Chapter 10

TI2BioP — Topological Indices to BioPolymers. A Graphical–Numerical Approach for Bioinformatics

Guillermin Agüero-Chapin, Reinaldo Molina-Ruiz, Gisselle Pérez-Machado, Vitor Vasconcelos, Zenaida Rodríguez-Negrin and Agostinho Antunes

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/61887

Abstract

We developed a new graphical–numerical method called TI2BioP (Topological Indices to BioPolymers) to estimate topological indices (TIs) from two-dimensional (2D) graphical approaches for the natural biopolymers DNA, RNA and proteins The methodology mainly turns long biopolymeric sequences into 2D artificial graphs such as Cartesian and four-color maps but also reads other 2D graphs from the thermodynamic folding of DNA/RNA strings inferred from other programs. The topology of such 2D graphs is either encoded by node or adjacency matrixes for the calculation of the spectral moments as TIs. These numerical indices were used to build up alignment-free models to the functional classification of biosequences and to calculate alignment-free distances for phylogenetic purposes. The performance of the method was evaluated in highly diverse gene/protein classes, which represents a challenge for current bioinformatics algorithms. TI2BioP generally outperformed classical bioinformatics algorithms in the functional classification of Bacteriocins, ribonucleases III (RNases III), genomic internal transcribed spacer II (ITS2) and adenylation domains (A-domains) of nonribosomal peptide synthetases (NRPS) allowing the detection of new members in these target gene/protein classes. TI2BioP classification performance was contrasted and supported by predictions with sensitive alignment-based algorithms and experimental outcomes, respectively. The new ITS2 sequence isolated from *Petrakia* sp. was used in our graphical–numerical approach to estimate alignment-free distances for phylogenetic inferences. Despite TI2BioP having been developed for application in bioinformatics, it can be extended to predict interesting features of other biopolymers than DNA and protein sequences. TI2BioP version 2.0 is freely available from http://ti2biop.sourceforge.net/.

Keywords: 2D graphs, Topological indices, Alignment-free models, Bioinformatics

1. Introduction

Graph theory has been successfully applied in several branches of science such as mathematics, physics, chemistry, biochemistry, biology and computer science to visualize complex relationships. A graph is a collection of vertices or nodes and a compilation of edges that connect pairs of vertices. They have been deeply studied to analyse pairwise relationships in a data collection [1].

Graph theory allowed the development of chemical graph theory (CGT) to explore the chemical molecular structure by combinatorial and topological approaches that lead to the calculation of mathematical descriptors [2]. The molecular topology is simplified in graphs where its vertices and edges represent the atoms and bonds, respectively. Thus, molecular descriptors from the graph representing an approximation of the molecular structure can be estimated to carry out quantitative-structure-activity/property relationship (QSAR/QSPR). These numerical indices have been traditionally used in QSAR/QSPR studies for drug discovery and design in medicinal chemistry [3, 4].

With the arrival of the genomics and proteomics era, the CGT has been extended to characterize long biopolymeric strings such as DNA/RNA and proteins, for comparative analyses without the use of sequence alignments. The monomers (nucleotides and amino acids) of the natural biopolymers can play the role of nodes while the edges of the graph are represented by covalent bonds, hydrogen bridges, electrostatic interactions, van der Waals bonds and so on [5-7]. Thus, the structure of complex biopolymers can be simplified into the topology of a graph to provide useful insights into such molecular systems. The graphs representing molecular systems may be described using numerical descriptors like the so-called topological indices (TIs) [8]. TIs encode information about the connections between atoms in the molecule and the properties for the connected atoms [9]. In this way, they can also be applied to characterize natural biopolymers like DNA, RNA and protein sequences [10].

The use of TIs to characterize numerically biosequences to perform massive analyses without alignments is an active research topic in bioinformatics [5-7]. To determine the TIs for the natural biopolymers, we build a graph as it was described previously. There are various types of TIs depending of the dimensionality (D) of the biopolymer representation. One-dimensional (1D) representation of biosequences depicts the linear sequence order, while two-dimensional (2D) and three-dimensional (3D) representations are related to sequence arrangement or geometry into these spaces [11-13]. The 2D biopolymer graphs have grabbed special attention due to fact that they have been very effective in exploring similarities/dissimilarities among DNA and protein sequences despite not representing their real structure [2]. So far, the 2D artificial representations for DNA and protein sequences with higher potentialities in bioinformatics are the spectrum-like, star-like, Cartesian-type and four-color maps [2, 14-17]. These DNA/RNA and protein maps can generally reveal higher-order useful information contained beyond the primary structure, i.e. nucleotide/amino acid distribution into a 2D space. Such graphical features can be quantified by the TIs to easily compare a great number of sequences/maps [18-21].

Regardless of the biopolymer representation type, the definition of an adjacency matrix is mandatory for the calculation of any TI. 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]. Later, several algorithms can be applied on the adjacency matrix to provide different TI types such as the Winner index (*W*) [24], first defined in a chemical context; and others like Randić invariant (χ) [25], Balaban index (*J*) [26], Broto–Moreau autocorrelation (ATSd) [27] and the spectral moments introduced by Estrada [28]. The spectral moments were defined as the sum of main diagonal entries of the different powers of the bond adjacency matrix [29]. Spectral moments were implemented in the TOPS-MODE (topological substructural molecular design) program [30] and have been widely validated by many authors to encode the structure of small molecules in QSAR studies [31-33]. Despite the versatility of the spectral moments in QSAR studies, they have been poorly used to describe biopolymers structures except when they promoted the arising of the Estrada folding index (*I*3) for proteins [34, 35] or when they were redefined as stochastic spectral moments by González-Diaz et al. to numerically characterize biopolymeric systems, i.e. the protein surface of human rhinoviruses [36], Arc repressors [37], kinases [38] and different types of biological complex networks [10]. The stochastic spectral moments are implemented in the MARCH-INSIDE (Markov chain invariants for network selection and design) methodology [39] and can be estimated from star-like and Cartesian-type representations for DNA and protein sequences [6, 40, 41]. Thus, the first reported alignment-free models based on a graphical–numerical approach to annotate biological functions in biosequences were built using the MARCH-INSIDE software [40, 42, 43]. However, such predicting models were more illustrative than practical for the bioinformatics. They were built and tested on small-sized datasets and generally without considering the degree of similarity among their members and data benchmarking to evaluate the TIs as alignment-free predictors [40, 42, 43].

On the other hand, stochastic spectral moment's calculation is mathematically more complicated than the original definitions by Estrada [28]. Stochastic spectral moments rely on defining Markov chain states over the starting node adjacency matrix that are later powered at different orders, while the original spectral moments are derived directly from the powering of the bond or edge adjacency matrix weighted with some bond property [30, 39].

Considering these shared previous experiences about the potentialities of the graphical–numerical methods for bioinformatics, we aim for the development of a new methodology called TI2BioP (Topological Indices to BioPolymers) to extend the original spectral moments as simple TIs to characterize numerically 2D artificial representations for the DNA/RNA and proteins structure [5, 44]. These TIs represent alignment-free predictors to detect functional signatures in members of gene and protein classes. Its practical importance for bioinformatics consisted of dealing with gene/protein classes sharing low sequence similarity and in estimating alignment-free distances for inferring phylogenetic relationships [21].

