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

Guillermín Agüero-Chapín,[†,‡] Humberto González-Díaz,*[,†,§,||] Gustavo de la Riva,[⊥] Edrey Rodríguez,[#] Aminael Sánchez-Rodríguez,[‡] Gianni Podda,[†] and Roberto I. Vazquez-Padrón[★]

Dipartimento Farmaco Chimico Tecnologico, Universitá Degli Studi di Cagliari, Cagliari, 09124, Italy, CAP, Faculty of Chemistry and Pharmacy, IBP, and CBQ, UCLV, Santa Clara 54830, Cuba, Unit for Bioinformatics & Connectivity Analysis (UBICA), Institute of Industrial Pharmacy and Department of Organic Chemistry, Faculty of Pharmacy, USC, Santiago de Compostela 15782, Spain, CINVESTAV-LANGEBIO, Irapuato, Guanajuato 36821, México, Caribbean Vitroplants, Santo Domingo 1464, Dominican Republic, and Vascular Biology Institute, School of Medicine, University of Miami, Miami, Florida 33136

Received August 29, 2007

The study of type III RNases constitutes an important area in molecular biology. It is known that the *pac1*$^+$ gene encodes a particular RNase III that shares low amino acid similarity with other genes despite having a double-stranded ribonuclease activity. Bioinformatics methods based on sequence alignment may fail when there is a low amino acidic identity percentage between a query sequence and others with similar functions (remote homologues) or a similar sequence is not recorded in the database. Quantitative structure−activity relationships (QSAR) applied to protein sequences may allow an alignment-independent prediction of protein function. These sequences of QSAR-like methods often use 1D sequence numerical parameters as the input to seek sequence-function relationships. However, previous 2D representation of sequences may uncover useful higher-order information. In the work described here we calculated for the first time the spectral moments of a Markov matrix (MMM) associated with a 2D-HP-map of a protein sequence. We used MMMs values to characterize numerically 81 sequences of type III RNases and 133 proteins of a control group. We subsequently developed one MMM-QSAR and one classic hidden Markov model (HMM) based on the same data. The MMM-QSAR showed a discrimination power of RNAses from other proteins of 97.35% without using alignment, which is a result as good as for the known HMM techniques. We also report for the first time the isolation of a new Pac1 protein ***(DQ647826)*** from *Schizosaccharomyces pombe* strain 428-4-1. The MMM-QSAR model predicts the new RNase III with the same accuracy as other classical alignment methods. Experimental assay of this protein confirms the predicted activity. The present results suggest that MMM-QSAR models may be used for protein function annotation avoiding sequence alignment with the same accuracy of classic HMM models.

1. INTRODUCTION

RNase III is a double-strand-specific ribonuclease (dsRNase) that usually makes staggered cuts in both strands of a double helical RNA, although in some cases it cleaves once in a single-stranded bulge in the helix.[1,2] The primary biological function of this system is the specific processing of rRNA and mRNA precursors,[3−5] but it has also been implicated in other diverse phenomena such as mRNA turnover,[6] conjugative DNA transfer,[7] antisense RNA-mediated regulation, and others.[8,9] For instance, Dicer and Drospha are type III RNases responsible for the generation of short interfering RNAs (siRNAs) from long double-stranded RNAs during RNA interference (RNAi). Also, the cellular processing of shRNAs shares common features with the biogenesis of naturally occurring miRNA such as cleavage by nuclear type III RNase Drosha, export from the nucleus, processing by a cytoplasmic type III RNase Dicer, and incorporation into the RNA-induced silencing complex (RISC). Each step has a crucial influence on the efficiency of RNAi.[10−13] It involves both RNase proteins in several important biological processes as for instance the function of Dicer on the vascular system regulating embryonic angiogenesis probably by processing miRNAs, which regulate the expression levels of some critical angiogenic regulators in the cell.[14] Recently, RNAi has moved from a purely experimental technique to the stage of potential clinical applications such as a possible use for the treatment of spinocerebellar ataxia or amyotrophic lateral sclerosis.[15] Many other dsRNases have been characterized from a variety of prokaryotic and eukaryotic sources, and RNase III from

* Corresponding author phone: +34-981-563100; fax: +34-981 594912; e-mail: gonzalezdiazh@yahoo.es or qohumbe@usc.es. Corresponding author address: Faculty of Pharmacy, University of Santiago de Compostela 15782, Spain.
[†] Universitá Degli Studi di Cagliari.
[‡] IBP, and CBQ, UCLV.
[§] Institute of Industrial Pharmacy, Faculty of Pharmacy, USC.
[||] Department of Organic Chemistry, Faculty of Pharmacy, USC.
[⊥] CINVESTAV-LANGEBIO.
[#] Caribbean Vitroplants.
[★] University of Miami.

10.1021/ci7003225 CCC: $40.75
Published on Web 02/07/2008

Escherichia coli is an archetype of this class of enzymes.[6,16,17] The RNase III family consists of a growing number of enzymes that includes at least 33 bacterial and 22 eukaryotic enzymes.[18] There have been numerous reports of dsRNase activities in eukaryotic cells, some of which exhibited properties consistent with a role in pre-rRNA processing.[19−21]

One of the best candidates for eukaryotic RNase III homologues is the Pac1 RNase from *Schizosaccharomyces pombe*.[22−24] The Pac1 product is derived from *Schizosaccharomyces pombe* pac1$^+$ gene expression, which is also involved in the regulation of sexual development,[25] possibly through a mechanism that involves the processing of certain small nucleolar RNAs (snRNAs).[26] Pac1 works in eukaryotes as dsRNase and shares a functional similarity to RNase III from *E. coli*. This fact was proved either by measuring the ability of Pac1 to degrade double-stranded RNA in vitro or by expressing *pac1*$^+$ in *E. coli*, where it produced an activity that converted dsRNA into acid-soluble products.[23] Despite these observations the Pac1 gene product shows low homology with other RNAse III enzymes, particularly with those ones belonging to bacteria. The homology between the different RNase III enzymes varies in the range 20−84% depending on their evolutionary distance, suggesting a low level of primary structure conservation.[27] It has been reported that antibodies prepared against Pacl RNase have failed to react with RNase III.[23] The Pac1 gene product from *Schizosaccharomyces pombe* belongs to subclass II of the RNase III family, which is characterized by the presence of an N-terminal extension and includes fungal RNase III.[27,28] This contains 363 amino acids (aa), and only its C-terminal 230 residues share 25% amino acid identity with the *Escherichia coli* ribonuclease III.[23]

Methods based on sequence alignment have revealed a low amino acidic identity (20−40%) for the *pac1*$^+$ gene product with other typical RNases III, either isolated from bacteria or even from species that are genetically close.[27,29] However, experimental observations show the Pac1 protein to be a dsRNAse enzyme. This relatively low degree of conservation probably reflects the species-specificity of RNase III, which prevents genetic complementation between members of the RNase III family.[30]

All of the facts discussed above hinder the prediction of the Pac1 gene product as an RNase III-like enzyme using computational methods based on sequence alignment. In fact, bioinformatics methods based on sequence alignment may fail in general for cases of low sequence homology between the query and the template sequences deposited in the data base. The lack of function annotation (defined biological function) for the sequences deposited in databases and used as templates for function prediction constitutes another weakness of alignment approaches.[31,32] Recently, a group of researchers published in *PROTEOMICS* (2006) a review[33] on the growing importance of machine learning methods for predicting a protein functional class independently of sequence similarity. In this review the authors make reference to various papers on the topic, including their own work.[34−45] These methods often use as the input 1D sequence numerical parameters specifically defined to seek sequence-function relationships. For instance, the so-called pseudoamino acid composition approach[46,47] based on 1D sequence coupling numbers has been widely used to predict subcellular localization, enzyme family class, and structural class as well as other attributes of proteins based on their sequence similarity.[45,48−74] Alternatively, some authors generalized molecular indices that are classically used for small molecules[75,76] to describe protein sequences, such as the generalization of Broto−Moreau indices by Caballero and Fernández et al.[77] On the other hand, many authors have introduced 2D or higher dimension representations of sequences prior to the calculation of numerical parameters. This constitutes an important step in order to uncover useful higher-order information not encoded by 1D sequence parameters.[78−98] In addition, 2D graphs have been used for proteins and DNA sequences by other researchers. For example, Zupan and Randić used spectral-like and zigzag representations. These authors suggested an algorithm for encoding long strings of building blocks (like 4 DNA bases, 20 natural amino acids, or all 64 possible base triplets) using "zigzag" or "spectrum-like" representations.[99] Hydrophobic cluster analysis (HCA) constitutes another well-known technique for the 2D representation of protein sequences.[100] Randić et al. ultimately approached protein representations by using 2D schemes based on nucleotide triplet codons or virtual genetic code.[101] Finally, we introduced hydrophobicity-polarity (HP) 2D Cartesian or latticelike representations for proteins related to plant metabolism.[93]

In this work, we propose to use the spectral moments of a Markov matrix (MMM) associated to a 2D-HP-graph to numerically characterize protein sequences and seek a QSAR model to predict type III RNAses without alignment. First, we derived hydrophobicity-polarity (HP) 2D Cartesian or latticelike representations (also called maps or graphs) for RNase III and control group protein sequences.[93] We then calculated the MMM values of order k (symbolized as ${}^{SR}\pi_k$) to characterize the protein sequence. Spectral moments for many kinds of graphs have been used before for quantitative structure−activity relationships (QSAR) studies on proteins.[102−112] We subsequently developed a classifier to connect protein sequence information (represented by the ${}^{SR}\pi_k$ values) with the classification of sequences as RNAse III or not. In general, different kinds of classifiers have been used to derive protein sequence QSAR models.[113,114] We selected a linear discriminant analysis (LDA), which is a simple but powerful technique.[115−121] The use of this MMM-QSAR model enabled us to predict a novel recombinant Pac1 (rPac1) protein as an RNase III-like enzyme from a new isolate of *Schizosaccharomyces pombe*. Prediction was also supported by profile Hidden markov model (HMM) analysis and submission to BLASTp and InterPro[122] servers and demonstrated by experimental evidence.

2. MATERIALS AND METHODS

2.1. Computational Methods. A Markov model (MM), also called MARCH-INSIDE, was used to codify information about 81 RNase III protein sequences belonging to prokaryote and eukaryote species downloaded from the GenBank database. Briefly, our methodology considers as states of the Markov chain (MC) any atom, nucleotide, or amino acid (aa) depending on the kind of molecule to be described.[123,124] Therefore, MM deals with the calculation of the probabilities (${}^{k}p_{ij}$) with which the charge distribution of aa moves from any aa in the vicinity i at time t_0 to another aa j along the protein backbone in discrete time periods until a stationary state is achieved.[125,126]

Table 1. Classification Results Derived from the Model for Training and Validation Series

	MMM training			MMM validation			
total%	97.35	RNases	control	RNases	control	100	total%
RNases	93.44	**57**	4	**20**	0	100	RNases
control	100	0	**90**	0	**43**	100	control
	MMM all sequences			**HMM classic**			
total%	98.1	RNases	control	RNases	control	97.50	total%
RNases	95.1	77	4	**80**	1	98.75	RNases
control	100	0	133	5	**128**	96.24	control

Table 2. Enzymatic Assay of Double-Stranded RNase Recombinant Pac1 ***DQ647826*** Extracted from *Schizosaccharomyces pombe* Strain 428-4-1

conc. rPac 1	1nM	10 nM	100 nM
EUV[a]	6.2×10^5	7.4×10^5	7.2×10^5
	6.0×10^5	6.8×10^5	7.3×10^5
	6.6×10^5	6.9×10^5	7.9×10^5
mean	6.4×10^5	7.0×10^5	7.5×10^5

[a] Enzymatic unit value for rPac 1 (U/mg).

Each RNase III sequence was labeled by its accession number; see Table 1 in the Supporting Information. The control group consists of 133 proteins, which were selected from 2184 high-resolution proteins in a structurally nonredundant subset of the Protein Data Bank (PDB); most of the data were published by other authors to distinguish enzymes and nonenzymes without alignment[127] (see Table 2 in the Supporting Information). Many researchers have demonstrated the possibility of predicting protein function from sequences,[128] and we used 2D-HP graphs to encode information about RNase III amino acid sequences.[93] We then calculated for the first time the ${}^{HP}\pi_k$ values for these graphs. As can be seen from the discussion above, we selected ${}^{HP}\pi_k$ based on the utility of other nonstochastic spectral moments[103−112] as well as other MMMs and other stochastic parameters.[102,129−131]

It is important to point out that this 2D graphical representation for proteins is similar to those previously reported for DNA,[92,96,97] but the 20 different amino acids are regrouped into HP classes instead of using 4 types of bases. These four groups characterize the HP physicochemical nature of the amino acids as polar, nonpolar, acidic, or basic.[132] The 2D-HP graph for the deduced amino acid sequence of rPac1 protein, obtained from *Schizosaccharomyces pombe* strain 428-4-1 (uploaded by our group with accession number ***DQ647826***), is shown in Figure 1. It is worth noting that 363 amino acids are rearranged in a 2D space compacting protein representation. 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) increase by +1 the abscissa axis coordinate for an acid amino acid (rightwards-step), or b) decrease by −1 the abscissa axis coordinate for a basic amino acid (leftwards-step), or c) increase by +1 the ordinate axis coordinate for a polar amino acid (upward-step), or d) decrease by −1 the ordinate axis coordinate for a nonpolar amino acid (downward-step).

