006). The few bioinformatics approaches developed to identify bacteriocins recognize putative open-reading frames (ORFs) based on sequence alignment (Dirix et al. 2004; Stein 2005) demanding the implementation of complex strategies due to the low conservation of the bacteriocin protein class. The use of topological indices (TIs), irrespective of sequence similarity, can be a promising alternative to predict proteinaceous bacteriocins (Estrada and Uriarte 2001; Gonzalez-Diaz et al. 2008; Gonzalez-Diaz et al. 2007c). Thus, we present Topological Indices to BioPolymers (TI2BioP) as an alignment-free approach inspired in both the Topological Substructural Molecular Design (TOPS-MODE) (Estrada 2000) and Markov Chain Invariants for Network Selection and Design (MARCH-INSIDE) methodology (González-Díaz et al. 2007) that calculates the spectral moments as simple TIs to obtain alignment-free models from quantitative sequence-function relationships (QSFR). This methodology takes advantage of the calculation of one-dimension (1D), two-dimension (2D), and three-dimension (3D) parameters based on the graphical representation of the chemical structure of biopolymers such as DNA, RNA, and proteins. We evaluated the TI2BioP accuracy to successfully identify proteinaceous bacteriocins in spite of its high sequence diversity. Since hydrophobicity and basicity are major criteria for the bactericide activity of bacteriocins (Fimland et al. 2002; Hammami et al. 2007), we derived the TIs from linear sequences plotting its amino acids (aas) into an 2D Cartesian Hydrophobicity-Polarity (2D-HP) lattice resembling a protein pseudo-secondary structure (see Figs. 1, 7). Thus, we calculated for the first time the spectral moments (${}^{HP}\mu_k$) of the edge matrix associated with such artificial secondary structures as TIs. The new spectral moments are based on the 2D spectral moments calculated by TOPS-MODE as well as on the 3D and HP-Lattice stochastic spectral moments calculated by MARCH-INSIDE (Agüero-Chapin et al. 2009; Gonzalez-Diaz et al. 2007a; Gonzalez-Diaz et al. 2007b; Munteanu et al. 2009), but have a different definition and contain new structural information. Its values characterized numerically 196 bacteriocin-like sequences and a control group made up of 200 representative CATH domains. Subsequently, several orders of ${}^{HP}\mu_k$ were used to develop an alignment-free QSFR model that allowed a 76.92% discrimination of bacteriocin proteins from other domains, a good result considering the high sequence diversity among the members of both groups. The model showed a prediction overall performance of 72.16%, specifically retrieving 66.7% of proteinaceous bacteriocins whereas the InterProScan classified just 60.2%. Our model further predicted successfully the cryptic bactericide function of the Cry 1Ab C-terminal domain from *Bacillus thuringiensis*'s endotoxin reported by Vazquez-Padrón et al. (Vazquez-Padron et al. 2004).

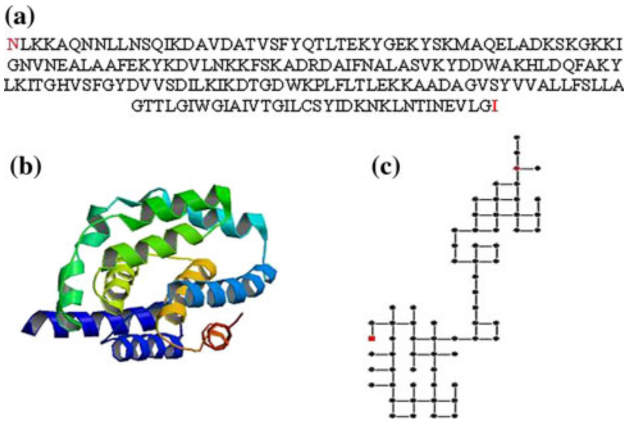

(a)
NLKKAQNNLLNSQIKDAVDATVSFYQTLTEKYGEKYSKMAQELADKSKGKKI
GNVNEALAAFEKYKDVLNKKFSKADRDAIFNALASVKYDDWAKHLDQFAKY
LKITGHVSFGYDVVSDILKIKDTGDWKPLFLTLEKKAADAGVSYVVALLFSLLA
GTTLGIWGIAIVTGILCSYIDKNKLNTINEVLGI
(b)
(c)

Fig. 1 Three structures for the colicin E1domain sequence. **a** Primary structure **b** three-dimensional structure **c** the pseudo-secondary Cartesian structure of hydropobicity and polarity

We conclude that the TI2BioP approach based on the higher-order encoding of the HP-spectral moments has a high accuracy that justifies its use as an alternative method to alignment approaches. TI2BioP retrieved successfully the screening of putative proteinaceous bacteriocins in spite of the high sequence diversity of this protein class. Furthermore, TI2BioP allowed the prediction of protein domains that have a cryptic bactericidal action, undetectable using alignment procedures. Finally, the alignment of 2D-HP protein maps offered a novel approach to explain evolutionary relationships between the Cry 1Ab C-terminal domain and the bacteriocin class.

Methods

Computational methods

An alignment-free methodology called "TI2BioP" is presented to codify the structural information of proteinaceous bacteriocins and a control group designed from 8,871 structurally non-redundant subset of the CATH database (Cuff et al. 2009). TI2BioP was built up on object-oriented Free Pascal IDE Tools (lazarus). The program can be run on Windows and Linux operating system. The user-friendly interface allows the users to access the sequence list introduction, selecting the representation type and calculations of TIs. It is based on the graph theory considering the "building blocks" of the biopolymers DNA, RNA, and protein as nodes or vertexes and the bonds between them as edges into a certain graph. Thus, the information contained in biopolymeric long strings is simplified in a graph considering some of its relevant features as the topology and properties of the monomers. These factors determine either the real secondary structure or the pseudo-folding of linear sequences into 2D-HP lattice. TI2BioP was developed on the basis of two well-known methodologies: "TOPS-MODE" (Estrada 2000) implemented in the "MODESLAB" software (Gutierrez and Estrada 2002) and the MARCH-INSIDE program (González-Díaz et al. 2007). TI2BioP shows a draw mode to represent automatically linear sequences of DNA, RNA and proteins as 2D graphs, but can also import files containing 2D structure inferred by other professional programs (Mathews 2006). The calculation of the topological indices from these 2D maps is performed following the TOPS-MODE approach (Estrada 1996; Estrada 2000). Finally, these TIs containing relevant information of the sequence are used to carry out a QSFR, which allow classifying gene and protein classes without the need to perform an alignment procedure.

We used the 2D-HP graphs to encode information about proteinaceous bacteriocin sequences following previous experiences achieved using the MARCH-INSIDE methodology (González-Díaz et al. 2007) in the prediction of protein function from linear sequences (Agüero-Chapin et al. 2008b; Agüero-Chapin et al. 2009; Gonzalez-Diaz et al. 2008).

