

QSAR for RNases and theoretic–experimental study of molecular diversity on peptide mass fingerprints of a new *Leishmania infantum* protein

Humberto González-Díaz · María A. Dea-Ayuela ·
Lázaro G. Pérez-Montoto · Francisco J. Prado-Prado ·
Guillermín Agüero-Chapín ·
Francisco Bolas-Fernández ·
Roberto I. Vazquez-Padrón · Florencio M. Ubeira

Received: 3 March 2009 / Accepted: 13 June 2009 / Published online: 4 July 2009
© Springer Science+Business Media B.V. 2009

Abstract The toxicity and low success of current treatments for Leishmaniasis determines the search of new peptide drugs and/or molecular targets in *Leishmania* pathogen species (*L. infantum* and *L. major*). For example, Ribonucleases (RNases) are enzymes relevant to several biologic processes; then, theoretical and experimental study of the molecular diversity of Peptide Mass Fingerprints (PMFs) of RNases is

useful for drug design. This study introduces a methodology that combines QSAR models, 2D-Electrophoresis (2D-E), MALDI-TOF Mass Spectroscopy (MS), BLAST alignment, and Molecular Dynamics (MD) to explore PMFs of RNases. We illustrate this approach by investigating for the first time the PMFs of a new protein of *L. infantum*. Here we report and compare new versus old predictive models for RNases based on Topological Indices (TIs) of Markov Pseudo-Folding Lattices. These group of indices called Pseudo-folding Lattice 2D-TIs include: Spectral moments $\pi_k(x,y)$, Mean Electrostatic potentials $\xi_k(x,y)$, and Entropy measures $\theta_k(x,y)$. The accuracy of the models (training/cross-validation) was as follows: $\xi_k(x,y)$ -model (96.0%/91.7%) > $\pi_k(x,y)$ -model (84.7/83.3) > $\theta_k(x,y)$ -model (66.0/66.7). We also carried out a 2D-E analysis of biological samples of *L. infantum* promastigotes focusing on a 2D-E gel spot of one unknown protein with $M < 20, 100$ and $pI < 7$. MASCOT search identified 20 proteins with Mowse score >30, but not one >52 (threshold value), the higher value of 42 was for a probable DNA-directed RNA polymerase. However, we determined experimentally the sequence of more than 140 peptides. We used QSAR models to predict RNase scores for these peptides and BLAST alignment to confirm some results. We also calculated 3D-folding TIs based on MD experiments and compared 2D versus 3D-TIs on molecular phylogenetic analysis of the molecular diversity of these peptides. This combined strategy may be of interest in drug development or target identification.

Electronic supplementary material The online version of this article (doi:10.1007/s11030-009-9178-0) contains supplementary material, which is available to authorized users.

H. González-Díaz (✉) · L. G. Pérez-Montoto · F. J. Prado-Prado ·
F. M. Ubeira
Department of Microbiology and Parasitology, and Department
of Organic Chemistry, Faculty of Pharmacy, USC,
15782 Santiago de Compostela, Spain
e-mail: gonzalezdiaz@yahoo.es

M. A. Dea-Ayuela
Department of Chemistry, Biochemistry and Molecular Biology,
Faculty of Experimental and Health Sciences, University Cardenal
Herrera-CEU, 46113 Moncada, Valencia, Spain

G. Agüero-Chapín
CBQ, Universidad Central de Las Villas (UCLV),
54830 Santa Clara, Cuba

G. Agüero-Chapín
CIMAR, Centro Interdisciplinar de Investigação Marinha e
Ambiental, Universidade do Porto, Rua dos Bragas, 177,
4050-123 Porto, Portugal

F. Bolas-Fernández
Department of Parasitology, Faculty of Pharmacy, Complutense
University, 28040 Madrid, Spain

R. I. Vazquez-Padrón
Department of Surgery Vascular Biology Institute, University of Miami,
Miami, FL 33101, USA

Keywords QSAR · Topological indices · Markov models ·
Protein folding · HP Lattice model · Ribonucleases ·
Leishmania · MALDI-TOF Mass Spectroscopy ·
2D-Electrophoresis · Sequence alignment · Molecular
dynamics

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
MD	Molecular Dynamics
MCM	Markov Chain Model
2DE	2D Electrophoresis
MS	Mass Spectroscopy

Introduction

Ribonucleases (RNases) are enzymes that usually make staggered cuts in both strands of a double helical RNA, although in some cases they cleave once in a single-stranded bulge in the helix. This fact becomes the exploration of the molecular diversity of RNases (or their peptide fragments that retain RNase activity) as an interesting source to search drug or drug-target candidates for drug development. For instance, Kimberly and Rosenberg [1] have recently reviewed and discussed the molecular diversity of the RNase A super-family that includes an extensive network of distinct and divergent gene lineages. 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, ostensibly to promote novel functions. The authors reviewed the literature on the evolution and biology of the RNase A lineages that have been characterized, specifically as involved in host defense including: (1) RNases 2 and RNases 3, also known as the eosinophil ribonucleases, which are rapidly evolving cationic proteins released from eosinophilic leukocytes, (2) RNase 7, an anti-pathogen ribonuclease identified in human skin, and (3) RNase 5, also known as angiogenin, another rapidly evolving RNase known to promote blood vessel growth with recently discovered antibacterial activity. Interestingly, some of the characterized anti-pathogen activities do not depend on RNase activity per se. The authors also discussed the ways in which the anti-pathogen activities characterized *in vitro* might translate into experimental confirmation *in vivo*. Then, they considered the possibility that other RNases, such as the dimeric bovine seminal RNase and the frog oocyte RNase, may have host defense functions and therapeutic value that remain unexplored. This therapeutic value was demonstrated by Onconase an RNase derived from the frog (*Rana pipiens*). However, this is the first and only RNase currently evaluated in clinical trials [2].

Conjugation or fusion of RNases to tumor-specific antibodies is a promising approach to further boost tumor cell killing of these compounds. In addition, Dicer and Drosha are

type III RNases responsible for the generation of short interfering RNAs (siRNAs) from long double-stranded RNAs during RNA interference (RNAi). It involves both RNase proteins in several important biological processes with high biological and molecular diversity. For instance, the function of Dicer on the vascular system regulating the embryonic angiogenesis probably by processing miRNAs, which regulates the expression levels of some critical angiogenic regulators in the cell [3]. The cellular processing of shRNAs shares common features with the biogenesis of naturally occurring miRNA, such as the cleavage by nuclear RNase Drosha, the translocation from the nucleus, processing by a cytoplasmic RNase Dicer, and the incorporation into the RNA-induced silencing complex (RISC). Each step has a crucial influence on the efficiency of RNAi and their consideration should be a part of a standard experimental design. The possible use of RNAi in the treatment of spinocerebellar ataxia or amyotrophic lateral sclerosis, with its advantages and pitfalls and possible extensions to other diseases has been discussed before [4]. More recently, a new RNase with tobacco mosaic virus inhibition was isolated and purified from *Bacillus cereus* ZH14. The inhibitory activity of the RNase in the purification process against tobacco mosaic virus was tested, and the percentage inhibition of the purified RNase (48 U/mL) reached 90% [5]. All the aspects above-mentioned becomes the isolation and prediction of new RNases (or peptides with RNase activity) a goal of the major importance for drug development and/or drug-target prediction.

One possibility to accomplish the study of molecular diversity is the use of proteomics techniques. For instance, some authors often use a combination of 2D-Electrophoresis (2D-E) and Mass Spectroscopy (MS) to isolate and characterize new sequences from biological samples [6]. Obtaining the peptide mass fingerprint (PMF) of a protein is a very useful procedure in this sense [7] and also for clinical purposes [8,9]. In these cases, we employ informatics tools, such as Sequest or MASCOT, to have the MS outcomes for some of the more important peptides of the more similar proteins [10,11]. It means that, for instance, MASCOT may provides a collection of MS signals and the corresponding sequence of peptides presented in known proteins matching with our MS input. In order to rank and select the best protein/peptide candidates, MASCOT uses the Mowse score [12]. If a template protein in the database has a high Mowse score (>52), this protein has a PMF very similar to the PMF of our query proteins, and we can detect a high sequence homology and perform the function annotation. However, there is still another situation that often appears in proteome research and do not coincide exactly with the two situations mentioned previously. We refer to this case, when you identify a new protein, perform the MS analysis of PMF, introduce it in MASCOT (or other MS and sequence database), and the software identify some template

candidates with an important Mowse score that is not sufficiently high to accurately annotate the query protein (>40). A previous study has reported an alternative to Mowse scoring with MASCOT and discussed the limits of accurate scoring [13]. Nevertheless, if this kind of situation persists you have neither the sequence of the query protein nor the sequence of a template protein with high homology but you have the PMFs of both the query and the template. We call this situation here as: the query sequence missing and Low-Mowse scoring case. Independently from the possibility of function annotation of Low-Mowse proteins this kind of PMFs are, in our opinion, ideal sources to fish interesting peptides with bioinformatics and/or data mining computational methods.

Many studies have indicated that computational modeling and various automated prediction methods developed recently [14], such as structural bioinformatics [15,16], molecular docking [17–19], molecular packing [20,21], pharmacophore modelling [22,23], Monte Carlo simulated annealing approach [24], diffusion-controlled reaction simulation [25], identification of membrane proteins and their types [26], identification of enzymes and their functional classes [27], identification of GPCR and their types [28,29], identification of proteases and their types [30,31], protein cleavage site prediction [32–34], and signal peptide prediction [35,36] can timely provide very useful information and insights for both basic research and drug design.

In general, the bioinformatics approaches used to annotate biological functions of nucleic acids and proteins, predict protein secondary structure, and exploring molecular diversity are based on sequence alignment procedures [37–40]. However, it has been noted that such procedures perform poorly in cases of low sequence homology between the query and template sequences deposited in the data base. Alignment techniques are also useless if there is a high query-template homology where we do not know the function of the template sequence deposited in the database [41]. One alternative is the application of alignment-free Machine Learning methods to predict protein functional class and explore molecular diversity based on structural parameters independently of sequence–sequence similarity [42–46]. For instance, the so-called pseudo-amino acid (PseAA) composition or PseAAC indices introduced by Chou to improve the prediction quality for protein subcellular localization and membrane protein type [47], as well as for enzyme functional class irrespective of sequence similarity [48]. The PseAA composition can be used to represent a protein sequence with a discrete model without completely losing its sequence-order information. Ever since the concept of Chou's PseAA composition was introduced, a variety of PseAAC approaches have been stimulated for enhancing the prediction quality of different protein features [30,49–57].

Using graphic approaches to study biological systems can also provide useful insights, as indicated by many previous studies on a series of important biological topics, such as enzyme-catalyzed reactions [58–64], protein folding kinetics [65], inhibition kinetics of processive nucleic acid polymerases, and nucleases [66–68], analysis of codon usage [69,70], and base frequencies in the anti-sense strands [71]. Moreover, graphical methods have been introduced for QSAR study [72–74] as well as utilized to deal with complicated network systems [75,76]. Recently, the “cellular automaton image” [77,78] has also been applied to study hepatitis B viral infections [79], HBV virus gene missense mutation [80], and visual analysis of SARS-CoV [8,9], as well as representing complicated biological sequences [81] and helping to identify protein attributes [29,82,83].

Authors such as Randic, Nandy, Liao, and others have introduced 2D or higher dimension graph representations of sequences prior to the calculation of numerical parameters, sometimes called Topological Indices (TIs). This constitutes an important step in order to uncover useful higher-order information not encoded by 1D sequence parameters [84–97]. Finally, these TIs or other type of parameters may be used as inputs to develop Quantitative Structure–Activity Relationship (QSAR) models in order to predict protein function and explore protein molecular diversity [98–101]. The idea behind this type of QSAR-like approach to protein molecular diversity is essentially the same reported by other authors on low-weight molecules QSAR/QSPR study, e.g., the important works of Roy et al. [101–108]. In fact, QSAR is one of the more important tools to explore molecular diversity nowadays [109–119].

In particular, for the case of proteins, the idea of describing them as networks is very interesting and has important advantages over computationally expensive methods (see, for instance, the interesting studies of Krishnan, Zibilut, and Giuliani et al. [120–125]). Specifically, different computational schemes have used charge and Hydrophobicity patterning along sequence to predict folding and mechanism and aggregation of proteins, Zibilut, and Giuliani et al. in proteome research [126]. Recently, our group have introduced Hydrophobicity–Polarity (HP) 2D Cartesian or lattice-like network representations for proteins [127]. We can use Markov Chains theory in order to calculate TIs of these lattices, which allow us to numerically encode higher-order sequence information. The method consists of the following steps, which can be applied to many different problems and have been revised in recent reviews [98,99,128]. First, we derived the Lattice-like representations (also called maps or graphs) for protein sequences. Next, we calculated the TIs values to characterize the protein sequence. Finally, we use these pseudo-folding TIs as inputs for QSAR or Clustering algorithms [95].

