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

Guillermín Agüero-Chapin, Gustavo A de La Riva, Reinaldo Molina-Ruiz, Aminael Sánchez-Rodríguez, Gisselle Pérez-Machado, Vítor Vasconcelos, Agostinho Antunes

www.elsevier.com/locate/yjtbi

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

Guillermín Agüero-Chapin, Gustavo A de la Riva, Reinaldo Molina-Ruiz, Aminael Sánchez-Rodríguez, Gisselle Pérez-Machado, Vítor Vasconcelos, Agostinho Antunes

PII: S0022-5193(10)00674-0
DOI: doi:10.1016/j.jtbi.2010.12.019
Reference: YJTBI 6288

To appear in: *Journal of Theoretical Biology*

Received date: 2 September 2010
Revised date: 15 November 2010
Accepted date: 13 December 2010

Cite this article as: Guillermín Agüero-Chapin, Gustavo A de la Riva, Reinaldo Molina-Ruiz, Aminael Sánchez-Rodríguez, Gisselle Pérez-Machado, Vítor Vasconcelos and Agostinho Antunes, Non-linear models based on simple topological indices to identify RNase III protein members, *Journal of Theoretical Biology*, doi:10.1016/j.jtbi.2010.12.019

Non-Linear models based on Simple Topological Indices to identify RNase III protein members

Guillermín 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] and 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, México*

[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*

***Corresponding author:** Antunes, A.: CIMAR, Rua dos Bragas, 177, 4050-123 Porto, Portugal, Tel: +351 22 3401 813, E-mail: aantunes@ciimar.up.pt or aantunes777@gmail.com

Abstract

Alignment-free classifiers are especially useful in the functional classification of protein classes with variable homology and different domain structures. Thus, the **TI2BioP** (**T**opological **I**ndices **to BioP**olymers) methodology [1] 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.

Key words: Alignment-free models/ Spectral Moments/ Clustering/ Decision Tree Models/ Artificial Neural Networks

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 [2]. While such available methodologies for sequence classification have a friendly interface for the normal users [3], its algorithms demand a high computational cost and in many cases require the implementation of stochastic process for building a predictive model [4]. Such procedures turn out to be less effective when the members of a certain gene [5] and protein [6] 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 [2; 7]. To date most of the alignment-free classifiers estimate 1D sequence parameters based on the amino acid composition to evaluate sequence-function relationships [8], predict protein-protein interactions [9] and protein attributes [10].

The introduction of 2D or higher dimension representations of sequences [11; 12] 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 **TI2BioP** (**T**opological **I**ndices **to BioP**olymers) [1] inspired in both the **TOPS-MODE** (**T**opological **S**ub-structural **Mo**lecular **De**sign) [13] and the **MARCH-INSIDE** (**M**arkov **C**hain **I**nvariants for **N**etwork **S**election & **D**esign) [14] 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 [15]. 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 [16]. 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 shows 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 [16].

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 to predict 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 [17; 18]. 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 [19], 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 [1] 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) [20]. Its DTM detection as a RNase III class member was remarkable 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) [1]. The program could be run on Windows and Linux operating system. 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 [21] or the artificial, but informative, folding of linear sequences [22]. 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¨ [13] implemented in the "MODESLAB" software [23] and the MARCH-INSIDE program [14]. The calculation of the spectral moments as TIs is performed according the TOPS-MODE approach [13] and the pseudo-secondary structures for the protein sequences were taken from the experiences achieved by using the MARCH-INSIDE methodology [22; 24]. We used the 2D lattice of Hydrophobicity (H) and Polarity (P) introduced

by our group to encode information about polygalacturonases enzymes [22] 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 [25]. 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 [26; 27; 28] that was extended later to classify protein families [22; 24] and structural RNA [29] using stochastic indices [30]. The **figure 1** shows how the new RNase III protein sequence from *Escherichia coli* BL21 substrain GG1108 is pseudo-folded into a HP-lattice or 2D-HP map that compact its linear sequence: its two major domains are highlighted in red (RNase III domain) and in blue (double-stranded RNA binding motif), respectively. 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 point out in black and red as a square and simple dot, respectively.

Figure 1 comes about here

All sequences are pseudo-folded into a HP-Cartesian lattice by TI2BioP. The original spectral moments (μ_k) introduced previously by Estrada [31; 32], that have been validated for many authors to encode the structure of small molecules in Quantitative Structure Activity Relationship (QSAR) studies [33; 34; 35], were applied to describe such protein 2D-HP maps ($^{HP}\mu_k$) to contain new structural information. The original adjacent matrix is modified according 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 the 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 bond adjacency matrix. **B** is 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 [36].