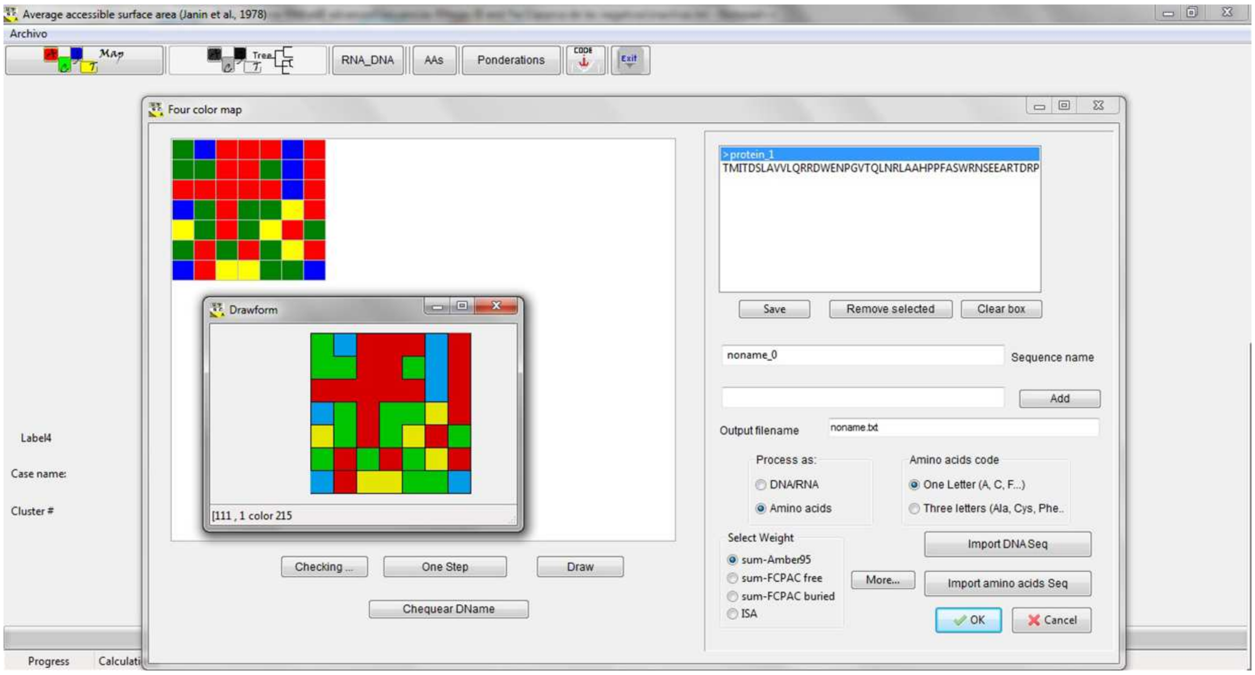
Traditionally, the prediction of the biological function, 2D and 3D structure of a query gene or protein has relied on similarity measures provided by alignment algorithms, to other recorded members of the family. All alignment-based methods, the dynamic programming algorithms implemented by Needleman–Wunsch [45] and Smith–Waterman [46], the heuristic algorithm for basic local alignment search tool (BLAST) [47] and the probabilistic hidden Markov models (HMM) [48, 49] have a friendly interface to search structural and functional sequence classifications, but they may fail in detecting gene/protein members that share low similarity to others of the family [21, 50, 51]. There are several evidences showing a low reliability for the biological functional prediction when protein families have pairwise sequence similarities below 50% [50, 52, 53]. In addition, inaccurate alignments for proteins that share less than 30% to 40% of identity, which is commonly called the ″twilight zone″ for the alignment algorithms, have been reported [50, 54]. Therefore, the reliability of phylogenetic inferences is also affected by failures of the multiple sequence alignment (MSA) algorithms when the taxa represented by sequences have greatly diverged [54]. Consequently, several alignment-independent approaches have been developed to overcome this limitation for an effective functional annotation [55, 56] and for reliable phylogenetic inferences in highly diverse gene/protein families [55, 57]. Most of the alignment-free classifiers have been based on amino acid composition to annotate protein functions [51, 55, 56]. It is very likely that the most popular alignment-free approach is Chou’s concept of pseudo–amino acid composition (PseAAC) that reflects the importance of the sequence order effect in addition to the amino acid composition to improve the prediction quality of protein cellular attributes [58]. On the other hand, the alignment-independent approaches reported for phylogenetic tree reconstruction have mostly been based on patterns discovered in unaligned sequences [59], amino acid composition [55] and a kernel approach for evolutionary sequence comparison [57].

While alignment methods have improved their sensitivity to detect functional and evolutionary signals in query sequences and species by using several strategies [60-62] and, on the other side, various alignment-free approaches have been reported to address the same drawback, there is still room for the development of new alignment-free biosequence descriptors. In this sense, graphical–numerical methods have been poorly explored as alignment-free tools in bioinformatics, to face current alignment algorithm limitations [2, 54]. Here, we summarize our experience in this subject through the application of TI2BioP to predict the functions of natural biopolymers (DNA and proteins) in classes representing a challenge for alignment algorithms as well as its introduction into the molecular evolutionary field.

2. Methods

2.1. TI2BioP software

TI2BioP was mainly developed from the TOPS-MODE methodology [30] for the estimation of the spectral moments series as TIs, but it takes advantage of the MARCH-INSIDE program platform [39]. It was built up on object-oriented Free Pascal IDE Tools (Lazarus) running on either a Windows or Linux operating system. TI2BioP has a friendly interface allowing users to introduce multiple fasta files containing either DNA or protein sequences to select the biopolymer 2D representation type and the calculation of TIs. We released version 2.0 of the software that can be freely downloaded from http://ti2biop.sourceforge.net/. This version contains two main types of 2D artificial representations, one based on Cartesian representation for DNA strings introduced by Nandy [63] and the other inspired by the four-color maps reported by Randic [64, 65] (Figure 1).



Average accessible surface area (Janin et al., 1978)

Four color map

>protein_1
TMITDSLAVVLQRRDWENPGVTQLNRLAAHPPFASWRNSEEARTDRP

Sequence name: noname_0

Output filename: noname.txt

Process as: ○ DNA/RNA ● Amino acids

Amino acids code: ● One Letter (A, C, F...) ○ Three letters (Ala, Cys, Phe..

Select Weight: ● sum-Amber95 ○ sum-FCPAC free ○ sum-FCPAC buried ○ ISA

Drawform

[111 , 1 color 215

Figure 1. TI2BioP window view of the (Topological Indices to BioPolymers) software for the representation of protein four-color maps

These two 2D artificial graphs implemented in TI2BioP can be applied to nucleotide and amino acid strings as well as to the spectral moments calculations for each type of 2D DNA and protein maps [44]. It is noteworthy that the 2D Cartesian representation was extended to proteins by our group [40] and protein four-color maps were modified according to the amino acid clustering proposed in ref. [40]. Such four-color map modifications allow the speeding up of graph-building and facilitates the calculation of spectral moments as TIs [66].

TI2BioP can also import files containing 2D structures inferred by other DNA/RNA folding algorithms, e.g. Mfold implemented in the RNA structure software [67], for the calculation of the spectral moments as TIs. TI2BioP automatically represents natural biopolymers as 2D graphs and straightforward calculates spectral moments series (TIs) to be used either for statistical classification techniques in building alignment-free models for functional classification or for deriving several alignment-free distance matrices, e.g. Euclidean, Jensen–Shannon, Hamming and Minkowsk for phylogenetic purposes (Figure 2).

2.2. Database

To evaluate the performance and efficacy of our graphical–numerical approach TI2BioP to detect DNA and protein signatures and to infer phylogenetic relationships, four gene/protein families having low sequence similarity among their members were selected. The gene/protein classes targeted were:

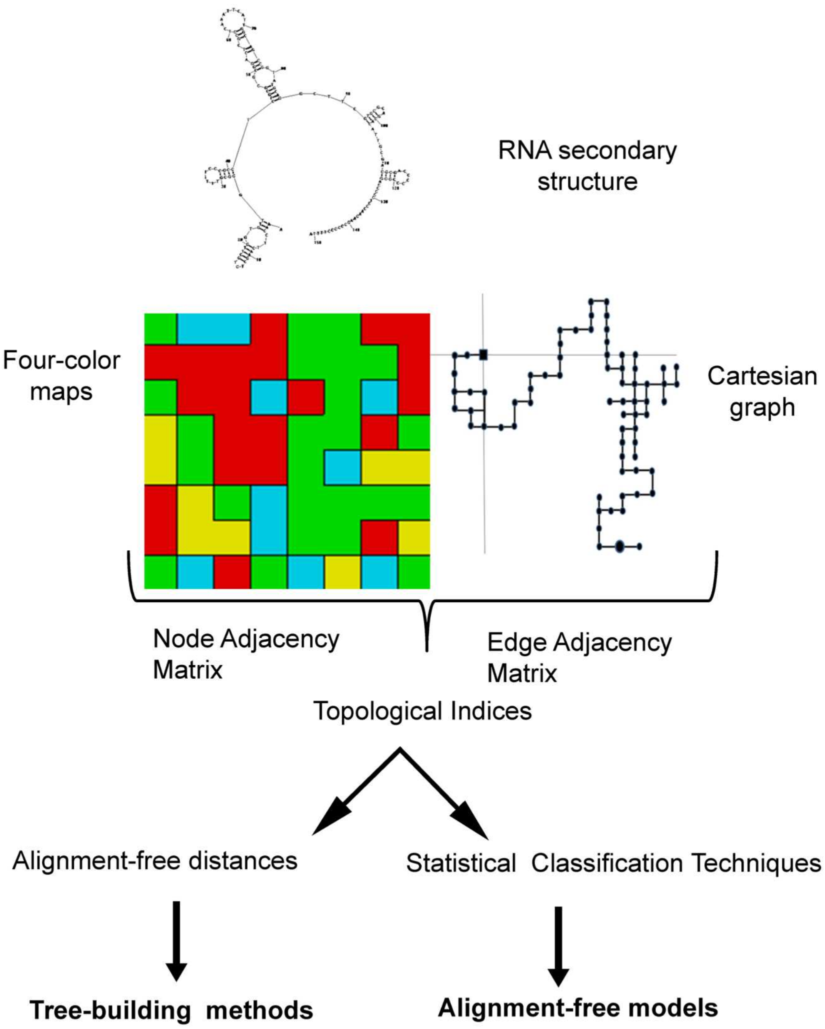


Figure 2. Workflow for the calculation of the topological indices by TI2BioP (Topological Indices to BioPolymers) from several 2D graphs for DNA, RNA and proteins.