On the other hand, Molecular Dynamics (MD) of peptides and proteins is central for drug and target discovery. Since, the pioneering article entitled “The Biological Functions of Low-Frequency Phonons” [129] was published in 1977, a series of investigations into biomacromolecules by means of dynamic avenues have been stimulated. It has been suggested through these studies that low-frequency (or terahertz frequency) collective motions do exist in proteins and DNA that hold a very high potential to reveal the profound dynamic mechanisms of many marvelous biological functions in biological systems (see, e.g., [130–143] and a comprehensive review [144]). Such inferences have been later observed by NMR [145], and applied in medical treatments [146, 147]. In view of this, to really understand the action mechanism of drugs with their receptors, we should consider not only their static structures but also their dynamical processes by simulating their interactions through a dynamic process. Thus, MD has become the foremost computational technique to investigate structure and function of peptides [148–153]. Consequently, we can use the 3D folded structures of the peptides obtained by MD to calculate 3D-TIs instead of pseudo-folding 2D-TIs.

The present study is aimed to develop a powerful computational approach for studying Peptide Mass Fingerprints of Ribonucleases by combining QSAR models, 2D-Electrophoresis (2D-E), MALDI-TOF Mass Spectroscopy (MS), BLAST alignment, and Molecular Dynamics (MD) in hopes that it may become a useful tool for drug development. We report two different experiments in order to introduce new Sequence and MD pseudo-folding TIs for the study of molecular diversity of PMFs. We also report new QSAR and Clustering analysis models based on these indices. In the first experiment (*Experiment 1*), we show the use in an experimental example to use 2D-Lattice electrostatic parameters to numerically characterize protein sequences and seek a model to predict RNase III function without relying on alignment. Different classes of 2D graphs representations of DNA, RNA, protein sequence, or proteomic maps have been used by other researchers [87, 91, 92, 154–164]. We subsequently developed three different classifiers (one for each type of TIs) to connect protein sequence information (represented by TIs 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 [165, 166]. We selected a Linear Discriminant Analysis (LDA), which is a simple but powerful technique [167]. In the other experiment (*Experiment 2*), we compared phylogenetic analysis of Peptides based on both folding 3D-TIs and pseudo-folding 2D-TIs. In both experiments, we illustrate the use of the new models in a practical example based on the analysis of the PMF of a new protein. As a result of this work we could characterize the PMF of the new protein and introduced at the same time new QSAR and Phyloge-

netic algorithms of general use for other peptides or proteins.

Materials and methods

2D-TIs of pseudo-folding lattices

The MARCH-INSIDE approach is used to calculate the Pseudo-Folding TIs of sequences. First, each aminoacid in the sequence is placed in a Cartesian 2D space $\mathbf{r}_2 = (x, y)$ starting with the first monomer at the (0, 0) coordinates. The coordinates of the successive aminoacids are calculated as follows: in a similar manner, then it can be used for a DNA [127]:

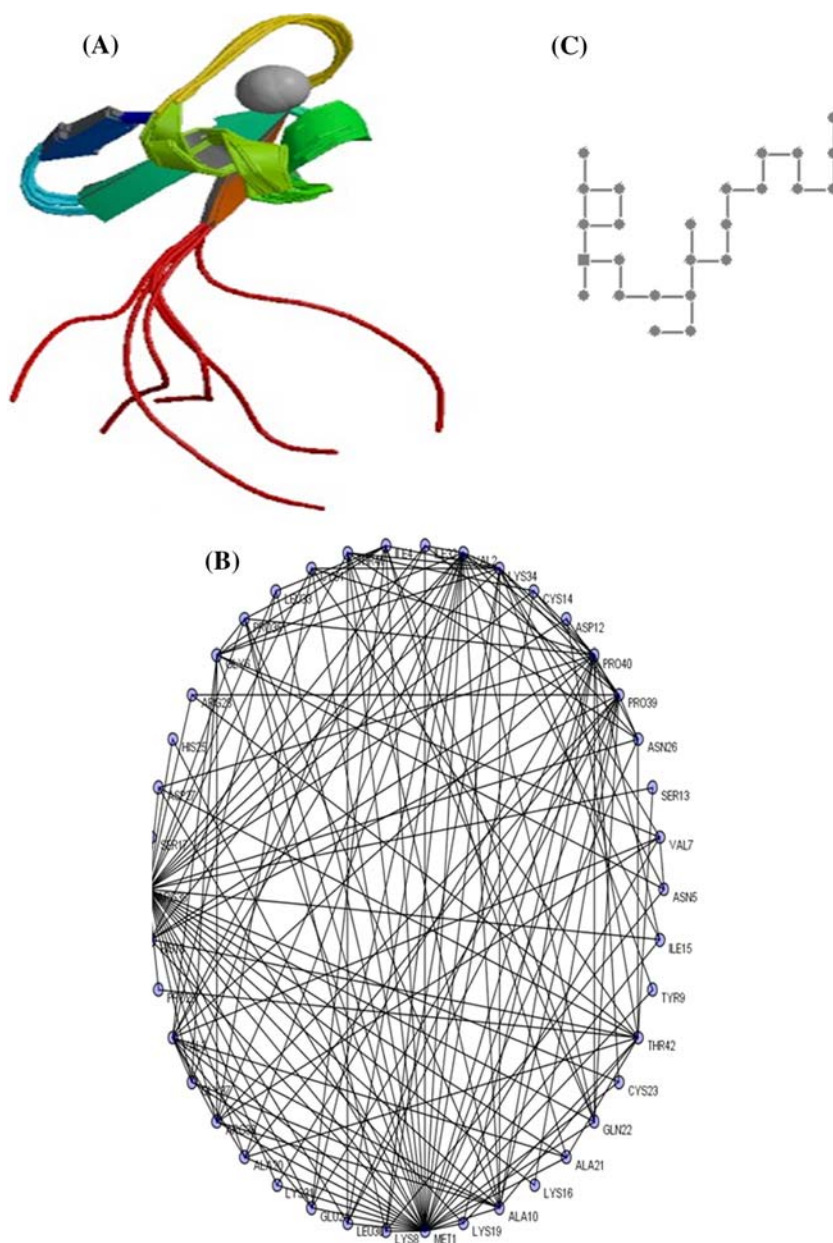
- Increases in +1 the x axe; coordinate for an acid aminoacid (rightwards-step),
- Decreases in -1 the x axe; coordinate for a basic aminoacid (leftwards-step),
- Increases in +1 the y axe; coordinate for a polar aminoacid (upwards-step), and
- Decreases in -1 the y axe; coordinate for a non-polar aminoacid (downwards-step).

Second, the method uses the Markov matrix ${}^1\Pi$, which is a squared matrix to characterize electrostatic interactions between aminoacids in the folded protein. Note that the number of nodes (n) in the graph may be equal or even smaller than the number of aminoacids. The matrix ${}^1\Pi$ contains the probabilities ${}^1p_{ij}(\mathbf{r}_2)$ of direct electrostatic interaction between two nodes placed at distance $y_k = 1$ within the lattice in \mathbf{r}_2 . The formula for ${}^1p_{ij}(\mathbf{r}_2)$ values is the following:

$$p_{ij}(\mathbf{r}_2) = \frac{\left(\frac{Q_j}{d_{ij}(\mathbf{r}_2)}\right)}{\sum_{m=1}^n \alpha_{il} \cdot \left(\frac{Q_l}{d_{il}(\mathbf{r}_2)}\right)}, \quad (1)$$

where Q_j is the charge of the node n_j (coincide with the sum of the charge for all aminoacids projected over the node), d_{ij} is the Euclidean distance between the nodes i and j , and α_{ij} equals to 1, if the nodes n_i and n_j are adjacent in the graph and equals to 0 otherwise. The charge of the node is equals to the sum of the charges of all aminoacids placed at this node. Afterward, we can calculate sequence pseudo-folding TIs in the form of different invariants of this matrix. In this study, we consider three different classes of pseudo-folding Electrostatic TIs: spectral moments $\pi_k(x, y)$, entropy values $\theta_k(x, y)$, and average electrostatic potentials $\xi_k(x, y)$. Using the Markov chain theory, we can calculate the values of

Fig. 1 3D model, 2D pseudo-folding lattice, and 3D structure network, for protein 1CO4



these parameters for all nodes placed a topological distance $k > 1$:

$$\pi_k(\mathbf{r}_2) = \sum_{i=j}^n k p_{ij}(\mathbf{r}_2) = Tr \left[\left({}^1\Pi \right)^k \right] \tag{2}$$

$$\theta_k(\mathbf{r}_2) = -k \cdot \sum_{j=1}^n \binom{k}{j} p_j(\mathbf{r}_2) \cdot \log \binom{k}{j} p_j(\mathbf{r}_2) \tag{3}$$

$$\xi_k(\mathbf{r}_2) = \sum_{j=1}^n k p_j(\mathbf{r}_2) \cdot Q_j, \tag{4}$$

where Tr is called the trace and points to the sum of all the values in the main diagonal of the matrices ${}^k\Pi = \left({}^1\Pi \right)^k$, calculated as natural powers of ${}^1\Pi$. The present 2D-TIs encode in a stochastic manner the interactions of charged nodes (one or more amino acids) placed at different distances not in the sequence (1D space), but in the 2D lattice embedded in \mathbf{r}_2 . Note that in Eqs. 3 and 4, we used absolute probabilities ${}^k p_j(\mathbf{r}_2)$ of interaction for a node with any other node placed at distance k instead of using directly the interaction probabilities ${}^k p_{ij}(\mathbf{r}_2)$. In protein QSAR, this kind of pseudo-folding lattices in $\mathbf{r}_2 = (x, y)$ may become an alternative, in terms of computational cost, to real folded structures in $\mathbf{r}_3 = (x, y, z)$. Figure 1 depicts both the pseudo-folding lattice network for

a protein in \mathbf{r}_2 and the aminoacid–aminoacid contact map network for the same protein in \mathbf{r}_3 . The calculation of the $k_{p_j}(\mathbf{r}_2)$ values has already been explained in detail in the literature, therefore, we do not cover this here [127, 168]. This theoretical description contains the essential elements to understand the work and the reader may also consult recent reviews that explain in detail the theory and applications of the MARCH-INSIDE approach [98, 99, 128].

Protein QSAR analysis

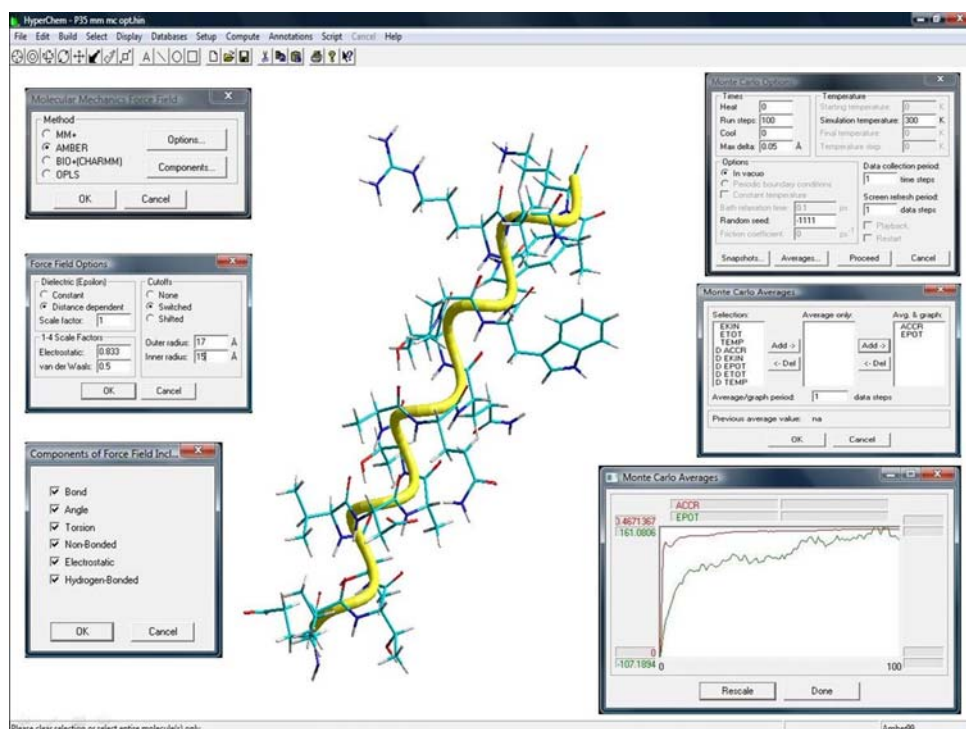
Linear Discriminant Analysis (LDA) was used to construct the QSAR classifier. LDA forward stepwise analysis was carried out for variable selection to build up the model [167]. All of the variables included in the model were standardized in order to bring them onto the same scale. Subsequently, a standardized linear discriminant equation that allows comparison of their coefficients was obtained [169]. The square of Canonical regression coefficient (Rc) and Wilk's statistics (U) were examined in order to assess the discriminatory power of the model ($U = 0$ perfect discrimination, being $0 < U < 1$), and the separation of the two group of proteins was statistically verified by the Fisher ratio (F) test with error level $p < 0.05$.

