Study of peptide fingerprints of parasite proteins and drug–DNA interactions with Markov-Mean-Energy invariants of biopolymer molecular-dynamic lattice networks

Lázaro Guillermo Pérez-Montoto [a,b], María Auxiliadora Dea-Ayuela [c], Francisco J. Prado-Prado [a,b], Francisco Bolas-Fernández [d], Florencio M. Ubeira [a], Humberto González-Díaz [a,*]

[a] Department of Microbiology and Parasitology, Faculty of Pharmacy, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain
[b] Department of Organic Chemistry, Faculty of Pharmacy, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain
[c] Departamento de Atención Sanitaria, Salud Pública y Sanidad Animal, Facultad CC Experimentales y de La Salud, Universidad CEU Cardenal Herrera, 46113 Moncada (Valencia), Spain
[d] Department of Parasitology, Faculty of Pharmacy, Complutense University, 28040 Madrid, Spain

ARTICLE INFO

ABSTRACT

Since the advent of Molecular Dynamics (MD) in biopolymers science with the study by Karplus et al. on protein dynamics, MD has become the by foremost well established, computational technique to investigate structure and function of biomolecules and their respective complexes and interactions. The analysis of the MD trajectories (MDTs) remains, however, the greatest challenge and requires a great deal of insight, experience, and effort. Here, we introduce a new class of invariants for MDTs based on the spatial distribution of Mean-Energy values $\xi_k(L)$ on a 2D Euclidean space representation of the MDTs. The procedure forces one MD trajectory to fold into a 2D Cartesian coordinates system using a step-by-step procedure driven by simple rules. The $\xi_k(L)$ values are invariants of a Markov matrix ($^1\mathbf{\Pi}$), which describes the probabilities of transition between two states in the new 2D space; which is associated to a graph representation of MDTs similar to the lattice networks (LNs) of DNA and protein sequences. We also introduce a new algorithm to perform phylogenetic analysis of peptides based on MDTs instead of the sequence of the polypeptide. In a first experiment, we illustrate this algorithm for 35 peptides present on the Peptide Mass Fingerprint (PMF) of a new protein of *Leishmania infantum* studied in this work. We report, by the first time, 2D Electrophoresis isolation, MALDI TOF Mass Spectroscopy characterization, and MASCOT search results for this PMF. In a second experiment, we construct the LNs for 422 MDTs obtained in DNA–Drug Docking simulations of the interaction of 57 anticancer furocoumarins with a DNA oligonucleotide. We calculated the respective $\xi_k(L)$ values for all these LNs and used them as inputs to train a new classifier with Accuracy = 85.44% and 84.91% in training and validation respectively. The new model can be used as scoring function to guide DNA–Drug Docking studies in drug design of new coumarins for PUVA therapy. The new phylogenetics analysis algorithms encode information different from sequence similarity and may be used to analyze MDTs obtained in Docking or modeling experiments for any classes of biopolymers. The work opens new perspective on the analysis and applications of MD in polymer sciences.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction

Computational approaches can timely provide very useful information and insights for both basic proteome research and drug design. Many line evidences of evidences such as structural bioinformatics [1], molecular docking [2], molecular packing [3], pharmacophore modeling [4], Monte Carlo simulated annealing approach [5], diffusion-controlled reaction simulation. In addition, Quantitative Structure–Activity Relationships (QSARs) [6,7], protein sub-cellular location prediction [8–10], protein structural class prediction [11], identification of membrane proteins [12], identification of enzymes and their functional classes [13,14], identification of GPCR [15], identification of proteases, protein cleavage site prediction [16], and signal peptide prediction [17] have indicated that they are widely welcome by science community.

* Corresponding author. Department of Microbiology and Parasitology, Faculty of Pharmacy, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain.
E-mail addresses: gonzalezdiazh@yahoo.es, humberto.gonzalez@usc.es (H. González-Díaz).

