U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

CIIMAR
CENTRO INTERDISCIPLINAR DE INVESTIGAÇÃO MARINHA E AMBIENTAL

# A Graphical and Numerical Approach for Functional Annotation and Phylogenetic Inference

**Guillermin Agüero Chapin**

Thesis Advisers

Ph.D. Agostinho Antunes, FCUP, University of Porto, Portugal

Ph.D. Vitor Vasconselos, FCUP, University of Porto, Portugal

Ph.D. Reinaldo Molina Ruiz, University Central of Las Villas, Cuba

Dissertation submitted in fulfillment of the requirements for the degree of Doctor in Biology of University of Porto

**2013**

## Acknowledgments

I acknowledge the Portuguese Fundação para a Ciência e Tecnologia (FCT) for financial support (SFRH/BD/47256/2008) to carry out this work. My gratitude to my research advisors, Ph.D. Agostinho Antunes, Ph.D. Vitor Vasconcelos and Ph.D. Reinaldo Molina, who provided scientific guidance and economical support for the successful culmination of this project. Thanks to my colleagues from LEGE, CIIMAR, for allowing me an easy integration with the group; they were very kind since my arrival. Gratefulness to colleagues, ex-colleagues and friends from the Molecular Simulation and Drug Design Group in Santa Clara, Cuba, for their valuable technical contributions, especially to Ph.D. Humberto González and Ph.D. Maykel Pérez for the support given while developing the **TI2BioP** software. Thanks to Ph.D. Aminael Sánchez for his orientation and help with the overall research, particularly for encouraging the learning of programming languages. I also acknowledge the revision performed by the Ph.D. Ricardo Medina that significantly improved the thesis writing.

Agradezco especialmente a:

*Mi familia por apoyar siempre mi superación profesional y por su amor incondicional*

*Mi primera y única amiga gallega, que me abrió las puertas de su casa, su familia y su corazón...Marta Teijeira Bautista*

*Mi amigo en Porto, en Galicia y espero que en todas partes y por siempre...Luis Cagide Fajin*

*Doris, la Dra pinareña y su esposo Luis, por su ayuda y hospitalidad desde el 2008 hasta la fecha*

*Angela por ser mi amiga,modelo y silenciosa compañia durante las tediosas compras*

*María por traer nuevamente el toque latino a la residencia y por resumir tan bien*

*Alius y Leticia por ser mis mejores amigas cubanas y ayudarme de alguna forma en la elaboración del trabajo*

*Mi suegri por quereme como un hijo y compatir mis éxitos y amarguras*

*Mi querida Gisselle por su total dedicación en la revisión, edición, discusión técnica de la tesis y sobretodo por ser la Chichi*

## Abstract

We developed a new graphical/numerical method called **TI2BioP** (**T**opological **I**ndices **to BioP**olymers) to estimate topological indices (TIs) from two-dimensional (2D) graphical approaches for DNA/RNA and protein data. TIs were used to build up alignment-free models to the functional classification and to infer evolutionary relationships without alignments in highly diverse gene/protein classes with relevance for the drug discovery process. The method was effective in the detection of new members and remote protein homologous compared to alignment procedures and was further confirmed by experimental evidences. The 2D Cartesian protein representation and its TIs were able to unravel the Cry 1Ab C-terminal domain from *Bacillus thuringiensis´*s endotoxin as a bacteriocin-like protein, which has not been detected by classical alignment methods. We registered in public databases new members of the internal transcribed spacer (ITS2) and of the RNase III protein classes using alignment-free models. The predictions of these two members were verified through enzymatic assay for the new RNase III member and by evaluating both queries against profiles Hidden Markov Models (HMM). The amino acid clustering strategy to build 2D Cartesian protein maps was extrapolated to generate a non-classical profile HMM with high prediction accuracy to detect RNase III members. Classical profiles HMM showed a lower classification performance on the ITS2 genomic class than alignment-free models in spite of the fact they were generated by multiple sequence alignment (MSA) algorithms improved for sets of low overall sequence similarity. The new ITS2 sequence isolated from *Petrakia* sp. was used by alignment-free based techniques applied for the first time to the estimation of phylogenetic inferences. The *Petrakia* sp. fungal isolate was placed inside the *Pezizomycotina* subphylum and the *Dothideomycetes* class. Finally, we demonstrated that the use of graphical/numerical-based models in cooperation with alignment sequence search methods provided the most reliable exploration of the Adenylation domains (A-domains) repertoire of Nonribosomal Peptide Synthetases (NRPS) in the *Microcystis aeruginosa* proteome. The knowledge of the complete A-domain repertoire in the proteome of a cyanobacteria species may allow unraveling new NRPS clusters for the discovery of novel natural products with important biological activities.



**Key words:** 2D graphs/ Topological Indices/ Sequence similarity/ Alignment-free models/ Functional annotation/ Phylogenetic analysis.

## Resumo

Neste trabalho foi desenvolvido um novo método gráfico-numérico **TI2BioP** (**T**opological **I**ndices **to BioP**olymers) para fornecer índices topológicos (ITs) a partir de aproximações gráficas a duas dimensões (2D), com o objectivo de construir modelos livres de alinhamentos para a classificação funcional de DNA, RNA e proteínas. O método permite igualmente inferir relações evolutivas em diversas classes de genes ou proteínas, directa ou indirectamente relacionadas com a descoberta de novos fármacos. O método foi efectivo na detecção de novos membros e de proteínas homólogas remotas, quando comparado com procedimentos que usam alinhamentos, tendo sido ainda confirmado por evidências experimentais. A representação cartesiana 2D das proteínas e seus ITs permitiu desvendar que o domínio Cry 1Ab C-terminal da endotoxina de *Bacillus thuringiensis* é uma proteína semelhante à bacteriocina, facto que não tinha sido possível detectar com métodos clássicos de alinhamento. Foram detectadas em bases de dados públicas novos membros das classes de proteínas RNase III e do espaçador interno transcrito (ITS2) usando modelos livres de alinhamentos. Essas predições foram verificadas através de ensaios enzimáticos, no caso do novo membro das RNase III, e através da avaliação de perfis HMM em ambos os casos. A estratégia de agrupamento de aminoácidos para construir os mapas cartesianos 2D de proteínas foi extrapolada para gerar um perfil HMM -clássico com alta precisão para a detecção de membros da classe RNase III. No entanto, os perfis clássicos HMM gerados por algoritmos de alinhamentos múltiplos de sequências (MSA), melhorados para conjuntos de sequências de baixa semelhança global, mostraram um pobre desempenho na classificação da classe genómica ITS2 relativamente aos modelos livres de alinhamentos. A nova sequência ITS2 isolada de *Petrakia* sp. foi utilizada pela primeira vez para fazer inferências filogenéticas com técnicas não baseadas em alinhamentos. O nosso grupo colocou o isolado fúngico de *Petrakia* sp. dentro do subphylum *Pezizomycotina* e a classe *Dothideomycetes*. Finalmente, demonstramos que o uso de modelos gráficos-numéricos juntamente com métodos de alinhamentos de pesquisa de sequências fornecem os resultados mais fiáveis nas explorações do repertório dos domínios de adenilação (domínios A) das sintetases de péptidos não ribosomais (NRPS) no proteoma de *Microcystis aeruginosa.* O conhecimento de todo o repertório de domínios A no proteoma das diversas espécies de cianobactérias poderá revelar novos grupos de NRPS, o que poderá potencialmente permitir a descoberta de novos produtos naturais com importantes actividades biológicas.



**Palavras-chave:** Gráficos 2D/ Índices Topológicos/ Similaridade de sequências/ Modelos livres de alinhamento/ Anotação funcional/ Análise filogenética.

# Index

# List of Figures

# List of Tables

## List of Abbreviations

| | |
|---|---|
| AAC | Amino acid composition |
| A | Adenylation |
| A-domains | Adenylation domains |
| ATSd | Broto–Moreau autocorrelation |
| AASA | AminoAcid Sequence Autocorrelation |
| ANN | Artificial Neural Networks |
| BLAST | Basic Local Alignment Search Tool |
| CGT | Chemical Graph Theory |
| CATH | Class, Architecture, Topology and Homology |
| C&RT | Classification and Regression Trees |
| CT | Classification Trees |
| CV | Cross-validation |
| C | Condensation |
| D | Dimensional |
| DTM | Decision Tree Models |
| Ed | Euclidean distance |
| FFT | Fast Fourier transform |
| GDA | General Discrimination Analysis |
| HMM | Profile Hidden Markov Models |
| ITS2 | Internal transcribed spacer |
| J | Balaban index |
| JC | Jukes-Cantor |
| K2P | Kimura-2-parameter |
| k-MCA | K-Means cluster analysis |
| MARCH-INSIDE | **Mar**kov **Ch**ain **I**nvariants for **N**etwork **S**election **& De**sign |
| MSA | Multiple Sequence Alignment |
| MAFFT | Multiple Alignment based onFFT |

| | |
|---|---|
| MCL | Maximum Composite Likelihood |
| MLP | Multilayer Layer Perceptron |
| NRPS | Nonribosomal Peptide Synthetases |
| NW | Needleman-Wunsch |
| NJ | Neighbour-joining |
| PINs | Protein interaction networks |
| PAM | Point Accepted Mutation |
| PDB | Protein Data Bank |
| PG | Polygalacturonase |
| PKS | Polyketide synthases |
| PseAAC | Pseudo amino acid composition |
| QSAR/QSPR | Quantitative-Structure-Activity/Property Relationship |
| ROC | Receiver Operating Characteristic |
| SVM | Support Vector Machine |
| SW | Smith-Waterman |
| S | Score |
| siRNAs | Short interfering RNAs |
| TIs | Topological Indices |
| TOPS-MODE | **To**pological **S**ubstructural **Mo**lecular **De**sign |
| **TI2BioP** | **T**opological **I**ndices **to BioP**olymers |
| UTRs | Untranslated regions |
| W | Winner index |

## List of Original Papers

This thesis is based on the following articles:

1. **Agüero-Chapin G**, Pérez-Machado G, Molina-Ruiz R, Pérez-Castillo Y, Morales-Helguera A, Vasconcelos V and Antunes A. **TI2BioP**: Topological Indices to BioPolymers. Its practical use to unravel cryptic bacteriocin-like domains. ***Amino Acids.*****2011; 40(2):431-42.**

2. **Agüero-Chapin G,** de la Riva GA, Molina-Ruiz R, Sánchez-Rodríguez A, Pérez-Machado G, Vasconcelos V and Antunes A. Non-linear modelsbasedon simple topological indices toidentifyRNase III protein members. ***Journal of Theoretical Biology***. **2011; 273(1):167-78**.

3. **Agüero-Chapin G,** Sánchez-Rodríguez A, Hidalgo-Yanes PI, Pérez-Castillo Y, Molina-Ruiz R, Marchal K, Vasconcelos V and Antunes A. An alignment-free approach for eukaryotic ITS2 annotation and phylogenetic inference. ***PLoS ONE*** **2011; 6(10).**

4. **Agüero-Chapin G,** Molina-Ruiz R, Maldonado E, de la Riva GA, Vasconcelos V and Antunes A. Exploring the adenylation domain repertoire of nonribosomal peptide synthetases using an ensemble of sequence-search methods. ***PLOS ONE*** **2013*;* 8(7).**

# INTRODUCTION

# 1. Introduction

Graphical approaches have been successfully used in several branches of science such as mathematics, physics, chemistry, biochemistry, biology and computer science to visualize complex relationships, including functional relations, outcomes of complicated processes and interactions, as well as to simplify scientific notation. A graph is a collection of vertices or nodes and a compilation of edges that connect pairs of vertices; they have been deeply studied in graph theory, a branch of discrete mathematics, to model pairwise relation between objects from a certain collection [1].

Graph theory has facilitated the development of Chemical Graph Theory (CGT) to allow combinatorial and topological exploration of the chemical molecular structure through the calculation of mathematical descriptors [2].The molecular topology is represented as graphs where atoms and bonds are considered as vertices and edges of the graph, respectively. It is possible to derive molecular descriptors from the graph representing an approximation of the molecular structure to carry out Quantitative-Structure-Activity/Property Relationship (QSAR/QSPR). Such mathematical descriptors have been traditionally used in QSAR/QSPR studies for small sized molecules. When these methodologies are applied to drugs, they can be used in medicinal chemistry for modeling drug design and drug-receptor relationships [3, 4].

More recently with the emergence of genomics and proteomics, the CGT is being extended to bioinformatics through the characterizations of DNA/RNA and proteins for comparative analysis without the use of sequence alignments. In genomics and proteomics, nucleotides, amino acids, proteins, electrophoresis spots, polypeptidic fragments, or more complex objects can play the role of nodes and the bonds or the relation either functional or geometrical between them are considered the edges of the graph [2].Thus, we can simplify complex biological systems like proteomes, metabolic networks and protein interaction networks (PINs) into the topology of a graph providing support to gain useful insights into such systems. All of these graphs or networks can be numerically described using the so-called Topological Indices (TIs) [5]. TIs are numerical indices derived from the graph-theoretical representation of a molecule as a whole and contain information about the connections between atoms in the molecule and the properties for the connected atoms [6]. Therefore, a topological index is the numerical representation of the information extracted from a chemical complex but it can be easily extended to characterize biological systems as mentioned above [7].

The use of TIs to characterize biosequences (DNA/RNA and proteins) to perform massive analyses without alignments is an active research topic in bioinformatics. To perform the numeric characterization of genes and proteins families through the calculation of TIs, we build a graph representing a DNA/RNA and protein sequence where nodes are represented by nucleotides or amino acids while the connections are normally represented by covalent bonds, hydrogen bridges, electrostatic interactions, van der Waals bonds and so on [8-10]. There are several types of TIs depending on the complexity of the biomolecule representation that could comprise several dimensional (D) spaces. Linear sequences (sequence order), is a one dimensional (1D) representation, while two-dimensional (2D) and three dimensional (3D) are related to sequence arrangement or geometry into these spaces [11-13]. Particular attention has been placed into the 2D representations which do not represent the "real structure" of the natural biopolymers but have been very effective in inspecting similarities/dissimilarities among biopolymers either by direct visualization or by numerical characterization [2]. Examples of 2D artificial representations for DNA and protein sequences with potentialities in bioinformatics include the spectrum-like, star-like, cartesian-type and four-color maps [2, 14-17]. These DNA/RNA and protein maps can generally unravel higher-order useful information contained beyond the primary structure, i.e. nucleotide/amino acid distribution into a 2D space. Their essence can be captured in a quantitative manner through TIs to easily compare a great number of sequences/maps [18-21].

Despite the complexity of the biomolecule representation, for the calculation of any TI is necessary the creation of the adjacency matrix. There are variants of the adjacency matrix e.g., node and edge adjacency matrix [22]. They translate the connectivity/adjacency relations between nodes or edges in the graph to a matrix arrangement [23]. The adjacency matrix is a square matrix where nodes or edges are numbered without specific order. The elements of the adjacency matrix $n_{ij}$ or $e_{ij}$ are equal to 1 if i and j are adjacent otherwise take the value of 0. When two nodes share a common edge are adjacent; the same consideration is applied for two edges sharing at least one node [24]. Once the adjacency matrix is built, there are several algorithms to calculate the TIs. One of the most common algorithms used in QSAR for TIs calculation can be represented according to the key vector-matrix-vector scheme. Several authors have reported TIs using this algorithm, such as the Winner index (W) [25], firstly defined in a chemical context; and others like Randić invariant (χ) [26], Balaban index (J) [27], Broto–Moreau autocorrelation (ATSd) [28] and the spectral moments introduced by Estrada [29].

The last ones are based on the method of moments developed in the 70'ś and applied in solid-state physics and chemistry. Estrada *et al..* extended these concepts to the use of

bond moments and included bond weights related to hydrophobic, electronic and steric molecular features. The spectral moments are defined as the sum of main diagonal entries of the different powers of the bond adjacency matrix [30]. This matrix is the same adjacency matrix described above where non-diagonal entries are 1 or 0 if the corresponding bonds share one atom or not but with the particularity that main diagonal entries are weighted with bond properties. Spectral moments were implemented in the **TOPS-MODE** (**To**pological **S**ubstructural **Mo**lecular **De**sign) program [31] and have been widely validated by many authors to encode the structure of small molecules in QSAR studies [32-34] including the characterization of the folding degree of proteins based on its dihedral angles [35, 36]. Despite, the versatility of the spectral moments in QSAR studies, they have been poorly used to describe biopolymers structures excepting when they promoted the arising of the Estrada folding index (*I*3) for proteins [37]. We extended the spectral moments to search structure-function relationships in DNA/RNA and proteins classes with no alignments, turning the spectral moments into alignment-free predictors applied to annotate biological functions of genes/proteins [8, 19, 38].

The prediction of the biological function, 2D and 3D structure of a query gene or protein has traditionally relied on similarity measures provided by alignment algorithms, to other recorded members of the family. The first tools to distinguish biologically significant relationships from chance similarities were based on dynamic programming algorithms; the Needleman-Wunsch [39] and Smith-Waterman algorithms [40]. Needleman-Wunsch algorithm was reported in 1970 to calculate global similarity scores between two sequences and it is more suitable when the homologies have been previously set, e.g. when evolutionary trees are built [39]. The Smith–Waterman procedure was proposed in 1981 to determine similar regions between two sequences instead of looking at the total sequence. This algorithm is able to detect sub-regions or sub-sequences with evolutionary conserved signals of similarity avoiding regions of low sequence similarity [40]. However the high computational cost of dynamic programming algorithms makes them impractical for searching large databases [41]. Rapid heuristic algorithms inspired on above-mentioned methods were developed to perform sequence searches against large databases in normal computers such the cases of FASTA [42] and Basic Local Alignment Search Tool (BLAST) programs [43]. FASTA algorithm was reported in 1985 by Lipman *et al.*. to find locally similar regions between two sequences based on identities but not gaps [42]. Such regions were rescored using a measure of similarity between residues, such as a PAM matrix [44]. BLAST was developed by Karlin and Altschul in 1990 to perform rapid sequences comparison

optimizing the local similarity as the maximal segment pair (MSP) score. BLAST can be applied to DNA and protein sequence database searches, motif searches, gene identification searches, and in the analysis of multiple regions of similarity in long DNA sequences. It is a simple and robust method that works faster than the other existing sequence comparison tools mentioned so far [43].

Today, the most powerful sequence-based comparison methods use sets of aligned sequences, either as profiles like HMM [45, 46] or position specific scoring matrices (PSSMs) [47]. Profile/HMM/PSSM methods are more sensitive than single-sequence comparison methods because they summarize the evolutionary history of a family, identifying more and less conserved positions within the protein [48]. PSSMs are the essence of the Position Specific Iterative-BLAST (PSI-BLAST), a very sensitive comparison tool that has revealed homologies between sequences that previously were recognized only from structure [47]. Profiles HMM are the core of the popular Protein family (Pfam) database made up for alignment profiles representing more than 3071 protein families [49]. Profiles HMM are sensitive tools to detect structural and functional protein signatures in large databases even when the sequence conservation is restricted.

In short, alignment methods have being improved their sensitivity to detect functional signals in query sequences by using several strategies like the substitution of the original similarity matrixes among the aligned sequences for PSSMs [47], the addition of other steps in BLAST searches [50-52] and the implementation of stochastic predictive models such as the case of profiles HMM for DNA/RNA and protein sequences [49, 53].

Similar efforts have been carried out to improve the quality of MSA in phylogenetic reconstruction accuracy [54-56]. Phylogenetic tree reconstruction is traditionally based on MSAs and heavily depends on the validity of this information bottleneck [57]. CLUSTALW was reported in 1994 to align any number of homologous nucleotide/protein sequences with an improved sensitivity and it is widely utilized since then due to its well performance in practice [58]. Later, other algorithms have tried to improve MSA on the accuracy of CLUSTALW. T-Coffee reported in 2000 employed a similar progressive strategy but achieved a higher accuracy alignment by combining information derived from global and local multiple alignments [56]. MAFFT reduced drastically the CPU time in respect to T-Coffee and CLUSTALW. Homologous regions are rapidly identified by fast Fourier transform (FFT) using a simplified scoring system. It provides increased alignment accuracy even for sequences having large insertions as well as distantly related sequences of similar length [59]. Edgar in 2004 implemented progressive and iterative refinement alignment strategies in MUSCLE. The speed and accuracy of MUSCLE are compared with T-Coffee, MAFFT and CLUSTALW [55]. More recently, DALIGN-TX was developed as a segment-based multiple

alignment tool improved for sets of low overall sequence similarity. On locally related sequences, DIALIGN-TX outperforms all other programs without increasing the CPU time [60].

Although the detection of functional signals in gene/protein classes has been improved as well as the quality of MSA to produce reliable phylogenetic trees; low sequence similarity or the similarity to genes/proteins lacking functional annotations represent a drawback for the performance of alignment algorithms [61]. It is still difficult to produce reliable alignments for proteins that share less than 30-40% of identity [61, 62]. Consequently, several alignment-independent approaches are being developed to overcome this limitation for an effective functional annotation and for reliable phylogenetic inferences in highly diverse gene/protein families.

Most of the alignment-free classifiers have been based on amino acid composition such as the one reported by Strope and Moriyama in 2007 to detect remote similarity of G-protein-coupled receptor superfamily members using support vector machines [63]. In the same year, alignment-free descriptors such as amino acid content and amino acid pair association rules were used along with routinely available classification methods to classify protein sequences [64]. The web-server Composition based Protein identification (COPid) was developed in 2008 by Kumar *et al.*. to exploit the full potential of protein composition to annotate the function of a protein from its composition using whole or part of the protein [65].

One of the most popular alignment-free approaches is the Chou"s concept of pseudo amino acid composition (PseAAC) introduced in 2001; it reflects the importance of the sequence order effect in addition to the amino acid composition to improve the prediction quality of protein cellular attributes [66]. This concept has been widely used to predict protein subcellular location [67], enzyme family classes [68], membrane protein types [69], protein quaternary structure [70] and many others protein attributes. A similar approach was developed by Caballero and Fernandez defined as Amino Acid Sequence Autocorrelation (AASA) vectors but instead of using a distance function (property difference) like in the PseAAC, they used autocorrelation (property multiplication) to predict the conformational stability of human lysozyme mutants [71]. AASA is an extension of the Broto-Moreau autocorrelation TIs used before in SAR studies to protein sequences [28]. Following the same philosophy González-Díaz and co-workers have scaled their Markovian descriptors and stochastic spectral moments to characterize protein sequences. They implemented 1D, 2D and 3D TIs into the **MARCH-INSIDE** (**Mar**kov **Ch**ain **I**nvariants for **N**etwork **S**election **&**

**De**sign) methodology to correlate biological properties with peptides and proteins structure [72].

Few alignment-free approaches have dealt with high sequence divergence [73] to increase the reliability of phylogenetic inferences [74]. There is evidence that at low divergence between sequences, the genetic code is a better indicator of the phylogenetic relationships while at high divergence better accuracy is achieved by focussing on amino acid properties [75]. The most relevant alignment-independent approaches reported for phylogenetic trees reconstruction have been based on patterns discovered in unaligned sequences [76], amino acid composition [65] and a kernel approach for evolutionary sequence comparison [74].

While several types of graphical/numerical methods have been developed to carry out comparative analyses for DNA/RNA and proteins with no alignments [11], very few studies have exploited its potentialities in bioinformatics to perform functional predictions and phylogenetic inferences in highly diverse gene/protein classes. Therefore, we aimed to develop a new graphical/numerical tool inspired on previous experiences achieved by other methodologies such as **TOPS-MODE** and **MARCH-INSIDE**, to face alignment problems. Such graphical/numerical tool was called **TI2BioP** (**T**opological **I**ndices **to BioP**olymers) because it allows the calculation of the original spectral moments as simple TIs from different 2D graphical approaches for DNA, RNA and proteins biopolymers. **TI2BioP** will be assessed by predicting functional classes and by inferring phylogenetic relationships at low sequence similarity [19].Thus, we placed the following hypothesis:

- Is **TI2BioP** methodology an effective tool to develop alignment-free models to predict biological functions for highly diverse gene/protein families and useful in the molecular evolutionary field to obtain reliable phylogenetic trees?

To validate this hypothesis we have selected four gene/protein families that share common features: (i) low sequence similarities among its members and (ii) relative involvement in the drug discovery process. Furthermore, the gene/protein family members studied belong to both prokaryotic and eukaryotic organisms.

The gene/protein families studied were the following:

1. Proteinaceous bacteriocins are toxins produced and exported by both gram-negative and gram-positive bacteria as a defense mechanism. 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 Bacteriocins can be applied as food preservatives and are of great interest for novel antibiotics development and as diagnostic agents for some cancers.

2. Ribonuclease III (RNase III) class shows variable homology and different domain structures. RNases III are useful for drug search or drug-target candidates for drug development because these enzymes are involved in several important biological processes.
3. ITS2 gene class shows a high sequence divergence, which has traditionally complicated ITS2 annotation and limited its use for phylogenetic inference at low taxonomical level analyses (genus and species level). ITS2 is the standard gene target for fungal taxonomical identification at the species level [77], being especially relevant to identify fungal species that cannot be cultured such as the potential producers of bioactive compounds [78].
4. A-domains from NRPS show low sequence similarity between its members. NRPS are megasynthetases composed by several domains organized in clusters for the synthesis of oligopeptides having different biological activities.

In summary, the objectives of this thesis are the following:

1. To develop the **TI2BioP** methodology based on the graph theory to generate alignment-free predictors (spectral moments).
2. To evaluate the ability of **TI2BioP** to detect functional signatures from the above-mentioned gene/protein classes and to identify new members of such families in cooperation with experimental evidences and alignment procedures.
3. To compare **TI2BioP** performance in detecting functional signatures among the gene/protein classes involved in this study against alignment algorithms like profiles HMM.
4. To demonstrate that our TIs are also useful for molecular evolutionary inferences.

# MATERIALS AND METHODS

# 2. Materials and Methods

## 2.1. General scheme of procedure

In this chapter, information about the general procedure and methods used in this study is provided. Alignment-free models for functional classification of highly diverse gene/protein families were developed using TIs derived from 2D graphical representations for DNA and protein sequences. The classification performance of such models were compared with the obtained by classical alignment procedures from the same database. Both methods were used in cooperation to annotate the function of new gene/protein members and to re-annotate proteomes (Figure 2.1).

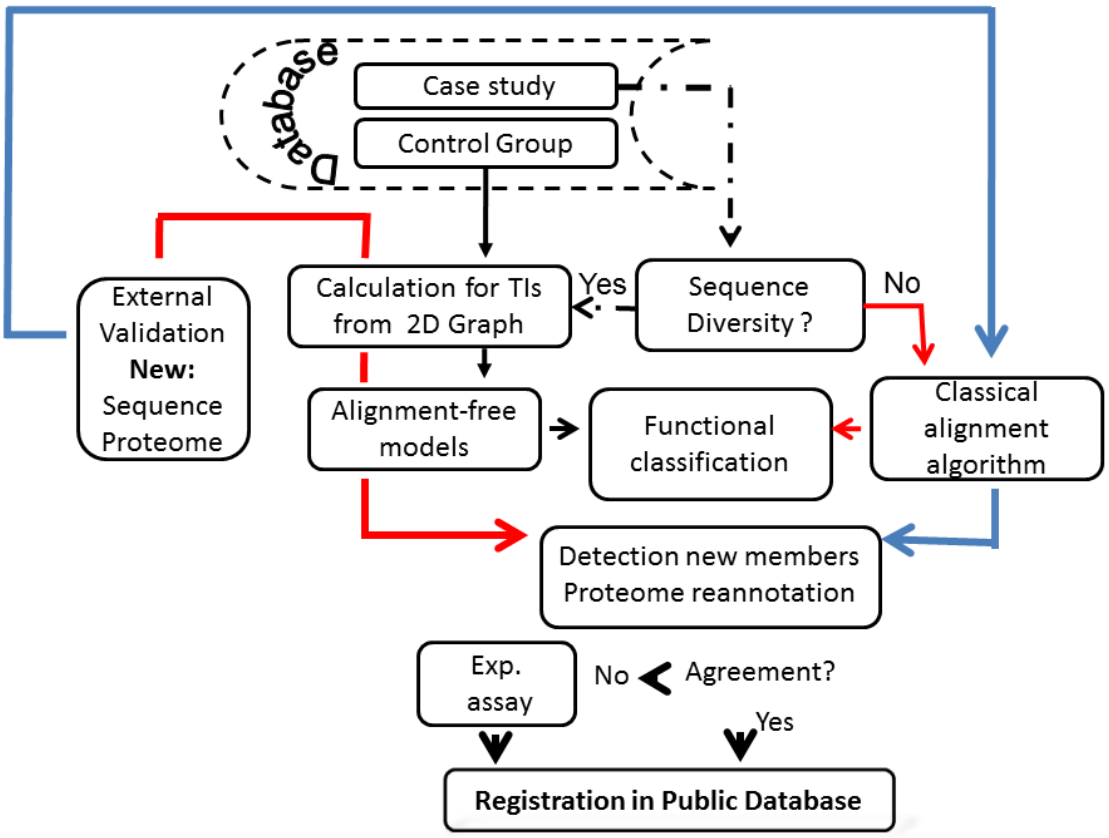



**Figure 2.1.** Flowchart of the general procedure to evaluate alignment-free models for functional classification of gene/protein classes involved in the study.

## 2.2. Database

To test the performance and efficacy of our alignment-free approach **TI2BioP** to detect DNA and protein signatures and to infer phylogenetic relationships, four gene/protein families having low sequence similarity among their members according to some reports were selected. In addition, such gene/protein families are relevant for drug discovery.

The gene/protein classes studied were:

1. *Bacteriocin protein class*: A total of 196 bacteriocin-like proteins sequences belonging to several bacterial species were collected from the two major bacteriocin databases, BAGEL [50] and BACTIBASE [79].
2. *Ribonuclease III class (RNase III)*: 206 RNase III protein sequences belonging to prokaryote and eukaryote species were downloaded from GenBank database gathering all RNAses III registered up to May of 2009**.**
3. *ITS2 class*: A total of 4 355 ITS2sequences from a wide variety of eukaryotic taxa (http://its2.bioapps.biozentrum.uni-wuerzburg.de) were used.
4. *Adenylation domains (A-domains)*: 138 A-domain sequences from NRPS were collected from the major NRPS–PKS database (http://www.nii. res.in/nrps-pks.html).

Because a negative set or control group to develop classification models is needed, three different control groups were selected according to some features: (1) structurally well-characterized sequences (2) high functional diversity among its members and (3) similar sequence lengths in respect to the study case.

*Protein control groups*:

1. Sequences from **C**lass, **A**rchitecture, **T**opology and **H**omology (**CATH**) domain database (version 3.2.0) (http://www.cathdb.info) sharing only 35% of sequence similarity were selected to provide a functional representation and avoid structural redundancy. This group was used as a control to develop the alignment-free models to recognize bacteriocin-like and A-domains sequences.
2. High-resolution proteins in a structurally non-redundant and representative subset from the Protein Data Bank (PDB) made up of enzymes and non-enzymes were also used. This subset was selected according to independent-sequence similarity criteria but using simple features such as secondary-structure content, amino acid

propensities, surface properties and ligands. Redundancy was removed by structural alignments to provide just representative structures. This protein subset was used as a control group to develop alignment-free models to detect RNase III enzymes [61].

*Gene control group*:

1. A non-redundant subset containing both 5‟- and 3‟-untranslated regions (UTRs) sequences from the fungi kingdom was selected from the eukaryotic mRNAs database: UTRdb (http://www.ba.itb.cnr.it/UTR/). It was selected as a control group to identify ITS2 members with no alignments because this class comprises diverse but structurally related genomic sequences to the ITS2 class. Similar to the ITS2, UTRs are non-coding regions with divergence among the eukaryotes but showing a more conserved secondary structure when are transcribed into RNAs [80].

### 2.2.1. Training and test subsets selection

After assessing the diversity of the study cases; both study case and control group were divided independently into training and test subsets. The selection of training and test members was carried at random when the size of the study case and the control group sets were originally balanced. Otherwise, both sets (study and control group) were balanced to avoid classification bias. K-Means cluster analysis (k-MCA)was performed to reduce representatively the size of control groups (generally, the control group is larger than the study case) [81].This procedure required a partition of the study case and its control group independently into several statistically representative K centers according the TIs values for each case (sequence). Each K center is the mean of the cases assigned to each cluster. Finally, clustering process was driven by structural features due to the intrinsic nature of TIs. Members that made up training and test sets were straightforwardly selected from such clusters.

### 2.2.2 Methods to explore database diversity

*Pairwise alignment*

Although a high dissimilarity among gene/protein members of each study case involved in this study was previously reported [52, 79, 82, 83]; the quantification of such diversity had not been recorded yet. It has been demonstrated that the reliability of the predicted biological function and the phylogenetic reconstruction dramatically decreases when gene/protein

families share pair-wise sequence similarities lower than 50% [57, 61, 84]. Thus, the Smith-Waterman (SW) [85] and Needleman-Wunsch (NW) [39] dynamic programming algorithms were used to perform local and global sequence similarities either between pairs of DNA or pairs of proteins (all *vs* all) to explore the sequence diversity of gene/protein families involved in the present study.

### 2.3. TI2BioP software

TIs were calculated by our in-house **TI2BioP** software from different 2D graphical approaches applied to DNA/RNA and proteins. **TI2BioP** is based on the graph theory considering the "building blocks" of DNA/RNA and protein biopolymers as nodes and the bonds between them as edges into the 2D graphs. Consequently, the information contained in biopolymeric long strings is simplified in the topology of 2D graphs that is eventually determined by the sequence order and the nucleotide/amino acid composition of these biopolymers.

**TI2BioP** was inspired mainly in TOPS-MODE [31] methodology for the calculation of the spectral moments as TIs but using the platform of **MARCH-INSIDE** program [72]. It was built up on object-oriented Free Pascal IDE Tools (lazarus) running either on Windows or Linux operating system. Its latest version (version 2.0) is freely available at http://**ti2biop**.sourceforge.net/. The friendly interface of **TI2BioP** allows the users to access the sequence list introduction, selecting the representation type and calculations of TIs. The software just needs an input data containing either DNA or protein sequences into a fasta format file to be represented as 2D artificial but informative graphs (Figure 2.3).

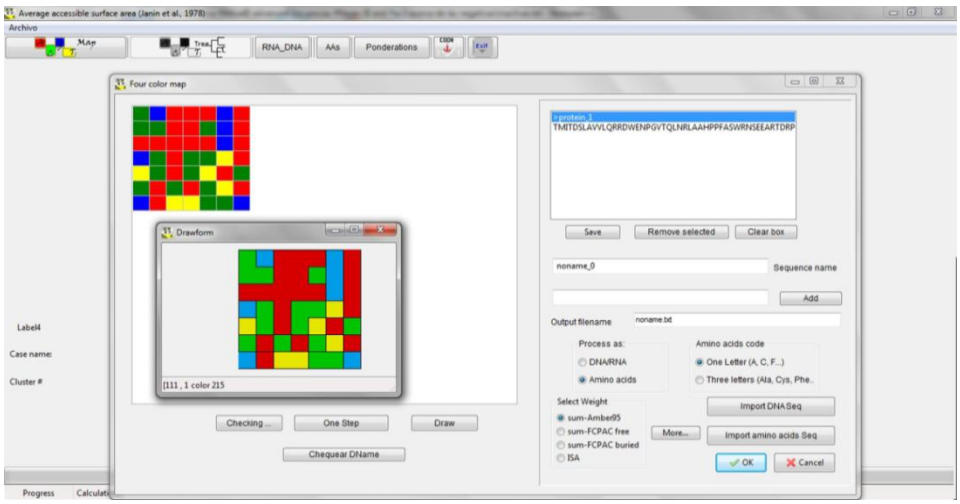



**Figure 2.3.**Window View of **TI2BioP** for the representation of protein four-color maps

Two main types of 2D artificial representations have been implemented so far in the software (i) one based on Cartesian representation for DNA strings introduced by Nandy [86] and the other inspired on the four-color maps reported by Randic [87]. These two 2D artificial graphs were implemented in **TI2BioP** to represent DNA and protein sequences as well as the spectral moments calculations for each type of 2D DNA and protein maps. It is important to highlight that **TI2BioP** can also import files containing 2D structure inferred by other professional programs e.g. the RNASTRUCTURE [88] for the calculation of the spectral moments as TIs.

### 2.3.1. 2D graphical representations of TI2BioP

*Cartesian representation for DNA and RNA*

We used the Cartesian representation reported by Nandy for the calculation of the spectral moments as TIs [86, 89]. This 2D representation is obtained by arranging a DNA primary sequence into a 2D Cartesian system following the sequential appearance of its bases. Purine and pyrimidine bases are placed on the Cartesian system by assigning them to X and Y axes, respectively. The representation is built by adding the k-th nucleotide of the DNA sequence to the coordinates (0, 0) of the Cartesian system. The first nucleotide is placed in the origin of the Cartesian system (0, 0) and it is represented as a square dot (figure 2.3.1). The value (1, 0) is assigned if the (k + 1)-th nucleotide is Guanine (rightwards-step); (-1, 0) if Adenine (leftwards-step); (0, 1) if Cytosine (upwards-step) or (0, -1) if the (k + 1)-th nucleotide is Thymine or Uracil (downwards-step) in the case of RNA sequences. The resulting path gives the overall graphical representation of DNA, though the information on the individual steps has been lost. The 2D Cartesian map for the sequence (AGCTG) is showed in the figure 2.3.1C; note that the central node contains both Guanine and Thymine nucleotides.

(A) $A_1$-$G_2$-$C_3$-$T_4$-$G_5$

(B)

| Nucleotide | Class | X position | Y position |
|---|---|---|---|
| $A_1$ | purine | 0 | 0 |
| $G_2$ | purine | -1 | 0 |
| $C_3$ | pyrimidine | -2 | 0 |
| $T_4$ | pyrimidine | -1 | 0 |
| $G_5$ | purine | -1 | 1 |

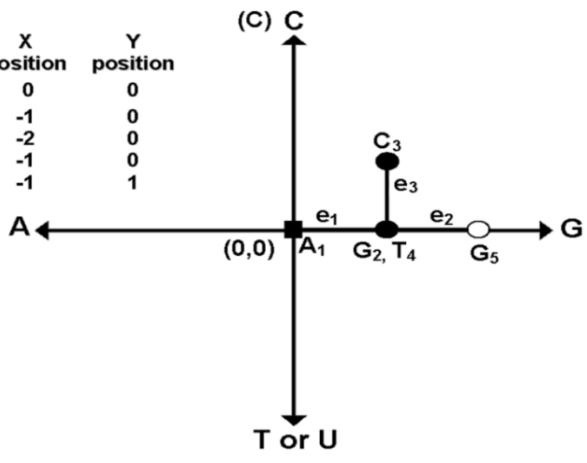



**Figure 2.3.1.** Building the 2D-Cartesian map for the (A) DNA fragment AGCTG. (B) The coordinates for each nucleotide in the Cartesian system. (C) The 2D-Cartesian map

*Cartesian representation for proteins*

This representation is based on the graphs introduced by Nandy for DNA sequences described above but instead of using the four nucleotides assigned to each axe direction in the Cartesian space, all twenty amino acids were grouped into four groups: acid, basic, polar and non-polar amino acids [90]. These four groups characterize the physicochemical nature of the amino acids in terms of hydrophobicity (H) and polarity (P) [91]. Each amino acid in the sequence is placed in a Cartesian 2D space starting with the first amino acid at the (0, 0) coordinates. The coordinates of the successive amino acids are calculated as follows:

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

Similar to 2D-Cartesian maps for DNA, more than one amino acid could be assigned to one node in the protein map. The information contained in the linear sequence is arranged into a 2D space of hydrophobicity and polarity (2D-HP) according to the sequence order and amino acid composition of proteins. 2D-HP protein maps provide a topology that depends on these two structural features, sequence order and amino acid composition, the same as to the real secondary structure.

The 2D-HP map of the sequence ($D_1$-$E_2$-$D_3$-$K_4$-$V_5$) is showed in the figure 2.3.2. Please note that the central node contains both E and K amino acids.

(a) **$D_1$-$E_2$-$D_3$-$K_4$-$V_5$**

| Amino acid | HP class | X position | Y position |
|---|---|---|---|
| $D_1$ | acid | 0 | 0 |
| $E_2$ | acid | -1 | 0 |
| $D_3$ | acid | -2 | 0 |
| $K_4$ | basic | -1 | 0 |
| $V_5$ | non-polar | -1 | 1 |

(b)

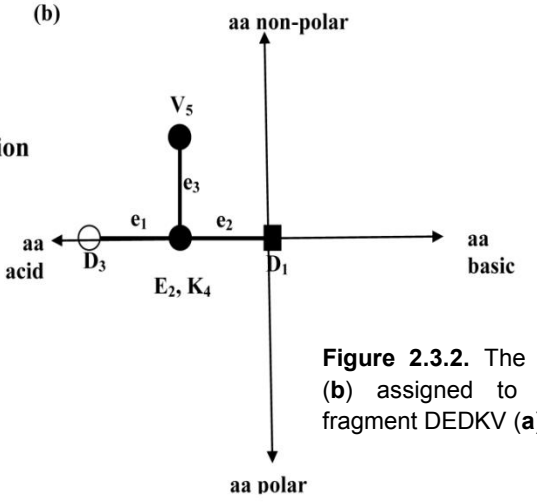



**Figure 2.3.2.** The 2D-HP map (**b**) assigned to the protein fragment DEDKV (**a**)

The figure 2.3.3 shows the complete 2D-HP map for the new RNase III member from *Escherichia coli* BL21 substrain GG1108. Its two major domains are highlighted in red (RNase3 domain) and in blue (double-stranded RNA binding motif), respectively.

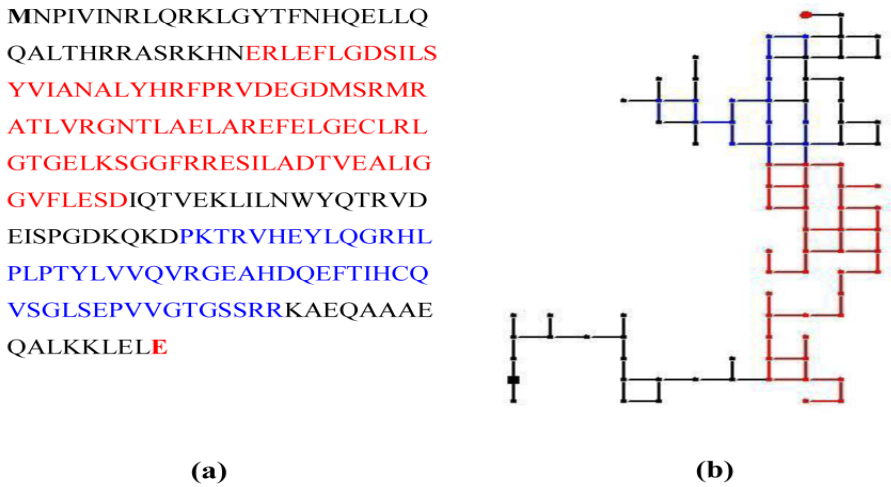



**Figure 2.3.3.** (**a**) RNase III protein sequence from *Escherichia coli* BL21 substrain GG1108 (**b**) 2D-HP map for the RNase III protein.