The spectral moments (μ_{k}) introduced previously by Estrada (Estrada 1996; Estrada 1997) were applied to describe protein 2D-maps. These TIs have been widely validated by many authors to encode the structure of small molecules in QSAR studies (González et al. 2006; Markovic et al. 2001) including the characterization of macro molecular chains based on dihedral angles by Estrada (Estrada 2007; Estrada and Hatano 2007). The original adjacent matrix is modified according the building of the 2HP-protein maps. The 20 different aas are clustered into 4 HP classes. These four groups characterize the HP physicochemical nature of the aas as polar, non-polar, acidic or basic (Jacchieri 2000). Each amino acid (aa) 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 aas are calculated as follows:

a. Decrease by -1 the abscissa axis coordinate for an acid aa (leftwards-step) or:
b. Increase by $+1$ the abscissa axis coordinate for a basic aa (rightwards-step) or:
c. Increase by $+1$ the ordinate axis coordinate for a non-polar aa (upwards-step) or:
d. Decrease by -1 the ordinate axis coordinate for a polar aa (downwards-step).

This 2D graphical representation for proteins is similar to those previously reported for DNA (Nandy 1994; Nandy 1996; Randic and Vracko 2000) and has been also useful for structural RNA classification (Agüero-Chapin et al. 2008a). The Fig. 1 shows the primary structure of the channel-forming domain of colicin E1 bacteriocin (a), the crystal structure of such domain (b) and its 2D-HP map (c). The 191 aas of the colicin E1 domain sequence are rearranged in a pseudo-secondary structure of hydrophobicity and polarity that compact its linear sequence. Note that a node (n) in the 2D-HP map could be made up for more than one aa. The N and C termini of the protein sequence in the 2D-HP map are labeled with a red square dot and simple dot, respectively.

We calculated for the first time the spectral moments (${}^{\mathrm{HP}}\mu_k$) values as TIs describing these proteins maps. The ${}^{\mathrm{HP}}\mu_k$ were selected based on the utility of μ_{k} to codify structural information in small molecules (Cabrera-Pérez et al. 2004; Estrada 2000) and also do to its relevance in Proteomics, when stochastically calculated (${}^{\mathrm{HP}}\pi_k$) using the Markov chain theory (Gonzalez-Diaz and Uriarte 2005; Gonzalez-Diaz et al. 2005).

Spectral moments for 2D-HP protein maps

After the representation of the sequences we assigned to each graph a bond matrix **B** for the computation of the spectral moments. These TIs are defined as the trace, i.e. sum of main diagonal entries of the different powers of the bond adjacency matrix. This matrix is a square symmetric matrix that its non-diagonal entries are ones or zeroes if the corresponding bonds share or not one aa. Thus, it set up connectivity relationships between the aa in the pseudo secondary structure (2D-HP map). The number of edges (e) in the graph is equal to the number of rows and columns in **B** but may be equal or even smaller than the number of peptide bonds in the sequence. Main diagonal entries can have bonds weights describing hydrophobic/polarity, electronic and steric features of the aas. Particularly, the main diagonal was weighted with the average of the electrostatic charge (Q) between two bound nodes that in turn are weighted with electrostatic charge (q) from

Amber 95 force field (Cornell et al. 1995). The q is equal to the sum of the charges of all aas placed in a node. Thus, it is easy to carry out the calculation of the spectral moments of **B** in order to numerically characterize the protein sequence.

$$^{HP}\mu_k = \mathrm{Tr}\left[(\mathrm{B})^k\right] \tag{1}$$

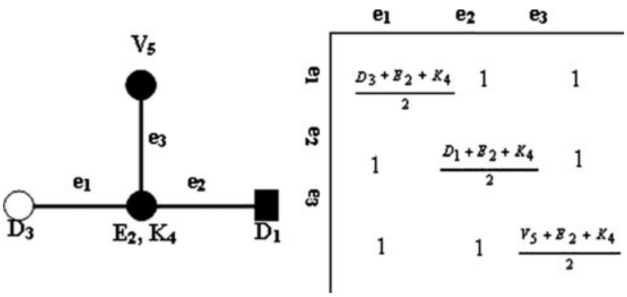
where Tr is called the trace and indicates the sum of all the values in the main diagonal of the matrices $(\mathbf{B})^k$, which are the natural powers of **B**.

In order to illustrate the calculation of the spectral moments, an example is described below. The 2D-HP map of the sequence (D_1-E_2-D_3-K_4-V_5) is showed in the Fig. 2 as well as its bond adjacency matrix. The calculation of the spectral moments up to the order $k = 3$ is also defined downstream of the Fig. 2. Please note in the graph that the central node contains both E, and K and the q values are represented in the matrix as the aa symbols ($E = 1.885$, $V = 2.24$, $K = 2.254$, $D = 1.997$).

Expansion of expression (1) for $k = 1$ gives the $^{HP}\mu_1$, for $k = 2$ the $^{HP}\mu_2$ and for $k = 3$ the $^{HP}\mu_{3.}$ The bond adjacency matrix derived from this linear graph is described for each case

$$^{HP}\mu_1 = \mathrm{Tr}[\mathrm{B}] = \mathrm{Tr}\left(\begin{bmatrix} 3.068 & 1 & 1 \\ 1 & 3.068 & 1 \\ 1 & 1 & 3.189 \end{bmatrix}\right) = 9.325 \tag{1a}$$

$$^{HP}\mu_2 = \mathrm{Tr}\left[(\mathrm{B})^2\right] = \mathrm{Tr}\left(\begin{bmatrix} 3.068 & 1 & 1 \\ 1 & 3.068 & 1 \\ 1 & 1 & 3.189 \end{bmatrix} \times \begin{bmatrix} 3.068 & 1 & 1 \\ 1 & 3.068 & 1 \\ 1 & 1 & 3.189 \end{bmatrix}\right) = (11.413)^2 + (11.413)^2 + (12.170)^2 \tag{1b}$$



$$\begin{array}{c|ccc} & e_1 & e_2 & e_3 \\ \hline e_1 & \frac{D_3 + E_2 + K_4}{2} & 1 & 1 \\ e_2 & 1 & \frac{D_1 + E_2 + K_4}{2} & 1 \\ e_3 & 1 & 1 & \frac{V_5 + E_2 + K_4}{2} \end{array}$$

Fig. 2 The 2D-HP map for the protein fragment DEDKV, aside the definition of its bond adjacency matrix. Note that all edges of the graph are adjacent, thus all non-diagonal entries are ones

$$^{HP}\mu_3 = \mathrm{Tr}\left[(\mathrm{B})^3\right] = \mathrm{Tr}\left(\begin{bmatrix} 3.068 & 1 & 1 \\ 1 & 3.068 & 1 \\ 1 & 1 & 3.189 \end{bmatrix}\right)^3 = (49.405)^3 + (49.405)^3 + (53.323)^3 \tag{1c}$$

The calculation of $^{HP}\mu_k$ values for protein sequences of both groups were carried out with our in-house software TI2BioP version 1.0®, including sequence representation (Molina et al. 2009). We proceeded to upload a row data table containing the sixteen $^{HP}\mu_k$ values for each sequence ($k = 1, 2, 3, \ldots 16$), two additional TIs defined as Edge Numbers and Edge Connectivity and the grouping variable (Bact-score) that indicates the bacteriocin-like proteins with value of 1 and −1 for control group sequences to statistical analysis software (Statsoft 2007). The overall methodology is represented schematically in order to improve the understanding of our approach (see Fig. 3).

Database

A total of 196 bacteriocin-like proteins sequences belonging to several bacterial species were collected from the two major bacteriocin databases, BAGEL (de Jong et al. 2006) and BACTIBASE (Hammami et al. 2007). A polypeptide or proteinaceous bacteriocin was considered according its sequence length (>100 bp). Each proteinaceous bacteriocin sequence retrieved was labeled respecting its original database ID code; see Table I in SM.