In this context, after a pioneer paper entitled 'The Biological Functions of Low-Frequency Phonons' [18] published in 1977, a series of investigations into biopolymers from dynamic point of view have been stimulated. These studies have suggested 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. [19–32] and a comprehensive review [33]). This kind of inferences has been later observed by NMR [34], and been further used for medical treatments [35,36]. In view of this, to understand really the interaction mechanism of drugs with proteins or DNA, we should consider not only the static structures concerned but also the dynamical information obtained by simulating their interactions through a dynamic process. The present study was attempted to address this problem from the angle of Molecular Dynamics (MD). In fact, since the advent of MD with the work of Karplus et al. [37–42], MD has become a computational technique to investigate structure and function of biomolecules and their respective complexes and interactions. It is also of high relevance taking into account that the previous structure of the polymeric double helix of DNA as well as the non-covalent binding (in dark) between DNA and drug has a strong influence on the subsequent photoreaction and therefore on their biological activity [43,44]. Consequently, MD studies of the biopolymers including polypeptides or polynucleotide DNA–Drug complexes are of the major relevance too [32,45]. In general, the analysis of the MD trajectories (MDTs) resulting from the integration of the equations of motions in MD remains, however, the greatest challenge and requires a great deal of insight, experience, and effort. In a recent and very important work, Hamacher [46] proposed a new, theoretical sound, and versatile analysis procedure that provide scientists with a semi-quantitative invariant measures to compare various scenarios of their respective simulations.

On the other side, using graphic approaches to study biological systems can provide useful insights. As indicated by many previous studies graph have been used on a series of important biological topics, such as enzyme-catalyzed reactions [47–49], protein folding kinetics [50], inhibition kinetics of processive nucleic acid polymerases and nucleases [51], analysis of codon usage [52], base frequencies in the anti-sense strands [53], analysis of DNA sequence [54]. Moreover, graphical methods have been introduced for QSAR study [55] as well as utilized to deal with complicated network systems [56,57]. Recently, the "cellular automaton image" [58] has also been applied to study hepatitis B viral infections [59], HBV virus gene miss-sense mutation [60], and visual analysis of SARS-CoV [61], as well as representing complicated biological sequences [62] and helping to identify protein attributes [63].

In this sense, several authors have used pseudo-folding lattice Hydrophobicity-Polarity (HP) models to simulate polymer folding making simulations to optimize the lattice structure and resemble real folding [64–71]. However, we can choose notably simpler polymer chain pseudo-folding rules to avoid optimization procedures and speed up notably the construction of the lattice. In this sense, useful graph representations of DNA, RNA and/or protein sequences have been introduced by Gates [72], Nandy [73], Leong [74], Randic et al. [75] based on 2D coordinate systems. We call these graph representations as polymer sequence pseudo-folding Lattice Networks (LNs) because they look like lattice structures and in fact we are forcing a sequence to fold in a way that not necessarily occurs in nature. In general, these LNs (as for other polymer graph representations) can be numerically characterized with Topological Indices (TIs), see for instance the previous paper series published by our group on Polymers [76–79]. These TIs describe the distribution of amino acids or nucleotides along the polymeric chain but also encode higher-order information or other type of

information on polymer mixtures. Thus lattice pseudo-folding TIs can be used in protein Quantitative Structure–Property Relationships (QSARs) [7,81] to connect polymeric structure with physico-chemical or biological properties. Our group has used the approach called MARCH-INSIDE to calculate these TIs of pseudo-folding lattice-like networks to predict diverse protein or DNA/RNA functions. For instance, we have used stochastic pseudo-folding spectral moments to predict Ribonucleases [82] and Dyneins [83]. The MARCH-INSIDE pseudo-folding TIs can be calculated when we sum the respective indices for each node of the graph. All the above-mentioned values were used recently to predict microbacterial promoters and compare entropies, spectral moments, and pseudo-folding electrostatic potentials [84]. The readers may see three recent reviews discussing the applications ranging from graph of small molecules to graph or network representation of protein sequences and 3D structure, DNA sequences, RNA secondary structure, or human blood proteome mass spectroscopy outcomes [7,81,85].

In any case, if we understand sequence as a type of input data we have not to limit the applications of the pseudo-folding lattice network method to proteins, DNA or RNA sequences. Elaborating this line of thinking we have proposed pseudo-folding lattice network representations of Mass Spectroscopy outcomes typical of blood Proteome samples containing many proteins. For instance we have constructed lattice network representations for mass spectroscopy results obtained from blood proteome samples typical of drugs causing cardiotoxicity [86]. After calculation of the sum of the TIs of each sample we used them to seek a new type of classifier. The model connects TIs values of the Mass Spectra of the blood proteome with the probability of appearance of drug cardiotoxicity [79,87]. We have used these lattice network TIs also to predict human prostate cancer [88].