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 [60] and BACTIBASE [68].

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 4355 ITS2 sequences 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 with respect to the study case.

Protein control groups:

1. Sequences from class, architecture, topology and homology (CATH) domain database (version 3.2.0) (http://www.cathdb.info) sharing only 35% sequence similarity were selected to provide a functional representation and avoid structural redundancy. This group was used as a control to develop alignment-free models to recognize bacteriocin-like and A-domain sequences.

2. High-resolution proteins in a structurally nonredundant and representative subset from the Protein Data Bank (PDB) made up of enzymes and nonenzymes were also used. This protein subset was used as a control group to develop alignment-free models to detect RNase III enzymes [50].

3. A nonredundant subset containing both 5′- and 3′-untranslated region (UTR) sequences from the eukaryotic mRNAs database: UTRdb (http://www.ba.itb.cnr.it/UTR/). It was selected as a control group to identify ITS2 members.

3. Results

This section summarizes the main results derived from the application of TI2BioP to the functional classification of protein bacteriocins [5], RNase III [69], ITS2 [21] and NRPS A-domains [66]. All these classes show high sequence divergence among their members, which represent a handicap for the good performance of alignment algorithms. In particular, the high sequence divergence among fungal ITS2 genomic fragments has been useful for fungi identification at the genus and species level. However, such sequence diversity is not suitable for reconstructing phylogenies at a higher taxonomical level. The use of simple alignment-free classifiers, like the topological indices (TIs), containing information about the sequence/structure of the natural biopolymers may reveal a useful approach for the gene/protein functional predictions and for assessing the phylogenetic relationships at high taxonomic levels in fungal species by using the ITS2 gene class.

The TI2BioP software provides TIs (spectral moments series) that are used as input predictors for statistical classification techniques and machine-learning methods to develop alignment-free models (Figure 2). Models were statistically tested by cross- and external-validation procedures. Their usefulness was proved by identifying new members belonging to each studied gene/protein class. Such alignment-free detections were either supported by experimental evidences or by sensitive alignment methods.

Table 1 displays the alignment-free models with the best performance for the functional classification of each target gene/protein family and the procedure carried out to achieve a consensus functional prediction of new members by such models. The functional annotation of the new members resulted from the prediction agreement among the graphical–numerical based models, experimental evidences and alignment algorithms. The 2D Cartesian protein representation and its derived TIs could unravel the Cry 1Ab C-terminal domain from *Bacillus thuringiensis′* endotoxin as a bacteriocin-like protein. The bactericide action of this domain was only confirmed by experimental evidences; no alignment algorithm could anticipate such activity [5]. In addition, new ITS2 and RNase III members were registered using alignment-free models based on the same biopolymer representation [21, 69]. The predictions of these two members were verified through enzymatic assay for the new RNase III member and by evaluating both queries against profiles HMM (Table 1). The amino acid clustering strategy according to their physicochemical properties to build 2D Cartesian protein maps was extended to generate a nonclassical profile HMM with higher prediction accuracy to detect RNase III members than classical profiles [69].

The effectiveness of the presented graphical–numerical approach in bioinformatics was also demonstrated by the introduction of protein four-color maps and TIs to detect A-domains despite their sequence diversity. A DTM based on this approach was chosen as the best alignment-free model to identify the A-domain signature (Table 1).

Gene/protein class	Control group	2D-graph type	Best-reported alignment-free model	Newly detected members	Prediction procedure
Protein Bacteriocins	CATH domains	Cartesian	GDA	Cry 1Ab C-terminal domain *Bacillus thuringiensis*	1. Alignment-free prediction 2. Experimental evidences
Genomic ITS2	5′and 3′UTRs	Cartesian and RNA Secondary Structure	ANN	ITS2 genomic *Petrakia* sp.	1. Alignment-free prediction 2. Alignment-based prediction
RNase III	Nonredundant subset PDB	Cartesian	DTM	RNase III *E. coli BL21*	1. Alignment-free prediction 2. Alignment-based prediction 3. Experimental evidences
A-domain NRPS	CATH domains	Four-color map	DTM	5 hits in the proteome of *Microcystis aeruginosa*	Pending registration

GDA, general discrimination analysis; ANN, artificial neural networks; DTM, decision tree models; ITS2, internal transcribed spacer; RNase III, ribonuclease III; A-domains, adenylation domains; CATH, class, architecture, topology and homology; PDB, Protein Data Bank; UTRs, untranslated regions.

Table 1. Best reported alignment-free models for the functional classification of each gene/protein family studied. Newly detected members of each gene/protein class and the procedure carried out for their definitive functional prediction

Its performance was contrasted to other different alignment-free approaches and homology-search methods in detecting A-domains on the same dataset. The Web server PseAAC (http://www.csbio.sjtu.edu.cn/bioinf/PseAAC/) was used to generate DTM based on other alignment-free features like amino acid composition (AAC) and pseudo amino acid composition (PseAAC) [70]. The DTM generated by four-color maps outperformed the DTM supported by AAC and PseAAC features (Table 2).

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 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 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	

CV, Cross-validation; DTM, decision tree model.

Accuracy (Acc), Sensitivity (Sv), Specificity (Sp) and *F*-score are classification quality measures. *F*-score values range between 0 and 1.

Table 2. 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

On the other hand, the alignment-free search of A-domains was also compared to homology-based methods such as single-template BLASTp, multitemplate BLASTp and profile HMM. These alignment-based algorithms show, by definition, different sensitivities to recognize distant homologs and therefore may provide different false-positive rates. Table 3 shows the classification results provided by different sequence-search methods including alignment-free (four-color maps, AAC and PseAAC) and homology-based (HMM, multitemplate BLASTp and BLASTp) approaches on the same dataset (138 A-domains + 8854 CATH domains). The DTM built from four-color maps TIs, HMM and multitemplate BLASTp identified all A-domains among the diversity of the dataset with no false-positives at nonstringent conditions (*E* value = 10).

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
Multitemplate 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

DTM, decision tree models; AAC, amino acid composition; PseAAC, pseudo–amino acid composition; HMM, hidden Markov models; BLAST, basic local alignment search tool; *E* value, a classification threshold related to the matches found by chances in alignment algorithms.

Table 3. True positives *vs.* false positives in the A-domain detection for different sequence-search methods among the overall dataset involved in the study

Considering the excellent performance of these three previous sequence-search methods, they were applied in cooperation to provide the most reliable exploration of the A-domain repertoire of NRPS in the *M. aeruginosa* proteome. DTM based on four-color map TIs detected two putative A-domain signatures among the proteome's hypothetical proteins while another three hypothetical proteins were detected as A-domains by the profile HMM. Sequence-search methods based on profiles (graphical and alignment) were able to detect five more hits than the 20 A-domains already annotated in the proteome, which were confirmed by the multitemplate BLASTp (Figure 3). These matches could reveal the presence of additional A-domain remote homologous, which would not have been detected by applying a single algorithm.