*Four-color maps for proteins*

Protein four-color maps are inspired on Randic"s DNA/RNA [92] and protein 2D graphical representations [87]; but instead of using the concept of virtual genetic code, we have constructed the spiral of square cells straightforward from the amino acid sequences [21]. The four colours are assigned to the four amino acids classes (polar, non-polar, acid and basic) used previously by our group in Nandy"s representation for proteins [38, 90] (Figure 2.3.4).

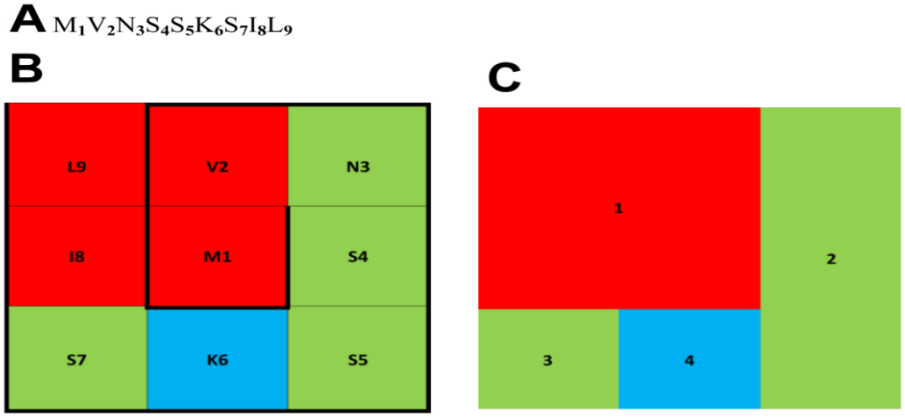



**Figure 2.3.4.** Steps for the four-color map construction for the first nine amino acids of 1 pdb AMU. (**A**) The first nine aminoacids of pdb 1AMU. (**B** and **C**) Building the four-color map for this protein fragment.

Figure 2.3.5 shows how the four-color map for the first A-domain structurally characterized (pdb 1AMU) is built up. The four colors are associated with each one of the amino acid groups: polar (green), non-polar (red), acid (yellow), basic (blue).

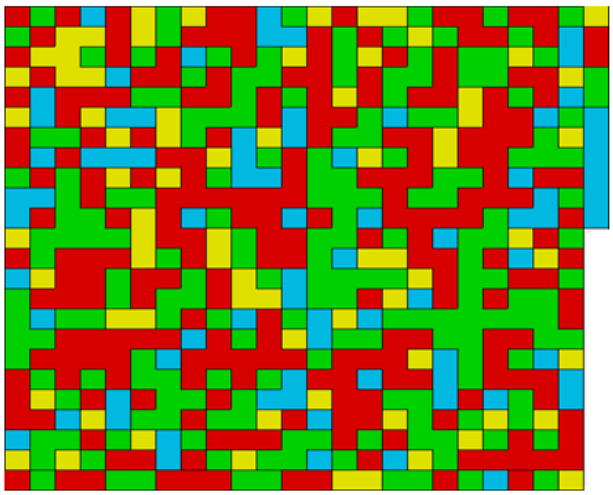


**Figure 2.3.5.** The final four-color map for pdb 1AMU.

### 2.3.2. Spectral moments calculation for different 2D graphical representations

*Spectral moments for 2D-Cartesian DNA and RNA maps*

In order to calculate the spectral moments, an edge adjacency matrix **B** was assigned to each graph after the representation of DNA/RNA sequences. **B** is a square symmetric matrix whose non-diagonal entries are either ones or zeroes if the corresponding edges (e) share or not one node in the 2D-Cartesian DNA map. These spectral moments are defined as the trace of **B** as is indicated in the equation below.

$$\mu_k = Tr\left[(B)^k\right] \qquad (1)$$

Where Tr is called the trace and indicates the sum of all values in the main diagonal of the matrices $^k\mathbf{B} = (\mathbf{B})^k$, which are the natural powers of **B** [29]. The different powers of **B** give rise to the spectral moments series ($\boldsymbol{\mu}_0$- $\boldsymbol{\mu}_{15}$). 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 nucleotide bonds in the sequence. The main diagonal was weighted with the average of the electrostatic charge (Q) between two bound nodes. The charge value q in a node is

equal to the sum of the charges of all nucleotides placed on it. The electrostatic charge of one nucleotide was derived from the Amber 95 force field [93].Thus, it sets up connectivity relationships between the nucleotides into the 2D Cartesian map but these interactions are characterized electronically providing a pseudo secondary structure for DNA sequences. When we apply equation 1 for the calculation of $\boldsymbol{\mu}_k$ values, we are considering the topology of DNA pseudo-folding and its electronic features to numerically characterize DNA and RNA sequences; therefore we use the term pseudo-folding spectral moment ($^{pf}\boldsymbol{\mu}_k$) (equation 2)

$$^{pf}\mu_k = Tr\left[(B)^k\right] \qquad (2)$$

In order to illustrate the calculation of the spectral moments for this type of representation, an example is developed in section 1.1 of the annexe 3.

*Spectral Moments for 2D-Cartesian protein maps*

We used the same philosophy for the calculation of spectral moments as TIs described for 2D Cartesian maps of DNA/RNA sequences. They were calculated from the adjacent matrix **B** which is modified according to the building of the 2D-Cartesian protein map; also called 2D-HP protein map due to the HP nature of the 2D Cartesian space. Thus, we used this nomenclature ($^{HP}\mu_k$) for the spectral moments series from the 2D Cartesian protein maps or 2D-HP protein maps. These TIs are analogously calculated as the trace of **B** (equation 2). We used the same property (electrostatic charge) to weight the main diagonal of **B** but using the values for amino acids.

*Spectral moments from inferred DNA/RNA secondary structures*

In addition to the development of artificial but informative representations for natural biopolymers; there are other approaches for inferring the secondary structure of DNA/RNA molecules. The Mfold algorithm implemented in the RNASTRUCTURE 4.0 software [94], which is based on the minimization of the folding energy (lowest ΔG), is one of the methods to infer DNA and RNA secondary structures. The algorithm generates 2D DNA/RNA maps topologies containing stems and loops formed by possible hydrogen bond interactions between nucleotides placed at middle and long range in the sequence. This graphical information is also provided through the connectivity table (ct files). Ct files contain information about the connection between nucleotides in the secondary structure generated with thermodynamic models [94]. Ct files containing topological information are imported by

**TI2BioP** to generate the bond adjacency matrix **B** for the calculation of spectral moments ($^{mf}\boldsymbol{\mu}_k$), based on folding thermodynamics parameters. The main diagonal was weighted with the average of the electrostatic charge (Q) between two bound nucleotides as described above.

*Spectral moments for protein four-color maps*

To calculate the spectral moments for protein four-color maps, we considered each map region as a node made up of the amino acids clustering; two adjacent regions of the map sharing at least one edge (not a vertex) are connected. **B** is calculated in a similar way but instead of considering the adjacency relationships between edges, it is set up between nodes. The number of nodes or clusters in the graph is equal to the number of rows and columns in **B**. Since a cluster is made up of several amino acids sharing similar physicochemical properties, the cluster is weighted with the sum of individual properties (e.g. electrostatic charge (q) [93]) of all amino acids placed in a cluster). The main diagonal of **B** was weighted with the average of the electrostatic charge (Q) between two adjacent clusters. The calculation of the spectral moments up to the order k = 3 from the four colours maps using the first nine amino acids of pdb 1AMU ($M_1V_2N_3S_4S_5K_6S_7I_8L_9$) is illustrated in details in the methods section of annex 4 [21].

### 2.3.3. Alignment-free models developed from TI2BioP's spectral moments

Spectral moments series were used as alignment-free predictors to develop classification models to detect gene/protein members and to calculate alignment-free distances in order to reconstruct phylogenetic relationships at low sequence similarities. The classification models were built up using the following statistical techniques:

*Linear Models. General Discrimination Analysis*

General Discrimination Analysis (GDA) was selected as the linear statistical technique to perform predictors selection (spectral moments from **TI2BioP**) and to develop alignment-free models [95-97]. Both, model and variable selection were based on the revision of Wilk"s ($\lambda$) statistics ($\lambda$ = 0 perfect discrimination, being $0<\lambda<1$). The Fisher ratio (F) was also inspected to indicate the contribution of one variable to the discrimination between groups with a probability of error (*p*-level) p(*F*) < 0.05.

*Non-linear models. Decision Tree Models (DTM)*

The application of DTM as alignment-free models to protein functional classification and to search for protein signatures was introduced for first time in this work. We used the Classification and Regression Trees (C&RT)-style univariate split selection from the Classification Trees (CT) module of the STATISTICA 8.0 for Windows [98].The *Gini* index was used as a measure of goodness of fit and the ¨Prune on misclassification error¨ was set up as an stopping rule to select the right-sized classification tree.

*Artificial Neural Networks (ANN)*

We used the Multilayer Layer Perceptron (MLP) network architecture as the most popular network architecture in use today. The selection of the subset of predictors that were most strongly related to the response variable was supported on the *Feature and Variable Selection* analysis of the ANN module from *STATISTICA* software [98]. The right complexity of the network was selected by testing different topologies to the MLP while checking the progress against a selection set to avoid over-fitting during the two-phase (back propagation/conjugate gradient descent) training algorithm [99].

*Evaluation of Model performance*

The performance of all alignment-free models was evaluated by several statistical measures commonly used for classification: accuracy, sensitivity, specificity and F-score (it reaches its best value at 1 and the worst score at 0). The prediction power of such models was evaluated on the test set (this subset was not used to train the model) and the same statistical parameters were applied to show the prediction performance. The area under the Receiver Operating Characteristic (ROC) curve, commonly known as AUC was also calculated for the training and test sets to evaluate the classifier"s performance (a value of AUC=1.0 means a perfect predictor and 0.5 a random predictor).

*Validation procedures*

The reliability of the classification models was verified by 10-fold cross-validation (CV) procedure on both training and test sets. The CV statistics for each of the ten samples were averaged to give a 10-fold estimate for the accuracy, sensitivity and specificity for both subsets [100]. The external validation was carried out using the test set as an external set to measure the predictability of the alignment-free models. In addition to this external validation, new gene/protein sequences (not registered in any database) and proteomes were used as real external cases to evaluate the prediction power of the alignment-free

models. The predictability (performance in the prediction) was always compared to alignment classification algorithms either for the test set or for new isolated sequences.

### 2.4. Alignment-based Methods for Functional Classification

*InterPro resource*

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. The sequence classification was carried out by the InterProScan tool [101] looking into the InterPro database [102].

*Profile Hidden Markov Models (HMM)*

Profiles HMM provide a classification model to predict structural and functional attributes for gene/protein families [103, 104]. The HMMER software containing the *hmmbuild* and *hmmsearch* programs, was used to obtain different profiles HMM [46]. *Hmmbuild* demands an MSA file obtained by any of the MSA methods that will be described below to build the profile HMM for a certain gene/protein family. The generated HMM profile is used for the hmmsearch to detect DNA/protein signatures against a database.

*Basic Local Alignment Search Tool* (*BLAST)*

BLAST is a widely used sequence search method to find matches to a query sequence within a large sequence database, such as Genbank. Although BLAST does not generate a predictive model as profiles HMM; it can be used for classification purposes. The similarity of a query sequence to others already annotated in a database is measured through a goodness score (S) and an estimate of the expected number of matches (*E*-value) with an equal or higher score than would be found by chance. Whether the query sequence is similar enough (positive) or not to others registered is decided based on the score and E-value threshold [47].

In this work we used two types of BLAST:

1. BLASTn search (*E*-value cutoff = $10e^{-10}$) against the NCBI database to contrast the annotation of a new ITS2 genomic sequence.
2. A multiple-template BLASTp reported by the NRPS-PKS database developers for NRPS (Adenylation (A), Condensation (C) and Thiolation (T)) domains searches was

used [52]. Multiple-template BLASTp consist in using each one of the (A, C and T) domain sequences as template to evaluate each query of a certain database by BLASTp (E-value=10) using BLOSUM62 scoring matrix and default values for gap penalties. The best matches under these conditions were retrieved. Namely, we used it to search the A-domain signature against the proteome of the cyanobacteria *Microcystis aeruginosa*.

### 2.5. Experimental validation

Although, the prediction performance of our alignment-free models in respect to alignment classification methods was assessed using a characterized test set, new isolated sequences were evaluated by both approaches as well as the re-annotation of the cyanobacteria *Microcystis aeruginosa* proteome. New sequences were registered at Genbank if at least two models (alignment-free models and alignment algorithms) agreed in the prediction; if they disagreed then an *in vitro* experimental test was used for a definitive functional annotation. A general experimental procedure was followed to isolate new gene/protein members that were used in this further external validation. A new member from the ITS2 genomic and the RNase III protein classes was isolated, respectively as follows:

1. DNA extraction from *Escherichia coli* and the fungus *Petrakia* sp.
2. PCR amplification of the ITS2 genomic sequence from *Petrakia* sp. and amplification of *E. coli* RNase III gene from *Escherichia coli BL 21* strain CG 1208
3. Sequencing of both PCR products
4. Purification and enzymatic assay of recombinant *E. coli* RNAse III

The isolation, expression and the cryptic bactericide function of the Cry 1Ab C-terminal domain from *Bacillus thuringiensis´*s endotoxin was indirectly included in this work because its sequence was used to evaluate the bacteriocin alignment-free model. Its bactericide function was unraveled by the author of this thesis under the supervision of Vázquez-Padrón R.I in 2004 [105].

### 2.6. Phylogenetic analysis

We have declared the hypothesis that our TIs are not only useful for functional classification of gene/protein classes but also to infer phylogenetic relationships. So, we used distance-based methods to reconstruct phylogenetic relations among different fungal classes applying the ITS2 biomarker. The taxon of interest was the fungus *Petrakia*sp. Phylogenetic trees were constructed using the Neighbour-joining (NJ) method. NJ trees

were generated from different sequence distance matrices: (1) alignment-based and (2) alignment-free distances

1. Alignment-based distances: Evolutionary distances computed using Jukes-Cantor (JC), Kimura2-parameter (K2P) and Maximum Composite Likelihood (MCL) substitution models were obtained using the MEGA4.

2. Alignment-free distances: Euclidean distances (Ed) were computed from TIs values of sequences involved in the analysis, whereas phylogenetic trees were obtained by applying hierarchic clustering methods (furthest neighbor linkage).

The validation of the classical NJ trees was mainly supported by bootstrap values and hierarchic clusters were validated by the cophenetic correlation coefficient. The topology consistency of both trees was also evaluated using other distances.

*Multiple sequence alignment (MSA) methods*

We used several MSA methods for different purposes; either to build profiles HMMs to search for a certain gene/protein signature in a sequence database or to carry out phylogenetic analyses. Depending on the sequence similarity degree of the gene/protein families involved in the study, different MSA methods were applied, such as:

1. CLUSTALW: It uses progressive alignment methods, then progressively more distant groups of sequences are aligned until a global alignment is obtained [58].
2. DIALIGN-TX: DALIGN-TX is a segment-based multiple alignment tool improved for sets of low overall sequence similarity [60].
3. MAFFT: Multiple Alignment (MA) based on fast Fourier transform (FFT), in which an amino acid sequence is turned into a sequence composed of volume and polarity values of each amino acid residue. It is suitable for sequences having large insertions or extensions as well as for distantly related sequences of similar lengths [106].

CLUSTALW and DIALIGN-TX were run using the default parameters. In the case of MAFFT, the iterative alignment option (L-INS-I) was used [40, 41].

# RESULTS

# 3. Results

In this section, the results are supported by the four articles published by the author during the thesis period. Each one of the author"s papers is attached under the heading ANNEXES at the end of this manuscript. They are organized from the **TI2BioP** version 1.0 implementation until its latest version 2.0 available at http://**ti2biop**.sourceforge.net/. Both software versions were applied to the gene/protein families in the same order given in the introduction. Pages belonging to the published works keep the actual journal numbering while thesis's sections follow the appropriate numeration system.

The results presented herein are related to the application of **TI2BioP** to the functional classification of proteinaceous bacteriocins, RNase III, ITS2 and NRPS A-domains. All these classes show high sequence divergence among their members and are related to the drug discovery process either by providing secondary metabolites with biological activities (bacteriocin and NRPS) or by representing drug targets (RNases III). The methodology was also useful to develop alignment-free based techniques applied to phylogenetic inference to complement the taxonomy of the *Petrakia* sp. fungal (putative producer of bioactive compounds) isolates using the ITS2 biomarker.

In general, all papers contain the development of alignment-free models using different statistical classification techniques and the TIs generated by **TI2BioP** as input predictors, for functional annotation of the families mentioned above. These models were built with linear and non-linear statistical techniques and validated by cross and external validation procedures. The applicability of the models was demonstrated by detecting new members belonging to each gene/protein class, which were supported by experimental evidences and by classical alignment methods. Table 3.1 shows the best reported alignment-free model for the functional classification of each gene/protein family involved in the study and the procedure carried out to achieve the functional annotation of new members by such models, in cooperation with experimental evidences and alignment procedures. The performance of the alignment-free models was always contrasted with classical alignment procedures such as InterPro and profiles HMM for functional detection of the chosen gene/protein classes (Table 3.2). The graphical approach **TI2BioP** was used to visualize functional relationships and also to allow an alignment-free molecular taxonomy driven by numerical indices.

Table 3.1. Best reported alignment-free models for the functional classification of each gene/protein family studied. New detected members of each gene/protein class and the procedure carried out for their functional annotation.

| Gene/protein class | Control Group | 2D-Graph Type | Best-Reported Alignment-Free Model | New Detected Members | Annotation Procedure |
|---|---|---|---|---|---|
| **Protein Bacteriocins** | CATH domains | Cartesian | GDA | Cry 1Ab C-terminal domain *Bacillus thuringiensis* | 1- Alignment-free prediction<br>2- Experimental evidences |
| **Genomic ITS2** | 5´and 3´UTRs | Cartesian and Mfold | ANN | ITS2 genomic *Petrakia* sp. | 1-Alignment-free prediction<br>2- Homology-based prediction |
| **RNase III** | Non-redundant subset (enzymes and non-enzymes) PDB | Cartesian | DTM | Rnase III *E coli BL21 substrain CG 1208* | 1- Alignment-free prediction<br>2- Homology-based prediction<br>3- Experimental evidences |
| **A-domains NRPS** | CATH domains | Four-color map | DTM | 5 hits in the proteome of *Microcystis aeruginosa* | No registration |

Table 3.2. Prediction performance measured through the sensitivity on the test set and the identification of the new member for the best reported alignment-free and alignment-based method. When alignment-based procedures achieved a sensitivity of 100%, complex algorithms were applied.

| | **Alignment-free models** | | | **Alignment-based procedures** | | |
|---|---|---|---|---|---|---|
| **Gene/protein class** | **Statistical Technique** | **Sensitivity Test set** | **New Member Prediction** | **Alignment algorithm** | **Sensitivity Test set** | **New Members Detection** |
| **Protein Bacteriocins** | GDA | 66.67% | Significant hit | InterPro | 60.2% | No-hit |
| **Genomic ITS2** | ANN | 92.59% | Significant hit | Profile HMM (MAFFT) | 66.66% | Significant hit |
| **RNase III** | DTM | 96.07% | Significant hit | Profile HMM (modified) | 100% | Significant hit |
| **A-domains NRPS** | DTM | 100% | Significant hits | profiles HMM | 100% | Significant hits |

# DISCUSSION

# 4. Discussion

This thesis was inspired on previous works carried out with the **MARCH-INSIDE** methodology to annotate biological functions in gene/protein classes from plants. We had previously reported the 2D-Cartesian representation for proteins and its numerical characterization through sequence coupling numbers calculated by such method graphical/numerical method [90]. Coupling sequence numbers are calculated in analogy to other stochastic molecular descriptors using the Markov Chain Theory. **MARCH-INSIDE** software provides stochastic molecular descriptors calculated from a node adjacency matrix whose elements are transition probabilities between the connected nodes [72]. Thus, we published the first alignment-free model built up with coupling numbers derived from this graphical representation to functionally classify polygalacturonase (PG) members from plants [90]. The PG family is the most intensively studied of all cell wall modifying enzymes expressed during the ripening process [107]. PG members were detected with high accuracy by our reported alignment-free model. Despite the fact that this study opened a door to the application of graphical methods in bioinformatics for the detection of functional signatures in protein families, its usage was still limited [18, 90]. Members of PG protein class show high sequence similarity and consequently classical alignment procedures provide an excellent performance to detect functional signatures at such high similarity level. Alignment algorithms are the most popular techniques in bioinformatics; they are based on similarity measures (number of nucleotides/amino acids matches) between a new gene/protein member against others registered in a database or against a sequence family profile to predict functional and structural attributes of new members [47, 108]. These methodologies (e.g. BLAST, Pfam, InterPro) have a friendly interface for undemanding users to search for sequences as well as for structural and functional classifications, but they show a low performance in detecting members belonging to highly diverse gene/protein families [19, 61, 63]. Several authors have provided evidence that the reliability of the predicted biological function dramatically decrease when protein families have pair-wise sequence similarities below 50% [57, 61, 84]. It has been also reported a twilight zone for the alignment algorithms where it is difficult to produce accurate alignments for proteins that share less than 30-40% of identity [61, 62]. Alignment-dependent algorithms ignore structural information beyond the linearity of the sequence, e.g. long-distance interactions. They are focused only on „positive" samples (protein family members) in the dataset without any contribution of „negative" samples (non-members) to the training of the algorithms. Other weakness of these methods

arises when a query sequence is similar to genes/proteins lacking functional annotations [109]. In addition, phylogenetic inferences relying on MSA methods are not reliable when gene/protein sequences show functional similarities but have greatly diverged [57].

Nevertheless, the first reported alignment-free model based on 2D-Cartesian protein maps using stochastic molecular descriptors was rather illustrative than useful to overcome the limitations of alignment algorithms [90]. Afterwards, the 2D-Cartesian protein representation was numerically characterized through stochastic spectral moments to detect a particular RNAse III member (Pac 1) among several different proteins of this class [110]. Methods based on sequence alignments have revealed a low amino acidic identity (20-40%) for the *pac1*+ gene product with other typical RNases III. However, experimental observations have shown RNAse activity for the Pac1 protein [111]. These evidences represented a motivation for developing alignment-free models based on graphical approaches as an alternative to traditional alignment procedures for functional annotation. The first results were encouraging because a simple linear equation could detect signatures of the RNAse III from a highly diverse dataset and with a higher accuracy than that achieved using alignment procedures [110]. This report provided some clues about the potential of graphical/numerical approaches as alternative tools to detect remote homologous due to their alignment-independence.

Considering these previous promising studies, we aimed to overcome such alignment limitations when using highly diverse gene/protein families through the creation of **TI2BioP** software as a new platform comprising several graphical approaches for DNA/RNA and proteins and its numerical characterization through simple TIs [112]. The TIs consist in the bond spectral moments introduced by Estrada which has been the inspiration for other stochastic molecular descriptors mentioned before [29] (annex 1). The use of different graphical approaches allows extracting different sequence information contained into these natural biopolymers and provides flexibility and diversity to the spectral moments calculations. These TIs were considered as alignment-free predictors for the development of classification models to annotate biological functions of gene/protein classes sharing low sequence similarity. They were also used to unravel functional and phylogenetic relationships using graphical and numerical sequence characterizations [8, 19, 38].

Although the gene/protein families selected to highlight the utility of **TI2BioP** show high sequence divergence among their members, they have also played an important role in the drug discovery process from natural sources. The bacteriocin protein class was the first

study case used to validate our methodology (annex 1). Bacteriocins are proteinaceous compounds of bacterial origin that are lethal to bacteria other than the producing strain. They have attracted attention as potential substitutes for, or as additions to, currently used antimicrobial compounds and also as probiotics [113].

The bacteriocin protein family is highly diverse in terms of size, method of killing, method of production, genetics, microbial target, immunity mechanisms and release, which contribute to its low pair-wise sequence similarity (23-50%). This makes bacteriocin classification a challenge based on alignment procedures [114], demanding the implementation of complex strategies [115, 116]. Since hydrophobicity and basicity are major criteria for the bactericide activity of bacteriocins, we clustered the 20 natural amino acids into four groups according to their hydrophobicity (H) and polarity (P) properties (polar, non-polar, acidic or basic amino acid [91]. Amino acids were placed into the 2D Cartesian system according to those (HP) properties arranging the sequence in a 2D-HP space. Using this artificial protein representation, spectral moments series ($^{HP}\mu_k$) from an adjacency matrix of edges were calculated for the first time. Using $^{HP}\mu_k$ series, a simple alignment-free model supported by linear statistical techniques was built up. Our model was effective to detect bacteriocin proteins from a highly diverse dataset made up of non-redundant CATH domains and bacteriocin sequences, retrieving 66.7% of the bacteriocin-like proteins from an external test set while the InterPro resource could just detect 60.2%. To our knowledge, this is the first report where an alignment-free model based on a graphical approach outperforms a popular alignment-based resource for functional sequence annotation. Due to the diversity of the protein bacteriocin class, the InterPro resource either had significant similarity matches to functional domains unrelated to the bactericidal function *per se* or did not find significant matches with any integrated sequence into this resource (results and discussion are detailed in annex 1).

On the other hand, clustering the 20 amino acids into four HP classes in a 2D Cartesian space provides graphical profiles which are generated by both the sequence order and the amino acid composition. Thus, such graphical profiles contain useful information beyond the primary structure. The calculated TIs captured numerically the essence of this artificial representation allowing the development of highly predictive models for bacteriocins detection. In fact, a remote bacteriocin homologous was detected in the Cry 1Ab C-terminal domain from *Bacillus thuringiensis´*s endotoxin, which had not been detected by classical alignment methods. Although bacteriocins and Cry 1Ab C-terminal domain share common biological features and function (explained in annex 1), their sequences are completely

different and consequently placed into two different protein classes by alignment procedures. The functional relationship between both protein classes had been assessed by experimental procedures in previous reports but no bioinformatic method was able to unravel it. **TI2BioP** methodology predicted successfully the bacteriocin function of the Cry 1Ab C-terminal either by using the generated alignment-free model or by the superposition of the 2D-Cartesian maps for Cry 1Ab C-terminal domain to other representative bacteriocins. This sort of 2D graphical alignment provided clues about the functional relation between them [8]. This graphical analysis has been useful to visualize similarities/dissimilarities between different protein classes in previous reports [90, 110].

The discovery of new bacteriocins contributes to the development and design of probiotics and antibiotics of narrow spectrum that prevent the arising of microbial resistance. Particularly, the bactericide Cry 1Ab C-terminal domain could be used in biotechnology either for developing positive selection cloning vectors or for regulating the growth of *E. coli* cultures.

Following the success of using 2D-HP maps for detecting distant homologues in the bacteriocin protein class, we applied them to detect RNAse III protein members through the calculation of the spectral moments (annex 2). RNase III family shows high diversity among its members regarding to the primary structure and domain organization. The homology among different RNase IIIs varies from 20 to 84% placing many of them into the twilight zone [117]. Additionally, the number and complexity of both ribonuclease and double stranded dsRNA binding domains in the RNase III architecture are also variable. In fact, this protein class is subdivided into four subclasses represented by four archetypes (bacterial RNase III, fungal RNase III, Dicer and Drosha) and their variable structural features also provide diversity in their biological role in the cell. Members from RNase III class have been characterized, specifically those involved in host defense promoting the release of cationic proteins from eosinophilic leukocytes [118]. Dicer and Drosha are responsible for the generation of short interfering RNAs (siRNAs) from long double-stranded RNAs during RNA interference (RNAi). Therefore, they are involved in several important biological processes with high biological and molecular diversity [119]. For instance, the function of Dicer on the vascular system regulates the embryonic angiogenesis probably by processing micro RNAs (miRNA), which regulate the expression levels of some critical angiogenic regulators in the cell [120]. Drosha activity is related to the processing of small nuclear RNAs sharing common features with the biogenesis of naturally occurring miRNA. Such miRNAs are likely

to be involved in most biological processes by affecting gene regulation. These facts show the relevance of RNase III as an interesting source to search drug or drug-target candidates for drug development [121].

Spectral moments ($^{HP}\mu_k$), derived from the 20 amino acids clustering into a 2D-HP Cartesian space, were used to develop three different non-linear approaches to detect RNase III protein signatures against a structurally non-redundant subset of the PDB. The electrostatic charge of each amino acid was used for the calculation of the spectral moments series to build alignment-free models. Two alignment-free models were obtained, Decision Tree Models (DTMs) and Artificial Neural Networks (ANNs), to predict RNase III members. Profiles HMM were additionally used as an alignment-dependent algorithm to compare the performance of these two previous models in the classification of RNase III class members. A non-classical profile HMM, inspired on the graphical clustering of the amino acids was built, to make a fair comparison among classification models. Such profile HMM consisted in reducing the amino acid alphabet, clustering them according to their electric charges.

Nonlinear classification models e.g. ANNs, Support Vector Machines (SVMs) and DTMs are more complex functions than linear discriminant based models. The relationship between predictors and the response variable in nonlinear functions is hard to interpret, contrary to the ability of linear models to easily measure the effect of predictors over the response variable. However, machine learning methods that use nonlinear functions like ANNs and Support Vector Machine (SVM) have been more frequently applied to the prediction of proteins structure and function [122-125] together with traditional alignment algorithms and probabilistic functions such as profiles HMM and Bayesian networks [49, 108, 124, 126]. While nonlinear models have been applied for several purposes in bioinformatics, DTMs have been poorly explored to annotate biological function of proteins despite their widespread use in other fields [127]. We reported for first time a simple and interpretable DTM to annotate the function of RNase III members using spectral moments as input predictors. The reported DTM showed a high predictive power (96.07%) using just one spectral moment at different splitting values while ANNs provide a lower predictability (92.15%). As mentioned before, ANNs were also evaluated as non-linear method for RNase III classification and its predictability was similar to the DTM but using a more complex function (see results in annex 2).

The non-classical profile HMM showed the best performance in the classification of proteins involved in this study. It reached the highest prediction rate (100%) for the RNase III

class in respect to ANN and DTM performance. Amino acid clustering according to its hydrophobic/charge properties was either effective at primary level to increase the sensitivity of the profile HMM to retrieve all RNase III members or at 2D level to develop high predictive alignment-free models like DTM.

The strategy of developing non-classical profiles HMM could be applied to other highly diverse protein families to retrieve remote homologous. In addition, it confirmed that amino acid clustering according to its physicochemical features into 2D protein maps bear a biochemical sense to annotate biological functions. Though, the non-classical profile HMM showed a slightly better performance than the alignment-free models, their generation demands programming skills while DTM search resulted the easiest way to detect the RNase III signature among the diversity of the dataset.

We have also validated the simplicity of DTM by predicting a new bacterial RNase III class member that was isolated and subsequently enzymatically tested and registered by our group (see annex 2). The efficiency of DTM as a sequence search procedure to screen a proteome in conjunction with TIs implemented in the **TI2BioP** software has been emphasized in annex 4 [21].

We have so far evaluated the TIs generated by **TI2BioP** as alignment-free classifiers in protein families. They should also be assessed in a gene family to prove their ability to characterize DNA sequences and for reconstructing phylogenies. For such purposes and with the aim to compare different graphical approaches, we used the original 2D Cartesian representation reported by Nandy for describing DNA sequences [86] and the secondary structure inferred by DNA folding algorithms (Mfold) [88], to derive two types of TIs for the ITS2 gene class (see annex 3).

The ITS2 eukaryotic gene class shows a high sequence divergence among its members, which have traditionally complicated ITS2 annotation and limited its use for phylogenetic inference at low taxonomical level analyses (genus and species level classifications). Despite its high sequence variability, the ITS2 secondary structure has been considerably conserved among all eukaryotes [128]. This fact was considered in the implementation of homology-based structure modelling approaches to improve the ITS2 annotation quality and to carry out phylogenetic analyses at higher classification levels or taxonomic ranks for eukaryotes [51, 82, 128]. Thus, the ITS2 database (http://its2.bioapps.biozentrum.uni-wuerzburg.de) was developed holding information about sequence, structure and taxonomic

classification of all ITS2 in GenBank [129]. Due to ITS2 sequence diversity, the annotation pipeline implemented in the aforementioned resource required the use of a specific score matrix in the BLAST search [129] and more recently, the use of profiles HMM based on conserved flanking regions (5.8S/28S rRNA) for the identification and delineation of ITS2 sequences [82, 130]. Although alignment based methods have been exploited to the top of its complexity to tackle the ITS2 annotation and phylogenetic inference [82, 129], no alignment-free approach has so far been able to successfully address these issues.

The use of the spectral moments series containing information about the sequence and structure of ITS2 was useful for ITS2 prediction and for phylogenetic reconstruction at high taxonomic levels in eukaryotes. Such TIs were derived from two types of DNA graphs by our **TI2BioP** methodology. The 2D-Cartesian representation for DNA sequences reported by Nandy and the 2D-DNA structure inferred by the Mfold algorithm were used to obtain the spectral moments from the edge adjacency matrix of both graph types (annex 3). Both 2D-DNA representations were previously used to derive stochastic molecular descriptors to develop alignment-free models for ribosomal and ACC oxidase RNA classes from *Psidium guajava*, respectively [131, 132]. These models obtained from graphical/numerical approaches represented the first ones to classify RNA sequences without alignments. However, they were evaluated in RNA classes having high conservation degree and using small sized datasets. Such reports and others related to protein classes were the background of our current work.

Linear and ANN-based models were developed as alignment-independent models using two types of spectral moments calculated by **TI2BioP**; one type derived from the 2D-Cartesian DNA representation [86, 133] and the other resulting from the Mfold 2D structure [88]. They both were used to classify ITS2 members among large datasets (4 355 ITS2 and 14 657 UTRs) and Mfold TIs were also applied to estimate phylogeny at higher taxonomic levels than genus and species in fungi. ANN-models provided a better performance to classify ITS2 genomic sequences than linear models in training and test sets for Nandy-like and Mfold structures, respectively. These results supported that the identification of gene signatures tend to be better when assessed with non-linear models. Although ANN-models built with TIs derived from Nandy-like and Mfold structures displayed an excellent performance to detect the ITS2 class; the Mfold graphical approach provided the best classification results. Mfold TIs contain structural information about DNA folding driven by thermodynamic rules, providing a more accurate description of the DNA/RNA structure. This

is the reason behind the application of Mfold TIs as an alignment-free approach to infer phylogenetic relationship to complement the taxonomy of a fungal isolate (see annex 3).

The Nandy-like representation is less accurate in the classification process due to its artificial nature even though it carries the sequence order information and the nucleotide composition, which are important features for the recognition at a genome scale of genes that do not encode for a protein [134, 135]. Thus, the utility of this easy structural approach is evidenced when the correct 2D structure is not available (i.e. the physiological structure that occurs on the cell) and can only be obtained by predictions based on free energy minimizations.

The performance of ANN-models derived from Nandy-like and Mfold structures were compared with several profiles HMM generated from MSA performed with CLUSTALW [58], DIALIGN-TX [60] and MAFFT [106] using different training sets, to classify the test set and to identify a new fungal member of the ITS2 class. Due to the low similarity level amongst the ITS2 sequences, we used DALIGN-TX and MAFFT that are expected to outperform CLUSTALW in such conditions. Performing a good alignment is a crucial step to generate a profile HMM with high classification power. The performance of alignment-free models was higher than the obtained by profiles HMM to classify the test set and to identify a new fungal member of the ITS2 class, even when they were built by MSA algorithms improved for sets of low overall sequence (annex 3, table 3). This new ITS2 sequence was isolated by our group (GenBank accession number FJ892749) from an endophytic fungus belonging to the genus *Petrakia.* Members of this fungal genus have been hard to be placed taxonomically and are potential producers of bioactive compounds [136]. We classified our fungal isolate as a mitosporic Ascomycota/*Petrakia sp.* according to its mycological culture features, as there is not a report with a detailed taxonomy about this genus, namely in the NCBI dedicated „Taxonomy" database. There is not specification about its subphylum and class [137] and the lack of other ITS2 sequences from different species of the genus *Petrakia* (with the exception of our sequence submission at the GenBank) precluded performing a phylogenetic analysis at the species level (low-level analysis). So, a higher-level phylogenetic study involving the Ascomycota phylum members sharing ITS2 sequence similarities with *Petrakia* may provide more details about its taxonomy into this phylum. We assumed that our fungal isolate belonged to the *Pezizomycotina* subphylum, the largest within Ascomycota phylum, according to a recent classification found in the "The dictionary of the Fungi" [138]. Consequently a higher-level phylogenetic analysis to elucidate the class

where *Petrakia sp.* belongs to was carried out using two different types of distance trees: (1) a traditional one based on multiple alignments of ITS2 sequences and (2) another irrespective of sequence similarity supported by Mfold TIs from the **TI2BioP** methodology. The alignment-free distances calculated from Mfold TIs provided similar phylogenetic relationships among the different classes of the Ascomycota phylum in respect to the traditional phylogenetic analysis (i.e. based on evolutionary distances derived from a multiple alignment of DNA sequences). Both phylogenetic analyses, the traditional and the alignment-free clustering, placed *Petrakia* isolate into the Dothideomycetes class. We concluded that our alignment-free approach is effective to construct hierarchical distance-trees containing relevant biological information with an evolutionary significance (see annex 3).

Up to now we have only used 2D Cartesian graphs to derive the spectral moments series to describe several gene/protein classes with high divergence degree between its members. Although these 2D graphs were useful to unravel functional relationships for distant homologous of these families, either by direct visualization or by development of classification models, there are other 2D graphical approaches reported for DNA and proteins that have been mostly unexplored in bioinformatics; such is the case of the four-color maps introduced by Randić [87, 92]. We implemented protein four-color maps in **TI2BioP** *version 2.0* to provide TIs as alignment-free predictors, which can cooperate with traditional homology search tools (e.g. BLAST, HMMs) to carry out an exhaustive exploration of functional signatures in highly diverse gene/protein families. These 2D graphs are inspired on Randić"s DNA/RNA and protein four-color maps, but with some modifications to speed up graph building and facilitate the calculation of spectral moments as TIs (see annex 4) [21].

A deep exploration of functional signatures in highly diverse gene/protein families should reveal the presence of remote homologous. Remote homologues are divergent gene/protein sequences that have conserved the same biological function in different organisms. They can be harvest in the alignment algorithms twilight zone (<30% of amino acid identity) and have been traditionally detected by the use of more sensitive alignment-based methods like PSI-BLAST [47] an profiles HMM [46]. Also complex alignment strategies have been adopted such as the ensemble of homology-search methods to overcome poor sequence similarity in gene/protein classes [50, 82].

The NRPS family can harbor remote homologous due to the high sequence divergence among its A-domains, ranging mostly from 10-40% of sequence identity. Consequently,

many of them are placed in the twilight zone (20-35% sequence identity) reported for the alignment methods [139-141] (figure 2, annex 4). In fact, A-domain members cannot be retrieved easily by BLASTp using a single template [52]. NRPSs are megasynthetases composed by several domains organized in clusters for the synthesis of oligopeptides with biological activities including natural drugs. A-domains are a mandatory component for each NRPS cluster, being responsible for the amino acid selection and its covalent fixation on the phospho-pantethein arm as thioester, through AMP-derivative intermediates during the production of oligopeptides via non-ribosomal biosynthesis [104]. To cope with the high sequence divergence of A-domains, we propose an ensemble of homology-search methods that integrates an alignment-free model that uses TIs derived from protein four-color maps [21].

Randić *et al.* introduced four-color maps to visualize similarities/dissimilarities between DNA sequences but also characterized numerically such DNA maps to gain a deeper insight into their similarities/dissimilarities when the number of differences among DNA sequences increased (possibly along with the length of the sequences). DNA four-color maps showed a high sensitivity to capture nucleotide changes but their numerical characterization was complicated, involving the definition of a distance matrix subdivided into 10 submatrices and finally providing a 10-dimensional vector for each DNA sequence [92]. Later, this author extended the same representation to inspect similarities/dissimilarities among protein sequences but turning protein alphabet into DNA sequences using the virtual genetic code concept. Their numerical characterizations were similar to those reported for DNA sequences; arriving to a 10-component vector but containing more structural information [87]. Although four-color maps were introduced with success to describe visually/numerically DNA and protein sequences, its application was illustrated for single sequences; the coding sequence of the exon 1 of the human b-globin gene and the A chain of human insulin, respectively. Their limited application in bioinformatics may be due to its complicated numerical characterization, especially for proteins demanding the use of a virtual genetic code and a 10-dimensional vector. We overcame this limitation by drawing the spiral of square cells straightforward from amino acid sequences avoiding the use of the virtual genetic code. The four colors are assigned to the four amino acids classes (polar, non-polar, acid and basic) used previously by our group in Nandy"s representation for proteins [38, 90]. Grouping the amino acids into four classes reduced the number of regions in the graph, making it simpler than if we would group them according to their 20 natural types. This same concept was used for the 2D-Cartesian representation for proteins discussed above. Both

protein representations (i.e. 2D-Cartesian and four-color maps) are similar to Nandy and Randić representations for DNA sequences but showing a major degeneracy degree since different amino acids can be placed in the same node or region of the graph, respectively. Such degeneracy produced by the amino acid clustering into four classes is useful to describe homologous sequences (replacement made with amino acids of similar properties) and remote homologous (important changes in the primary structure but still retaining the same biological function). While small changes in the sequence do not affect the topology of the map, this kind of amino acid substitution produces implicit numerical changes in the calculation of TIs making sequences differentiation possible. When an amino acid exchange occurs between different physicochemical groups of amino acids, this change affects the topology of the map and consequently affects significantly the TIs values estimation. The topology of the 2D-Cartesian and protein four-color maps is determined by the sequence order and its amino acid composition (amino acid content according to the four groups mentioned above) and its essence is numerically captured by their respective spectral moments (annex 4) [21].

Under this scenario we evaluated in annex 4, the potential of the spectral moments derived from protein four-color maps to generate high predictive alignment-free models to detect the A-domain NRPS signature among a benchmark dataset. We used a dataset made up of CATH domains sharing less than 30% of sequence identity and cleaned from any A-domain signal and A-domains from experimentally characterized NRPS clusters that share 10-40% of sequence identity. The spectral moments series were used to develop several alignment-free models using linear and nonlinear statistical techniques. Nonlinear models showed a better performance in classifying A-domains in respect to linear models, supporting previous results. DTM was selected among the nonlinear models due to its excellent performance and its simple way to detect A-domains in a highly diverse dataset. These results agreed with those obtained for the RNase III class, where a DTM showed the best classification performance among all alignment-free models. Additionally, the DTM based on our graphical/numerical method was contrasted to other different alignment-free approaches and homology-search methods in detecting A-domains on the same dataset. The Webserver PseAAC (http://www.csbio.sjtu.edu.cn/bioinf/PseAAC/) was used to generate alignment-free DTM based on amino acid composition (AAC) and pseudo amino acid composition (PseAAC) [142]. On the other hand, homology-based methods for A-domains detection were performed by single-template BLASTp, multi-template BLASTp and

profile HMM. These alignment-based methods show by definition different sensitivity to recognize distant homologs.