2.2. 2D-HP Graph MMMs Used as Sequence Numerical Descriptors. After the representation of the sequences we assigned to each graph a stochastic matrix ${}^1\Pi$. Note that

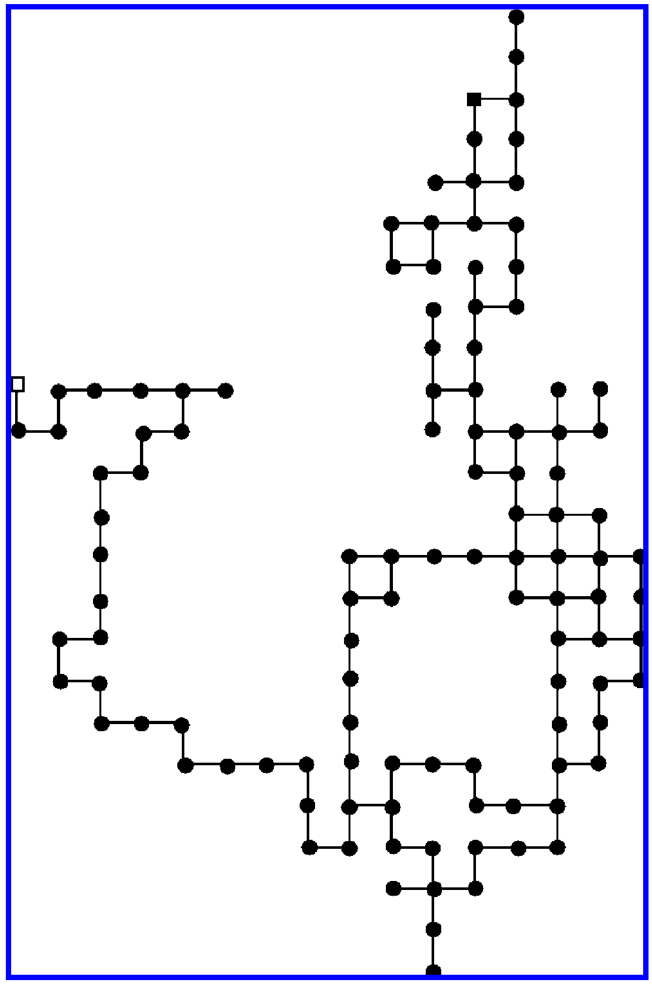

Figure 1. 2D Cartesian representation for the amino acid sequence of the rPac1 protein *Schizosaccharomyces pombe* strain 428-4-1, GenBank accession number ***DQ647826***. Note that a node may contain more than one amino acid, which ensures graph compactness.

the number of nodes (n) in the graph is equal to the number of rows and columns in ${}^1\Pi$ but may be equal to or even smaller than the number of amino acids or DNA bases in the sequence. The elements of ${}^1\Pi$ are the probabilities ${}^1p_{ij}$ of reaching a node n_i with charge Q_i moving through a walk of length $k = 1$ from another node n_j with charge Q_j[133]

$$p_{ij} = \frac{Q_j}{\sum_{m=l}^{n} \alpha_{il} \cdot Q_l} \tag{1}$$

where α_{ij} equals 1 if the nodes n_i and n_j are adjacent in the graph and equal to 0 otherwise. Q_j is equal to the sum of the electrostatic charges of all amino acids placed at this node. It then becomes straightforward to carry out the calculation of the spectral moments of ${}^1\Pi$ in order to numerically characterize the protein sequence

$$\mathrm{MMM}_k = {}^{SR}\pi_k = \sum_{i=j}^{n} {}^k p_{ij} = \mathrm{Tr}[({}^1\Pi)^k] \tag{2}$$

where Tr is called the trace and indicates that we sum all the values in the main diagonal of the matrices ${}^k\Pi = ({}^1\Pi)^k$, which are the natural powers of $^{1}\Pi$. The present class of MMMs encodes in a stochastic manner the distribution of the amino acid properties (charge) through all of the nodes placed at different distances in the 2D-HP lattice. Expansion of expression 2 for $k = 0$ gives the order zero MMM_0 ($^{HP}\pi_0$); for $k = 1$ the short-range MMM_1 ($^{HP}\pi_1$), for $k = 2$ the middle-range MMM_2 ($^{HP}\pi_2$), and for $k = 3$ the long-range MMMs. This extension is illustrated for the linear graph n_1-n_2-n_3, which is characteristic of the sequence (Asp-Glu-Asp-Lys); please note that the central node contains both Glu and Asp:[93]

$$^{HP}\pi_0 = \mathrm{Tr}[(^{1}\Pi)^0] = \mathrm{Tr}\left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right) = 3 \tag{2a}$$

$$^{HP}\pi_1 = \mathrm{Tr}[(^{1}\Pi)^1] = \mathrm{Tr}\left(\begin{bmatrix} ^{1}p_{11} & ^{1}p_{12} & 0 \\ ^{1}p_{21} & ^{1}p_{22} & ^{1}p_{23} \\ 0 & ^{1}p_{32} & ^{1}p_{33} \end{bmatrix}\right) = {}^{1}p_{11} + {}^{1}p_{22} + {}^{1}p_{22} \tag{2b}$$

$$^{HP}\pi_2 = \mathrm{Tr}[(^{1}\Pi)^1] = \mathrm{Tr}\left(\begin{bmatrix} ^{1}p_{11} & ^{1}p_{12} & 0 \\ ^{1}p_{21} & ^{1}p_{22} & ^{1}p_{23} \\ 0 & ^{1}p_{32} & ^{1}p_{33} \end{bmatrix} \cdot \begin{bmatrix} ^{1}p_{11} & ^{1}p_{12} & 0 \\ ^{1}p_{21} & ^{1}p_{22} & ^{1}p_{23} \\ 0 & ^{1}p_{32} & ^{1}p_{33} \end{bmatrix}\right) = {}^{2}p_{11} + {}^{2}p_{22} + {}^{2}p_{22} \tag{2c}$$

$$^{HP}\pi_2 = \mathrm{Tr}[(^{1}\Pi)^2] = \mathrm{Tr}\left(\begin{bmatrix} ^{1}p_{11} & ^{1}p_{12} & 0 \\ ^{1}p_{21} & ^{1}p_{22} & ^{1}p_{23} \\ 0 & ^{1}p_{32} & ^{1}p_{33} \end{bmatrix} \cdot \begin{bmatrix} ^{1}p_{11} & ^{1}p_{12} & 0 \\ ^{1}p_{21} & ^{1}p_{22} & ^{1}p_{23} \\ 0 & ^{1}p_{32} & ^{1}p_{33} \end{bmatrix} \cdot \begin{bmatrix} ^{1}p_{11} & ^{1}p_{12} & 0 \\ ^{1}p_{21} & ^{1}p_{22} & ^{1}p_{23} \\ 0 & ^{1}p_{32} & ^{1}p_{33} \end{bmatrix}\right) = {}^{3}p_{11} + {}^{3}p_{22} + {}^{3}p_{22}. \tag{2d}$$

All calculations of $^{HP}\pi_k$ values for protein sequences of both groups were carried out with our in-house software MARCH-INSIDE, version 2.0, including sequence representation.[134] We proceeded to upload a row data table with eleven $^{HP}\pi_k$ values for each sequence ($k = 0, 1, 2, ...10$) and grouping variable RNaseIII-score = 1 (for RNAses) and −1 (for control group sequences) to statistical analysis software.[135] The overall methodology is represented schematically in order to improve the understanding of our approach (see Figure 2).

2.3. Statistical Analysis. K-Means Cluster Analysis. The negative group was selected from 2184 proteins with diverse functions (enzymes and nonenzymes) recorded in the PDB, as mentioned before. Our negative subset was designed according to K-means cluster analysis **(**k-MCA).[136] The method consists of carrying out a partition of the starting group made up by a non-RNase III series of proteins into several statistically representative clusters of sequences. Thus, one may select the members to conform to the negative subset from all of these clusters. This procedure ensures that the main protein classes (as determined by the clusters derived from k-MCA) will be represented in the model control group, thus allowing the representation of the entire 'experimental universe'. The spectral moment series was

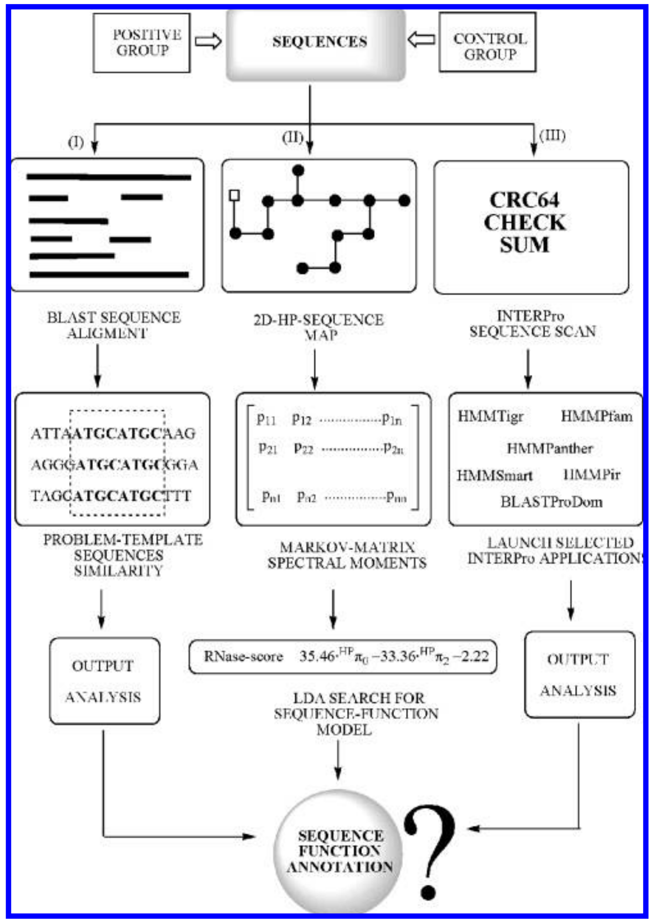

Figure 2. Schematic representation of the steps given in this work.

explored as clustering variables in order to carry out k-MCA. The procedure described above is represented graphically in Figure 3, where a cluster analysis was carried out to select a representative sample for the control group.

2.4. Linear Discriminant Analysis. LDA forward stepwise analysis was carried out for variable selection to build up the model.[115−121] All of the variables included in the model were standardized in order to bring them onto the same scale. Subsequently, a standardized linear discriminant equation that allows comparison of their coefficients was obtained.[137] The square of Mahalanobis's distance (D^2) and Wilk's (λ) statistic ($\lambda = 0$ perfect discrimination, being $0 < \lambda < 1$) were examined in order to assess the discriminatory power of the model. Pac1 protein was submitted to BLASTp to show graphically the similarity of the sequence compared to other RNases III. Each sequence presented in this study was also submitted to the InterPro server[122] in order to compare our methodology with other classical sources of predictive functional annotation. InterPro consists of a database of protein families, domains, and functional sites in which identifiable features found in known proteins can be applied to unknown protein sequences.

3. EXPERIMENTAL SECTION

3.1. Strains and Culture Media. The *Schizosaccharomyces pombe* strain 428-4-1 was routinely grown in yeast extract (YEB) medium at 30 °C during 12 h. Bacterial strain *Escherichia coli* DH5α was grown in luria broth (LB). Transformed bacteria were recovered in the same LB

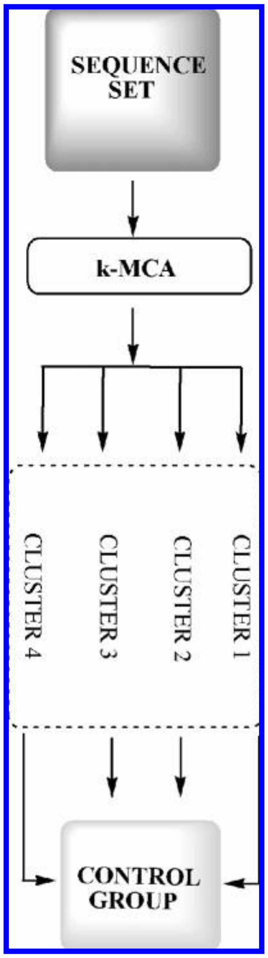

Figure 3. k-MCA procedure for control group design.

medium but supplemented with carbenicillin at 100 μg/mL. Media were also supplemented with bacteriological agar when required.

3.2. Total DNA Extraction. A colony from *Schizosaccharomyces pombe* strain 428-4-1 was inoculated in 5 mL of YEB medium and grown at 30 °C during 12 h until $OD_{600} = 0.5$. From this culture, 250 μL was transferred to 50 mL of the same medium and grown overnight at the same temperature. When the $OD_{600} = 0.8$, cells were collected by centrifugation and broken using small glass pearls. A cellular pellet was resuspended in 500 μL of 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). Total DNA solution was measured at 260 nm in a GENESYS 10 spectrophotometer, reaching a concentration of 3.8 μg/μL. The solution was also run on agarose gel (0.8%), and high integrity was seen.