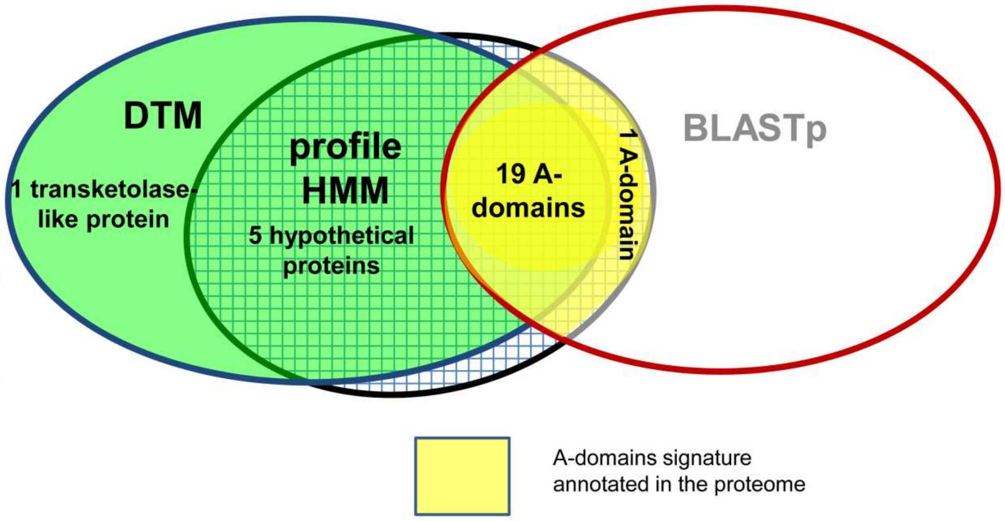

DTM
1 transketolase-like protein
profile HMM
5 hypothetical proteins
19 A-domains
1 A-domain
BLASTp
A-domains signature annotated in the proteome

Figure 3. Reannotation of the A-domains in the proteome of *M. aeruginosa* by using an ensemble of algorithms (DTM based on four-color maps, profile HMM and multitemplate BLASTp).

The performance of DTM built from four-color map TIs was contrasted with sensitive alignment procedures like multitemplate BLASTp and HMM. Similarly, the other alignment-free models such as GDA and ANN relying on Cartesian and thermodynamic TIs were also compared to InterPro and HMM profiles for functional detection of the chosen gene/protein classes (Tables 1 and 4).

	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

GDA, general discrimination analysis; ANN, artificial neural networks; DTM, decision tree models; HMM, hidden Markov models; MAFFT, multiple alignment based on fast Fourier transform; InterPro, a web resource that combines different protein signature recognition methods.

Table 4. 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

The TIs supplied by TI2BioP are also used to estimate alignment-free distances that can be introduced into tree-building methods, e.g. unweighted pair group method with arithmetic mean (UPGMA), neighbor joining method (NJ) and minimum evolution (ME) to infer evolutionary relationships (Figure 2).

The newly predicted ITS2 sequence, isolated from the fungus *Petrakia* sp., was used by clustering techniques applied for the first time to the alignment-free estimation of phylogenetic inferences (Table 1). The *Petrakia* sp. fungal isolate was placed inside the *Pezizomycotina* subphylum and the *Dothideomycetes* class by the inference agreement of classical genetic distances and the alignment-free distances based on TIs (Figure 4). We concluded that our graphical–numerical approach is effective to construct distance-trees containing relevant biological information with an evolutionary significance [21].

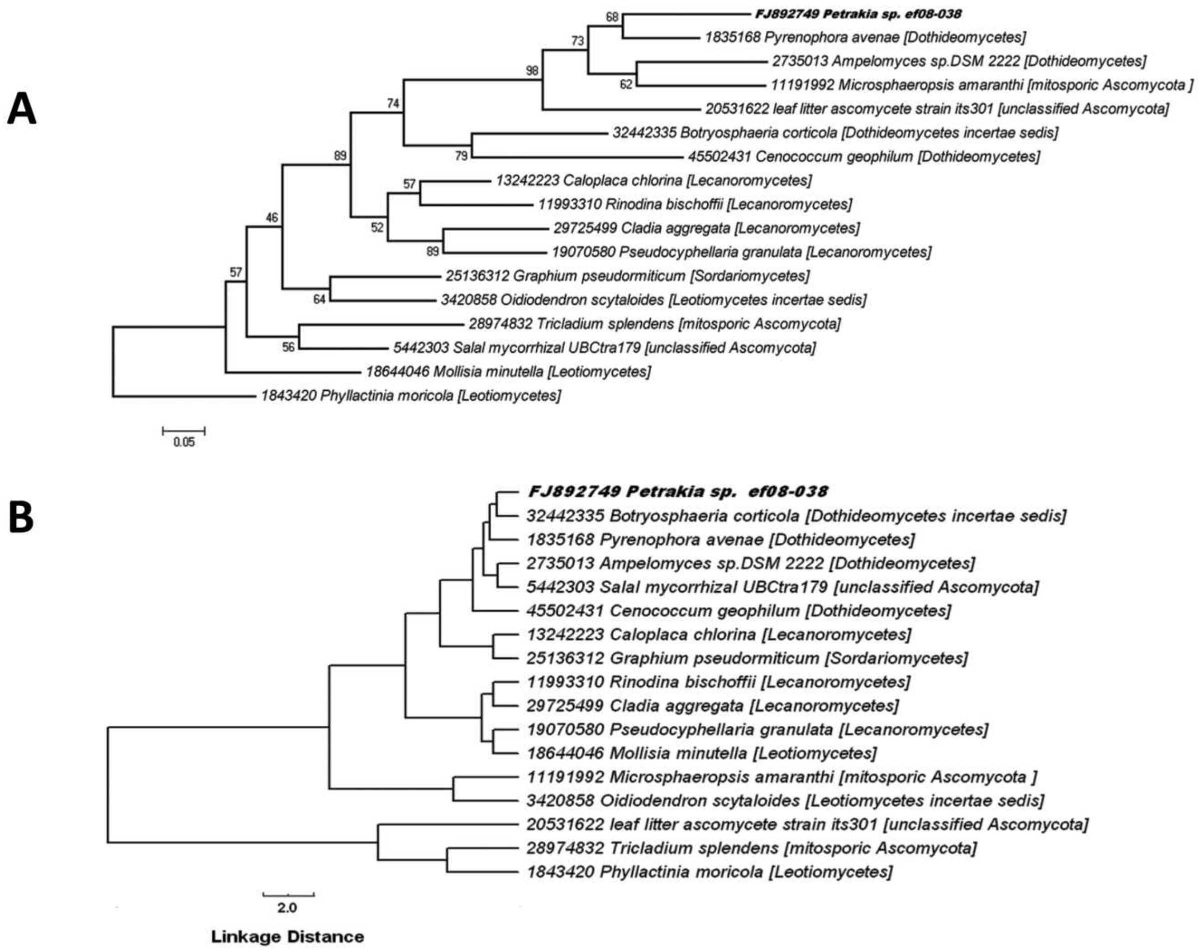

A
68
73
98
62
74
79
89
57
52
89
46
57
64
56
FJ892749 Petrakia sp. ef08-038
1835168 Pyrenophora avenae [Dothideomycetes]
2735013 Ampelomyces sp.DSM 2222 [Dothideomycetes]
11191992 Microsphaeropsis amaranthi [mitosporic Ascomycota]
20531622 leaf litter ascomycete strain its301 [unclassified Ascomycota]
32442335 Botryosphaeria corticola [Dothideomycetes incertae sedis]
45502431 Cenococcum geophilum [Dothideomycetes]
13242223 Caloplaca chlorina [Lecanoromycetes]
11993310 Rinodina bischoffii [Lecanoromycetes]
29725499 Cladia aggregata [Lecanoromycetes]
19070580 Pseudocyphellaria granulata [Lecanoromycetes]
25136312 Graphium pseudormiticum [Sordariomycetes]
3420858 Oidiodendron scytaloides [Leotiomycetes incertae sedis]
28974832 Tricladium splendens [mitosporic Ascomycota]
5442303 Salal mycorrhizal UBCtra179 [unclassified Ascomycota]
18644046 Mollisia minutella [Leotiomycetes]
1843420 Phyllactinia moricola [Leotiomycetes]
0.05
B
FJ892749 Petrakia sp. ef08-038
32442335 Botryosphaeria corticola [Dothideomycetes incertae sedis]
1835168 Pyrenophora avenae [Dothideomycetes]
2735013 Ampelomyces sp.DSM 2222 [Dothideomycetes]
5442303 Salal mycorrhizal UBCtra179 [unclassified Ascomycota]
45502431 Cenococcum geophilum [Dothideomycetes]
13242223 Caloplaca chlorina [Lecanoromycetes]
25136312 Graphium pseudormiticum [Sordariomycetes]
11993310 Rinodina bischoffii [Lecanoromycetes]
29725499 Cladia aggregata [Lecanoromycetes]
19070580 Pseudocyphellaria granulata [Lecanoromycetes]
18644046 Mollisia minutella [Leotiomycetes]
11191992 Microsphaeropsis amaranthi [mitosporic Ascomycota]
3420858 Oidiodendron scytaloides [Leotiomycetes incertae sedis]
20531622 leaf litter ascomycete strain its301 [unclassified Ascomycota]
28974832 Tricladium splendens [mitosporic Ascomycota]
1843420 Phyllactinia moricola [Leotiomycetes]
2.0
Linkage Distance

Figure 4. Higher-level phylogenetic analysis to infer the class of the fungal isolate *Petrakia* sp. (A) Neighbor-joining tree based on ITS2 sequences using the substitution Kimura 2-parameter (K2P). (B) Neighbor-joining tree clustering based on the Euclidean distance calculated from the four-color map TI values.