MD study of PMFs of the new protein

The Molecular Dynamics Trajectories (MDTs) or energetic profiles of all the starting structure of peptides were also

obtained by means of the Monte Carlo (MC) method, using the HyperChem package [170, 171]. In this sense, the AMBER94 force field [172] was used with distant-dependent dielectric constant (scale factor 1), electrostatic and Van der Waals values by default and cutoffs shifted with outer radius of 14 Å (see Fig. 2). All the components of the force field were included and the atom type was recalculated keeping their current charges. Previous to MC simulation, the geometry of all the structures of peptides were optimized with this same force field. Finally, the simulation was executed in vacuo at 300 K and 100 optimization steps obtaining MDTs with 100 potential energy dE_j ($j = 1, 2, 3, \dots, 100$) values each one. We obtained 22 MDTs for 19 peptides. In order to obtain realistic MDTs, there is an additional parameter that we monitor in MD algorithms, which is known as the acceptance ratio (ACCR). It appears as ACCR on the list of possible selections in the MC Averages dialog box of HyperChem (see Fig. 2). The ACCR is a running average of the ratio of the number of accepted moves to attempted moves. Varying the step size can produce a large effect on the ACCR value. The step size ($\Delta\mathbf{r}_3$) is the maximum allowed atomic displacement used in the generation of trial configurations. The default value of \mathbf{r}_3 in HyperChem is 0.05 Å [170]. For most organic molecules, this will result in ACCR of about 0.5 Å, which means that about 50% of all moves are accepted. Increasing the size of the trial displacements may lead to more complete searching of configuration space, but the acceptance ratio will, in general, decrease. Smaller displacements generally lead to higher acceptance ratios but result in more limited sampling.

Fig. 2 Hyperchem interface showing MD study of a peptide



There has been little research to date on what the optimum value of the acceptance ratio should be.

3D-TIs of structures folding determined with MD

The method may also use the Markov matrix ${}^1\Pi$, which is a squared matrix to characterize electrostatic interactions between aminoacids in the folded 3D structure of the peptide obtained by MD. The matrix ${}^1\Pi$ contains the probabilities ${}^1p_{ij}$ of direct electrostatic interaction between two nodes placed at distance lower than cut-off within the 3D space of coordinates $\mathbf{r}_3 = (x,y,z)$:

$$p_{ij}(\mathbf{r}_3) = \frac{\left(\frac{Q_j}{d_{ij}(\mathbf{r}_3)}\right)}{\sum_{m=l}^n \alpha_{il} \cdot \left(\frac{Q_l}{d_{il}(\mathbf{r}_3)}\right)}, \quad (5)$$

where Q_j is the charge of the node n_j (coincide with the sum of the charge for all aminoacids projected over the node), d_{ij} is the Euclidean distance between the nodes i and j , and α_{ij} equals to 1, if the nodes n_i and n_j are adjacent in the graph and equals to 0 otherwise. Afterward, we can calculate sequence pseudo-folding TIs in the form of different invariants of this matrix. In this study, we consider three different classes of real folding 3D-TIs: spectral moments $\pi_k(\mathbf{r}_3)$, entropy values $\theta_k(\mathbf{r}_3)$, and average electrostatic potentials $\xi_k(\mathbf{r}_3)$. Using the Markov chain theory, we can calculate the values of these parameters for all nodes placed a topological distance $k > 1$:

$$\pi_k(\mathbf{r}_3) = \sum_{i=j}^n {}^k p_{ij}(\mathbf{r}_3) = Tr \left[\left({}^1\Pi \right)^k \right] \quad (6)$$

$$\theta_k(\mathbf{r}_3) = -k \cdot \sum_{j=1}^n \left({}^k p_j(\mathbf{r}_3) \right) \cdot \log \left({}^k p_j(\mathbf{r}_3) \right) \quad (7)$$

$$\xi_k(\mathbf{r}_3) = \sum_{j=1}^n {}^k p_j(\mathbf{r}_3) \cdot Q_j(\mathbf{r}_3). \quad (8)$$

2D versus 3D-TIS phylogenetic analysis of PMFs

In principle, we can use different distance functions, here, we select only the Euclidean distance due to the Euclidean nature of the Cartesian of both the space used to derive the pseudo-folding lattices \mathbf{r}_2 and the real folding space \mathbf{r}_3 . Using the Tree Joining Cluster (TJC) analysis, algorithm implemented on the software Statistica, we were able to construct, visualize, and compare the phylogentic trees based on both 2D and 3D-TIs. The molecules used in this study were the same 19 peptides found on the PMF of the new protein. In general, in the phylogentic analysis, we can calculate here $(3 \text{ type of indices}) \times (2 \text{ type of graphs}) = 6$ different Euclidean distances. In order to give a general notation for all these

equations, we use the symbol ${}^p TI_k(\mathbf{r}_d)$, which take the values $TI = \theta, \xi, \text{ or } \pi$ and the dimension of the space $d = 2$ for $\mathbf{r}_2 = (x,y)$ or $d = 3$ for $\mathbf{r}_3 = (x,y,z)$. The equation that describes the formula may used to calculate the nine types of Euclidean distances, mentioned above or alternatively, we can group all the TIs of the same \mathbf{r}_d :

$${}^{TI} D_{pq}(\mathbf{r}_d) = \sqrt{\sum_{k=0}^5 ({}^p TI_k(\mathbf{r}_d) - {}^q TI_k(\mathbf{r}_d))^2}. \quad (9)$$

Experimental methods

Cell culture of parasites

Promastigotes of the Leishmania strain LEM75 were grown in Schneider medium supplemented to a final concentration of 0.4 g/L NaHCO₃, 4 g/L HEPES, 100 mg/L penicillin and streptomycin, and 10% fetal bovine serum (Gibco), pH 6.8 and 26 °C.

Sample preparation

Mid-log promastigotes were recovered on day 7 post-inoculum (p.i.) and the parasites were centrifuged at 3,000 rpm for 10 min at 4 °C. The resulting pellet was washed five times with Tris-HCl pH 7.8, and resuspended in 0.1 mL of this same buffer. The sample was sonicated for 10 s with a virsonic 5 (virTis, NY, USA) set at 70% output power on ice bath. The homogenate was extracted in 5 mM Tris-HCl buffer pH 7.8 containing 1 mM phenylmethylsulfonyl fluoride (PMSF) as a protease inhibitor, at 4 °C overnight and, subsequently, centrifuged at 10,000g for 1 h at 4 °C (Biofuge 17RS: Heraeus Sepatech, Gmb, Osterode, Denmark). The supernatant was dialyzed overnight at 4 °C in 0.5 mM Tris-HCl buffer. Proteins were precipitated by 20% TCA (trichloroacetic acid) in acetone with 20 mM DTT for 1 h at −20 °C, added 1:1 to the homogenate. Then, the sample was centrifuged at 10,000 rpm for 15 min and the pellet was washed with cold acetone containing 20 mM DTT. Residual acetone was removed by air drying. In order to achieve a well-focused first-dimension separation, sample proteins must be completely disaggregated and fully solubilized, in a sample buffer containing 7 M urea, 2 M thiourea, 4% CHAPS, DeStreak buffer (Amersham Bioscience), 5 mM Co₃K₂, 2% IPG buffer (Amersham Bioscience), and incubated at room temperature for 30 min. Following clarification by centrifugation at room temperature (12,000 rpm, 10 min) the supernatant were stored frozen.

2D-Electrophoresis (2D-E)

In total 340 μL of rehydration buffer were added to promastigotes solubilized extracts (7 M urea, 2 M thiourea, 2% CHAPS, 0.75% IPG buffer 4–7, and bromophenol blue) and immediately were adsorbed onto 18 cm immobilized pH 4–7 gradient (IPG) strips (Amersham Biosciences) [173]. Optimal IEF was carried out at 20 °C, with an active rehydration step of 12 h (50 V), and then focused on an IPGphor IEF unit (Amersham Biosciences) by using the following program: 150 V for 2 h, 500 V for 1 h, 1,000 V for 1 h, 1,000–2,000 V for 1 h, and 8,000 V for 6 h. After focusing, IPG strips were equilibrated for 15 min in 10 mL of 50 mM Tris-HCl, pH 8.8, 6 M urea, 30% v/v glycerol, 2% w/v SDS, traces of bromophenol blue containing 100 mg of DTT, and further incubated for 25 min in the same buffer replacing DTT by 300 mg of iodoacetamide. After equilibration, the IPG strips were placed onto 12.5% SDS-polyacrylamide gels and sealed with 0.5% (w/v) agarose. SDS-PAGE was run at 15 mA/gel. The 2D gels were stained with silver staining mass spectrometry compatible. Briefly, the gels were fixed in 40% ethanol (v/v), 10% (v/v) acetic acid overnight, then sensitized with sodium acetate 0.68 % (w/v) and 0.05% sodium thiosulfate for 30 min, and washed with desionized water thrice for 5 min. The gels were incubated in 0.25% (w/v) silver nitrate for 30 min. After incubation, it was rinsed with desionized water twice for 50 s followed by adding the developing solution, which contained 2.5 % (w/v) sodium carbonate with 0.04% (v/v) formaldehyde until intensity desired. Development was terminated by adding 1.5 % (w/v) EDTA.

MALDI-TOF Mass Spectrometry (MS)

Spots of interest were manually excised from silver-stained 2D-E gels after being destained as described by Gharahdaghi et al. [174]. Then, gel pieces were incubated with 12.5 ng/ μL sequencing grade trypsin (Roche Molecular Biochemicals) in 25 mM AMBIC overnight at 4 °C. After digestion, the supernatants (crude extracts) were separated. Peptides were extracted from the gel pieces first into 50% ACN, 1% trifluoroacetic acid and then into 100% ACN. Then, 1 μL of each sample and 0.4 μL of 3 mg/mL α -cyano-4-hydroxycinnamic acid matrix (Sigma) in 50% ACN, 0.01% trifluoroacetic acid were spotted onto a MALDI target. MALDI-TOF MS analyzes were performed on a Voyager-DE STR mass spectrometer (PerSeptive Biosystems, Framingham, MA, USA). The following parameters were used: cysteine as S-carbamidomethyl derivative and methionine in oxidized form. Spectra were acquired over the m/z range of 700–4500 Da. Tryptic, monoisotopic peptide mass lists were generated and exploited for database searching. MS/MS sequencing analysis were carried out using the MALDI-tandem time-of-flight mass spectrometer 4700 Proteomics Analyzer (Applied Bio-

systems, Framingham, MA). The MS study was performed at the University Complutense de Madrid Proteome Facility platform.

MASCOT database search

The peptide mass fingerprinting data obtained from MALDI-TOF analyses were used to search for protein candidates using MASCOT software program [10]. The MASCOT search parameters were adjusted according to the MS experiment carried out and the above description as follows: Type of search: Sequence Query, Enzyme: Trypsin, Fixed modifications: Carbamidomethyl (C), Variable modifications: Oxidation (M), Mass values: MONOISOTOPIC, Protein Mass: Unrestricted, Peptide Mass Tolerance: ± 100 ppm, Fragment Mass Tolerance: ± 0.4 Da, Max Missed Cleavages: 1, and Instrument type: MALDI-TOF-TOF. We introduced the MS signals correspondent to one of the unidentified 2D-E spots (protein) into the MASCOT analysis system. The sample was recorded in this web page with the search title: Sample Set ID: 1122, Analysis ID: 1466, MALDI Well ID: 17500, Spectrum ID: 7971, and Path = \040519\Leishmania\New Analysis 2. The database used was Leishmania 290703 (with 7,467 sequences and 4,469,604 residues).

BLAST search

The more relevant peptide fragments of the new protein were submitted to BLASTP to show graphically the similarity of the sequence compared to other RNases [175]. The BLAST procedure was carried out using as query database the non-redundant NCI database and allowing BLAST to search for conserved domains through the CD-search tool [176].

Results and discussion

Experiment 1

Pseudo-folding 2D-TIs QSAR models for RNases

The search for tools to explore molecular diversity that complement or improve classical alignment tools like BLAST with information from gene ontology, RNA secondary structure prediction, partial ordering, or other sources constitutes a goal of major importance [177–180]. In particular, different structural parameters have been used to mining the molecular diversity of peptides. For instance, Jacchieri have investigated structural propensities, co-localization of peptide fragments in protein sequences, interactions between peptide fragments in close structural proximity and the participation of physical chemical profiles in the distribution of structural motifs among peptide fragments in the Protein

Data Bank (PDB) and the SwissProt databases [181]. In this study, we calculated three families of TIs that can be used as inputs for the QSAR study of the molecular diversity of RNase proteins and peptides. We selected TIs instead of other indices due to their fast calculation and high accuracy demonstrated in QSAR studies of molecular diversity [116, 182–185]. This calculation was carried out for two groups of protein sequences, one made up of RNase-like enzymes and the other formed by heterogeneous proteins. A simple LDA was developed to classify a novel sequence as RNase or not using as inputs the above-mentioned parameters. The best equation found was:

$$S(\xi) = 20.15 \times \xi_1(\mathbf{r}_2) - 15.8 \times \xi_2(\mathbf{r}_2) - 112.3 \quad (10)$$

$$R = 0.87, \quad U = 0.24, \quad F = 231.9, \quad p < 0.001.$$

The statistical parameters for the above equation were: Canonical Regression Coefficient (R), Wilk's statistic (U), Fisher ratio (F), and error level (p -level), which have to be <0.05 [186]. In this equation, as well as in the two other QSAR (see below) the variable $S(\text{TI})=S(\xi)$, $S(\pi)$, or $S(\theta)$ are the outputs of the models. These are real valued scores assigned by the model to the propensity with which a given protein is predicted as RNase. This discriminant function presented excellent results both in training and external cross-validation series carried out with an external set made up of RNase proteins and diverse no-RNase proteins not used to train the model (see Table 1). In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling test, and jackknife test [187]. However, as elucidated by [188] and demonstrated in [189], among the three cross-validation methods, the jackknife test is deemed the most objective that can always yield a unique result for a given benchmark data set, and hence has been increasingly used by investigators to examine the accuracy of various prediction models (see, e.g., [30, 49–52, 190, 191]). In the current study, for reducing computational time as done by many other investigators, we used independent data set test for cross-validation. Its results are remarkable in comparison to results obtained by other researchers on using the LDA method in QSAR studies [192].