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

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

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 the **figure 2**. The calculation of the spectral moments up to the order k = 3 is also defined downstream of the **figure 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, D= 1.997)

Figure 2 comes about here

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

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

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

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

The calculation of $^{HP}\mu_k$ values for protein sequences of both groups were carried out with our in-house software **TI2BioP** *version 1.0* ®, including sequence representation [37]. We proceeded to upload a row data table containing the sixteen $^{HP}\mu_k$ values for each sequence (k = 1, 2, 3,…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 [38], 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 GenBank database gathering RNAses III registered up to May of 2009**.** Each RNase III sequence was labelled 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 [39], see File I of SM. The selection of such subset was determined using a K-Means cluster analysis **(**k-MCA) [40]. 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 [40]. 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 the Euclidean distance.

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 [40; 41]. The procedure described above is represented graphically in **figure 3** for both groups.

Figure 3 comes about here

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 [38]. 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 C&RT (Classification and Regression Trees)-style univariate split selection method was used since it 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 ¨Prune on misclassification error¨ was set as 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 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 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 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) [42]. In 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) ; 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 30ºC 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 100µg/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 LB medium and grown at 30ºC during 12 hours 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 standard procedure. Cellular pellet was resuspended in 300µL 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 using for PCR amplification of *Escherichia coli* RNase type III were designed based on the previously reported *E. coli* RNAse type III coding sequence [43; 44]: forward primer (RNaseIII5') 5'-c**ccATGG**ACCCCATCGTAATTAATCGGC–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, 1mM of each dNTP, 1.5mM $MgCl_2$, 2µM of each PAC5' and PAC3' primers, in a total volume of 50µL, 1x 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, 1000 ω). Transformation was plated onto of LB medium supplemented with 40 µL of 20 µg/mL X-gal solution and 4µL of isopropylthio-β-D-galactoside from 200 µg/mL IPTG solution per plate and grown overnight at 37ºC. 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 [45].

3.6 Synthesis and preparation of dsRNA substrate for enzymatic assay.

The synthesis and preparation of dsRNA substrate for enzymatic assay of recombinant *E. coli* RNAse III was conceived according to create optimized dsRNA structure for measurement of enzymatic activity [16]. One of the T7 substrates, named R1.1 RNA (109 nt), was used for 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 [46]. The DNA fragment encoding for 109 nt R1.1 RNA were 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 [α^{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 [α^{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 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 purified the recombinant protein as described [45] in the form of 6x(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 [47]. The *E. coli* assay was carried using the following conditions: 30 mM Tris-HCl (pH7.6), 1 mM DTT, 10 mM of $MgCl_2$, 10 nM of dsRNA substrate and 100 mM polydifferent 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 0.1V of 50 mM MgCl2, incubated at 30ºC for 10 minutes and stopped by addition of 500 µl of 5% ice-cooled TCA followed by 15 minutes on ice. The aliquots were centrifuged at 16 000g during 5 minutes in Spin-X filter unit (Costar). The soluble fractions (filtrate) were quantified by 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 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, respectively. 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 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 **TI2BioP** methodology [1]. We explored the standard deviation between and within clusters, the respective Fisher ratio and their p-level of significance [40]. 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 **(figure 4)**. They

described three and four statistically homogeneous clusters for the RNase III class and the control group, respectively.

Table 1 comes about here

Figure 4 comes about here

Such division of the RNase III protein sequences into three clusters according our TIs is a close approximation to the structure-based characterization reported by Lamontagne and Elela for this family [16], which divided it into 4 sub-classes. However, our three groups coincided perfectly with another subdivision based on the biological activity [48].

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 super-families, to date no DTM have been developed to differentiate 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 [49; 50; 51]. 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 262126 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 black) and three terminal nodes (outlined in gray) summing up a total of five nodes. In the graph, the numbers of the nodes are labelled 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 control group, as is indicated by the control group label

(-1) placed in the top-right corner of the root node. Sequences from 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 less than or equal to 422.6 are sent to node number 2 and tentatively classified as RNase III enzymes, and the protein sequences with $^{HP}\mu_1$ values greater than 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 lesser than or equal to 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 greater than 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 to evaluate 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 **(figure 5)**. All the information in the tree graph is also available in the tree structure shown in **Table 2**.

Figure 5 and Table 2 come about here

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 [52]. In this case, $^{HP}\mu_1$ is clearly the most important predictor to discriminate the RNase III class from other protein signatures (**figure 6**).

Figure 6 comes about here

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 computational cost, the DTM was validated using 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 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 predictability performance. Both cases showed values less than 0.5, which is an excellent result for the misclassification of the model.

Table 3 comes about here

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 group. 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.

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 [53]. 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 the 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**). The **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.