3.3. Primer Design. Forward (PAC5′) 5′-ccc***ATGG***GACGGTTTAAGAGGCATC-3′ and reverse (PAC3′) 5′-gtgg***ggttaa***cgggcaaac**TTA**G-3′ primers were designed based on the previously reported *pac1*$^+$ coding sequence from *Schizosaccharomyces pombe* mutant snm1-1. The primer sequences show the restriction sites Nco1 and Kpn1 introduced at the 3′ and 5′ ends, i.e., the first ATG and the stop TTA codon. The coding regions are shown in capital letters.[138]

3.4. PCR Amplifications. Amplification of the *pac1*$^+$ gene from *Schizosaccharomyces pombe* was performed by standard PCR from its total DNA. The reaction mixture containing 10 ng of template, 1 mM of each dNTP, 1.5 mM $MgCl_2$, 2 μM of each PAC5′ and PAC3′ primers, 1× buffer Taq Pol (Gibco BRL), and 2.5 U Taq Pol (Gibco) was completed to a total volume of 50 μL. The PCR was carried out using a thermocycler (Perkin-Elmer 2400) programmed as follows: 5 min initial template denaturation at 94 °C, cycle steps—1 min template denaturation at 94 °C, 2 min primer annealing at 45 °C, 2 min primer extension at 72 °C for 30 cycles, plus a final extension step at 72 °C for 5 min.[29,30,138] PCR reaction showed a band coinciding with the size of the reported *pac1*$^+$ ORF.[138]

3.5. Plasmid Construction and Sequencing. The PCR amplification product was purified using a GEL Band Purification kit (*AmershamPharmaciaBiotech*) and ligated to pMOS-Blue T-vector (*AmershamPharmaciaBiotech*). The ligation was transformed into electrocompetent *E. coli* DH5α by electroporation in 0.2 mm cuvettes using a Gene Pulser Machine (BioRad) (12.5 kV, 25 μF, 1000 ω). The transformation was plated onto LB medium supplemented with 40 μL of 20 μg/mL X-gal solution and 4 μL of isopropylthio-β-D-galactoside from a 200 μg/mL IPTG solution per plate and allowed to grow overnight at 37 °C. White colonies—presumably carrying the recombinant pac1 gene inserted in pMOS-Blue T-vector, named pRSPac1—were selected, and plasmid DNA was extracted for analysis of the cloned fragment by restriction enzymes. Sequencing of the cloned fragment was performed using an ABI 3700 sequencer (Applied Biosystems),[139] and this showed a product of 1.111 Kb.

3.6. Purification of Recombinant Pac1. A single colony of *E. coli* DH5α with pRSPac1 was grown overnight at 30 °C in 5 mL of LB medium supplemented with carbencillin at 100 μg/mL. 250 μL of culture was then inoculated to 250 mL of the same medium supplemented with carbenicillin (100 μg/mL) and grown under the same culture conditions until $OD_{600} = 0.8$; at this point 50 μL of 200 μg/mL IPTG solution was added to the culture. Three hours after induction, cells were harvested by centrifugation and washed with 15 mL of 50 mM tris-HCl (pH 8), 100 mM NaCl, and 1 mM EDTA. Cells were collected by centrifugation and stored at −70 °C overnight. Around 3 g of frozen cells was resuspended in 15 mL of lysis buffer (1% NP40, 0.5% sodium deoxycholate, 0.1 M NaCl, 30 mM Tris-HCl (pH 8), 1 mM EDTA), and 5 mM $MgCl_2$ and DNase1 (10 μg/mL) were added. The cell suspension was incubated on ice for 10 min. Inclusion bodies were collected by washing four times with lysis buffer and twice with 50 mM Tris-HCl 5 mM (pH 8), 1 mM DTT. Finally, the sample was dissolved in 5 mL of loading buffer and boiled in a water bath for 10 min. The total volume of extract was divided into five preparative PAGE electrophoresis samples containing 1 mL of protein extract, which were run in 12% gel. The component corresponding to 45.5 kDa recombinant Pac1 protein was visualized by staining with an aqueous solution of 0.05% Coomassie brilliant blue R250. In each case the recombinant protein was excised from polyacrylamide gel, recovered by electroelution, combined, concentrated with Centricon-10 (Amicon) to 0.5 mL, and diluted to 1.5 mL with a storage buffer to a final composition of 500 mM NaCl, 20 mM sodium phosphate (ph 7.4), 67 mM imidazole, 1 mM DTT, 1 mM EDTA, and 30% glycerol. The recPac1 preparation was stored at −20 °C.[29,30,138]

3.7. Synthesis and Preparation of Complementary RNA Strands. The enzymatic assay of recombinant Pac1 was carried out according to the optimized conditions described by Rotondo and Frendewey.[29] In a previous experiment (data not shown) we amplified by PCR a fragment corresponding to the fourth intron of *Schizosaccharomyces pombe* β-tubuline from its total DNA and inserted the amplified fragment into pBluescript II KS (−) for further in vitro transcription purposes. The integrity of the amplified sequence and transcriptional fusion was tested by sequencing. We reproduced exactly the described assay to compare the activity of our recombinant enzyme with the results from other reports. This construction was used as a template for the PCR of fragments corresponding to transcriptional-fusion suitable for the synthesis of both complementary strands of dsRNA substrate for an in vitro transcription reaction. For this purpose the following primers were synthesized: a) 5′- gctcggaattaaccctcactaag↓ggaacGTAGGTTTTTTGCTTTC-3′ (T3 promoter in lower case, 5′ end of the *Schizosaccharomyces pombe* β-tubuline fourth intron in upper case) and b) 5′-ggtacctaatacgactcactatag↓ggagaCTACAGTCGTCAGTAC-3′ (T7 promoter in lower case, complement of the 3′ end of the *Schizosaccharomyces pombe* β-tubuline fourth intron in upper case).

The arrows indicate the transcription initiation site. The PCR products were purified, and 50 ng of each was used to synthesize both complementary strands of the dsRNA Pac1 substrate. The transcription reactions were prepared in a final volume of 20 μL containing 40 mM Tris-HCl (pH 7.9), 6 mM $MgCl_2$, 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 T3 or T7 RNA polymerase (Amersham Pharmacia Biotech). In the case of the transcription reaction driven by the T3 promoter, the addition of 50 mM NaCl to the reaction mixture was required. In all cases the reactions were prepared on ice and were then incubated at 37 °C during 10 min. The resulting transcripts were treated with DNAse I (Promega), phenol extraction and precipitation with 2.5 V/V of absolute ethanol were carried out, and the samples were stored overnight at −70 °C. The complementary RNA strands were collected by centrifugation at 16 000*g* during 10 min at 4 °C. Finally, the pellets were washed with 70% ethanol, dried, resuspended in diethyl pyrocarbonate treated with distilled water, and stored at −70 °C.

3.8. Preparation of dsRNA Substrate for Pac1 Enzymatic Assay. Equimolar quantities of both complementary strands were mixed in diethyl pyrocarbonate and treated with distilled water to give a final volume of 50 μL. The mixture was heated during 10 min at 100 °C in a water bath. The whole bath was then firmly closed and placed into thermal box overnight to allow annealing of both complementary strands into the dsRNA substrate. The unpaired ends and 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. The substrate for the Pac1 assay consisted of 101 bp dsRNA, identical to the substrate used by Rotondo and Frendewey.[29]

3.9. Enzymatic Assay of Recombinant Pac1. The Pac1 assay was carried out using the following conditions: 30 mM Tris-HCl (pH7.6), 1 mM DTT, 5 mM of $MgCl_2$, 10 nM of dsRNA substrate, and different quantities (0, 1, 10, 100 nM) of purified recombinant Pac1 enzyme. Enzymatic reactions were completed on ice, started by the addition of 0.1 V of 50 mM $MgCl_2$, incubated at 30 °C for 10 min, and stopped by the addition of 500 μL of 5% ice-cooled TCA followed by 15 min on ice. The aliquots were centrifuged at 16 000*g* during 5 min in a 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 Pac1 enzyme. The procedure was repeated three times with three repetitions per experiment.[29,30,138]

4. RESULTS AND DISCUSSION

4.1. MMM-QSAR Model To Predict Type III RNAses without Alignment. Many different parameters can be used to encode protein sequence information and further assign or predict the function or physical properties of proteins and their mutants.[140,141] The present approach involves the calculation of different sequence parameters based on MMs, which can be applied to different kinds of molecular graphs[131] including DNA, RNA, and proteins.[93,142] MMs have been applied successfully to genomics and proteomics and represent an important tool for analyzing biological sequence data. In particular, MMs have been used for protein folding recognition[143] and the prediction of protein signal sequences.[144,145] MMs have also been applied to predict alpha turns[146] and beta turns[147] as well as other tight turns and their types.[148] Particularly, MMs have been further used to predict the specificity of GalNAc-transferase[149] and cleavage sites in proteins by proteases,[150−153] greatly stimulating the development for drug design against AIDS and SARS.[154−162] In this work we calculated MMMs (${}^{HP}\pi_k$) of the stochastic matrix that describe the distribution of the amino acids of the protein sequence in the 2D-HP graph. This calculation was carried out for two groups of protein sequences, one made up of RNase III-like enzymes and the other formed by heterogeneous proteins. This last group contains 133 members, and these were selected as follows:

Original data were submitted to k-MCA as described previously.[136] The k-MCA divided the data into four clusters containing 439, 684, 592, and 469 members, respectively. Selection was based on the distance from each member with respect to the cluster center (Euclidean distance). We selected the closer cases to the center in order to ensure the inclusion of representative members of each cluster in the control group. Depending on the cluster size, a proportional number of proteins were set; 27 cases were taken from the first cluster, 42 from the second, 36 from the third, and 28 from the fourth to give a total of 133 members in the control group. We always bore in mind the principle of discriminant analysis in terms of balancing the size of the control group with respect to the RNase III group. A simple MMM-QSAR was then developed to classify a novel sequence as RNase III or not. The best equation found for this purpose was

$$\text{RNaseIII} - \text{score} = 35.46 \times {}^{HP}\pi_0 - 33.36 \times {}^{HP}\pi_2 - 2.22 \quad (3)$$

The statistical parameters for the above equation were Wilk's statistic ($\lambda = 0.18$), Mahalanobis's distance ($D^2 =$

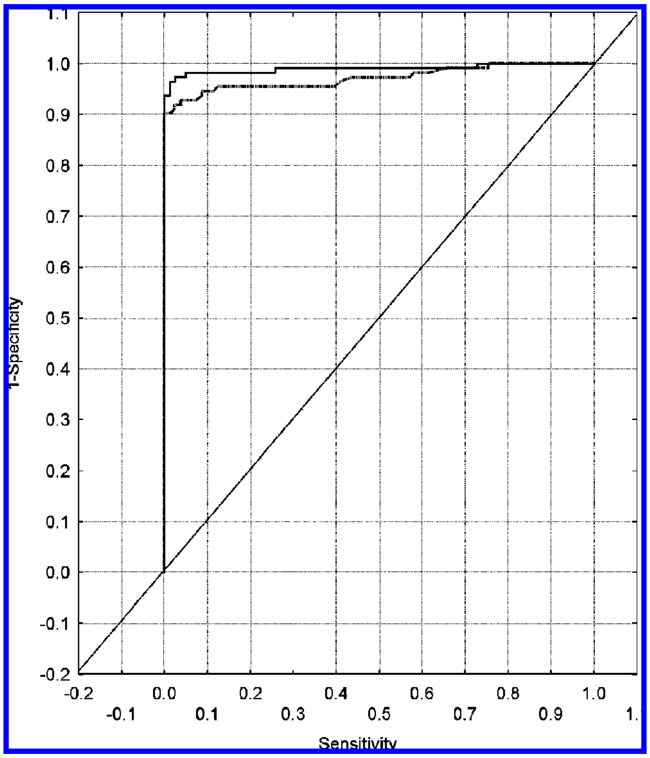

Figure 4. Receiver operating characteristic curve (ROC-curve) for training (dark line), validation (dot line), and random classifier (tight line) with areas under the curve of 0.99, 0.97, and 0.5, respectively.

16.36), and error level (p-level < 0.001).[163] This discriminant function misclassified only four cases out of 214 proteins used in both the training and validation series, reaching a high level of accuracy of 98.13%. More specifically, the model classified correctly 77/81 (95.06%) of RNase III-like enzymes and 100% of the control group. The respective classification matrices for training and cross-validation are depicted in Table 1. Analyzing the definition of MMMs in above equations, it is important to highlight that the combination of a positive contribution of ${}^{HP}\pi_0$ and a negative contribution of ${}^{HP}\pi_2$ in eq 3 points to a HP folding rule for the biological activity. The higher the number of HP-folded nodes (evaluated by ${}^{HP}\pi_0$) and the lower the number of middle-range self-return HP-folding walk (${}^{HP}\pi_2$) the higher will be the assigned activity score. Thus, highest scores for classification of RNase III-class are proportionally related not only to the content of hydrophobic residues in the protein sequence but also to the middle-range patterns formed by the backbone in the HP-lattice folding process. Otherwise, the folding patterns at long-range ($MMM_{3...}$) within the lattice do not directly affect the final assigned biological activity score.