The negative group was selected from 8,871 protein downloaded from the CATH domain database of protein

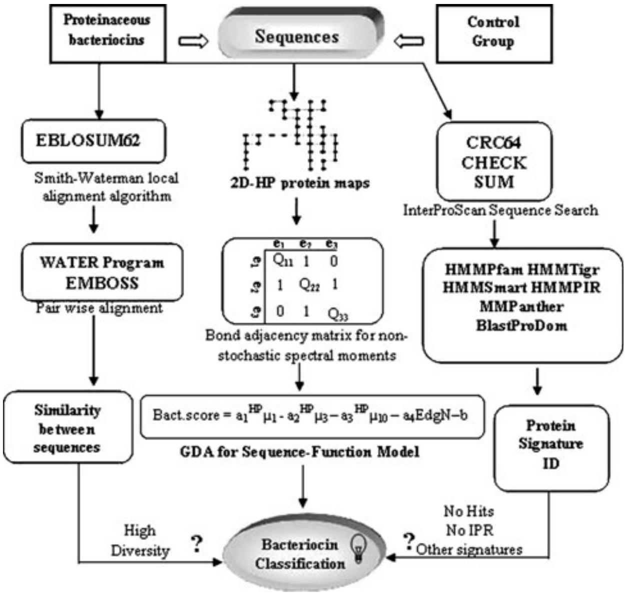

Proteinaceous bacteriocins
Sequences
Control Group
EBLOSUM62
Smith-Waterman local alignment algorithm
WATER Program EMBOSS
Pair wise alignment
Similarity between sequences
2D-HP protein maps
e1 e2 e3
e1 Q11 1 0
e2 1 Q22 1
e3 0 1 Q33
Bond adjacency matrix for non-stochastic spectral moments
Bact.score = a1 HPμ1 - a2 HPμ3 – a3 HPμ10 – a4EdgN–b
GDA for Sequence-Function Model
CRC64 CHECK SUM
InterProScan Sequence Search
HMMPfam HMMTigr HMMSmart HMMPIR MMPanther BlastProDom
Protein Signature ID
High Diversity
?
Bacteriocin Classification
?
No Hits
No IPR
Other signatures

Fig. 3 The overall procedure followed for the classification of bacteriocins

structural families (version 3.2.0) (http://www.cathdb.info) (Cuff et al. 2009). Particularly, we used the FASTA sequence database for all CATH domains (based on COMBS sequence data) sharing just the 35% of sequence similarity as the starting group. The COMBS sequences provide the full sequence instead of only the residues present in the ATOM records (Brandt et al. 2008). The FASTA database is made non-redundant case-sensitively and IDs are concatenated. The 200 members of the final control subset were selected using a k-means cluster analysis (k-MCA) (Mc Farland and Gans 1995a). CATH domains IDs that make up the control group are also showed in Table Ia of SM. Training and predicting series of the bacteriocin database were designed following the same procedure.

Statistical analysis: k-means cluster analysis (k-MCA)

This method has been applied before in QSAR to design the training and predicting series (Kowalski and Marcoin 2001; Mc Farland and Gans 1995b). The method requires a partition of the bacteriocin and the starting control group independently into several statistically representative clusters of sequences. The members to conform the control group are selected from all of these clusters and afterwards the sequences of the training and predicting series. This procedure ensures that the main protein classes will be considered in the control group allowing the representation of the entire 'experimental universe'. The spectral moment series were explored as clustering variables in order to carry out k-MCA. The procedure described above is represented graphically in Fig. 4 for both groups.

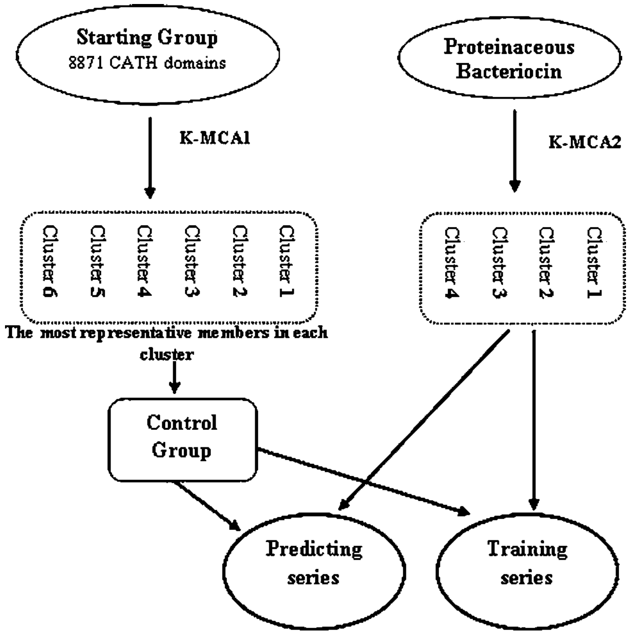


Fig. 4 Scheme describing the design of training and predicting series using k-MCA for both bacteriocins and control group

General discriminant analysis (GDA)

The starting control group was reduced following the k-MCA to balance both groups according to the GDA requirements; then training and predicting series were selected from 200 CATH members. The GDA best subset was carried out for variable selection to build up the model (Marrero-Ponce et al. 2005; Marrero-Ponce et al. 2004; Meneses-Marcel et al. 2005; Ponce et al. 2004). The STATISTICA software reviewed all the variable predictors for finding the "best" possible sub model. The variables were standardized in order to bring them onto the same scale. Subsequently, a standardized linear discriminant equation that allows comparison of their coefficients was obtained (Kutner et al. 2005). The model selection was based on the revision of Wilk's (λ) statistic ($\lambda = 0$ perfect discrimination, being $0 < \lambda < 1$) in order to assess the discriminatory power of the model. We also inspected the Fisher ratio (F), value of a variable indicating its statistical significance in the discrimination between groups, which is a measure of the extent of how a variable makes an unique contribution to the prediction of group membership with a probability of error (p level) $p(F) < 0.05$.

Applicability domain

A simple method to investigate the applicability domain of a prediction model is to carry out a leverage plot (plotting residuals vs. leverage of proteins used in the training set) (Eriksson et al. 2003; Niculescu et al. 2004). The leverage (h) of a sequence in the original variable space which measures its influence on the model is defined as

$$h_i = x_i^T (X^T X)^{-1} x (i = 1, \ldots, n)$$

where x_i is the descriptor vector of the considered sequence and $\boldsymbol{X}$ is the model matrix derived from the training set descriptor values. The warning leverage $h*$ is defined as follows:

$$h^* = 3 \times p'/n$$

where n is the number of training sequences and p' is the number of model adjustable parameters.

Alignment procedures

The Smith–Waterman algorithm was used to perform local sequence alignment for determining similar regions between pairs of bacteriocin protein sequences (all vs. all) (Smith and Waterman 1981). The water program was downloaded from the European Molecular Biology Open

Software Suite (EMBOSS) (http://www.ebi.ac.uk/Tools/emboss) and run on Linux Ubuntu 8.04. Water uses the Smith–Waterman algorithm (modified for speed enhancements) to calculate the local alignment. EBLOSUM62 was set as the substitution matrix and gap penalties values were taken by default.