Table 4 should come here

This ANN model showed an overall classification in training, selection and test of 93.89, 93.34 and 90.65 %, respectively, which are quite good results taking into account the classification values reported for protein families with a higher degree of conservation [24]. The classification results derived from our alignment-free approach to classify RNase III membership is showed in **Table 5** and in File IISM for more details.

Table 5 comes about here

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. The Figure 7 depicts the network map for the best MLP model.

Figure 7 come about here

To validate the ANN model, we constructed the Receiver Operating Characteristic (ROC) curve for the training, selection and test subsets. In each case, the 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 (**figure 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 Fernandez and Caballero [54; 55; 56].

Figure 8 comes about here

4.4 Non-classical HMM in RNase III classification

In order to compare with other non-linear methodologies based on 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 homologues 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 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 recognized 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 previous reports the application of classical HMM on RNase III classification showed a major failing rate on a similar control subset [47].

Our free-alignment approach TI2BioP provide 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 to carry 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.

Table 6 comes about here

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 3.8μg/μL. It was also run on agarose gel 0.8% visualizing high integrity. 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 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 dataset. 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 classify easily 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 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 5.5×10^{-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. 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 require the building of a modified HMM-profile based on 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 information easily. On the other hand, we demonstrated that our strategy of 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. The **figure 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 CG1208 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 [57]. Enzymatic activity showed values of 5.858×10^5, 6.017×10^5, and 6.177×10^5 U/mg, respectively for each assay, and the mean value was 6.017×10^5 U/mg **(see Table7)**.

Figure 9 and Table 7 come about here

5. Conclusions

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

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/AA-AMB/104983/2008.

Supplementary Material

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:

References

[1] G. Agüero-Chapin, G. Pérez-Machado, R. Molina-Ruiz, Y. Pérez-Castillo, A. Morales-Helguera, V. Vasconcelos, A. Antunes, TI2BioP: Topological Indices to BioPolymers. Its practical use to unravel cryptic bacteriocin-like domains, Amino Acids. Doi:10.1007/s00726-010-0653-9 (2010).

[2] P.K. Strope, E.N. Moriyama, Simple alignment-free methods for protein classification: a case study from G-protein-coupled receptors, Genomics. 89 (2007) 602-12.

[3] S.F. Altschul, 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. 25 (1997) 3389-3402.

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

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

[6] A. de Jong, S.A. van Hijum, J.J. Bijlsma, J. Kok, O.P. Kuipers, BAGEL: a web-based bacteriocin genome mining tool, Nucleic Acids Res. 34 (2006) W273-9.

[7] S. Deshmukh, S. Khaitan, D. Das, M. Gupta, P.P. Wangikar, An alignment-free method for classification of protein sequences, Protein Pept Lett. 14 (2007) 647-57.

[8] M. Kumar, V. Thakur, G.P. Raghava, COPid: composition based protein identification, In Silico Biol. 8 (2008) 121-8.

[9] S. Roy, D. Martinez, H. Platero, T. Lane, M. Werner-Washburne, Exploiting amino acid composition for predicting protein-protein interactions, PLoS ONE. 4 (2009) e7813.

[10] K.C. Chou, Automated prediction of protein attributes and its impact to biomedicine and drug discovery, in: G. Alterovitz, R. BensonM.F. Ramoni, (Eds.), Automation in Proteomics and Genomics: An Engineering Case-Based Approach (Harvard-MIT interdisciplinary special studies courses), Wiley & Sons, UK, 2009, pp. 97-143.

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

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

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

[14] 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.

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

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

[17] M. Punta, B. Rost, Neural networks predict protein structure and function, Methods Mol Biol. 458 (2008) 203-30.

[18] R. Nair, B. Rost, Protein subcellular localization prediction using artificial intelligence technology, Methods Mol Biol. 484 (2008) 435-63.

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

[20] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, E.W. Sayers, GenBank, Nucleic Acids Res. 37 (2009) D26-31.

[21] D.H. Mathews, RNA secondary structure analysis using RNAstructure, Curr Protoc Bioinformatics. Chapter 12 (2006) Unit 12 6.

[22] G. Aguero-Chapin, H. Gonzalez-Diaz, R. Molina, J. Varona-Santos, E. Uriarte, Y. Gonzalez-Diaz, 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 (2006) 723-30.

[23] Y. Gutierrez, E. Estrada, MODESLAB 1.0 (Molecular DEScriptors LABoratory) for Windows, (2002).

[24] G. Agüero-Chapin, J. Varona-Santos, G.d.l. Riva, A. Antunes, T. González-Villa, E. Uriarte, H. González-Díaz, Alignment-Free Prediction of Polygalacturonases with Pseudofolding Topological Indices: Experimental Isolation from Coffea arabica and prediction of a New Sequence, J Proteome Res. 8 (2009) 2122-2128.