The success of this strategy encouraged us to consider other classes of sequence data and solve different problems. For instance, the MDTs referred in previous paragraphs are time series obtained from simulation runs that constitute another type of sequential data. Considering that the Mean Values of a Markov Chain associated to LN are also sequence invariants we decided to explore here the use of these indices to describe MDTs. In the present paper, we adapted LN representations for the study of MDTs obtained in both DNA–Drug Docking and Peptide structure optimization experiments. In this sense, we report two different experiments: in one we report a new phylogenetic analysis for MDTs of Peptides (Experiment 1) in the other we obtain a new scoring function for DNA–Drug Docking studies (Experiment 2).

2. Materials and methods

2.1. General description of experiments

In Experiment 1: we shall deal with the following questions: (a) adaptation of LN to represent MDTs obtained in peptide optimization procedures; (b) calculation of a new class of TIs for peptide MDT networks; and (c) introduction of a new phylogenetic approach to compare the MDTs of different peptides found in the Peptide Mass Fingerprint (PMF) of protein. For it, we are going to use as example a real experiment we describe here by the first time. Here we isolate with 2D gel Electrophoresis and characterize with MALDI TPF MS all the peptides found on the PMF of a protein expressed on *Leishmania infantum*. The study of PMFs of new proteins may become an interesting source to fish new peptides with potential use as drug, in vaccine design, or as disease biomarkers. In particular, *Leishmania* parasitic species are the causal agents of Leishmaniosis one of the most important parasitic diseases [89,90]. The toxicity and inefficacy of actual organic drugs

against Leishmaniosis justify research projects to find new drugs or drug molecular targets in *Leishmania* species including *L. infantum* (*L. infantum*) and *Leishmania major* (*L. major*), both important pathogens [83,91–93]. In this sense the bioinformatics studies of *Leishmania* gene and proteins become a significant goal [94,95].

One possibility to accomplish the former goal is the use of proteome research techniques. For instance, in proteome research authors often use a combination of 2D Electrophoresis (2DE) and Mass Spectroscopy (MS) to isolate and characterize new sequences from biological samples [96]. Obtaining the PMF of the protein is a very useful procedure in this sense [97]. 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 [98,99]. It means that, for instance, MASCOT may give the collection of MS signals and the corresponding sequence of peptides present in known proteins that match with our MS input. In order to rank and select the best protein/peptide candidates MASCOT use the Mowse score (M_s) [100]. If a template protein in the database has a high M_s (>51), this protein has a PMF very similar to the PMF of our query proteins and we can detect a high sequence homology and perform function annotation. However, there is still another situation that often appears in proteome research and do not coincide exactly with the two situations above-mentioned in the first paragraph. We refer to the 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 M_s that is not sufficiently high to accurately annotate the query protein (>40). In an excellent work have been reported an alternative to M_s and discussed the limits of accurate scoring [101]. 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 methods. Anyhow, from these facts we can conclude that the method used to compare different peptides in this search is very important.

Bioinformatics methods based on Sequence alignment and similarity measures are very useful to perform sequence function annotation. Some authors have referred however that an alignment procedure may fail in cases of low sequence homology between the query and the template sequences deposited in the database. Alignment techniques are also useless if there is high query-template homology but we do not know the function of the template sequence deposited in the database [102]. On the other hand, some authors mentioned in the previous paragraphs have introduced 2D or higher dimension graph representations of sequences prior to the calculation of TIs. These representations are associated to algebraic structures that have been extended also to genetic codes [103–105]. This constitutes an important step in order to uncover useful higher-order information not encoded by 1D sequence parameters [73,106–118]. In any case, we can use either the sequence directly or the graph parameter to develop phylogenetic trees in order to compare different peptides. Phylogenetic analysis often relays on tree graph construction. Applications of graphs and networks in phylogenetics are too broad a topic for detailed treatment here. The reader can consult reviews and compilations on this topic for an overview of this area [120]. Phylogenetics is commonly used to predict which amino acid residues are critical for the function of a given protein [121]. However, such approaches do have inherent limitations, such as the requirement for the identification of multiple homologs of the protein under consideration. Thibert and Bredesen [122] reported a study of cancer proteins in an extensive human PIN constructed by