A validation procedure was subsequently performed in order to assess the model predictability. This validation was carried out with an external series of 20 RNase III-like proteins and a further 43 diverse proteins (see Table 1). The present model showed an average predictability of 100% for each group, which is remarkable in comparison to results obtained by other researchers on using the LDA method in QSAR studies.[164−167] These results are also consistent with many others we have recently reviewed in-depth and published in the form of a review article on the uses of different networklike indices in small-sized, nucleic acids, and proteins QSAR.[167] In addition, we carried out a clas-

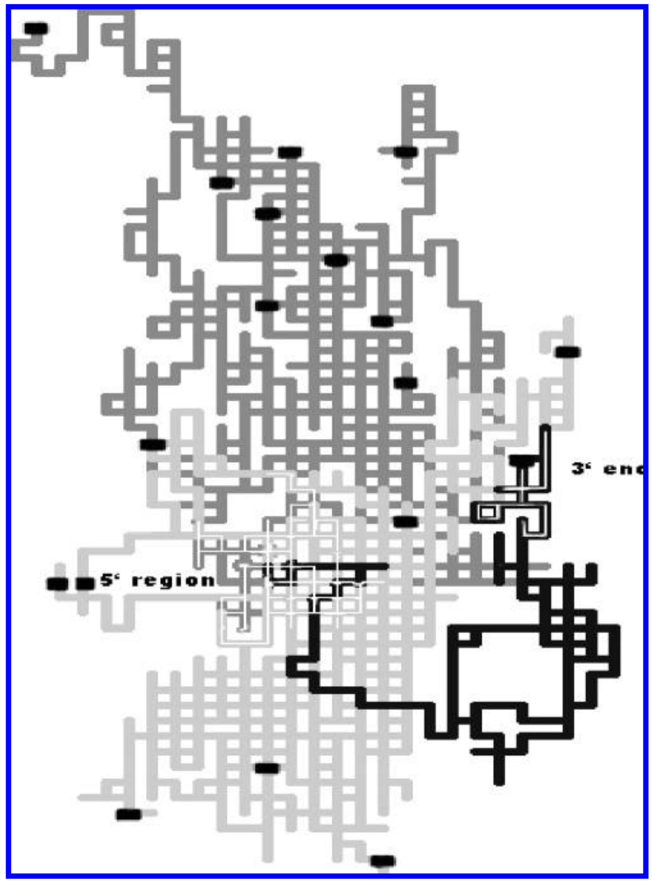

Figure 5. 2D-HP map superposition of RNases from prokaryotes (dark gray), eukaryotes (in light gray), and rPac1 ***DQ647826*** from *Schizosaccharomyces pombe* strain 428-4-1 (in black).

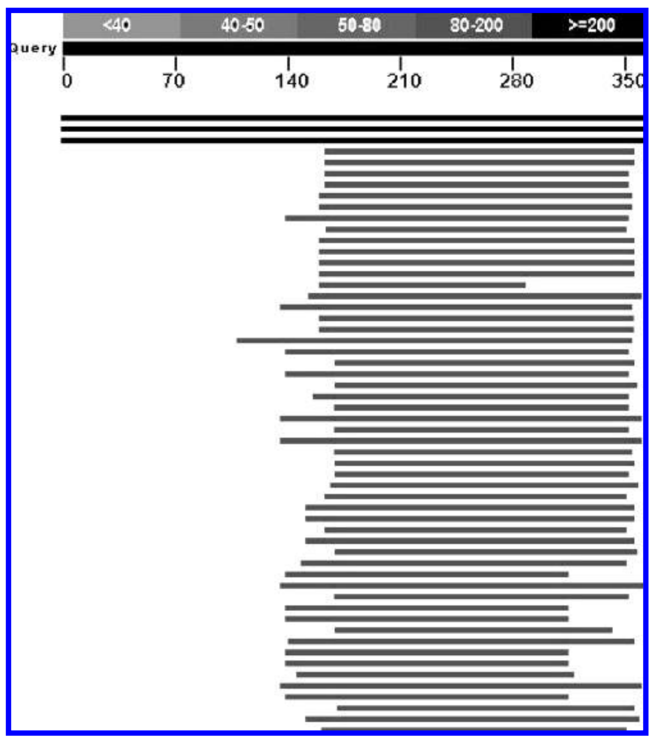

Figure 6. BLASTp analysis for rPac1 protein sequence ***DQ647826***. Note that the scale of scoring is progressive in darkness. Sequence names are not depicted.

sification analysis with all of the proteins included. These results provide further evidence of the robustness of the

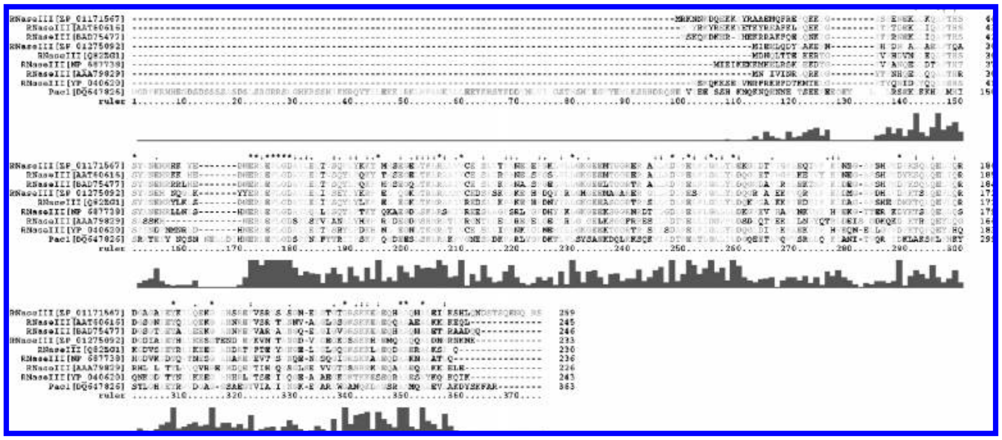

Figure 7. Clustal X sequence alignment involving RNase III-like enzymes. Each sequence is represented by its accession to the GenBank Database Protein. Sequences used in the alignment were represented previously in the Cartesian 2D system **(**Figure 3**)**. We use sequences from bacteria and rPac1 from *S. pombe*. [ZP_01171567] *Bacillus sp*. NRRL B 14911, [AAT60616] *Bacillus thuringiensis*, [BAD75477] *Geobacillus kaustophilus* HTA426, [ZP_01275092] *Lactobacillus reuteri*, [Q82ZG1] *Enterococcus faecalis*, [NP_687738] *Streptococcus agalactiae* 2603VR, [AAA79829] *E. coli*, [YP_040620] *Staphylococcus aureus* MRSA252, [DQ647826] *S. pombe* strain 428-4-1.

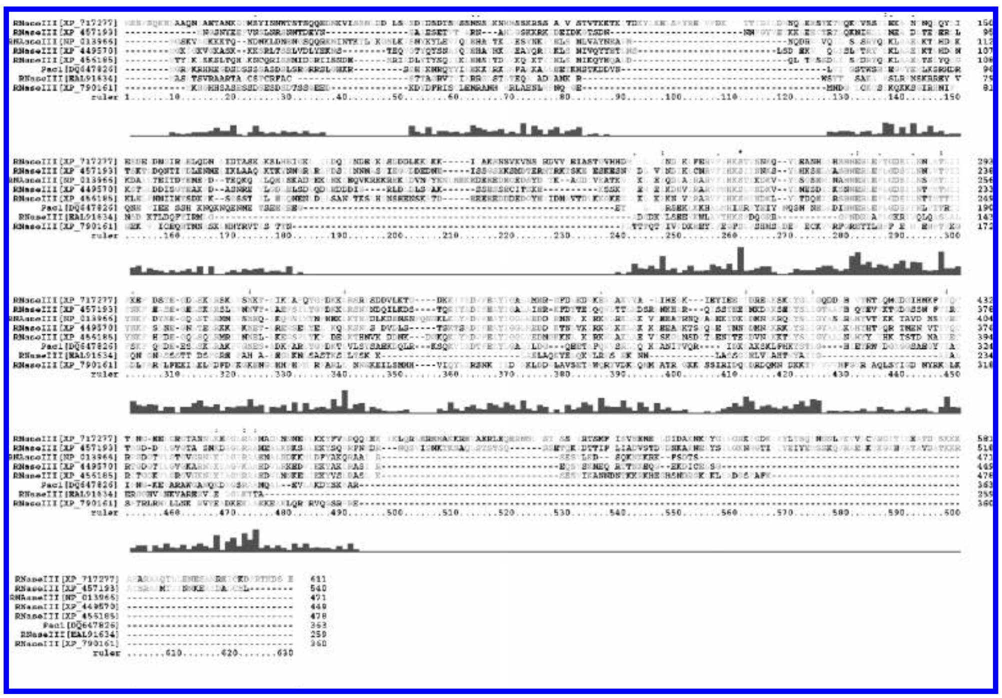

Figure 8. Clustal X sequence alignment involving RNase III-like enzymes. Each sequence is represented by its accession to the GenBank Database Protein. Sequences used in the alignment were represented previously in the Cartesian 2D system **(**Figure 3**)**. We use sequences from some representative eukaryotes and rPac1 from *S. pombe*. [XP_717277] *Candida albicans* SC5314, [XP_457193] *Debaryomyces hansenii* CBS767, [NP_013966] *Saccharomyces cerevisiae*, [XP_449570] *Candida glabrata* CBS138, [XP_456185] *Kluyveromyces lactis*, [DQ647826] *S. pombe* strain 428-4-1, [EAL91634] *Aspergillus fumigatus* Af293, [XP_790161] *Strongylocentrotus purpuratus*.

results obtained. The receiver operating characteristic (ROC) curve was also constructed for the training and validation series. Notably, the curve presented a pronounced curvature (convexity) with respect to the $y = x$ line for both series (see Figure 4). This result confirms that the present model is a significant classifier, having areas of 0.99 (training) and 0.97 (validation)—i.e. markedly higher than 0.5, which is the value for a random classifier.[168]

4.2. Isolation, Prediction, and Assay of a Novel Pac1 from *Schizosaccharomyces pombe* Strain 428-4-1. *4.2.1. Isolation.* In this work we isolated, cloned, and expressed a new Pac1 DNA sequence from *Schizosaccharomyces pombe* strain 428-4-1, and its nucleotide and amino acid sequence was recorded on the GenBank database with accession number ***DQ647826***. The theoretical prediction of its translated ORF as an RNase III-like enzyme was performed by the present alignment-independent approach instead of traditional alignment methods. The theoretical prediction of rPac1 as a double-stranded RNase was confirmed experimentally by in vitro assays.

4.2.2. Prediction. Our Pac1 protein sequence was analyzed using the MMM-QSAR methodology with the aim of recognizing the rPac1 gene product as a eukaryotic RNase III homologue. The sequence was represented in a Cartesian 2D system and calculated including the whole data set. This particular case was included in the validation subset in order to make a prediction. The MMM-QSAR model although very simple (two variables) allowed the correct classification of the rPac1 product as an RNase III-like enzyme with the maximum probability ($p = 1$). In order to make a graphical comparison between our methodology and alignment methods like BLASTp,[169−172] several representative RNase III protein sequences from prokaryotes and eukaryotes were selected together with rPac1 for representation in a 2D-mapping system (see Figure 5).

The 2D-HP map protein representation introduced here revealed a significant separation for the groups consisting of dsRNases from prokaryotes (in dark gray) and eukaryotes (in light gray). The rPac1 protein (in black) is placed between the two groups, acting as a sort of link between the RNase III families. This representation possibly supports evolutionary relationships between double-stranded RNase protein sequences. Since the Cartesian 2D protein representation is mainly based on amino acid composition, we can highlight a major region from rPac1 matching eukaryote sequences (in light gray) and another small region that lies within the prokaryote region (in dark gray). There is also a nonmatching region specific for rPac1 in *Schizosaccharomyces pombe* that does not exist in other eukaryotes. However, matching regions in the graph made a significant contribution to calculation of the spectral moments, thus allowing successful recognition of rPac1 as RNase III. These results coincide with the use of different variants of 2D-HP folding maps for proteins.[173−175]

We also performed an alignment between the previously selected sequences and our rPac1 product using the Clustal W program, version 1.81 (see Figures 7 and 8). Alignment results coincide with those obtained in previous studies reported by other authors. The rPac1 showed low amino acid identity percentages in comparison to dsRNase sequences from other eukaryote organisms, even for those belonging to yeast-related species. Short and less frequent regions match along the protein sequences, especially toward the N-terminal region (see Figure 8). The comparison with prokaryote sequences showed a matching region toward the protein's C-terminal part, from the 170 up to the 260 amino acid position. This region corresponds with the RNase III C-terminal domain (RIBOc), which is conserved in eukaryotic, bacterial, and archeal RNase III and is associated with the catalytic activity. There is a significant N-terminal region in the Pac1 product that does not appear in the RNase III prokaryote family—a finding consistent with other reports (see Figure 7).[29] A BLAST[169−172] analysis was carried out on the translated rPac1 DNA sequence (see Figure 5). This method recognized successfully our query sequence as a Pac1 ribonuclease, reaching up to 98% of amino acid identity with others already recorded from *Schizosaccharomyces pombe* strains. Although this analysis showed lower scores (close to 80%) in comparison to other typical dsRNases, the approach still enabled protein query recognition as RNase III. With the aim of comparing different methods, it is possible to set an equivalence for the score value (80%) from BLAST with our predicted probability, $p = 1$, for rPac1 to act as an RNase III-like enzyme. BLAST also revealed low amino acid identity (<40%) toward the C-terminal portion despite this representing the highest conserved region in the four existing RNase III subclasses.[169−172]

On the other hand, as mentioned previously, each sequence included in the study was submitted to InterProt. All cases (100%) from the RNAse III group matched significantly with RNase III domains (*IPR000999*), allowing the total recognition as dsRNases (see Table ISM). In the case of the control group, six cases did not have InterProt identification, and three of them did not have any hits reported (95, 50% of predictability) (see Table IISM). These results confirm that our model replaces neither classical methods for protein function annotation like BLAST[169−172] and InterProt[122,176] nor new alignment techniques based on partial order, secondary structure, or gene ontology[177−180] but becomes an interesting alternative tool—especially due to its alignment-independence and simplicity. It is also important to highlight that our methodology can be considered as a good classifier, despite its simplicity, as it gives rise to a linear equation with two variables at most.[181] Thus, once the whole database has been screened and proteins having the desired function are recognized, it would be worthwhile to assess results obtained using our approach using other methodologies.