In order to compare the previous model with other methodologies based on MM, we developed two additional MARCH-INSIDE models. These models were based on spectral moments and entropy invariants. The equations of these models and their more important statistic parameters are depicted bellow:

$$S(\pi) = 0.59 \times \pi_0(\mathbf{r}_2) - 1.99 \times \pi_2(\mathbf{r}_2) - 21.58 \quad (11)$$

$$R = 0.66, \quad U = 0.56, \quad F = 56.6, \quad p < 0.001,$$

$$S(\theta) = 8.29 \times \theta_0(\mathbf{r}_2) - 16.73 \quad (12)$$

$$R = 0.26, \quad U = 0.93, \quad F = 10.5, \quad p = 0.002.$$

Both equations perform a statistically significant separation of two groups of proteins ($p < 0.05$). The equation based on π_k is essentially the same model that was previously reported by our group but, we incorporate it here in order to perform a comparative study [193]. However, the accuracy of the models is notably lower than the accuracy of model 1 (10). Note that the values of Canonical Regression coefficients are R model 1 > R model 2 (11) > R model 3 (12) and, correspondingly, the inverse tendency is observed for the Wilk's statistics of group separation (U model 1 < U model 2 < U model 3). Detailed information on the classification performance of these models was reported in Table 1. From these results, we can expect that the models based on different families of indices will present different accuracy in predictions. In this case, we should select the ξ -model represented by Eq. 10 as the better option with respect to the π -model and the θ -model. These results are consistent with those obtained in our previous reports, in which we used 2D pseudo-folding electrostatic parameters as sequence descriptors for function annotation of other classes of proteins [127].

2-DE isolation of a novel sequence

In this section, we present a comparative study of molecular phylogenetic trees, useful for molecular diversity characterization, which are based on Pseudo-folding lattice 2D-TIs versus other trees that use Folding 3D-TIs values. We illustrate the comparison with a practical case: comparison of peptides found in the PMF of a new query protein reported here. In Fig. 3, we illustrate an overall view of the 2D-E map obtained from the *L. infantum* promastigote homogenate. In this figure, we have done a zooming in the left-to-down corner to highlight an area of high density of spots, which apparently corresponds to protein fragments of low MW and low pI. Our interest in this area derived from the fact that these spots remained unchanged from gel to gel repetitions and might correspond to relevant proteins of this parasite. In order to start investigation on the nature of these proteins, initially, we marked the spot with an arrow and encircled in the zoom image for this area, see Fig. 3.

MS results for new query protein

The protein contained in each spot was submitted to in-gel trypsin digestion and the mass of the resulting PMF, which is expression of the molecular diversity of the parasite protein, was obtained from MALDI-TOF MS analysis. We have studied before other proteins on the same region [194]. However, we focus our attention in this study on the protein corresponding to one spot not investigated before. Once we have obtained the data from MALDI-TOF MS analysis for this

Table 1 Classification results for RNase QSAR models based on $\pi_k(\text{SL})$, $\xi_k(\text{SL})$, and $\theta_k(\text{SL})$

Parameter	%	Group	No-RNases	RNases
$\xi_k(\text{SL})$ -model train				
Specificity	95.0	No-RNases	76	4
Sensitivity	97.1	RNases	2	68
Accuracy	96.0			
$\xi_k(\text{SL})$ -model validation				
Specificity	84.0	No-RNases	21	4
Sensitivity	100.0	RNases	0	23
Accuracy	91.7			
$\pi_k(\text{SL})$ -model train				
Specificity	76.3	No-RNases	61	19
Sensitivity	94.3	RNases	4	66
Accuracy	84.7			
$\pi_k(\text{SL})$ -model validation				
Specificity	72.0	No-RNases	18	7
Sensitivity	95.7	RNases	1	22
Accuracy	83.3			
$\theta_k(\text{SL})$ -model train				
Specificity	58.8	No-RNases	47	33
Sensitivity	74.3	RNases	18	52
Accuracy	66.0			
$\theta_k(\text{SL})$ -model validation				
Specificity	56.0	No-RNases	14	11
Sensitivity	78.3	RNases	5	18
Accuracy	66.7			

spot, the more relevant MS signals were introduced into the MASCOT search engine [195, 196]. We selected in MASCOT the *L. major* database of annotated proteins with MS recorded due to its similarity to *L. Infantum* [197]. The MASCOT search of MS signals does not match to any template hit with Ms higher than 51 ($p < 0.05$) (see Table 2). However, we found a relatively high score of Ms=42 for an RNase I with MASCOT accession code CHR16-22_tmp.17 and molecular weight Mw=108,096. The two following match founds (Ms=40 and Ms=39) correspond to template proteins CHR16-22_tmp.27 and L344.4 with Mw=30,867 and 52,863, but unknown function.

In any case, almost all relative interesting matches found have been also recorded for unknown function or hypothetical proteins. These aspects make difficult the assignation of sequence and function for the new protein. But, at the same time, increase our interest on the PMF of this new query protein that do not match to known templates. As we mentioned in the introduction of this report the PMF of this type of protein may be of high interest. In Table 3, we give detailed information on the results of the MS analysis of the PMF of

the new protein using MALDI-TOF technique and MASCOT search engine. Similar combination have been successfully used in the past to study *Trichinella* antigens [173] and possible *Leshmania* dynein proteins [194]. In this table, we have shown only the 22 more interesting peptides matching with the MS of other proteins on the MASCOT search. We calculated the three type of pseudo-folding lattice 2D-TIs for these peptides.

In Table 4, we summarized the results obtained after the QSAR-based exploration of the molecular diversity of the PMF of the new protein. We depict in this table, the pseudo-folding lattices for some peptides with higher Mw. We also predicted the contribution to RNase activity (see in Table 4 score values) using the two best QSAR models reported on this experiment (previous section). Both QSAR models coincide very well on the prediction of RNase scores for the new peptides. We found a regression coefficient of $R = 0.88$ between the RNase score of the QSAR based on $\xi_k(\mathbf{r}_2)$ values versus the model based on $\theta_k(\mathbf{r}_2)$ indices.

The QSAR study predicted the higher RNase scores for peptides P07, P08, P09, and P14. The first three peptides

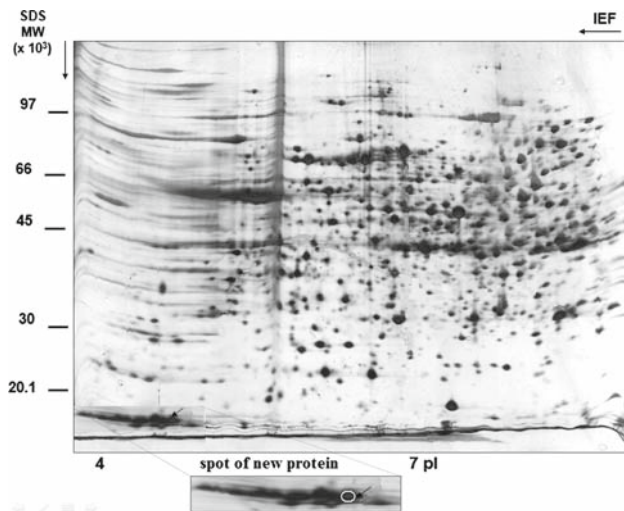


Fig. 3 2-DE analysis of proteins from *L. infantum*

Table 2 First 20 proteins found with MASCOT with weak similarity to the new *L. infantum* query

Protein	Accession number ^a	Mw ^b	Ms ^c	Annotation ^d
1	CHR16-22_tmp.17	108,096	42	RNase I
2	CHR16-22_tmp.27	30,867	40	–
3	L344.4	52,863	39	–
4	CHR7-11_tmp.271	16,228	38	Ubiquinone biosynthesis
5	CHR33_tmp.03c	88,054	36	–
6	CHR16-22_tmp.55	60,009	36	CG2839 protein
7	LmjF36.2130	63,492	36	Putative RNA helicase
8	L3856.03	60,300	35	Probable t-complex
9	LmjF36.0340	7,448	35	Nop10p
10	CHR7-11_tmp.74	62,638	34	Organizing protein
11	LM24.98	168,810	34	–
12	LmjF31.2850c	22,350	34	Ribosomal protein
13	LmjF25.1840c	72,441	32	Transcriptional regulator
14	P1408.05	25,483	32	–
15	CHR7-11_tmp.109	47,282	32	Flagellar protofilament
16	CHR16-22_tmp.74	92,715	31	Heat shock protein
17	CHR27_tmp.171	160,251	31	–
18	CHR28_tmp.22c	73,148	31	p450 reductase
19	CHR7-11_tmp.678	25,290	30	–
20	CHR7-11_tmp.616	25,290	30	–

^a Refer to the accession number used by MASCOT

^b Mw is the molecular weight

^c Ms is the MASCOT score

^d Function annotation predicted by MASCOT using alignment procedures

match with template 1, a protein previously described as RNase I. The last peptide P14 matches, however, with a template protein of unknown function. Taking into consideration, the possible interest of the peptides found on PMFs of the new

Table 3 Summary of MASCOT analysis of PMF for three best protein candidates

Peptide	Mw ^a _{obs}	Mw ^b _{expt}	Mw ^c _{calc}	Mw ^d _{dif}	sequence
Protein 1					
P01	773.46	772.46	772.41	0.05	ngvlnek
P02	789.4	788.39	788.41	−0.02	reesir
P03	927.53	926.52	926.44	0.08	ahaaaaamr
P04	999.58	998.57	998.6	−0.02	qvvtalgr
P05	1537.93	1536.92	1536.77	0.15	vmpvimgmatslqk
P06	2163.06	2162.05	2162.03	0.02	kmnvntgvvtgeeaaeeasr
P07	2223.01	2222	2222.07	−0.08	gsntnaiqmslglgqqlfdgr
P08	2238.97	2237.96	2238.12	−0.16	vmpvimgmatslqkefvpgr
Protein 2					
P09	773.46	772.46	772.46	0	tdllrr
P10	813.37	812.36	812.43	−0.07	mhisglr
P11	817.42	816.41	816.35	0.06	tgaveedp
P12	2185.03	2184.02	2184.19	−0.17	altvagdtgllasvevntarar
P13	833.43	832.43	832.38	0.05	aveeeek
Protein 3					
P14	779.46	778.45	778.4	0.05	slsgypr
P15	789.4	788.39	788.4	−0.01	dplttsr
P16	795.41	794.41	794.38	0.03	hangspgr
P17	877.47	876.46	876.44	0.02	rcllcr
P18	921.52	920.51	920.5	0.02	avaglesfk
P19	965.53	964.53	964.45	0.08	mgescllr

^a Mw_{obs}: Observed Molecular weight

^b Mw_{exp}: Experimental Molecular weight

^c Mw_{calc}: Calculated Molecular weight

^d Mw_{dif}: Difference between Mw_{calc} and Mw_{exp}

protein for the design of new RNases, we decided to confirm the predicted scores with a BLAST alignment search. In Table 5, we summarized the result of this search. The BLAST score was adjusted considering that we use here short peptides chains of <20 aa length and not full protein sequences. We selected this approach, since BLAST-like method, such as PSI-BLAST, and other methods have been used to confirm and/or complement predictive algorithms before [39]. In Table 5, we can note that in fact both QSAR and BLAST predict a positive RNase score for these peptides. This may be relevant, as we are using alternative methods that complement each other (QSAR is alignment-free whereas BLAST rely upon alignment) [127, 198–201].

Experiment 2

MD simulation for the PMF of the new protein

It can be noted in Table 4 that in this type of representation some aminoacids (aa) overlap on the same nodes resulting that the number of aa is higher than the number of nodes in


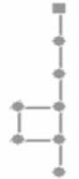


the lattice (see *Experiment 1*). This aspect plus the pseudo-folding procedure used to obtain lattices (not real folding) have given rise to the question about the structural accuracy versus computational cost, when we compare 2D-TIs to 3D-TIs. The problem is relevant and not only restricted to lattices 2D-TIs but also any kind of 2D-TIs [202]. In this sense, we decided to investigate in which extension the pseudo-folding lattice 2D-TIs are able to capture information present on 3D structure. For it, we first need the 3D structures of the peptides in order to calculate the 3D folding versions of the same type of pseudo-folding TIs. Then, we need to compare the higher dimension $\pi_k(\mathbf{r}_3)$, $\theta_k(\mathbf{r}_3)$, and $\xi_k(\mathbf{r}_3)$ values with the lower dimension $\pi_k(\mathbf{r}_2)$, $\theta_k(\mathbf{r}_2)$, and $\xi_k(\mathbf{r}_2)$ indices. For this study, we used the same 19 peptides found on the PMF of the new protein. Unfortunately, we have only the sequences of the peptides but not the 3D structures. Consequently, we obtained first, the optimal 3D folded structures using a MD search for the 19 peptides (see Fig. 2). In Table 6, we have summarized the results of MD simulation of these peptides. In this table, we reported the initial energy (E_0) and energy gradient (δ_0) based on the starting structure constructed with standard parameters for α -helixes (bond distances, angles,

and dihedral angles) set as default on the sequence editor of Hyperchem [170, 171]. We also reported the (E_1) and energy gradient (δ_1) obtained after optimization of the structure with AMBER force field obtained by MC method applied to MD simulation. Finally, we report in Supplementary material file sm3 the ACCR values for the MDT of the 19 peptides. In the MD study, most researchers tend to try for an average ACCR value around 0.5 and smaller values may be appropriate when longer runs are acceptable, and more extensive sampling is necessary. In the present study, all the ACCR values were lower than 5.0, in consequence, we can accept the MD results as valid [170, 171].