The DTM generated by four-color maps outperformed the DTM supported by AAC and PseAAC (annex 4, table 4). Although A-domains share 10-40% of sequence identity with several members placed in the twilight zone, it was possible to retrieve all of them using four-color maps. The other two left alignment-free models (AAC and PseAAC) showed lower sensitivity but they did not provide many false positives (annex 4, table 5). It was also demonstrated the effect of the sequence order besides the AAC on the prediction quality; when the PseAAC concept was applied, there was an improvement in all standard classification measures (annex 4, table 4). Regarding homology-based methods sensitivity, classification results agreed with the fact that multi-template BLASTp and profile HMM are more sensitive than simple BLASTp. Both multi-template BLASTp and profile HMM easily retrieved all A-domain members at expectation values (E-value≤10) without reporting any false positive (annex 4, table 5)**.** However, the BLASTp search using a single template provided false positives (significant matches) among CATH domains at both high (E-value=10) and relatively stringent cut-offs (E-values<0.05), which is considered statistically significant and useful for filtering easily identifiable homologs pairs [43, 143] (annex 4, table 5). False positives came up in simple BLASTp searches despite we had cleaned the negative set (CATH domains) from any A-domain signal (by the use of profile HMM-based searches). In contrast to multi-template BLASTp and profile HMM searches, the single-BLASTp search sensitivity did not show stability in identifying the A-domain signal among a benchmark dataset (CATH domains) when the classification parameter (E-value cut-off) was changed. Thus, due to the A-domain diversity, it is less reliable to extrapolate or apply BLASTp searches using a single A-domain template to an unknown test dataset such as an entire proteome. Therefore, the easy and reliable identification of A-domains by multi-template BLASTp, profile HMM and four-color maps have been combined to explore the A-domain repertoire in the proteome of the cyanobacteria *Microcystis aeruginosa* NIES-843 [21].

The cyanobacteria *Microcystis aeruginosa* contains NRPS proteins as hybrids with polyketide synthases (PKS) being its proteome a good test set to explore the complete A-domain repertoire leading to the detection of new A-domain variants. We carried out for first time an alignment-free search for the A-domain signature in combination to homology-search methods against the *Microcystis aeruginosa* proteome. Thus, multiple-template

BLASTp and profile HMM searches for A-domains were also performed to deeply explore the repertoire of this protein signature in such proteome. The knowledge of the complete repertoire of A-domains in the proteome of cyanobacteria species may allow unraveling new NRPS clusters for the discovery of novel natural products with important biological activities. Interestingly, DTM detected two putative A-domain signatures among proteome's hypothetical proteins while another three hypothetical proteins were detected as A-domains by the profile HMM (annex 4, figure 4). Sequence search methods based on profiles (graphical and alignment) were able to detect more hits than the 20 A-domains already annotated in the proteome, which were confirmed by the multi-template BLASTp. Hypothetical proteins are greatly expanded in cyanobacteria and have been placed into the diversity of the nuclease superfamily by homology inference. Probably the graphical and HMM profiles detected signals of the A-domain signature among the highly diverse hypothetical proteins, leading us to the discovery of new A-domains variants.

Both methods detected different additional hits as A-domains but they were found among the hypothetical proteins, which is a good clue for the presence of A-domains remote homologues in the proteome of *Microcystis aeruginosa*. The use of an ensemble of sequence search methods provides a more exhaustive description of certain protein class since each method extracts different features from protein sequences; their integration provides a higher yield for the detection of remote protein homologous with more confidence. This is the first report where a graphical-based method worked well in cooperation with alignment procedures to search for remote homologous, a whole challenge for current bioinformatics [21].

The main goal of this thesis is to provide evidence of the potential use of graphical/numerical approaches to characterize DNA/RNA and proteins. This new tool is not in competition with currently available “tools” such as BLAST, profile HMM, FASTA and other computer software, but instead in cooperation with existing methodologies, as well as with experimentation procedures required to overcome hard comparative studies of DNA, RNA, proteins, and even proteomes [2].

# CONCLUSIONS

# 5. Conclusions

1. The Cartesian representation and the four-color maps reported for DNA/RNA sequences were extended to characterize graphically protein sequences. These graphical approaches allowed the successful application of the spectral moments to encode numerically relevant structural information of the natural biopolymers (DNA/RNA and proteins). Such graphical/numerical extension was implemented in the **TI2BioP** software.
2. Spectral moments were useful to develop reliable alignment-free models to detect functional signatures from highly diverse gene/protein classes. A new bacteriocin, ITS2 and RNase III member were detected by alignment-free models in cooperation either to experimental evidences or to alignment procedures.
3. Alignment-independent models outperformed alignment algorithms in detecting accurately all members of the ITS2 and bacteriocin class. They showed a similar performance for other classes (RNase III and NRPS A-domain) in respect to improved alignment strategies.
4. Remote homologous were detected for the bacteriocin protein class and for NRPS A-domains in cooperation with experimental evidences and advanced alignment procedures, respectively.
5. TIs calculated from graphical approaches were introduced for the first time to construct hierarchical distance-trees with similar topology to classic distance-trees. This fact gives clues about the relevant biological information that bear such TIs and its potential use for molecular evolution.

## 5.1. Future Directions

1. Further characterization of the hypothetical proteins that were detected with the A-domain NRPS signature in the proteome of *Microcystis aeruginosa* by assembling sequence-search methods. To extend such methodology to other proteomes with the aim to spot new NRPS clusters.
2. The implementation of other 2D representation for DNA and proteins in **TI2BioP** software for structural and functional classification purposes.
3. To assess the suitability of our TIs derived from the 2D graphs in estimating phylogenetic distances by using a benchmark dataset (simulated sequences).

4. To improve our evolutionary approach providing alignment-free distances defined within an evolutionary framework.

# REFERENCES

# 6. References

1. Biggs, N., E. Lloyd, and R. Wilson, eds. *Graph Theory*. 1986, Oxford University Press. 1736–1936.

2. Randic, M., *et al.*, *Graphical representation of proteins.* Chem Rev, 2011. **111**(2): p. 790-862.

3. Gonzalez-Diaz, H., *et al*., *Medicinal chemistry and bioinformatics--current trends in drugs discovery with networks topological indices.* Curr Top Med Chem, 2007. **7**(10): p. 1015-29.

4. Estrada, E. and E. Uriarte, *Recent advances on the role of topological indices in drug discovery research.* Curr Med Chem, 2001. **8**(13): p. 1573-88.

5. Gonzalez-Diaz, H., *et al.*, *Proteomics, networks and connectivity indices.* Proteomics, 2008. **8**(4): p. 750-78.

6. Estrada, E. and I. Gutman, *A Topological Index Based on Distances of Edges of Molecular Graphs.* J Chem Inf Comput Sci, 1996. **36**: p. 850-853.

7. Riera-Fernandez, P., *et al*., *From QSAR models of drugs to complex networks: state-of-art review and introduction of new Markov-spectral moments indices.* Curr Top Med Chem, 2012. **12**(8): p. 927-60.

8. Aguero-Chapin, G., *et al.*, ***TI2BioP****: Topological Indices to BioPolymers. Its practical use to unravel cryptic bacteriocin-like domains.* Amino Acids, 2011. **40**(2): p. 431-42.

9. Perez-Bello, A., *et al.*, *Alignment-free prediction of mycobacterial DNA promoters based on pseudo-folding lattice network or star-graph topological indices.* J Theor Biol, 2009. **256**(3): p. 458-66.

10. Ortega-Broche, S.E., *et al.*, *TOMOCOMD-CAMPS and protein bilinear indices--novel bio-macromolecular descriptors for protein research: I. Predicting protein stability effects of a complete set of alanine substitutions in the Arc repressor.* FEBS J, 2010. **277**(15): p. 3118-46.

11. Gonzalez-Diaz, H., *et al.*, *Generalized lattice graphs for 2D-visualization of biological information.* J Theor Biol, 2009. **261**(1): p. 136-47.

12. Concu, R., *et al.*, *Computational chemistry study of 3D-structure-function relationships for enzymes based on Markov models for protein electrostatic, HINT, and van der Waals potentials.* J Comput Chem, 2009. **30**(9): p. 1510-20.

13. Marrero-Ponce, Y., *et al.*, *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, 2005. **13**(8): p. 3003-15.

14. Randic´, M., *Graphical representation of DNA as a 2-D map.* Chem. Phys. Lett., 2004(386): p. 468–471.

15. Randic, M., J. Zupan, and D. Vikic-Topic, *On representation of proteins by star-like graphs.* J Mol Graph Model, 2007. **26**(1): p. 290-305.

16. Randic, M. and J. Zupan, *Highly compact 2D graphical representation of DNA sequences.* SAR QSAR Environ Res, 2004. **15**(3): p. 191-205.

17. Nandy, A., *Recent investigations into global characteristics of long DNA sequences.* Indian J Biochem Biophys, 1994. **31**(3): p. 149-55.

18. Aguero-Chapin, G., *et al*., *Alignment-free prediction of polygalacturonases with pseudofolding topological indices: experimental isolation from Coffea arabica and prediction of a new sequence.* J Proteome Res, 2009. **8**(4): p. 2122-8.

19. Aguero-Chapin, G., *et al.*, *An alignment-free approach for eukaryotic ITS2 annotation and phylogenetic inference.* PLoS ONE 2011. **6**(10).

20. Cruz-Monteagudo, M., *et al.*, *3D-MEDNEs: an alternative "in silico" technique for chemical research in toxicology. 2. quantitative proteome-toxicity relationships (QPTR) based on mass spectrum spiral entropy.* Chem Res Toxicol, 2008. **21**(3): p. 619-32.

21. Aguero-Chapin, G., *et al.*, *Exploring the Adenylation Domain Repertoire of Nonribosomal Peptide Synthetases Using an Ensemble of Sequence-Search Methods.* PLoS One, 2013. **8**(7).

22. Katritzky, A.R. and E.V. Gordeeva, *Traditional topological indices vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research.* J Chem Inf Comput Sci, 1993. **33**(6): p. 835-57.

23. Randic, M. and J. Zupan, *On interpretation of well-known topological indices.* J Chem Inf Comput Sci, 2001. **41**(3): p. 550-60.

24. Estrada, E., *Edge Adjacency Relationships and a Novel Topological Index Related to Molecular Volume.* J Chem Inf Comput Sci, 1995. **35**: p. 31-33.

25. Wiener, H., *Structural Determination of Paraffin Boiling Points.* J. Am. Chem. Soc, 1947. **69**: p. 17-20.

26. Randic, M., *Graph theoretical approach to structure-activity studies: search for optimal antitumor compounds.* Prog Clin Biol Res, 1985. **172A**: p. 309-18.

27. Balaban, A.T., *et al.*, *Four new topological indices based on the molecular path code.* J Chem Inf Model, 2007. **47**(3): p. 716-31.

28. Moreau, G. and P. Broto, *The Autocorrelation of a topological structure. A new molecular descriptor.* Nouv J Chim, 1980. **4**: p. 359-360.

29. Estrada, E., *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, 1996. **36**: p. 844-849.

30. Estrada, E., *Spectral Moments of the Edge-Adjacency Matrix of Molecular Graphs. 2. Molecules Containing Heteroatoms and QSAR Applications.* J Chem Inf Comput Sci, 1997. **37**: p. 320-328.

31. Estrada, E., *On the topological sub-structural molecular design (TOSS-MODE) in QSPR/QSAR and drug design research.* SAR QSAR Environ Res, 2000. **11**(1): p. 55-73.

32. Markovic, S., Z. Markovic, and R.I. McCrindle, *Spectral moments of phenylenes.* J Chem Inf Comput Sci, 2001. **41**(1): p. 112-9.

33. González, M.P., C. Teran, and M. Teijeira, *A topological function based on spectral moments for predicting affinity toward $A_3$ adenosine receptors.* Bioorg Med Chem Lett, 2006. **16**(5): p. 1291-6.

34. Morales, A.H., M.P. González, and J.R. Briones, *TOPS-MODE approach to predict mutagenicity in dental monomers.* Polymer, 2004. **45**: p. 2045-2050.

35. Estrada, E. and N. Hatano, *A Tight-Binding "Dihedral Orbitals" Approach to Electronic Communicability in Protein Chains.* Chemical Physics Letters, 2007. **449**: p. 216-220.

36. Estrada, E., *A Tight-Binding "Dihedral Orbitals" Approach to the Degree of Folding of Macromolecular Chains.* Journal of Physical Chemistry B, 2007. **111**: p. 13611-13618.

37. Estrada, E., *Characterization of the folding degree of proteins.* Bioinformatics, 2002. **18**(5): p. 697-704.

38. Aguero-Chapin, G., *et al.*, *Non-linear models based on simple topological indices to identify RNase III protein members.* J Theor Biol, 2011. **273**(1): p. 167-178.

39. Needleman, S.B. and C.D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins.* J Mol Biol, 1970. **48**(3): p. 443-53.

40. Smith, T.F. and M.S. Waterman, *Identification of common molecular subsequences.* J Mol Biol, 1981. **147**(1): p. 195-7.

41. Gotoh, O., *An improved algorithm for matching biological sequences.* J Mol Biol, 1982. **162**(3): p. 705-8.

42. Pearson, W.R. and D.J. Lipman, *Improved tools for biological sequence comparison.* Proc Natl Acad Sci U S A, 1988. **85**(8): p. 2444-8.

43. Altschul, S.F., *et al.*, *Basic Local Alignment Search Tool.* J. Mol. Biol., 1990. **215**: p. 403-410.

44. Dayhoff, M.O., R.M. Schwartz, and B.C. Orcutt, *A model of evolutionary change in proteins*, in *Atlas of Protein Sequence and Structure*, M.O. Dayhoff, Editor 1978, Nat. Biomed. Res. Found: Washington, DC. p. 345 - 352.

45. Krogh, A.B., M.; Mian, I. S.; Sjeander, K.; Haussler, D., *Hidden Markov models in computational biology. Applications to protein modeling.* J Mol Biol, 1994. **235**(5): p. 1501-31.

46. Eddy, S.R., *A new generation of homology search tools based on probabilistic inference.* Genome Inform, 2009. **23**(1): p. 205-11.

47. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J, *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucl. Acids Res, 1997. **25**: p. 3389-3402.

48. Park, J., *et al.*., *Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods.* J Mol Biol, 1998. **284**: p. 1201-1210.

49. Finn, R.D., *et al.*, *The Pfam protein families database.* Nucleic Acids Res, 2009.

50. de Jong, A., *et al.*, *BAGEL: a web-based bacteriocin genome mining tool.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W273-9.

51. Selig, C., *et al.*, *The ITS2 Database II: homology modelling RNA structure for molecular systematics.* Nucleic Acids Res, 2008. **36**(Database issue): p. D377-80.

52. Ansari, M.Z., *et al*., *NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W405-13.

53. Griffiths-Jones, S., *et al*., *Rfam: an RNA family database.* Nucleic Acids Res, 2003. **31**(1): p. 439-41.

54. Katoh, K., *et al.*, *MAFFT version 5: improvement in accuracy of multiple sequence alignment.* Nucleic Acids Res, 2005. **33**(2): p. 511-8.

55. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 2004. **32**(5): p. 1792-7.

56. Notredame, C., D.G. Higgins, and J. Heringa, *T-Coffee: a novel method for fast and accurate multiple sequence alignment.* J. Mol.Biol, 2000. **302**: p. 205-217.

57. Schwarz, R.F., *et al.*, *Evolutionary Distances in the Twilight Zone—A Rational Kernel Approach.* PLoS ONE, 2010. **5**(12).

58. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucleic Acids Res, 1994. **22**(22): p. 4673-80.

59. Katoh, K., *et al*., *MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.* Nucleic Acids Res, 2002. **30**(14): p. 3059-66.

60. Subramanian, A.R., M. Kaufmann, and B. Morgenstern, *DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment.* Algorithms Mol Biol, 2008. **3**: p. 6.

61. Dobson, P.D. and A.J. Doig, *Distinguishing Enzyme Structures from Non-enzymes Without Alignments.* J. Mol. Biol., 2003. **330**: p. 771–783.

62. Pearson, W.R. and M.L. Sierk, *The limits of protein sequence comparison?* Current Opinion in Strctural Biology, 2005. **15**: p. 254-260.

63. Strope, P.K. and E.N. Moriyama, *Simple alignment-free methods for protein classification: a case study from G-protein-coupled receptors.* Genomics, 2007. **89**(5): p. 602-12.

64. Deshmukh, S., *et al.*, *An alignment-free method for classification of protein sequences.* Protein Pept Lett, 2007. **14**(7): p. 647-57.

65. Kumar, M., V. Thakur, and G.P. Raghava, *COPid: composition based protein identification.* In Silico Biol, 2008. **8**(2): p. 121-8.

66. Chou, K.C., *Prediction of protein cellular attributes using pseudo-amino acid composition.* Proteins, 2001. **43**(3): p. 246-55.

67. Chou, K.C. and Y.D. Cai, *Prediction of protein subcellular locations by GO-FunD-PseAA predictor.* Biochem Biophys Res Commun, 2004. **320**(4): p. 1236-9.

68. Chou, K.C. and Y.D. Cai, *Predicting enzyme family class in a hybridization space.* Protein Sci, 2004. **13**(11): p. 2857-63.

69. Cai, Y.D. and K.C. Chou, *Predicting membrane protein type by functional domain composition and pseudo-amino acid composition.* J Theor Biol, 2005.

70. Chou, K.C. and Y.D. Cai, *Predicting protein quaternary structure by pseudo amino acid composition.* Proteins, 2003. **53**(2): p. 282-9.

71. Caballero, J., *et al.*, *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, 2006. **46**(3): p. 1255-68.

72. González-Díaz H, Molina-Ruiz R, and Hernandez I, *MARCH-INSIDE v3.0 (**MAR**kov **CH**ains **IN**variants for **SI**mulation & **DE**sign)*, 2007. p. Windows supported version under request to the main author contact email: gonzalezdiazh@yahoo.es.

73. Hohl, M. and M.A. Ragan, *Is multiple-sequence alignment required for accurate inference of phylogeny?* Syst Biol, 2007. **56**(2): p. 206-21.

74. Schwarz, R.F., *et al.*, *Evolutionary distances in the twilight zone--a rational kernel approach.* PLoS One, 2010. **5**(12): p. e15788.

75. Tomii, K. and M. Kanehisa, *Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins.* Protein Eng, 1996. **9**(1): p. 27-36.

76. Hohl, M., I. Rigoutsos, and M.A. Ragan, *Pattern-based phylogenetic distance estimation and tree reconstruction.* Evol Bioinform Online, 2006. **2**: p. 359-75.

77. Nilsson, R.H., *et al.*, *Intraspecific ITS variability in the kingdom fungi as expressed in the international sequence databases and its implications for molecular species identification.* Evol Bioinform Online, 2008. **4**: p. 193-201.

78. Sieber, T.N., *Endophytic fungi in forest trees: are they mutualists?* Fungal Biology Reviews. **21**(2-3): p. 75-89.

79. Hammami, R., *et al.*, *BACTIBASE: a new web-accessible database for bacteriocin characterization.* BMC Microbiology 2007. **7**: p. 89

80. Pesole, G., *et al.*, *UTRdb and UTRsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs.* Nucleic Acids Res, 2000. **28**(1): p. 193-6.

81. Mc Farland, J.W. and D.J. Gans, *Cluster Significance Analysis. In Method and Principles in Medicinal Chemistry*. van Waterbeemd, H ed, ed. R. Manhnhold, L. Krogsgaard, and V. Timmerman. Vol. 2. 1995, Weinheim, Germany: VCH. 295-307.

82. Koetschan, C., *et al.*, *The ITS2 Database III--sequences and structures for phylogeny.* Nucleic Acids Res, 2009.

83. Macrae, I.J. and J.A. Doudna, *Ribonuclease revisited: structural insights into ribonuclease III family enzymes.* Curr Opin Struct Biol, 2007. **17**(1): p. 138-45.

84. Rost, B., *Enzyme function less conserved than anticipated.* J. Mol. Biol., 2002. **318**: p. 595-608.

85. Smith, T.F., M.S. Waterman, and W.M. Fitch, *Comparative biosequence metrics.* J Mol Evol, 1981. **18**(1): p. 38-46.

86. Nandy, A., *Two-dimensional graphical representation of DNA sequences and intron-exon discrimination in intron-rich sequences.* Comput Appl Biosci, 1996. **12**(1): p. 55-62.

87. Randic, M., *et al.*, *Graphical representation of proteins as four-color maps and their numerical characterization.* J Mol Graph Model, 2009. **27**(5): p. 637-41.

88. Mathews, D.H., *RNA secondary structure analysis using RNAstructure.* Curr Protoc Bioinformatics, 2006. **Chapter 12**: p. Unit 12 6.

89. Nandy, A., *A new graphical representation and analysis of DNA sequence structure. Methodology and application to globin genes. .* Curr. Sci., 1994. **66**: p. 309-313.

90. Aguero-Chapin, G., *et al.*, *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, 2006. **580**(3): p. 723-30.

91. Jacchieri, S.G., *Mining combinatorial data in protein sequences and structures.* Molecular Diversity, 2000(5): p. 145–152.

92. Randic, M., *et al.*, *Four-color map representation of DNA or RNA sequences and their numerical characterization.* Chemical Physics Letters 2005. **407**: p. 205-208.

93. Cornell, W.D., *et al.*, *A second generation force field for the simulation of proteins, nucleic acids, and organic molecules.* J. Am. Chem. Soc., 1995. **117**(19): p. 5179-5197.

94. Mathews, D.H., *Predicting a set of minimal free energy RNA secondary structures common to two sequences.* Bioinformatics, 2005. **21**(10): p. 2246-53.

95. Marrero-Ponce, Y., *et al.*, *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, 2005. **13**(4): p. 1005-20.

96. Marrero-Ponce, Y., *et al.*, *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, 2004. **12**(20): p. 5331-42.

97. Ponce, Y.M., *et al.*, *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, 2004. **12**(20): p. 5331-42.

98. Statsoft, *STATISTICA 8.0 (data analysis software system for windows)*, 2008.

99. The MathWorks, I., ed. *Neural network toolbox user´s guide for use with MATLAB*. 2004, The Mathworks Inc: Massachusetts.

100. Rivals, I. and L. Personnaz, *On cross validation for model selection.* Neural Comput, 1999. **11**(4): p. 863-70.

101. Quevillon, E., *et al.*, *InterProScan: protein domains identifier.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W116-20.

102. Hunter, S., *et al.*, *InterPro: the integrative protein signature database.* Nucleic Acids Res, 2009. **37**(Database issue): p. D211-5.

103. Ansari, M.Z., *et al.*, *In silico analysis of methyltransferase domains involved in biosynthesis of secondary metabolites.* BMC Bioinformatics, 2008. **9**: p. 454.

104. Jenke-Kodama, H. and E. Dittmann, *Bioinformatic perspectives on NRPS/PKS megasynthases: advances and challenges.* Nat Prod Rep, 2009. **26**(7): p. 874-83.

105. Vazquez-Padron, R.I., *et al.*, *Cryptic endotoxic nature of Bacillus thuringiensis Cry1Ab insecticidal crystal protein.* FEBS Lett, 2004. **570**(1-3): p. 30-6.

106. Katoh, K., *et al.*, *Improvement in the accuracy of multiple sequence alignment program MAFFT.* Genome Inform, 2005. **16**(1): p. 22-33.

107. Jain, N., *et al.*, *Biochemistry of fruit ripening of guava (Psidium guajava L.): compositional and enzymatic changes.* Plant Foods Hum Nutr, 2003. **58**(4): p. 309-15.

108. Finn, R.D., J. Clements, and S.R. Eddy, *HMMER web server: interactive sequence similarity searching.* Nucleic Acids Res, 2011. **39**(Web Server issue): p. W29-37.

109. Davies, M.N., *et al.*, *Alignment-Independent Techniques for Protein Classification.* Current Proteomics, 2008. **5**(4): p. 217-223.

110. Aguero-Chapin, G., *et al.*., *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, 2008. **48**(2): p. 434-48.

111. Lamontagne, B. and S.A. Elela, *Evaluation of the RNA determinants for bacterial and yeast RNase III binding and cleavage.* J Biol Chem, 2004. **279**(3): p. 2231-41.

112. Molina, R., G. Agüero-Chapin, and M.P. Pérez-González, ***TI2BioP*** *(Topological Indices to BioPolymers) version 2.0.*2011: Molecular Simulation and Drug Design (MSDD), Chemical Bioactives Center, Central University of Las Villas, Cuba.

113. Gillor, O., A. Etzion, and M.A. Riley, *The dual role of bacteriocins as anti- and probiotics.* Appl Microbiol Biotechnol, 2008. **81**(4): p. 591-606.

114. Cotter, P., C. Hill, and R. Ross, *What's in a name? Class distinction for bacteriocins.* Nature Reviews Microbiology, 2006. **4**(2).

115. Dirix, G., *et al*., *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, 2004. **25**(9): p. 1425-40.

116. Stein, T., *Bacillus subtilis antibiotics: structures, syntheses and specific functions.* Mol Microbiol, 2005. **56**(4): p. 845-57.

117. Lamontagne B and Elela S.A, *Evaluation of the RNA determinants for bacterial and yeast RNase III binding and cleavage.* J Biol Chem, 2004. **279**(3): p. 2231-41.

118. Dyer, K.D. and H.F. Rosenberg, *The RNase a superfamily: generation of diversity and innate host defense.* Mol Divers, 2006. **10**(4): p. 585-97.

119. Court, D., *RNA processing and degradation by RNase III. In: Control of mRNA stability*. Brawerman, G., Belasco, J. ed1993, New York: Academic Press. 70–116.

120. Lee, Y., *et al.*, *The nuclear RNase III Drosha initiates microRNA processing.* Nature, 2003. **425**: p. 415-419.

121. Gonzalez-Diaz, H., *et al.*, *QSAR for RNases and theoretic-experimental study of molecular diversity on peptide mass fingerprints of a new Leishmania infantum protein.* Mol Divers, 2010. **14**(2): p. 349-69.

122. Punta, M. and B. Rost, *Neural networks predict protein structure and function.* Methods Mol Biol, 2008. **458**: p. 203-30.

123. Nair, R. and B. Rost, *Protein subcellular localization prediction using artificial intelligence technology.* Methods Mol Biol, 2008. **484**: p. 435-63.

124. Fernandez, M., *et al.*, *Protein radial distribution function (P-RDF) and Bayesian-Regularized Genetic Neural Networks for modeling protein conformational stability: chymotrypsin inhibitor 2 mutants.* J Mol Graph Model, 2007. **26**(4): p. 748-59.

125. Cai, Y.D., *et al.*, *Application of SVM to predict membrane protein types.* J Theor Biol, 2004. **226**(4): p. 373-6.

126. Burden, F.R. and D.A. Winkler, *Predictive Bayesian neural network models of MHC class II peptide binding.* J Mol Graph Model, 2005. **23**(6): p. 481-9.

127. Ripley, B., *Pattern Recognition and Neural Networks*1996, Cambridge, UK: Cambridge University Press.

128. Schultz, J., *et al.*, *A common core of secondary structure of the internal transcribed spacer 2 (ITS2) throughout the Eukaryota.* RNA 2005. **11**: p. 361-364.

129. Schultz, J., *et al.*, *The internal transcribed spacer 2 database--a web server for (not only) low level phylogenetic analyses.* Nucleic Acids Research, 2006. **34**.

130. Keller, A., *et al.*, *5.8S-28S rRNA interaction and HMM-based ITS2 annotation.* Gene, 2009. **430**(1-2): p. 50-7.

131. Gonzalez-Diaz, H., *et al.*, *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, 2005. **15**(11): p. 2932-7.

132. Aguero-Chapin, G., *et al.*, *Comparative study of topological indices of macro/supramolecular RNA complex networks.* J Chem Inf Model, 2008. **48**(11): p. 2265-77.

133. Nandy, A., *Empirical relationship between intra-purine and intra-pyrimidine differences in conserved gene sequences.* PLoS One, 2009. **4**(8): p. e6829.

134. Schattner, P., *Searching for RNA genes using base-composition statistics.* Nucleic Acids Res, 2002. **30**(2): p. 2076-82.

135. Wong, T.K., *et al*., *Adjacent nucleotide dependence in ncRNA and order-1 SCFG for ncRNA identification.* PLoS One, 2010. **5**(9).

136. Qi, F.H., *et al.*, *Fungal endophytes from Acer ginnala Maxim: isolation, identification and their yield of gallic acid.* Lett Appl Microbiol, 2009. **49**(1): p. 98-104.

137. Bisby, F., *et al*., *Species 2000 & ITIS Catalogue of Life: 2007 Annual Checklist Taxonomic Classification. CD-ROM; Species 2000: Reading, U.K.* 2007.

138. Kirk, P.M., P.F. Cannon, and J.A. Stalpers, *The dictionary of the Fungi*. 10th ed, ed. Paul M Kirk, *et al.*.2008, UK: CABI. 784.

139. Rost, B., *Twilight zone of protein sequence alignments.* Protein Eng, 1999. **12**(2): p. 85-94.

140. Hobohm, U. and C. Sander, *A sequence property approach to searching protein databases.* J Mol Biol, 1995. **251**(3): p. 390-9.

141. Wass, M.N. and M.J. Sternberg, *ConFunc--functional annotation in the twilight zone.* Bioinformatics, 2008. **24**(6): p. 798-806.

142. Shen, H.B. and K.C. Chou, *PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition.* Anal Biochem, 2008. **373**(2): p. 386-8.

143. Boekhorst, J. and B. Snel, *Identification of homologs in insignificant blast hits by exploiting extrinsic gene properties.* BMC Bioinformatics, 2007. **8**(356).

# ANNEXES

# 7.Annexes

| No. | Ref. | |
|---|---|---|
| 1 | 8 | **Agüero-Chapin G**, Pérez-Machado G, Molina-Ruiz R, Pérez-Castillo Y, Morales-Helguera A, Vasconcelos V and Antunes A . **TI2BioP**: Topological Indices to BioPolymers. Its practical use to unravel cryptic bacteriocin-like domains. ***Amino Acids.*** **2011; 40(2):431-42.** |
| 2 | 38 | **Agüero-Chapin G,** de la Riva GA, Molina-Ruiz R, Sánchez-Rodríguez A, Pérez-Machado G, Vasconcelos V and Antunes A. Non-linear models based on simple topological índices to identify RNase III protein members. ***Journal of Theoretical Biology***. **2011; 273(1):167-78**. |
| 3 | 19 | **Agüero-Chapin G,** Sánchez-Rodríguez A, Hidalgo-Yanes PI, Pérez-Castillo Y, Molina-Ruiz R, Marchal K, Vasconcelos V and Antunes A. An alignment-free approach for eukaryotic ITS2 annotation and phylogenetic inference. ***PLoS ONE*** **2011; 6(10).** |
| 4 | 21 | **Agüero-Chapin G,** Molina-Ruiz R, Maldonado E, de la Riva GA, Vasconcelos V and Antunes A. Exploring the adenylation domain repertoire of nonribosomal peptide synthetases using an ensemble of sequence-search methods. ***PLOS ONE*** **2013*;* 8(7).** |

# ANNEX 1

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

**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

## Introduction

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

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).

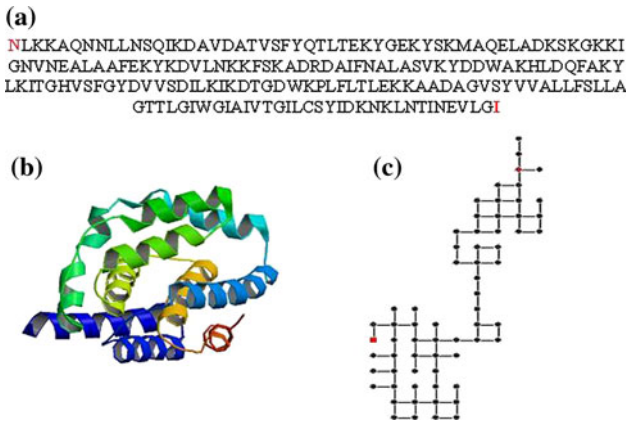



**Fig. 1** Three structures for the colicin E1domain sequence. **a** Primary structure **b** three-dimensional structure **c** the pseudo-secondary Cartesian structure of hydropobicity and polarity

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 ($\mu_k$) 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 ($^{HP}\mu_k$) values as TIs describing these proteins maps. The $^{HP}\mu_k$ were selected based on the utility of $\mu_k$ 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 ($^{HP}\pi_k$) 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.

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

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

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

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

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

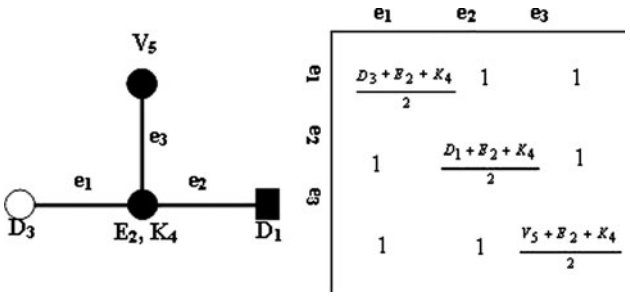



**Fig. 2** The 2D-HP map for the protein fragment DEDKV, aside the definition of its bond adjacency matrix. Note that all edges of the graph are adjacent, thus all non-diagonal entries are ones

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

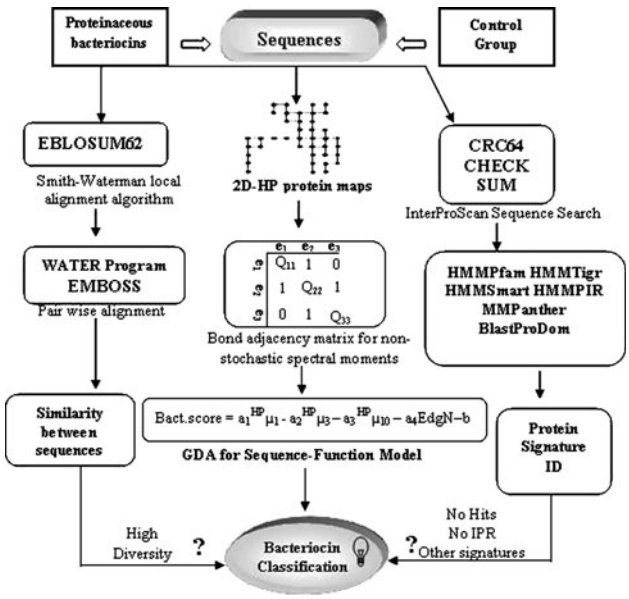



**Fig. 3** The overall procedure followed for the classification of bacteriocins

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.

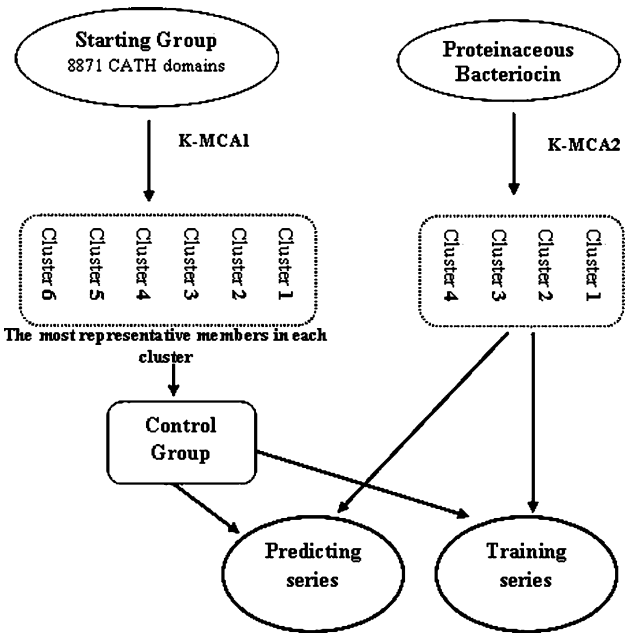



**Fig. 4** Scheme describing the design of training and predicting series using k-MCA for both bacteriocins and control group

### 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 ($\lambda$) 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 ($0 \leq$ score $\leq 1$). The best classification function equation found for the bacteriocin group after GDA analyses was:

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

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 ($\lambda$) 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% | $\lambda$ | 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

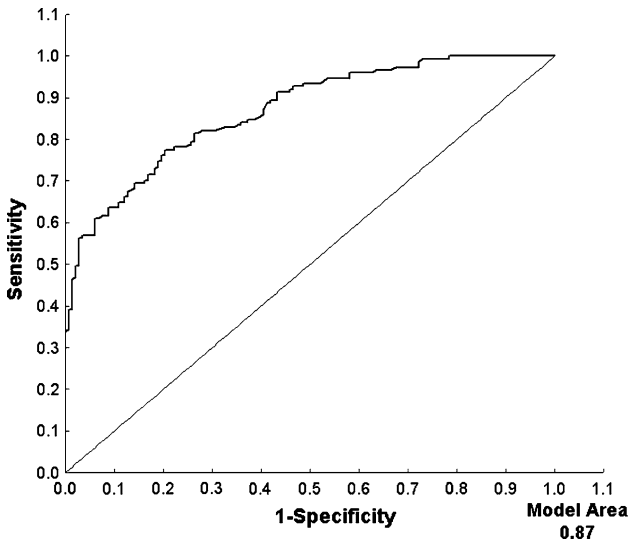



**Fig. 5** Receiver Operating Characteristic curve (ROC-curve) for the bacteriocin model (*dark line*) and random classifier (*tight line*) with areas under curve of 0.87, and 0.5, respectively

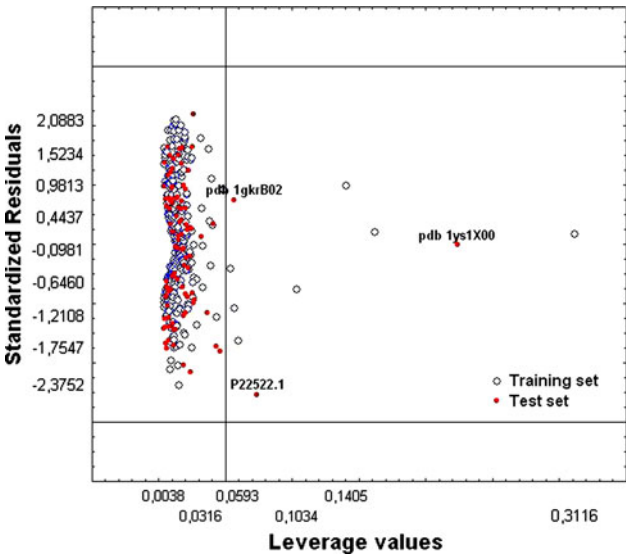



**Fig. 6** Graphical representation for the applicability domain of the model

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.

## 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 $\delta$ 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 $\mu_1$, $\mu_3$, $\mu_{10}$ 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.

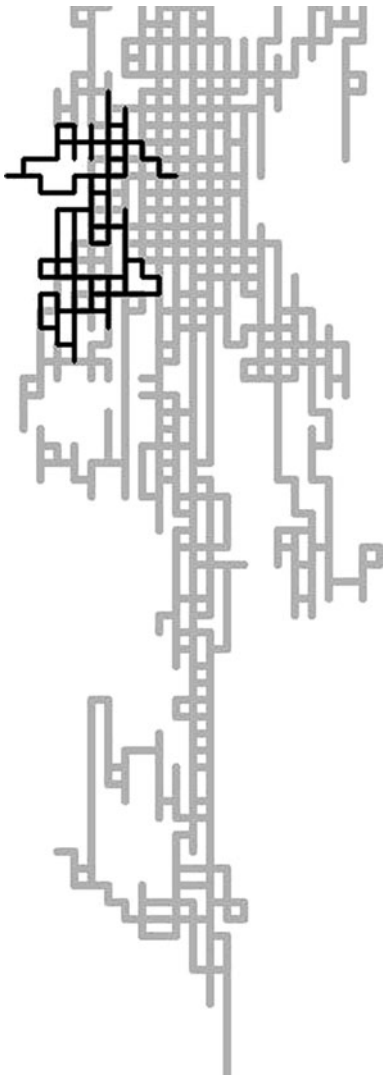


**Fig. 7** Pseudo-folding of the Cry 1Ab C-terminal domain sequence (in *black*) into the bacteriocins 2D-HP lattice (in *gray*)

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

# ANNEX 2

Journal of Theoretical Biology 273 (2011) 167–178

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

# Non-linear models based on simple topological indices to identify RNase III protein members

Guillermin Agüero-Chapin [a,b], Gustavo A de la Riva [c], Reinaldo Molina-Ruiz [b], Aminael Sánchez-Rodríguez [d], Gisselle Pérez-Machado [b], Vítor Vasconcelos [a,e], Agostinho Antunes [a,*]

[a] CIMAR/CIIMAR, Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Rua dos Bragas, 177, 4050-123 Porto, Portugal
[b] Molecular Simulation and Drug Design (CBQ), Central University of Las Villas, Santa Clara 54830, Cuba
[c] Departamento de Biología, Instituto Superior Tecnológico de Irapuato (ITESI), Irapuato, Guanajuato 36821, Mexico
[d] CMPG, Department of Microbial and Molecular Systems, KU Leuven, Kasteelpark Arenberg 20, B-3001 Leuven, Belgium
[e] Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Portugal

ARTICLE INFO

*Article history:*
Received 2 September 2010
Received in revised form
15 November 2010
Accepted 13 December 2010
Available online 28 December 2010

*Keywords:*
Alignment-free models
Spectral moments
Clustering
Decision tree models
Artificial neural networks

ABSTRACT

Alignment-free classifiers are especially useful in the functional classification of protein classes with variable homology and different domain structures. Thus, the **T**opological **I**ndices **to** **BioP**olymers (**TI2BioP**) methodology (Agüero-Chapin et al., 2010) inspired in both the **TOPS-MODE** and the **MARCH-INSIDE** methodologies allows the calculation of simple topological indices (TIs) as alignment-free classifiers. These indices were derived from the clustering of the amino acids into four classes of hydrophobicity and polarity revealing higher sequence-order information beyond the amino acid composition level. The predictability power of such TIs was evaluated for the first time on the RNase III family, due to the high diversity of its members (primary sequence and domain organization). Three non-linear models were developed for RNase III class prediction: Decision Tree Model (DTM), Artificial Neural Networks (ANN)-model and Hidden Markov Model (HMM). The first two are alignment-free approaches, using TIs as input predictors. Their performances were compared with a non-classical HMM, modified according to our amino acid clustering strategy. The alignment-free models showed similar performances on the training and the test sets reaching values above 90% in the overall classification. The non-classical HMM showed the highest rate in the classification with values above 95% in training and 100% in test. Although the higher accuracy of the HMM, the DTM showed simplicity for the RNase III classification with low computational cost. Such simplicity was evaluated in respect to HMM and ANN models for the functional annotation of a new bacterial RNase III class member, isolated and annotated by our group.



## 1. Introduction

There are many software tools for searching sequences into databases, but all use some measure of similarity between sequences to annotate the biological function of certain gene or protein (Strope and Moriyama, 2007). While such available methodologies for sequence classification have a friendly interface for the normal users (Altschul et al., 1997), its algorithms demand a high computational cost and in many cases require the implementation of stochastic process for building a predictive model (Finn et al., 2009). Such procedures turn out to be less effective when the members of a certain gene (Selig et al., 2008) and protein (de Jong et al., 2006) class diverge and show different domain structures; then much more expensive alignment strategies in time and memory are required to improve the classification accuracy. Thus, the development of effective and less costly classification methods based on alignment-free classifiers is important as a complement to alignment-dependent algorithms (Strope and Moriyama, 2007; Deshmukh et al., 2007). To date, most of the alignment-free classifiers estimate 1D sequence parameters based on the amino acid composition to evaluate sequence-function relationships (Kumar et al., 2008), predict protein–protein interactions (Roy et al., 2009) and protein attributes (Chou, 2009).