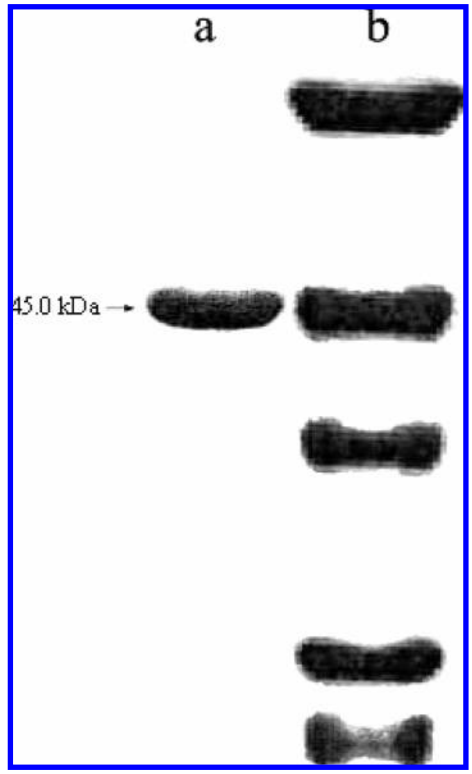

Figure 9. Electrophoresis of rPac1 protein. 45 kDa rPac1 was purified and loaded in 12.5% PAGE-SDS and stained with Coomassie brilliant: (a) band corresponding to rPac1 purified and (b) molecular weight marker—66.2 kDa, 45.7 kDa, 31 kDa, 21.5 kDa, and 14.4 kDa (unstained SDS-PAGE standards broad range, BioRad).

In order to compare the MMM-QSAR approach reported here with other methodologies based on MM, training and negative (non-RNAses sequences) sets were scored with a classic HMMs.[176,182−190] Classification driven by an HMM built on the original training set resulted in an accuracy of 98.75% for the positive sequences (training set) and 96.24% for the negative sequences (see Table 2). Our query sequence ***DQ647826*** was also successfully predicted with the maximum score by the HMM.

4.2.3. Experimental Evidence for RNase III Activity. Recombinant Pac1 protein from *Schizosaccharomyces pombe* strain 428-4-1 was purified in order to measure its double-stranded RNase activity in vitro. The corresponding product size (45.5 kDa) coincided with the reported size for the native protein (see Figure 9). Double-stranded activity was measured in vitro by following the protocol described above.

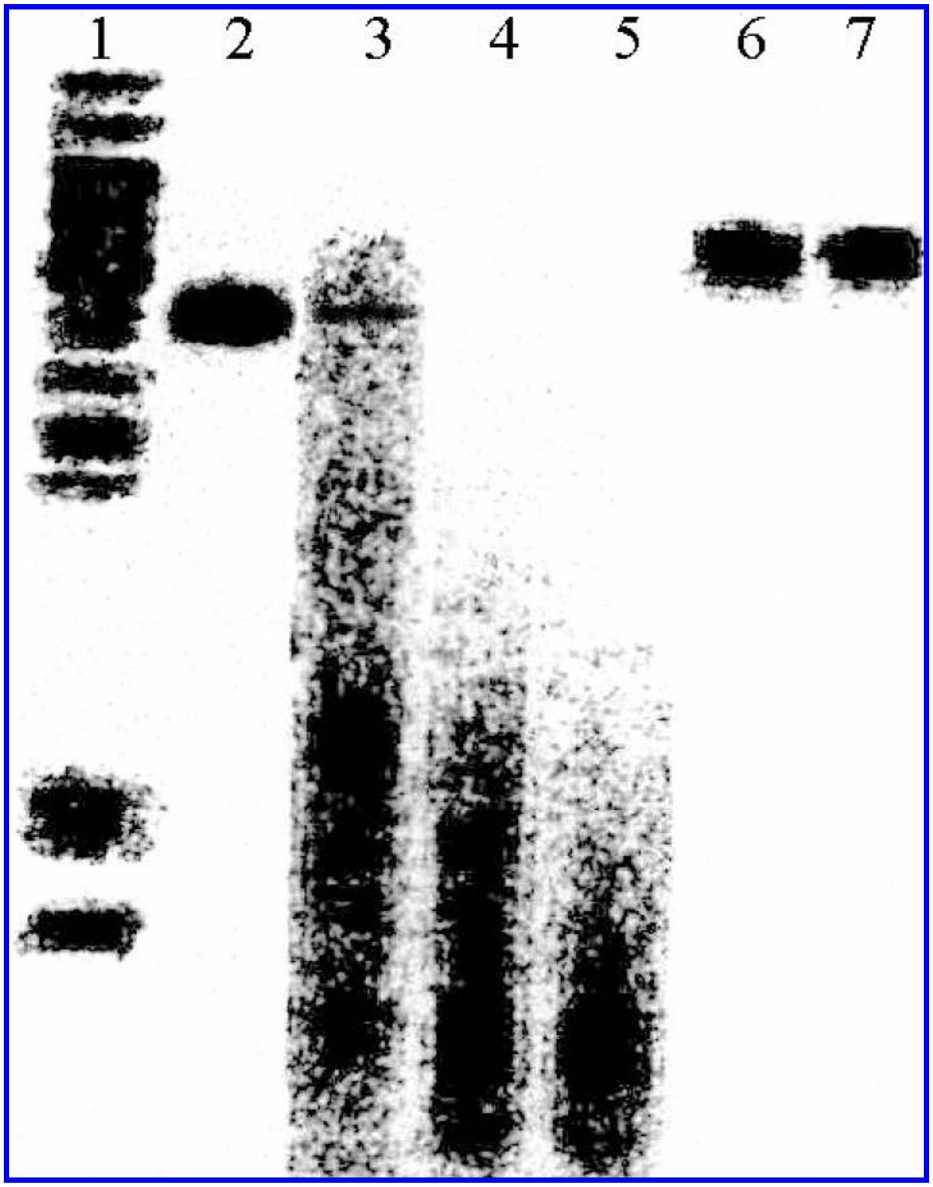

Figure 10. Autoradiography of rPac1 enzymatic assay. To visualize the cleavage activity of the dsRNA substrate generated by T3/T7 "in vitro" transcription, aliquots of enzymatic assay were taken at 2, 5, and 10 min, loaded in 12.5% PAGE/7 M urea followed by autoradiography. Lane 1 is pBR322 digested by Msp1. Lane 2 is the intact dsRNA substrate. Lanes 3−5 are the results of rPac1 enzymatic activity at 2, 5, and 10 min of reaction at 30 °C. Lanes 6 and 7 are the T3 and T7 ssRNA obtained by "in vitro" transcription too which are not degradated by RNAse activity of Pac1.

The unit definition for all RNase III types is the amount of enzyme able to solubilize 1 nmol of acid-precipitable radioactivity *per hour*.[17] Pac1 activity showed values comparable to other results (5×10^5 U/mg) obtained for a recombinant Pac1 product from *Schizosaccharomyces pombe* by Rotondo and Frendewey.[29] Results derived from the enzymatic activity assay are shown in Table 3 for each experiment; the mean value was 6.96×10^5 U/mg.

The kinetic enzymatic reaction of rPac1 by monitoring dsRNA integrity (lanes 2−5) is illustrated in Figure 10. These last experiments for the ***DQ647826*** sequence were not carried out to validate the MMM-QSAR model but to show how to use it for predicting RNase III-like protein function annotation. We recall that the validation of the MMM-QSAR model was assessed with the external prediction series as recommended for any QSAR studies (see previous sections).[191]

5. CONCLUSIONS

The work described here introduces a new approach to predict RNase type III function from protein sequences irrespective of sequence alignment using the MMMs associated with a 2D sequence representation as the input for an LDA classifier. This MMM-QSAR classifier successfully discriminates between RNase-like sequences and a control group. The Pac1 gene product was chosen as a representative example of a sequence with low amino acid identity compared to other enzymes with similar activity. The present methodology achieves high classification scores similar to bioinformatics tools based on sequence alignment (BLASTp) and comparable results to other predicting protein function annotation methods like InterProt and HMMs. The predictions made by the present model coincide with outcomes from experimental isolation, expression, and enzymatic activity measurement of a novel *pac1*$^+$ gene sequence ***DQ647826*** isolated from a new isolate *Schizosaccharomyces pombe* strain 428-4-1. The work promotes the use of the experience accumulated in small-molecules QSAR with spectral moments[192] and other kinds of indices in new types of proteins QSAR studies, now a focus of interest for many researchers worldwide.[193−197]

List of Abbreviations. HP − hydrophobicity and polarity; RNases − ribonucleases; QSAR − quantitative structure−activity relationships; dsRNase − double-strand-specific ribonuclease; snRNAs − small nucleolar RNA; LDA − linear discriminant analysis; ORF − open reading frame; MM − Markov model; HMM − hidden Markov model; ROC curve − receiver operating characteristic curve.

ACKNOWLEDGMENT

The authors are grateful to Associate Editor Dr. Dusanka Janezic and unknown referees for the kind attention and useful revision of the work. The authors are also indebted to Dr. Miriam Klibansky (ICIDCA) for providing the *Schizosaccharomyces pombe* strain. H.G.-D. is grateful for the financial support of the Concellería de Innovación e Industria de Galicia and the Programa Operativo Integrado de Galicia del Fondo Social Europeo (F.S.E.), for a tenure-track research position at the USC (Programa Isidro Parga Pondal), and a 1-year (2007) postdoctoral position at the Dipartimento Farmaco Chimico Tecnologico, Universitá Degli Studi di Cagliari, Italy.

Supporting Information Available: Detailed information on the proteins used in this study including organism, accession number, protein definition, values of the stochastic spectral moments, and scores (Tables ISM and IISM). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Robertson, H. *Escherichia coli* ribonuclease III cleavage sites. *Cell* **1982**, *30*, 669−672.

(2) Chelladurai, B.; Li, H.; Zhang, K.; Nicholson, A. W. Mutational analysis of a ribonuclease III processing signal. *Biochemistry* **1993**, *32*, 7549−7558.

(3) Abeyrathne, P. D.; Nazar, R. N. Parallels in rRNA processing: conserved features in the processing of the internal transcribed spacer 1 in the pre-rRNA from Schizosaccharomyces pombe. *Biochemistry* **2005**, *44*, 16977−87.

(4) Dunn, J. J.; Studier, F. W. T7 early RNAs and Escherichia coli ribosomal RNAs are cut from large precursor RNAs in vivo by ribonuclease III. *Proc. Natl. Acad. Sci. U.S.A.* **1973**, *70*, 3296−3300.

(5) Young, R. A.; Steitz, J. A. Complementary sequences 1700 nucleotides apart from a ribonuclease III cleavege site in Escherichia coli ribosomal precursor RNA. *Proc. Natl. Acad. Sci. U.S.A.* **1978**, *75*, 3593−3597.

(6) Court, D. RNA processing and degradation by RNase III. In *Control of mRNA stability*; Brawerman, G., Belasco, J., Ed.; Academic Press: New York, 1993; p 70−116.

(7) Koraimann, G.; Schroller, C.; Graus, H.; Angerer, D.; Teferle, K.; Hogenauer, G. Expression of gene 19 of the conjugative plasmid R1 is controlled by RNAse III. *Mol. Microbiol.* **1993**, *9*, 717−727.

(8) Gerdes, K.; Nielsen, A.; Thorsted, P.; Wagner, E. G. Mechanism of killer gene activation. Antisense RNA-dependent RNAse III cleavage ensures rapid turn-over of the stable hok, srnB and pndA effector messenger RNAs. *J. Mol. Biol.* **1992**, *226*, 637−649.