2D-TI versus 3D-TI phylogentic study of PMF for new protein

Using information about the distribution of aminoacids in the sequence of the protein has been the major tendency on molecular phylogentic analysis [203]. In the introduction, we discussed the importance of new molecular phylogentic approaches for protein based on other types of molecular structure information. In materials and methods, we outlined

Table 4 Summary of QSAR Data Mining exploration of 19 peptides found on the PMF of query protein

Inputs ^a				Score ^d		Inputs ^a				Score ^d	
Peptide	Sequence	aa ^b	n ^c	S(ξ)	S(θ)	Peptide	Sequence	aa ^b	n ^c	S(ξ)	S(θ)
P01	ngvlnek	7	4	6.6	0.8	P04	qvvtalrgr	9	7	32.7	4.6
P02	reesir	6	4	2.9	0.8	P05	vmpvimgmatslqk	14	8	56.1	5.4
P03	aheaaaamr	9	8	33.7	5.4	P9	tdllrr	6	6	14.3	3.7
P06		22	13	101.6	8.0	P07		22	8	91.4	5.4
 kmnvntgvvtgeeaaeeaaasr						 gsntnaiqmslglgqqldgr					
P08		21	10	89.2	6.6	P12		23	12	112.5	7.6
 vmpvimgmatslqkefvpggr						 altvagdtgllasvevntarar					
P10	mhisglr	7	5	12.1	2.5	P16	hangspgr	8	5	17.7	2.5
P11	tgaveedp	8	7	23.8	4.6	P17	rcllcr	6	4	1.1	0.8
P13	aveeeek	7	6	17.3	3.7	P18	avaglesfk	9	5	24.4	2.5
P14	smsgypr	7	5	14.7	2.5	P19	mgescllr	8	5	18.1	2.5
P15	dpltsr	7	5	14.9	2.5						

^a Information related to the input lattice graphs and/or peptides

^b aa is the number of aminoacids

^c n is the number of nodes in the lattice graph

^d Scores predicted with the QSAR models

Table 5 Summary of MASCOT, QSAR, and BLAST RNase scores of some relevant peptides in PMF

MASCOT scores		
Template 1	Protein	Template 2
DNA-directed RNA polymerase I	Function	Hypothetical protein
CHR16-22_tmp.17	ID	CHR16-22_tmp.27
108 096	Mass	30 867
42	Mowse	40

Protein 1

Protein 2

BLAST vs. QSAR Scoring for some peptides		
P06		P08
kmnvntgvvtgeaaeeaaar	sequence	vmpvimgmatslqkefvpgr
36.7	BLAST	69.4
101.6	QSAR	89.2

P07		P12
gsntnaiqmslglgqldgr	sequence	altvagdtgllasvevntarar
68.5	BLAST	33.3
91.4	QSAR	112.5

the possibility of construction of a phylogenetic tree for the PMFs of the new protein using TIs based on folded \mathbf{r}_3 structure or pseudo-folded structures in \mathbf{r}_2 . In the previous section, we recalled that the first type of TIs gives a more realistic picture of the protein structure, but the second-one are easier to calculate, which is important to scale the method up for large databases [202]. In this sense, it is important to compare the different TIs and the subsequent phylogenetic trees generated. For it, we have calculated first, the $TI_k(\mathbf{r}_d)$ values for the 19 peptides and then the peptide-peptide distance using Eq. 9. We calculated only the $TI_k(\mathbf{r}_d)$ that have some relevance for RNase activity according to the QSAR Eqs. 10, 11,

and 12. It means that, we calculated the pseudo-folding indices $\xi_1(\mathbf{r}_2)$, $\xi_2(\mathbf{r}_2)$, $\pi_0(\mathbf{r}_2)$, $\pi_2(\mathbf{r}_2)$, and $\theta_0(\mathbf{r}_2)$. In Table 6, we reported the values of all these $TI_k(\mathbf{r}_d)$ for the 19 peptides.

In Fig. 4, we illustrated with a Two-way joining analysis that the indices calculated at different structural levels have typical values and forming structural clusters. In fact, Two-way joining analysis can detect automatically the 2D-pseudo-folding cluster and the cluster for 3D-folding TIs. It demonstrates that the method presents variations on the results depending on the detail level selected to describe the protein structure. In order to reaffirm this, we calculated the TIs using 3D-folded structure considering all atoms in the protein and

Table 6 Some $\xi_k(\mathbf{r}_2)$, $\theta_k(\mathbf{r}_2)$, and $\pi_k(\mathbf{r}_2)$ values for 19 peptides found on the PMF of the new protein

Peptide	Sequence	$\theta_0(\mathbf{r}_2)$	$\pi_0(\mathbf{r}_2)$	$\pi_2(\mathbf{r}_2)$	$\xi_1(\mathbf{r}_2)$	$\xi_2(\mathbf{r}_2)$
P01	ngvlnek	0.60	4	1.00	6.75	6.42
P02	reesir	0.60	4	1.19	4.72	4.06
P03	aheaaaaamr	0.90	8	2.40	7.00	5.90
P04	qvvtalgr	0.85	7	2.04	7.17	6.05
P05	vmpvimgmatslqk	0.90	8	2.00	10.54	9.00
P06	kmnvnvtgvtgeeaaeeaaasr	1.11	13	3.30	15.66	12.99
P07	gsntnaiqmslglgqqlfdgr	0.90	8	1.40	14.58	11.92
P08	vmpvimgmatslqkefvpggr	1.00	10	2.20	14.01	11.50
P09	tdllrr	0.78	6	1.93	4.84	4.08
P10	mhisglr	0.70	5	1.01	5.07	4.28
P11	tgaveedp	0.85	7	2.04	5.99	5.11
P12	altvagdgtgllasvevntarar	1.08	12	2.49	14.75	11.09
P13	aveeeek	0.78	6	1.86	5.64	4.91
P14	slsgypr	0.70	5	1.31	5.00	4.03
P15	dpltsr	0.70	5	1.47	5.38	4.50
P16	hangspgr	0.70	5	1.48	5.90	4.99
P17	rcllcr	0.60	4	1.02	4.98	4.50
P18	avaglesfk	0.70	5	1.33	6.76	5.66
P19	mgescllr	0.70	5	1.42	5.87	4.92

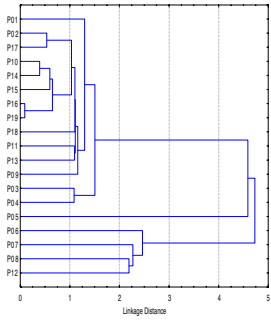
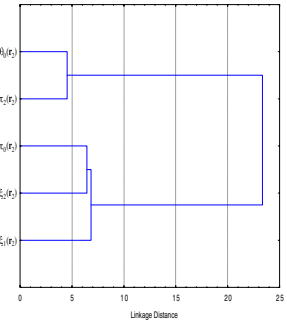
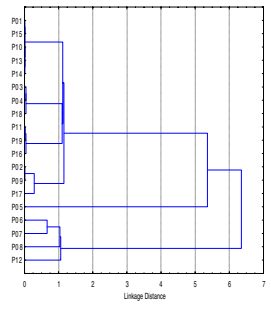
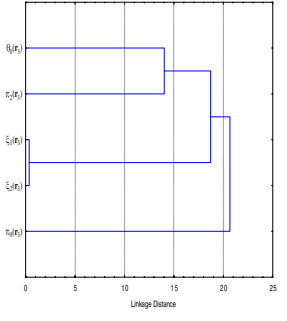
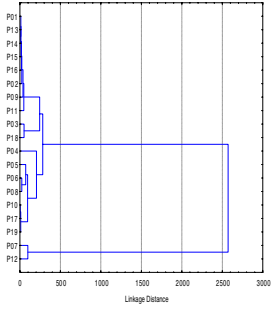
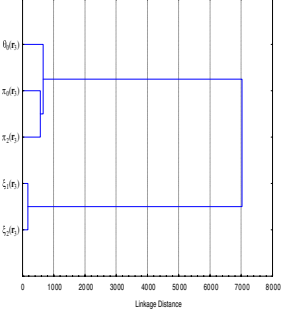
3D-TI aa-C α -only scheme (PDB file)						3D-TI all-atoms scheme (HIN file)				
Pept.	$\theta_0(\mathbf{r}_3)$	$\pi_0(\mathbf{r}_3)$	$\pi_2(\mathbf{r}_3)$	$\xi_1(\mathbf{r}_3)$	$\xi_2(\mathbf{r}_3)$	$\theta_0(\mathbf{r}_3)$	$\pi_0(\mathbf{r}_3)$	$\pi_2(\mathbf{r}_3)$	$\xi_1(\mathbf{r}_3)$	$\xi_2(\mathbf{r}_3)$
P01	7.03	7	4.19	0.57	0.63	16.9	107	171.0	1221.5	1248.1
P02	6.47	6	4.22	0.45	0.51	16.9	108	180.2	1252.3	1285.8
P03	7.93	9	4.40	0.58	0.66	17.5	126	186.4	1449.4	1485.7
P04	7.93	9	4.43	0.60	0.68	18.0	147	198.8	1679.7	1722.5
P05	9.53	14	5.49	0.73	0.81	19.5	222	0.0	1653.0	1689.0
P06	10.99	21	7.03	1.18	1.31	20.5	294	0.0	1653.0	1689.0
P07	10.99	21	7.69	1.15	1.27	20.6	304	295.5	3479.0	3551.0
P08	10.82	20	7.07	1.05	1.15	20.8	318	0.0	1653.0	1689.0
P09	6.47	6	4.22	0.46	0.51	17.0	112	174.7	1279.6	1310.4
P10	7.03	7	4.19	0.56	0.62	17.2	117	0.0	1653.0	1689.0
P11	7.51	8	4.18	0.67	0.74	16.7	103	167.6	1179.0	1198.9
P12	11.16	22	7.30	1.28	1.42	20.7	311	294.0	3547.4	3620.8
P13	7.03	7	4.20	0.55	0.62	16.9	108	171.7	1234.3	1255.3
P14	7.03	7	4.20	0.55	0.61	16.9	107	171.6	1230.8	1255.1
P15	7.03	7	4.19	0.57	0.64	16.9	108	171.9	1241.6	1266.9
P16	7.51	8	4.22	0.68	0.75	16.8	105	171.2	1204.9	1240.4
P17	6.47	6	4.48	0.37	0.43	17.0	112	0.0	1653.0	1689.0
P18	7.93	9	4.36	0.58	0.65	17.6	130	187.1	1489.9	1515.7
P19	7.51	8	4.18	0.68	0.76	17.5	127	0.0	1653.0	1689.0

not only C α atoms as many researchers use to. The results show that we can detect certain hierarchy in the cluster organization of the indices (see Fig. 4).

However, in cluster analysis, we can easily note that even (see Table 6) the three classes of indices have different values and form different clusters. The overall variability for

all the indices is very similar in each peptide and somehow peptide specific. It means that peptide-to-peptide variations are more notable than structural level variations. In fact, the results of the phylogenetic tree analysis demonstrated relatively larger variations on the alternative clustering of the 19 peptides than on the alternative clustering of TIs using \mathbf{r}_2 ,

Table 7 Comparative study of pseudo-folding, folding and all-atoms folding schemes

		2D Pseudo-folding			
Peptide Phylogenetic tree	Statistics		TI Phylogenetic tree	Statistics	
	Peptide	Mean ^a		TI	Mean
	P01	3.8		$\theta_0(\mathbf{r}_2)$	0.8
	P02	2.9		$\pi_0(\mathbf{r}_2)$	6.7
	P03	4.8		$\pi_2(\mathbf{r}_2)$	1.7
	P04	4.6		$\xi_1(\mathbf{r}_2)$	7.9
	P05	6.1		$\xi_2(\mathbf{r}_2)$	6.6
	P06	9.2			
	P07	7.4			
	P08	7.7			
	P09	3.5			
	P10	3.2			
	P11	4.2			
	P12	8.3			
3D Folding					
	P01	3.9		$\theta_0(\mathbf{r}_3)$	8.1
	P02	3.5		$\pi_0(\mathbf{r}_3)$	10.6
	P03	4.5		$\pi_2(\mathbf{r}_3)$	5.0
	P04	4.5		$\xi_1(\mathbf{r}_3)$	0.7
	P05	6.1		$\xi_2(\mathbf{r}_3)$	0.8
	P06	8.3			
	P07	8.4			
	P08	8.0			
	P09	3.5			
	P10	3.9			
	P11	4.2			
	P12	8.6			
3D Folding All-atoms					
	P01	552.9		$\theta_0(\mathbf{r}_3)$	18.0
	P02	568.6		$\pi_0(\mathbf{r}_3)$	161.4
	P03	653.0		$\pi_2(\mathbf{r}_3)$	133.8
	P04	753.2		$\xi_1(\mathbf{r}_3)$	1653.0
	P05	716.7		$\xi_2(\mathbf{r}_3)$	1689.0
	P06	731.3			
	P07	1530.0			
	P08	736.2			
	P09	578.7			
	P10	695.2			
	P11	533.0			
	P12	1558.8			