Bacteriocin classification using classical methods

Each bacteriocin protein sequence presented in this study was also submitted to InterProScan for its classification (Quevillon et al. 2005). Sequences in FASTA format were analyzed one by one at the http://www.ebi.ac.uk/Tools/InterProScan looking into the InterPro database (Hunter et al. 2009).

Results and discussion

Prediction of proteinaceous Bacteriocins using 2D-HP TIs

We calculated spectral moments (${}^{HP}\mu_k$) of the bond adjacency matrix that describe electronically the connection between the aas in the pseudo secondary structure or 2D-HP map of the protein sequence. This calculation was carried out for two groups of protein sequences, one made up of bacteriocin-like proteins and the other formed by heterogeneous CATH domains. The members of both groups were selected as follows: (1) the bacteriocin group contained 196 members in total; (2) the members of the training and predicting series were chosen according to the k-means cluster analysis (k-MCA); (3) the k-MCA divided the data into four clusters containing 75, 78, 27, and 16 members, respectively; (4) the selection was based on the distance from each member with respect to the cluster center (Euclidean distance); (5) the members of the external validation subset were selected uniformly in respect to Euclidean distance taking out the 25% in each cluster; and (6) the remaining cases were used to train the model.

To set up the final control group, the original data of 8,871 proteins were reduced to 200 members in order to balance the two groups as required by general discriminant analysis (GDA). Data selection was also carried out using the k-MCA to ensure the inclusion of representative protein domains of each cluster in the control group. The original data were split into six statistically representative clusters of sequences made up by: 1,553, 1,416, 1,754, 1,863, 1,339, and 946 members. Afterwards, the members comprising the training and predicting subsets were selected following the same procedure described above.

Cluster of cases were carried out using the TIs computed in TI2BioP methodology. We have explored the standard deviation between and within clusters, the respective Fisher ratio and their p level of significance (Mc Farland and Gans 1995b). All variables were used to construct the clusters but only the combination from the ${}^{HP}\mu_{12}$ to ${}^{HP}\mu_{16}$ showed p levels <0.05 for Fisher test, as depicted in Table 1. Four statistically homogeneous clusters of proteinaceous bacteriocins were described coinciding with the existence of four proteinaceous subclasses described by Cotter et al. (Cotter et al. 2006).

Table 1 Main results of the k-MCA for the proteinaceous bacteriocins class and the control group

Protein descriptors	Between SS[a]	Within SS[b]	Fisher ratio (F)	p level[c]
Variance analysis bacteriocins-like proteins				
${}^{HP}\mu_{12}$	161.61	33.39	309.72	<0.001
${}^{HP}\mu_{13}$	160.19	34.81	294.48	<0.001
${}^{HP}\mu_{14}$	161.44	33.56	307.91	<0.001
${}^{HP}\mu_{15}$	159.49	35.51	287.43	<0.001
${}^{HP}\mu_{16}$	162.20	32.80	316.52	<0.001
Control group				
${}^{HP}\mu_{12}$	8347.91	522.09	28349.40	<0.001
${}^{HP}\mu_{13}$	8319.39	550.61	26788.96	<0.001
${}^{HP}\mu_{14}$	8334.12	535.88	27574.19	<0.001
${}^{HP}\mu_{15}$	8313.42	556.58	26482.71	<0.001
${}^{HP}\mu_{16}$	8336.86	533.14	27725.03	<0.001

[a] Variability between groups

[b] Variability within groups

[c] Level of significance

The k-MCA based on TI2BioP structural indices revealed a high diversity between bacteriocins-like proteins sequences, which was further supported by the pair-wise alignment results performed between its 196 proteinaceous members using the Smith–Waterman local algorithm. The Smith–Waterman procedure is able to obtain correct alignments in regions of low similarity between distantly related biological sequences. Thus, it is possible to detect sub regions or sub-sequences with an evolutionary conserved signal of similarity. Bacteriocins are good candidates to perform this procedure, because aa similarity percentages can be as low as 25.7%. The 85% of the sequences pairs aligned (16,240 pairs) showed similarity percentage below 50% and the 23% sequences pairs (4,375 pairs) showed similarity below 35% in just short sub regions. These outcomes are consistent with the high diversity of bacteriocins and with the distinct performance of the classification methods (see the Smith–Waterman results in Table II of SM).

Once we performed a representative selection of the training set for both groups, the discrimination functions can be determined. Thus, we choose the functions with higher statistical significance but with few parameters as possible. Each discriminant function expresses in probability terms the tendency or propensity of a given aa sequence to belong to the bacteriocin-like protein class. The model classifies the sequences according to its biological function providing a predicted probability as a numerical score ($0 \leq \text{score} \leq 1$). The best classification function equation found for the bacteriocin group after GDA analyses was:

$$\begin{aligned} \text{Bact-score } &= 6.86 \times {}^{\text{HP}}\mu_1 - 2.06 \times {}^{\text{HP}}\mu_3 - 2.39 \\ &\times {}^{\text{HP}}\mu_{10} - 2.34 \times \text{EdgeN} - 0.08 \\ &\times N = 299 \quad \lambda = 0.63 \quad F = 53.22 \quad p < 0.001 \qquad (3) \end{aligned}$$

Where, N is the number of proteins used to seek the corresponding classification models, which discriminate between proteinaceous bacteriocins and representative CATH domains. The statistics parameters of the above equation are the same usually shown for QSAR linear discriminant models (Santana et al. 2006; Vilar et al. 2005), including Wilk's statistical (λ) and Fisher ratio (F) with a probability of error (p level) $p(F)$. The value of $p(F)$ shows significance, rejecting the null hypothesis (H_0) (no difference between two groups).

This discriminant function (equation 3) classified correctly 230 out of 299 proteins used in the training series (level of accuracy of 76.92%). More specifically, the model correctly classified 122/148 (82.43%) sequences of proteinaceous bacteriocins and 108/151 (71.52%) of the control group. A validation procedure was subsequently performed in order to assess the model predictability.

We used the subsampling test to examine the prediction accuracy of our model. This validation procedure is easier to implement and provides reliable results in the validation of a predictive model at low computational cost (Rivals and Personnaz 1999). Thus, we took out randomly subsamples representing the 25% of the training set to assess the model predictability. The procedure was repeated ten times varying the composition of the subsamples. Afterwards the mean values for the Wilk's statistical, accuracy, sensitivity and specificity in training and external validation subsets were calculated. The respective classification matrices for training and cross-validation are depicted in Table 2. The classification results derived from the sub-sample test were very similar to those achieved from the member's selection using k-MCA; notice that the Wilk's statistical remained almost invariant showing the robustness of the model.

An external validation was also performed using the predicting series derived from the k-MCA. It is important to highlight that this external set was not used to build the model. This procedure was carried out with an external series of 48 bacteriocin-like proteins and 49 CATH domains (see Table 2). The model showed a prediction overall performance of 72.16%, able to predict 32/48 (66.7%) of proteinaceous bacteriocins and 38/49 (77.5%) of the functionally diverse domains. This result is remarkable relatively to other QSAR studies using 2D stochastic indices to classify protein classes with higher degree of sequence conservation among its members (Agüero-Chapin et al. 2009; Vilar et al. 2008). The classification of each protein sequence (bacteriocins and CATH domains) is shown in more details in Table I and Ia of SM.