[25] S.G. Jacchieri, Mining combinatorial data in protein sequences and structures, Molecular Diversity (2000) 145–152.

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

[27] M. Randic, M. Vracko, On the similarity of DNA primary sequences, J Chem Inf Comput Sci. 40 (2000) 599-606.

[28] A. Nandy, Recent investigations into global characteristics of long DNA sequences, Indian J Biochem Biophys. 31 (1994) 149-55.

[29] G. Aguero-Chapin, A. Antunes, F.M. Ubeira, K.C. Chou, H. Gonzalez-Diaz, Comparative study of topological indices of macro/supramolecular RNA complex networks, J Chem Inf Model. 48 (2008) 2265-77.

[30] Z. Yuan, Prediction of protein subcellular locations using Markov chain models., FEBS Lett. 451 (1999) 23-6.

[31] E. Estrada, 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 (1996) 844-849.

[32] E. Estrada, Spectral Moments of the Edge-Adjacency Matrix of Molecular Graphs. 2. Molecules Containing Heteroatoms and QSAR Applications, J Chem Inf Comput Sci. 37 (1997) 320-328.

[33] H. Gonzalez-Diaz, M. Cruz-Monteagudo, D. Vina, L. Santana, E. Uriarte, E. De Clercq, 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 (2005) 1651-7.

[34] S. Markovic, Z. Markovic, R.I. McCrindle, Spectral moments of phenylenes, J Chem Inf Comput Sci. 41 (2001) 112-9.

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

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

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

[38] Statsoft, STATISTICA 7.0 (data analysis software system for windows), 2007.

[39] P.D. Dobson, A.J. Doig, Distinguishing Enzyme Structures from Non-enzymes Without Alignments, J. Mol. Biol. 330 (2003) 771–783.

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

[41] M. Cruz-Monteagudo, H. Gonzalez-Diaz, Unified drug-target interaction thermodynamic Markov model using stochastic entropies to predict multiple drugs side effects, Eur J Med Chem. 40 (2005) 1030-41.

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

[43] T. Date, W. Wickner, Isolation of the Escherichia coli leader peptidase gene and effects of leader peptidase overproduction in vivo, Proc Natl Acad Sci U S A. 78 (1981) 6106-10.

[44] P.E. March, J. Ahnn, M. Inouye, The DNA sequence of the gene (rnc) encoding ribonuclease III of Escherichia coli, Nucleic Acids Res. 13 (1985) 4677-85.

[45] A.K. Amarasinghe, I. Calin-Jageman, A. Harmouch, W. Sun, A.W. Nicholson, Escherichia coli ribonuclease III: affinity purification of hexahistidine-tagged enzyme and assays for substrate binding and cleavage, Methods Enzymol. 342 (2001) 143-58.

[46] K. Zhang, A.W. Nicholson, Regulation of ribonuclease III processing by double-helical sequence antideterminants, Proc Natl Acad Sci U S A. 94 (1997) 13437-41.

[47] G. Aguero-Chapin, H. Gonzalez-Diaz, G. de la Riva, E. Rodriguez, A. Sanchez-Rodriguez, G. Podda, R.I. Vazquez-Padron, 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 (2008) 434-48.

[48] A.W. Nicholson, Ribonucleases: Structures and Functions, Academic Press, Michigan 1997.

[49] M. Cruz-Monteagudo, H. Gonzalez-Diaz, E. Uriarte, Simple stochastic fingerprints towards mathematical modeling in biology and medicine 2. Unifying Markov model for drugs side effects, Bull Math Biol. 68 (2006) 1527-54.

[50] M. Cruz-Monteagudo, C.R. Munteanu, F. Borges, M.N. Cordeiro, E. Uriarte, H. Gonzalez-Diaz, 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 (2008) 9684-93.

[51] Y. Marrero-Ponce, M.T. Khan, G.M. Casanola Martin, A. Ather, M.N. Sultankhodzhaev, F. Torrens, R. Rotondo, Prediction of tyrosinase inhibition activity using atom-based bilinear indices, ChemMedChem. 2 (2007) 449-78.

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

[53] D.E. Rumelhart, J.L. McClelland, Parallel distributed processing: Explorations in the microstructure of cognition, MIT Press, Cambridge, MA, 1986.

[54] J. Caballero, M. Fernandez, 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 (2008) 1580-605.

[55] L. Fernandez, J. Caballero, J.I. Abreu, M. Fernandez, Amino acid sequence autocorrelation vectors and Bayesian-regularized genetic neural networks for modeling protein conformational stability: gene V protein mutants, Proteins. 67 (2007) 834-52.