The introduction of 2D or higher dimension representations of sequences (Liao et al., 2006; Randic and Zupan, 2004) previous to the calculation of such numerical parameters allows uncovering higher-order useful information not encoded by 1D sequence parameters. Thus, we cluster the amino acids of protein sequences according to its charge or its hydrophobic features into a 2D representation or map that provides higher sequence-order information beyond the amino acid composition level. This approach is one of the applications of our methodology **T**opological **I**ndices **to** **BioP**olymers (**TI2BioP**) (Agüero-Chapin et al., 2010) inspired in both the **T**opological **S**ub-structural

* Corresponding author. Tel.: +351 22 3401 813.
*E-mail addresses:* aantunes@ciimar.up.pt, aantunes777@gmail.com (A. Antunes).

0022-5193/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2010.12.019

**Mo**lecular **De**sign (**TOPS-MODE**) (Estrada, 2000) and the **M**arkov **C**hain **I**nvariants for **N**etwork **S**election & **D**esign (**MARCH-INSIDE**) (González-Díaz et al., 2007) methodologies. **TI2BioP** allows the calculation of the spectral moments as simple topological indices (TIs) from different structural representation of biopolymers (DNA, RNA and proteins) that can be used for the prediction of functional classes irrespective of sequence similarity.

The RNase III family was selected as a case of study to assess the predictability power of our alignment-free classifiers (TIs), due to the high diversity among its members (primary sequence and domain organization). This protein class belongs to a super-family that includes an extensive network of distinct and divergent gene lineages (Dyer and Rosenberg, 2006). Although all RNases of this super-family share invariant structural and catalytic elements and some degree of enzymatic activity, the primary sequences have diverged significantly. In fact, the RNase III family can be divided into four subclasses (Lamontagne and Elela, 2004). Class 1 consists of bacterial enzymes with a minimal RNase III domain and a single dsRNA binding domain (dsRBD). Class 2 includes fungal enzymes, with an extra N-terminal region without any recognizable motif. Class 3 comprise the Drosha orthologs found in animals, which has two RNase III domains and one dsRBD in the C-terminal half and a proline-rich domain and an arginine rich (R-rich) domain in the N-terminal half of the protein. Class 4 RNase III enzymes contain the Dicer homologs expressed in *S. pombe*, plants and animals. Their C-terminal half appears similar to Drosha, but the N-terminal half features show different domain structures. The homology among the different RNase IIIs may vary from 20% to 84% depending on their evolutionary distance, suggesting a low level of primary structure conservation (Lamontagne and Elela, 2004).

The electric charge clustering of the amino acids was used to develop three different non-linear models: Classification Trees (CT), Artificial Neural Networks (ANNs) and Hidden Markov Models (HMM), which allowed predicting the RNase III membership of a query sequence. CT and ANN-based models are alignment-free approaches obtained, using our TIs as input predictors. These models were compared with a traditional alignment algorithm to recognize protein signatures: HMM, which was modified by using a non-classical alignment profile based on the clustering of amino acids according to their charges values.

The ANNs have been more frequently applied to the prediction of protein structure and function than the CTs (Punta and Rost, 2008; Nair and Rost, 2008). Although, the CTs are widely used in applied fields as diverse as medicine (diagnosis), computer science (data structures), botany (classification) and psychology (decision theory), due to its easy interpretation based on a graphical representation (Ripley, 1996), they have been poorly explored in proteomics, namely to annotate the biological function of proteins. In this sense, we showed its novel application into the proteomics field by allowing the identification of RNase III-like sequences, using simple TIs as alignment-free classifiers. The simple procedure to search an RNase III protein among the available protein molecular diversity was compared with the classification performance obtained using other artificial intelligent methods, such as ANNs and HMMs.

We showed that the spectral moments were useful as input predictors to develop non-linear models (DTM, ANN) to classify the RNase III family irrespective of sequence similarity. A simple and interpretable alignment-free Decision Tree Model (DTM) was built to detect RNase III-like members, using just one TI at two different levels. However, the ANN-based model used 18 TIs as input predictors and demanded a more complex topology to retrieve similar results in the RNase III family classification. Although, the HMM based on our clustering strategy provided an optimal performance in the prediction of the test set, it is not a practical procedure for a normal user. Therefore, we recommend the easy use of the DTM based on the spectral moments calculated by the TI2BioP methodology (Agüero-Chapin et al., 2010) for the RNase III classification. The performance of the three non-linear models was also compared for the prediction of a new bacterial member of the RNase III class. This sequence was isolated, characterized and annotated by our group at the GenBank Database (accession number GU190214) (Benson et al., 2009). Its DTM detection as a RNase III class member was remarkably simple and required low computational cost relatively to the HMM and ANN models.

## 2. Methods

### 2.1. Computational methods

**TI2BioP** was built up on object-oriented Free Pascal IDE Tools (lazarus) (Agüero-Chapin et al., 2010). The program could be run on Windows and Linux operating systems. The user friendly interface allows the users to access to 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 approximated secondary structure (Mathews, 2006) or the artificial, but informative, folding of linear sequences (Aguero-Chapin et al., 2006). TI2BioP allows the calculation of the spectral moments derived from such inferred and artificial 2D structures of DNA, RNA and proteins. Consequently, it 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). The calculation of the spectral moments as TIs is performed according to the TOPS-MODE approach (Estrada, 2000) and the pseudo-secondary structures for the protein sequences were taken from experiences achieved by using the MARCH-INSIDE methodology (Aguero-Chapin et al., 2006; Agüero-Chapin et al., 2009). We used the 2D lattice of hydrophobicity (H) and polarity (P) introduced by our group to encode an information about polygalacturonases enzymes (Aguero-Chapin et al., 2006) to obtain the protein pseudo-secondary structures.

The 20 different amino acids are regrouped into four HP classes. These four groups characterize the HP physicochemical nature of the amino acids as polar, non-polar, acidic or basic (Jacchieri, 2000). Each amino acid 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 amino acids are calculated as follows:

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

This 2D graphical representation for proteins is similar to those previously reported for DNA (Nandy, 1996; Randic and Vracko, 2000; Nandy, 1994) that was extended later to classify protein families (Aguero-Chapin et al., 2006; Agüero-Chapin et al., 2009) and structural RNA (Aguero-Chapin et al., 2008), using stochastic indices (Yuan, 1999). Fig. 1 shows how the new RNase III protein sequence from *Escherichia coli* BL 21 substrain GG1108 is pseudo-folded into a HP-lattice or 2D-HP map that compacts its linear sequence: its two major

domains are highlighted in red (RNase III domain) and in blue (double-stranded RNA binding motif). Note that a node ($n$) in the 2D-HP map could be made up for more than one amino acid. The N and C termini residues are pointed out in black and red as a square and simple dot, respectively.

All sequences are pseudo-folded into a HP-Cartesian lattice by TI2BioP. The original spectral moments ($\mu_k$) introduced previously by Estrada (Estrada, 1996; Estrada, 1997), which have been validated for many authors to encode the structure of small molecules in Quantitative Structure Activity Relationship (QSAR) studies (Gonzalez-Diaz et al., 2005; Markovic et al., 2001; González et al., 2006), were applied to describe such protein 2D-HP maps ($^{HP}\mu_k$) to contain the new structural information. The original adjacent matrix is modified according to the building of the 2D-HP protein maps described above.

### 2.2. Building an electronic bond matrix for 2D-HP protein maps. Calculation of TIs irrespective of sequence similarity

After the representation of sequences we assigned to each graph, a bond adjacency matrix **B** for the computation of the TIs. They are called "spectral moments", defined as the trace of **B** consisting in the sum of main diagonal entries, of the different powers of the bond

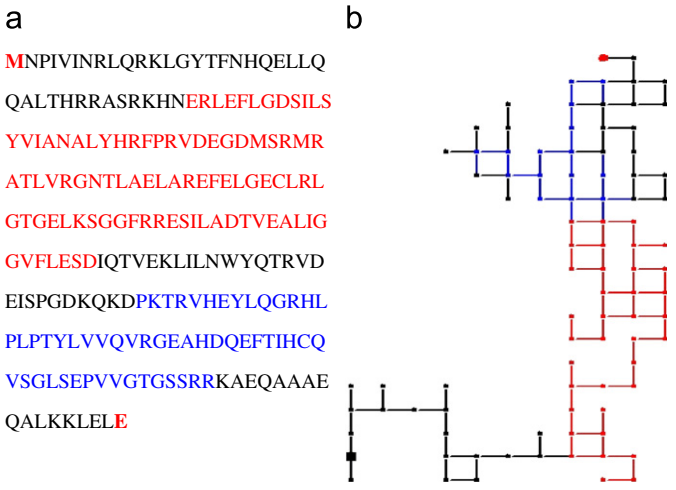



**Fig. 1.** (a) RNase III protein sequence from *Escherichia coli* BL 21 substrain GG1108 and (b) pseudo folding of this sequence into a 2D-HP-lattice.

adjacency matrix. **B** is the square symmetric matrix, where its non-diagonal entries are ones or zeroes if the corresponding bonds or edges share or not one amino acid. Thus, it set up connectivity relationships between the amino acid in the artificial 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 be bonds weights describing hydrophobic/polarity, electronic and steric features of the amino acids. In particular, the main diagonal was weighted with the average of the electrostatic charge ($Q$) between two bound nodes. The charge value $q$ in a node is equal to the sum of the charges of all amino acids placed on it. The $q$ value for each amino acid was derived from the Amber 95 force field (Cornell et al., 1995).

Thus, it is easy to carry out the calculation of the spectral moments of **B** in order to numerically characterize the protein sequence.

$$^{HP}\mu = Tr[(\mathbf{B})^k] \tag{1}$$

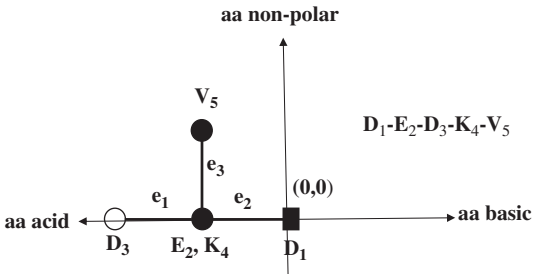
where $Tr$ is the operator "trace" that indicates the sum of all the values in the main diagonal of the matrices $^k\mathbf{B}=(\mathbf{B})^k$, which are the natural powers of **B**.

In order to illustrate the calculation of the spectral moments, an example is developed below. The building of the 2D-HP map on the Cartesian system for the protein fragment ($D_1$–$E_2$–$D_3$–$K_4$–$V_5$), the coordinates for each one of its amino acids and the definition of its bond adjacency matrix are depicted in Fig. 2. The calculation of the spectral moments up to the order $k=3$ is also defined downstream of Fig. 2. Please note in the graph that the central node contains both $E$ and $K$ and $q$ values are represented in the matrix as the amino acid symbols ($E=1.885$, $V=2.24$, $K=2.254$ and $D=1.997$).

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

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

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




a

| Amino acid | HP class | X position | Y position |
|---|---|---|---|
| $D_1$ | acid | 0 | 0 |
| $E_2$ | acid | -1 | 0 |
| $D_3$ | acid | -2 | 0 |
| $K_4$ | basic | -1 | 0 |
| $V_5$ | non-polar | -1 | 1 |

b

| | $e_1$ | $e_2$ | $e_3$ |
|---|---|---|---|
| $e_1$ | $\frac{D_3+E_2+K_4}{2}$ | 1 | 1 |
| $e_2$ | 1 | $\frac{D_1+E_2+K_4}{2}$ | 1 |
| $e_3$ | 1 | 1 | $\frac{V_5+E_2+K_4}{2}$ |

**Fig. 2.** Building the 2D-HP map on Cartesian axes for the protein fragment DEDKV: (a) the coordinates for each amino acid in Cartesian system and (b) the definition of the bond adjacency matrix derived from the 2D-HP map. Note that all edges of the graph are adjacent, thus all non-diagonal entries are ones.

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

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 the 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 a grouping variable (Group) that indicates the RNase III-like proteins with value of 1 and $-1$ for the control group (CG) sequences to statistical analysis software (Statsoft, STATISTICA 7.0, 2007), see File I of supplementary materials (SM).

### 2.3. Database

A total of 206 RNase III protein sequences belonging to prokaryote and eukaryote species were downloaded from the GenBank database gathering RNases III registered up to May of 2009. Each RNase III sequence was labeled by its accession number. The control group was selected from 2015 high-resolution proteins in a structurally non-redundant subset of the Protein Data Bank (PDB); such data were published by other authors to distinguish enzymes and non-enzymes without alignment (Dobson and Doig, 2003), see File I of an SM. The selection of such subset was determined, using a *K*-means cluster analysis (*k*-MCA) (Mc Farland and Gans, 1995). This same procedure was carried out to design the training and predicting series in both groups.

### 2.4. Selection of training and predicting series. K-means cluster analysis (k-MCA)

The selection of members to conform training and predicting series was carried out by *k*-MCA (Mc Farland and Gans, 1995). This method requires a partition of the RNase III group and the 2015 high-resolution proteins independently into several statistically representative clusters of sequences. The RNase III members that conform the training and predicting series were selected straightforward from its clusters according to Euclidean distance.

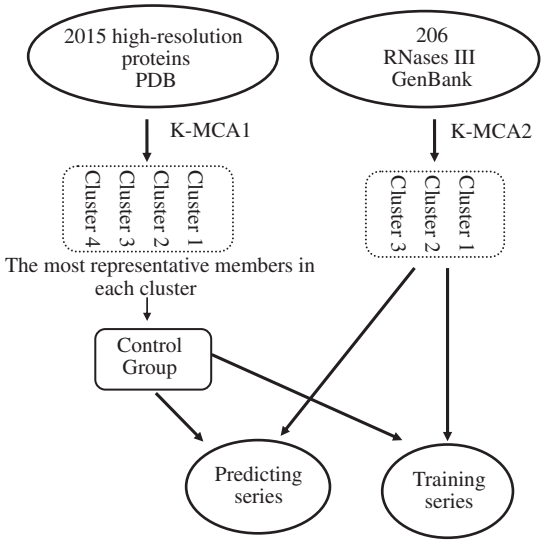



**Fig. 3.** Scheme describing the design of training and predicting series, using *k*-MCA for both RNase III and control group.

A representative sample of 224 non-redundant proteins was set as the control group. This subset was selected from the partition of the 2015 proteins into representative clusters following the same procedure, which ensures the main protein classes will be represented in the control group. Finally the control group was further partitioned in training and prediction series. The spectral moment series was explored as clustering variables, in order to carry out *k*-MCA. This method has been widely applied before in QSAR to design the training and predicting series (Mc Farland and Gans, 1995; Cruz-Monteagudo and Gonzalez-Diaz, 2005). The procedure described above is represented graphically in Fig. 3 for both groups.

### 2.5. Non-linear methods for RNase III classification. Decision tree models

A series of eighteen TIs, consisting in sixteen spectral moments (${}^{HP}\mu_k$), edge numbers and edge connectivity, calculated for protein sequences from training and predicting series were used as ordered predictors to build a DTM, using the CT module of the STATISTICA 7.0 for Windows (Statsoft, STATISTICA 7.0, 2007). A categorical variable that assign the value of 1 to the RNase III class and $-1$ to the control group was set as dependent variable. CT is a technique that builds a classification rule to predict the class membership on the basis of feature information. CT is a data-analysis method for relating a categorical dependent variable ($Y$) to one or more independent variables ($X$), in order to uncover or simply understand the elusive relationship, $Y=f(X)$. The result of CT is a "graph" that divides the study sample into smaller samples (every subsample is called a node) according to whether a particular selected predictor is above of a chosen cutoff value or not. In the development of the DTM, the Classification and Regression Trees (C&RT)-style univariate split selection method was used since it examines all possible splits for each predictor variable at each node to find the split producing the largest improvement in goodness of fit. The prior probabilities were estimated for both groups with equal misclassification cost. The *Gini* index was used as a measure of goodness of fit and the "Prune on misclassification error" was set as the stopping rule to select the right-sized classification tree.

The prediction capacity of the classification model was verified by a cross-validation (CV) procedure. Ten random sub-samples were selected from the learning sample. The classification tree of the specified size is computed ten times, each time leaving out one of the subsamples from the computations, and using such sub-sample as a test sample for cross-validation. The CV costs computed for each of the ten test samples are then averaged to give the 10-fold estimate of the CV costs.

### 2.6. Artificial neural networks (ANN) for RNase III classification

We used an ANN as another non-linear method for RNase III classification, using the same series of TIs as input variables and only one output variable (RNase III membership). We used the multilayer layer perceptron (MLP), due to its ability to model functions of almost arbitrary complexity showing a simple interpretation as a form of an input–output model. As starting point we used one hidden layer, with the number of units equal to half the sum of the number of input and output units. To select the right complexity of network, we tested different topologies to the MLP, but checking the progress against an independent data set to avoid over-fitting during the back propagation training method. The selection set was extracted by *k*-MCA from the training set used to build the DTM, the test set to asses an ANN predictability was the same.

### 2.7. Building a non-classical HMM for the RNase III family

A non-classical profile HMMs for this family were constructed based on the training set, using the HMMer software package (release 2.3.2) (Krogh et al., 1994). In the first place, a HMM representing the appearance probabilities of amino acids charges was obtained. For this purpose, amino acids were grouped according to their charges values as follows.

class-I=(A, S, G); class-II=(M, L, I, V); class-III=(K, R, T, H); class-IV=(N, D, E, Q); and class-V=(F, Y, W). Based on this regrouping, sequences in the training set were modified according to the following criteria: amino acids belonging to the same class were substituted by the same character identifier of each group. Regardless their charge characteristics, proline and cysteine remained unchangeable, due to its biological meaning. The modified training set was then aligned. The *HMM-build* program was used to create a new profile HMM based on the alignment of the set of sequences. Finally, the *HMM-search* was used to score test sequences against the non-classical HMM.

## 3. Experimental section

### 3.1. Strains and culture media

*Escherichia coli BL 21* strain CG 1208 was routinely grown in Luria Broth (LB) medium at 30 °C during 12 h. Bacterial strains *Escherichia coli BL 21* strain CG 1208 and DH5α was grown in Luria Broth (LB). Transformed bacteria were recovered in the same LB medium, but supplemented with carbencillin at 100 μg/ml. Media were also supplemented with bacteriological agar when it was required.

### 3.2. Total DNA extraction

A colony from *Escherichia coli BL 21* strain CG 1208 was inoculated in 5 ml of an LB medium and grown at 30 °C during 12 h until $OD_{600}=0.5$. From this culture, 250 μl were transferred to 50 ml of the same medium and grown overnight at the same temperature. When $OD_{600}=0.8$, cells were collected by centrifugation and broken using the standard procedure. Cellular pellet was resuspended in 300 μl sterile water at 50 °C and the extract was separated from cellular debris by centrifugation. Total DNA was purified using a total DNA extraction kit (Qiagen GmbH, Germany).

### 3.3. Primers design

The primers used for PCR amplification of *Escherichia coli* RNase type III were designed based on the previously reported *E. coli* RNase type III coding sequence (Date and Wickner, 1981; March et al., 1985): forward primer (RNaseIII5′) 5′-c**ccATGG**ACCC-CATCGTAATTAATCGGC-3′ and reverse primer (RNase III3′) 5′-caataaatccgc**ggatcc**ttttatcgatgcTCA-3′. In both, primer sequences are shown the restriction sites Nco1 and BamH1 introduced at 5′ and 3′ ends, the start ATG and the stop TGA codon. The coding regions are also shown in capital letters.

### 3.4. PCR amplifications

Amplification of *E. coli* RNase III gene from *Escherichia coli BL 21* strain CG 1208 was performed by standard PCR from its total DNA. The reaction mixture containing 10 ng of template, 1 mM of each dNTP, 1.5 mM $MgCl_2$, 2 μM of each PAC5′ and PAC3′ primers, in a total volume of 50 μl, 1 × Taq Pol (Gibco BRL) and 2.5 U Taq Pol (Gibco) was completed. The PCR was carried out using thermo-cycler (Perkin Elmer 2400) programmed as follows: 5 min previous template denaturation at 94 °C, cycle steps: 1 min template denaturation at 94 °C, 2 min primer annealing at 45 °C, 1 min primer extension at 72 °C for 30 cycles; plus a final extension step at 72 °C for 5 min. PCR product was visualized by electrophoresis on 1% TBE agarose gel.

### 3.5. Plasmid construction and sequencing

PCR amplification product was purified using GEL Band Purification kit (*Amersham Pharmacia Biotech*) and ligated to pMOS-Blue *T*-vector (*Amersham Pharmacia Biotech*). The ligation was transformed into electrocompetent *E. coli* DH5α by electroporation in 0.2 cm cubettes and Gene Pulser Machine (BioRad) (12.5 kV, 25 μF, 1000ω). Transformation was plated onto an LB medium supplemented with 40 μl of 20 μg/ml X-gal solution and 4 μl of isopropylthio-β-D-galactoside from 200 μg/ml IPTG solution per plate and grown overnight at 37 °C. White colonies, presumable carrying the recombinant *E. coli* RNase III gene inserted in pMOS-Blue *T*-vector, named pREC1, were selected and plasmid DNA extracted for analysis of cloned fragments. Sequencing of cloned fragment was performed using the ABI 3700 sequencer (Applied Biosystems). The cloned gene was properly manipulated for further purification and enzymatic assay purposes as described (Amarasinghe et al., 2001).

### 3.6. Synthesis and preparation of dsRNA substrate for an enzymatic assay

The synthesis and preparation of dsRNA substrate for enzymatic assay of recombinant *E. coli* RNase III was conceived according to create an optimized dsRNA structure for the measurement of enzymatic activity (Lamontagne and Elela, 2004). One of the T7 substrates, named R1.1 RNA (109nt), was used for the biological assay of recombinant enzyme. This short RNA forms hairpin structures containing the recognition and cleavage sites by *E. coli* RNase type III and have been extensively studied (Zhang and Nicholson, 1997). The DNA fragment encoding for 109nt R1.1 RNA was synthesized chemically, purified by denaturing gel-electrophoresis and cloning into pBluescript II KS (-) for further T7 polymerase transcription. The integrity of the cloned fragment was verified by sequencing. The RNA transcripts were generated by T7 polymerases using oligonucleotides as templates and the reactions were carried in the presence of [$\alpha^{32}$P] UTP. The transcription reactions were prepared in a final volume of 20 μl containing 40 mM Tris–HCl (pH 7.9), 6 mM MgCl2, 2 mM spermedine, 10 mM DTT, 0.5 mM of each ribonucleoside (Amersham Pharmacia Biotech), 50 μCi [$\alpha^{32}$P] UTP (800 Ci/mmol), 20 U RNAsin (Promega), and 20 U T7 RNA polymerase (Amersham Pharmacia Biotech). The unpaired RNA strands were removed by an RNase A (Promega) treatment. The dsRNA substrate was purified (PAGE-TBE 15% gel) and stored in diethyl pyrocarbonate (DEPC) treated distilled water at −70 °C and purified for the enzymatic assay.

### 3.7. Enzymatic assay of recombinant E. coli RNase III

The *E. coli* RNase III gene was properly cloned within Nco1 and BamH1 of pIVEX2.4a (Roche Applied Science, Indianapolis, IN 46250, United States) to produce and purify the recombinant protein as described (Amarasinghe et al., 2001) in the form of 6 × (His)-RNase III. Double-stranded RNase activity of recombinant protein form was performed basically with the same method we used for *S. pombe* strain 428-4-1, but with minor variations

(Aguero-Chapin et al., 2008). The *E. coli* assay was carried using the following conditions: 30 mM Tris–HCl (pH 7.6), 1 mM DTT, 10 mM of $MgCl_2$, 10 nM of dsRNA substrate and 100 mM poly- different quantities (0, 1, 10, 100 nM) of purified recombinant *E. coli* RNase type III enzyme. Enzymatic reactions were completed on ice and started by the addition of 0.1 V of 50 mM $MgCl_2$, incubated at 30 °C for 10 min and stopped by addition of 500 μl of 5% ice-cooled TCA followed by 15 min on ice. The aliquots were centrifuged at 16,000 g during 5 min in Spin-X filter unit (Costar). The soluble fractions (filtrate) were quantified by the liquid scintillation counting. The counting data represent the amount of acid precipitable polynucleotide phosphorus (dsRNA) substrate transformed into acid soluble cleavage products by *E. coli* RNase type III enzyme. The procedure was repeated three times with three repetitions per experiment.

## 4. Results and discussion

### 4.1. Predicting type III RNase activity irrespective of sequence alignment

In this work, we calculated the spectral moments ($^{HP}\mu_k$) of the bond adjacency matrix between the amino acids of protein sequences pseudo-folded into the 2D-HP-lattice. Such TIs describe electronically the amino acids connectivity at different orders in a pseudo-secondary structure that is determined by the hydrophobic and polarity features of the amino acids. The calculation was carried out for two groups of protein sequences, one made up of 206 RNase III-like enzymes and other conformed by 224 non-redundant enzymes and non-enzymes as the control group.

The members of the training and predicting series for the RNase III class were selected according to the *k*-MCA, which divided the data into three clusters containing 53, 77 and 76 members. Selection was based on the distance from each member to the cluster center (Euclidean distance). The members of the external validation subset were selected uniformly in respect to Euclidean distance taking out the 25% in each cluster. The remainder of the cases was used to train the model.

To set up the final control group, the original data of 2015 proteins (enzymes and non-enzymes) were reduced to 224 members in order to balance both the groups. 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 was split into four statistically representative clusters of sequences made up by: 267, 430, 655 and 663 members. Afterwards, the members to constitute the training and predicting subsets were selected following the same procedure described for the RNase III class.

Clustering of cases was carried out by using the TIs computed in the **TI2BioP** methodology (Agüero-Chapin et al., 2010). We explored the standard deviation between and within clusters, the respective Fisher ratio and their p-level of significance (Mc Farland and Gans, 1995). All variables were used to construct the clusters, but only the combination from the $^{HP}\mu_{10}$ to $^{HP}\mu_{14}$ showed *p*-levels $< 0.05$ for Fisher test, as depicted in Table 1. We also obtained different mean values for these five variables that produce an evident separation between the clusters (Fig. 4). They described three and four statistically homogeneous clusters for the RNase III class and the control group, respectively.

Such division of the RNase III protein sequences into three clusters according to our TIs is a close approximation to the structure-based characterization reported by Lamontagne and Elela for this family (Lamontagne and Elela, 2004), which divided it into four sub-classes. However, our three groups coincided perfectly with another subdivision based on the biological activity (Nicholson, 1997).

**Table 1**
Main results of the *k*-MCA for the RNase III class and the control group.

| Protein descriptors | Between SS[a] | Within SS[b] | Fisher ratio (*F*) | *p*-Level[c] |
|---|---|---|---|---|
| **Variance analysis RNase III-like proteins** | | | | |
| $^{HP}\mu_{10}$ | 134.49 | 70.51 | 193.60 | $< 0.001$ |
| $^{HP}\mu_{11}$ | 142.75 | 62.25 | 232.75 | $< 0.001$ |
| $^{HP}\mu_{12}$ | 143.97 | 61.03 | 239.44 | $< 0.001$ |
| $^{HP}\mu_{13}$ | 146.00 | 58.99 | 251.23 | $< 0.001$ |
| $^{HP}\mu_{14}$ | 141.02 | 63.98 | 223.73 | $< 0.001$ |
| **Control group** | | | | |
| $^{HP}\mu_{10}$ | 1716.57 | 297.43 | 3868.72 | $< 0.001$ |
| $^{HP}\mu_{11}$ | 1763.70 | 250.30 | 4723.45 | $< 0.001$ |
| $^{HP}\mu_{12}$ | 1760.22 | 253.78 | 4649.43 | $< 0.001$ |
| $^{HP}\mu_{13}$ | 1770.23 | 243.77 | 4867.85 | $< 0.001$ |
| $^{HP}\mu_{14}$ | 1767.92 | 246.08 | 4815.87 | $< 0.001$ |

[a] Variability between groups.
[b] Variability within groups.
[c] Level of significance.

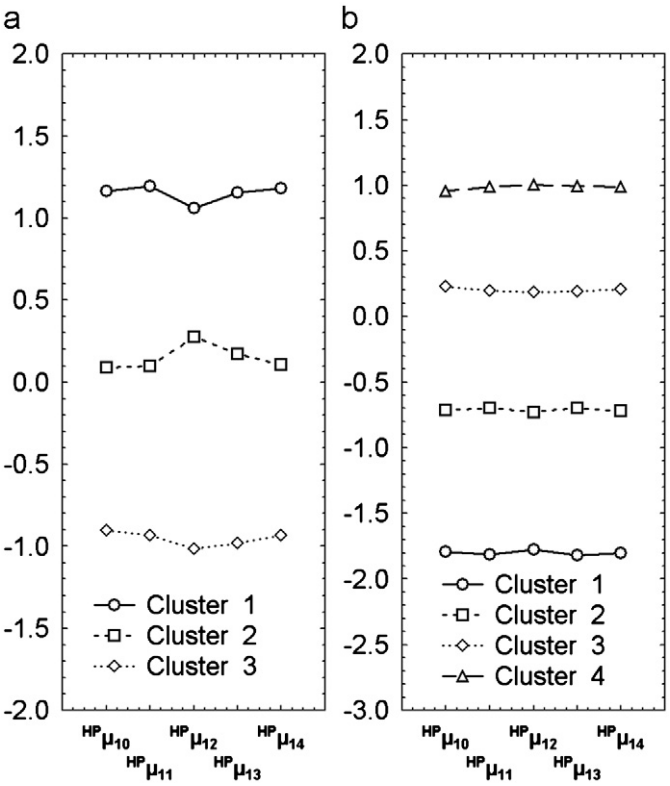



**Fig. 4.** Plot of the TIs's means for each cluster: (a) division of the RNase III group into three clusters and (b) division of the control group into four clusters.

### 4.2. Prediction based on DTM using TIs

Although different alignment-free methods have been reported for improving the classification accuracy in protein classes and superfamilies, to date no DTM have been developed to differentiate the protein classes. We select the RNase III class to assess the DTM predictability, due to its diversity in sequence similarity and domain organization between its members representing different subclasses. Thus, we used the CT as an exploratory technique to obtain a DTM to differentiate the RNase III class from a non-redundant subset of

enzymes and non-enzymes, as linear traditional methods failed to succeed on that goal. We carried out previously a general discrimination analysis (GDA) for variable selection to build up a linear model (Cruz-Monteagudo et al., 2006; Cruz-Monteagudo et al., 2008; Marrero-Ponce et al., 2007). Eighteen variables, including a series of sixteen ${}^{HP}\mu_k$ calculated by **TI2BioP** methodology, were reviewed for finding the "best" possible sub model with the *STATISTICA* software. The best sub model selected from 262,126 models showed a Wilk's statistic of 0.86, indicating little separation between the two groups. All predictors entered significantly into the model, but just provided an overall classification of 62.32%. In contrast, the development of DTM based on C&RT-style exhaustive search for univariate splits showed excellent results on the RNase III classification.

The method found the ${}^{HP}\mu_1$ predictor as the splitting variable to produce two decision splits at different values showing the largest improvement in goodness of fit, therefore an effective classification was developed. The tree structure was very simple, two decision nodes (outlined in blue) and three terminal nodes (outlined in red) summing up a total of five nodes. In the graph, the numbers of the nodes are labeled on its top-left-corner. All 323 training sequences are assigned to the root node (first node) and tentatively classified as non-RNase III enzymes or the control group, as is indicated by the control group label $(-1)$ placed in the top-right-corner of the root node. Sequences from the control group are chosen as the initial classification, because they are slightly more than RNase III enzymes (1), as is indicated by the histogram plotted within the root node.

The root node is split, forming two new nodes. The text below the root node describes the split. It indicates that protein sequences with ${}^{HP}\mu_1$ values $\leq 422.6$ are sent to node number 2 and tentatively classified as RNase III enzymes, and the protein sequences with ${}^{HP}\mu_1$ values $> 422.6$ are assigned to node number 3 and classified as non-RNase III enzymes or other non-enzymatic proteins. Similarly, node 2 is subsequently split taking the decision that sequences with ${}^{HP}\mu_1$ values $\leq 339.69$ are sent to node number 4 to be classified in the control group (59 cases). The remaining 160 proteins with ${}^{HP}\mu_1$ values of $> 339.69$ are sent to node number 5 to be classified as RNase III enzymes.

The tree graph presents all this information in a simple and straightforward way allowing evaluating the information in much less time. The histograms plotted within the tree's terminal nodes show that the classification tree classifies the RNase III enzymes from the control group quite efficiently (Fig. 5). All the information in the tree graph is also available in the tree structure shown in Table 2.

When univariate splits are performed, the predictor variables can be ranked on a 0–100 scale in terms of their potential importance in accounting for responses on the dependent variable (Breiman et al., 1984). In this case, ${}^{HP}\mu_1$ is clearly the most important predictor to discriminate the RNase III class from other protein signatures (Fig. 6).

The DTM classified correctly 296 out of the 323 proteins used in the training series (level of accuracy of 91.64%). More specifically, the model correctly classified 144/155 (92.90%) of RNase III-like sequences and 152/168 (90.48%) of the control group. In order to minimize the computational cost, the DTM was validated using the 10-fold cross-validation method. For this purpose, we took out randomly 65 sequences representing the 20% of the training set to examine the prediction accuracy of the model. The procedure was repeated 10 times varying the composition of the sub-samples. The mean values for the accuracy, sensitivity and specificity obtained in the 10-fold cross-validation on the training sample were very similar to those achieved from the data partition, using *k*-MCA showing the

**Table 2**
Tree structure in details, child nodes, observed class *n*'s, predicted class and split condition for each node.

| Node | Left branch | Right branch | *n* In control (−1) | *n* In RNase III class (1) | Predicted class | Split constant | Split variable |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 168 | 155 | −1 | −422.602 | ${}^{HP}\mu_1$ |
| 2 | 4 | 5 | 66 | 153 | 1 | −339.687 | ${}^{HP}\mu_1$ |
| **3** | | | **102** | 2 | −1 | | |
| **4** | | | **50** | 9 | −1 | | |
| **5** | | | 16 | **144** | 1 | | |

Numbers in bold highlight the well-classified cases and the terminal nodes.

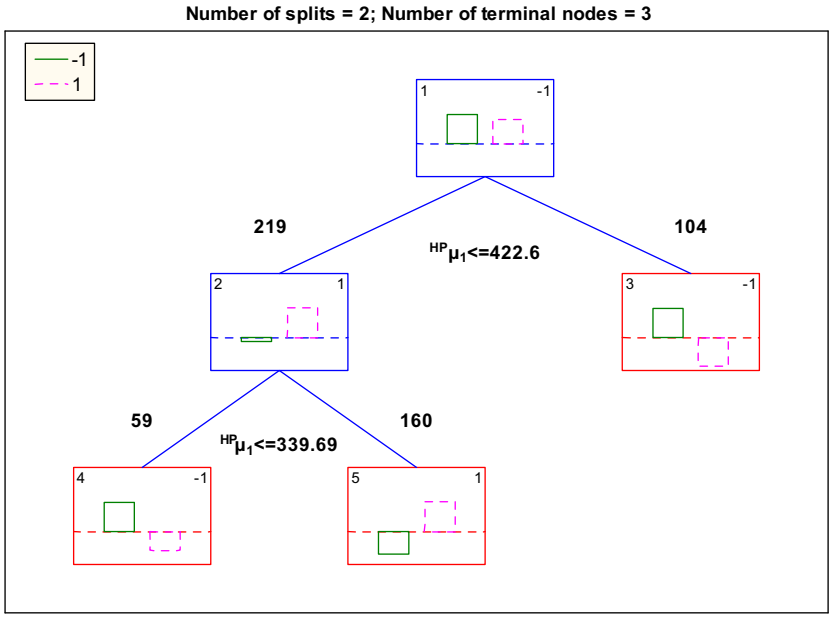



**Fig. 5.** The architecture of the DTM: decision nodes are represented in blue and terminal nodes in red. The RNase III class is labeled with 1, using an intermittent line. Otherwise the control group is signed with −1 using a continuous line. Numbers at the right-corner of the nodes indicates tentative membership to one group. Numbers at the left-corner represent the node's number. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

robustness of the DTM. The classification matrices for training and cross-validation are depicted in Table 3.

An external validation was also performed using the same cross-validation method mentioned above on 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 107 protein sequences, 51 RNase III-like proteins and 56 proteins from the control group (**see** Table 3 **and File IISM**). The model showed a prediction overall performance of 92.52%, being able to predict 49/51 (96.07%) of the ribonucleases III and 50/56 (89.28%) of the functionally diverse proteins. The cross-validation cost (CV cost) and standard deviation (SD) in misclassification were also explored for the two validation procedures to evaluate the predictability performance. Both cases showed values $<0.5$, which is an excellent result for the misclassification of the model.

The retrieved DTM structure is very simple and its graphical display makes easier the interpretation of the data classification. Particularly, the spectral moment $^{HP}\mu_1$ is the split condition at two levels to predict membership of protein sequences in the RNase III class or in other structural and functional different groups. This fact points out that proteins sequences pseudo-folded into 2D-HP maps with values of $339.69 \leq {}^{HP}\mu_1 \leq 422.6$ are more likely to present double-stranded ribonuclease activity.

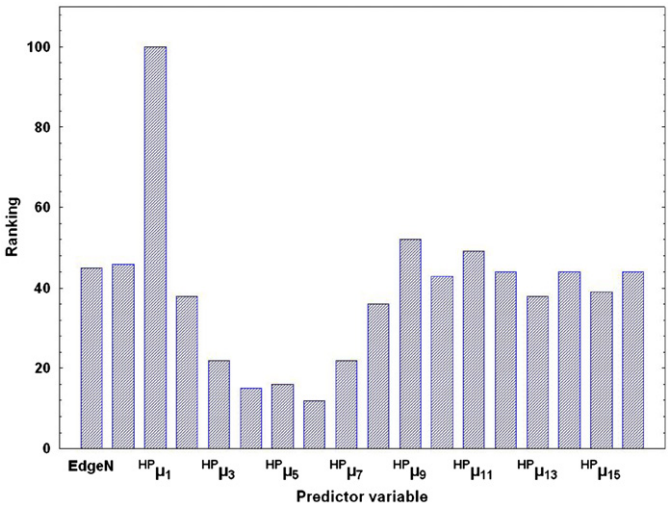



**Fig. 6.** Predictor variable importance rankings, rankings on scale from 0=low importance to 100=high importance.

**Table 3**
Classification results derived from CT for the training and the validation series. Predicted class (row) × observed class *n*'s (column).

| | Training sample *k*-MCA ($N$=323) | | | Cross-validation 10-fold | | | |
|---|---|---|---|---|---|---|---|
| | **Class (%)** | **RNase III class** | **Control group** | **Class (%)** | **RNase III class** | **Control group** | **CV cost** |
| **RNase III class** | 92.90 | **144** | 16 | 92.90 | **144** | 20 | 0.095 |
| **Control Group** | 90.48 | 11 | **152** | 88.10 | 11 | **148** | **SD** |
| **Total** | **91.64** | 155 | 168 | **90.40** | 155 | 168 | 0.016 |
| | **External validation** ($N$=107) | | | | | | |
| | **Class (%)** | | **RNase III class** | | **Control group** | **CV cost** | |
| **RNase III class** | 96.07 | | **49** | | 6 | 0.07 | |
| **Control group** | 89.28 | | 2 | | **50** | **SD** | |
| **Total** | **92.52** | | 51 | | 56 | 0.025 | |

Numbers in bold highlight the well-classified cases.

### 4.3. Artificial neural networks (ANN) in the prediction of the RNase III class

The complexity of DTM as a non-linear statistical method to predict the RNase III class, using our TIs was evaluated in respect to another non-linear method: ANN. The multilayer layer perceptron (MLP) was selected as the most popular ANN architecture in use today (Rumelhart and McClelland, 1986). The MLP was tested at different topologies using the 18 predictors calculated by the TI2BioP methodology as input variables. From the same training set used to develop the DTM, an independent data set (the selection set) was selected to keep an independent check on the progress of the back propagation algorithm used for training. Such selection set was chosen by *k*-MCA to take out a representative subset of 61 sequences that were not used in the back propagation algorithm. Thus, 262 cases were used for the training and the same test subset made up of 107 cases was evaluated on the external validation (**File IISM**). Table 4 shows the different MLP topologies used to select the right complexity of network, the performance on training, selection and test progress were examined as well as its errors. The best model was the MLP profile number 7 (highlighted in bold), which showed an excellent performance on training, selection and test sets, minimizing its respective errors.

This ANN model showed an overall classification in training, selection and test of 93.89%, 93.34% and 90.65%, respectively,

**Table 4**
Different topologies for the MLP on the RNase III classification. Performance and error on training, selection and test sets.

| **Model summary report** | | | | | | | |
|---|---|---|---|---|---|---|---|
| | **MLP profile** | **Train perf.** | **Select perf.** | **Test perf.** | **Train error** | **Select error** | **Test error** |
| **1** | 18:18-10-1:1 | 0.885 | 0.967 | 0.850 | 0.303 | 0.226 | 0.325 |
| **2** | 18:18-9-1:1 | 0.946 | 0.934 | 0.869 | 0.214 | 0.226 | 0.336 |
| **3** | 18:18-8-1:1 | 0.954 | 0.934 | 0.887 | 0.216 | 0.223 | 0.343 |
| **4** | 18:18-7-1:1 | 0.893 | 0.918 | 0.897 | 0.291 | 0.278 | 0.334 |
| **5** | 18:18-6-1:1 | 0.923 | 0.885 | 0.869 | 0.281 | 0.311 | 0.338 |
| **6** | 18:18-5-1:1 | 0.904 | 0.901 | 0.850 | 0.284 | 0.291 | 0.351 |
| **7** | **18:18-4-1:1** | **0.938** | **0.934** | **0.906** | **0.240** | **0.244** | **0.294** |
| **8** | 18:18-3-1:1 | 0.908 | 0.885 | 0.831 | 0.288 | 0.291 | 0.363 |
| **9** | 18:18-2-1:1 | 0.541 | 0.524 | 0.626 | 0.459 | 0.459 | 0.461 |
| **10** | 18:18-2-1:1 | 0.923 | 0.918 | 0.869 | 0.264 | 0.284 | 0.345 |

**Table 5**
Classification results derived from an ANN (MPL-7) for training, selection and test series.

| | **Train (1)** | **Train (−1)** | **Selection (1)** | **Selection (−1)** | **Test (1)** | **Test (−1)** |
|---|---|---|---|---|---|---|
| **RNase III class (1)** | **119** | 9 | **28** | 3 | **47** | 6 |
| **Control group (−1)** | 7 | **127** | 1 | **29** | 4 | **50** |
| **Total** | 126 | 136 | 29 | 32 | 51 | 56 |
| **Good class (%)** | 94.44 | 93.38 | 96.55 | 90.62 | 92.15 | 89.28 |
| **Overall class (%)** | 93.89 | | 93.34 | | 90.65 | |

Numbers in bold highlight the well-classified cases.