(9) Blomberg, P.; Wagner, E. G. H.; Nordström, K. Control of replication of plasmid R1: The duplex between the antisense RNA, CopA, and its target, CopT, is processed specifically in vivo and in vitro by RNAse III. *EMBO J.* **1990**, *9*, 2331−2340.

(10) Gregory, R. I.; Chendrimada, T. P.; Shiekhattar, R. MicroRNA biogenesis: isolation and characterization of the microprocessor complex. *Methods Mol. Biol.* **2006**, *342*, 33−47.

(11) Carmell, M. A.; Hannon, G. J. RNase III enzymes and the initiation of gene silencing. *Nat. Struct. Mol. Biol.* **2004**, *11*, 214−218.

(12) Meister, G.; Landthaler, M.; Peters, L.; Chen, P. Y.; Urlaub, H.; Luhrmann, R.; Tuschl, T. Identification of novel argonaute-associated proteins. *Curr. Biol.* **2005**, *15*, 2149−2155.

(13) Han, J.; Lee, Y.; Yeom, K. H.; Kim, Y. K.; Jin, H.; Kim, V. N. The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev.* **2004**, *18*, 3016−3027.

(14) Yang, W. J.; Yang, D. D.; Na, S.; Sandusky, G. E.; Zhang, Q.; Zhao, G. Dicer is required for embryonic angiogenesis during mouse development. *J. Biol. Chem.* **2005**, *280*, 9330−9335.

(15) Pekarik, V. Design of shRNAs for RNAi-A lesson from pre-miRNA processing: possible clinical applications. *Brain Res. Bull.* **2005**, *68*, 115−120.

(16) Robertson, H. D.; Webster, R. E.; Zinder, N. D. Purification and properties of ribonuclease III from *Escherichia coli. J. Biol. Chem* **1968**, *243*, 82−91.

(17) Dunn, J. J. Ribonulcease III. In *The Enzymes*; Boyer, P. D., Ed.; Academic Press: New York, 1982; Vol. 15 (Part B), pp 485−499.

(18) Nicholson, A. W. Structure, reactivity and biology of double-stranded RNA. *Prog. Nucl. Acid Res. Mol. Biol.* **1996**, *52*, 1−65.

(19) Grummt, I.; Hall, S. H.; Crouch, R. J. Localisation of an endonuclease specific for double-stranded RNA within the nucleolus and its implication in processing ribosomal transcripts. *Eur. J. Biochem.* **1979**, *94*, 437−443.

(20) Ohtsuki, K.; Groner, Y.; Hurwitz, J. Isolation and purification of double-stranded ribonuclease from calf thymus. *J. Biol. Chem.* **1977**, *252*, 483−491.

(21) Lipardi, C.; Baek, H. J.; Wei, Q.; Paterson, B. M. Analysis of short interfering RNA function in RNA interference by using Drosophila embryo extracts and schneider cells. *Methods Enzymol.* **2005**, *392*, 351−371.

(22) Xu, H.-P.; Riggs, M.; Rodgers, L.; Wigler, M. A gene from *S. Pombe* with homology to *E. Coli* RNAse III blocks conjugation and sporulation when overexpressed in wild type cells. *Nucleic Acids Res.* **1990**, *18*, 5304.

(23) Iino, Y.; Sugimoto, A.; Yamamoto, M. *S. pombe pac1*$^+$, whose overexpression inhibits sexual development, encodes a ribonuclease III-like RNAse. *EMBO J.* **1991**, *10*, 221−226.

(24) Qian, Z.; Xuan, B.; Hong, J.; Hao, Z.; Wang, L.; Huang, W. Expression and purification of the carboxyl terminus domain of Schizosaccharomyces pombe dicer in Escherichia coli. *Protein Pept. Lett.* **2005**, *12*, 311−314.

(25) Xu, H. P.; Riggs, M.; Rodgers, L.; Wigler, M. A gene from S. pombe with homology to E. coli RNAse III blocks conjugation and sporulation when overexpressed in wild type cells. *Nucleic Acids Res.* **1990**, *18*, 5304.

(26) Potashkin, J. F. D. A mutation in a single gene of Schizosaccharomyces pombe affects the expression of several snRNAs and causes defects in RNA processing. *EMBO J.* **1990**, *9*, 525−534.

(27) Lamontagne, B.; Elela, S. A. Evaluation of the RNA determinants for bacterial and yeast RNase III binding and cleavage. *J. Biol. Chem.* **2004**, *279*, 2231−2241.

(28) Saunders, L. R.; Barber, G. N. The dsRNA binding protein family: critical roles, diverse cellular functions. *FASEB J.* **2003**, *17*, 961−983.

(29) Rotondo, G.; Frendewey, D. Purification and characterization of the Pac1 ribonuclease of Schizosaccharomyces pombe. *Nucleic Acids Res.* **1996**, *24*, 2377−2386.

(30) Rotondo, G.; Huang, J. Y.; Frendewey, D. Substrate structure requirements of the Pac1 ribonuclease from Schizosaccharmyces pombe. *RNA* **1997**, *3*, 1182−1193.

(31) Dobson, P. D.; Doig, A. J. Predicting enzyme class from protein structure without alignments. *J. Mol. Biol.* **2005**, *345*, 187−199.

(32) Dobson, P. D.; Cai, Y. D.; Stapley, B. J.; Doig, A. J. Prediction of protein function in the absence of significant sequence similarity. *Curr. Med. Chem.* **2004**, *11*, 2135−2142.

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

(34) Lin, H. H.; Han, L. Y.; Zhang, H. L.; Zheng, C. J.; Xie, B.; Chen, Y. Z. Prediction of the functional class of lipid binding proteins from sequence-derived properties irrespective of sequence similarity. *J. Lipid. Res.* **2006**, *47*, 824−831.

(35) Cui, J.; Han, L. Y.; Cai, C. Z.; Zheng, C. J.; Ji, Z. L.; Chen, Y. Z. Prediction of functional class of novel bacterial proteins without the use of sequence similarity by a statistical learning method. *J. Mol. Microbiol. Biotechnol* **2005**, *9*, 86−100.

(36) Han, L. Y.; Cai, C. Z.; Ji, Z. L.; Cao, Z. W.; Cui, J.; Chen, Y. Z. Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. *Nucleic Acids Res.* **2004**, *32*, 6437−6344.

(37) Han, L. Y.; Cai, C. Z.; Ji, Z. L.; Chen, Y. Z. Prediction of functional class of novel viral proteins by a statistical learning method irrespective of sequence similarity. *Virology* **2005**, *331*, 136−143.
(38) Han, L. Y.; Cai, C. Z.; Lo, S. L.; Chung, M. C.; Chen, Y. Z. Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. *RNA* **2004**, *10*, 355−368.
(39) Chen, J.; Liu, H.; Yang, J.; Chou, K. C. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. *Amino Acids* **2007**, *33*, 423−428.
(40) Guo, Y. Z.; Li, M.; Lu, M.; Wen, Z.; Wang, K.; Li, G.; Wu, J. Classifying G protein-coupled receptors and nuclear receptors based on protein power spectrum from fast Fourier transform. *Amino Acids* **2006**, *30*, 397−402.
(41) Chou, K. C.; Elrod, D. W. Prediction of enzyme family classes. *J. Proteome Res.* **2003**, *2*, 183−190.
(42) Elrod, D. W.; Chou, K. C. A study on the correlation of G-protein-coupled receptor types with amino acid composition. *Protein Eng.* **2002**, *15*, 713−715.
(43) Sun, X. D.; Huang, R. B. Prediction of protein structural classes using support vector machines. *Amino Acids* **2006**, *30*, 469−475.
(44) Wen, Z.; Li, M.; Li, Y.; Guo, Y.; Wang, K. Delaunay triangulation with partial least squares projection to latent structures: a model for G-protein coupled receptors classification and fast structure recognition. *Amino Acids* **2006**, *32*, 277−283.
(45) Zhang, S. W.; Pan, Q.; Zhang, H. C.; Shao, Z. C.; Shi, J. Y. Prediction protein homo-oligomer types by pseudo amino acid composition: Approached with an improved feature extraction and naive Bayes feature fusion. *Amino Acids* **2006**, *30*, 461−468.
(46) Chou, K. C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* **2005**, *21*, 10−19.
(47) Chou, K. C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* **2001**, *43* (3), 246−255.
(48) Chou, K. C.; Shen, H. B. Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. *Biochem. Biophys. Res. Commun.* **2006**, *347* (1), 150−157.
(49) Xiao, X.; Shao, S.; Ding, Y.; Huang, Z.; Huang, Y.; Chou, K. C. Using complexity measure factor to predict protein subcellular location. *Amino Acids* **2005**, *28*, 57−61.
(50) Xiao, X.; Shao, S.; Ding, Y.; Huang, Z.; Chen, X.; Chou, K. C. Using cellular automata to generate image representation for biological sequences. *Amino Acids* **2005**, *28*, 29−35.
(51) Gao, Y.; Shao, S.; Xiao, X.; Ding, Y.; Huang, Y.; Huang, Z.; Chou, K. C. Using pseudo amino acid composition to predict protein subcellular location: Approached with Lyapunov index, Bessel function, and Chebyshev filter. *Amino Acids* **2005**, *28*, 373−376.
(52) Wang, S. Q.; Yang, J.; Chou, K. C. Using stacked generalization to predict membrane protein types based on pseudo-amino acid composition. *J. Theor. Biol.* **2006**, *242*, 941−946.
(53) Shen, H. B.; Chou, K. C. Ensemble classifier for protein fold pattern recognition. *Bioinformatics* **2006**, *22*, 1717−1722.
(54) Chou, K. C.; Shen, H. B. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-Nearest Neighbor classifiers. *J. Proteome Res.* **2006**, *5*, 1888−1897.
(55) Chou, K. C.; Shen, H. B. Large-scale predictions of gram-negative bacterial protein subcellular locations. *J. Proteome Res.* **2006**, *5*, 3420−3428.
(56) Shen, H. B.; Chou, K. C. Virus-PLoc: a fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. *Biopolymers* **2007**, *85*, 233−240.
(57) Shen, H. B.; Chou, K. C. Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. *Biochem. Biophys. Res. Commun.* **2007**, *355*, 1006−1011.
(58) Xiao, X.; Shao, S. H.; Huang, Z. D.; Chou, K. C. Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. *J. Comput. Chem.* **2006**, *27*, 478−482.
(59) Xiao, X.; Shao, S.; Ding, Y.; Huang, Z.; Chou, K. C. Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. *Amino Acids* **2005**.
(60) Chou, K. C.; Shen, H. B. MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem. Biophys. Res. Commun.* **2007**, *360*, 339−345.
(61) Chou, K. C.; Shen, H. B. Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *J. Proteome Res.* **2007**, *6*, 1728−1734.
(62) Pu, X.; Guo, J.; Leung, H.; Lin, Y. Prediction of membrane protein types from sequences and position-specific scoring matrices. *J. Theor. Biol.* **2007**, *247*, 259−265.
(63) Zhou, X. B.; Chen, C.; Li, Z. C.; Zou, X. Y. Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. *J. Theor. Biol.* **2007**, *248*, 546−551.
(64) Pan, Y. X.; Zhang, Z. Z.; Guo, Z. M.; Feng, G. Y.; Huang, Z. D.; He, L. Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach. *J. Protein Chem.* **2003**, *22*, 395−402.
(65) Chen, C.; Tian, Y. X.; Zou, X. Y.; Cai, P. X.; Mo, J. Y. Using pseudo-amino acid composition and support vector machine to predict protein structural class. *J. Theor. Biol.* **2006**, *243*, 444−448.
(66) Chen, C.; Zhou, X.; Tian, Y.; Zou, X.; Cai, P. Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. *Anal. Biochem.* **2006**, *357*, 116−121.
(67) Du, P.; Li, Y. Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physico-chemical features of segmented sequence. *BMC Bioinformatics* **2006**, *7*, 518.
(68) Mondal, S.; Bhavna, R.; Babu, R. M.; Ramakumar, S. Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. *J. Theor. Biol.* **2006**, *243*, 252−260.
(69) Chen, Y. L.; Li, Q. Z. Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition. *J. Theor. Biol.* **2007**, *248*, 377−381.
(70) Chen, Y. L.; Li, Q. Z. Prediction of the subcellular location of apoptosis proteins. *J. Theor. Biol.* **2007**, *245*, 775−783.
(71) Lin, H.; Li, Q. Z. Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant. *Biochem. Biophys. Res. Commun.* **2007**, *354*, 548−551.
(72) Lin, H.; Li, Q. Z. Using Pseudo Amino Acid Composition to Predict Protein Structural Class: Approached by Incorporating 400 Dipeptide Components. *J. Comput. Chem.* **2007**, *28*, 1463−1466.
(73) Shi, J. Y.; Zhang, S. W.; Pan, Q.; Cheng, Y.-M.; Xie, J. Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. *Amino Acids* **2007**, *33*, 69−74.
(74) Kurgan, L. A.; Stach, W.; Ruan, J. Novel scales based on hydrophobicity indices for secondary protein structure. *J. Theor. Biol.* **2007**, *248*, 354−366.
(75) Du, Q.; Mezey, P. G.; Chou, K. C. Heuristic molecular lipophilicity potential (HMLP): a 2D-QSAR study to LADH of molecular family pyrazole and derivatives. *J. Comput. Chem.* **2005**, *26*, 461−470.
(76) Du, Q. S.; Huang, R. B.; Wei, Y. T.; Du, L. Q.; Chou, K. C. Multiple field three dimensional quantitative structure-activity relationship (MF-3D-QSAR). *J. Comput. Chem.* **2007**, in press, doi 10.1002/jcc.20776.
(77) Caballero, J.; Fernandez, L.; Abreu, J. I.; Fernandez, M. Amino Acid Sequence Autocorrelation vectors and ensembles of Bayesian-Regularized Genetic Neural Networks for prediction of conformational stability of human lysozyme mutants. *J. Chem. Inf. Model.* **2006**, *46*, 1255−68.
(78) Liao, B.; Ding, K. Graphical approach to analyzing DNA sequences. *J. Comput. Chem.* **2005**, *26*, 1519−1523.
(79) Liao, B.; Wang, T. M. Analysis of similarity/dissimilarity of DNA sequences based on nonoverlapping triplets of nucleotide bases. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1666−1670.
(80) Liao, B.; Wang, T. M. New 2D graphical representation of DNA sequences. *J. Comput. Chem.* **2004**, *25*, 1364−1368.
(81) Liao, B.; Xiang, X.; Zhu, W. Coronavirus phylogeny based on 2D graphical representation of DNA sequence. *J. Comput. Chem.* **2006**, *27*, 1196−1202.
(82) Yu-Hua, Y.; Liao, B.; Tian-Ming, W. A 2D graphical representation of RNA secondary structures and the analysis of similarity/dissmilarity based on it. *J. Mol. Struct.: THEOCHEM* **2005**, *755*, 131−136.
(83) Liao, B.; Wang, T. A 3D Graphical representation of RNA secondary structure. *J. Biomol. Struct. Dyn.* **2004**, *21*, 827−832.
(84) Liao, B.; Ding, K.; Wang, T. On A Six-Dimensional Representation of RNA Secondary Structures. *J. Biomol. Struct. Dyn.* **2005**, *22*, 455−464.
(85) Liao, B.; Wang, T.; Ding, K. On A Seven-Dimensional Representation of RNA Secondary Structures. *Mol. Simul.* **2005**, *31*, 1063−1071.
(86) Liao, B.; Luo, J.; Li, R.; Zhu, W. RNA Secondary structure 2D graphical representation without degeneracy. *Int. J. Quantum Chem.* **2006**, *106*, 1749−1755.
(87) Zhu, W.; Liao, B.; Ding, K. A condensed 3D Graphical representation of RNA secondary structures. *J. Mol. Struct.: THEOCHEM* **2005**, *757*, 193−198.
(88) Song, J.; Tang, H. A new 2-D graphical representation of DNA sequences and their numerical characterization. *J. Biochem. Biophys. Methods* **2005**, *63*, 228−239.
(89) Randic, M.; Zupan, J. Highly compact 2D graphical representation of DNA sequences. *SAR QSAR Environ. Res.* **2004**, *15*, 191−205.
(90) Randic, M.; Balaban, A. T. On a four-dimensional representation of DNA primary sequences. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 532−539.