Table 2 Classification results derived from the model for training and validation series

Training set (k-MCA)				External validation (k-MCA)			
Total%	76.92	Bact.	Control	Bact.	Control	72.17	Total%
Bact.	82.43	**122**	26	**32**	16	66.67	Bact.
Control	71.52	43	**108**	11	**38**	77.5	Control

Cross-validation (Training set)

Training subset				Validation subset		
Bact.	Control	Total%	λ	Bact.	Control	Total%
82.36	71.76	77.02	0.629	81.55	70.54	75.94

Numbers in bold highlight the well-classified cases

As can be seen from the model equation, the spectral moment ${}^{\text{HP}}\mu_1$ is the major predictor that contributes positively to the bacteriocin classification. However, the rest of the predictors (${}^{\text{HP}}\mu_3$, ${}^{\text{HP}}\mu_{10}$ and EdgeN) affect bacteriocin identification in a negative way. This fact points out that the increase of higher-order spectral moments and edge numbers affects negatively the bacteriocin identification. Proteins sequences pseudo folded into 2D-HP maps with few edges numbers and high values of ${}^{\text{HP}}\mu_1$ are more likely to present the antibiotic action on other bacteria. Edge numbers are associated directly with the length of the linear sequence but in our pseudo secondary structure are also influenced by the composition of its acid, basic, polar and non-polar aas. Thus, bacteriocins proteinaceous sequences pseudo folded in a more compact 2D-HP map show a balance of hydrophobicity due to its amino acidic composition. This fact agrees with the amphiphatic properties of mature bacteriocins, which form domains or helices having hydrophobic and hydrophilic regions; an essential structural feature to perform its antibiotic action (Kaur et al. 2004). It also supports the fact that naturally-occurring antimicrobial agents are often peptide-like bacteriocins rather than proteins (Sang and Blecha 2008).

The protein classification based solely on linear sequence homology can perform poorly when the sequence diversity is high, as in the case of bacteriocins. By contrast, the classification based on higher structural organization is much more effective because during the evolution of protein families often its secondary and tertiary structure remained more conserved than the primary sequence. Our TIs reveal hidden but relevant information contained in the primary sequence, as the hydrophobicity/polarity features of its aas, which are important properties for the secondary structure fold of bacteriocins (Hammami et al. 2007). Consequently, the 2D-HP TIs are useful to determine with more accuracy the biological function when higher structural levels are not available (e.g. 2D and 3D information). This fact makes the TIs very useful to carry out easily the screening of large protein databases, such as entire proteomes, by considering information beyond the primary level.

In addition to validation procedures, the receiver operating characteristic (ROC) curve was also constructed for our model. Notably, the curve presented a convexity with respect to the $y = x$ line for the training series (see Fig. 5). This result confirms that the present model is a significant classifier having an area under the curve above 0.8. According to the ROC curve theory, random classifiers have an area of only 0.5, which clearly differentiate our classifiers from those working at random (Swets 1988).

Sequences with $h > h^* = 0.05$ are out of the model's applicability domain. As observed in Fig. 6, most of the sequences in training and test set lies within the model's applicability domain; just seven training and three validation sequences are out.

Particularly, it is important for the model predictability to recognize sequences used in the test set that are outside of its applicability domain. Thus, sequences like pdb1gkrB02 ($h = 0.056$), pdb1ys1X00 ($h = 0.223$) and P22522.1 ($h = 0.073$) should not be predicted as proteinaceous bacteriocins using this model or at least be considered cautiously. Considering such analysis, these three last cases will remain out of the external set increasing slightly the overall prediction percentage from 72.17 up to 72.34%. The new classification results after the removal of such cases from the external set are shown in Table 3.

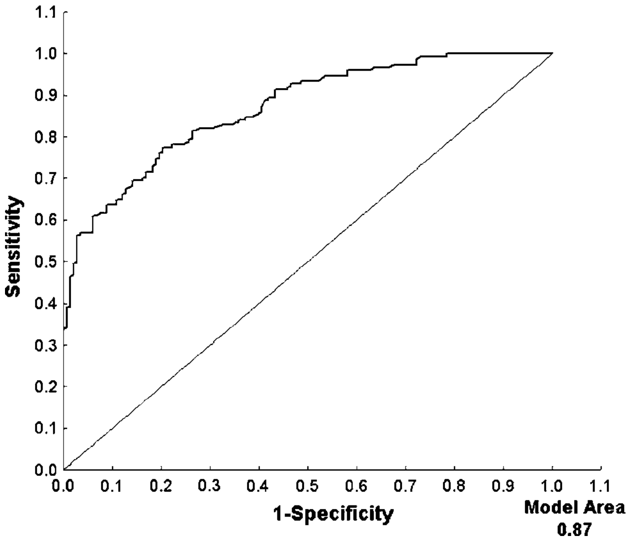


Fig. 5 Receiver Operating Characteristic curve (ROC-curve) for the bacteriocin model (*dark line*) and random classifier (*tight line*) with areas under curve of 0.87, and 0.5, respectively

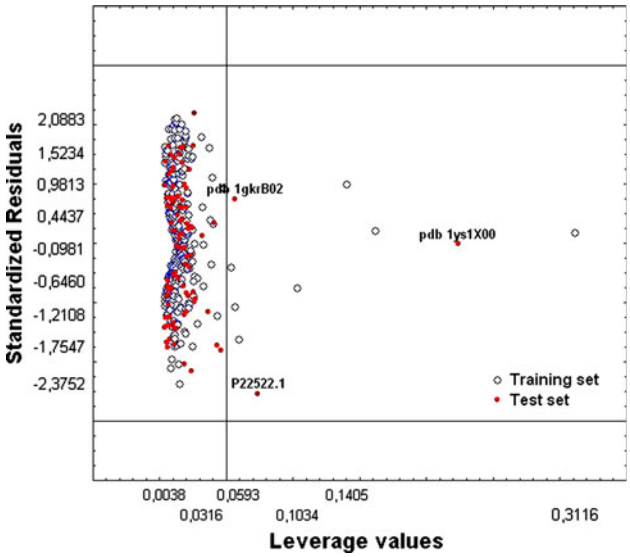


Fig. 6 Graphical representation for the applicability domain of the model

Bacteriocins prediction using classical methodologies

In order to compare the methodology reported here with classical predictive sources of functional annotation, the 196 proteinaceous bacteriocins used in this study were submitted to InterPro analysis using its InterProScan tool (Quevillon et al. 2005). This tool combines different protein signature recognition methods native to the InterPro member databases into one resource with look up of corresponding InterPro and Gene Ontology annotation. Protein signature databases have become vital tools for identifying distant relationships in novel sequences and hence are used for the classification and function deduction of protein sequences. Most of the protein signature recognition methods implemented in InterPro rely up to certain

Table 3 Results of the external set prediction after determining the model's applicability domain

External set	Classification percentage	Control Group	Bacteriocins
Control Group	76.59	**36**	11
Bacteriocins	68.09	15	**32**
Total	72.34	47	47

Numbers in bold highlight the well-classified cases extend on alignment procedures, which justify why we have selected it to carry out a comparative study using our alignment-free approach. In this sense, InterProScan tool did not classify 40 protein sequences out of a total of 196. Out of these 40 non-classified sequences, 16 did not retrieve any hits and the remaining sequences did not have integrated signatures on InterPro database, thus just being classifying 79.6% of the data. In addition, 38 proteinaceous bacteriocins were recognized by InterPro as having other protein signatures, unrelated to bacteriocins-like sequences, thus decreasing the good classification percentage to 60.2% (see Table III of SM). Despite the simplicity of our alignment-free approach, the QSFR model showed a general classification of 78.6% (154/196). This result is not distant from the InterPro's performance considering the unclassified bacteriocins (79.6%), but is considerably higher than the general classification percentage provided by the InterPro (60.2%). Therefore, the identification of proteinaceous bacteriocins using alignment approaches is not a simple task considering the high diversity in its primary structure. Bacteriocins-like sequences could have significant similarities to functional domains unrelated with the bactericidal function per se or may not match any recorded sequence, as suggested by this study. The use of alignment methods will also make difficult the detection of a putative bactericide function in polypeptides or domains that have been traditionally classified in another class. Thus, independent domains belonging to proteins with completely different functions from the bacteriocin class might never be detected unless if using experimental procedures. In this sense, the development of alternative methods not relying on sequence similarity to detect bactericidal function in proteins, polypeptides (domains) and internal domains could be a solution.