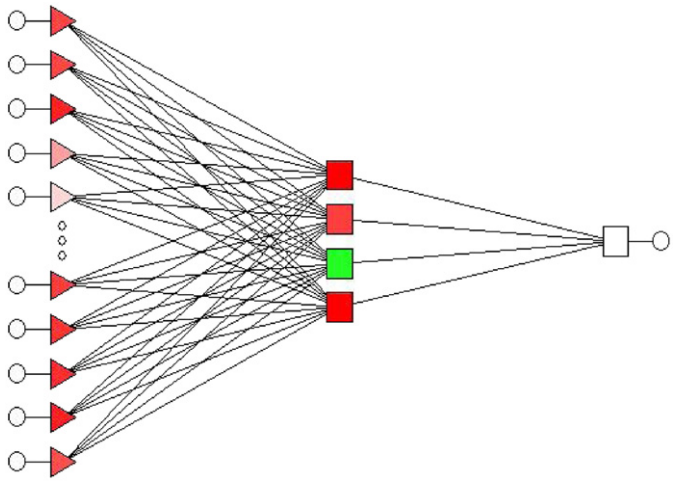


**Fig. 7.** The architecture of the MLP profile 7. It represents several input variables, four neurons in a one layer and only one output variable (from the left to the right).

**Table 6**
Classification results on an RNase III class derived from the three classification algorithms used in the study. DTM, ANN-MLP and HMM modified for training and test series in the RNase III class and control group (CG).

| | DTM | | ANN-MLP | | HMM modified | |
|---|---|---|---|---|---|---|
| | **RNase III** | **CG** | **RNase III** | **CG** | **RNase III** | **CG** |
| **Training** | 92.90 | 90.48 | 94.44 | 93.38 | 94.83 | 100 |
| **Overall** | **91.64** | | **93.89** | | **96.11** | |
| **Test** | 96.07 | 89.28 | 92.15 | 89.28 | 100 | 100 |
| **Overall** | **92.52** | | **90.65** | | **100** | |

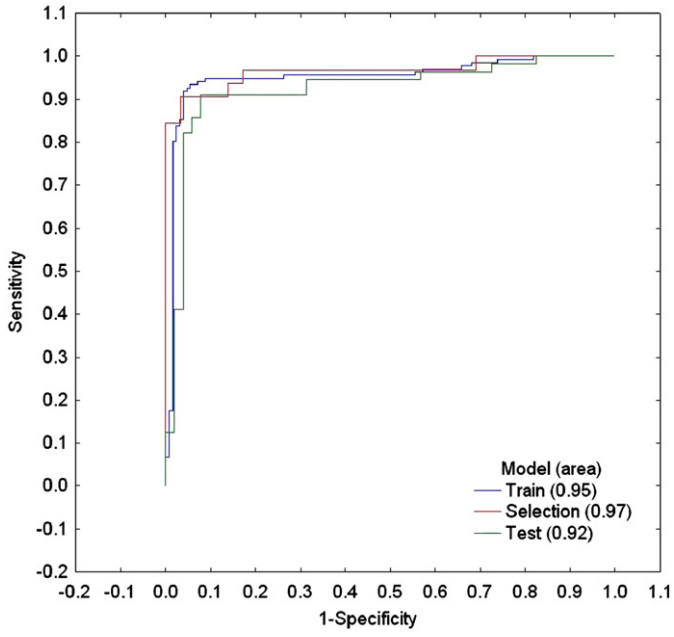



**Fig. 8.** Receiver operating characteristic curve (ROC-curve) for the ANN-based model in training (blue line), selection (red line) and test (green line) sets with areas under curves of 0.95, 0.97 and 0.92, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

which are quite good results taking into account the classification values reported for protein families with a higher degree of conservation (Agüero-Chapin et al., 2009). The classification results derived from our alignment-free approach to classify RNase III membership are showed in Table 5 and in File IISM for more details.

Although the excellent results obtained, the method is based on a non-linear function of high complexity implemented in the MLP classifier. ANN-based models are complex non-linear functions that are unknown, therefore hard to interpret. In addition, the 18 predictors entered in the ANN model, using one hidden layer made up of four neurons representing a more complex architecture to face the RNase III classification in contrast with the simplicity of the DTM. Fig. 7 depicts the network map for the best MLP model Table 6.

To validate the ANN model, we constructed the receiver operating characteristic (ROC) curve for the training, selection and test subsets. In each case, curve presented an area higher than 0.5 reaching values of 0.95, 0.97 and 0.92 for training, selection and test sets, respectively (Fig. 8). According to the ROC-curve theory, random classifiers have an area of only 0.5. This result confirms that the present model is a significant classifier relatively to those working at random. The validity of this type of procedures in developing ANN-QSAR models have been demonstrated before, namely by Caballero and Fernandez (2008), Fernandez et al. (2007) and Caballero et al. (2007).

### 4.4. Non-classical HMM in RNase III classification

In order to compare with other non-linear methodologies based on the sequence alignment, the training and the test set from the RNase III class and control group were scored against a non-classical HMMs profile. We constructed a modified training set representing the electrical properties of the amino acids to add sense to the comparison with the **TI2BioP** methodology. The retrieved HMM represents the occurrence probabilities of amino acids charge groups. As this modification has an implicit generalization step, we expect this model to perform better in detecting remote homologs than classical HMMs. Since our TIs encode information of the complete sequence, we present the classification results for the whole sequences. The HMM performance on an RNase III training set was 94.83%, 147 out of 155 satisfied the *E*-value cut off, while the test set was successfully predicted at 100% (51/51). In the case of the control group coming from a high-resolution non-redundant subset from PDB, the HMM did not recognize any RNase III sequence in the training and the test sets of this group, showing a classification of 100% (**see File IIISM**). We consider a better general performance of the modified HMMs, due to the hydrophobic clustering in the alignment profile according to the amino acids charges. In fact, in the previous reports, the application of classical HMM on RNase III classification showed a major failing rate on a similar control subset (Aguero-Chapin et al., 2008).

Our free-alignment approach TI2BioP provides simplicity to non-linear methods like DTM that can be used as an alternative classification method for the RNase III class allowing a simple screening of a large set of proteins and at low computational cost. It just requires carrying out the calculation of ${}^{HP}\mu_1$ values for the 2D-HP protein maps (automatically represented and calculated by the TI2BioP methodology). On the other hand, the basis of our graphical approach inspired the building of a non-classical HMM-profile to increase the prediction accuracy in the recognition of double-stranded ribonucleases. Although maximal prediction percentages were attained, its main drawback stems from its hard implementation for non-specialized researches. The prediction of a completely new putative RNase III type sequence (unregistered previously in a public database) represents another way of validating the DTM simplicity in respect to the HMM and the ANN models.

### 4.5. Isolation, prediction and biological activity for a new RNase III member

#### 4.5.1. Isolation and sequencing

We isolated, cloned and expressed a new putative RNase type III DNA sequence from *Escherichia coli BL 21* strain CG 1208. Total DNA solution was measured at 260 nm in a spectrophotometer reaching a concentration of 3.8 μg/μl. It was also run on an agarose gel 0.8% visualizing high integrity. The PCR reaction showed a band coinciding with the size of the predicted ORF (data not showed). Sequencing retrieved a product of 681 kB, and its nucleotide and amino acid sequence from a genomic-cloned gene was recorded at the GenBank database with the accession number GU190214. Before submission to GenBank, this new RNase III member was also predicted using our three non-linear models and further tested enzymatically as a ribonuclease.

#### 4.5.2. Prediction of GU190214 using non-linear models. A comparative study

We analyzed our new RNase III sequence GU190214 using TI2BioP methodology to predict its protein open reading frame (ORF) as a member of the RNase III class. Its deduced protein ORF was automatically pseudo-folded into a hydrophobicity and polarity lattice as performed previously for the whole data set. Afterwards, its ${}^{HP}\mu_1$ value was calculated according to the TI2BioP methodology. It showed a ${}^{HP}\mu_1$ value of 422.38, which was further evaluated on the DTM. Following the tree graph representing the DTM, we can easily classify our query sequence. Accordingly to the first decision on the node two, it is classified as an RNase III; then after a second decision, the classification was reaffirmed being submitted it to the terminal node number five. The prediction of our query sequence, using the other alignment-free non-linear model, was also carried out. This particular case was included in the validation subset to be predicted, using the ANN-based model. Finally, the MLP also classified it in the group of the RNase III class supporting that the identification of protein signatures tend to be better assessed with non-linear models.

In order to compare the prediction with classical alignment procedures based on the non-linear functions, our protein query sequence was coded according to the amino acid charge clustering and assessed against the non-classical RNase III HMM-profile. The *HMM-search* predicted it with a high score of 154.7, highly significant (*E*-value of $5.5 \times 10^{-47}$) in the recognition of the ribonuclease III domain. All three models showed a good performance in the classification of the query sequence. However, the simplicity of DTM to classify a protein sequence based only on two values of one predictor is remarkable in respect to the others procedures. An ANN-based model retrieved a similar performance in the classification, but it was built on the basis of 18 predictors and its model architecture is much more complex than DTM. Although the non-classical HMM showed the best performance for the query sequence and the whole database, its implementation requires the building of a modified HMM-profile based on an amino acid charge clustering, the codification of the query sequence and running the HMM-search program, which demand a much higher computational cost. All these steps hinder its practicality for a normal user that wants to retrieve an information easily. On the other hand, we demonstrated that our strategy of an amino acid clustering according to their charge or to hydrophobic features can increase the accuracy in the classification of protein families with divergent members either using classical procedures or alignment-free models.

#### 4.5.3. Enzymatic assay of the recombinant RNase III

The recombinant enzyme was expressed in *E. coli* DH5α strain and purified, as we described previously. Fig. 9 shows the results of the expression and the purification assays. The double-stranded RNase activity of the recombinant protein from the *E. coli* strain BL 21 CG 1208 was measured in vitro following the protocol described above. The unit definition for all RNase III types is the amount of enzyme able to solubilize 1 nmol of acid precipitable per hour (Dunn and Ribonulcease III, 1982). Enzymatic activity showed values of $5.858 \times 10^5$, $6.017 \times 10^5$ and $6.177 \times 10^5$ U/mg, for each assay, and the mean value was $6.017 \times 10^5$ U/mg **(see** Table 7**)**.

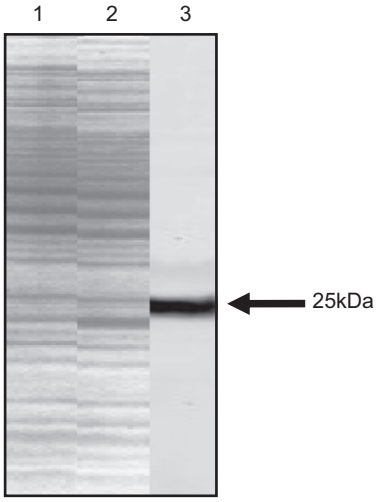



**Fig. 9.** Electrophoresis of the 25 kDa recombinant *E. coli* RNase III from *E. coli* DH5α: pREC1 loaded in 12.5% PAGE-SDS and stained with coomassie brilliant. Lane 1: crude extract from non-induced bacteria; Lane 2: crude extract from an induced bacteria; and Lane 3: purified recombinant *E. coli* RNase III.

**Table 7**
Assay of biological activity of recombinant bacterial RNase III using 10 nM of dsRNA substrate and polydifferent quantities of recombinant enzyme: 0, 1, 10 and 100 nM. The procedure consisted in three independent experiments with three repetitions per experiment.

| Enzyme | Enzymatic activity | | |
|---|---|---|---|
| nM | Experiment 1 $10^5$ U/mg | Experiment 2 $10^5$ U/mg | Experiment 3 $10^5$ U/mg |
| | 0.011 | 0.042 | 0.021 |
| 0.0 | 0.023 | 0.016 | 0.011 |
| | 0.012 | 0.014 | 0.013 |
| | 6.200 | 6.015 | 6.312 |
| 0.1 | 6.512 | 5.912 | 6.801 |
| | 6.011 | 6.108 | 6.709 |
| | 6.701 | 5.519 | 6.089 |
| 1.0 | 6.603 | 5.808 | 5.816 |
| | 6.415 | 5.901 | 5.588 |
| | 6.211 | 6.009 | 6.131 |
| 10.0 | 6.112 | 6.221 | 6.674 |
| | 6.221 | 6.325 | 6.415 |
| | 6.306 | 6.119 | 6.201 |
| 100.0 | 6.614 | 6.201 | 5.803 |
| | 6.507 | 6.067 | 5.587 |
| Average | **5.858** | **6.017** | **6.177** |

## 5. Conclusions

The amino acid clustering in a protein sequence according to the hydrophobic features or to charge properties at the primary

level and higher sequence-orders is effective to produce non-linear functions with high prediction power for the RNase III class. When this clustering is projected into a 2D protein map, it is possible to calculate simple TIs characterizing the protein sequence. Thus, TIs can be used to develop alignment-free approaches based on DTM and ANN, being of great utility for the classification of functional protein classes with low sequence similarity. Although, the non-classical HMM provided a higher accuracy in the prediction on the RNase III class, the use of DTM based on the TI2BioP methodology also showed excellent results in the detection of molecular diverse members of this protein class with low computational and procedure costs.

## Acknowledgments

The authors acknowledge the Portuguese Fundação para a Ciência e a Tecnologia (FCT) for financial support to GACH (SFRH/BD/47256/2008) and the project PTDC/BIA-BDE/69144/2006 and PTDC/AAC-AMB/104983/2008.

## Appendix A. Supporting information

Detailed information on the protein sequences used in the study is supplied in the online **Supplementary Materials** including IDs or accession numbers, training and prediction series, values of the TIs predictors, cluster members (**File ISM).** Classification results derived from DTM and ANN-model on the test set (**File IISM**). HMM classification results on training and test sets are also showed in **File IIISM**. This information is available free of charge via the Internet at: doi:10.1016/j.jtbi.2010.12.019.

## References

Agüero-Chapin, G., Pérez-Machado, G., Molina-Ruiz, R., Pérez-Castillo, Y., Morales-Helguera, A., Vasconcelos, V., Antunes, A., 2010. TI2BioP: topological indices to biopolymers. Its practical use to unravel cryptic bacteriocin-like domains. Amino Acids. doi:10.1007/s00726-010-0653-9.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J, 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Aguero-Chapin, G., Gonzalez-Diaz, H., Molina, R., Varona-Santos, J., Uriarte, E., Gonzalez-Diaz, Y., 2006. Novel 2D maps and coupling numbers for protein sequences. The first QSAR study of polygalacturonases; isolation and prediction of a novel sequence from *Psidium guajava* L. FEBS Lett. 580, 723–730.

Agüero-Chapin, G., Varona-Santos, J., Riva, G.d.l., Antunes, A., Gonza´lez-Villa, T., Uriarte, E., Gonza´lez-Di´az, 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.

Aguero-Chapin, G., Antunes, A., Ubeira, F.M., Chou, K.C., Gonzalez-Diaz, H., 2008. Comparative study of topological indices of macro/supramolecular RNA complex networks. J. Chem. Inf. Modeling 48, 2265–2277.

Amarasinghe, A.K., Calin-Jageman, I., Harmouch, A., Sun, W., Nicholson, A.W., 2001. *Escherichia coli* ribonuclease III: affinity purification of hexahistidine-tagged enzyme and assays for substrate binding and cleavage. Methods Enzymol. 342, 143–158.

Aguero-Chapin, G., Gonzalez-Diaz, H., de la Riva, G., Rodriguez, E., Sanchez-Rodriguez, A., Podda, G., Vazquez-Padron, R.I., 2008. 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. Modeling 48, 434–448.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., 2009. GenBank. Nucleic Acids Res. 37, D26–D31.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. University of California, California, San Diego.

Chou, K.C., 2009. Automated prediction of protein attributes and its impact to biomedicine and drug discovery. In: Alterovitz, G., Benson., R., Ramoni, M.F. (Eds.), Automation in Proteomics and Genomics: an Engineering Case-Based Approach (Harvard-MIT Interdisciplinary Special Studies Courses). Wiley & Sons, UK, pp. 97–143.

Cornell, W.D., Cieplak, P., IBayly, C., Gould, I.R., Merz, K.W.J., Ferguson, D.M., CSpellmeyer, P.A., Fox, T., Caldwell, J.W., Kollman, P.A., 1995. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J. Am. Chem. Soc. 117, 5179–5197.

Cruz-Monteagudo, M., Gonzalez-Diaz, H., 2005. Unified drug–target interaction thermodynamic Markov model using stochastic entropies to predict multiple drugs side effects. Eur. J. Med. Chem. 40, 1030–1041.

Cruz-Monteagudo, M., Gonzalez-Diaz, H., Uriarte, E., 2006. Simple stochastic fingerprints towards mathematical modeling in biology and medicine 2. Unifying Markov model for drugs side effects. Bull. Math. Biol. 68, 1527–1554.

Cruz-Monteagudo, M., Munteanu, C.R., Borges, F., Cordeiro, M.N., Uriarte, E., Gonzalez-Diaz, H., 2008. Quantitative proteome–property relationships (QPPRs). Part 1: finding biomarkers of organic drugs with mean Markov connectivity indices of spiral networks of blood mass spectra. Bioorg. Med. Chem. 16, 9684–9693.

Caballero, J., Fernandez, M., 2008. Artificial neural networks from MATLAB in medicinal chemistry. Bayesian-regularized genetic neural networks (BRGNN): application to the prediction of the antagonistic activity against human platelet thrombin receptor (PAR-1). Curr. Top. Med. Chem. 8, 1580–1605.

Caballero, J., Zampini, F.M., Collina, S., Fernandez, M., 2007. Quantitative structure–activity relationship modeling of growth hormone secretagogues agonist activity of some tetrahydroisoquinoline 1-carboxamides. Chem. Biol. Drug Des. 69, 48–55.

Deshmukh, S., Khaitan, S., Das, D., Gupta, M., Wangikar, P.P., 2007. An alignment-free method for classification of protein sequences. Protein Pept. Lett. 14, 647–657.

Dyer, K.D., Rosenberg, H.F., 2006. The RNase a superfamily: generation of diversity and innate host defense. Mol. Divers. 10, 585–597.

Dobson, P.D., Doig, A.J., 2003. Distinguishing enzyme structures from non-enzymes without alignments. J. Mol. Biol. 330, 771–783.

Date, T., Wickner, W., 1981. Isolation of the *Escherichia coli* leader peptidase gene and effects of leader peptidase overproduction in vivo. In: Proc. Natl. Acad. Sci. USA 78, 6106–6110.

Dunn, J., Ribonulcease III, J., 1982. In: The Enzymes. Academic Press, New York.

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., 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.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E.L., Eddy, S.R., Bateman, A., 2009. The Pfam protein families database. Nucleic Acids Res..

Fernandez, L., Caballero, J., Abreu, J.I., Fernandez, M., 2007. Amino acid sequence autocorrelation vectors and Bayesian-regularized genetic neural networks for modeling protein conformational stability: gene V protein mutants. Proteins 67, 834–852.

González-Díaz H, Molina-Ruiz R, Hernandez I, MARCH-INSIDE v3.0 (**MAR**kov **CH**ains **IN**variants for **SI**mulation & **DE**sign), 2007, 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.

Gonzalez-Diaz, H., Cruz-Monteagudo, M., Vina, D., Santana, L., Uriarte, E., De Clercq, E., 2005. 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, M.P., Teran, C., Teijeira, M., 2006. A topological function based on spectral moments for predicting affinity towards $A_3$ adenosine receptors. Bioorg. Med. Chem. Lett. 16, 1291–1296.

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

Kumar, M., Thakur, V., Raghava, G.P., 2008. COPid: composition based protein identification. Silico Biol. 8, 121–128.

Krogh, A.B., Brown, M., Mian, I.S., Sjeander, K., Haussler, D., 1994. Hidden Markov models in computational biology. Applications to protein modeling. J. Mol. Biol. 235, 1501–1531.

Liao, B., Luo, J., Li, R., Zhu, W., 2006. RNA secondary structure 2D graphical representation without degeneracy. Int. J. Quantum Chem. 106, 1749–1755.

Lamontagne, B, Elela, S.A, 2004. Evaluation of the RNA determinants for bacterial and yeast RNase III binding and cleavage. J. Biol. Chem. 279, 2231–2241.

Mathews, D.H., 2006. RNA secondary structure analysis using RNA structure. Curr. Protocols Bioinformatics. Chapter 12 Unit 12 6.

Markovic, S., Markovic, Z., McCrindle, R.I., 2001. Spectral moments of phenylenes. J. Chem. Inf. Comput. Sci. 41, 112–119.

Molina, R., Agüero-Chapin, G., Pérez-González, M.P., 2009. TI2BioP (Topological Indices to BioPolymers) *version 1.0*, Molecular Simulation and Drug Design (MSDD). Chemical Bioactives Center, Central University of Las Villas, Cuba.

J.W. Mc Farland, D.J. Gans, Cluster Significance Analysis. In Method and Principles in Medicinal Chemistry, VCH, Weinheim, Germany, 1995.

March, P.E., Ahnn, J., Inouye, M., 1985. The DNA sequence of the gene (rnc) encoding ribonuclease III of *Escherichia coli*. Nucleic Acids Res. 13, 4677–4685.

Marrero-Ponce, Y., Khan, M.T., Casanola Martin, G.M., Ather, A., Sultankhodzhaev, M.N., Torrens, F., Rotondo, R., 2007. Prediction of tyrosinase inhibition activity using atom-based bilinear indices. Chem. Med. Chem. 2, 449–478.

Nair, R., Rost, B., 2008. Protein subcellular localization prediction using artificial intelligence technology. Methods Mol. Biol. 484, 435–463.
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., 1994. Recent investigations into global characteristics of long DNA sequences. Indian J. Biochem. Biophys. 31, 149–155.
Nicholson, A.W., 1997. Ribonucleases: Structures and Functions. Academic Press, Michigan.
Punta, M., Rost, B., 2008. Neural networks predict protein structure and function. Methods Mol. Biol. 458, 203–230.
Roy, S., Martinez, D., Platero, H., Lane, T., Werner-Washburne, M., 2009. Exploiting amino acid composition for predicting protein–protein interactions. PLoS ONE 4, e7813.
Randic, M., Zupan, J., 2004. Highly compact 2D graphical representation of DNA sequences. SAR QSAR Environ. Res. 15, 191–205.
Ripley, B., 1996. Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge, UK.
Randic, M., Vracko, M., 2000. On the similarity of DNA primary sequences. J. Chem. Inf. Comput. Sci. 40, 599–606.
Rumelhart, D.E., McClelland, J.L., 1986. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT Press, Cambridge, MA.
Strope, P.K., Moriyama, E.N., 2007. Simple alignment-free methods for protein classification: a case study from G-protein-coupled receptors. Genomics 89, 602–612.
Selig, C., Wolf, M., Muller, T., Dandekar, T., Schultz, J., 2008. The ITS2 database II: homology modelling RNA structure for molecular systematics. Nucleic Acids Res. 36, D377–D380.
Statsoft, STATISTICA 7.0, 2007. (Data analysis software system for windows), .
Yuan, Z., 1999. Prediction of protein subcellular locations using Markov chain models. FEBS Lett. 451, 23–26.
Zhang, K., Nicholson, A.W., 1997. Regulation of ribonuclease III processing by double-helical sequence antideterminants. Proc. Natl. Acad. Sci. USA 94, 13437–13441.
de Jong, A., van Hijum, S.A., Bijlsma, J.J., Kok, J., Kuipers, O.P., 2006. BAGEL: a web-based bacteriocin genome mining tool. Nucleic Acids Res. 34, W273–W279.

# ANNEX 3

OPEN ACCESS Freely available online

PLoS one

# An Alignment-Free Approach for Eukaryotic ITS2 Annotation and Phylogenetic Inference

**Guillermin Agüero-Chapin[1,2,3☯], Aminael Sánchez-Rodríguez[4☯], Pedro I. Hidalgo-Yanes[2,5], Yunierkis Pérez-Castillo[2], Reinaldo Molina-Ruiz[2], Kathleen Marchal[4], Vítor Vasconcelos[1,3], Agostinho Antunes[1,3]***

**1** CIMAR/CIIMAR, Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Porto, Portugal, **2** Molecular Simulation and Drug Design (CBQ), Universidad Central "Marta Abreu" de Las Villas (UCLV), Santa Clara, Cuba, **3** Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Porto, Portugal, **4** CMPG, Department of Microbial and Molecular Systems, KU Leuven, Leuven, Belgium, **5** Area of Microbiology, University of León, León, Spain

## Abstract

The ITS2 gene class shows a high sequence divergence among its members that have complicated its annotation and its use for reconstructing phylogenies at a higher taxonomical level (beyond species and genus). Several alignment strategies have been implemented to improve the ITS2 annotation quality and its use for phylogenetic inferences. Although, alignment based methods have been exploited to the top of its complexity to tackle both issues, no alignment-free approaches have been able to successfully address both topics. By contrast, the use of simple alignment-free classifiers, like the topological indices (TIs) containing information about the sequence and structure of ITS2, may reveal to be a useful approach for the gene prediction and for assessing the phylogenetic relationships of the ITS2 class in eukaryotes. Thus, we used the **TI2BioP** (**T**opological **I**ndices **to** **BioP**olymers) methodology [1,2], freely available at http://ti2biop.sourceforge.net/ to calculate two different TIs. One class was derived from the ITS2 artificial 2D structures generated from DNA strings and the other from the secondary structure inferred from RNA folding algorithms. Two alignment-free models based on Artificial Neural Networks were developed for the ITS2 class prediction using the two classes of TIs referred above. Both models showed similar performances on the training and the test sets reaching values above 95% in the overall classification. Due to the importance of the ITS2 region for fungi identification, a novel ITS2 genomic sequence was isolated from *Petrakia* sp. This sequence and the test set were used to comparatively evaluate the conventional classification models based on multiple sequence alignments like Hidden Markov based approaches, revealing the success of our models to identify novel ITS2 members. The isolated sequence was assessed using traditional and alignment-free based techniques applied to phylogenetic inference to complement the taxonomy of the *Petrakia* sp. fungal isolate.

**Citation:** Agüero-Chapin G, Sánchez-Rodríguez A, Hidalgo-Yanes PI, Pérez-Castillo Y, Molina-Ruiz R, et al. (2011) An Alignment-Free Approach for Eukaryotic ITS2 Annotation and Phylogenetic Inference. PLoS ONE 6(10): e26638. doi:10.1371/journal.pone.0026638

**Editor:** Jonathan H. Badger, J. Craig Venter Institute, United States of America

**Received** May 16, 2011; **Accepted** September 29, 2011; **Published** October 26, 2011



**Funding:** The authors acknowledge the Portuguese Fundação para a Ciência e a Tecnologia (FCT) for financial support to GACH (SFRH/BD/47256/2008), and the project PTDC/BIA-BDE/69144/2006 and PTDC/AA-AMB/104983/2008. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

* E-mail: aantunes@ciimar.up.pt

☯ These authors contributed equally to this work.

## Introduction

Standard alignment methods are less effective for the functional prediction of gene and protein classes that show a high primary sequence divergence between their members [3]. Thus, the implementation of stochastic models [4], the modification of the original similarity matrixes among the aligned sequences, and the addition of other steps in the alignment procedures [5,6], have been strategies adopted to improve the classification of divergent gene/protein functional classes. On the other hand, several alignment-free methods have been developed as an alternative to traditional alignment algorithms for gene/protein classification at low sequence similarity level [1,7,8].

The internal transcribed spacer 2 (ITS2) eukaryotic gene class is one of the cases showing a higher sequence divergence among its members, which have traditionally complicated ITS2 annotation and limited its use for phylogenetic inference at low taxonomical level analyses (genus and species level classifications). Despite the ITS2 high sequence variability, the ITS2 structure has been considerably conserved among all eukaryotes [9]. This fact has been considered for the implementation of homology-based structure modelling approaches to improve the ITS2 annotation quality and also as a tool for eukaryote phylogenetic analyses at higher classification levels or taxonomic ranks [6,9,10]. Thus, the ITS2 database (http://its2.bioapps.biozentrum.uni-wuerzburg.de) was developed holding information about sequence, structure and taxonomic classification of all ITS2 in GenBank [11]. However, due to ITS2 high sequence variability, the annotation pipeline implemented in the aforementioned resource requires the use of a specific score matrix in the BLAST search [11] and more recently, the use of HMM for the identification and delineation of the ITS2 sequences [10,12]. Although alignment based methods have been exploited to the top of its complexity to tackle the ITS2 annotation and phylogenetic inference [10,11], no alignment-free approach has been able to successfully address these issues so far. The use of simple alignment-free classifiers like the topological indices (TIs) containing also information about the sequence and structure of ITS2 can be another useful approach for the prediction and

phylogenetic analyses of the ITS2 class in eukaryotes. Such TIs are determined by our methodology entitled **T**opological **I**ndices **to BioP**olymers **"TI2BioP"** where the spectral moments are calculated from different graphical approaches representing the structure of the biopolymers: DNA, RNA and proteins [1,2]. TI2BioP is now available at http://ti2biop.sourceforge.net/ as a public tool for the calculation of two different TIs, one class derived from the ITS2 artificial 2D structures generated from DNA strings (Nandy structures) [13,14] and the other class resulting from the secondary structure inferred with RNA folding algorithms (Mfold) [15]. These alignment-free classifiers were used to build linear and Artificial Neural Networks (ANN)-models for classifying the ITS2 members among positive and negative sets and also to estimate the ITS2 phylogeny at higher classification levels.

The ANN-models provided the highest classification accuracy (95.9 and 97.5%) during the training step compared to the linear models for Nandy-like and Mfold structures, respectively. A very similar ANN performance was obtained for the test set for both structural representations. These results support that the identification of gene signatures tend to be better when assessed with non-linear models. We also showed the utility of the artificial secondary structure when the correct 2D structure is not available (i.e. the physiological structure that occurs on the cell) and can only be obtained by predictions based on free energy minimizations.

The performance of our two alignment-free models based on ANN was also compared with several profile Hidden Markov Models (HMMs) generated from alignments performed with CLUSTALW [16], DIALIGN-TX [17] and MAFFT [18] using different training sets, to classify the test set and to identify a new fungal member of the ITS2 class. Moreover, a BLASTn search against NCBI was carried out to give more reliability to the gene annotation and to assess taxonomically related hits to our query fungal sequence. ITS2 is the standard gene target for fungal identification and taxonomy at the species level [19]. This new ITS2 sequence was isolated by our group (GenBank accession number FJ892749) from an endophytic fungus belonging to the genus *Petrakia*. Members of this fungal genus have been hard to be placed taxonomically and are potential producers of bioactive compounds [20]. The *Petrakia sp.* strain was morphologically identified and its ITS2 sequence was used to carry out traditional and alignment-free phylogenetic analyses to support its taxonomic characterization.

The alignment-free models identified the new query sequence as a member of the ITS2 class with high significance, while the profile HMMs showed a poor performance in doing so. We demonstrated that our TIs are useful not only in sequence identification but also in molecular evolutionary inferences. The alignment-free tree built based on TIs provided similar phylogenetic relationships among the different classes of the Ascomycota phylum in respect to the traditional phylogenetic analysis (i.e. based on evolutionary distances derived from a multiple alignment of DNA sequences). Both analyses placed the *Petrakia* genus inside the *Pezizomycotina* subphylum and the *Dothideomycetes* class.

## Methods

### 1. Computational methods. Topological Indices to BioPolymers (TI2BioP)

TI2BioP allows the calculation of the spectral moments derived from inferred and artificial 2D structures of DNA, RNA and proteins [21]. Consequently, it is feasible to carry out a structure-function correlation using such sequence/structure numerical indices. The calculation of the spectral moments as sequence descriptors is performed according to the TOPS-MODE approach [22] implemented in the "MODESLAB" software [23] and the draw mode for sequence representation was retrieved from the MARCH-INSIDE methodology [24,25,26]. TI2BioP can also import files containing 2D structure inferred by other professional softwares like the RNASTRUCTURE [15]. We propose for the first time to fold the ITS2 genomic sequences into an artificial secondary structure based on Nandy's representation for DNA strings [13]. This graph groups purine and pyrimidine bases on a Cartesian system assigning to X and Y axes each nucleotide-type, respectively. The representation was carried out by adding to the coordinates (0, 0) of the Cartesian system the k-th nucleotide of the DNA sequence. The value (1, 0) if the (k+1)-th nucleotide is Guanine (rightwards-step); (−1, 0) if Adenine (leftwards-step); (0, 1) if Cytosine (upwards-step) or (0, −1) if the (k+1)-th nucleotide is Thymine or Uracil (downwards-step).

**Figure 1** depicts the 2D Cartesian representation of the 558 bp genomic DNA fragment from *Petrakia sp.* ef08-038 (accession number FJ892749) comprising the ITS2 with its boundaries (**fig. 1A**) and only the ITS2 (**fig. 1B**). The figure also shows the ITS2 sequence (without its boundaries) folded as DNA (**fig. 1C**) and RNA (**fig. 1D**) by the Mfold program.

In the study, a total of 4,355 out of the original 5,092 ITS2 sequences from a wide variety of eukaryotic taxa (http://its2.bioapps.biozentrum.uni-wuerzburg.de) shared similar secondary structure features and were used as positive dataset.

The negative set or control group comprises diverse but structurally related genomic sequences to the ITS2 class: the untranslated regions (UTRs) of eukaryotic mRNAs. They are non-coding regions with divergence among the eukaryotes but showing a more conserved secondary structure when are transcribed into RNAs [27]. A non-redundant subset containing 6,529 and 8,128 of the 5′- and 3′-UTRs' sequences from the fungi kingdom, respectively, was selected from the eukaryotic mRNAs database: UTRdb (http://www.ba.itb.cnr.it/UTR/). The sequence diversity among the ITS2 and UTRs datasets was explored comparatively using the Needleman-Wunsch (NW) [28] and Smith-Waterman (SW) [29] algorithms. See in supporting information (S) the NW & SW analyses (**File S1 and figure S1**).

Training and test series were randomly selected. The members of the test set were selected taking out at random the 20% of the overall data (19,012 cases). The remainder of the cases was used to train the model. Sequences with ambiguity at least in one nucleotide position were removed from both groups. Each ITS2 and UTR sequence retrieved was labeled respecting its original database ID code; see File S2.

All sequences (positive and negative sets) were pseudo-folded into a Cartesian system by TI2BioP to obtain the artificial secondary structures as it was described above. On the other hand, they were also used to infer optimized DNA secondary structures by the Mfold algorithm implemented in the RNASTRUCTURE 4.0 software [30] (**fig. 1C**). The structural optimization is based on the minimization of the folding energy (lowest $\Delta G$). Spectral moments ($\mu_k$) introduced previously by Estrada et al. (1996) [31,32] were applied to codify the new structural information contained into the artificial secondary structures and into the inferred secondary structures of both the ITS2 and UTRs sequences.

**1.1. Calculation of TIs irrespective of sequence similarity.** The topological indices called "spectral moments" were calculated as the sum of the entries placed in the main diagonal of the bond adjacency matrix (**B**) for the DNA/RNA sequences. **B** is a square matrix of n×n row and column where its non-diagonal entries are ones or zeroes if the corresponding bonds

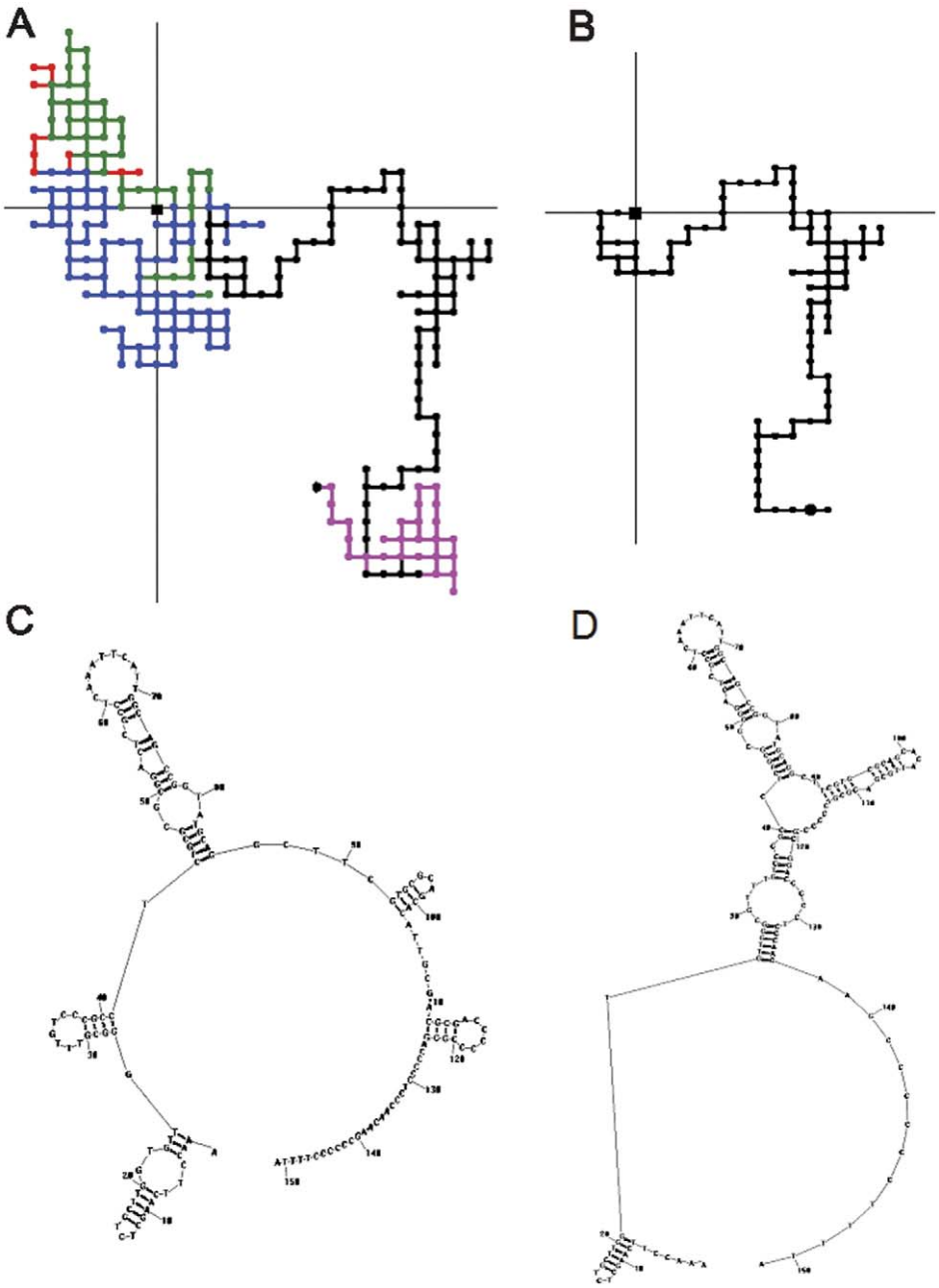



**Figure 1. The ITS2 region (in black) with its boundaries ordered 5′upstream: a short end corresponding to the 18S rDNA (in red), the ITS1 (in green), the 5.8S rDNA and 3′downstream: a short fragment of the 28S rDNA (in pink) (A).** The ITS2 region pseudo-folded into the 2D-Cartesian system (B). The ITS2 sequence folded as a DNA and RNA structure by the Mfold program, respectively (C and D).
doi:10.1371/journal.pone.0026638.g001

or edges share or not one nucleotide. Thus, it set up connectivity relationships between the nucleotides in certain DNA/RNA graph. The different powers of **B** give the spectral moments of higher order.

In the DNA/RNA artificial secondary structure, 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 bonds in the nucleotide sequence. The main diagonal entries of **B** were weighted with the average of the electrostatic charge (Q) between two bound nodes. The charge value q in a node is equal to the sum of the charges of all nucleotide placed on it. The electrostatic charge of one nucleotide was derived from the Amber 95 force field [33]. Thus, it is easy to carry out the calculation of the spectral moments of **B** in order to numerically characterize the pseudo-folding ($^{pf}\boldsymbol{\mu}_k$) of DNA/RNA sequences.

$$^{pf}\mu_k = Tr\left[(B)^k\right] \quad (1)$$

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

In order to illustrate the calculation of the spectral moments, an example is developed below. The 2D Cartesian network of the

sequence (AGCTG) is showed in the figure 2D and its bond adjacency matrix is depicted in the figure 2C; note that the central node contains both Guanine and Thymine nucleotides. The calculation of the spectral moments up to the order k = 3 is also defined below on the **figure 2**. The q values are represented in the matrix as the nucleotides symbols (G = 0.24, A = 0.22, C = 0.19, T and U = 0.21).

Expansion of expression (1) for k = 1 gives the ${}^{pf}\mu_1$, for k = 2 the ${}^{pf}\mu_2$ and for k = 3 the ${}^{pf}\mu_3$. The calculation of the spectral moments up to order three from this DNA graph is described below.

$$ {}^{pf}\mu_1 = Tr[B] = Tr\left(\begin{bmatrix} 0.335 & 1 & 1 \\ 1 & 0.345 & 1 \\ 1 & 1 & 0.320 \end{bmatrix}\right) = 1.0 \qquad (1a) $$

$$ \begin{aligned} {}^{pf}\mu_2 &= Tr\left[(\mathrm{B})^2\right] \\ &= Tr\left(\begin{bmatrix} 0.335 & 1 & 1 \\ 1 & 0.345 & 1 \\ 1 & 1 & 0.320 \end{bmatrix} \times \begin{bmatrix} 0.335 & 1 & 1 \\ 1 & 0.345 & 1 \\ 1 & 1 & 0.320 \end{bmatrix}\right) \qquad (1b) \\ &= (2.11)^2 + (2.12)^2 + (2.10)^2 \end{aligned} $$

$$ \begin{aligned} {}^{pf}\mu_3 &= Tr\left[(\mathrm{B})^3\right] \\ &= Tr\left(\begin{bmatrix} 0.335 & 1 & 1 \\ 1 & 0.345 & 1 \\ 1 & 1 & 0.320 \end{bmatrix}\right)^3 \qquad (1c) \\ &= (2.038)^3 + (2.041)^3 + (2.033)^3 \end{aligned} $$

**TI2BioP** *version 1.0* ® arrange automatically the DNA/RNA sequences into a 2D Cartesian network [21] and also import the connectivity table (ct files) generated by the RNASTRUCTURE 4.0 software. Ct files contain information about the connection between nucleotides in the secondary structure generated with thermodynamic models [30]. Thus, it is possible to perform the calculation of the spectral moments (${}^{mf}\boldsymbol{\mu}_k$) based on folding thermodynamics parameters for the positive and negative sets. Another two additional TIs defined as Edge Numbers and Edge Connectivity were introduced for these two DNA/RNA structural approaches; see File S2 for more details.

## 2. Building up alignment free-models with TIs

**2.1. Variable screening.** We used the *Feature Selection and Variable Screening* module of the *Data Mining* menu from *STATISTICA* software [34] to select a subset of predictors that is most strongly related to the dependent (outcome) variable of interest regardless of whether that relationship is simple (linear) or complex (nonlinear). The algorithm for selecting those variables is not biased in favor of a single method for subsequent analyses; further post-processing algorithms were applied, based on linear and non-linear modeling methods.

**2.2. Alignment-free models for ITS2 classification. Linear models.** The General Discrimination Analysis (GDA) was carried out for building up linear models for ITS2 alignment-free identification [35,36,37,38]. The most significant predictors obtained from the variable screening method for each structural approach were used to fit linear discriminant functions. Both subsets of TIs were standardized in order to become equally scaled to allow an effective comparison between the regression coefficients [39]. The model performance was evaluated by several statistical measures: accuracy, area under the Receiver Operating Characteristic (ROC) curve, commonly known as AUC with a value of 1.0 for a perfect predictor and 0.5 for a random predictor and the F-score (it reaches its best value at 1 and worst score at 0) [40].