computational methods. They compared a couple of phylogenetic approaches to several different network-based methods. Another interesting direction is the use of TIs of the DNA, RNA and/or protein graph representations described above to construct phylogenetic trees in an alignment-independent way. For instance, Zupan and Randič [123] studied Spectrum-like and Zig–Zag representations of the beta-globin gene for different species and also obtained phylogenetic trees without alignment. In another paper, Liao proposed a 2D graphical representation of a DNA sequence [124]. Liao et al. [125] used this representation as a basis to compute the similarities between 11 mitochondrial sequences belonging to different species and used the elements of the similarity matrix to construct the phylogenic tree. Among all above-mentioned, Liao, Randic, Basak, Vackro, Nandy and Wang [108,118] associated a DNA sequence having n bases with $n \times n$ non-negative real symmetric matrix A with elements a_{ij} and use its leading eigenvalue to characterize the DNA sequence in phylogenetic studies. These matrices have been derived from 2DD representations and different TIs calculated [126]. On the other hand, Zhang et al. [127] very recently introduced TIs referred to as Zinv for 3DD curves and used them to analyze the phylogenetic relationships for the seven HA (H5N1) sequences of avian influenza virus. The equations used in these methods to calculate the indices Inv(A) and Zinv(A) as well as the phylogenetic distances between two peptides p and q are given as follows [127], see Equations (1)–(4). In closing, in the Experiment 2 we recognize the necessity of new peptide phylogenetic methods, the success of TIs of LN to encode sequence information, and the importance of MDT studies. Consequently, based on these facts we introduce a new phylogenetic method based on $\xi_k(L)$ values.

$$\mathrm{Inv}(A) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\sum_{j=1}^{n} a_{ij} \right) \tag{1}$$

$$\mathrm{Zinv}(A) = \frac{1}{n - \left(\frac{1}{n}\right)} \sum_{i=1}^{n} \left(\sum_{j=1}^{n} a_{ij} \right) \tag{2}$$

$$D_{\mathrm{pq}}(\mathrm{Inv}) = \sqrt{\sum_{k=1}^{m} [\mathrm{Inv}_k(\mathrm{p}) - \mathrm{Inv}_k(\mathrm{q})]^2} \tag{3}$$

$$D_{\mathrm{pq}}(\mathrm{Zinv}) = \sqrt{\sum_{k=1}^{m} [\mathrm{Zinv}_k(\mathrm{p}) - \mathrm{Zinv}_k(\mathrm{q})]^2} \tag{4}$$

On the other hand, in Experiment 2: we shall deal with the following questions: (a) generalization of LN to represent DNA–Drug MDTs, (b) calculation a new class of TIs for MDT networks based on the MARCH-INSIDE approach, and (c) development of a new scoring function for DNA–Drug Docking. The TIs introduced here are the Mean values of a Markov Chain associated to the LN of an MDT. These mean values may be average values of atomic electronegativities, amino acid electrostatic potentials, intensity of Mass/charge signals in Mass Spectroscopy or other parameters. The type of parameter obtained depends on the type of systems under study (molecules, proteins, Mass Spectra), the parts of the system (atoms, amino acids, MS signals), and the property used to describe these parts (electronegativities, electrostatic charge, signal intensity). For instance, in other works, we used Markov chain pseudo-folding electrostatic potentials to found models that predict Polygalacturonases [117] or human colon and breast cancer biomarkers [129]. In this experiment we found a model that can be used as scoring function to evaluate DNA–Drug Docking search. These models belong to a general class of