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

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

FIGURES LEGENDS

Figure 1. (a) RNase III protein sequence from *Escherichia coli* BL21 substrain GG1108 (b) Pseudo folding of this sequence into a 2D-HP-lattice.

Figure 2. Building the 2D-HP map on the Cartesian axes for the protein fragment DEDKV. (a) The coordinates for each amino acid in the Cartesian system. (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.

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

Figure 4. Plot of the TIs's Means for Each Cluster (a) Division of the RNase III group into three clusters (b) Division of the Control Group into four clusters.

Figure 5. The architecture of the DTM. Decision Nodes are represented in black and terminal nodes in gray. 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.

Figure 6. Predictor Variable Importance Rankings, Rankings on scale from 0=low importance to 100=high importance.

Figure. 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).

Figure 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 curve of 0.95, 0.97 and 0.92, respectively.

Figure 9. Electrophoresis of the 25 kDa recombinant *E. coli* RNAse III from E.coli DH5 α: 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 induced bacteria; Lane 3: purified recombinant *E. coli* RNAse III.

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

Variance analysis RNase III-like proteins				
Protein Descriptors	**Between SS[a]**	**Within SS[b]**	**Fisher ratio *(F)***	***p*-Level[c]**
$^{HP}\mu_{10}$	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.

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)	Predict. 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.

Table 3. Classification results derived from CT for the training and the validation series. Predicted Class (row) x 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.

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 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.

Table 6. Classification results on 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).

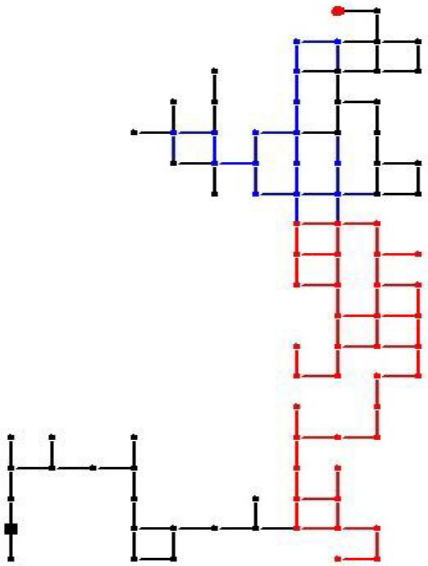
	DTM		**ANN-MLP**		**HMM modified**	
	RNase III	**CG**	**RNase III**	**CG**	**RNase III**	**CG**
Training	92.90	90.48	94.44	93.38	94.83	100
Overall	**91.64**		**93.89**		**96.11**	
Test	96.07	89.28	92.15	89.28	100	100
Overall	**92.52**		**90.65**		**100**	

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.

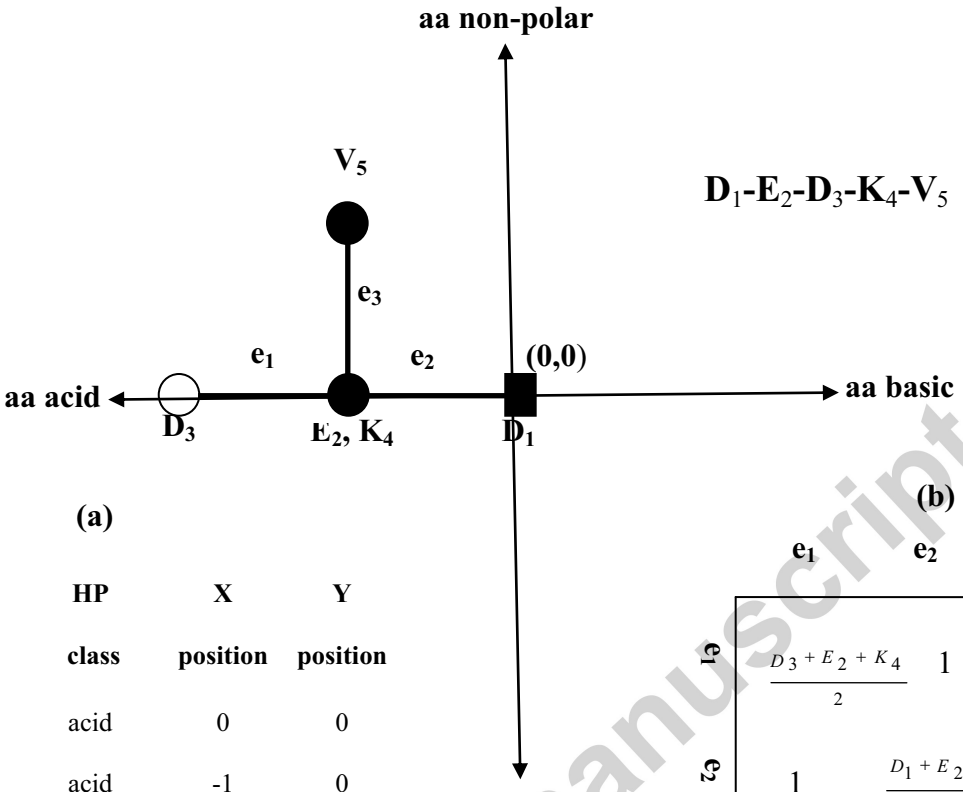
Enzyme	Enzymatic Activity		
nM	Experiment 1 10^5 U/mg	Experiment 2 10^5 U/mg	Experiment 3 10^5 U/mg
0.0	0.011	0.042	0.021
	0.023	0.016	0.011
	0.012	0.014	0.013
0.1	6.200	6.015	6.312
	6.512	5.912	6.801
	6.011	6.108	6.709
1.0	6.701	5.519	6.089
	6.603	5.808	5.816
	6.415	5.901	5.588
10.0	6.211	6.009	6.131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