4. Discussion

The results shown in the previous section were motivated by related works carried out with the MARCH-INSIDE methodology. We have previously reported the 2D-Cartesian representation for proteins and its numerical characterization through stochastic sequence descriptors calculated by MARCH-INSIDE [40], to annotate biological functions in gene/protein classes. Thus, we published the first alignment-free model built up with stochastic TIs to functionally classify polygalacturonase (PG) members from plants [40]. Despite the fact that PG members were detected with high accuracy by our reported alignment-free model, classical alignment procedures also did this due to PG proteins showing a high sequence similarity, while not representing a challenge for alignment algorithms. This study opened a door to the application of graphical methods in bioinformatics for the detection of functional signatures in protein families; however, its application was more illustrative than useful to overcome the limitations of alignment algorithms [18, 40].

The 2D-Cartesian protein representation was also numerically characterized through stochastic spectral moments to detect a particular RNAse III member from *Schizosaccharomyces pombe* (Pac 1) among the diversity of this class [71]. Although alignment algorithms have demonstrated a low amino acid identity (20%–40%) between Pac 1 and other typical RNases III, the Pac 1 protein shows a remarkable ribonuclease activity [72]. This fact prompted the application of our graphical–numerical method as an alternative to traditional alignment procedures for functional annotation. Thus, an alignment-free model was developed by linear statistical techniques to successfully detect the RNase III signature among a highly diverse dataset including the Pac 1 member. The model showed a higher accuracy and a similar sensitivity in detecting the RNase III signature than that achieved using alignment procedures [71]. This report provided some clues about the potential of graphical–numerical approaches as alternative tools to detect remote homologous due to their alignment-independence essence.

Considering such promising studies, we aimed to overcome alignment handicaps to face functional detections in highly diverse gene/protein families through the creation of TI2BioP software [5, 44]. The utility of TI2BioP was proved in classifying four gene/protein classes having great sequence divergence among their members (see Section 2.2). The alignment-free model's performance was always compared to alignment algorithms since we are pursuing an alternative tool to such methods. Alignment algorithms are the most popular techniques in bioinformatics; they basically score similarity measures at a predefined biological significance between a query gene/protein against others already registered or against a family profile to predict the structural and functional class [73, 74]. Although alignment-dependent algorithms have been improved through years of use, they do not consider structural information beyond the primary sequence, e.g. long-distance interactions and also ignore the important contribution of a negative set (nonmembers of the family), especially for the alignment algorithms building a profile-based model. Another weakness of this method arises when a query sequence is similar to genes/proteins lacking functional annotations [75]. In addition, phylogenetic inferences relying on MSA methods are not reliable when gene/protein sequences show functional similarities but have greatly diverged [52]. Consequently, such handicaps motivate the arising of alignment-free methods that exploit extra-information hidden in the linearity of the sequence, e.g. amino acid pseudo amino acid composition.

To validate our graphical–numerical methodology implemented in the TI2BioP software, we applied it to detect proteinaceous bacteriocins. Bacteriocins are small proteins of bacterial origin that are lethal to bacteria other than the producing strain. They have found applications in the pharmaceutical and food industry as potential antimicrobial agents and food preservatives, respectively [76]. 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 has contributed to its low pair-wise sequence similarity (23%–50%). These family features represent a challenge for alignment procedures in the identification of protein bacteriocins [77], demanding the implementation of complex strategies [78, 79]. However, we built an effective alignment-free model based on the 2D-Cartesian protein representation and its derived TIs to detect bacteriocin proteins among the diversity of a dataset made up of nonredundant CATH domains and bacteriocin sequences. The model retrieved 66.7% of the bacteriocin-like proteins from an external test set while the InterPro resource could just detect 60.2% (Table 4). This is the first report where an alignment-free model based on a graphical approach entirely outperforms a popular alignment-based resource for functional sequence annotation [5].

The other bioinformatics utility of our graphical–numerical method consisted of the detection of a remote bacteriocin homologous in the Cry 1Ab C-terminal domain from *Bacillus thuringiensis´* endotoxin, which had not been detected by classical alignment methods. Although the functional relationship between bacteriocins and Cry 1Ab C-terminal domain classes have been assessed by experimental procedures in previous reports [80], their sequences are completely different and consequently placed into two different protein classes by alignment procedures. TI2BioP could successfully detect the bactericide function of Cry 1Ab C-terminal, just corroborated by experimental assays, either by scoring the query sequence for the alignment-free model or by the graphical superposition of the 2D-Cartesian maps for Cry 1Ab C-terminal domain to other representative bacteriocins [5]. This graphical analysis has been useful to visualize similarities/dissimilarities between different classes of natural biopolymers [40, 71].

After having success in detecting distant homologous among the protein bacteriocin family using TIs derived from 2D-Cartesian maps, the RNase III enzymatic class was selected as the second target to evaluate TI2BioP performance. The RNase III protein class contains members having great variability regarding the primary structure and domain organization. The similarities among different RNase IIIs varies from 20% to 84%, placing many of them in the twilight zone [81]. In addition, this diversity is also influenced by differences in the domain architectures of RNase III, which have led to a subdivision of the enzymatic class into four subclasses represented by four archetypes (bacterial RNase III, fungal RNase III, Dicer and Drosha).

Spectral moments derived from the 20 amino acids clustering according to their physicochemical properties into a 2D-Cartesian space (2D Cartesian maps) were used to develop three different nonlinear approaches to detect RNase III protein signatures among the diversity of

206 RNases III and a structurally nonredundant subset of the PDB made up of enzymes and no enzymes. Two alignment-free models based on decision tree models (DTMs) and artificial neural networks (ANNs) were built from TIs provided by TI2BioP to identify RNase III members. Additionally, a nonclassical profile HMM, inspired on the graphical clustering of the amino acids was developed, to make a fair comparison among alignment-free models and alignment algorithms [82].

While 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 [83-86], DTMs have been poorly explored in bioinformatics despite their widespread use in other fields [87]. We reported, for the first time, a simple and interpretable DTM to identify 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%) (Table 4) [69, 82].

The nonclassical 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 with respect to the performance of ANN and DTM (Table 4). Amino acid clustering according to its hydrophobic/charge properties was either effective at the primary level to increase the sensitivity of the profile HMM or at the 2D level to develop highly predictive DTM. Although the nonclassical profile HMM showed a slightly better performance than the alignment-free models, its generation demands programming skills while DTM search resulted in the easiest way to detect the RNase III signature among the diversity of the dataset [69]. The usability of DTM was also shown by predicting a new bacterial RNase III class member that was isolated and subsequently enzymatically tested and registered by our group (Table 1). The efficiency of DTM as a sequence search procedure to screen a proteome in conjunction with the TIs implemented in the TI2BioP software will be seen below [20].

Up to now, the TIs generated by TI2BioP have successfully been applied as alignment-free predictors in protein families but their classification performance should also be evaluated in highly diverse gene families, as well as their ability for reconstructing phylogenies. In this sense, the original 2D Cartesian representation reported by Nandy for describing DNA sequences [63] and the secondary structure inferred by DNA/RNA folding algorithms (Mfold) [67] were used to derive two types of TIs, for the ITS2 gene class.

The ITS2 eukaryotic gene class shows a high sequence divergence among its members, which has traditionally been exploited in low-level taxonomical analyses, especially for the unequivocal classification of fungal species. However, such sequence variability has complicated the ITS2 annotation and its use in phylogenetic analyses at higher taxonomic ranks. In this sense, the ITS2 secondary structure which has been conserved among all eukaryotes has been 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 [61, 88-90]. Although alignment-based methods have been exploited to the top of its complexity to tackle the ITS2 annotation and phylogenetic inference [88, 91], no alignment-free approach has so far been able to successfully address these issues.