methods known as QSARs are devoted to unravel structural and physicochemical requirements for biological activity in a great variety of compounds [130]. The classic QSAR studies connect information of the chemical structure of the molecule, expressed by means of numbers, with the biological activity [55]. However, QSAR-like procedures are not restricted to drugs and biological activity but other systems and properties, such as proteins or DNA/RNA may be predicted [132–135]. One special class of indices used in QSAR are the TIs of molecular graphs; which indicates the presence of vertices or nodes (atoms) and connections or edges between nodes (chemical bonds) [136–139]. Nevertheless, we can use TIs of different types of graph representations or networks may be used. In these networks, amino acids, nucleotides, enzymes, microorganisms, cerebral cortex regions, web pages, social groups, etc., may play the role of nodes and electrostatic interactions, mutations, metabolic reactions, host–parasite relationships, brain region co-activations, links, diseases propagations, etc. may play the role of edges [7,140–145]. Many authors prefer to use the term Quantitative Structure-Binding affinity Relationship (QSBR) when one use QSAR-like procedures to predict drug–target binding affinity and 3D structural information [146]. In any case, both approaches QSAR and QSBR diverge in some degree on the type of measure (activity or binding) and sometimes on how detailed we need to know the chemical structure (2D or 3D) but both use essentially the same algorithm. In addition to predicting drug activity we can use 3D drug-target QSAR/QSBR models as scoring function to guide the search of optimal drug–target interaction geometries in drug–target Docking studies [147–149]. Almost all QSAR/QSBR or other types of Docking-scoring functions are aimed to predict protein-drug interactions. For instance, Wang et al. [150] reported a comparative study of eleven whereas Ferrara et al. [151] studied nine different Docking-scoring functions all for Protein–drug interactions. Conversely, DNA–Drug and RNA–Drug Docking are generally less investigated. In particular, we did not found a QSBR scoring function for DNA–Furocoumarin Docking. The furocoumarins are a class of natural or synthetic compounds with very interesting pharmacological properties [152]. Commonly used in the treatment of skin diseases such as psoriasis and mycosis fungoides [153]. This treatment called PUVA consists in a therapy that combines the use of both chemicals and long-wave ultraviolet light (UV-A) [154]. The molecular base of PUVA is connected with the highly specific photo damage in DNA of epidermal cells. This damage interferes with the DNA replication, producing an inhibition of DNA synthesis which reduces or blocks the cell duplication [155]. Although the lineal furocoumarins (psoralens) are able to form the three adduct types, the geometry of the angular ones (angelicins) only allows them to form monoadducts with the DNA. On the other hand, it is well known that the side effects observed in PUVA therapy, such as skin phototoxicity and risk of skin cancer are strictly connected with the bi-functional lesions in DNA [156]. These facts points to the stability DNA–Drug complex as a central factor in the activity of anticancer drugs in general including furocoumarins. In closing, in the Experiment 1 we recognize the necessity of new scoring functions for DNA–Drug Docking methods, the importance of furocoumarins in PUVA therapy, the success of TIs of LN to encode sequence information, and the importance of MDT studies. Consequently, based on these facts we introduce a new DNA–Drug scoring function for furocoumarins based on $\xi_k(L)$ values.

2.2. Markov-Mean values for 2D lattice representation of MDTs

The MARCH-INSIDE approach is extended here to the study of LN representations for MDTs obtained in DNA–Drug Docking studies. In Fig. 1, we illustrate an example of LN for an MDT. The key of the method we propose is the regrouping into four groups of the Energy values ^dE_s obtained for different steps (s) of one MD
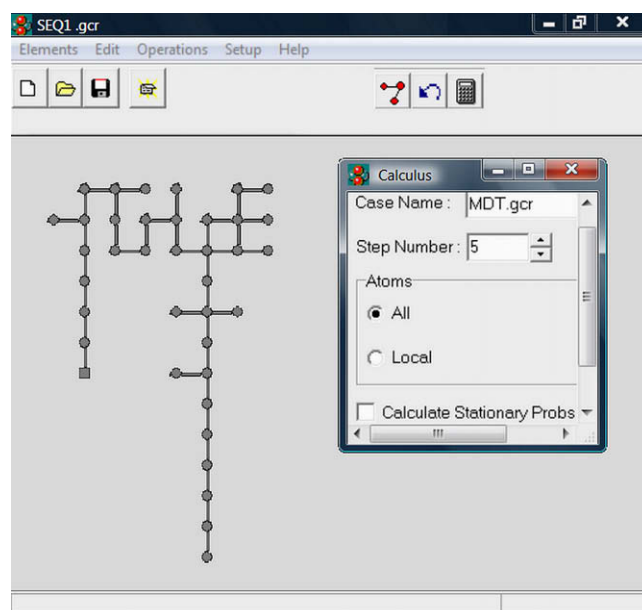
Fig. 1. Snapshot of a MARCH-INSIDE view for an LN representation of an MDT.

trajectory after docking one drug with DNA. These four groups characterize the deviation of the energy value ^dE_s from the average energy of the same MDT at different steps (MD-average); or the deviation from average energy values in same step for other MDTs (Step-average). First, the values of energy for an MDT are placed in a Cartesian 2D space starting with the first energy value at the coordinates (0, 0). The coordinates of the successive energy values are calculated as follows, in a similar manner than it can be used for a DNA or proteins [117]:

(a) Increases in +1 the x axes; if $^dE_s >$ MD-average and $^dE_s >$ Step-average (upwards-step) or:
(b) Decreases in −1 the x axes; if $^dE_s >$ MD-average and $^dE_s <$ Step-average (rightwards-step) or:
(c) Increases in +1 the y axes; if $^dE_s <$ MD-average and $^dE_s >$ Step-average (leftwards-step) or:
(d) Decreases in −1 the y axes; if $^dE_s <$ MD-average and $^dE_s <$ Step-average (downwards-step).