MNPIVINRLQRKLGYTFNHQELLQ
QALTHRRASRKHNERLEFLGDSILS
YVIANALYHRFPRVDEGDMSRMR
ATLVRGNTLAELAREFELGECLRL
GTGELKSGGFRRESILADTVEALIG
GVFLESDIQTVEKLILNWYQTRVD
EISPGDKQKDPKTRVHEYLQGRHL
PLPTYLVVQVRGEAHDQEFTIHCQ
VSGLSEPVVGTGSSRRKAEQAAAE
QALKKLELE

(a)



(b)



(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	$\frac{V_5 + E_2 + K_4}{2}$

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

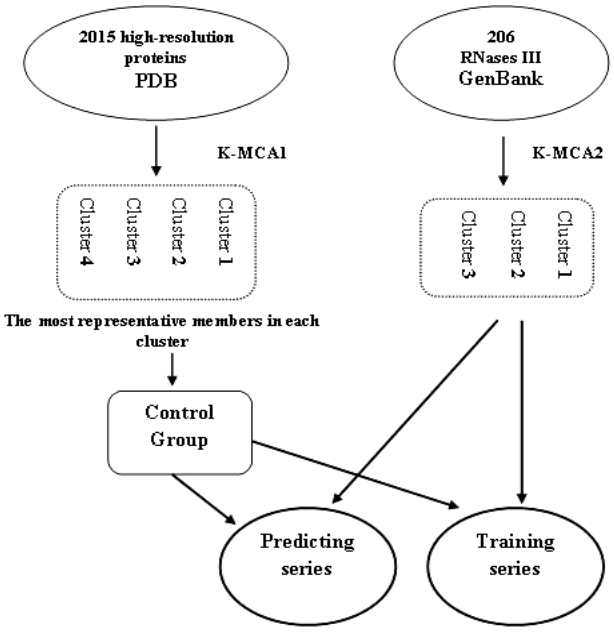


Figure 4. Plot of the TIs's Means for Each Cluster (A) Division of the RNase III group into three clusters (B) Division of the Control Group into four clusters

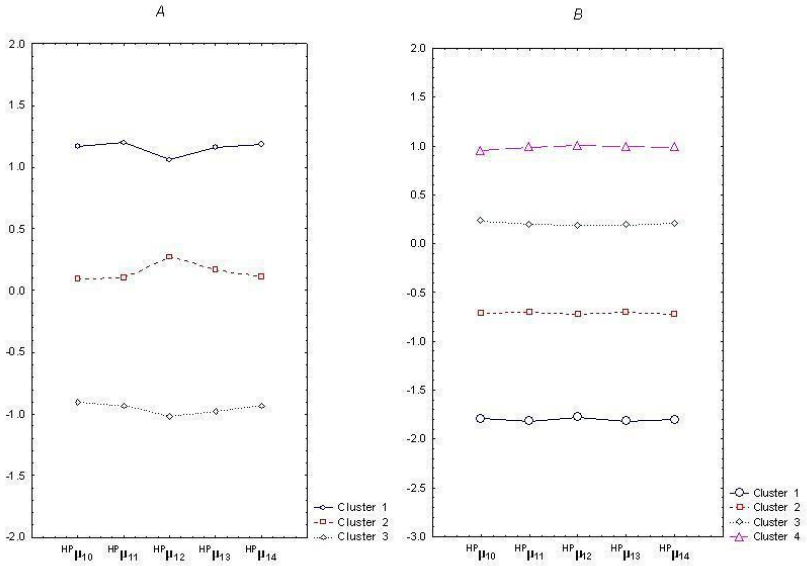


Figure 5. The architecture of the DTM. Decision Nodes are represented in blue and terminal nodes in red. The RNase III class is labeled with 1 using an intermittent line. Otherwise the control group is signed with -1 using a continuous line. Numbers at the right-corner of the nodes indicates tentative membership to one group. Numbers at the left-corner represent the node's number.