(91) Liu, Y.; Guo, X.; Xu, J.; Pan, L.; Wang, S. Some notes on 2-D graphical representation of DNA sequence. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 529−33.
(92) Randic, M.; Vracko, M. On the similarity of DNA primary sequences. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 599−606.
(93) Agüero-Chapin, G.; González-Díaz, H.; Molina, R.; Varona-Santos, J.; Uriarte, E.; González-Díaz, 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*, 723−730.
(94) Randič, M.; Vračko, M.; Nandy, A.; Basak, S. C. On 3-D Graphical Representation of DNA Primary Sequences and Their Numerical Characterization. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1235−1244.
(95) Randic, M.; Vracko, M.; Nandy, A.; Basak, S. C. On 3-D graphical representation of DNA primary sequences and their numerical characterization. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1235−44.
(96) Nandy, A. Two-dimensional graphical representation of DNA sequences and intron-exon discrimination in intron-rich sequences. *Comput. Appl. Biosci.* **1996**, *12*, 55−62.
(97) Nandy, A. Recent investigations into global characteristics of long DNA sequences. *Ind. J. Biochem. Biophys.* **1994**, *31*, 149−155.
(98) Nandy, A.; Basak, S. C. Simple numerical descriptor for quantifying effect of toxic substances on DNA sequences. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 915−919.
(99) Zupan, J.; Randic, M. Algorithm for coding DNA sequences into "spectrum-like" and "zigzag" representations. *J. Chem. Inf. Model.* **2005**, *45*, 309−13.
(100) Woodcock, S.; Mornon, J. P.; Henrissat, B. Detection of secondary structure elements in proteins by hydrophobic cluster analysis. *Protein Eng.* **1992**, *5*, 629−35.
(101) Randic, M. 2-D graphical representation of proteins based on virtual genetic code. *SAR QSAR Environ. Res.* **2004**, *15*, 147−157.
(102) González-Díaz, H.; Uriarte, E.; Ramos de Armas, R. Predicting stability of Arc repressor mutants with protein stochastic moments. *Bioorg. Med. Chem.* **2005**, *13*, 323−331.
(103) Vilar, S.; Santana, L.; Uriarte, E. Probabilistic neural network model for the in silico evaluation of anti-HIV activity and mechanism of action. *J. Med. Chem.* **2006**, *49*, 1118−1124.
(104) Vilar, S.; Estrada, E.; Uriarte, E.; Santana, L.; Gutierrez, Y. In silico studies toward the discovery of new anti-HIV nucleoside compounds through the use of TOPS-MODE and 2D/3D connectivity indices. 2. Purine derivatives. *J. Chem. Inf. Model.* **2005**, *45*, 502−514.
(105) Gonzalez, M. P.; del Carmen Teran Moldes, M. A TOPS-MODE approach to predict affinity for A1 adenosine receptors. 2-(Arylamino)adenosine analogues. *Bioorg. Med. Chem.* **2004**, *12*, 2985−2993.
(106) Gonzalez, M. P.; Helguera, A. M.; Cabrera, M. A. Quantitative structure-activity relationship to predict toxicological properties of benzene derivative compounds. *Bioorg. Med. Chem.* **2005**, *13*, 1775−1781.
(107) Gonzalez, M. P.; Moldes del Carmen Teran, M. A TOPS-MODE approach to predict adenosine kinase inhibition. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 3077−3079.
(108) Gonzalez, M. P.; Teran, C.; Teijeira, M. A topological function based on spectral moments for predicting affinity toward A3 adenosine receptors. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 1291−1296.
(109) 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.
(110) Estrada, E. Characterization of the folding degree of proteins. *Bioinformatics* **2002**, *18*, 697−704.
(111) Estrada, E.; Gonzalez, H. What are the limits of applicability for graph theoretic descriptors in QSPR/QSAR? Modeling dipole moments of aromatic compounds with TOPS-MODE descriptors. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 75−84.
(112) Estrada, E.; Uriarte, E.; Vilar, S. Effect of Protein Backbone Folding on the Stability of Protein-Ligand Complexes. *J. Proteome Res.* **2006**, *5*, 105−111.
(113) Du, Q. S.; Wei, Y. T.; Pang, Z. W.; Chou, K. C.; Huang, R. B. Predicting the affinity of epitope-peptides with class I MHC molecule HLA-A*0201: an application of amino acid-based peptide prediction. *Protein Eng. Des. Sel.* **2007**, *20*, 417−423.
(114) Du, Q. S.; Huang, R. B.; Wei, Y. T.; Wang, C. H.; Chou, K. C. Peptide reagent design based on physical and chemical properties of amino acid residues. *J. Comput. Chem.* **2007**, *28*, 2043−2050.
(115) Meneses-Marcel, A.; Marrero-Ponce, Y.; Machado-Tugores, Y.; Montero-Torres, A.; Pereira, D. M.; Escario, J. A.; Nogal-Ruiz, J. J.; Ochoa, C.; Aran, V. J.; Martinez-Fernandez, A. R.; Garcia Sanchez, R. N. A linear discrimination analysis based virtual screening of trichomonacidal lead-like compounds: outcomes of *in silico* studies supported by experimental results. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3838−3843.
(116) Marrero-Ponce, Y.; Castillo-Garit, J. A.; Olazabal, E.; Serrano, H. S.; Morales, A.; Castanedo, N.; Ibarra-Velarde, F.; Huesca-Guillen, A.; Sanchez, A. M.; Torrens, F.; Castro, E. A. Atom, atom-type and total molecular linear indices as a promising approach for bioorganic and medicinal chemistry: theoretical and experimental assessment of a novel method for virtual screening and rational design of new lead anthelmintic. *Bioorg. Med. Chem.* **2005**, *13*, 1005−1020.
(117) Marrero-Ponce, Y.; Diaz, H. G.; Zaldivar, V. R.; Torrens, F.; Castro, E. A. 3D-chiral quadratic indices of the 'molecular pseudograph's atom adjacency matrix' and their application to central chirality codification: classification of ACE inhibitors and prediction of sigma-receptor antagonist activities. *Bioorg. Med. Chem.* **2004**, *12*, 5331−5342.
(118) Garcia-Garcia, A.; Galvez, J.; de Julian-Ortiz, J. V.; Garcia-Domenech, R.; Munoz, C.; Guna, R.; Borras, R. New agents active against Mycobacterium avium complex selected by molecular topology: a virtual screening method. *J. Antimicrob. Chemother.* **2004**, *53*, 65−73.
(119) Gozalbes, R.; Galvez, J.; Garcia-Domenech, R.; Derouin, F. Molecular search of new active drugs against Toxoplasma gondii. *SAR QSAR Environ. Res.* **1999**, *10*, 47−60.
(120) Ponce, Y. M.; González-Díaz, H.; Zaldivar, V. R.; Torrens, F.; Castro, E. A. 3D-chiral quadratic indices of the 'molecular pseudograph's atom adjacency matrix' and their application to central chirality codification: classification of ACE inhibitors and prediction of sigma-receptor antagonist activities. *Bioorg. Med. Chem.* **2004**, *12*, 5331−5342.
(121) Santana, L.; Uriarte, E.; González-Díaz, H.; Zagotto, G.; Soto-Otero, R.; Mendez-Alvarez, E. A QSAR model for in silico screening of MAO-A inhibitors. Prediction, synthesis, and biological assay of novel coumarins. *J. Med. Chem.* **2006**, *49*, 1149−56.
(122) Zdobnov, E. M.; R. A. InterProScan-an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **2001**, *17*, 847−848.
(123) González-Díaz, H.; Vina, D.; Santana, L.; de Clercq, E.; Uriarte, E. Stochastic entropy QSAR for the in silico discovery of anticancer compounds: Prediction, synthesis, and in vitro assay of new purine carbanucleosides. *Bioorg. Med. Chem.* **2006**, *14*, 1095−1107.
(124) González-Díaz, H.; Molina, R. R.; Uriarte, E. Stochastic molecular descriptors for polymers. 1. Modelling the properties of icosahedral viruses with 3D-Markovian negentropies. *Polymer* **2003**, *45*, 3845−3853.
(125) González-Díaz, H.; Aguero, G.; Cabrera, M. A.; Molina, R.; Santana, L.; Uriarte, E.; Delogu, G.; Castanedo, N. Unified Markov thermodynamics based on stochastic forms to classify drugs considering molecular structure, partition system, and biological species: distribution of the antimicrobial G1 on rat tissues. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 551−557.
(126) González-Díaz, H.; de Armas, R. R.; Molina, R. Vibrational Markovian modelling of footprints after the interaction of antibiotics with the packaging region of HIV type 1. *Bull. Math. Biol.* **2003**, *65*, 991−1002.
(127) Dobson, P. D.; Doig, A. J. Distinguishing Enzyme Structures from Non-enzymes Without Alignments. *J. Mol. Biol.* **2003**, *330*, 771−783.
(128) Yuan, Z. Prediction of protein subcellular locations using Markov chain models. *FEBS Lett.* **1999**, *451*, 23−26.
(129) González-Díaz, H.; Cruz-Monteagudo, M.; Vina, D.; Santana, L.; Uriarte, E.; De Clercq, E. 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.* **2005**, *15*, 1651−1657.
(130) Ramos de Armas, R.; González-Díaz, H.; Molina, R.; Perez Gonzalez, M.; Uriarte, E. Stochastic-based descriptors studying peptides biological properties: modeling the bitter tasting threshold of dipeptides. *Bioorg. Med. Chem.* **2004**, *12*, 4815−4822.
(131) González-Díaz, H.; Uriarte, E.; Biopolymer stochastic moments. I. Modeling human rhinovirus cellular recognition with protein surface electrostatic moments. *Biopolymers* **2005**, *77*, 296−303.
(132) Jacchieri, S. G. Mining combinatorial data in protein sequences and structures. *Mol. Diversity* **2000**, *5*, 145−152.
(133) Ramos de Armas, R.; González-Díaz, H.; Molina, R.; Uriarte, E. Markovian Backbone Negentropies: Molecular, descriptors for protein research. I. Predicting protein stability in Arc repressor mutants. *Proteins* **2004**, *56*, 715−723.
(134) González-Díaz, H.; Molina R.; Hernández, I. *MARCH-INSIDE 2.0*; 2005. The present software is a preliminary version; the final version will be free upon request to the corresponding author H. González-Díaz at gonzalezdiazh@yahoo.es or qohumbe@usc.es.
(135) *STATISTICA 6.0*; Statsoft Inc.: 2001.
(136) Mc Farland, J. W.; Gans, D. J. Cluster Significance Analysis. In *Method and Principles in Medicinal Chemistry*; van Waterbeemd, H,. Ed.; VCH: Weinheim, Germany, 1995; Vol. 2, pp 295−307.