Secondly, the method uses the matrix $^1\mathbf{\Pi}$, which is a squared matrix to characterize the MDT embedded into the LN. Please, note that the number of nodes (n) in the graph may be equal or even smaller than the number of steps given to obtain the MD profile. The same happens for amino acids or DNA bases in the polymeric chain. Accordingly, the matrix $^1\mathbf{\Pi}$ contains the probabilities $^1p_{ij}$ to reach a node n_i moving throughout a walk of length $k = 1$ from other node n_j [83,129]:

$$p_{ij} = \frac{\left(\frac{1}{D_{0j}}\right) \cdot \left(\sum_{s \in j} {}^dE_s\right)}{\sum_{m=s}^{n} \alpha_{il} \cdot \left(\frac{1}{D_{0s}}\right) \cdot \left(\sum_{s \in j} {}^dE_s\right)} = \frac{\left(\frac{{}^dE_j}{D_{0j}}\right)}{\sum_{m=l}^{n} \alpha_{il} \cdot \left(\frac{{}^dE_l}{D_{0s}}\right)} \tag{5}$$

where, ^dE_j is the sum of all energy values of the steps ^dE_s that overlap on the same node j. The parameter α_{ij} equals to 1 if the nodes n_i and n_j are adjacent in the graph and equals to 0 otherwise. The value D_{0j} gives the geometric location of the node and represents the Euclidean distance between the node and the center of coordinates. Let be, the vector of initial probabilities $\pi_0 = [{}^0p_1, {}^0p_2,$
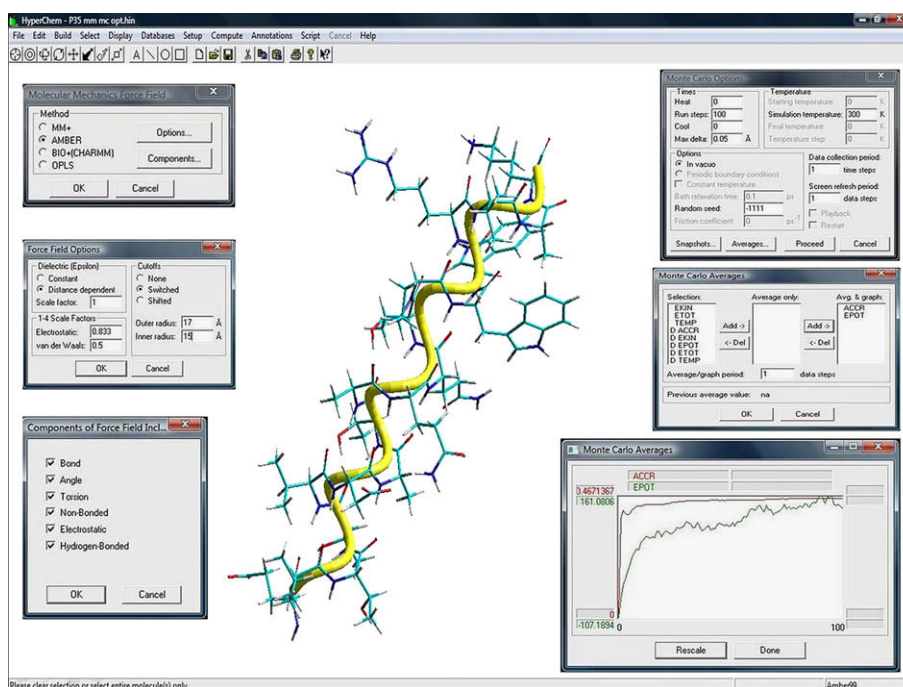
Fig. 2. Snapshot of Hyperchem's interface illustrating one peptide of the new protein.