^a Mean distance from this peptide to the other 11 peptides

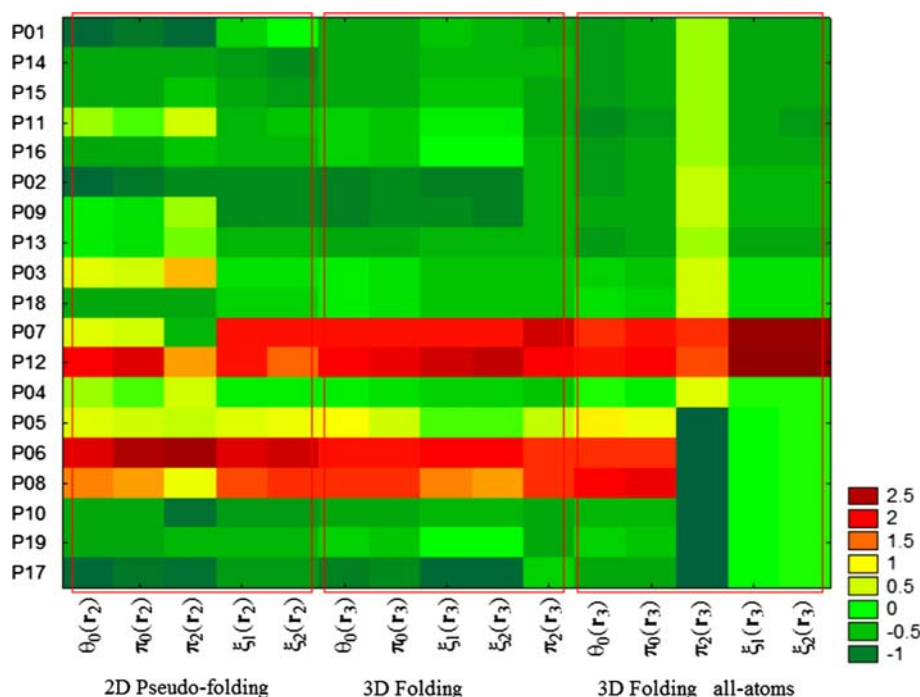
\mathbf{r}_3 for C_α only, or all-atoms \mathbf{r}_3 TIs. In Table 7, we depict the final results obtained for the phylogenetic tree analysis of either peptides or TIs. This results show that, in principle, the distance $^{TI}D_{pq}(\mathbf{r}_d)$ between a peptide p and other q based on $TI_k(\mathbf{r}_2)$ is structurally sensitive and codify sufficient structural information with respect to more detailed structural level. Actually, an inspection of a simple correlation matrix demonstrated that all the TIs calculated have correlations are significant at $p < 0.05$ except for $\pi_k(\mathbf{r}_3)$ based on all atoms, which seems to be the more structurally sensitive TI calculated in this study. We can conclude that pseudo-fold-

ing $TI_k(\mathbf{r}_2)$ phylogenetic algorithms may become a fast and efficient alternative to $TI_k(\mathbf{r}_3)$ methods, as well as a higher structurally detail complement to traditional sequence-only methods.

Conclusions

In this study, we demonstrate that it is possible to develop and compare alignment-free QSAR models using sequence pseudo-folding TIs (based on Markov matrices). In addition, we compared this indices with similar indices based on 3D

Fig. 4 Two-way joining study of folding TIs for different structural levels



structures obtained by MD simulation. We also show with a practical example, the use of these QSAR and Molecular Phylogenetic models to predict RNase activity and explore the molecular diversity of peptides found on the PMFs of the new query protein isolated here by the first time from *L. infantum*.

Acknowledgments We thank editors G. A. Morales and K. Roy for kind invitation to submit this work. H. Gonzalez-Díaz acknowledges program Isidro Parga Pondal of the Xunta de Galicia and European Union (F. S. E.) by funding a research contract position at the Faculty of Pharmacy, USC, Spain. Authors from University of Porto acknowledge the Portuguese Fundação para a Ciência e a Tecnologia (FCT) (SFRH/BD/47256/2008) for financial support.

References

- Dyer KD, Rosenberg HF (2006) The RNase a superfamily: generation of diversity and innate host defense. *Mol Divers* 10:585–597
- Schirrmann T, Krauss J, Arndt MA, Rybak SM, Dubel S (2009) Targeted therapeutic RNases (ImmunoRNases). *Expert Opin Biol Ther* 9:79–95
- Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Radmark O, Kim S, Kim VN (2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature* 425:415–419
- Pekarik V (2005) Design of shRNAs for RNAi—a lesson from pre-miRNA processing: possible clinical applications. *Brain Res Bull* 68:115–120
- Zhou WW, Niu TG (2009) Purification and some properties of an extracellular ribonuclease with antiviral activity against tobacco mosaic virus from *Bacillus cereus*. *Biotechnol Lett* 31:101–105
- Aksu S, Scheler C, Focks N, Leenders F, Theuring F, Salmikow J, Jungblut PR (2002) An iterative calibration method with prediction of post-translational modifications for the construction of a two-dimensional electrophoresis database of mouse mammary gland proteins. *Proteomics* 2:1452–1463
- Tebbe A, Klein C, Bisle B, Siedler F, Scheffer B, Garcia-Rizo C, Wolfertz J, Hickmann V, Pfeiffer F, Oesterhelt D (2005) Analysis of the cytosolic proteome of *Halobacterium salinarum* and its implication for genome annotation. *Proteomics* 5:168–179
- Gao L, Ding YS, Dai H, Shao SH, Huang ZD, Chou KC (2006) A novel fingerprint map for detecting SARS-CoV. *J Pharm Biomed Anal* 41:246–250
- Wang M, Yao JS, Huang ZD, Xu ZJ, Liu GP, Zhao HY, Wang XY, Yang J, Zhu YS, Chou KC (2005) A new nucleotide-composition based fingerprint of SARS-CoV with visualization analysis. *Med Chem* 1:39–47
- Perkins DN, Pappin DJC, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20:3551–3567
- Resing KA, Meyer-Arendt K, Mendoza AM, Aveline-Wolf LD, Jonscher KR, Pierce KG, Old WM, Cheung HT, Russell S, Wattawa JL, Goehle GR, Knight RD, Ahn NG (2004) Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal Chem* 76:3556–3568
- Savitski MM, Nielsen ML, Kjeldsen F, Zubarev RA (2005) Proteomics-grade de novo sequencing approach. *J Proteome Res* 4:2348–2354
- Savitski MM, Nielsen ML, Zubarev RA (2005) New data base-independent, sequence tag-based scoring of peptide MS/MS data validates Mowse scores, recovers below threshold data, singles out modified peptides, and assesses the quality of MS/MS techniques. *Mol Cell Proteomics* 4:1180–1188
- Chou KC (2009) Automated prediction of protein attributes and its impact to biomedicine and drug discovery. In: Alterovitz

- G, Benson R, Ramoni MF (eds) Automation in proteomics and genomics. Wiley, UK, pp 97–143
15. Chou KC (2004) Structural bioinformatics and its impact to biomedical science. *Curr Med Chem* 11:2105–2134
 16. Chou KC (2004) Molecular therapeutic target for type-2 diabetes. *J Proteome Res* 3:1284–1288
 17. Chou KC, Wei DQ, Zhong WZ (2003) Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS. *Biochem Biophys Res Commun* 308:148–151
 18. Li Y, Wei DQ, Gao WN, Gao H, Liu BN, Huang CJ, Xu WR, Liu DK, Chen HF, Chou KC (2007) Computational approach to drug design for oxazolidinones as antibacterial agents. *Med Chem* 3:576–582
 19. Wang JF, Wei DQ, Chen C, Li Y, Chou KC (2008) Molecular modeling of two CYP2C19 SNPs and its implications for personalized drug design. *Protein Pept Lett* 15:27–32
 20. Chou KC, Nemethy G, Scheraga HA (1984) Energetic approach to packing of α -helices: 2. General treatment of non-equivalent and nonregular helices. *J Am Chem Soc* 106:3161–3170
 21. Chou KC, Maggiora GM, Nemethy G, Scheraga HA (1988) Energetics of the structure of the four- α -helix bundle in proteins. *Proc Natl Acad Sci USA* 85:4295–4299
 22. Sirois S, Wei DQ, Du Q, Chou KC (2004) Virtual screening for SARS-CoV protease based on KZ7088 pharmacophore points. *J Chem Inf Comput Sci* 44:1111–1122
 23. Chou KC, Wei DQ, Du QS, Sirois S, Zhong WZ (2006) Progress in computational approach to drug development against SARS. *Curr Med Chem* 13:3263–3270
 24. Chou KC (1992) Energy-optimized structure of antifreeze protein and its binding mechanism. *J Mol Biol* 223:509–517
 25. Chou KC, Zhou GP (1982) Role of the protein outside active site on the diffusion-controlled reaction of enzyme. *J Am Chem Soc* 104:1409–1413
 26. Chou KC, Shen HB (2007) MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem Biophys Res Commun* 360:339–345
 27. Shen HB, Chou KC (2007) EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. *Biochem Biophys Res Commun* 364:53–59
 28. Chou KC (2005) Prediction of G-protein-coupled receptor classes. *J Proteome Res* 4:1413–1418
 29. Xiao X, Wang P, Chou KC (2009) GPCR-CA: a cellular automation image approach for predicting G-protein-coupled receptor functional classes. *J Comput Chem* 30:1414–1423
 30. Chou KC, Shen HB (2008) ProIdent: a web server for identifying proteases and their types by fusing functional domain and sequential evolution information. *Biochem Biophys Res Commun* 376:321–325
 31. Shen HB, Chou KC (2009) Identification of proteases and their types. *Anal Biochem* 385:153–160
 32. Chou KC (1993) A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. *J Biol Chem* 268:16938–16948
 33. Chou KC (1996) Prediction of human immunodeficiency virus protease cleavage sites in proteins. *Anal Biochem* 233:1–14
 34. Shen HB, Chou KC (2008) HIVcleave: a web-server for predicting HIV protease cleavage sites in proteins. *Anal Biochem* 375:388–390
 35. Chou KC, Shen HB (2007) Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. *Biochem Biophys Res Commun* 357:633–640
 36. Shen HB, Chou KC (2007) Signal-3L: a 3-layer approach for predicting signal peptides. *Biochem Biophys Res Commun* 363:297–303
 37. Tamiya T, Fujimi TJ (2006) Molecular evolution of toxin genes in Elapidae snakes. *Mol Divers* 10:529–543
 38. Lajoix AD, Gross R, Akinin C, Dietz S, Granier C, Laune D (2004) Cellulose membrane supported peptide arrays for deciphering protein–protein interaction sites: the case of PIN, a protein with multiple natural partners. *Mol Divers* 8:281–290
 39. Song J, Burrage K, Yuan Z, Huber T (2006) Prediction of cis/trans isomerization in proteins using PSI-BLAST profiles and secondary structure information. *BMC Bioinformatics* 7:124
 40. Balakrishnan R, Christie KR, Costanzo MC, Dolinski K, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hong EL, Nash R, Oughtred R, Skrzypek M, Theesfeld CL, Binkley G, Dong Q, Lane C, Sethuraman A, Weng S, Botstein D, Cherry JM (2005) Fungal BLAST and model organism BLASTP best hits: new comparison resources at the Saccharomyces Genome Database (SGD). *Nucleic Acids Res* 33:D374–D377
 41. Han L, Cui J, Lin H, Ji Z, Cao Z, Li Y, Chen Y (2006) Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity. *Proteomics* 6:4023–4037
 42. Lin HH, Han LY, Zhang HL, Zheng CJ, Xie B, Chen YZ (2006) Prediction of the functional class of lipid binding proteins from sequence-derived properties irrespective of sequence similarity. *J Lipid Res* 47:824–831
 43. Lin HH, Han LY, Cai CZ, Ji ZL, Chen YZ (2006) Prediction of transporter family from protein sequence by support vector machine approach. *Proteins* 62:218–231
 44. Han LY, Cai CZ, Ji ZL, Cao ZW, Cui J, Chen YZ (2004) Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. *Nucleic Acids Res* 32:6437–6444
 45. Han LY, Cai CZ, Ji ZL, Chen YZ (2005) Prediction of functional class of novel viral proteins by a statistical learning method irrespective of sequence similarity. *Virology* 331:136–143
 46. Fontaine F, Pastor M, Gutierrez-de-Teran H, Lozano JJ, Sanz F (2003) Use of alignment-free molecular descriptors in diversity analysis and optimal sampling of molecular libraries. *Mol Divers* 6:135–147
 47. Chou KC (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* 43:246–255
 48. Chou KC (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* 21:10–19
 49. Chou KC, Shen HB (2007) Large-scale plant protein subcellular location prediction. *J Cell Biochem* 100:665–678
 50. Zhang GY, Fang BS (2008) Predicting the cofactors of oxidoreductases based on amino acid composition distribution and Chou's amphiphilic pseudo amino acid composition. *J Theor Biol* 253:310–315
 51. Lin H (2008) The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. *J Theor Biol* 252:350–356
 52. Jiang X, Wei R, Zhang TL, Gu Q (2008) Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy. *Protein Pept Lett* 15:392–396
 53. Chou KC, Shen HB (2006) Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. *Biochem Biophys Res Commun* 347:150–157
 54. Zhou XB, Chen C, Li ZC, Zou XY (2007) Using Chou's amphiphilic pseudo-amino acid composition and support vector