An alignment-free prediction to a cryptic bacteriocin-like domain

We provide a practical example of our approach to detect putative bacteriocin-like sequences in internal domains of proteins unrelated with the bactericidal function—the case of the C-terminal domain of the Cry1Ab endotoxin from *Bacillus thuringiensis* subsp. *kurstaki* (Vazquez-Padron et al. 2004). Cry1Ab is one of the most studied insecticidal proteins produced by *B. thuringiensis* as crystalline inclusion body during sporulation (Bravo et al. 2004; Padilla et al. 2006; Pardo-Lopez et al. 2006). Consequently, its nucleotide and amino acid sequence have been recorded for a large number of *B. thuringiensis* strains. Several sequences are nearly identical, and have been designated as variations of the same gene. The crystal protein (Cry) genes specify a family of related insecticidal proteins (Bravo 1997).

Although the Cry 1Ab C-terminal domain is not exported to the medium due to its internal location into a crystal protein, it shares relevant features to bacteriocins such as (1) it is produced by a Gram-positive bacteria (*B. thuringiensis*), (2) inhibits the growth of other bacteria genera like *A. tumefaciens* and *E. coli*, both being Gram-negative bacteria and showing a broad range of bactericidal action, (3) it presents an immunity mechanism to its original host *B. thuringiensis* by binding to the N-terminal portion of the δ endotoxin, and (4) it is encoded by a large *B. thuringiensis* plasmid despite others being chromosomally encoded.

According to these evidences, the C-terminal domain of 549 aa is a bacteriocin-like sequence. However, the sequence is recognized by alignment methods like Basic Local Search Alignment Tool (BLAST) as a delta-endotoxin from *B. thuringiensis*. The InterProScan showed "no hits" meaning no possible classification among the protein classes recorded in the InterPro database. Therefore, the use of alignment-free procedures as TI2BioP represents a complementary alternative to the classical methodologies. The Cry 1Ab C-terminal domain was pseudo folded in a 2D-HP lattice, afterwards the calculation of its TIs (spectral moments) were carried out and the values of $\mu_{1,}$ $\mu_{3,}$ μ_{10} and Edge numbers were evaluated in our classification function. The discriminant equation predicted that the Cry 1Ab C-terminal domain was a bacteriocin-like sequence with a high score of 0.97. The QSFR model prediction is consistent with our experimental observations (Vazquez-Padron et al. 2004).

Moreover, we also applied the water program to find maximal local similarities between the Cry 1Ab C-terminal domain and all proteinaceous bacteriocins used in our study. We investigated common structural features accounting for the cryptic bactericidal action of the C-ter domain. The pair-wise local alignment showed similarities below 50% to the 88.8% of the bacteriocins, with 43.37% of them sharing less than 35% of sequence similarities with our query. That is the case of Q88LD6 classified as a bacteriocin production protein reaching the maximal similarity (80%) in a short region of 15 aa with a low aa identity percentage (see Table IV of SM).

2D-HP maps insight into the bactericide function and evolution of Cry 1Ab C-terminal domain

Alignment procedures based on linear homology are limited to search structural relationships between proteins with similar biological functions but low conservation at the primary level. However, exploiting sequence features beyond the primary level can be insightful in the characterization of a certain protein class. We selected the most representative sequences (the closest ones to the cluster centroid) into the four clusters of proteinaceous bacteriocins divided by the $^{HP}\mu_k$ to perform a bi-dimensional alignment. The 2D alignment of the pseudo-secondary structures of bacteriocins and Cry 1Ab C-terminal domain provided graphical evidence that both are functionally related. Starting from the coordinates (0, 0), a clear superposition between the C-ter domain and the HP-lattice conformed by bacteriocin sequences are shown in Fig. 7. The matching region is evident in contrast with the low-similarity percentages obtained by the Smith–Waterman procedure. This fact supports the relevance of the hydrophobicity and basicity to characterize functionally a bacteriocin-like sequence and the cryptic bactericide action of this Cry 1Ab portion.

The *cry* genes are mostly found in large conjugative plasmids. Such plasmids also contain coding sequences to other proteins being the gene cluster involved in the production and exportation of antibiotic peptides, one of the most amazing determiners (Bravo et al. 2007). For instance, the sequencing on the coding plasmid (*pBtoxis*) to *Bt* subsp. *israelensis* toxins showed the presence of toxin short sequences with homology to central and C-terminal regions of Cry proteins (Berry et al. 2002)**.** These apparent remainders have suggested that during the *pBtoxis* evolution, its ancestors have been host of other toxins that were then lost. Considering that these genes are also characterized for their mobilization by transposition either into this species or in-between others (Barloy et al. 1998; Yokoyama et al. 2004), an evolutionary hypothesis to the finding of the bactericide function of the Cry 1Ab C-ter from the *Bt* subsp. *kurstaki* could be proposed. We believe that such fragment belongs to an ancestral bacteriocin that could have lost its mechanism of exportation.

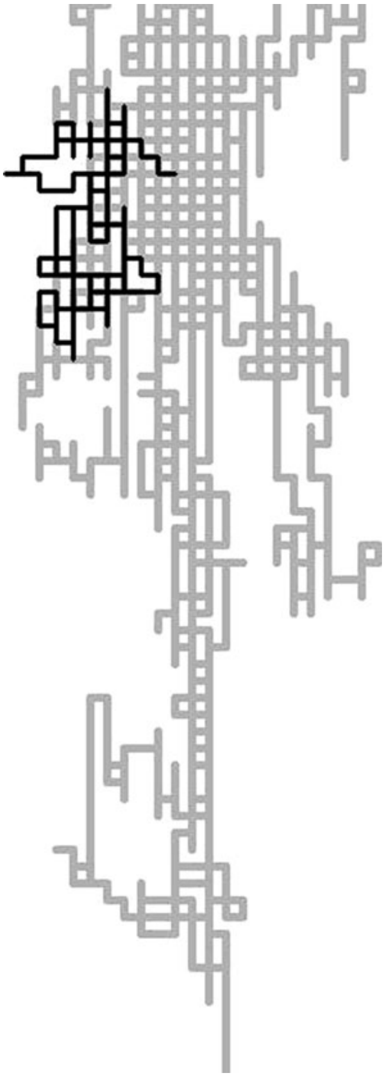


Fig. 7 Pseudo-folding of the Cry 1Ab C-terminal domain sequence (in *black*) into the bacteriocins 2D-HP lattice (in *gray*)

These results confirmed the utility of our alignment-independent method to recognize cryptic bacteriocins that are difficult to identify if using solely alignment procedures. Our method is also effective because it allows the use of graphical procedures to find functional and evolutionary relationships among very distant protein classes. The simplicity and advantage of our approach make it suitable for complementing classical alignment tools, which can be of particular relevance to screen bacterial proteomes for new polypeptides with antibiotic action.