The use of TIs containing information about the sequence and structure of ITS2 can be an alignment-free solution to improve the ITS2 prediction and for phylogenetic reconstruction at high taxonomic levels in eukaryotes. Alignment-independent approaches are represented by two ANN-based models for ITS2 classification among a large and diverse dataset, one built with 2D-Cartesian TIs [63, 92] and the other resulting from the Mfold 2D structure TIs [67]. Although ANN models built with both TI types (Cartesian and Mfold) 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 why the Mfold TIs were applied as an alignment-free approach to infer phylogenetic relationship to complement the taxonomy of a fungal isolate.

The performance of both ANN models were compared to several profiles HMM generated from MSA performed with CLUSTALW [93], DIALIGN-TX [94] and MAFFT [95] to classify the test set and to identify a new fungal member of the ITS2 class. Alignment-free models outperformed profiles HMM in classifying the test set and in identifying the new fungal member of the ITS2 class, even when HMMs were built by MSA algorithms improved for sets of low overall sequence.

The 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 are potential producers of bioactive compounds but they have been hard to place taxonomically [96]. In fact, the NCBI dedicated "taxonomy" database does not have clear information about its genus and that there is no specification about its subphylum and class [97]. On the other hand, the lack of other registered ITS2 sequences from different species of the genus *Petrakia* precluded performing a phylogenetic analysis at the species level (low-level analysis). Then, assuming that our fungal isolate belongs to the *Pezizomycotina* subphylum following a recent classification found in the "The dictionary of the Fungi" [98], a higher-level phylogenetic analysis to elucidate the class of *Petrakia* sp. was carried using two different types of distance trees: (1) a traditional NJ-tree based on multiple alignments of ITS2 sequences and (2) another tree irrespective of sequence similarity built from Mfold TIs. The alignment-free distances calculated from Mfold TIs provide similar phylogenetic relationships among the different classes of the Ascomycota phylum regarding 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 in the Dothideomycetes class (Figure 4). We concluded that our alignment-free approach was effective for constructing hierarchical distance trees containing relevant biological information with an evolutionary significance [21].

So far, the 2D Cartesian graphs have been used to derive a TI series with the aim of being applied in bioinformatics. However, there are other 2D graphical approaches reported for DNA and proteins that have been mostly unexplored in this field; such is the case of the four-color maps introduced by Randić [64, 65]. Consequently, the four-color maps for DNA and protein sequences were implemented in the latest version of TI2BioP in order estimate new alignment-free predictors that 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.

The NRPS family can harbor remote homologous due to the high sequence divergence among its A-domains, ranging mostly from 10% to 40% of sequence identity. Consequently, many of them are placed in the twilight zone (20%–35% sequence identity) reported for the alignment methods [99]. In fact, A-domain members cannot be retrieved easily by BLASTp using a single template [62]. 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 [20].

The four-color map TIs were used to develop several alignment-free models using linear and nonlinear mathematical functions. Nonlinear models outperformed linear models in classifying A-domains confirming previous outcomes. DTM was the model of choice due to its excellent performance and its simple way to detect A-domains in a highly diverse dataset [20]. The DTM built up with four-color map TIs overdid other alignment-free concepts like ACC and PseACC, providing the highest sensitivity (Table 2) and no false-positives in A-domain identification (Table 3). In addition, it showed a similar performance to sensitive alignment algorithms like profile HMM and multitemplate BLASTp (Table 3).

As a result of comparing methods to detect A-domains, we can conclude that classification results among homology-based methods agreed with the fact that multitemplate BLASTp and profile HMM are more sensitive than simple BLASTp. Both multitemplate BLASTp and profile HMM easily retrieved all A-domain members at expectation values (E value ≤ 10) without reporting any false-positive (Table 3). However, the BLASTp search using a single template provided false-positives (significant matches) among the negative set (CATH domains) at both high (E value = 10) and relatively stringent cutoffs (E values < 0.05), which is considered statistically significant and useful for filtering easily identifiable homologs pairs [47, 100].

Because of the single-template BLASTp sensitivity did not show stability in identifying the A-domain signal at different classification stringency (E value); it was considered less reliable to perform sequence searches on unknown test datasets such as an entire proteome. Therefore, the easy and reliable identification of A-domains in the proteome of the cyanobacteria *M. aeruginosa* NIES-843 [20] was carried out by the combination of multitemplate BLASTp, profile HMM and four-color maps. Profiles HMM and four-color maps found additional hits as A-domains among the hypothetical proteins, giving clues for the presence of A-domain's remote homologous in the proteome of *M. aeruginosa* (Figure 3). Hypothetical proteins have not been definitively annotated and can be reannotated by applying more sensitive strategies. The assembling of sequence-search methods encoding different features from protein sequences can provide a better description of the proteome and therefore, remote protein homologous can be detected with more confidence [20]. Thus, we are introducing a new sensitive approach to search for remote homologous by integrating graphical–numerical methods with alignment procedures.

In summary, the presented graphical–numerical method implemented in the TI2BioP software does not suffer from many of the alignment algorithm limitations. Particularly, the artificial

2D graphs and the TIs encode higher-order useful information contained beyond the primary structure of the natural biopolymers allowing the building-up of effective alignment-free models. By contrast, our graphical–numerical approach has some handicaps stemming from the artificial nature of the 2D graphs which do not represent the real secondary structure of the biopolymers. Many of these 2D graphs bear some redundancy that leads to the loss of sequence information. On the other hand, spectral moment (TIs) estimation by powering matrixes, from thousands of graphs or maps, still demands a high computational cost.

5. Conclusions

We provided several evidences of the potential use of graphical–numerical approaches to characterize DNA/RNA and proteins that can be extended to other biopolymers. This new software called TI2BioP is not in competition with currently available bioinformatics tools, but instead works in cooperation with existing methodologies, as well as with experimental procedures required to overcome hard comparative studies of the natural biopolymers.

Acknowledgements

The authors acknowledge the Portuguese Fundação para a Ciência e a Tecnologia (FCT) for financial support to GACH (SFRH/BPD/92978/2013). AA was partially supported by the Strategic Funding UID/Multi/04423/2013 through national funds provided by FCT and European Regional Development Fund (ERDF) in the framework of the programme PT2020, and the FCT projects PTDC/AAC-AMB/121301/2010 (FCOMP-01-0124-FEDER-019490) and PTDC/AAG-GLO/6887/2014. The funders had no role in the study's design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author details

Guillermin Agüero-Chapin[1,2], Reinaldo Molina-Ruiz[2], Gisselle Pérez-Machado[2],
Vitor Vasconcelos[1,3], Zenaida Rodríguez-Negrin[2] and Agostinho Antunes[1,3*]

*Address all correspondence to: aantunes@ciimar.up.pt

1 CIMAR/CIIMAR, Centro Interdisciplinar de Investigação Marinha e Ambiental,
Universidade do Porto, Porto, Portugal

2 Centro de Bioactivos Químicos, Universidad Central "Marta Abreu" de Las Villas (UCLV),
Santa Clara, Cuba

3 Faculdade de Ciências, Departamento de Biologia, Universidade do Porto, Porto, Portugal

References

[1] Biggs N, Lloyd E, Wilson R, editors. Graph Theory: Oxford University Press; 1986. 1736–936.

[2] Randic M, Zupan J, Balaban AT, Vikic-Topic D, Plavsic D. Graphical representation of proteins. Chem Rev. 2011; 111: 790-862. DOI: 10.1021/cr800198j

[3] Gonzalez-Diaz H, Vilar S, Santana L, Uriarte E. Medicinal chemistry and bioinformatics--current trends in drugs discovery with networks topological indices. Curr Top Med Chem. 2007; 7: 1015-29.

[4] Estrada E, Uriarte E. Recent advances on the role of topological indices in drug discovery research. Curr Med Chem. 2001; 8: 1573-88.

[5] Aguero-Chapin G, Perez-Machado G, Molina-Ruiz R, Perez-Castillo Y, Morales-Helguera A, Vasconcelos V, et al. TI2BioP: Topological Indices to BioPolymers. Its practical use to unravel cryptic bacteriocin-like domains. Amino Acids. 2011; 40: 431-42.

[6] Perez-Bello A, Munteanu CR, Ubeira FM, De Magalhaes AL, Uriarte E, Gonzalez-Diaz H. Alignment-free prediction of mycobacterial DNA promoters based on pseudo-folding lattice network or star-graph topological indices. J Theor Biol. 2009; 256: 458-66.