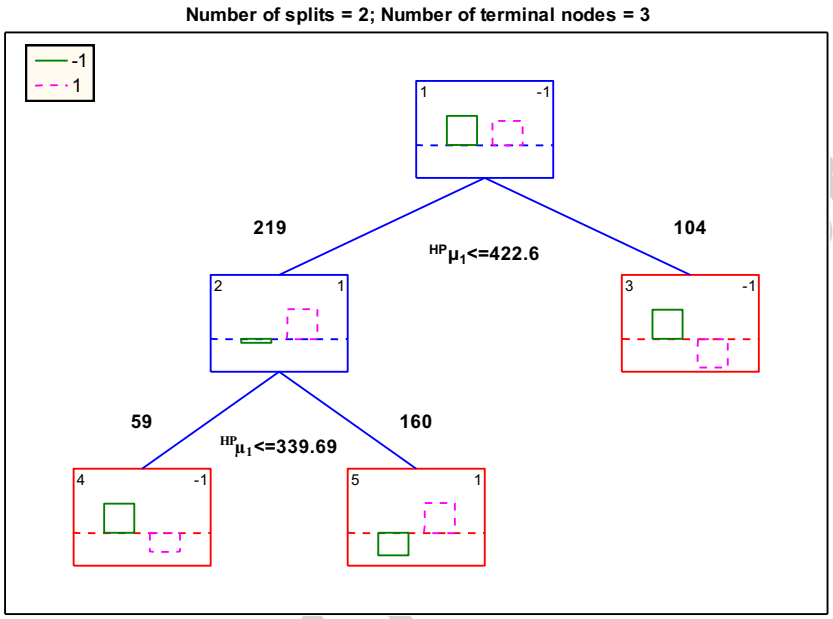


Figure 6. Predictor Variable Importance Rankings, Rankings on scale from 0=low importance to 100=high importance

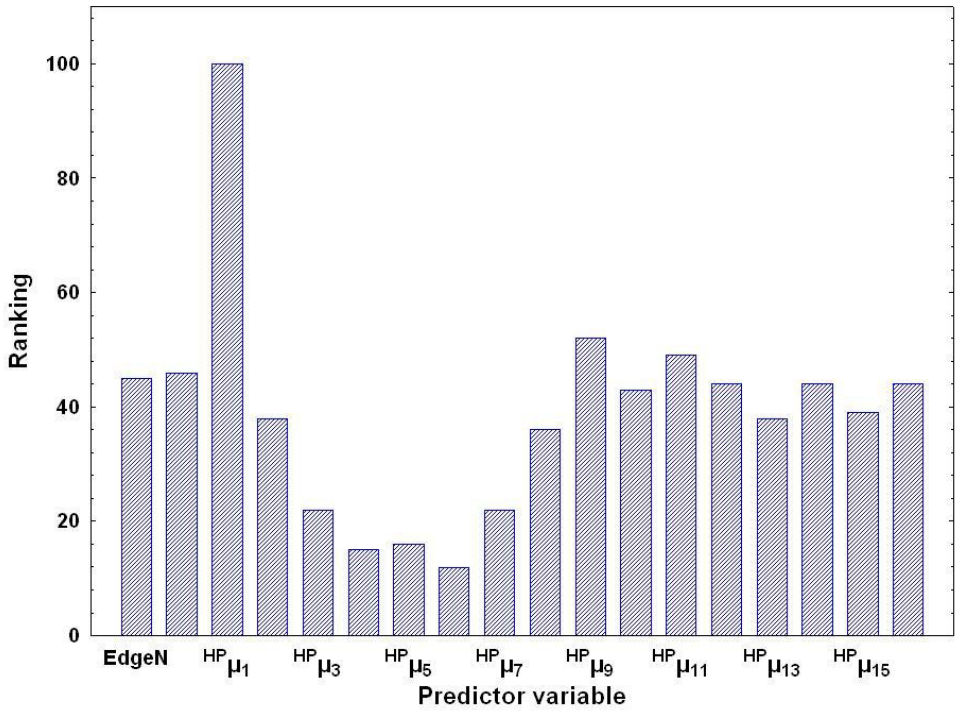


Figure. 7 The architecture of the MLP profile 7. It represents several input variables, four neurons in a one layer and only one output variable (from the left to the right)