- machine for prediction of enzyme subfamily classes. *J Theor Biol* 248:546–551
55. Zhang GY, Fang BS, Li HC (2008) Predicting lipase types by improved Chou's pseudo-amino acid composition. *Protein Pept Lett* 15:1132–1137
 56. Lin H, Ding H, Guo FBF-B, Zhang AY, Huang J (2008) Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. *Protein Pept Lett* 15:739–744
 57. Ding YS, Zhang TL (2008) Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier. *Pattern Recognit Lett* 29:1887–1892
 58. Chou KC, Jiang SP, Liu WM, Fee CH (1979) Graph theory of enzyme kinetics: 1. Steady-state reaction system. *Sci Sinica* 22:341–358
 59. Chou KC, Forsen S (1980) Graphical rules for enzyme-catalysed rate laws. *Biochem J* 187:829–835
 60. Chou KC (1981) Two new schematic rules for rate laws of enzyme-catalysed reactions. *J Theor Biol* 89:581–592
 61. Zhou GP, Deng MH (1984) An extension of Chou's graphical rules for deriving enzyme kinetic equations to system involving parallel reaction pathways. *Biochem J* 222:169–176
 62. Myers D, Palmer G (1985) Microcomputer tools for steady-state enzyme kinetics. *Comput Appl Biosci* 1:105–110
 63. Andraos J (2008) Kinetic plasticity and the determination of product ratios for kinetic schemes leading to multiple products without rate laws: new methods based on directed graphs. *Can J Chem* 86:342–357
 64. Chou KC (1989) Graphic rules in steady and non-steady state enzyme kinetics. *J Biol Chem* 264:12074–12079
 65. Chou KC (1990) Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady-state systems. *Biophys Chem* 35:1–24
 66. Althaus IW, Chou JJ, Gonzales AJ, Deibel MR, Chou KC, Kezdy FJ, Romero DL, Aristoff PA, Tarpley WG, Reusser F (1993) Steady-state kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-87201E. *J Biol Chem* 268:6119–6124
 67. Althaus IW, Gonzales AJ, Chou JJ, Romero DL, Deibel MR, Chou KC, Kezdy FJ, Resnick L, Busso ME, So AG et al (1993) The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. *J Biol Chem* 268:14875–14880
 68. Chou KC, Kezdy FJ, Reusser F (1994) Steady-state inhibition kinetics of processive nucleic acid polymerases and nucleases. *Anal Biochem* 221:217–230
 69. Chou KC, Zhang CT (1992) Diagrammatization of codon usage in 339 human immunodeficiency virus proteins and its biological implication. *AIDS Res Hum Retroviruses Nat Protoc* 8:1967–1976
 70. Zhang CT, Chou KC (1994) A graphic approach to analyzing codon usage in 1562 *Escherichia coli* protein coding sequences. *J Mol Biol* 238:1–8
 71. Chou KC, Zhang CT, Elrod DW (1996) Do "antisense proteins" exist?. *J Protein Chem* 15:59–61
 72. Gonzalez-Diaz H, Sanchez-Gonzalez A, Gonzalez-Diaz Y (2006) 3D-QSAR study for DNA cleavage proteins with a potential anti-tumor ATCUN-like motif. *J Inorg Biochem* 100:1290–1297
 73. Prado-Prado FJ, Gonzalez-Diaz H, de la Vega OM, Ubeira FM, Chou KC (2008) Unified QSAR approach to antimicrobials. Part 3: first-tasking QSAR model for input-coded prediction, structural back-projection, and complex networks clustering of antiprotozoal compounds. *Bioorg Med Chem* 16:5871–5880
 74. Gonzalez-Diaz H, Bonet I, Teran C, De Clercq E, Bello R, Garcia MM, Santana L, Uriarte E (2007) ANN-QSAR model for selection of anticancer leads from structurally heterogeneous series of compounds. *Eur J Med Chem* 42:580–585
 75. Gonzalez-Diaz H, Gonzalez-Diaz Y, Santana L, Ubeira FM, Uriarte E (2008) Proteomics, networks and connectivity indices. *Proteomics* 8:750–778
 76. Gonzalez-Diaz H, Vilar S, Santana L, Uriarte E (2007) Medicinal chemistry and bioinformatics—current trends in drugs discovery with networks topological indices. *Curr Top Med Chem* 7:1015–1029
 77. Wolfram S (1984) Cellular automation as models of complexity. *Nat Protoc* 311:419–424
 78. Wolfram S (2002) A new kind of science. Wolfram Media, Champaign, IL
 79. Xiao X, Shao SH, Chou KC (2006) A probability cellular automaton model for hepatitis B viral infections. *Biochem Biophys Res Commun* 342:605–610
 80. Xiao X, Shao S, Ding Y, Huang Z, Chen X, Chou KC (2005) An application of gene comparative image for predicting the effect on replication ratio by HBV virus gene missense mutation. *J Theor Biol* 235:555–565
 81. Xiao X, Shao S, Ding Y, Huang Z, Chen X, Chou KC (2005) Using cellular automata to generate image representation for biological sequences. *Amino Acids* 28:29–35
 82. Xiao X, Shao SH, Ding YS, Huang ZD, Chou KC (2006) Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. *Amino Acids* 30:49–54
 83. Xiao X, Chou KC (2007) Digital coding of amino acids based on hydrophobic index. *Protein Pept Lett* 14:871–875
 84. Liao B, Ding K (2005) Graphical approach to analyzing DNA sequences. *J Comput Chem* 26:1519–1523
 85. Liao B, Wang TM (2004) Analysis of similarity/dissimilarity of DNA sequences based on nonoverlapping triplets of nucleotide bases. *J Chem Inf Comput Sci* 44:1666–1670
 86. Liao B, Wang TM (2004) New 2D graphical representation of DNA sequences. *J Comput Chem* 25:1364–1368
 87. Liao B, Xiang X, Zhu W (2006) Coronavirus phylogeny based on 2D graphical representation of DNA sequence. *J Comput Chem* 27:1196–1202
 88. Yu-Hua Y, Liao B, Tian-Ming W (2005) A 2D graphical representation of RNA secondary structures and the analysis of similarity/dissimilarity based on it. *J Mol Struct Theochem* 755:131–136
 89. Liao B, Wang T (2004) A 3D Graphical representation of RNA secondary structure. *J Biomol Struct Dyn* 21:827–832
 90. Liao B, Ding K, Wang T (2005) On a six-dimensional representation of RNA secondary structures. *J Biomol Struct Dyn* 22:455–464
 91. Liao B, Wang T, Ding K (2005) On a seven-dimensional representation of RNA secondary structures. *Mol Simulat* 31:1063–1071
 92. Liao B, Luo J, Li R, Zhu W (2006) RNA secondary structure 2D graphical representation without degeneracy. *Int J Quantum Chem* 106:1749–1755
 93. Zhu W, Liao B, Ding K (2005) A condensed 3D graphical representation of RNA secondary structures. *J Mol Struct Theochem* 757:193–198
 94. Randic M, Vracko M (2000) On the similarity of DNA primary sequences. *J Chem Inf Comput Sci* 40:599–606
 95. Agüero-Chapin G, González-Díaz H, Molina R, Varona-Santos J, Uriarte E, Gonzalez-Diaz Y (2006) Novel 2D maps and coupling numbers for protein sequences. The first QSAR study of polygalacturonases: isolation and prediction of a novel sequence from *Psidium guajava* L. *FEBS Lett* 580:723–730

96. Randić M, Vračko M, Nandy A, Basak SC (2000) On 3-D graphical representation of DNA primary sequences and their numerical characterization. *J Chem Inf Comput Sci* 40:1235–1244
97. Nandy A (1996) Two-dimensional graphical representation of DNA sequences and intron-exon discrimination in intron-rich sequences. *Comput Appl Biosci* 12:55–62
98. González-Díaz H, González-Díaz Y, Santana L, Ubeira FM, Uriarte E (2008) Proteomics, networks and connectivity indices. *Proteomics* 8:750–778
99. González-Díaz H, Vilar S, Santana L, Uriarte E (2007) Medicinal chemistry and bioinformatics: current trends in drugs discovery with networks topological indices. *Curr Top Med Chem* 7:1025–1039
100. Li W, Lin K, Feng K, Cai Y (2008) Prediction of protein structural classes using hybrid properties. *Mol Divers* 12:171–179
101. Du QS, Huang RB, Wei YT, Du LQ, Chou KC (2008) Multiple field three dimensional quantitative structure–activity relationship (MF-3D-QSAR). *J Comput Chem* 29:211–219
102. Leonard JT, Roy K (2005) QSAR by LFER model of HIV protease inhibitor mannitol derivatives using FA-MLR, PCRA, and PLS techniques. *Bioorg Med Chem* 13:2967–2973
103. Roy K, Leonard JT (2005) QSAR analyses of 3-(4-benzylpiperidin-1-yl)-N-phenylpropylamine derivatives as potent CCR5 antagonists. *J Chem Inf Model* 45:1352–1368
104. Bhattacharya P, Roy K (2005) QSAR of adenosine A3 receptor antagonist 1,2,4-triazolo[4,3-a]quinoxalin-1-one derivatives using chemometric tools. *Bioorg Med Chem Lett* 15:3737–3743
105. Bhattacharya P, Leonard JT, Roy K (2005) Exploring 3D-QSAR of thiazole and thiadiazole derivatives as potent and selective human adenosine A3 receptor antagonists. *J Mol Model* 11:516–524
106. Roy K (2004) Topological descriptors in drug design and modeling studies. *Mol Divers* 8:321–323
107. Roy K, Mandal AS (2009) Predictive QSAR modeling of CCR5 antagonist piperidine derivatives using chemometric tools. *J Enzyme Inhib Med Chem* 24:205–223
108. Du Q, Mezey PG, Chou KC (2005) Heuristic molecular lipophilicity potential (HMLP): a 2D-QSAR study to LADH of molecular family pyrazole and derivatives. *J Comput Chem* 26:461–470
109. Pasha FA, Srivastava HK, Singh PP (2005) Semiempirical QSAR study and ligand receptor interaction of estrogens. *Mol Divers* 9:215–220
110. Golbraikh A, Tropsha A (2002) Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Mol Divers* 5:231–243
111. Ghafourian T, Cronin MT (2004) Comparison of electrotopological-state indices versus atomic charge and superdelocalisability indices in a QSAR study of the receptor binding properties of halogenated estradiol derivatives. *Mol Divers* 8:343–355
112. Gao H, Bajorath J (1998) Comparison of binary and 2D QSAR analyses using inhibitors of human carbonic anhydrase II as a test case. *Mol Divers* 4:115–130
113. Estrada E, Quincoces JA, Patlewicz G (2004) Creating molecular diversity from antioxidants in Brazilian propolis. Combination of TOPS-MODE QSAR and virtual structure generation. *Mol Divers* 8:21–33
114. Douali L, Villemin D, Zyad A, Cherqaoui D (2004) Artificial neural networks: non-linear QSAR studies of HEPT derivatives as HIV-1 reverse transcriptase inhibitors. *Mol Divers* 8:1–8
115. Besalu E, Ponc R, de Julian-Ortiz JV (2003) Virtual generation of agents against *Mycobacterium tuberculosis*: a QSAR study. *Mol Divers* 6:107–120
116. Balaban AT, Basak SC, Beteringhe A, Mills D, Supuran CT (2004) QSAR study using topological indices for inhibition of carbonic anhydrase II by sulfanilamides and Schiff bases. *Mol Divers* 8:401–412
117. Agrawal VK, Srivastava S, Khadikar PV (2004) QSAR study on phosphoramidothioate (Ace) toxicities in housefly. *Mol Divers* 8:413–419
118. Afantitis A, Melagraki G, Sarimveis H, Koutentis PA, Markopoulos J, Igglessi-Markopoulou O (2006) A novel simple QSAR model for the prediction of anti-HIV activity using multiple linear regression analysis. *Mol Divers* 10:405–414
119. Du QS, Huang RB, Wei YT, Pang ZW, Du LQ, Chou KC (2009) Fragment-based quantitative structure–activity relationship (FB-QSAR) for fragment-based drug design. *J Comput Chem* 30:295–304
120. Krishnan A, Giuliani A, Zbilut JP, Tomita M (2008) Implications from a network-based topological analysis of ubiquitin unfolding simulations. *PLoS ONE* 3:e2149
121. Krishnan A, Zbilut JP, Tomita M, Giuliani A (2008) Proteins as networks: usefulness of graph theory in protein science. *Curr Protein Pept Sci* 9:28–38
122. Krishnan A, Giuliani A, Zbilut JP, Tomita M (2007) Network scaling invariants help to elucidate basic topological principles of proteins. *J Proteome Res* 6:3924–3934
123. Krishnan A, Giuliani A, Tomita M (2007) Indeterminacy of reverse engineering of Gene Regulatory Networks: the curse of gene elasticity. *PLoS ONE* 2:e562
124. Palumbo MC, Colosimo A, Giuliani A, Farina L (2007) Essentiality is an emergent property of metabolic network wiring. *FEBS Lett* 581:2485–2489
125. Tun K, Dhar PK, Palumbo MC, Giuliani A (2006) Metabolic pathways variability and sequence/networks comparisons. *Bio Med Chem* 7:24
126. Zbilut JP, Giuliani A, Colosimo A, Mitchell JC, Colafranceschi M, Marwan N, Webber CL Jr, Uversky VN (2004) Charge and hydrophobicity patterning along the sequence predicts the folding mechanism and aggregation of proteins: a computational approach. *J Proteome Res* 3:1243–1253
127. Agüero-Chapin G, Gonzalez-Diaz H, Molina R, Varona-Santos J, Uriarte E, Gonzalez-Diaz Y (2006) Novel 2D maps and coupling numbers for protein sequences. The first QSAR study of polygalacturonases: isolation and prediction of a novel sequence from *Psidium guajava* L. *FEBS Lett* 580:723–730
128. González-Díaz H, Prado-Prado F, Ubeira FM (2008) Predicting antimicrobial drugs and targets with the MARCH-INSIDE approach. *Curr Top Med Chem* 8:1676–1690
129. Chou KC, Chen NY (1977) The biological functions of low-frequency phonons. *Sci Sinica* 20:447–457
130. Chou KC, Chen NY, Forsen S (1981) The biological functions of low-frequency phonons. 2. Cooperative effects. *Sci Sinica* 18:126–132
131. Chou KC (1983) Low-frequency vibrations of helical structures in protein molecules. *Biochem J* 209:573–580
132. Chou KC (1983) Identification of low-frequency modes in protein molecules. *Biochem J* 215:465–469
133. Chou KC (1984) Biological functions of low-frequency vibrations (phonons). III. Helical structures and microenvironment. *Biophys J* 45:881–889
134. Chou KC (1984) The biological functions of low-frequency vibrations (phonons). 4. Resonance effects and allosteric transition. *Biophys Chem* 20:61–71
135. Chou KC (1984) Low-frequency vibrations of DNA molecules. *Biochem J* 221:27–31
136. Chou KC (1985) Low-frequency motions in protein molecules. Beta-sheet and beta-barrel. *Biophys J* 48:289–297
137. Chou KC (1987) The biological functions of low-frequency vibrations (phonons). VI. A possible dynamic mechanism of allosteric transition in antibody molecules. *Biopolymers* 26:285–295