**2.3. Alignment-free models for ITS2 classification. Non-linear models. Artificial Neural Networks (ANN).** We used ANN method for ITS2 classification using the same series of TIs as

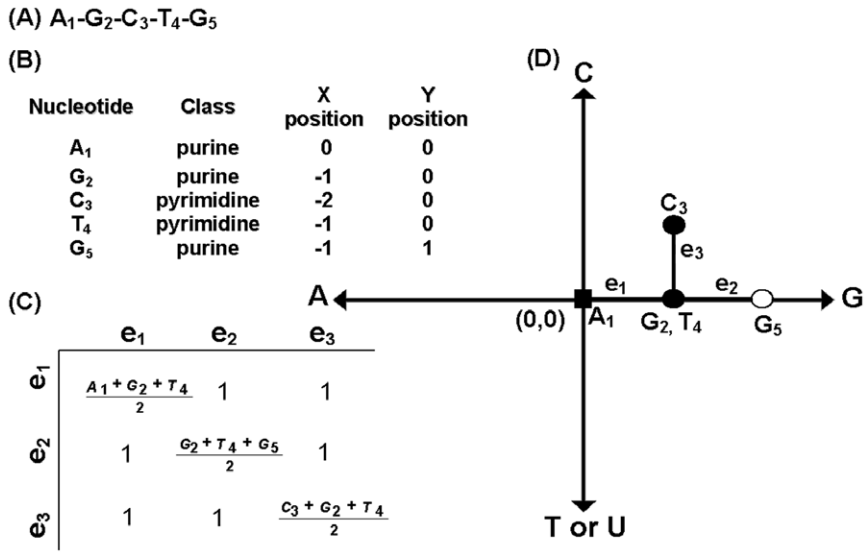



**Figure 2. Building the 2D-Cartesian map for the (A) DNA fragment AGCTG.** (B) The coordinates for each nucleotide in the Cartesian system. (C) The definition of the bond adjacency matrix derived from (D) the 2D-Cartesian map. Note that all edges of the graph are adjacent, thus all non-diagonal entries are ones.
doi:10.1371/journal.pone.0026638.g002

input variables and only one output variable (ITS2 membership). We used the Multilayer Layer Perceptron (MLP) due to its ability to model functions of almost arbitrary complexity showing a simple interpretation as a form of input-output model. To select the right complexity of the network, we tested different topologies to the MLP while checking the progress against a selection set to avoid over-fitting during the two-phase (back propagation/ conjugate gradient descent) training algorithm [41]. The selection set was extracted at random from the training set (10%) by also generating random numbers. The test set was the same used for GDA representing an external subset (not used during training algorithms) to check the final network performance.

The optimal cutoff for ITS2 gene classification for ANN-models was defined by determining on the ROC-curve the model's parameter values ('accept' and 'reject' classification thresholds) giving the nearest point (optimal operating point) to the (0,1) coordinates. This point constitutes the ideal condition for ITS2 classification (most balanced solution where both specificity and sensitivity are maximized). The optimal operating point was determined by computing the slope S that considers the misclassification costs for each class. The point was found by moving the straight line with slope S from the upper left corner of the ROC plot (0, 1) down and to the right until it intersects the ROC curve.

## 3. Alignment-based models for ITS2 classification. Profile Hidden Markov Models (HMM)

Three training subsets were selected to build up several profile HMMs for ITS2 gene classification: (i) 134 sequences extracted representatively from the original training set (2802 ITS2 sequences) to represent evenly the whole range of sequence similarity while retaining representative members from all the eukaryotic taxa within the training set (this sampling was based on the sequence similarity clustering carried out in File S1); (ii) 80 sequences representative of the fungal kingdom selected following a similar procedure as described in (i); and (iii) 2802 ITS2 sequences used to train the alignment-free models. In addition, three different multiple sequence alignments (MSA) algorithms were used to align these subsets: CLUSTALW [16], DIALIGN-TX [17] and MAFFT [18]. Due to the low similarity level amongst the ITS2 sequences, we have used DALIGN-TX and MAFFT that are expected to outperform CLUSTALW in such conditions. DALIGN-TX is a segment-based multiple alignment tool improved for sets of low overall sequence similarity and the MAFFT program is able to identify homologous regions among distantly related sequences. Performing a good alignment is a crucial step to generate a profile HMM with high classification power.

CLUSTALW and DIALIGN-TX were run using the default parameters. In the case of MAFFT the iterative alignment option (L-INS-I) was used [29,42].

Alignments were edited in every case as follows: aligned positions were removed from both ends until gaps were observed in less than 10% of the aligned sequences. Thus, we removed non-informative positions from the multiple alignments that could deteriorate the resulting HMM. Edited alignments were used as input for *hmmbuild* release 2.3.2 [43], which generated the profile HMMs. During the profile HMMs generation step the *fast* option of the *hmmbuild* program was used with a default value equal to 0.5. This option assigns the *insert* state to every column in the alignment containing gaps in at least half of the sequences. In this way, the resulting HMMs do not make an explicit use of the sequence distribution (i.e. nucleotides frequencies) of positions with high amount of gaps but rather consider them as insertion states.

The obtained profile HMMs allowed to classify members of the test set, as well as the newly isolated ITS2 sequence from *Petrakia* sp. (see below) using *hmmsearch*. An optimal cutoff for the ITS2 classification was determined by running each profile HMM at 20 different E-values (0.1–10). The E-value that maximizes both sensitivity and specificity was selected as the optimal classification cutoff. The performance of these models at the optimal classification cutoff was compared to that of the alignment-free models described above (sections 2.2.2 and 2.2.3).

## 4. Phylogenetic analyses

We defined an empirical threshold of ITS2 representatives with more than 60% of sequence similarity with our query fungus (Petrakia sp. ef08-038) among the members of the Ascomycota phylum for the phylogenetic analysis. This allowed the retrieval of an ITS2 subset comprising 16 sequences that encompassed several classes from the subphyla Pezizomycotina (Dothideomycetes, Lecanoromycetes, Leotiomycetes and Sordariomycetes), while the remaining cases were either taxonomically characterized as mitosporic Ascomycotas (asexual species that produce conidia namely mitospores) or unclassified Ascomycotas. The 16 ITS2 sequences plus our query sequence (FJ892749) were aligned with the CLUSTAL W setting a Gap Open Penalty (GOP) of 20 and a Gap Extension Penalty (GEP) of 10. The final alignment was edited removing end gaps and the phylogenetic analyses were conducted in MEGA4 software [19]. Neighbour-joining (NJ) trees were generated from different sequence distance matrices from (1) alignment and (2) alignment-free approaches:

1. NJ trees based on different evolutionary distances computed using Jukes-Cantor (JC), Kimura 2-parameter (K2P) and Maximum Composite Likelihood (MCL) substitution models were obtained using the MEGA4. In addition, the Minimum Evolution (ME) method was assessed on the JC and K2P distance matrices. The bootstrap support (BS) values for nodes were computed from 1000 replicates.
2. A NJ tree was built based on the hierarchic clustering that uses the Euclidean distance matrix as a multidimensional measure to form the sequences clusters. Euclidean distance (Ed) was computed from the TIs values of the same seventeen ITS2 sequences mentioned above and the complete linkage or furthest neighbor was used as cluster method.

$$\text{Euclidean distance } (x,y) = \left\{ \sum_i (x_i - y_i)^2 \right\}^{1/2} \qquad (2)$$

The quality of this numerical taxonomy was tested (i) performing the Joining Tree Clustering with different distance metrics (City-block, Chebychev, and Power distance), (ii) using other cluster methods (Single linkage, Unweighted pair-group average and the Ward's method), and (iii) calculating the cophenetic correlation coefficient.

## 5. Experimental section

*Petrakia* strain was isolated from leaves of *Acer psedoplatanus*. The plant material was collected in Kaiserslautern, Germany. It was cut and surface-sterilized by immersion in 70% ethanol for 1 min, 5% NaOCl for 3 min and 70% ethanol for 1 sec followed by a wash in sterile distilled water. Samples were then cut into small fragments and plated onto 2% malt agar with penicillin G and

streptomycin sulfate (each 200 mg/l). The mycelial culture was deposited in the culture collection of the Institute of Biotechnology and Drug Research (IBWF), Kaiserslautern.

DNA extraction was performed as described previously by Sacks [44]. The entire ITS (ITS1, 5.8S rDNA, and ITS2) region was amplified for ITS sequence analysis. The primers used for amplification were ITS5 (5′-GGAAGTAAAAGTCGTAACA-AGG) and ITS4 (5′- TCCTCCGCTTATTGATATGC) according to White et al. [45]. Their method was used with slight modifications: A GeneAmp PCR System 9700 was employed (Applied Biosystem, Foster City, CA, USA). The PCR amplification cycle consisted of 30 s at 94°C, 1 min at 50°C, and 1 min at 72°C. PCR products were sequenced by MWG Biotech (Ebersberg, Germany) with the same primers used for the amplification. Each sequence was obtained in duplicate from each of two separate PCR amplifications.

## Results and Discussion

### 6. Predicting eukaryotic ITS2 sequences with alignment-free classifiers

Two classes of predictors comprising 18 TIs each were calculated by the TI2BioP methodology for 19,012 genomic sequences (4,355 ITS2 and 14,657 UTRs): the spectral moments series ($\boldsymbol{\mu}_0$- $\boldsymbol{\mu}_{15}$) of the bond adjacency matrix between the nucleotides arranged into the Cartesian space ($^{pf}\boldsymbol{\mu}_k$) and between the nucleotides connected into the Mfold structures ($^{mf}\boldsymbol{\mu}_k$). Other two additional TIs were computed (the Edge Numbers and the Edge Connectivity) for each class. The spectral moments are structural-based TIs that describe electronically the nucleotide connectivity at different orders in these two structural approaches. The Nandy-like structure is determined by the sequence order and DNA/RNA nucleotide composition. The 2D structure obtained by the Mfold software depends also of the primary sequence but its folding is driven by the optimization of thermodynamics parameters (lowest folding free energy-$\Delta G^0$).

In order to select the most significant predictors for both datasets (Nandy-like and Mfold structures), we carried out a feature selection as a preliminary variable screening method before the model building. We found that the four most significant variables ($p<0.01$) were the Edge Connectivity, the $^{pf}\mu_0$, $^{pf}\mu_1$, and $^{pf}\mu_2$ for Nandy's structures and for Mfold structures the $^{mf}\mu_0$, $^{mf}\mu_5$, $^{mf}\mu_7$ and $^{mf}\mu_{15}$ (**figure 3**).

These two sets of four variables were used as input predictors to build classification linear models based on the GDA implemented in the *STATISTICA* software [34]. The alignment-free classifiers based on Nandy-like structures provided classification accuracy in training and test of 84.87 and 84.95%, respectively. The AUC and F-score for the test set were of 0.919 and 0.687, respectively. In contrast, the TIs derived from the Mfold structures showed a better classification performance. Its accuracy level was notably higher in training (94.17%) and in the test subset (94.26%). The same was true for the AUC and F-score statistics that reach values of 0.983 and 0.960, respectively. These facts point out that the TIs calculated from the 2D topology predicted by folding thermodynamics rules are more effective classifiers than the TIs derived from artificial structures. However, the former takes much more computational and procedure cost than for the TIs obtained from the Cartesian graphical approach. The 2D Cartesian TIs have been useful for protein and RNA structure descriptors when higher structural levels are not available [46,47,48]. Thus, we evaluate non-linear methods on both data sets with the aim to improve the classification performance, especially for the pseudo-folding TIs. The Artificial Neural Networks (ANN), particularly the Multilayer Layer Perceptron (MLP) was selected as the most popular ANN architecture in use today [49].

**6.1 Artificial Neural Networks (ANN) in the prediction of the ITS2 class.** The MLP was tested at different topologies using the four predictors already selected for each secondary structural approach as input variables. From the same training set used to develop the discriminant function, an independent data set (the selection set) was selected. This subset was chosen randomly taking out the 20% of the training set being not used in the back propagation algorithm. Thus, 12,168 cases were used for the training, 3,042 represented the selection subset and the 3,802 cases were evaluated in external validation to set the comparison.

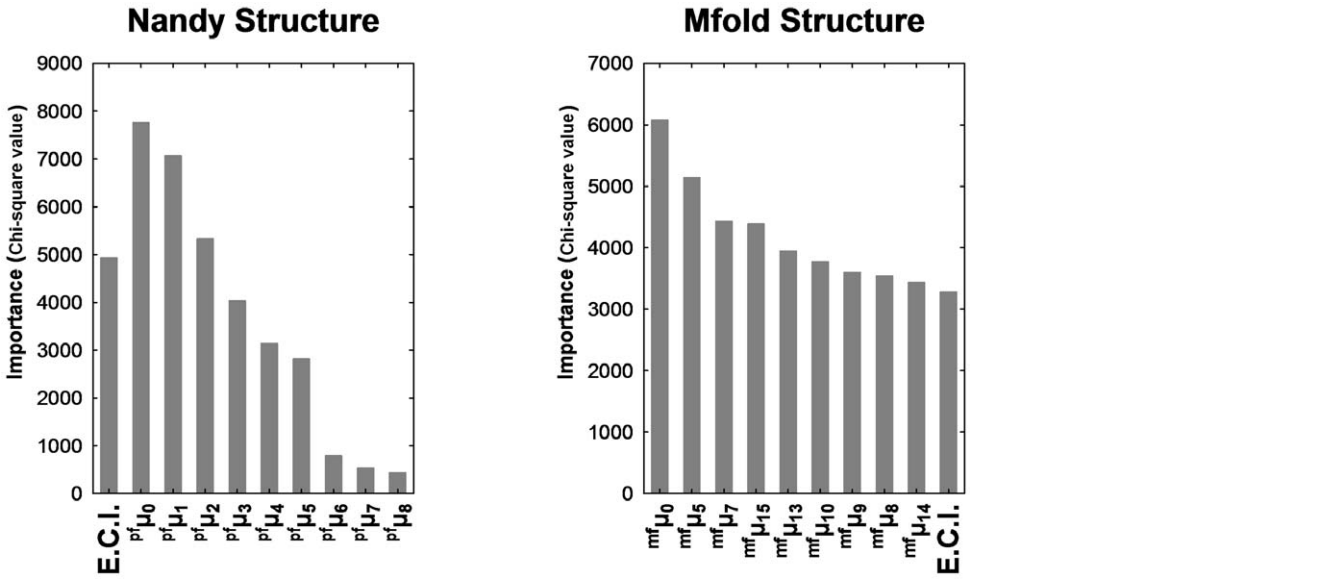



**Figure 3. Predictor importance according the variable screening analysis for the Nandy and Mfold structures.** E.C.I. (Edege Connectivity Index).
doi:10.1371/journal.pone.0026638.g003

The **Table 1** shows the different MLP topologies used to select the right complexity of the ANN in both datasets, the performance on training, selection and test progress were examined as well as its errors. The best models were the MLP profiles number 3 and 1 (highlighted in bold) for Nandy and Mfold datasets, respectively, which showed the best accuracy on training, selection and test sets, minimizing its respective errors. These ANN-models showed a higher accuracy level in classifying the training and test sets in respect to the linear models. The TIs calculated from the Mfold structures provided a better ANN performance on the data classification than when derived from the Nandy graphical approach. Although, ANN-based models showed an analogue behaviour in respect to the linear models (Mfold > Nandy); the classification performances of both structural approaches are more similar and higher when a non-linear function is applied (**Table 1**). This suggests that the identification of gene signatures tend to be better assessed with non-linear models and we further showed the utility of the artificial but informative folding of the biopolymeric sequences for gene/protein class identification [24,50,51].

The classification results derived from our two best alignment-free approaches to classify ITS2 membership is showed in **Table 2 and File S3**. The structural TIs based on the folding thermodynamics rules provide a more accurate description of the DNA/RNA structure, which is supported by the classification results (**Table 2**). The 2D topology of these molecules is affected by the primary information and by the possible hydrogen interactions between nucleotides forming the stems and loops; therefore a better functional classification performance is achieved. Although the Nandy-like representation is less accurate in the classification due to its artificial nature, it takes into account the sequence order information and the nucleotide composition, which are important features for the recognition at a genome scale of genes that do not encode a protein [52,53]. Thus, the utility of this easy structural approach is reflected in the excellent discrimination achieved between these two distinct DNA/RNA functional classes with divergence among its members but sharing common structural features.

We carried out a 10-fold cross validation to examine the classification performance of our alignment-free models. This validation procedure is easier to implement and provides reliable results in the validation of a predictive model at low computational cost [54]. Thus, the original data set was divided at random into 10 subsets containing the same number of cases. Of the 10 subsets, a single subset was retained as a prediction subsample for testing the model, and the remaining nine subsets were used as the training data. Since a selection subset is also needed to check the training algorithm, it was selected from the training set at random (10%). The cross-validation procedure is then repeated 10 folds or rounds using each of the 10 subsets for prediction exactly once, in such way ensures that all cases were predicted and used in training. Afterwards the average values for the accuracy, sensitivity, specificity for training and test sets, as well as the AUC were calculated to provide a single estimation from the 10 folds (**Table 2**).

We plotted the ROC curve for each fold from the cross-validation procedure on the test set. In each fold or round, the curve presented an area higher than 0.5 (**figure 4**). According to the ROC curve theory random classifiers have an area of only 0.5. This result confirms that the present model is a significant classifier relatively to those working at random. In the plotting, the ROC curves for the ANN-models (MLP-1 and 3) on the test set were included to show visually its classification performance similarity

**Table 1.** Testing different topologies for the MLP on the ITS2 classification using TIs from Nandy and Mfold DNA structures.

| | Profile | Train Accuracy | Selection Accuracy | Test Accuracy | Train Error | Selection Error | Test Error |
|---|---|---|---|---|---|---|---|
| **Nandy structure** | | | | | | | |
| **1** | MLP 4:4-4-1:1 | 0.946 | 0.948 | 0.946 | 0.232 | 0.226 | 0.230 |
| **2** | MLP 4:4-3-1-1:1 | 0.946 | 0.949 | 0.945 | 0.225 | 0.219 | 0.224 |
| **3** | **MLP 4:4-2-2-1:1** | **0.959** | **0.958** | **0.956** | **0.178** | **0.180** | **0.187** |
| **4** | MLP 4:4-1-3-1:1 | 0.949 | 0.950 | 0.948 | 0.199 | 0.198 | 0.200 |
| **5** | MLP 4:4-3-1:1 | 0.946 | 0.948 | 0.946 | 0.232 | 0.226 | 0.230 |
| **6** | MLP 4:4-2-1-1:1 | 0.772 | 0.769 | 0.768 | 0.419 | 0.422 | 0.422 |
| **7** | MLP 4:4-1-2-1:1 | 0.946 | 0.949 | 0.945 | 0.216 | 0.210 | 0.215 |
| **8** | MLP 4:4-2-1:1 | 0.946 | 0.948 | 0.946 | 0.232 | 0.225 | 0.230 |
| **9** | MLP 4:4-1-1:1 | 0.946 | 0.949 | 0.945 | 0.233 | 0.226 | 0.231 |
| **Mfold structure** | | | | | | | |
| **1** | **MLP 4:4-4-1:1** | **0.976** | **0.975** | **0.973** | **0.140** | **0.138** | **0.145** |
| **2** | MLP 4:4-3-1-1:1 | 0.968 | 0.968 | 0.967 | 0.158 | 0.155 | 0.162 |
| **3** | MLP 4:4-2-2-1:1 | 0.942 | 0.954 | 0.943 | 0.207 | 0.196 | 0.204 |
| **4** | MLP 4:4-1-3-1:1 | 0.941 | 0.955 | 0.943 | 0.206 | 0.194 | 0.203 |
| **5** | MLP 4:4-3-1:1 | 0.969 | 0.970 | 0.967 | 0.159 | 0.155 | 0.162 |
| **6** | MLP 4:4-2-1-1:1 | 0.957 | 0.961 | 0.960 | 0.176 | 0.170 | 0.172 |
| **7** | MLP 4:4-1-2-1:1 | 0.943 | 0.955 | 0.944 | 0.205 | 0.193 | 0.202 |
| **8** | MLP 4:4-2-1:1 | 0.943 | 0.956 | 0.944 | 0.205 | 0.193 | 0.202 |
| **9** | MLP 4:4-1-1:1 | 0.941 | 0.940 | 0.945 | 0.209 | 0.211 | 0.199 |

*Accuracy and error rates on training, selection and test sets.*
doi:10.1371/journal.pone.0026638.t001

**Table 2.** Classification results derived from the ANN-models (MLP-3 and 1) for Nandy and Mfold structures respectively in training, selection and test series.

| **Nandy structure** | **Training** | | | **Selection** | | **Test** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **ITS2** | | **CG** | **ITS2** | **CG** | **ITS2** | | **CG** | |
| ITS2 class | **2434** | | 128 | **575** | 31 | **770** | | 38 | |
| Control Group (CG) | 368 | | **9238** | 87 | **2349** | 121 | | **2863** | |
| Total | 2802 | | 9366 | 662 | 2380 | 891 | | 2911 | |
| Sensitivity (Sv) (%) | 86.86 | | | 86.85 | | 86.42 | | | |
| Specificity (Sp) (%) | 98.63 | | | 98.70 | | 98.35 | | | |
| Accuracy (Acc) (%) | 95.95 | | | 96.12 | | 95.58 | | | |
| AUC | 0.984 | | | 0.985 | | 0.980 | | | |
| F-score | | | | | | 0.939 | | | |
| **10-fold CV** | **Sv** | **Sp** | **Acc** | | | **Sv** | **Sp** | **Acc** | **AUC** |
| Average | 84.79 | 98.85 | 95.64 | | | 84.59 | 98.87 | 95.59 | 0.978 |
| **Mfold structure** | **Training** | | | **Selection** | | **Test** | | | |
| | **ITS2** | | **CG** | **ITS2** | **CG** | **ITS2** | | **CG** | |
| ITS2 class | **2592** | | 102 | **604** | 19 | **825** | | 35 | |
| Control Group (CG) | 210 | | **9264** | 58 | **2361** | 66 | | **2876** | |
| Total | 2802 | | 9366 | 662 | 2380 | 891 | | 2911 | |
| Sensitivity(Sv) (%) | 92.50 | | | 91.24 | | 92.59 | | | |
| Specificity (%) | 98.91 | | | 99.20 | | 98.79 | | | |
| Accuracy (%) | 97.57 | | | 97.53 | | 97.31 | | | |
| AUC | 0.994 | | | 0.995 | | 0.994 | | | |
| F-score | | | | | | 0.960 | | | |
| **10-fold CV** | **Sv** | **Sp** | **Acc** | | | **Sv** | **Sp** | **Acc** | **AUC** |
| Average | 92.37 | 99.01 | 97.50 | | | 92.26 | 98.97 | 97.44 | 0.993 |

10-folds Cross Validation (CV) procedure on training and test sets.
*Numbers in bold highlight well-classified cases.*
doi:10.1371/journal.pone.0026638.t002

with the 10-fold cross validation (**figure 4**). Thus, the similarity in the prediction performance between the 10-fold cross validation procedure and the reported ANN-models shows the robustness of our models. The validity of this type of procedures in structure-function relationship studies based on ANN-models has been demonstrated before [55,56,57].

We found an optimum cutoff for ITS2 gene classification using an "acceptance" threshold of 0.475 that provides a sensitivity of 0.929 and a specificity of 0.986 for our best predictive model (based on M-fold' TIs). Moreover, for the other alignment-free model that used Nandy-like's TIs, the "acceptance" classification threshold was 0.529 showing a sensitivity of 0.838 and a specificity of 0.988.

Although ANN-based models are more complex than linear functions, the architecture of these networks is rather simple since they use just four predictors and one hidden layer made up of four neurons for the case of the TIs calculated from Mfold structures and two layers with the same amount of neurons for the Nandy structural approach (**figure 5**). Thus, the ANN-models based on the TI2BioP methodology are effective and simple tools to search an ITS2 sequences among the diversity of this DNA/RNA class in a wide variety of eukaryotic taxa.

## 7. Hidden Markov Models in the classification of the ITS2 class. A comparative study

Hidden Markov Models (HMM) has been widely used for classification purposes of DNA and protein sequences [58]. Their simplicity and high performance have made them the core of the popular database Pfam [4]. Profile HMMs generates predictive models in which classification performance can be easily evaluated in terms of accuracy, sensitivity and specificity. Nine profile HMMs from members of the ITS2 class were built up using three MSA algorithms (CLUSTALW, DIALIGN-TX and MAFFT) with different training sets. The classification measures for both the profile HMMs and the alignment-free models are shown in **Table 3**.

As shown in **Table 3**, all the profile HMMs obtained for the ITS2 classification provide a lower performance in respect to the alignment-free approaches. Nevertheless, we obtained generally some improvements in the sensitivity on the ITS2 classification when the E-value cutoff was increased (**File S6**) and when the profile HMMs based on improved MSA algorithms was applied. The use of a wider training set comprising 2802 ITS2 sequences also improved the classification performance for the profile HMMs based on DIALIGN-TX and MAFFT algorithms since this dataset better captures the vast diversity of the ITS2 class. However, the ITS2 query sequence from *Petrakia* sp. was identified with a higher significance level when a fungi-specific dataset aligned with MAFFT was considered for building the models (**Table 3**).

We provide information about the MSA handled with CLUSTALW, DIALIGN-TX and MAFFT (**File S4**) and the ITS2 profile HMMs generated with the aforementioned MSA algorithms on the three training sets described in section 2.3 (**File S5**).

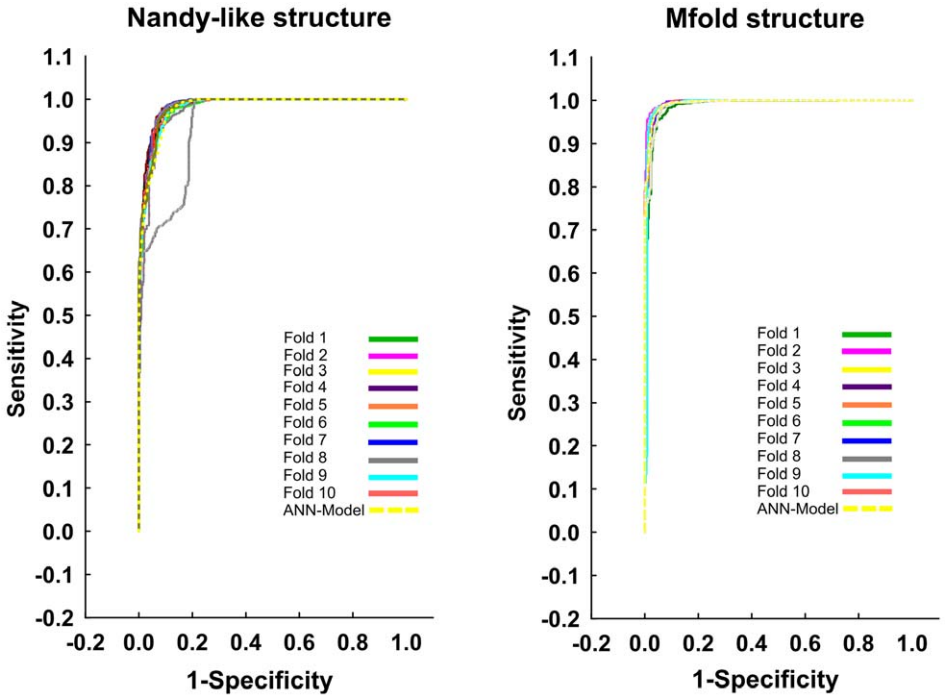



**Figure 4. ROC-curves for the 10-fold cross validation procedure of both ANN-models (Nandy and Mfold structures) on the test set.** The curve for the reported model in each case is represented by a yellow discontinuous line.
doi:10.1371/journal.pone.0026638.g004

We explain the low performance of the profile HMMs on the poorly informative multiple alignments used for its creation. Neither the use of a specific nor of an extended training set aligned with an improved MSA (e.g. MAFFT) assures a good classification; the maximum sensitivity obtained on the test set was only 66.66% (**Table 3**). This result is in line with the one previously obtained by developers of the ITS2 database [10], which reported the use of more conserved 5.8S and 28S rRNAs adjacent to the ITS2 in order to obtain an useful profile HMM. All together, these results reinforce the usability of our alignment-free models that additionally require less sequence information compared to classical alignment-based approaches.

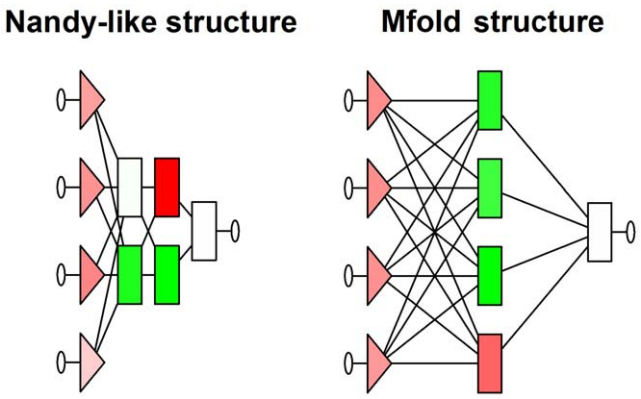



**Figure 5. The architecture of the ANN-models (MLP-3 and MLP-1) for Nandy and Mfold structures, respectively.** It represents four input variables, four neurons in two layers (Nandy) and four in one layer (Mfold) and only one output variable (from the left to the right).
doi:10.1371/journal.pone.0026638.g005

As a practical validation, a novel ITS2 genomic sequence was isolated from a fungal isolate as a part of its taxonomic characterization. This ITS2 sequence was used to evaluate the ability of the ANN-models and the profile HMMs to identify a novel member of this gene class and also its use into the traditional and alignment-free phylogenetic assessment.

## 8. Experimental results. Annotation of a novel ITS2 member using several predictive models

We selected the fungal genus *Petrakia* that lives inside plants of the *genus Acer*, which can be a latent pathogen agent of these plants and a potentially producer of bioactive compounds [59]. Members of the *Petrakia* genus are placed inside the Ascomycota phylum despite the absence of a defined ascus (a microscopic sexual structure in which nonmotile spores, called ascospores, are formed). These fungi that produce conidia (mitospores) instead of ascospores were previously described as mitosporic Ascomycota [53]. However, its taxonomy identification has been a problem at the species level. Thus, a polyphasic approach involving mycological culture with molecular detection [60] to determine the presence of fungi in plants is needed.

Our fungal isolate showed all morphological characteristics of a mitosporic Ascomycota/ genus *Petrakia* such as: aerial mycelium, cover entire plate of Malt Extract Agar medium, conidiophores forming dark sporodochium, conidia pigmented, many-celled, muriform, with several cylindrical projections [61] (**figure 6A**). However, the species could not be unequivocally determined and therefore an attempt to perform a low level-phylogenetic analysis supported on the ITS2 biomarker was required to complement the fungus detection.

We isolated a genomic DNA fragment of 558 bp comprising the entire (ITS1, 5.8S rDNA, and ITS2) region with shorts ends at 5′and 3′positions corresponding to the 18S and 28S rDNA conserved genes, respectively (**figure 6B**). The PCR product was

**Table 3.** Comparative analysis for the classification performance on the test set and *Petrakia sp.* ITS2 sequence using nine profile-HMMs built up with CLUSTALW, DALIGN-TX and MAFFT algorithms with different training sets.

| **ALIGNMENT BASED MODELS** | | | | |
|---|---|---|---|---|
| **Training set (source and number of sequences)** | **Sequence Alignment (processing) Method** | **Optimal Classification Cutoff (E-value)** | **Sensitivity/Specificity (%)** | **Prediction on the ITS2 *Petrakia* sp.*** |
| Representative fungi (80 sequences) | CLUSTALW | 2.0 | 15.82/100 | No significant hit |
| | DALIGN-TX | 9.0 | 18.18/100 | No significant hit |
| | MAFFT | 5.0 | 20.20/100 | 0.02 |
| Representative eukaryotes (134 sequences) | CLUSTALW | 2.0 | 13.92/100 | No significant hit |
| | DALIGN-TX | 0.1 | 6.95/100 | No significant hit |
| | MAFFT | 2.0 | 3.59/100 | No significant hit |
| Eukaryotes (2802 sequences) | CLUSTALW | 8.0 | 12.69/100 | No significant hit |
| | DALIGN-TX | 0.8 | 35.58/100 | No significant hit |
| | MAFFT | 4.0 | 66.66/100 | 1.0 |
| **ALIGNMENT-FREE MODELS** | | | | |
| **Training set (source and number of sequences)** | **2D Structural Approach** | **Optimal Classification Cutoff (Accept/Reject)** | **Sensitivity/Specificity (%)** | **Prediction on the ITS2 *Petrakia* sp.** |
| Eukaryotes (12168 sequences) | **Nandy structure** | Accept $>$ 0.529 | 83.80/98.80 | 0.990 |
| | **Mfold structure** | Accept $>$ 0.475 | 92.90/98.60 | 0.996 |

The classification results of our alignment-free models (Mfold and Nandy) when using an optimal cutoff are also provided.
**Classification performance at optimal cutoff in every case (E-value).*
doi:10.1371/journal.pone.0026638.t003

sequenced and registered at the GenBank Database (accession number FJ892749). The ITS2 region was delineated by alignment methods [62] using the conserved 5.8S and 28S rDNA flanking fragments. Then, the ITS2 region was selected to evaluate the predictability of our alignment-free models based on the TI2BioP methodology and also by predictive alignment procedures.

We selected the ANN-based models for the ITS2 classification since they show the highest classification rate for both structural approaches. Both alignment-free models allowed a successfully prediction of the *Petrakia* ITS2 sequence with a confidence level of 0.996 and 0.990 for the Mfold and Nandy-like structures, respectively (**Table 3**). Despite the high divergence among the ITS2 sequences, the models were able to identify a new fungal ITS2 sequence from a dataset made up of divergent UTR sequences with similar structural features but functional different. We also demonstrated that Nandy-like structures provide patterns

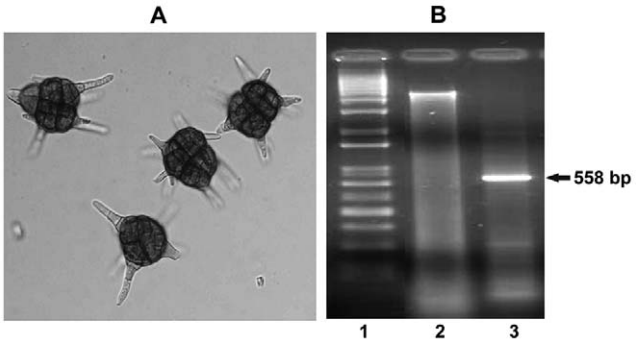



**Figure 6. Conidia of Petrakia sp. from 7 days culture on Malt Extract Agar (×400) (A).** Isolation of a novel ITS2 genomic sequence from *Petrakia* sp. (1) 1 Kb ladder (Gibco BLR), (2) Genomic DNA from the *Petrakia* isolate, (3) PCR reaction with the ITS5 and ITS4 primers (B).
doi:10.1371/journal.pone.0026638.g006

that are useful for gene class discrimination. These 2D artificial maps for DNA/RNA provides information about the connectivity of the nucleotides, but also accounts for the content of purines (GA) and pyrimidine (CT) in the rDNA molecules, which can be observed in the tendency of occupying certain quadrant in the Cartesian system (**figure 1**). The variations in the content of nucleotides have been also used in the genomic recognition of non-protein-coding RNAs [52].

By contrast, profile HMMs generated with different MSA algorithms and different training sets showed in general a poor classification performance on the ITS2 sequence of *Petrakia* sp. Only the profile HMMs based on MAFFT classified it correctly (**Table 3**). Despite that the alignment-free methods and the profile HMMs based on MAFFT recognized our query ITS2 sequence with significance, a BLASTn search (*E*-value cutoff = $10e^{-10}$) against the NCBI database was carried out to support the annotation of the newly isolated sequence by looking for hits belonging or related to the *Petrakia* genus. We retrieved the second best hit (HQ433006) from an uncultured fungus from the Ascomycota phylum. The score (172) and sequence similarity (89%) between our query and this hit were significant (*E*-value = 4e-$^{40}$). However, the BLAST search did not find any hit from the *Petrakia* genus except our own submission (first hit). This confirms that *Petrakia* genus is not well-represented at NCBI and has not been deeply studied yet either taxonomically or as a source of novel secondary metabolites.

## 9. A comparative phylogenetic analysis

The lack of other ITS2 sequences from different species of the genus *Petrakia* (with the exception of our sequence submission at the GenBank) precluded performing a phylogenetic analysis at the species level (low-level analysis). We classified our fungal isolate as a mitosporic Ascomycota/*Petrakia sp.* according to its mycological culture features, as there is not a report with a detailed taxonomy about this genus namely in the NCBI dedicated 'Taxonomy'

database (http://www.ncbi.nlm.nih.gov/tanonomy). Furthermore, there is no specification about its subphylum and class [63]. These fungal species was initially placed into a separate artificial phylum "the Deuteromycota" along with asexual species from other fungal taxa but currently asexual ascomycetes are identified and classified based on morphological or physiological similarities to ascus-bearing taxa, as well as based on phylogenetic analyses of DNA sequences [64]. So, a higher-level phylogenetic study involving Ascomycota members haring ITS2 sequence similarities with *Petrakia* may complement its taxonomy relatively to the ascus-bearing taxa. First, we assumed that our fungal isolate belonged to the *Pezizomycotina* subphylum, the largest within Ascomycota phylum. Our inference agree with a recent classification found in the "The dictionary of the Fungi" [65].

Two different types of distance trees were built: (1) a traditional one based on multiple alignments of ITS2 sequences and (2) another irrespective of sequence similarity supported by the TI2BioP methodology. Both phylogenetic analyses, the traditional and the alignment-free clustering, showed that the *Petrakia* isolate is similar to the Dothideomycetes class members (**figure 7 and 8**). Dothideomycetes is the largest and most diverse class of ascomycete fungi. They are often found as pathogens, endophytes or epiphytes of living plants sharing some morphological features described above for the *Petrakia* genus [66]. In addition, *Petrakia sp.* was placed by the two different computational taxonomic approaches near to the mitosporic Ascomycota *Ampelomyces sp.DSM 2222* supporting the mycological characterization of the query fungus. *Ampelomyces sp.DSM 2222* is taxonomically placed among the Dothideomycetes class and inside the mitosporic Leptosphaeriaceae family producing conidia as *Petrakia sp.* We only show the NJ-tree based on the K2P substitution model to illustrate the tree topology and the BS values for each node that support our phylogenetic inferences (**figure 7**). Similar tree topologies and BS support were obtained with other evolutionary distance matrices and the ME method (see section 2.4) (**figure S2**).

Furthermore, we evaluate the stability of our results on the NJ-tree clustering: (i) by measuring the influence of several alignment-free distances (City-block, Chebychev, and Power distance) in addition to the Euclidean distance, (ii) by assessing other clustering methods (Single linkage, Uweighted pair-group average and the Ward's method) and (iii) by calculating the cophenetic correlation coefficient for the clustering depicted in the **figure 8**. The topologies of the alignment-free trees based on different distance metrics are quite similar as well as the positions of the taxa in respect of our query fungus along the four trees (**figure S3**). Similar outcomes were obtained when different clustering methods were computed using the Euclidean distance to plot the trees (**figure S4**). These two facts support the consistency of our original alignment-free clustering despite the difficulty to perform a statistical significance testing, as unlike many other statistical procedures, cluster analysis methods are mostly used when we do not have any *a priori* hypotheses. One way to measure the validity of the cluster information generated by the linkage function is to compare it with the original proximity data generated by the pairwise distance (Euclidean) function. If the clustering is valid, the linking of objects in the cluster tree should have a strong correlation with the distances between objects in the distance vector. The cophenet function compares these two sets of values and computes their correlation, returning a value called the cophenetic correlation coefficient (ccc) [18]. We retrieve a ccc value for the furthest-neighbor clustering of 0.87 showing an strong correlation (the closer the value of the ccc is to 1, the better the clustering solution). The cophenet function was used to evaluate the clustering method using the other distance metrics mentioned above. The ccc values for the City-block, Chebychev, and Power distances were 0.84, 0.82 and 0.80, respectively, showing consistency in the clustering solution.

The tree topologies obtained for both approaches are somewhat similar as well as the sub-topologies within the Ascomycota classes, specially the Petrakia's location among the Dothideomycetes. Moreover, *Phyllactinia moricola* (outgroup) is placed far from the rest of the members (inner group). Therefore, the NJ clustering based on the Euclidean distance matrix computed from our alignment-free indices largely agrees with the traditional NJ distance tree, which have a phylogenetic meaning since is based on evolutionary distances.