Conclusions

We presented TI2BioP methodology as a successful alternative approach relatively to alignment procedures to identify proteinaceous bacteriocins from domain sequences. Its usefulness stems from the use of 2D-HP protein maps to calculate the spectral moments as TIs. Such TIs condense the hydrophobicity and polarity information of the sequences and were used to develop a simple QSFR classifier. Despite the bacteriocins high diversity, this QSFR model could discriminate successfully the bacteriocin-like sequences among representative CATH domains and showed a good predictability. TI2BioP provided several advantages for the bacteriocins classification relatively to classical protein function annotation methods like InterPro. Moreover, the predictions made by our model for the Cry 1Ab C-ter domain coincided with its cryptic bactericidal action demonstrated in practical experiments. Thus, overlapping of protein pseudo-secondary structures can be an useful alternative to reveal functional and evolutionary relationships of orthologous proteins.

Acknowledgments The authors acknowledge the Portuguese Fundação para a Ciência e a Tecnologia (FCT) for financial support to GACH (SFRH/BD/47256/2008), AMH (SFRH/BPD/63946/2009) and the project PTDC/BIA-BDE/69144/2006 and PTDC/AAC-AMB/104983/2008. GACH acknowledges the Assistant Professor Roberto I. Vázquez-Padrón from the University of Miami, USA for providing useful information on the Cry 1Ab endotoxin from *Bt* subsp. *kurstaki*.

References

Agüero-Chapin G, Antunes A, Ubeira FM, Chou KC, Gonzalez-Diaz H (2008a) Comparative study of topological indices of macro/supramolecular RNA complex networks. J Chem Inf Model 48:2265–2277

Agüero-Chapin G, Gonzalez-Diaz H, de la Riva G, Rodriguez E, Sanchez-Rodriguez A, Podda G, Vazquez-Padron RI (2008b) MMM-QSAR recognition of ribonucleases without alignment: comparison with an HMM model and isolation from Schizosaccharomyces pombe, prediction, and experimental assay of a new sequence. J Chem Inf Model 48:434–448

Agüero-Chapin G, Varona-Santos J, de la Riva G, Antunes A, González-Villa T, Uriarte E, González-Díaz H (2009) Alignment-free prediction of polygalacturonases with pseudofolding topological indices: experimental isolation from coffea arabica and prediction of a new sequence. J Proteome Res 8:2122–2128

Barloy F, Lecadet MM, Delecluse A (1998) Distribution of clostridial cry-like genes among *Bacillus thuringiensis* and *Clostridium* strains. Curr Microbiol 36:232–237

Berry C, O'Neil S, Ben-Dov E, Jones AF, Murphy L, Quail MA, Holden MT, Harris D, Zaritsky A, Parkhill J (2002) Complete sequence and organization of pBtoxis, the toxin-coding plasmid of *Bacillus thuringiensis* subsp. *israelensis*. Appl Environ Microbiol 68:5082–5095

Brandt BW, Heringa J, Leunissen JA (2008) SEQATOMS: a web tool for identifying missing regions in PDB in sequence context. Nucleic Acids Res 36:W255–W259

Bravo A (1997) Phylogenetic relationships of *Bacillus thuringiensis* delta-endotoxin family proteins and their functional domains. J Bacteriol 179:2793–2801

Bravo A, Gomez I, Conde J, Munoz-Garay C, Sanchez J, Miranda R, Zhuang M, Gill SS, Soberon M (2004) Oligomerization triggers binding of a *Bacillus thuringiensis* Cry1Ab pore-forming toxin to aminopeptidase N receptor leading to insertion into membrane microdomains. Biochim Biophys Acta 1667:38–46

Bravo A, Gill SS, Soberon M (2007) Mode of action of *Bacillus thuringiensis* Cry and Cyt toxins and their potential for insect control. Toxicon 49:423–435

Cabrera-Pérez MA, Bermejo Sanz M, Ramos-Torres L, Grau-Ávalos R, Pérez-González M, González-Díaz H (2004) A topological sub-structural approach for predicting human intestinal absorption of drugs. Eur J Med Chem 39:905–916

Cornell WD, Cieplak P, Bayly C, Gould IR, Merz KM Jr, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 117:5179–5197

Cotter P, Hill C, Ross R (2005) Bacteriocins: developing innate immunity for food. Nat Rev Microbiol 3:777–788

Cotter P, Hill C, Ross R (2006) What's in a name? Class distinction for bacteriocins. Nat Rev Microbiol 4

Cruz-Chamorro L, Puertollano MA, Puertollano E, de Cienfuegos GA, de Pablo MA (2006) In vitro biological activities of magainin alone or in combination with nisin. Peptides 27:1201–1209

Cuff AL, Sillitoe I, Lewis T, Redfern OC, Garratt R, Thornton J, Orengo CA (2009) The CATH classification revisited-architectures reviewed and new ways to characterize structural divergence in superfamilies. Nucleic Acids Res 37:D310–D314

de Jong A, van Hijum SA, Bijlsma JJ, Kok J, Kuipers OP (2006) BAGEL: a web-based bacteriocin genome mining tool. Nucleic Acids Res 34:W273–W279

Dirix G, Monsieurs P, Dombrecht B, Daniels R, Marchal K, Vanderleyden J, Michiels J (2004) Peptide signal molecules and bacteriocins in Gram-negative bacteria: a genome-wide in silico screening for peptides containing a double-glycine leader sequence and their cognate transporters. Peptides 25:1425–1440

Eriksson L, Jaworska J, Worth AP, Cronin MT, McDowell RM, Gramatica P (2003) Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ Health Perspect 111:1361–1375

Estrada E (1996) Spectral moments of the edge adjacency matrix in molecular graphs. 1. Definition and applications to the prediction of physical properties of alkanes. J Chem Inf Comput Sci 36:844–849

Estrada E (1997) Spectral moments of the edge-adjacency matrix of molecular graphs. 2. Molecules containing heteroatoms and QSAR applications. J Chem Inf Comput Sci 37:320–328

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 (2007) A tight-binding "Dihedral Orbitals" approach to the degree of folding of macromolecular chains. J Phys Chem B 111:13611–13618

Estrada E, Hatano N (2007) A tight-binding "Dihedral Orbitals" approach to electronic communicability in protein chains. Chem Phys Lett 449:216–220

Estrada E, Uriarte E (2001) Recent advances on the role of topological indices in drug discovery research. Curr Med Chem 8:1573–1588

Fimland G, Eijsink VG, Nissen-Meyer J (2002) Mutational analysis of the role of tryptophan residues in an antimicrobial peptide. Biochemistry 41:9508–9515

Gillor O, Nigro L, Riley M (2005) Genetically engineered bacteriocins and their potential as the next generation of antimicrobials. Curr Pharm Des 11:1067–1075