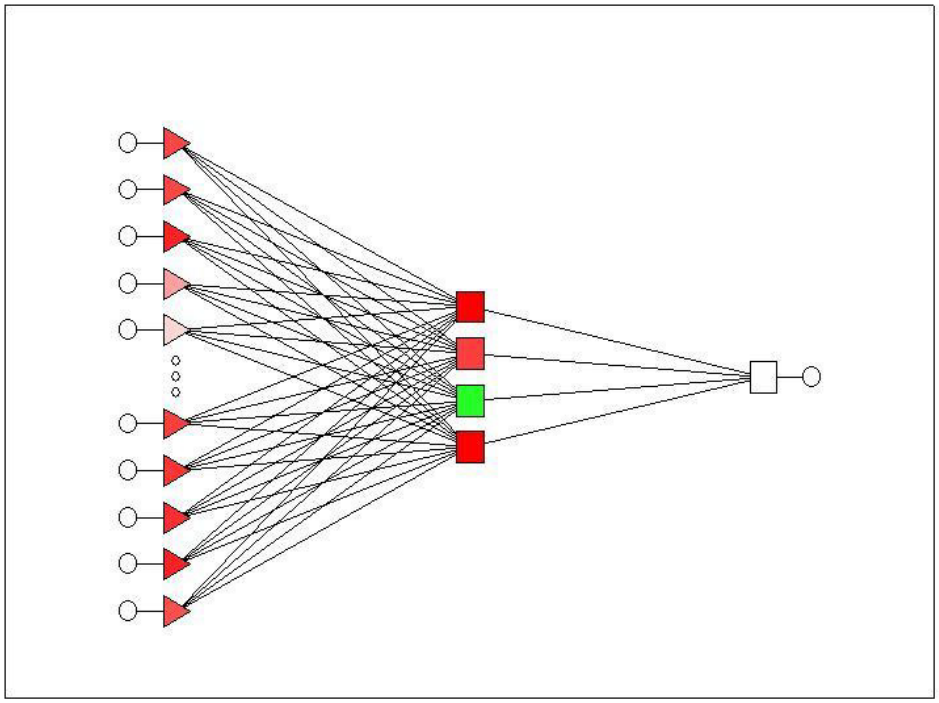
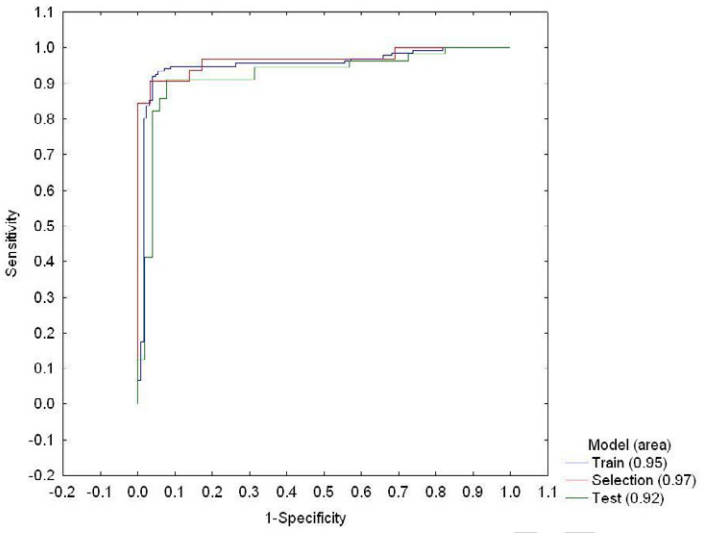


Figure 8. Receiver Operating Characteristic curve (ROC-curve) for the ANN-based model in training (blue line), selection (red line) and test (green line) sets with areas under curve of 0.95, 0.97 and 0.92, respectively.



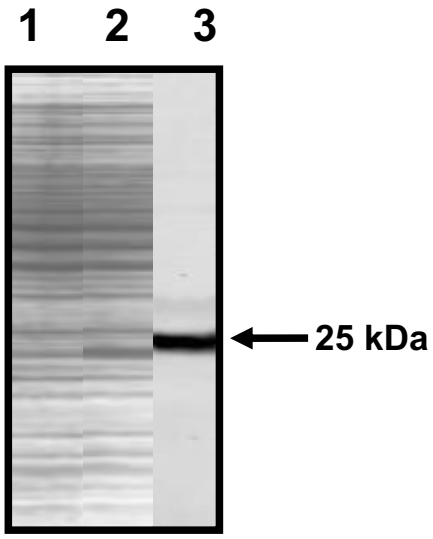


Figure 9. Electrophoresis of the 25 kDa recombinant *E. coli* RNAse III from E.coli DH5 α: pREC1 loaded in 12.5% PAGE-SDS and stained with coomassie brilliant. Lane 1: crude extract from non induced bacteria; Lane 2: crude extract from induced bacteria; Lane 3: purified recombinant *E. coli* RNAse III.