[7] Ortega-Broche SE, Marrero-Ponce Y, Diaz YE, Torrens F, Perez-Gimenez F. TOMO-COMD-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: 3118-46. DOI: 10.1111/j.1742-4658.2010.07711.x

[8] Gonzalez-Diaz H, Gonzalez-Diaz Y, Santana L, Ubeira FM, Uriarte E. Proteomics, networks and connectivity indices. Proteomics. 2008; 8: 750-78. DOI: 10.1002/pmic.200700638 [doi]

[9] Estrada E, Gutman I. A Topological Index Based on Distances of Edges of Molecular Graphs. J Chem Inf Comput Sci. 1996; 36: 850-3.

[10] Riera-Fernandez P, Martin-Romalde R, Prado-Prado FJ, Escobar M, Munteanu CR, Concu R, 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: 927-60.

[11] Gonzalez-Diaz H, Perez-Montoto LG, Duardo-Sanchez A, Paniagua E, Vazquez-Prieto S, Vilas R, et al. Generalized lattice graphs for 2D-visualization of biological information. J Theor Biol. 2009; 261: 136-47.

[12] Concu R, Podda G, Uriarte E, Gonzalez-Diaz H. 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: 1510-20.

[13] Marrero-Ponce Y, Medina-Marrero R, Castillo-Garit JA, Romero-Zaldivar V, Torrens F, Castro EA. 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: 3003-15.

[14] Randic´ M. Graphical representation of DNA as a 2-D map. Chem Phys Lett. 2004; 468–71.

[15] Randic M, Zupan J, Vikic-Topic D. On representation of proteins by star-like graphs. J Mol Graph Model. 2007; 26: 290-305.

[16] Randic M, Zupan J. Highly compact 2D graphical representation of DNA sequences. SAR QSAR Environ Res. 2004; 15: 191-205.

[17] Nandy A. Recent investigations into global characteristics of long DNA sequences. Indian J Biochem Biophys. 1994; 31: 149-55.

[18] Aguero-Chapin G, Varona-Santos J, de la Riva GA, Antunes A, Gonzalez-Vlla T, Uriarte E, 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: 2122-8.

[19] Cruz-Monteagudo M, Gonzalez-Diaz H, Borges F, Dominguez ER, Cordeiro MN. 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: 619-32. DOI: 10.1021/tx700296t [doi]

[20] Aguero-Chapin G, Molina-Ruiz R, Maldonado E, De la Riva GA, Sanchez-Rodriguez A, Vasconcelos V, et al. Exploring the Adenylation Domain Repertoire of Nonribosomal Peptide Synthetases Using an Ensemble of Sequence-Search Methods. PLoS One. 2013; 8: DOI: 10.1371/journal.pone.0065926

[21] Aguero-Chapin G, Sanchez-Rodriguez A, Hidalgo-Yanes PI, Perez-Castillo Y, Molina-Ruiz R, Marchal K, et al. An alignment-free approach for eukaryotic ITS2 annotation and phylogenetic inference. PLoS One. 2011; 6: DOI: 10.1371/journal.pone.0026638

[22] Katritzky AR, Gordeeva EV. Traditional topological indices vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research. J Chem Inf Comput Sci. 1993; 33: 835-57.

[23] Randic M, Zupan J. On interpretation of well-known topological indices. J Chem Inf Comput Sci. 2001; 41: 550-60.

[24] Wiener H. Structural Determination of Paraffin Boiling Points. J Am Chem Soc. 1947; 69: 17-20.

[25] Randic M. Graph theoretical approach to structure-activity studies: search for optimal antitumor compounds. Prog Clin Biol Res. 1985; 172A: 309-18.

[26] Balaban AT, Beteringhe A, Constantinescu T, Filip PA, Ivanciuc O. Four new topological indices based on the molecular path code. J Chem Inf Model. 2007; 47: 716-31. DOI: 10.1021/ci6005068 [doi]

[27] Moreau G, Broto P. The Autocorrelation of a topological structure. A new molecular descriptor. Nouv J Chim. 1980; 4: 359-60.

[28] 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: 844-9.

[29] 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: 320-8.

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

[31] Markovic S, Markovic Z, McCrindle RI. Spectral moments of phenylenes. J Chem Inf Comput Sci. 2001; 41: 112-9.

[32] González MP, Teran C, Teijeira M. A topological function based on spectral moments for predicting affinity toward A_3 adenosine receptors. Bioorg Med Chem Lett. 2006; 16: 1291-6.

[33] Morales AH, González MP, Briones JR. TOPS-MODE approach to predict mutagenicity in dental monomers. Polymer. 2004; 45: 2045-50.

[34] Estrada E. Characterization of the folding degree of proteins. Bioinformatics. 2002; 18: 697-704.

[35] Estrada E, Hatano N. A Tight-Binding "Dihedral Orbitals" Approach to Electronic Communicability in Protein Chains. Chemical Physics Letters. 2007; 449: 216-20.

[36] Gonzalez-Diaz H, Uriarte E. Biopolymer stochastic moments. I. Modeling human rhinovirus cellular recognition with protein surface electrostatic moments. Biopolymers. 2005; 77: 296-303.

[37] Gonzalez-Diaz H, Uriarte E, Ramos de Armas R. Predicting stability of Arc repressor mutants with protein stochastic moments. Bioorg Med Chem. 2005; 13: 323-31.

[38] Gonzalez-Diaz H, Saiz-Urra L, Molina R, Gonzalez-Diaz Y, Sanchez-Gonzalez A. Computational chemistry approach to protein kinase recognition using 3D stochastic

van der Waals spectral moments. J Comput Chem. 2007; 28: 1042-8. DOI: 10.1002/jcc.20649 [doi]

[39] González-Díaz H, Molina-Ruiz R, Hernandez I. MARCH-INSIDE v3.0 (MARkov CHains INvariants for SImulation & DEsign). 3.0 ed2007. p. Windows supported version under request to the main author contact email: gonzalezdiazh@yahoo.es.

[40] Aguero-Chapin G, Gonzalez-Diaz H, Molina R, Varona-Santos J, Uriarte E, Gonzalez-Diaz Y. 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: 723-30. DOI: 10.1016/j.febslet.2005.12.072 [doi]

[41] Aguero-Chapin G, Antunes A, Ubeira FM, Chou KC, Gonzalez-Diaz H. Comparative study of topological indices of macro/supramolecular RNA complex networks. J Chem Inf Model. 2008; 48: 2265-77. DOI: 10.1021/ci8001809 [doi]

[42] Gonzalez-Diaz H, Aguero-Chapin G, Varona-Santos J, Molina R, de la Riva G, Uriarte E. 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: 2932-7.

[43] Gonzalez-Diaz H, Aguero-Chapin G, Varona J, Molina R, Delogu G, Santana L, et al. 2D-RNA-coupling numbers: a new computational chemistry approach to link secondary structure topology with biological function. J Comput Chem. 2007; 28: 1049-56. DOI: 10.1002/jcc.20576 [doi]

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

[45] Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970; 48: 443-53.

[46] Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981; 147: 195-7. DOI: 0022-2836(81)90087-5 [pii]

[47] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool. J Mol Biol. 1990; 215: 403-10.

[48] Krogh AB, M.; Mian, I. S.; Sjeander, K.; Haussler, D. Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol. 1994; 235: 1501-31.

[49] Eddy SR. A new generation of homology search tools based on probabilistic inference. Genome Inform. 2009; 23: 205-11.

[50] Dobson PD, Doig AJ. Distinguishing Enzyme Structures from Non-enzymes Without Alignments. J Mol Biol. 2003; 330: 771–83.

[51] Strope PK, Moriyama EN. Simple alignment-free methods for protein classification: a case study from G-protein-coupled receptors. Genomics. 2007; 89: 602-12. DOI: 10.1016/j.ygeno.2007.01.008

[52] Schwarz RF, Fletcher W, Förster F, Merget B, Wolf M, Schultz J, et al. Evolutionary Distances in the Twilight Zone—A Rational Kernel Approach. PLoS ONE. 2010; 5:

[53] Rost B. Enzyme function less conserved than anticipated. J Mol Biol. 2002; 318: 595-608.

[54] Pearson WR, Sierk ML. The limits of protein sequence comparison? Current Opinion in Strctural Biology. 2005; 15: 254-60.

[55] Kumar M, Thakur V, Raghava GP. COPid: composition based protein identification. In Silico Biol. 2008; 8: 121-8.