González MP, Teran C, Teijeira M (2006) A topological function based on spectral moments for predicting affinity toward A_3 adenosine receptors. Bioorg Med Chem Lett 16:1291–1296

Gonzalez-Diaz H, Uriarte E (2005) Biopolymer stochastic moments. I. Modeling human rhinovirus cellular recognition with protein surface electrostatic moments. Biopolymers 77:296–303

Gonzalez-Diaz H, Uriarte E, Ramos de Armas R (2005) Predicting stability of Arc repressor mutants with protein stochastic moments. Bioorg Med Chem 13:323–331

Gonzalez-Diaz H, Perez-Castillo Y, Podda G, Uriarte E (2007a) Computational chemistry comparison of stable/nonstable protein mutants classification models based on 3D and topological indices. J Comput Chem 28:1990–1995

Gonzalez-Diaz H, Saiz-Urra L, Molina R, Gonzalez-Diaz Y, Sanchez-Gonzalez A (2007b) Computational chemistry approach to protein kinase recognition using 3D stochastic van der Waals spectral moments. J Comput Chem 28:1042–1048

Gonzalez-Diaz H, Vilar S, Santana L, Uriarte E (2007c) Medicinal chemistry and bioinformatics—current trends in drugs discovery with networks topological indices. Curr Top Med Chem 7:1015–1029

Gonzalez-Diaz H, Gonzalez-Diaz Y, Santana L, Ubeira FM, Uriarte E (2008) Proteomics, networks and connectivity indices. Proteomics 8:750–778

González-Díaz H, Molina-Ruiz R, Hernandez I (2007) MARCH-INSIDE v3.0 (markov chains invariants for simulation & design), pp Windows supported version under request to the main author contact email: gonzalezdiazh@yahoo.es.

Gutierrez Y, Estrada E (2002) MODESLAB 1.0 (Molecular descriptors laboratory) for Windows.

Hammami R, Zouhir A, Hamida JB, Fliss I (2007) BACTIBASE: a new web-accessible database for bacteriocin characterization. BMC Microbiol 7:89

Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJ, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C (2009) InterPro: the integrative protein signature database. Nucleic Acids Res 37:D211–D215

Jacchieri SG (2000) Mining combinatorial data in protein sequences and structures. Molecular Diversity, pp 145–152

Kaur K, Andrew LC, Wishart DS, Vederas JC (2004) Dynamic relationships among type IIa bacteriocins: temperature effects on antimicrobial activity and on structure of the C-terminal amphipathic alpha helix as a receptor-binding region. Biochemistry 43:9009–9020

Kowalski WJ, Marcoin W (2001) Estimation of bioavailability of selected magnesium organic salts by means of molecular modelling. Boll Chim Farm 140:322–328

Kutner MH, Nachtsheim CJ, Neter J, Li W (2005) Standardized multiple regression model applied linear statistical models. McGraw Hill, New York, pp 271–277

Markovic S, Markovic Z, McCrindle RI (2001) Spectral moments of phenylenes. J Chem Inf Comput Sci 41:112–119

Marrero-Ponce Y, Diaz HG, Zaldivar VR, Torrens F, Castro EA (2004) 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 JA, Olazabal E, Serrano HS, Morales A, Castanedo N, Ibarra-Velarde F, Huesca-Guillen A, Sanchez AM, Torrens F, Castro EA (2005) 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

Mathews DH (2006) RNA secondary structure analysis using RNAstructure. Curr Protoc Bioinformatics chap 12 (Unit 12.6)

Mc Farland JW, Gans DJ (1995a) Cluster significance analysis. In: van Waterbeemd H (ed) Method and principles in medicinal chemistry. VCH, Weinheim

Mc Farland JW, Gans DJ (1995b) Cluster significance analysis. In: Manhnhold R, Krogsgaard-Larsen P, Timmerman V, Van Waterbeemd H (eds) Method and principles in medicinal chemistry, VCH, Weinhiem 2:295–307

Meneses-Marcel A, Marrero-Ponce Y, Machado-Tugores Y, Montero-Torres A, Pereira DM, Escario JA, Nogal-Ruiz JJ, Ochoa C, Aran VJ, Martinez-Fernandez AR, Garcia Sanchez RN (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

Molina R, Agüero-Chapin G, Pérez-González MP (2009) TI2BioP (Topological indices to biopolymers) version 1.0. Molecular simulation and drug design (MSDD). Chemical Bioactives Center, Central University of Las Villas, Cuba

Munteanu CR, Vazquez JM, Dorado J, Sierra AP, Sanchez-Gonzalez A, Prado-Prado FJ, Gonzalez-Diaz H (2009) Complex network spectral moments for ATCUN Motif DNA cleavage: first predictive study on proteins of human pathogen parasites. J Proteome Res 8:5219–5228

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

Niculescu SP, Atkinson A, Hammond G, Lewis M (2004) Using fragment chemistry data mining and probabilistic neural networks in screening chemicals for acute toxicity to the fathead minnow. SAR QSAR Environ Res 15:293–309

Padilla C, Pardo-Lopez L, de la Riva G, Gomez I, Sanchez J, Hernandez G, Nunez ME, Carey MP, Dean DH, Alzate O, Soberon M, Bravo A (2006) Role of tryptophan residues in toxicity of Cry1Ab toxin from *Bacillus thuringiensis*. Appl Environ Microbiol 72:901–907

Pardo-Lopez L, Gomez I, Munoz-Garay C, Jimenez-Juarez N, Soberon M, Bravo A (2006) Structural and functional analysis of the pre-pore and membrane-inserted pore of Cry1Ab toxin. J Invertebr Pathol 92:172–177

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R (2005) InterProScan: protein domains identifier. Nucleic Acids Res 33:W116–W120

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

Rivals I, Personnaz L (1999) On cross validation for model selection. Neural Comput 11:863–870

Sand SL, Haug TM, Nissen-Meyer J, Sand O (2007) The bacterial peptide pheromone plantaricin A permeabilizes cancerous, but not normal, rat pituitary cells and differentiates between the outer and inner membrane leaflet. J Membr Biol 216:61–71

Sang Y, Blecha F (2008) Antimicrobial peptides and bacteriocins: alternatives to traditional antibiotics. Anim Health Res Rev 9:227–235

Santana L, Uriarte E, González-Díaz 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

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197

Statsoft (2007) STATISTICA 7.0 (data analysis software system for windows)

Stein T (2005) Bacillus subtilis antibiotics: structures, syntheses and specific functions. Mol Microbiol 56:845–857

Swets JA (1988) Measuring the accuracy of diagnostic systems. Science 240:1285–1293

Vazquez-Padron RI, de la Riva G, Aguero G, Silva Y, Pham SM, Soberon M, Bravo A, Aitouche A (2004) Cryptic endotoxic nature of *Bacillus thuringiensis* Cry1Ab insecticidal crystal protein. FEBS Lett 570:30–36

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, Gonzalez-Diaz H, Santana L, Uriarte E (2008) QSAR model for alignment-free prediction of human breast cancer biomarkers based on electrostatic potentials of protein pseudofolding HP-lattice networks. J Comput Chem 29:2613–2622

Yokoyama T, Tanaka M, Hasegawa M (2004) Novel cry gene from *Paenibacillus lentimorbus* strain Semadara inhibits ingestion and promotes insecticidal activity in Anomala cuprea larvae. J Invertebr Pathol 85:25–32