These findings support the importance of including ITS2 structural information when assessing the phylogenetic relationships at higher levels in eukaryote evolutionary comparisons. Although the Euclidean distance is simply a sort of geometric

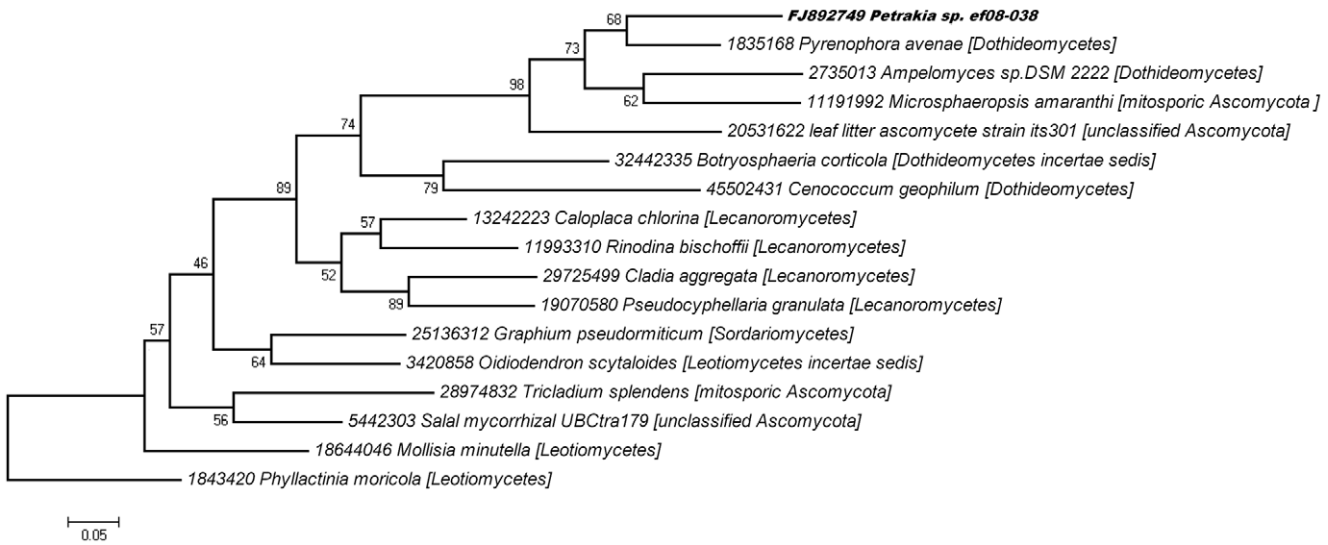



**Figure 7. Neighbor-joining tree based on the ITS2 sequences using the substitution Kimura 2-parameter (K2P).**
doi:10.1371/journal.pone.0026638.g007

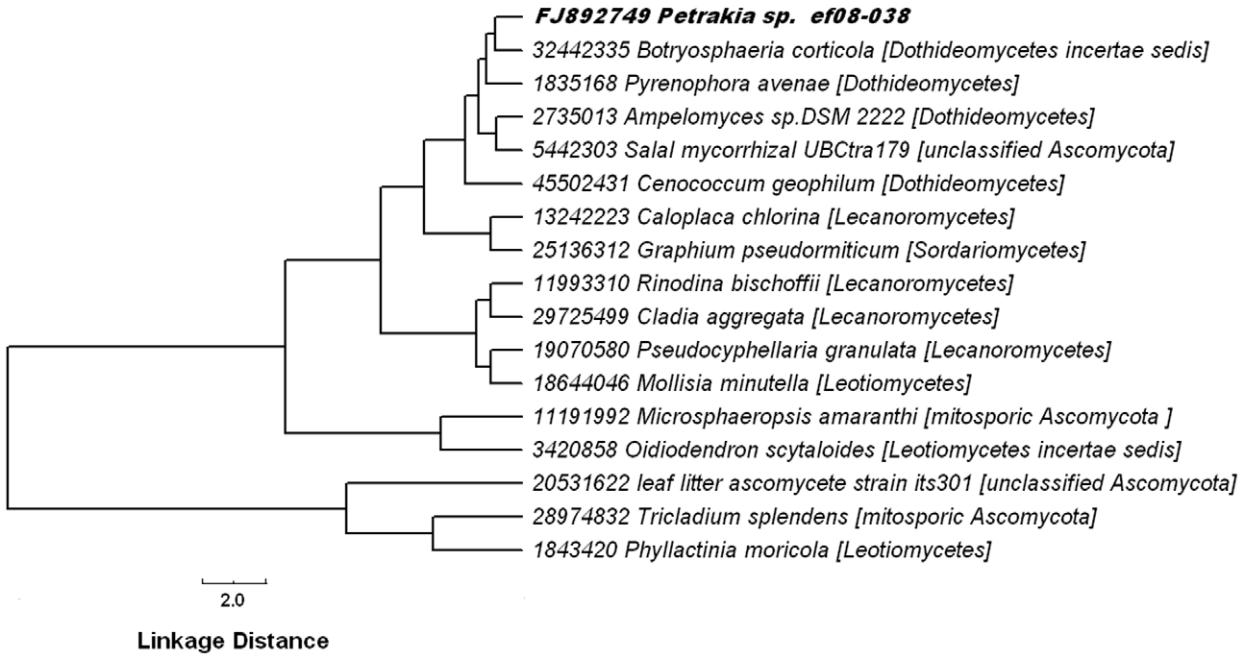



**Figure 8. Neighbor-joining tree clustering based on the Euclidean distance calculated from the TIs values.**
doi:10.1371/journal.pone.0026638.g008

distance in a multidimensional space with no phylogenetic meaning, it led to an effective hierarchical biological clustering with an evolutionary approach because it was derived from the TIs containing both sequence and structural information.

## Conclusions

Topological indices containing information about ITS2 sequences and structures are effective to produce ANN-models with a high prediction power despite the sequence diversity of this class. The use of artificial but informative DNA/RNA secondary structures is a less-costly alternative for the ITS2 classification when higher structural levels are not available or the correct structure is only rarely found by standard RNA folding algorithms. TI2BioP provided simplicity and reliability to ANN-models to search a novel ITS2 member, performing even better than the profile HMMs built up with optimized MSA algorithms for low overall sequence similarity. In addition, our alignment-free approach is effective to construct hierarchical distance-trees containing relevant biological information with an evolutionary significance.

## Supporting Information

**File S1** Exploring ITS2 and UTRs sequence diversity by Needleman-Wunsch and Smith-Waterman procedures.
(DOC)

**File S2** IDs, training and prediction series, values of the TIs predictors for the ITS2 and UTR sequences.
(XLS)

**File S3** Classification results derived from ANN-models on the training, selection and test set for the two structural approaches.
(XLS)

**File S4** MSA performed by several algorithms (CLUSTALW, DIALIGN-TX and MAFFT) using three different training sets (File S4.1–4.9).
(RAR)

**File S5** ITS2 profile HMMs generated with the MSA showed in File S4 (File S4.1–4.9).
(RAR)

**File S6** ROC analysis for each profile HMM at 20 different E-values (0.1–10).
(XLS)

**Figure S1** Pair wise comparison (all *vs* all) for the ITS2 and UTRs sequences evaluated in this study using the Needleman-Wunsch (NW) (in light gray) and Smith-Waterman (SW) (in dark gray) alignment algorithms.
(TIF)

**Figure S2** Neighbor-joining trees based on JC (in black) and MCL (in red) substitution models and ME trees based on the JC (in green) and K2P (in blue) evolutionary distances.
(TIF)

**Figure S3** Neighbour-joining trees built with different alignment-free distance metrics: Euclidean (in black), City-block (in blue), Chebychev (in red) and Power (in green) distances. Each taxa is labeled for a number as follow: (1) FJ892749 *Petrakia sp. ef08-038*, (2) 1835168 *Pyrenophora avenae* [Dothideomycetes], (3) 2735013 *Ampelomyces sp.DSM 2222* [Dothideomycetes], (4) 11191992 *Microsphaeropsis amaranthi* [mitosporic Ascomycota], (5) 20531622 *leaf litter ascomycete strain its301* [unclassified Ascomycota], (6) 32442335 *Botryosphaeria corticola* [Dothideomycetes incertae sedis], (7) 45502431 *Cenococcum geophilum* [Dothideomycetes], (8) 13242223 *Caloplaca chlorina* [Lecanoromycetes], (9) 11993310 *Rinodina bischoffii* [Lecanoromycetes], (10) 29725499 *Cladia aggregata* [Lecanoromycetes], (11) 19070580 *Pseudocyphellaria granulata* [Lecanoromycetes], (12) 25136312 *Graphium pseudormiticum* [Sordariomycetes], (13) 3420858 *Oidiodendron scytaloides* [Leotiomycetes incertae sedis], (14) 28974832 *Tricladium splendens* [mitosporic Ascomycota], (15) 5442303 *Salal mycorrhizal UBCtra179* [unclassified Ascomycota], (16) 18644046 *Mollisia minutella* [Leotiomycetes], (17) 1843420 *Phyllactinia moricola* [Leotiomycetes].
(TIF)

**Figure S4** Joining-tree clustering using different methods for the linkage of the Euclidean distance: Complete linkage (in black), single linkage (in blue), unweighted pair-group average (in red) and the Ward's method (in green). Taxa are labeled by numbers as in the figure S3.
(TIF)

## Acknowledgments

GACH would like to thank Dr. Leticia Arco-García from the CEI/UCLV for her collaborative work on the validation of hierarchic clusters. Special thanks to Prof. Dr. Timm Anke from Department of Biotechnology, University of Kaiserslautern, for allowing Hidalgo-Yanes P.I to work in his research group. Comments made by the Associate Editor, Jonathan H. Badger, and two anonymous referees improved a previous version of this manuscript.

## Author Contributions

Conceived and designed the experiments: AA GACH ASR. Performed the experiments: GACH PIHY RMR YPC ASR. Analyzed the data: GACH ASR AA. Contributed reagents/materials/analysis tools: AA VV RMR KM. Wrote the paper: GACH ASR AA.

## References

1. Aguero-Chapin G, Perez-Machado G, Molina-Ruiz R, Perez-Castillo Y, Morales-Helguera A, et al. (2011) TI2BioP: Topological Indices to BioPolymers. Its practical use to unravel cryptic bacteriocin-like domains. Amino Acids 40: 431–442.
2. Aguero-Chapin G, de la Riva GA, Molina-Ruiz R, Sanchez-Rodriguez A, Perez-Machado G, et al. (2011) Non-linear models based on simple topological indices to identify RNase III protein members. J Theor Biol 273: 167–178.
3. Strope PK, Moriyama EN (2007) Simple alignment-free methods for protein classification: a case study from G-protein-coupled receptors. Genomics 89: 602–612.
4. Finn RD, Mistry J, Tate J, Coggill P, Heger A, et al. (2009) The Pfam protein families database. Nucleic Acids Res.
5. 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: W273–279.
6. Selig C, Wolf M, Muller T, Dandekar T, Schultz J (2008) The ITS2 Database II: homology modelling RNA structure for molecular systematics. Nucleic Acids Res 36: D377–380.
7. Chou KC (2009) Automated prediction of protein attributes and its impact to biomedicine and drug discovery. In: Alterovitz G, Benson R, Ramoni MF, eds. Automation in Proteomics and Genomics: An Engineering Case-Based Approach (Harvard-MIT interdisciplinary special studies courses). UK: Wiley & Sons. pp 97–143.
8. Perez-Bello A, Munteanu CR, Ubeira FM, De Magalhaes AL, Uriarte E, et al. (2009) Alignment-free prediction of mycobacterial DNA promoters based on pseudo-folding lattice network or star-graph topological indices. J Theor Biol 256: 458–466.
9. Schultz J, Maisel S, Gerlach D, Müller T, Wolf M (2005) A common core of secondary structure of the internal transcribed spacer 2 (ITS2) throughout the Eukaryota. RNA 11: 361–364.
10. Koetschan C, Forster F, Keller A, Schleicher T, Ruderisch B, et al. (2009) The ITS2 Database III–sequences and structures for phylogeny. Nucleic Acids Res.
11. Schultz J, Müller T, Achtziger M, Seibel P, Dandekar T, et al. (2006) The internal transcribed spacer 2 database–a web server for (not only) low level phylogenetic analyses. Nucleic Acids Research 34.
12. Keller A, Schleicher T, Schultz J, Muller T, Dandekar T, et al. (2009) 5.8S–28S rRNA interaction and HMM-based ITS2 annotation. Gene 430: 50–57.
13. Nandy A (1996) Two-dimensional graphical representation of DNA sequences and intron-exon discrimination in intron-rich sequences. Comput Appl Biosci 12: 55–62.
14. Nandy A (2009) Empirical relationship between intra-purine and intra-pyrimidine differences in conserved gene sequences. PLoS One 4: e6829.
15. Mathews DH (2006) RNA secondary structure analysis using RNAstructure. Curr Protoc Bioinformatics Chapter 12: Unit 12 16.
16. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680.
17. Subramanian AR, Kaufmann M, Morgenstern B (2008) DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. Algorithms Mol Biol 3: 6.
18. Katoh K, Kuma K, Miyata T, Toh H (2005) Improvement in the accuracy of multiple sequence alignment program MAFFT. Genome Inform 16: 22–33.
19. Nilsson RH, Kristiansson E, Ryberg M, Hallenberg N, Larsson KH (2008) Intraspecific ITS variability in the kingdom fungi as expressed in the international sequence databases and its implications for molecular species identification. Evol Bioinform Online 4: 193–201.
20. Qi FH, Jing TZ, Wang ZX, Zhan YG (2009) Fungal endophytes from Acer ginnala Maxim: isolation, identification and their yield of gallic acid. Lett Appl Microbiol 49: 98–104.
21. Molina R, Agüero-Chapin G, Pérez-González MP TI2BioP (Topological Indices to BioPolymers) *version 1.0*: Molecular Simulation and Drug Design (MSDD), Chemical Bioactives Center, Central University of Las Villas, Cuba.
22. 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.
23. Gutierrez Y, Estrada E (2002) MODESLAB 1.0 (Molecular DEScriptors LABoratory) for Windows.
24. Aguero-Chapin G, Gonzalez-Diaz H, Molina R, Varona-Santos J, Uriarte E, et al. (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.
25. Agüero-Chapin G, Varona-Santos J, Riva Gdl, Antunes A, González-Villa T, et al. (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.
26. González-Díaz H, Molina-Ruiz R, Hernandez I (2007) MARCH-INSIDE v3.0 (**MAR**kov **CH**ains **IN**variants for **SI**mulation & **DE**sign). 3.0 ed. pp. Windows supported version under request to the main author contact email: gonzalezdiazh@yahoo.es.
27. Pesole G, Liuni S, Grillo G, Licciulli F, Larizza A, et al. (2000) UTRdb and UTRsite: specialized databases of sequences and functional elements of 5′ and 3′ untranslated regions of eukaryotic mRNAs. Nucleic Acids Res 28: 193–196.
28. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48: 443–453.
29. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147: 195–197.
30. Mathews DH (2005) Predicting a set of minimal free energy RNA secondary structures common to two sequences. Bioinformatics 21: 2246–2253.
31. 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.
32. 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.
33. Cornell WD, Cieplak P, IBayly C, Gould IR, Merz KWJ, et al. (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 117: 5179–5197.
34. Statsoft (2007) STATISTICA 7.0 (data analysis software system for windows). version 7.0 ed.
35. Meneses-Marcel A, Marrero-Ponce Y, Machado-Tugores Y, Montero-Torres A, Pereira DM, et al. (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.
36. Marrero-Ponce Y, Castillo-Garit JA, Olazabal E, Serrano HS, Morales A, et al. (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.
37. 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.
38. Ponce YM, 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.
39. Kutner MH, Nachtsheim CJ, Neter J, Li W (2005) Standardized Multiple Regression Model. Applied Linear Statistical Models. Fifth ed. New York: McGraw Hill. pp 271–277.
40. Sokolova M, Japkowicz N, Szpakowicz S. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. 1015–1021.
41. Zupan J, Gasteiger J (1999) Neural Networks in Chemistry and Drug Design: An Introduction. Weinheim: Wiley-VCH. 483 p.
42. Gotoh O (1982) An improved algorithm for matching biological sequences. J Mol Biol 162: 705–708.

43. Krogh ABM, Mian IS, Sjeander K, Haussler D (1994) Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol 235: 1501–1531.
44. Sacks W, Nurnberger T, Hahlbrock K, Scheel D (1995) Molecular characterization of nucleotide sequences encoding the extracellular glycoprotein elicitor from Phytophthora megasperma. Mol Gen Genet 246: 45–55.
45. White TJ, Bruns T, Lee S, Taylor J (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis MA, Gelfand DH, Sninsky JJ, White TJ, eds. PCR Protocols: A Guide to Methods and Applications. San Diego: Academic Press. pp 315–322.
46. Aguero-Chapin G, Varona-Santos J, de la Riva GA, Antunes A, Gonzalez-Vlla T, et al. (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.
47. Gonzalez-Diaz H, Aguero-Chapin G, Varona J, Molina R, Delogu G, et al. (2007) 2D-RNA-coupling numbers: a new computational chemistry approach to link secondary structure topology with biological function. J Comput Chem 28: 1049–1056.
48. Aguero-Chapin G, Gonzalez-Diaz H, de la Riva G, Rodriguez E, Sanchez-Rodriguez A, et al. (2008) 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.
49. Rumelhart DE, McClelland JL (1986) Parallel distributed processing: Explorations in the microstructure of cognition. Cambridge, MA: MIT Press.
50. Gonzalez-Diaz H, Gonzalez-Diaz Y, Santana L, Ubeira FM, Uriarte E (2008) Proteomics, networks and connectivity indices. Proteomics 8: 750–778.
51. Gonzalez-Diaz H, Aguero-Chapin G, Varona-Santos J, Molina R, de la Riva G, et al. (2005) 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.
52. Schattner P (2002) Searching for RNA genes using base-composition statistics. Nucleic Acids Res 30: 2076–2082.
53. Wong TK, Lam TW, Sung WK, Yiu SM (2010) Adjacent nucleotide dependence in ncRNA and order-1 SCFG for ncRNA identification. PLoS One 5.
54. Rivals I, Personnaz L (1999) On cross validation for model selection. Neural Comput 11: 863–870.
55. Caballero J, Fernandez M (2008) Artificial neural networks from MATLAB in medicinal chemistry. Bayesian-regularized genetic neural networks (BRGNN): application to the prediction of the antagonistic activity against human platelet thrombin receptor (PAR-1). Curr Top Med Chem 8: 1580–1605.
56. Fernandez M, Caballero J, Fernandez L, Abreu JI, Garriga M (2007) Protein radial distribution function (P-RDF) and Bayesian-Regularized Genetic Neural Networks for modeling protein conformational stability: chymotrypsin inhibitor 2 mutants. J Mol Graph Model 26: 748–759.
57. Fernandez M, Caballero J, Fernandez L, Sarai A (2010) Genetic algorithm optimization in drug design QSAR: Bayesian-regularized genetic neural networks (BRGNN) and genetic algorithm-optimized support vectors machines (GA-SVM). Mol Divers.
58. Winters-Hilt S (2006) Hidden Markov model variants and their application. BMC Bioinformatics 7 Suppl 2: S14.
59. Sieber TN. Endophytic fungi in forest trees: are they mutualists? Fungal Biology Reviews 21: 75–89.
60. Nagano Y, Elborn JS, Millar BC, Walker JM, Goldsmith CE, et al. (2009) Comparison of techniques to examine the diversity of fungi in adult patients with cystic fibrosis. Med Mycol. pp 1–12.
61. Von Arx JA (1981) The Genera of Fungi Sporulating in Pure Culture: Lubrecht & Cramer Ltd. 315 p.
62. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res 25: 3389–3402.
63. Bisby F, Roskov Y, Ruggiero M, Orrell T, Paglinawan L, et al. (2007) Species 2000 & ITIS Catalogue of Life: 2007 Annual Checklist Taxonomic Classification. CD-ROM; Species 2000: Reading, U.K.
64. Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33: 511–518.
65. Kirk PM, Cannon PF, Stalpers JA (2008) The dictionary of the Fungi; Paul M Kirk, Paul F Cannon, David W Minter, Stalpers aJA, eds. UK: CABI. 784 p.
66. Arx von J, Müller E (1975) A re-evaluation of the bitunicate ascomycetes with keys to families and genera. Studies in Mycology 9: 1–159.

# ANNEX 4

OPEN ACCESS Freely available online

PLOS ONE

# Exploring the Adenylation Domain Repertoire of Nonribosomal Peptide Synthetases Using an Ensemble of Sequence-Search Methods

**Guillermin Agüero-Chapin[1,2,3], Reinaldo Molina-Ruiz[2], Emanuel Maldonado[1], Gustavo de la Riva[4], Aminael Sánchez-Rodríguez[5], Vitor Vasconcelos[1,3], Agostinho Antunes[1,3]***

**1** CIMAR/CIIMAR, Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Porto, Portugal, **2** Molecular Simulation and Drug Design (CBQ), Universidad Central¨Marta Abreü de Las Villas (UCLV), Santa Clara, Cuba, **3** Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Porto, Portugal, **4** Departamento de Biología, Instituto Tecnológico Superior de Irapuato (ITESI), Carretera Irapuato-Silao Km. 12.5, El Copal, Irapuato, Guanajuato, México, **5** CMPG, Department of Microbial and Molecular Systems, KU Leuven, Leuven, Belgium

## Abstract

The introduction of two-dimension (2D) graphs and their numerical characterization for comparative analyses of DNA/RNA and protein sequences without the need of sequence alignments is an active yet recent research topic in bioinformatics. Here, we used a 2D artificial representation (four-color maps) with a simple numerical characterization through topological indices (TIs) to aid the discovering of remote homologous of Adenylation domains (A-domains) from the Nonribosomal Peptide Synthetases (NRPS) class in the proteome of the cyanobacteria *Microcystis aeruginosa*. Cyanobacteria are a rich source of structurally diverse oligopeptides that are predominantly synthesized by NPRS. Several A-domains share amino acid identities lower than 20 % being a possible source of remote homologous. Therefore, A-domains cannot be easily retrieved by BLASTp searches using a single template. To cope with the sequence diversity of the A-domains we have combined homology-search methods with an alignment-free tool that uses protein four-color-maps. **TI2BioP** (**T**opological **I**ndices **to** **Bio**Polymers) *version 2.0*, available at http://ti2biop.sourceforge.net/ allowed the calculation of simple TIs from the protein sequences (four-color maps). Such TIs were used as input predictors for the statistical estimations required to build the alignment-free models. We concluded that the use of graphical/numerical approaches in cooperation with other sequence search methods, like multi-templates BLASTp and profile HMM, can give the most complete exploration of the repertoire of highly diverse protein families.

**Citation:** Agüero-Chapin G, Molina-Ruiz R, Maldonado E, de la Riva G, Sánchez-Rodríguez A, et al. (2013) Exploring the Adenylation Domain Repertoire of Nonribosomal Peptide Synthetases Using an Ensemble of Sequence-Search Methods. PLoS ONE 8(7): e65926. doi:10.1371/journal.pone.0065926

**Editor:** Christos A. Ouzounis, The Centre for Research and Technology, Hellas, Greece

**Received** November 21, 2012; **Accepted** May 1, 2013; **Published** July 16, 2013



**Funding:** The authors acknowledge the Portuguese Fundação para a Ciência e a Tecnologia (FCT) for financial support to GACH (SFRH/BD/47256/2008), and the projects PTDC/AAC-AMB/104983/2008 (FCOMP-01-0124-FEDER-008610), PTDC/AAC-CLI/116122/2009 (FCOMP-01-0124-FEDER-014029), and PesT-C/MAR/LA0015/2011 to AA. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

* E-mail: aantunes@ciimar.up.pt

## Introduction

The Chemical Graph Theory (CGT) consists in the application of the graph theory to perform combinatorial and topological exploration of the chemical molecular structure. Currently, the CGT is being extended to bioinformatics through the introduction of two-dimensional (2D) graphs for comparative analyses of DNA/RNA and proteins without the use of sequence alignments. These 2D graphs or maps do not represent the "real structure" of the natural biopolymers but they have been very effective to inspect similarities/dissimilarities among them, either by direct visualization or by numerical characterization [1]. Examples of 2D artificial representations of DNA and protein sequences with potentialities in bioinformatics include the spectrum-like, star-like, cartesian-type and four-color maps [1–5]. These DNA/RNA and protein maps can generally unravel higher-order useful information contained beyond the primary structure, i.e. nucleotide/amino acid distribution into a 2D space. Their essence can be captured in a quantitative manner through numerical indices to easily compare a great number of sequences/maps [6–8]. One of the simplest numerical characterizations of sequences comprehends the use of topological indices. Topological Indices (TIs) are based on the connectivity between the elements composing the 2D graph in terms of whether they are connected or not [9,10]. While several types of 2D maps have been developed for DNA/RNA and proteins, including their numerical characterization [11], the four-color maps application in bioinformatics has been mostly unexplored, being limited to illustrative examples on the comparative characterization of DNA and protein sequences [12]. However, the use of the four-color maps and its numerical characterization can cooperate with traditional homology search tools (e.g. BLAST, HMMs) to carry out an exhaustive exploration of functional signatures in highly diverse gene/protein families. Such exploration is effective when all family members are retrieved including remote homologs. Remotes homologues are divergent gene/protein sequences that have conserved the same biological function in different organisms. They can be harvest in the alignment algorithms twilight zone (<30% of amino acid identity) and have been traditionally detected by the use of more

sensitive alignment-based methods like PSI-BLAST [13] and profiles Hidden Markov Models (HMM) [14]. The Nonribosomal Peptide Synthetases (NRPS) family can harbor remote homologous due to the high sequence divergence among its Adenylation domains (A-domains). In fact, all A-domain members cannot be retrieved easily by BLASTp using a single template [15]. NRPS are megasynthetases composed by several domains organized in clusters for the synthesis of oligopeptides with biological activities. A-domains are mandatory in each NRPS cluster being responsible for the amino acid selection and its covalent fixation on the phospho-pantethein arm as thioester, through AMP-derivative intermediate during the production of oligopeptides via non-ribosomal [16]. Cyanobacteria are a rich source of structurally diverse oligopeptides that are predominantly synthesized by NRPS. In *Microcystis*, a common cyanobacteria genus in eutrophic freshwaters, numerous bioactive peptides have been identified that can be mostly classified as aeruginosins, microginins, microcystins, cyanopeptolins, and anabaenopeptins [17]. In the present work we aim to annotate the A-domain repertoire in the proteome of *Microcystis aeruginosa* as a strategy to spot NRPS clusters. To handle the high sequence diversity of A-domains we used an ensemble of homology-search methods, including an alignment-free model that integrates the four-color-maps for proteins. **TI2BioP** (**T**opological **I**ndices **to BioP**olymers) *version 2.0*, available at http://ti2biop.sourceforge.net/ allows the calculation of TIs from the four-color maps for protein sequences [18]. Such TIs were used as input predictors for statistical techniques to build alignment-free models. We concluded that the use of an ensemble of sequence search methods (homology-based and alignment-free) can give the best exploration of the repertoire of highly diverse protein classes, such as the NRPS represented by its A-domains. The graphical method rendered a Decision Tree Model (DTM) that detected signatures of 22 A-domains in the proteome of *Microcystis aeruginosa* matching 19 out of 20 hits previously annotated as A-domains. The multiple-template BLASTp found exactly the 20 A-domain signatures annotated in the proteome, while the profile HMM detected the same 20 hits plus three additional ones. DTM and profile HMM identified, respectively, two and three A-domain signatures not found by multi-template BLASTp among the hypothetical proteins. The consensus detection of additional hits by the two sequence search methods provides clues for the presence of further A-domains remote homologues. The new A-domain variants found in the proteome of *Microcystis aeruginosa* could unravel the presence of novel NRPS clusters.

## Results

### Alignment-free model selection

We computed 17 TIs that consist in spectral moment series ($^{fc}\mu_0$-$^{fc}\mu_{16}$) derived from four-color maps representing 8892 protein domains (138 A-domains and 8854 CATH domains) using **TI2BioP** (described in Methods and Database). The $^{fc}\mu_0$-$^{fc}\mu_{16}$ series were used as input predictors to build classification linear models as the simplest relation between the response variable and the predictors. General Discrimination Analysis (GDA) best subset implemented in the *STATISTICA* software was used for such purposes [19]. We select the best subset of predictors that accounts for the more effective discrimination between A and CATH domains through plotting the $\lambda$ variation against the number of predictors in the set of models. A parsimonious linear model was selected at the point where the $\lambda$ start to decrease smoothly **(Figure 1).**

We found a linear classification function (**see equation below**) with four significant predictors ($^{fc}\mu_1$, $^{fc}\mu_2$, $^{fc}\mu_9$, $^{fc}\mu_{12}$) describing the topology of the four-color maps at short range ($^{fc}\mu_1$, $^{fc}\mu_2$) and at long range ($^{fc}\mu_9$, $^{fc}\mu_{12}$) interactions.

$$\begin{aligned}&\text{AvsCATHdomains} = 54.83^{HP}\mu_1 - 20.94^{HP}\mu_2 \oplus 68.70^{HP}\mu_9 \\ &-62.0^{HP}\mu_{12} - 252.69 \qquad (1)\\ &N = 6750 \qquad \lambda = 0.11 \qquad F = 1556.7 \qquad p < 0.05\end{aligned}$$

Where, $N$ is the number of domain sequences used to train the classification model and the statistics parameters commonly used to evaluate linear functions (Wilk's statistical ($\lambda$) and Fisher ratio (F) with a probability of error ($p$-level)) [20,21]. They provided values indicating a good power of discrimination ($\lambda = 0.11$) with significance ($p(F)<0.05$).

The model classification performance is shown in **Table 1** together with the classification results from other alignment-free models developed with non-linear techniques.

GDA provides good classification results in detecting A-domains despite the members of this class ranged mostly between 10–40% of sequence identity **(Figure 2A)** and the CATH domains share less than 35% of sequence identity. Pair-wise identity is the most common cutoff used to decide the twilight zone for alignment algorithms [22]. Sequence alignments unambiguously distinguish between protein pairs of similar and non-similar functional and structural signals when the pairwise sequence identity is high (>40%). The signal gets blurred in the twilight zone of 20-35% sequence identity [22–24]. Particularly, the test set was made up of A-domains mostly sharing between 20 to 30% of amino acid identity **(Figure 2B)** and CATH domains with the diversity above-mentioned. Such test set matches into the twilight zone where generally remote homologous can be harvested.

The prediction power on the test set could be improved using non-linear models like Decision Tree Models (DTM) and Artificial Neural Networks (ANN) as can be seen below.

Although several alignment-free methods have been reported for improving classification accuracy in protein classes and superfamilies [25–27], DTM have been poorly explored to differentiate protein classes [28]. We used Classification Trees (CT) as an exploratory technique to obtain a DTM as predictive tools to detect A-domain signatures. The method found the $^{fc}\mu_1$ and $^{fc}\mu_2$ predictors as splitting variables to produce two decisions split at different values, respectively. The tree structure was very simple, two decision nodes (outlined in blue) and three terminal nodes (outlined in red) summing up a total of five nodes. The numbers of the nodes are labelled on its top-left corner and on the top-right corner are placed the label of the predicted class (A or CATH domain). The 6750 training sequences are assigned to the root node (first node) and tentatively classified as CATH domains or control set. CATH domains are chosen as the initial classification because they are numerically superior to A-domains.

The root node is split, forming two new nodes. The text below the root node describes the split. It indicates that protein sequences with $^{fc}\mu_1$ values higher than or equal to 3817 are sent to node number 3 and tentatively classified as A-domains, by contrary domain sequences with $^{fc}\mu_1$ values lesser than this value are assigned to node number 2 and classified in the control set (CATH domains). Similarly, node 3 is subsequently split taking the decision that sequences with $^{fc}\mu_2$ values lesser than or equal to 11.12 are sent to node number 4 to be classified as A domains (109 cases). The remaining domain sequence with $^{fc}\mu_2$ value greater than 11.12 are sent to node number 5 to be classified as CATH domains reaching 6641 cases well classified (100%).

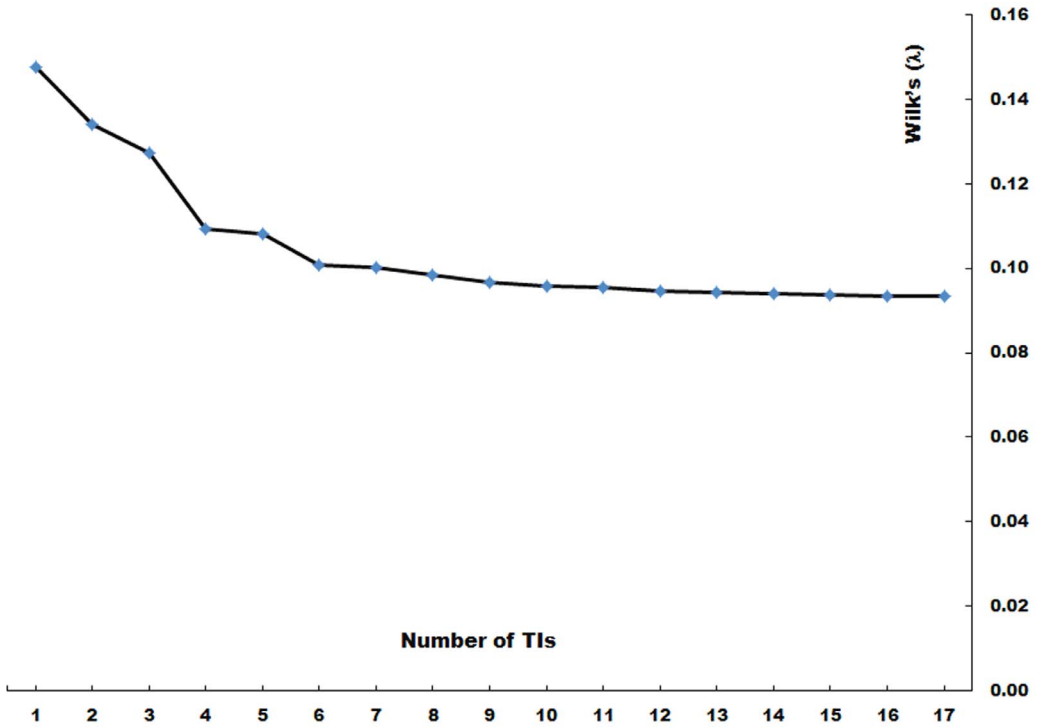



**Figure 1. Assessing the relationship between the number of TIs entered in each model and the Wilk's (λ) values obtained for each one.**
doi:10.1371/journal.pone.0065926.g001

The tree graph presents all this information in a simple and straightforward way allowing processing the information easily. The histograms plotted within the tree's terminal nodes show the excellent performance of the DTM for the recognition of A-domain signatures (**Figure 3**). The information from the tree plot is also available in **Table 2**.

The classification results from the DTM development to recognize A-domain signatures on training and test sets are shown in **Table 1** as well as the results for the 10-fold CV procedure on the training set and the predictability on the test set. The classification improvement is remarkable in respect to the linear models.

ANN is one of the most popular non-linear modelling techniques in use today and has been frequently applied into bioinformatics [29–31]. The selection of input variables is a critical part of neural network design. We use the combination of our own experience and several feature selection algorithms (Forward, Backward and Genetic Algorithm Selection) based on Multilayer Perceptrons (MLP) available in the *STATISTICA Neural Networks* module for variable selection [19]. The $^{fc}\mu_0$ and $^{fc}\mu_1$ predictors were selected by consensus from the three methods. Then, a good starting point to set the topology of the MLP is to use one hidden layer, with the number of units equal to half the sum of the number of input and output units.

The **Table 3** shows the different MLP topologies used to select the right complexity of the ANN. The performance on training, selection and test progress were examined as well as its errors. The best model was the MLP profile highlighted in bold in **Table 3**, which showed the best accuracy on training, selection and test sets, minimizing its respective errors.

The classification results derived from the best MLP profile to classify A-domains are shown in **Table 1**. This ANN-model also showed a higher accuracy level in classifying the training and test sets in respect to the linear model but a very similar performance in comparison to the DTM. However, according to the statistics from the 10-fold CV procedure carried out for each alignment-free model, the DTM shows the best statistics average (**Table 1**) being the most robust model reported among them. Therefore, DTM was the selected model to perform A-domains search among the proteome of *Microcystis aeruginosa*.

## Alignment-free approaches vs. homology-search methods in the detection of A-domains

We carried out a comparatively analysis to evaluate the sensitivity of other different alignment-free approaches and homology-search methods in respect to our graphical/numerical model to detect A-domains among the overall dataset (138 A-domains and 8 854 CATH domains) included in study. Such comparison was addressed to inspect the ability of our alignment-free approach to detect distant A-domains members (A-domains placed in the twilight zone) in the selected dataset. The Webserver PseAAC (http://www.csbio.sjtu.edu.cn/bioinf/PseAAC/) was used to generate alignment-free approaches based on amino acid composition (AAC) and pseudo amino acid composition (PseAAC) [32]. Both approaches provided classifiers to build up DTM under the same statistical parameters reported for our graphical/numerical-based model. Amino acids were weighted with their hydrophobicity values, similarly to the physicochemical property used for the four-color maps and λ values that reflect the sequence order effect was set to 0 if the AAC is only considered and 1 if we take into account the sequence order [33].

Most of the alignment-free classifiers have been based on AAC to predict protein cellular attributes and biological functions including remote homologs detection [26,34]. One of the most popular alignment-free approaches is the Chou's concept of PseAAC that reflects the importance of the sequence order effect

**Table 1.** Classification results for the three alignment-free models (GDA, DTM and ANN) in A-domains detection.

| | Training | | | Test | |
|---|---|---|---|---|---|
| **GDA** | **A-domain** | | **CATH domain** | **A- domain** | **CATH domain** |
| A-domain | **102** | | 0 | **24** | 0 |
| CATH domain | 7 | | **6641** | 5 | **2213** |
| Total | 109 | | 6641 | 29 | 2213 |
| Sensitivity (Sv) (%) | 93.58 | | | 82.76 | |
| Specificity (Sp) (%) | 100 | | | 100 | |
| Accuracy (Acc) (%) | 99.89 | | | 99.78 | |
| F-score | | | | 0.99 | |
| **10-fold CV** | **Sv** | **Sp** | **Acc** | | |
| Average | 93.58 | 100 | 99.89 | | |

| | Training | | | Test | |
|---|---|---|---|---|---|
| **DTM** | **A-domain** | | **CATH domain** | **A-domain** | **CATH domain** |
| A-domain | **109** | | 0 | **29** | 0 |
| CATH domain | 0 | | **6641** | 0 | **2213** |
| Total | 109 | | 6641 | 29 | 2213 |
| Sensitivity (%) | 100 | | | 100 | |
| Specificity (%) | 100 | | | 100 | |
| Accuracy (%) | 100 | | | 100 | |
| F-score | | | | 1.0 | |
| **10-fold CV** | **Sv** | **Sp** | **Acc** | | |
| Average | 98.16 | 99.98 | 99.95 | | |

| | Training | | Selection | | Test | |
|---|---|---|---|---|---|---|
| **ANN** | **A-domain** | **CATH domain** | **A-domain** | **CATH domain** | **A-domain** | **CATH domain** |
| A-domain | **87** | 0 | **21** | 0 | **28** | 0 |
| CATH domain | 0 | **5313** | 1 | **1328** | 1 | **2213** |
| Total | 87 | 5313 | 22 | 1328 | 29 | 2213 |
| Sensitivity (%) | 100 | | 95.45 | | 96.55 | |
| Specificity (%) | 100 | | 100 | | 100 | |
| Accuracy (%) | 100 | | 99.92 | | 99.95 | |
| F-score | | | | | 1.0 | |
| **10-fold CV** | **Sv** | **Sp** | **Acc** | | | |
| Average | 80.24 | 79.91 | 79.92 | | | |

doi:10.1371/journal.pone.0065926.t001

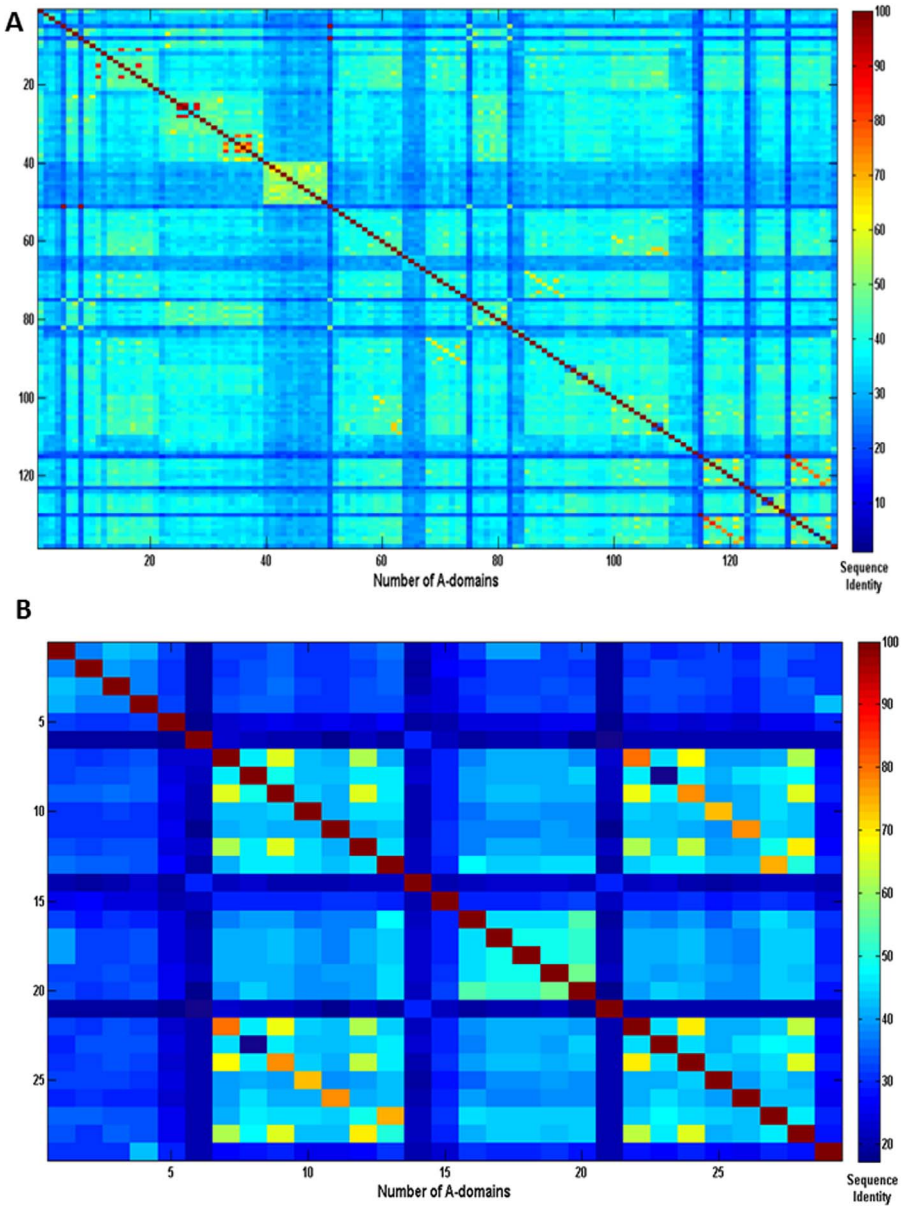



**Figure 2. Dot plot for the global sequence identity matrix obtained by Needleman-Wunsch algorithm for A-domains.** (A) All A-domains involved in the study. (B) A-domains of the test set.
doi:10.1371/journal.pone.0065926.g002

in addition to the AAC to improve the prediction quality to detect protein attributes [33,35]. Classification trees were selected as the statistical technique to generate alignment-free models due to its simplicity and reliability to recognize the A-domain signature among the overall dataset (**Table 1**).

On the other hand, homology-based searches for A-domains were performed by single-template BLASTp, multi-template BLASTp and profile HMM. These methods that show by definition different sensitivity to recognize distant homologs were evaluated considering their ability to retrieve all A-domains (close and distant members).

Our alignment-free model (DTM) generated by four-color maps outperformed alignment-free models (DTM) supported by AAC and PseAAC (**Table 4**). Although A-domains share 10–40% of sequence identity with several members placed in the twilight zone, it was possible to retrieve all of them using four-color maps. In spite of the fact that the other two left alignment-free methods (AAC and PseAAC) showed lower sensitivity, they did not provide many false positives (**Table 5**). It was also demonstrated the effect of the sequence order besides the AAC on the prediction quality; when $\lambda$ was increased from 0 to 1, there was an improvement in all standard classification measures (**Table 4**).