[56] Deshmukh S, Khaitan S, Das D, Gupta M, Wangikar PP. An alignment-free method for classification of protein sequences. Protein Pept Lett. 2007; 14: 647-57.

[57] Schwarz RF, Fletcher W, Forster F, Merget B, Wolf M, Schultz J, et al. Evolutionary distances in the twilight zone--a rational kernel approach. PLoS One. 2010; 5: e15788. DOI: 10.1371/journal.pone.0015788

[58] Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins. 2001; 43: 246-55.

[59] Hohl M, Rigoutsos I, Ragan MA. Pattern-based phylogenetic distance estimation and tree reconstruction. Evol Bioinform Online. 2006; 2: 359-75.

[60] de Jong A, van Hijum SA, Bijlsma JJ, Kok J, Kuipers OP. BAGEL: a web-based bacteriocin genome mining tool. Nucleic Acids Res. 2006; 34: W273-9.

[61] Selig C, Wolf M, Muller T, Dandekar T, Schultz J. The ITS2 Database II: homology modelling RNA structure for molecular systematics. Nucleic Acids Res. 2008; 36: D377-80.

[62] Ansari MZ, Yadav G, Gokhale RS, Mohanty D. NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. Nucleic Acids Res. 2004; 32: W405-13. DOI: 10.1093/nar/gkh359

[63] Nandy A. Two-dimensional graphical representation of DNA sequences and intron-exon discrimination in intron-rich sequences. Comput Appl Biosci. 1996; 12: 55-62.

[64] Randic M, Mehulic K, Vukicevic D, Pisanski T, Vikic-Topic D, Plavsic D. Graphical representation of proteins as four-color maps and their numerical characterization. J Mol Graph Model. 2009; 27: 637-41. DOI: 10.1016/j.jmgm.2008.10.004

[65] Randic M, Lers N, Plavšić D, Basak S, Balaban A. Four-color map representation of DNA or RNA sequences and their numerical characterization. Chemical Physics Letters 2005; 407: 205-8.

[66] Aguero-Chapin G, Molina-Ruiz R, Maldonado E, de la Riva G, Sanchez-Rodriguez A, Vasconcelos V, et al. Exploring the adenylation domain repertoire of nonribosomal peptide synthetases using an ensemble of sequence-search methods. PLoS One. 2013; 8: e65926. DOI: 10.1371/journal.pone.0065926

[67] Mathews DH. RNA secondary structure analysis using RNAstructure. Curr Protoc Bioinformatics. 2006; Chapter 12: Unit 12 6. DOI: 10.1002/0471250953.bi1206s13 [doi]

[68] Hammami R, Zouhir A, Hamida JB, Fliss I. BACTIBASE: a new web-accessible database for bacteriocin characterization. BMC Microbiology 2007; 7: 89 DOI: 10.1186/1471-2180-7-89

[69] Aguero-Chapin G, de la Riva GA, Molina-Ruiz R, Sanchez-Rodriguez A, Perez-Machado G, Vasconcelos V, et al. Non-linear models based on simple topological indices to identify RNase III protein members. J Theor Biol. 2011; 273: 167-78.

[70] Shen HB, Chou KC. PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. Anal Biochem. 2008; 373: 386-8.

[71] Aguero-Chapin G, Gonzalez-Diaz H, de la Riva G, Rodriguez E, Sanchez-Rodriguez A, Podda 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: 434-48. DOI: 10.1021/ci7003225 [doi]

[72] Lamontagne B, Elela SA. Evaluation of the RNA determinants for bacterial and yeast RNase III binding and cleavage. J Biol Chem. 2004; 279: 2231-41. DOI: 10.1074/jbc.M309324200

[73] Altschul SF, 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: 3389-402.

[74] Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011; 39: W29-37. DOI: 10.1093/nar/gkr367

[75] Davies MN, Secker A, Freitas AA, Timmis J, Clark E, Flower DR. Alignment-Independent Techniques for Protein Classification. Current Proteomics. 2008; 5: 217-23.

[76] Gillor O, Etzion A, Riley MA. The dual role of bacteriocins as anti- and probiotics. Appl Microbiol Biotechnol. 2008; 81: 591-606.

[77] Cotter P, Hill C, Ross R. What's in a name? Class distinction for bacteriocins. Nature Reviews Microbiology. 2006; 4:

[78] Dirix G, Monsieurs P, Dombrecht B, Daniels R, Marchal K, Vanderleyden J, 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: 1425-40.

[79] Stein T. Bacillus subtilis antibiotics: structures, syntheses and specific functions. Mol Microbiol. 2005; 56: 845-57.

[80] Vazquez-Padron RI, de la Riva G, Aguero G, Silva Y, Pham SM, Soberon M, et al. Cryptic endotoxic nature of Bacillus thuringiensis Cry1Ab insecticidal crystal protein. FEBS Lett. 2004; 570: 30-6. DOI: 10.1016/j.febslet.2004.06.021

[81] Lamontagne B, Elela S.A. Evaluation of the RNA determinants for bacterial and yeast RNase III binding and cleavage. J Biol Chem. 2004; 279: 2231-41.

[82] Chapin GA. A Graphical and Numerical Approach for Functional Annotation and Phylogenetic Inference: University of Porto; 2013.

[83] Punta M, Rost B. Neural networks predict protein structure and function. Methods Mol Biol. 2008; 458: 203-30.

[84] Nair R, Rost B. Protein subcellular localization prediction using artificial intelligence technology. Methods Mol Biol. 2008; 484: 435-63. DOI: 10.1007/978-1-59745-398-1_27 [doi]

[85] Cai YD, Ricardo PW, Jen CH, Chou KC. Application of SVM to predict membrane protein types. J Theor Biol. 2004; 226: 373-6.

[86] Fernandez M, Caballero J, Fernandez L, Abreu JI, Garriga M. 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: 748-59. DOI: 10.1016/j.jmgm.2007.04.011

[87] Ripley B. Pattern Recognition and Neural Networks. Cambridge, UK: Cambridge University Press; 1996

[88] Koetschan C, Forster F, Keller A, Schleicher T, Ruderisch B, Schwarz R, et al. The ITS2 Database III--sequences and structures for phylogeny. Nucleic Acids Res. 2009;

[89] Schultz J, Maisel S, Gerlach D, Müller T, and Wolf M. A common core of secondary structure of the internal transcribed spacer 2 (ITS2) throughout the Eukaryota. RNA 2005; 11: 361-4.

[90] Schultz J, Muller T, Achtziger M, Seibel PN, Dandekar T, Wolf M. The internal transcribed spacer 2 database--a web server for (not only) low level phylogenetic analyses. Nucleic Acids Res. 2006; 34: W704-7.

[91] Schultz J, Müller T, Achtziger M, Seibel P, Dandekar T, Wolf M. The internal transcribed spacer 2 database--a web server for (not only) low level phylogenetic analyses. Nucleic Acids Research. 2006; 34: DOI: 10.1093/nar/gkl129

[92] Nandy A. Empirical relationship between intra-purine and intra-pyrimidine differences in conserved gene sequences. PLoS One. 2009; 4: e6829. DOI: 10.1371/journal.pone.0006829

[93] Thompson JD, Higgins DG, Gibson TJ. 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: 4673-80.

[94] Subramanian AR, Kaufmann M, Morgenstern B. DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. Algorithms Mol Biol. 2008; 3: 6. DOI: 10.1186/1748-7188-3-6

[95] Katoh K, Kuma K, Miyata T, Toh H. Improvement in the accuracy of multiple sequence alignment program MAFFT. Genome Inform. 2005; 16: 22-33.

[96] Qi FH, Jing TZ, Wang ZX, Zhan YG. Fungal endophytes from Acer ginnala Maxim: isolation, identification and their yield of gallic acid. Lett Appl Microbiol. 2009; 49: 98-104.

[97] Bisby F, Roskov Y, Ruggiero M, Orrell T, Paglinawan L, Brewer P, et al. Species 2000 & ITIS Catalogue of Life: 2007 Annual Checklist Taxonomic Classification. CD-ROM; Species 2000: Reading, U.K. 2007;

[98] Kirk PM, Cannon PF, Stalpers JA. The dictionary of the Fungi. 10th ed. UK: CABI; 2008. 784.

[99] Rost B. Twilight zone of protein sequence alignments. Protein Eng. 1999; 12: 85-94.

[100] Boekhorst J, Snel B. Identification of homologs in insignificant blast hits by exploiting extrinsic gene properties. BMC Bioinformatics. 2007; 8.