138. Chou KC, Mao B (1988) Collective motion in DNA and its role in drug intercalation. *Biopolymers* 27:1795–1815
139. Chou KC (1989) Low-frequency resonance and cooperativity of hemoglobin. *Trends Biochem Sci* 14:212–213
140. Chou KC, Maggiora GM, Mao B (1989) Quasi-continuum models of twist-like and accordion-like low-frequency motions in DNA. *Biophys J* 56:295–305
141. Martel P (1992) Biophysical aspects of neutron scattering from vibrational modes of proteins. *Prog Biophys Mol Biol* 57:129–179
142. Chou KC, Zhang CT, Maggiora GM (1994) Solitary wave dynamics as a mechanism for explaining the internal motion during microtubule growth. *Biopolymers* 34:143–153
143. Sinkala Z (2006) Soliton/exciton transport in proteins. *J Theor Biol* 241:919–927
144. Chou KC (1988) Low-frequency collective motion in biomacromolecules and its biological functions. *Biophys Chem* 30:3–48
145. Chou JJ, Li S, Klee CB, Bax A (2001) Solution structure of Ca^{2+} -calmodulin reveals flexible hand-like properties of its domains. *Nat Struct Biol* 8:990–997
146. Gordon G (2007) Designed electromagnetic pulsed therapy: clinical applications. *J Cell Physiol* 212:579–582
147. Gordon G (2008) Extrinsic electromagnetic fields, low frequency (phonon) vibrations, and control of cell function: a non-linear resonance system. *J Biomed Sci Eng* 1:152–156
148. McCammon JA, Gelin BR, Karplus M (1977) Dynamics of folded proteins. *Nature* 267:585–590
149. Karplus M, McCammon JA (2002) Molecular dynamics simulations of biomolecules. *Nat Struct Biol* 9:646–652
150. McCammon JA, Karplus M (1977) Internal motions of antibody molecules. *Nature* 268:765–766
151. Navarro E, Tejero R, Fenude E, Celda B (2001) Solution NMR structure of a D, L-alternating oligonucleotide as a model of beta-helix. *Biopolymers* 59:110–119
152. Navarro E, Fenude E, Celda B (2004) Conformational and structural analysis of the equilibrium between single- and double-strand beta-helix of a D, L-alternating oligonucleotide. *Biopolymers* 73:229–241
153. Navarro E, Fenude E, Celda B (2002) Solution structure of a D, L-alternating oligonucleotide as a model of double-stranded antiparallel beta-helix. *Biopolymers* 64:198–209
154. Woodcock S, Mornon JP, Henrissat B (1992) Detection of secondary structure elements in proteins by hydrophobic cluster analysis. *Protein Eng* 5:629–635
155. Randic M (2004) 2-D graphical representation of proteins based on virtual genetic code. *SAR QSAR Environ Res* 15:147–157
156. Randic M, Zupan J, Vikić-Topić D (2007) On representation of proteins by star-like graphs. *J Mol Graph Model* 26:290–305
157. Randic M (2006) Quantitative characterizations of proteome: dependence on the number of proteins considered. *J Proteome Res* 5:1575–1579
158. Zupan J, Randic M (2005) Algorithm for coding DNA sequences into “spectrum-like” and “zigzag” representations. *J Chem Inf Model* 45:309–313
159. Randic M, Lers N, Vukicević D, Plavšić D, Gute BD, Basak SC (2005) Canonical labeling of proteome maps. *J Proteome Res* 4:1347–1352
160. Randic M, Estrada E (2005) Order from chaos: observing hormesis at the proteome level. *J Proteome Res* 4:2133–2136
161. Randic M, Lers N, Plavšić D, Basak SC (2004) On invariants of a 2-D proteome map derived from neighborhood graphs. *J Proteome Res* 3:778–785
162. Randic M, Nović M, Vracko M (2002) On characterization of dose variations of 2-D proteomics maps by matrix invariants. *J Proteome Res* 1:217–226
163. Liao B, Tan M, Ding K (2005) A 4D representation of DNA sequences and its application. *Chem Phys Lett* 402:380–383
164. Liao B (2005) A 2D graphical representation of DNA sequence. *Chem Phys Lett* 401:196–199
165. Hua S, Sun Z (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* 17:721–728
166. Chou KC (2002) Prediction of protein signal sequences. *Curr Protein Pept Sci* 3:615–622
167. Santana L, Uriarte E, González-Díaz H, Zagotto G, Soto-Otero R, Méndez-Alvarez E (2006) A QSAR model for in silico screening of MAO-A inhibitors. Prediction, synthesis, and biological assay of novel coumarins. *J Med Chem* 49:1149–1156
168. González-Díaz H, Agüero-Chapin G, Varona J, Molina R, De-logu G, Santana L, Uriarte E, Gianni P (2007) 2D-RNA-coupling numbers: a new computational chemistry approach to link secondary structure topology with biological function. *J Comput Chem* 28:1049–1056
169. Kutner MH, Nachtsheim CJ, Neter J, Li W (2005) Standardized multiple regression model. In: Kutner MH, Nachtsheim CJ, Neter J, Li W (eds) *Applied linear statistical models*. 5th edn. McGraw Hill, New York, pp 271–277
170. Froimowitz M (1993) HyperChem: a software package for computational chemistry and molecular modeling. *BioTechniques* 14:1010–1013
171. HyperChem (TM) (2002) Hypercube, Inc., Gainesville, Florida, USA
172. Liu Y, Beveridge DL (2002) Exploratory studies of ab initio protein structure prediction: multiple copy simulated annealing, AMBER energy functions, and a generalized born/solvent accessibility solvation model. *Proteins* 46:128–146
173. Dea-Ayuela MA, Bolás-Fernández F (2005) Two-dimensional electrophoresis and mass spectrometry for the identification of species-specific *Trichinella* antigens. *Vet Parasitol* 132:43–49
174. Gharahdaghi F, Weinberg CR, Meagher DA, Imai BS, Mische SM (1999) Mass spectrometric identification of proteins from silver-stained polyacrylamide gel: a method for the removal of silver ions to enhance sensitivity. *Electrophoresis* 20:601–605
175. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:389–402
176. Marchler-Bauer A, Bryant SH (2004) CD-Search: protein domain annotations on the fly. *Nucleic Acids Res* 32:W327–W331
177. Jones CE, Baumann U, Brown AL (2005) Automated methods of predicting the function of biological sequences using GO and BLAST. *BMC Bioinformatics* 6:272
178. Zehetner G (2003) OntoBlast function: from sequence similarities directly to potential functional annotations by ontology terms. *Nucleic Acids Res* 31:3799–3803
179. Yang AS (2002) Structure-dependent sequence alignment for remotely related proteins. *Bioinformatics* 18:1658–1665
180. Lee C, Grasso C, Sharlow MF (2002) Multiple sequence alignment using partial order graphs. *Bioinformatics* 18:452–464
181. Jacchieri SG (2000) Mining combinatorial data in protein sequences and structures. *Mol Divers* 5:145–152
182. Ghosh P, Thanadath M, Bagchi MC (2006) On an aspect of calculated molecular descriptors in QSAR studies of quinolone antibacterials. *Mol Divers* 10:415–427
183. Gonzalez MP, Helguera AM, Collado IG (2006) A topological substructural molecular design to predict soil sorption coefficients for pesticides. *Mol Divers* 10:109–118
184. Milicević A, Nikolić S, Trinajstić N (2004) On reformulated Zagreb indices. *Mol Divers* 8:393–399
185. Torrens F (2004) Valence topological charge-transfer indices for dipole moments. *Mol Divers* 8:365–370

186. Van Waterbeemd H (1995) Chemometric methods in molecular design. Wiley, New York
187. Chou KC, Zhang CT (1995) Prediction of protein structural classes. *Crit Rev Biochem Mol Biol* 30:275–349
188. Chou KC, Shen HB (2008) Cell-PLoc: a package of web-servers for predicting subcellular localization of proteins in various organisms. *Nat Protoc* 3:153–162
189. Chou KC, Shen HB (2007) Recent progress in protein subcellular location prediction. *Anal Biochem* 370:1–16
190. Chou KC, Shen HB (2007) Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *J Proteome Res* 6:1728–1734
191. Li FM, Li QZ (2008) Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. *Protein Pept Lett* 15:612–616
192. Marrero-Ponce Y, Medina-Marrero R, Castillo-Garit JA, Romero-Zaldivar V, Torrens F, Castro EA (2005) 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* 13:3003–3015
193. Agüero-Chapín G, González-Díaz H, de la Riva G, Rodríguez E, Sánchez-Rodríguez A, Podda G, Vazquez-Padrón RI (2008) MMM-QSAR recognition of ribonucleases without alignment: comparison with HMM model and isolation from *Schizosaccharomyces pombe*, prediction, and experimental assay of a new sequence. *J Chem Inf Mod* 48:434–448
194. Dea-Ayuela MA, Perez-Castillo Y, Meneses-Marcel A, Ubeira FM, Bolas-Fernandez F, Chou KC, Gonzalez-Diaz H (2008) HP-Lattice QSAR for dynein proteins: experimental proteomics (2D-electrophoresis, mass spectrometry) and theoretic study of a *Leishmania infantum* sequence. *Bioorg Med Chem* 16:7770–7776
195. Lei Z, Elmer AM, Watson BS, Dixon RA, Mendes PJ, Sumner LW (2005) A two-dimensional electrophoresis proteomic reference map and systematic identification of 1367 proteins from a cell suspension culture of the model legume *Medicago truncatula*. *Mol Cell Proteomics* 4:1812–1825
196. Giddings MC, Shah AA, Gesteland R, Moore B (2003) Genome-based peptide fingerprint scanning. *Proc Natl Acad Sci USA* 100:20–25
197. Arakaki T, Le Trong I, Phizicky E, Quartley E, DeTitta G, Luft J, Lauricella A, Anderson L, Kalyuzhniy O, Worthey E, Myler PJ, Kim D, Baker D, Hol WG, Merritt EA (2006) Structure of Lmaj006129AAA, a hypothetical protein from *Leishmania major*. *Acta Crystallograph Sect F Struct Biol Cryst Commun* 62:175–179
198. Sternberg MJ, King RD, Lewis RA, Muggleton S (1994) Application of machine learning to structural molecular biology. *Philos Trans R Soc Lond B Biol Sci* 344:365–371
199. Han L, Cui J, Lin H, Ji Z, Cao Z, Li Y, Chen Y (2006) Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity. *Proteomics* 6:4023–4037
200. González-Díaz H, Agüero-Chapin G, Varona-Santos J, Molina R, de la Riva G, Uriarte E (2005) 2D RNA-QSAR: assigning ACC oxidase family membership with stochastic molecular descriptors; isolation and prediction of a sequence from *Psidium guajava* L. *Bioorg Med Chem Lett* 15:2932–2937
201. Agüero-Chapin G, Antunes A, Ubeira FM, Chou KC, Gonzalez-Diaz H (2008) Comparative study of topological indices of macro/supramolecular RNA complex networks. *J Chem Inf Model* 48:2265–2277
202. González-Díaz H, Pérez-Castillo Y, Podda G, Uriarte E (2007) Computational chemistry comparison of stable/nonstable protein mutants classification models based on 3D and topological indices. *J Comput Chem* 28:1990–1995
203. Puslednik L, Serb JM (2008) Molecular phylogenetics of the Pectinidae (Mollusca: Bivalvia) and effect of increased taxon sampling and outgroup selection on tree topology. *Mol Phylogenet Evol* 48:1178–1188