Regarding homology-based methods sensitivity, classification results agreed with the fact that multi-template BLASTp and profile HMM are more sensitive than simple BLASTp. Both multi-template BLASTp and profile HMM easily retrieved all A-domain members at expectation values (E-value$\leq$10) without reporting any false positive (**Table 5**). However, the BLASTp search using a single template provided false positives (significant matches) among CATH domains at both high (E-value = 10) and relatively stringent cut-offs (E-values$<$0.05) (**Files S1–S5**), which is considered statistically significant and useful for filtering easily

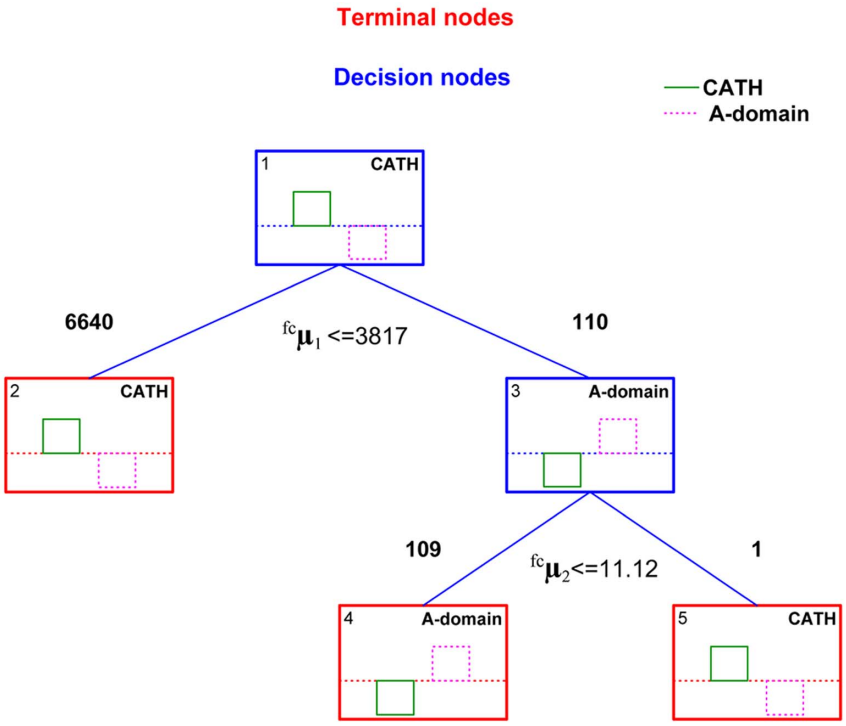



**Figure 3. Architecture for the DTM. Decision Nodes are represented in blue and terminal nodes are in red.** A-domains are labeled using an intermittent line. Otherwise CATH domains are signed by a continuous line. Labels at the right-corner of the nodes indicate tentative membership to A or CATH domain class. Numbers at the left-corner represent the node's number.
doi:10.1371/journal.pone.0065926.g003

identifiable homologs pairs [36,37] (**Table 5**). False positives came up in simple BLASTp searches despite we had cleaned the negative set (CATH domains) from any A-domain signal (by the use of profile HMM-based searches). In contrast to multi-template BLASTp and profile HMM searches, the single-BLASTp search sensitivity did not show stability in identifying the A-domain signal among a benchmark dataset (CATH domains) when the classification parameter (E-value cut-off) was changed. Thus, due to the A-domain diversity, it is less reliable to extrapolate or apply BLASTp searches using a single A-domain template to an unknown test dataset such as an entire proteome. The multi-template BLAST reported by the PKS-NRPS developers was not only useful to detect A-domains with correct boundaries [15]; it also provided more sensitivity (no false positive) and reliability in the identification of this domain class from no stringent conditions (**File S6**). In addition, both the profile HMM described in the methods section (**File S7**) and the DTM built up from four-color maps profiles reached the top in classifying the positive and negative sets. These facts support that profile-based methods are more effective to deal with remote protein homology unless a multi-template BLASTp strategy or PSI-BLAST is conducted. The easy and reliable identification of A-domains by multi-template BLASTp, profile HMM and four-color maps in contrast to a simple BLASTp search and other alignment-free methods provided real clues about the ability of the four-color maps to identify A-domain members in the twilight zone given the evaluated dataset.

## An ensemble of methods to explore the repertoire of NRPS A-domains in Microcystis aeruginosa

The potentialities of the four-color maps and its numerical characterization to detect A-domains in the twilight zone are promising, as we showed previously. Detecting A-domains remote homologues with reliability in a proteome that contains a large diversity of proteins is a challenge for any sequence search method. As several homology-search methods have been assembled into a

**Table 2.** Tree structure in details, child nodes, observed class n's, predicted class, and split condition for each node.

| Node | Left branch | Right branch | CATH | A-domain | Predicted class | Split constant | Split variable |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 6641 | 109 | CATH | −3817.00 | $^{fc}\mu_1$ |
| **2** | | | **6640** | 0 | CATH | −11.13 | $^{fc}\mu_2$ |
| 3 | 4 | 5 | 1 | **109** | A-domain | | |
| **4** | | | 0 | **109** | A-domain | | |
| **5** | | | **1** | 0 | CATH | | |

Numbers in bold highlight the well-classified cases and the terminal nodes.
doi:10.1371/journal.pone.0065926.t002

**Table 3**. Testing different topologies for the MLP on the A-domain classification using TIs from four-color maps.

| Performance Summary for ANN | | | | | | |
|---|---|---|---|---|---|---|
| | **MLP Topologies** | **Train Accuracy** | **Selection Accuracy** | **Test Accuracy** | **Train Error** | **Select Error** | **Test Error** |
| **1** | **MLP 2:2–1–1:1** | **1.000** | **0.999** | **0.999** | **0.000** | **0.027** | **0.021** |
| **2** | MLP 2:2–2–1:1 | 0.756 | 0.757 | 0.758 | 0.001 | 0.024 | 0.020 |
| **3** | MLP 2:2–1–1–1:1 | 0.755 | 0.763 | 0.759 | 0.001 | 0.038 | 0.024 |
| **4** | MLP 2:2–3–1:1 | 0.756 | 0.755 | 0.760 | 0.016 | 0.033 | 0.035 |
| **5** | MLP 2:2–1–2–1:1 | 0.755 | 0.762 | 0.757 | 0.013 | 0.025 | 0.026 |
| **6** | MLP 4:2–2–1–1:1 | 0.756 | 0.757 | 0.759 | 0.006 | 0.022 | 0.020 |

Accuracy performance and error on training, selection and test sets.
doi:10.1371/journal.pone.0065926.t003

certain annotation resource to retrieve accurately all members from highly diverse gene/protein families [38,39], we used our graphical alignment-free method not in competition but in cooperation with alignment procedures to explore the whole repertoire of A-domains, including the detection of new variants (remote homologous), in the proteome of *Microcystis aeruginosa*.

The proteome of the *Microcystis aeruginosa* NIES-843 (http://genome.kazusa.or.jp/cyanobase) is encoded from a 5.8Mbp genome with 6 311 annotated genes; some of them codifying NRPS proteins as hybrids with polyketide synthases (PKS) representing a good target to evaluate the detection of A-domains. DTM was selected among the alignment-free models due to its excellent performance at low sequence identity and its simple way to recognize A-domains. We just calculate the TIs for a proteome and select A-domain signatures according to the DTM rule ($^{fc}\mu_1 \geq 3817$ and $^{fc}\mu_2 \leq 11.12$) (**File S8**). DTM search detected 19 A-domain signatures that coincided with the previously annotation inferred for these genes in the proteome. Three additional cases were also detected as A-domains, but these cases have been previously predicted to be other protein signatures unrelated to NRPS A-domains in the proteome, namely a transketolase-like protein and the other two were hypothetical proteins. The putative hits with some remote relation to A-domains are probably found among the hypothetical proteins due to its unclear annotation. To increase the confidence and quality of the A-domains re-annotation, two sensitive homology-search methods were evaluated on the same proteome. We carried out multi-template BLASTp and profile HMM searches for A-domains in the proteome

**Table 4.** Classification results for alignment-free DTM based on four-color maps, amino acid composition (AAC) and pseudo-amino acid composition (PseAAC) in the A-domains detection.

| **Four-color maps DTM** | **Training** | | | **Test** |
|---|---|---|---|---|
| Sensitivity (Sv) (%) | 100 | | | 100 |
| Specificity (Sp) (%) | 100 | | | 100 |
| Accuracy (Acc) (%) | 100 | | | 100 |
| F-score | | | | 1.0 |
| **10-fold CV** | **Sv** | **Sp** | **Acc** | |
| Average | 98.16 | 99.98 | 99.95 | |
| **AAC ($\lambda = 0$) DTM** | **Training** | | | **Test** |
| Sensitivity (%) | 53.70 | | | 3.44 |
| Specificity (%) | 100 | | | 99.68 |
| Accuracy (%) | 99.25 | | | 98.44 |
| F-score | | | | 0.07 |
| **10-fold CV** | **Sv** | **Sp** | **Acc** | |
| Average | 21.73 | 100 | 98.78 | |
| **PseAAC ($\lambda = 1$) DTM** | **Training** | | | **Test** |
| Sensitivity (%) | 67.89 | | | 20.68 |
| Specificity (%) | 99.80 | | | 99.77 |
| Accuracy (%) | 99.30 | | | 98.75 |
| F-score | | | | 0.40 |
| **10-fold CV** | **Sv** | **Sp** | **Acc** | |
| Average | 21.73 | 100 | 98.78 | |

doi:10.1371/journal.pone.0065926.t004

**Table 5**. True positives *vs*. false positives in the A-domain detection for different sequence-search methods among the overall dataset involved in the study.

| Sequence-search method | True positive | False Positive |
|---|---|---|
| DTM (Four-color maps) | 138 | 0 |
| DTM (AAC) | 59 | 7 |
| DTM (PseAAC) | 80 | 18 |
| HMM (E-value = 10) | 138 | 0 |
| Multi-template BLASTp (E-value = 10) | 138 | 0 |
| BLASTp (E-value = 10) | 138 | 6033 |
| BlASTp (E-value = 0.05) | 138 | 122 |
| BLASTp (E-value = 0.01) | 138 | 24 |
| BLASTp (E-value = 0.001) | 138 | 4 |
| BLASTp (E-value = 0.0001) | 138 | 0 |

doi:10.1371/journal.pone.0065926.t005

according to procedures described in the Methods section, respectively. Multi-template-BLASTp found 20 significant hits coinciding perfectly with the number of A-domains signatures in the annotated genome (**File S9**). The profile HMM detected 23 significant matches for the A-domain signature in the cyanobacteria proteome (**File S10**). Twenty out of these 23 matches agreed with the multi-template BLASTp results and therefore with the current proteome annotation. The remaining three detected hits by the profile HMM were found among the hypothetical proteins, similarly to the alignment-free search (**Figure 4**). These five hits retrieved by the use of two different sequence search methods among the hypothetical proteins could reveal the presence of additional A-domains remote homologues.

## Discussion

The potential usefulness of several graphical/numerical approaches to characterize genes and proteins for comparative analyses without the use of alignments has been recently reported by Randić *et al* [1,40,41]. We have extended this philosophy through the **TI2BioP** tool to characterize graphically and numerically large sequences databases [18]. The 2D Cartesian representation for genes and proteins and its simple numerical characterization were implemented in **TI2BioP** *version 1.0*, especially to deal with functional classification problems at low sequence similarity [8,28,42]. Our alignment-free models predictions based on graphical profiles have generally been used in cooperation with profile HMMs and experimental evidences [8,28].

In this work we highlighted a practical utility of the four-color maps accompanied with sensitive alignment procedures to detect a functional signal among a highly diverse protein domains dataset including a proteome. The four-color maps construction was based on a similar procedure carried out to the building of 2D Cartesian maps for protein sequences, previously used with success to detect functional signatures at low homology level [28,42].

Proteins four-color maps were modified by clustering the amino acids according to their physicochemical properties in four groups (polar, non-polar, acid and basic) labeled in the map with four colors. The numerical characterization of the four-color maps can describe homologous sequences (replacement between amino acids of similar properties) and remote homologous (important changes in the primary structure but still retaining the same biological function). While small changes in the sequence do not affect the topology of the map, this kind of amino acid substitution produces implicit numerical changes in the calculation of the TIs making possible the differentiation of the sequences. When an amino acid exchange occurs between different physicochemical groups of amino acids, this change affects the topology of the map and consequently affects significantly the TIs values estimation.

The TIs consist in the spectral moments series ($^{fc}\mu_0$-$^{fc}\mu_{16}$) describing the protein four-color maps. The topology of the protein four-color maps is determined by the sequence order and its amino acid composition (amino acid content according to the above-mentioned four groups). These two sequence features define the number and composition of the clusters formed in the map. The spectral moments series codify a range of information about the protein four-color maps that comprise the number of formed clusters in the map ($^{fc}\mu_0$) until the connectivity between the clusters in the map at different range ($^{fc}\mu_1$-$^{fc}\mu_{16}$). Our approach has a similar conceptual framework to the PseAAC introduced by Chou [33] but instead of using linear information (amino acid composition and sequence order) to get a vector representing the protein, four-color maps are built following similar rules but containing higher order information beyond the linearity of the

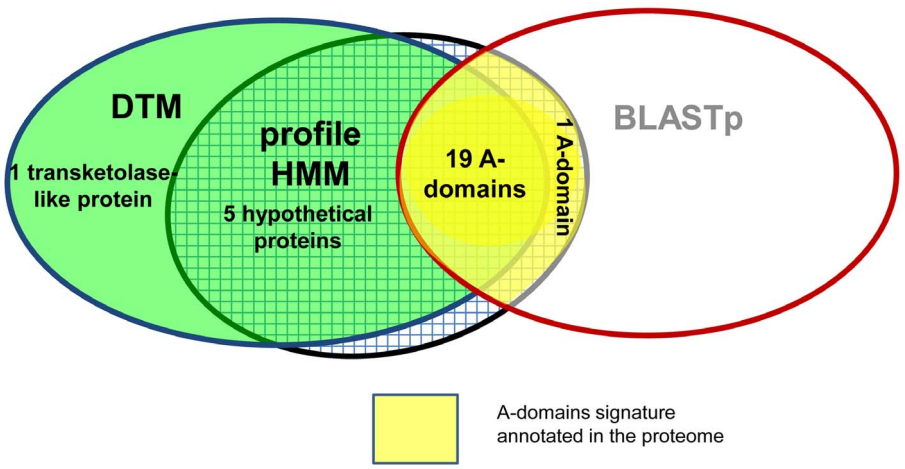



**Figure 4. Re-annotation of the A-domains in the proteome of *Microcystis aeruginosa* by using an ensemble of algorithms.**
doi:10.1371/journal.pone.0065926.g004

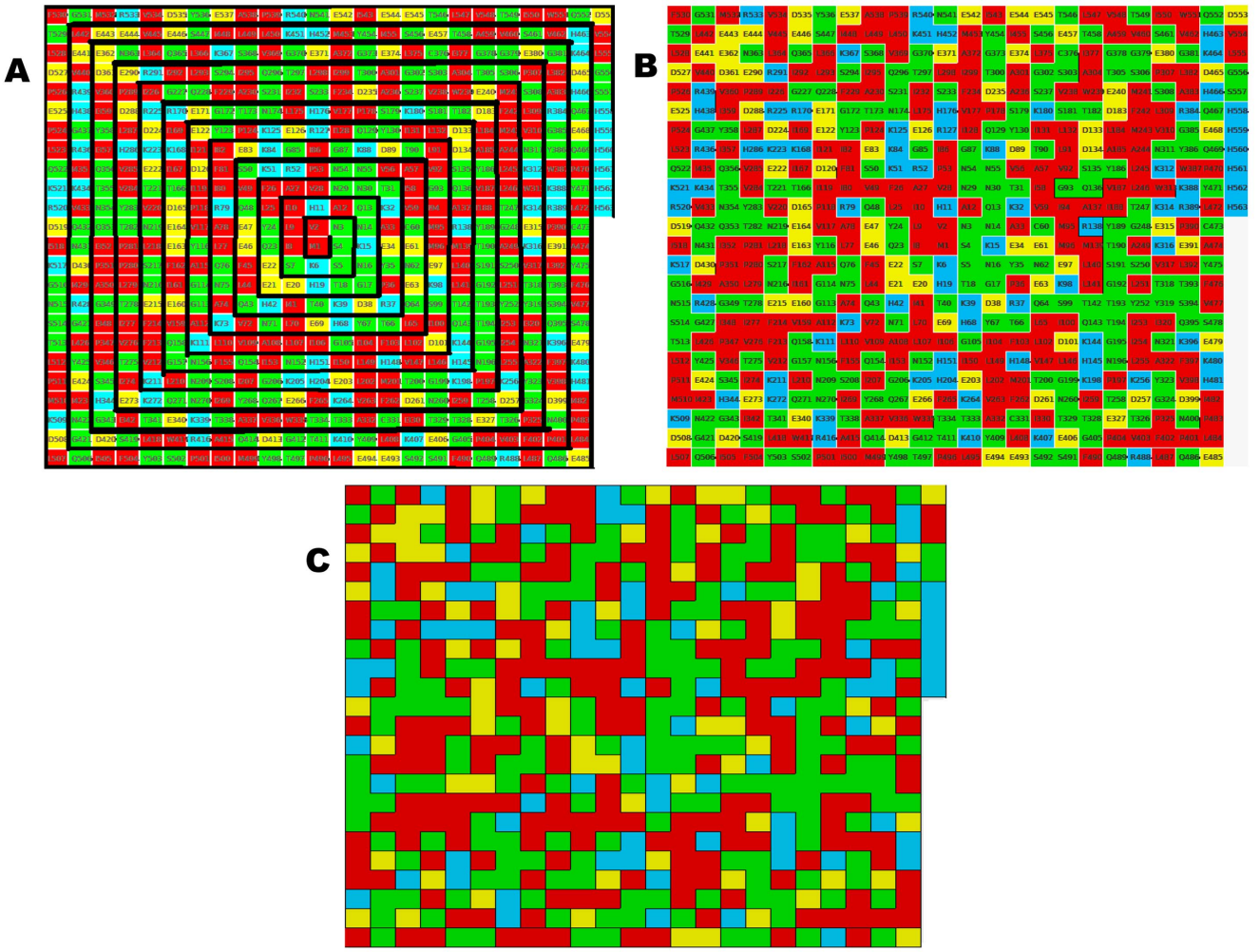



**Figure 5. Steps for the four-color map construction of 1 pdb AMU.** (A) Arranging the protein sequence into a square spiral. (B) Making up the clusters according to the amino acids properties: polar (green), non-polar (red), acid (yellow), basic (blue). (C) The final four-color map for pdb 1AMU.
doi:10.1371/journal.pone.0065926.g005

sequence. Afterwards, the topology of such 2D graphs is described by node adjacency matrices used to calculate the spectral moments series as TIs.

The spectral moments series ($^{fc}\mu_0$-$^{fc}\mu_{16}$) were used to develop several alignment-free models with linear and non-linear statistical techniques. DTM and ANN showed a better performance in classifying A-domains in respect to linear models supporting that the identification of protein signatures are better assessed with non-linear models. DTM was the best-reported alignment-free model due to the reasons given in the previous section. Consequently, it was applied to get other alignment-free models based on AAC and PseAAC to inspect their sensitivity to retrieve all A-domains members. Such DTM displayed lower classification rates than those reached by the four-color maps based models (**Table 4**). It seems that higher order patterns providing by the four-color maps are more effective in the detection of A-domains than linear sequence features driven by AAC and PseAAC. Therefore, the DTM based on four-color map patterns was selected to perform the alignment-free search for A-domains in the proteome of the cyanobacteria *Microcystis aeruginosa*.

Interestingly, DTM detected in the proteome two putative hits of A-domain signatures among the hypothetical proteins and later another three hypothetical proteins were detected as A-domains by the profile HMM (**Figure 4**). The sequence search methods based on profiles (graphical and alignment) were able to detect more hits than the 20 A-domains already annotated in the proteome, which were also detected by the multi-template BLASTp. Hypothetical proteins are greatly expanded in cyanobacteria and have been placed into the diversity of the nuclease superfamily by homology inference. Probably the graphical and HMM profiles detected signals of the A-domain signature among the diversity of the hypothetical proteins leading us to new variants of A-domains.

Both methods detected different additional hits as A-domains but they were found among the hypothetical proteins, which is a good clue for the presence of further A-domains remote homologues in the proteome of *Microcystis aeruginosa*. The use of an ensemble of methods provides more confidence to the predictions since each method exploits different features of the protein sequences. Four-color maps generate graphical patterns using the sequence order and the amino acid composition arranged into a 2D space. These graphical profiles are numerically described in a wide range of information by series of TIs, which characterize individually the sequences. Consequently, such TIs are flexible to be used for different classification problems (from high sequence identities until the twilight zone).

On the other hand, the profile HMM is based on amino acid positions conserved at low range through multiple sequence alignments in linear sequences. HMM profiles are proved sensitive tools for remote protein homology detection even when the sequence conservation is restricted to short motifs, as is the case of A-domains [16,43].

The ensemble of the three sequence search algorithms (DTM, multi-template BLASTp and profile HMM) provided the best solution for the search of remote homologues among a highly diverse protein class.

## Methods

### Computational methods

**TI2BioP** software *version 2.0* was used for the calculation of spectral moments as TIs associated with the protein four-color maps depicted below **(Figure 5).** Protein four-color maps are inspired on the Randic's DNA/RNA [44] and protein 2D graphical representations [12]; but instead of using the concept of virtual genetic code, we construct the spiral of square cells straightforward from the amino acid sequences. The four colors are assigned to the four amino acids classes (polar, non-polar, acid and basic) used previously by our group in Nandy's representation for proteins [28,45]. A node adjacency matrix is defined to calculate the spectral moments to describe the topology of these proteins colored maps **(Figure 6)**.

**Figure 5** shows how the four-color map for the first A-domain structurally characterized is built up. It belongs to the Gramicidin Synthetase cluster isolated from *Brevibacillus brevis* (pdb 1AMU). Each of the four colors is associated with each one of the amino acid groups: polar (green), non-polar (red), acid (yellow), basic (blue).

### Database

**Positive set.** 109 A-domain sequences from NRPS were collected from the major NRPS–PKS database (http://www.nii.res.in/nrps-pks.html) to conform the training set. The test set was made up of 29 A-domain sequences independently gathered from the subset of the NRPS-PKS hybrids (http://www.nii. res.in/nrps-pks.html). The sequence diversity among A-domains was explored comparatively using the Needleman-Wunsch (NW) algorithm.

**Negative set.** The starting group was made up for 8 871 protein sequences downloaded from the **CATH** (**C**lass, **A**rchitecture, **T**opology and **H**omology) domain database of protein structural families (version 3.2.0) (http://www.cathdb.info). We select the FASTA sequence database for all CATH domains sharing just the 35% of sequence similarity (<35% of sequence identity). The starting data was reduced to 8 854 CATH domains:

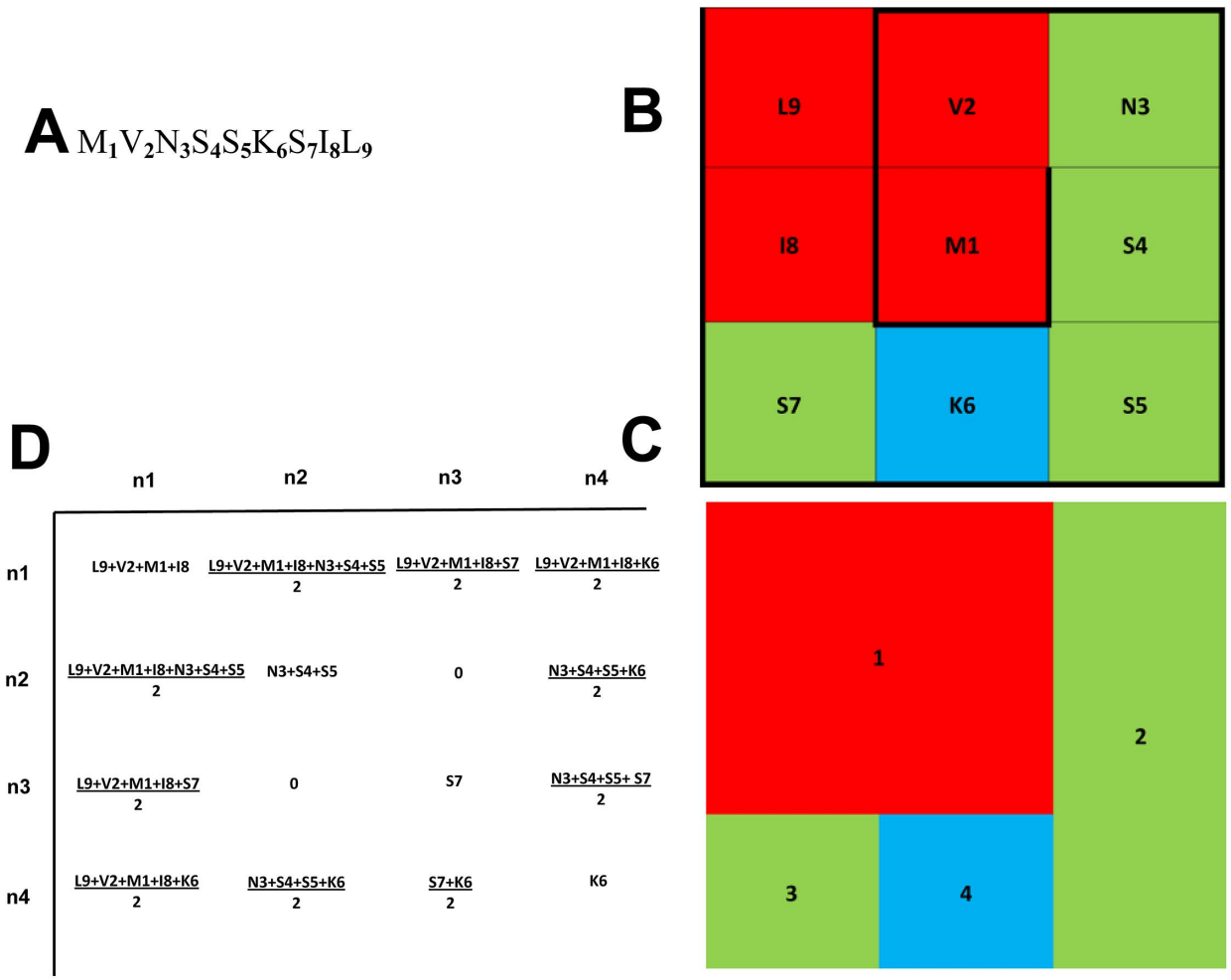



**Figure 6. From the protein sequence to its numerical characterization.** (A) The first nine aminoacids of pdb 1AMU. (B and C) Building the four-color map for A. (D) The definition of the node adjacency matrix derived from C the four-color map.
doi:10.1371/journal.pone.0065926.g006

17 cases were removed because they showed the A-domain signature when an *hmmsearch* was performed against the AMP-binding profile HMM (PF00501). The members of the test set (2 213 sequences) were selected taking out at random the 20% from the 8 854 CATH domains; the rest 6641 CATH domains were used to train the models.

Each A-domain and CATH domain sequence retrieved was labeled respecting its original database ID code (**File S11**).

## Numerical characterization of protein four-color maps through the spectral moments

The spectral moments are TIs calculated as the sum of the entries placed in the main diagonal of the bond adjacency matrix (**B**) between atoms for the small organic molecules. **B** is a square matrix of n x n row and column where its non-diagonal entries are ones or zeroes if the corresponding bonds or edges (n) share or not one atom. The different powers of **B** give the spectral moments of higher order to obtain the spectral moments series ($\mu_0$- $\mu_{15}$).

$$\mu_k = Tr\left[(B)^k\right] \quad (2)$$

Where Tr is called the trace and indicates the sum of all the values in the main diagonal of the matrices ${}^{k}\mathbf{B} = (\mathbf{B})^{k}$, which are the natural powers of **B** [46].

For the calculation of the spectral moments from the protein four-color maps, we consider each region of the map as a node made up for the amino acids clustering; two adjacent regions of the map sharing at least one edge (not a vertex) are connected. **B** is calculated in a similar way but instead of considering the adjacency relationships between bonds or edges, it is set between nodes. The number of nodes or clusters in the graph is equal to the number of rows and columns in **B**. Since a cluster is made up for several amino acids sharing similar physicochemical properties, the cluster is weighted with the sum of the individual properties (e.g. electrostatic charge (q)) of all amino acids placed in the cluster). The main diagonal of **B** was weighted with the average of the electrostatic charge (Q) between two adjacent clusters. The q values were taken from Amber 95 force field [47]. The calculation of the spectral moments up to the order k = 3 from the four colours maps is illustrated (downstream **figure 6**) using the first nine amino acids of pdb 1AMU ($M_1V_2N_3S_4S_5K_6S_7I_8L_9$). The **figure 6** represents the four-color map built up for these nine amino acids, as well as its cluster adjacency matrix. q values are represented in the matrix as the amino acids symbols (M = 1.91, V = 2.24, N = 2.07, S = 2.09, K = 2.254, I = 2.02, L = 1.91).

Expansion of expression (2) for k = 0 gives the ${}^{fc}\mu_0$, for k = 1 the ${}^{fc}\mu_1$ and for k = 2 the ${}^{fc}\mu_2$. The node adjacency matrix derived from this 2D map is described for each case

$${}^{fc}\mu_0 = Tr\left[(\mathrm{B})^0\right] = Tr\left(\begin{bmatrix} 8.09 & 7.17 & 5.09 & 5.17 \\ 7.17 & 6.25 & 0 & 4.25 \\ 5.09 & 0 & 2.09 & 2.17 \\ 5.17 & 4.25 & 2.17 & 2.25 \end{bmatrix}^0\right) = 4.0 \quad (2a)$$

$${}^{fc}\mu_1 = Tr\left[(\mathrm{B})^1\right] = Tr\left(\begin{bmatrix} 8.09 & 7.17 & 5.09 & 5.17 \\ 7.17 & 6.25 & 0 & 4.25 \\ 5.09 & 0 & 2.09 & 2.17 \\ 5.17 & 4.25 & 2.17 & 2.25 \end{bmatrix}^1\right) \quad (2b)$$

$$= 8.09 + 6.25 + 2.09 + 2.25$$

$${}^{fc}\mu_2 = Tr\left[(\mathrm{B})^2\right]$$

$$= Tr\left(\begin{bmatrix} 8.09 & 7.17 & 5.09 & 5.17 \\ 7.17 & 6.25 & 0 & 4.25 \\ 5.09 & 0 & 2.09 & 2.17 \\ 5.17 & 4.25 & 2.17 & 2.25 \end{bmatrix} \times \begin{bmatrix} 8.09 & 7.17 & 5.09 & 5.17 \\ 7.17 & 6.25 & 0 & 4.25 \\ 5.09 & 0 & 2.09 & 2.17 \\ 5.17 & 4.25 & 2.17 & 2.25 \end{bmatrix}\right)$$

$$= \begin{bmatrix} 169.5 & 124.8 & 63.0 & 95.0 \\ 124.8 & 108.5 & 45.7 & 73.2 \\ 63.0 & 45.7 & 34.9 & 35.7 \\ 95.0 & 73.2 & 35.7 & 54.5 \end{bmatrix} \quad (2c)$$

$$= 169.5 + 108.5 + 34.9 + 54.5$$

**TI2BioP** *version 2.0* arranges automatically all domain sequences (positive and negative sets) into four-colour maps and allows the calculation of spectral moments series (${}^{fc}\mu_k$). **File S12** shows the calculation of these indices to the positive and negative sets.

## Alignment-free models development with four-color maps TIs for A-domains detection

**Linear models. General Discrimination Analysis.** The General Discrimination Analysis (GDA) best subset was carried out for variable selection to build up the linear models [48–50]. All variable predictors were reviewed for finding the "best" possible sub model. The predictors 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 [51]. The model and variable selection was based on the revision of Wilk's ($\lambda$) statistic ($\lambda = 0$ perfect discrimination, being $0<\lambda<1$). The Fisher ratio (F) was also inspected to indicate the contribution of one variable to the discrimination between groups with a probability of error ($p$-level) $p(F)<0.05$.

**Non-linear methods. Decision Tree Models (DTM).** The development of the DTM was performed using the C&RT (Classification and Regression Trees)-style univariate split selection from the Classification Trees (CT) module of the STATISTICA 8.0 for Windows [19]. The C&RT examine all possible splits for each predictor variable at each node to find the split producing the largest improvement in goodness of fit. The prior probabilities were estimated for both groups with equal misclassification cost. The *Gini* index was used as a measure of goodness of fit and the FACT-style direct stopping was set to 0.1 as stopping rule to select the right-sized classification tree.

**Artificial Neural Networks (ANN).** We used the Multilayer Layer Perceptron (MLP) network architecture as the most popular network architecture in use today. The selection of the subset of predictors that is most strongly related to the response variable was supported on the *Feature and Variable Selection* analysis of the ANN module from *STATISTICA* software [19]. The right complexity of the network was selected by testing different topologies to the MLP while checking the progress against a selection set to avoid over-fitting during the two-phase (back propagation/conjugate gradient descent) training algorithm. The selection set was randomly extracted (10%) from the training set. The test set was the same used for GDA and DTM representing an external subset (not used during training algorithms) to check the final network performance [52].

### Evaluation of models' performance and validation procedure

The performance of the all alignment-free models was evaluated by several statistical measures commonly used for classification: accuracy, sensitivity, specificity and F-score (it reaches its best value at 1 and worst score at 0). The robustness of the classification model was verified by a 10-fold cross-validation (CV) procedure on the training set. The CV statistics for each of the ten samples were averaged to give the 10-fold estimate for the accuracy, sensitivity and specificity [53]. In addition, a test set made up for 2242 domains was selected to evaluate the prediction power of each model.

### Ensemble of methods for re-annotation of A-domains NRPS in the proteome Microcystis aeruginosa

We used an ensemble of three methods for the re-annotation of the *Microcystis aeruginosa* proteome considering its repertoire of A-domains signatures.

1. The graphical method represented by the alignment-free model (DTM) to perform the A-domain search in the proteome. Spectral moments series from the four-color maps were calculated for the proteome of *Microcystis aeruginosa* NIES-843 (6 311 annotated genes) and later a simple rule was applied to detect A-domain signatures ($^{fc}\mu_1 \geq 3817$ and $^{fc}\mu_2 \leq 11.12$).
2. A profile HMM for whole A-domain sequences was built as follows: (i) the 109 A-domain sequences used in training the alignment-free models were aligned by CLUSTALW [54], (ii) alignment was edited by Gblock software [55] to increase the alignment quality (iii), edited alignment was used as input for *hmmbuild* release 2.3.2 [14]. The generated profile HMM is used to search A-domains in the proteome of *Microcystis aeruginosa*.
3. The multiple-template BLASTp reported by the NRPS-PKS database developers for A-domain searches was used [15]. Multiple-template BLASTp consist in using each one of the 109 A-domains from the training set as template to evaluate each query of the proteome by BLASTp. BLOSUM62 scoring matrix, default values for gap penalties and E-value = 10 were set as BLASTp parameters and just the best matches were retrieved.

## Conclusions

The utility of graphical approaches in bioinformatics has been demonstrated by the introduction of the four-color maps and the TIs as a cooperative tool for detecting remote homologous of A-domains in the proteome of *Microcystis aeruginosa*. Since each sequence search method extract different features from the protein sequences, their integration allow a more exhaustive description of certain protein class and therefore provide a higher yield for the detection of remote protein homologous. The knowledge of the complete repertoire of A-domains in the proteome of cyanobacteria species may allow unraveling new NRPS clusters for the discovery of novel natural products with important biological activities.

## Supporting Information

**File S1 BLASTp (E-value = 10) search for A-domains using a single template against the whole dataset.**
(TXT)

**File S2 BLASTp (E-value = 0.05) search for A-domains using a single template against the whole dataset.**
(TXT)

**File S3 BLASTp (E-value = 0.01) search for A-domains using a single template against the whole dataset.**
(TXT)

**File S4 BLASTp (E-value = 0.001) search for A-domains using a single template against the whole dataset.**
(TXT)

**File S5 BLASTp (E-value = 0.0001) search for A-domains using a single template against the whole dataset.**
(TXT)

**File S6 BLASTp (E-value = 10) search for A-domains using multiple templates against the whole data set.**
(XLS)

**File S7 Profile HMM (E-value = 10) search for A-domains against the whole data set.**
(TXT)

**File S8 Alignment-free search for A-domain signatures in the proteome of *Microcystis aeruginosa*.**
(XLS)

**File S9 Multi-template BLASTp search for A-domains in the proteome of *Microcystis aeruginosa*.**
(XLS)

**File S10 HMM profile search for A-domain signatures in the proteome of *Microcystis aeruginosa*.**
(TXT)

**File S11 Database used in the study. Fasta files for training and test series of A and CATH domains.**
(ZIP)

**File S12 IDs, training and prediction series, values of the TIs for A and CATH domains.**
(XLS)

## Acknowledgments

Comments made by the Academic Editor Christos A. Ouzounis and the reviewers improved a previous version of this manuscript. Special thanks to Prof. Milan Randić to support the use of graphical/numerical approaches in bioinformatics.

## Author Contributions

Conceived and designed the experiments: GAC AA. Performed the experiments: GAC RMR EM. Analyzed the data: GAC RMR ASR AA. Contributed reagents/materials/analysis tools: AA VV RMR GdlR. Wrote the paper: GAC AA.

## References

1. Randic M, Zupan J, Balaban AT, Vikic-Topic D, Plavsic D (2011) Graphical representation of proteins. Chem Rev 111: 790–862.
2. Randic M (2004) Graphical representation of DNA as a 2-D map. Chem Phys Lett: 468–471.
3. Randic M, Zupan J, Vikic-Topic D (2007) On representation of proteins by star-like graphs. J Mol Graph Model 26: 290–305.
4. Randic M, Zupan J (2004) Highly compact 2D graphical representation of DNA sequences. SAR QSAR Environ Res 15: 191–205.
5. Nandy A (1994) Recent investigations into global characteristics of long DNA sequences. Indian J Biochem Biophys 31: 149–155.
6. Aguero-Chapin G, Varona-Santos J, de la Riva GA, Antunes A, Gonzalez-Vlla T, et al. (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.
7. Cruz-Monteagudo M, Gonzalez-Diaz H, Borges F, Dominguez ER, Cordeiro MN (2008) 3D-MEDNEs: an alternative "in silico" technique for chemical research in toxicology. 2. quantitative proteome-toxicity relationships (QPTR) based on mass spectrum spiral entropy. Chem Res Toxicol 21: 619–632.
8. Aguero-Chapin G, Sanchez-Rodriguez A, Hidalgo-Yanes PI, Perez-Castillo Y, Molina-Ruiz R, et al. (2011) An alignment-free approach for eukaryotic ITS2 annotation and phylogenetic inference. PLoS One 6: e26638.
9. Randic M, Zupan J (2001) On interpretation of well-known topological indices. J Chem Inf Comput Sci 41: 550–560.
10. Estrada E, Uriarte E (2001) Recent advances on the role of topological indices in drug discovery research. Curr Med Chem 8: 1573–1588.
11. Gonzalez-Diaz H, Perez-Montoto LG, Duardo-Sanchez A, Paniagua E, Vazquez-Prieto S, et al. (2009) Generalized lattice graphs for 2D-visualization of biological information. J Theor Biol 261: 136–147.
12. Randic M, Mehulic K, Vukicevic D, Pisanski T, Vikic-Topic D, et al. (2009) Graphical representation of proteins as four-color maps and their numerical characterization. J Mol Graph Model 27: 637–641.
13. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
14. Eddy SR (2009) A new generation of homology search tools based on probabilistic inference. Genome Inform 23: 205–211.
15. Ansari MZ, Yadav G, Gokhale RS, Mohanty D (2004) NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. Nucleic Acids Res 32: W405–413.
16. Jenke-Kodama H, Dittmann E (2009) Bioinformatic perspectives on NRPS/PKS megasynthases: advances and challenges. Nat Prod Rep 26: 874–883.
17. Welker M, von Dohren H (2006) Cyanobacterial peptides – nature's own combinatorial biosynthesis. FEMS Microbiol Rev 30: 530–563.
18. Molina R, Agüero-Chapin G, Pérez-González MP (2011) TI2BioP (Topological Indices to BioPolymers) *version 2.0.*: Molecular Simulation and Drug Design (MSDD), Chemical Bioactives Center, Central University of Las Villas, Cuba.
19. Statsoft (2008) STATISTICA 8.0 (data analysis software system for windows). version 8.0 ed.
20. Santana L, Uriarte E, González-Díaz H, Zagotto G, Soto-Otero R, et al. (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.
21. 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.
22. Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng 12: 85–94.
23. Hobohm U, Sander C (1995) A sequence property approach to searching protein databases. J Mol Biol 251: 390–399.
24. Wass MN, Sternberg MJ (2008) ConFunc--functional annotation in the twilight zone. Bioinformatics 24: 798–806.
25. Concu R, Podda G, Ubeira FM, Gonzalez-Diaz H (2010) Review of QSAR models for enzyme classes of drug targets: Theoretical background and applications in parasites, hosts, and other organisms. Curr Pharm Des 16: 2710–2723.
26. Strope PK, Moriyama EN (2007) Simple alignment-free methods for protein classification: a case study from G-protein-coupled receptors. Genomics 89: 602–612.
27. Ding YS, Zhang TL, Chou KC (2007) Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. Protein Pept Lett 14: 811–815.
28. Aguero-Chapin G, de la Riva GA, Molina-Ruiz R, Sanchez-Rodriguez A, Perez-Machado G, et al. (2011) Non-linear models based on simple topological indices to identify RNase III protein members. J Theor Biol 273: 167–178.
29. Cai YD, Liu XJ, Chou KC (2003) Prediction of protein secondary structure content by artificial neural network. J Comput Chem 24: 727–731.
30. Cai YD, Liu XJ, Chou KC (2002) Artificial neural network model for predicting protein subcellular location. Comput Chem 26: 179–182.
31. Cai YD, Liu XJ, Chou KC (2001) Artificial neural network model for predicting membrane protein types. J Biomol Struct Dyn 18: 607–610.
32. Shen HB, Chou KC (2008) PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. Anal Biochem 373: 386–388.
33. Chou KC (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 43: 246–255.
34. Kumar M, Thakur V, Raghava GP (2008) COPid: composition based protein identification. In Silico Biol 8: 121–128.
35. Chou KC (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21: 10–19.
36. Boekhorst J, Snel B (2007) Identification of homologs in insignificant blast hits by exploiting extrinsic gene properties. BMC Bioinformatics 8.
37. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic Local Alignment Search Tool. J Mol Biol 215: 403–410.
38. 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: W273–279.
39. Koetschan C, Forster F, Keller A, Schleicher T, Ruderisch B, et al. (2009) The ITS2 Database III--sequences and structures for phylogeny. Nucleic Acids Res.
40. Randic M (2012) Very efficient search for protein alignment–VESPA. J Comput Chem 33: 702–707.
41. Randic M (2013) Very efficient search for nucleotide alignments. J Comput Chem 34: 77–82.
42. Aguero-Chapin G, Perez-Machado G, Molina-Ruiz R, Perez-Castillo Y, Morales-Helguera A, et al. (2011) TI2BioP: Topological Indices to BioPolymers. Its practical use to unravel cryptic bacteriocin-like domains. Amino Acids 40: 431–442.
43. Ansari MZ, Sharma J, Gokhale RS, Mohanty D (2008) In silico analysis of methyltransferase domains involved in biosynthesis of secondary metabolites. BMC Bioinformatics 9: 454.
44. Randic M, Lers N, Plavšić D, Basak S, Balaban A (2005) Four-color map representation of DNA or RNA sequences and their numerical characterization. Chemical Physics Letters 407: 205–208.
45. Aguero-Chapin G, Gonzalez-Diaz H, Molina R, Varona-Santos J, Uriarte E, et al. (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.
46. 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.
47. Cornell WD, Cieplak P, IBayly C, Gould IR, Merz KWJ, et al. (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 117: 5179–5197.
48. Marrero-Ponce Y, Castillo-Garit JA, Olazabal E, Serrano HS, Morales A, et al. (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.
49. 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.
50. Ponce YM, 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.
51. Kutner MH, Nachtsheim CJ, Neter J, Li W (2005) Standardized Multiple Regression Model. Applied Linear Statistical Models. Fifth ed. New York: McGraw Hill. 271–277.
52. The MathWorks I, editor (2004) Neural network toolbox users guide for use with MATLAB. Massachusetts: The Mathworks Inc.
53. Rivals I, Personnaz L (1999) On cross validation for model selection. Neural Comput 11: 863–870.
54. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680.
55. Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56: 564–577.