(137) Kutner, M. H.; Nachtsheim, C. J.; Neter, J.; Li, W. Standardized Multiple Regression Model. In *Applied Linear Statistical Models*, 5th ed.; McGraw Hill: New York, 2005; pp 271−277.
(138) Rotondo, G.; Gillespie, M.; Frendewey, D. Rescue of the fission yeast snRNA synthesis mutant snm1 by overexpression of the double-strand-specific Pac1 ribonuclease. *Mol. Gen. Genet.* **1995**, *247*, 698−708.
(139) Sambrook, J.; Fritsh, E. F.; Maniatis, T. *Molecular Cloning: A Laboratory Manual*, 2nd ed.; Cold Spring Harbor Laboratory Press: U.S.A., 1989.
(140) Pawar, A. P.; Dubay, K. F.; Zurdo, J.; Chiti, F.; Vendruscolo, M.; Dobson, C. M. Prediction of "aggregation-prone" and "aggregation-susceptible" regions in proteins associated with neurodegenerative diseases. *J. Mol. Biol.* **2005**, *350*, 379−392.
(141) Chiti, F.; Stefani, M.; Taddei, N.; Ramponi, G.; Dobson, C. M. Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature* **2003**, *424*, 805−808.
(142) González-Díaz, 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*, 2932−2937.
(143) Di Francesco, V. M. P. J.; Garnier, J. FORESST: fold recognition from secondary structure predictions of proteins. *Bioinformatics* **1999**, *15*, 131−140.
(144) Chou, K. C. Prediction of signal peptides using scaled window. *Peptides (N. Y.)* **2001**, *22*, 1973−1979.
(145) Chou, K. C.; Shen, H. B. Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. *Biochem. Biophys. Res. Commun.* **2007**, *357*, 633−640.
(146) Chou, K.-C. Prediction and classification of alpha-turn types. *Biopolymers* **1997**, *42*, 837−53.
(147) Chou, K. C. Prediction of beta-turns. *J. Pept. Res.* **1997**, *49*, 120−144.
(148) Chou, K. C. Prediction of tight turns and their types in proteins. *Anal. Biochem.* **2000**, *286*, 1−16.
(149) Chou, K. C. A sequence-coupled vector-projection model for predicting the specificity of GalNAc-transferase. *Protein Sci.* **1995**, *4*, 1365−1383.
(150) Chou, K. C. A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. *J. Biol. Chem.* **1993**, *268*, 16938−16948.
(151) Chou, J. J. A formulation for correlating properties of peptides and its application to predicting human immunodeficiency virus protease-cleavable sites in proteins. *Biopolymers* **1993**, *33*, 1405−1414.
(152) Chou, J. J. Predicting cleavability of peptide sequences by HIV protease via correlation-angle approach. *J. Protein Chem.* **1993**, *12*, 291−302.
(153) Chou, K. C. Prediction of human immunodeficiency virus protease cleavage sites in proteins. *Anal. Biochem.* **1996**, *233*, 1−14.
(154) Chou, K. C.; Wei, D. Q.; Zhong, W. Z. Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS. *Biochem. Biophys. Res. Commun.* **2003**, *308*, 148−151.
(155) Chou, K. C. Structural bioinformatics and its impact to biomedical science. *Curr. Med. Chem.* **2004**, *11*, 2105−2134.
(156) Sirois, S.; Sing, T.; Chou, K. C. HIV-1 gp120 V3 loop for structure-based drug design. *Curr. Protein Pept. Sci.* **2005**, *6*, 413−422.
(157) Chou, K. C.; Wei, D. Q.; Du, Q. S.; Sirois, S.; Zhong, W. Z. Progress in computational approach to drug development against SARS. *Curr. Med. Chem.* **2006**, *13*, 3263−3270.
(158) Du, Q.; Wang, S.; Jiang, Z.; Gao, W.; Li, Y.; Wei, D.; Chou, K. C. Application of bioinformatics in search for cleavable peptides of SARS-CoV M(pro) and chemical modification of octapeptides. *Med. Chem.* **2005**, *1*, 209−213.
(159) Du, Q. S.; Sun, H.; Chou, K. C. Inhibitor design for SARS coronavirus main protease based on "distorted key theory". *Med. Chem.* **2007**, *3*, 1−6.
(160) Wei, D. Q.; Zhang, R.; Du, Q. S.; Gao, W. N.; Li, Y.; Gao, H.; Wang, S. Q.; Zhang, X.; Li, A. X.; Sirois, S.; Chou, K. C. Anti-SARS drug screening by molecular docking. *Amino Acids* **2006**, *31*, 73−80.
(161) Zhang, R.; Wei, D. Q.; Du, Q. S.; Chou, K. C. Molecular modeling studies of peptide drug candidates against SARS. *Med. Chem.* **2006**, *2*, 309−314.
(162) Wang, S. Q.; Du, Q. S.; Zhao, K.; Li, A. X.; Wei, D. Q.; Chou, K. C. Virtual screening for finding natural inhibitor against cathepsin-L for SARS therapy. *Amino Acids* **2007**, *33*, 129−135.
(163) Van Waterbeemd, H. *Chemometric methods in molecular design*; Wiley-VCH: New York, 1995; Vol. 2.
(164) Morales Helguera, A. C. P. M. A.; Pérez González, M.; Molina Ruiz, R.; González-Díaz, H. A topological substructural approach applied to the computational prediction of rodent carcinogenicity. *Bioorg. Med. Chem.* **2005**, *13*, 2477−2488.
(165) Marrero-Ponce, Y.; Medina-Marrero, R.; Castillo-Garit, J. A.; Romero-Zaldivar, V.; Torrens, F.; Castro, E. A. 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*, 3003−3015.
(166) Marrero-Ponce, Y.; Medina-Marrero, R.; Torrens, F.; Martinez, Y.; Romero-Zaldivar, V.; Castro, E. A. Atom, atom-type and total nonstochastic and stochastic quadratic fingerprints: a promising approach for modeling of antibacterial activity. *Bioorg. Med. Chem.* **2005**, *13*, 2881−2899.
(167) González-Díaz, 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*, 1025−1039.
(168) Swets, J. A. Measuring the accuracy of diagnostic systems. *Science* **1988**, *240*, 1285−1293.
(169) Dieterich, G.; Karst, U.; Wehland, J.; Jansch, L. MineBlast: a literature presentation service supporting protein annotation by data mining of BLAST results. *Bioinformatics* **2005**, *21*, 3450−3451.
(170) Balakrishnan, R.; Christie, K. R.; Costanzo, M. C.; Dolinski, K.; Dwight, S. S.; Engel, S. R.; Fisk, D. G.; Hirschman, J. E.; Hong, E. L.; Nash, R.; Oughtred, R.; Skrzypek, M.; Theesfeld, C. L.; Binkley, G.; Dong, Q.; Lane, C.; Sethuraman, A.; Weng, S.; Botstein, D.; Cherry, J. M. Fungal BLAST and Model Organism BLASTP Best Hits: new comparison resources at the Saccharomyces Genome Database (SGD). *Nucleic Acids Res.* **2005**, *33*, D374−377.
(171) Zhou, Y.; Huang, G. M.; Wei, L. UniBLAST: a system to filter, cluster, and display BLAST results and assign unique gene annotation. *Bioinformatics* **2002**, *18*, 1268−1269.
(172) Zhang, J.; Madden, T. L. PowerBLAST: a new network BLAST application for interactive or automated sequence analysis and annotation. *Genome Res.* **1997**, *7*, 649−656.
(173) Berger, B.; Leighton, T. Protein folding in the hydrophobic-hydrophilic (HP) model is NP complete. *J. Comput. Biol.* **1998**, *5*, 27−40.
(174) Gupta, A.; Manuch, J.; Stacho, L. Structure-approximating inverse protein folding problem in the 2D HP model. *J. Comput. Biol*. **2005**, *12*, 1328−1345.
(175) Chen, M.; Huang, W. Q. A branch and bound algorithm for the protein folding problem in the HP lattice model. *Gen. Prot. Bioinform.* **2005**, *3*, 225−230.
(176) Bateman, A.; Haft, D. H. HMM-based databases in InterPro. *Brief Bioinform.* **2002**, *3*, 236−245.
(177) Song, J.; Burrage, K.; Yuan, Z.; Huber, T. Prediction of cis/trans isomerization in proteins using PSI-BLAST profiles and secondary structure information. *BMC Bioinform.* **2006**, *7*, 124.
(178) Zehetner, G. OntoBlast function: From sequence similarities directly to potential functional annotations by ontology terms. *Nucleic Acids Res.* **2003**, *31*, 3799−3803.
(179) Yang, A. S. Structure-dependent sequence alignment for remotely related proteins. *Bioinformatics* **2002**, *18*, 1658−1665.
(180) Lee, C.; Grasso, C.; Sharlow, M. F. Multiple sequence alignment using partial order graphs. *Bioinformatics* **2002**, *18*, 452−464.
(181) Marrero-Ponce, Y.; Medina-Marrero, R.; Castro, A. E.; Ramos de Armas, R.; González-Díaz, H.; Romero-Zaldivar, V.; Torrens, F. Protein Quadratic Indices of the "Macromolecular Pseudograph's α-Carbon Atom Adjacency Matrix". 1. Prediction of Arc Repressor Alanine-mutant's Stability. *Molecules* **2004**, *9*, 1124−1147.
(182) Xu, D.; Liu, H. J.; Wang, Y. F. BSS-HMM3s: an improved HMM method for identifying transcription factor binding sites. *DNA Sequence* **2005**, *16*, 403−411.
(183) Soding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **2005**, *21*, 951−960.
(184) Azad, R. K.; Borodovsky, M. Probabilistic methods of identifying genes in prokaryotic genomes: connections to the HMM theory. *Brief Bioinform.* **2004**, *5*, 118−130.
(185) Chen, Y.; Yu, P.; Luo, J.; Jiang, Y. Secreted protein prediction system combining CJ-SPHMM, TMHMM, and PSORT. *Mamm. Genome* **2003**, *14*, 859−865.
(186) Delorenzi, M.; Speed, T. An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. *Bioinformatics* **2002**, *18*, 617−625.
(187) Gough, J.; Chothia, C. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res.* **2002**, *30*, 268−272.
(188) Bystroff, C.; Thorsson, V.; Baker D. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *J. Mol. Biol.* **2000**, *301*, 173−190.
(189) Krogh, A. Using database matches with for HMMGene for automated gene detection in Drosophila. *Genome Res.* **2000**, *10*, 523−528.

(190) Lukashin, A. V.; Borodovsky, M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* **1998**, *26*, 1107−1115.
(191) Cruz-Monteagudo, M.; González-Díaz, H.; Agüero-Chapin, G.; Santana, L.; Borges, F.; Domínguez, R. E.; Podda, G.; Uriarte, E. Computational Chemistry Development of a Unified Free Energy Markov Model for the Distribution of 1300 Chemicals to 38 Different Environmental or Biological Systems. *J. Comput. Chem.* **2007**, *28*, 1909−1922.
(192) González-Díaz, H.; Pérez-Castillo, Y.; Podda, G.; Uriarte, E. Computational Chemistry Comparison of Stable/Nonstable Protein Mutants Classification Models Based on 3D and Topological Indices. *J. Comput. Chem.* **2007**, *28*, 1990−1995.
(193) Yao, Y. H.; Nan, X. Y.; Wang, T. M. A class of 2D graphical representations of RNA secondary structures and the analysis of similarity based on them. *J. Comput. Chem.* **2005**, *26*, 1339−1346.
(194) González-Díaz, H.; Agüero-Chapin, G.; Varona, J.; Molina, R.; Delogu, G.; Santana, L.; Uriarte, E.; Podda, G. 2D-RNA-Coupling Numbers: A New Computational Chemistry Approach to Link Secondary StructureTopology with Biological Function. *J. Comput. Chem.* **2007**, *28*, 1049−1056.
(195) González-Díaz, H.; Saiz-Urra, L.; Molina, R.; Santana, L.; Uriarte, E. A Model for the Recognition of Protein Kinases Based on the Entropy of 3D van der Waals Interactions. *J. Proteome Res.* **2007**, *6*, 904−908.
(196) González-Díaz, 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*, 1042−1048.
(197) Zhang, X.; Luo, J.; Yang, L. New Invariant of DNA Sequence Based on 3DD-Curves and Its Application on Phylogeny. *J. Comput. Chem.* **2007**, *28*, 2342−2346.

CI7003225