$^0p_j,\ldots {}^0p_n]$ and the vector of node energies $\varepsilon = [^dE_1, {}^dE_2, {}^dE_j,\ldots {}^dE_n]$, the calculation of $\xi_k(L)$ values is straightforward to realize be means of Chapman–Kolgomorov equations; these indices can be interpreted as Mean-Energy values for on the 2D Euclidean space representation for MDTs:

$$\xi_k(L) = \boldsymbol{\pi}_0^t \cdot {}^k\varPi \cdot \boldsymbol{\varepsilon} = \boldsymbol{\pi}_0^t \cdot (^1\varPi)^k \cdot \boldsymbol{\varepsilon} = \sum_{i=j}^n {}^kp_{ij} \cdot {}^dE_j \qquad (6)$$

2.3. Cell culture of parasites

Promastigotes of the *Leishmania* strain LEM75 were grown in medium Schneider supplemented to a final concentration of 0.4 g/L NaHCO$_3$, 4 g/L HEPES, 100 mg/L penicillin and 100 mg/L streptomycin and 10% fetal bovine serum (Gibco), pH 6.8 and 26 °C [83].

2.4. Sample preparation

Mid-log promastigotes were recovered on the seventh day post-inoculum (p.i.) and the parasites were centrifuged at 3000 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 in 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, GmbH, Osterode, Denmark). The supernatant was dialysed overnight at 4 °C in 0.5 mM Tris–HCl buffer. Proteins were precipitated with 20% TCA (trichloroacetic acid) in acetone with 20 mM DTT for 1 h at −20 °C, added 1:1 to the homogenated. Then, the sample was centrifuged at 10,000g 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 Biosciences), 5 mM CO$_3$K$_2$, 2% IPG buffer (Amersham Biosciences) and incubated at room temperature for 30 min. Following clarification by centrifugation at room temperature (12,000g, 10 min) the supernatant was stored frozen at −20 °C [83].

2.5. 2DE experiments

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 pH 4–7, bromophenol blue) and were immediately adsorbed onto 18 cm immobilized pH 4–7 gradient (IPG) strips (Amersham Biosciences) [157]. 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, 1000 V for 1 h, 1000–2000 V for 1 h and 8000 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, and 100 mg of DTT. Then, the strips were incubated for 25 min in the same buffer but 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. 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 were sensitized with sodium acetate 0.68% (w/v) and 0.05% sodium thiosulfate for 30 min and washed with deionized water 3 times for 5 min. The gels were incubated in 0.25% (w/v) silver nitrate for 30 min. After incubation, it was rinsed with deionized water 2 times for 50 s followed by adding the developing solution, which contained 2.5% (w/v) sodium carbonate with 0.04% (v/v) formaldehyde until the desired intensity range. Development was finished by adding 1.5% (w/v) EDTA.
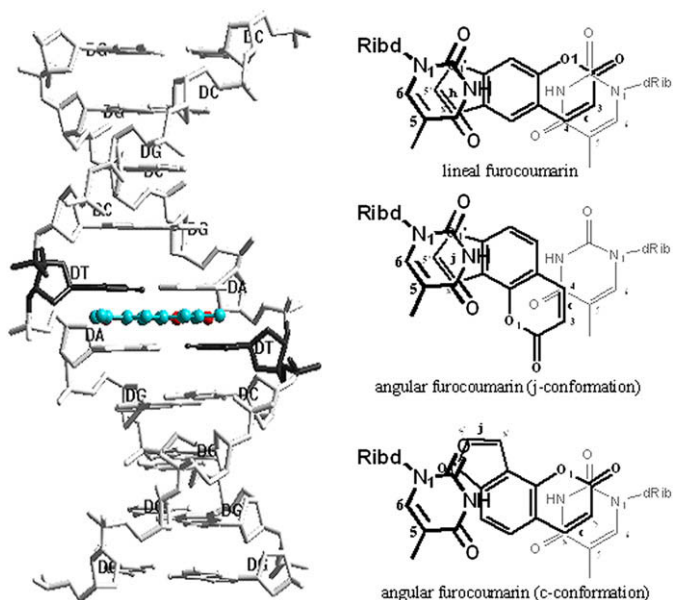
Fig. 3. DNA–Drug complex and some views of drug intercalation.

2.6. MALDI-TOF MS

Spots of interest were manually excised from silver stained 2-DE gels after being de-stained, as described by Gharahdaghi et al. [158]. Then, gel pieces were incubated with 12.5 ng/μl sequencing grade trypsin (Roche Molecular Biochemicals) in 25 mM AmBic, overnight, at 37 °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, one microliter 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 analyses 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.

2.7. MASCOT search

The PMF data, obtained from MALDI-TOF MS analyses, were used to search for protein candidates in two sequence databases: SWISS-PROT/TrEMBL non-redundant protein database (www.expasy.ch/sprot) and a complete genomic database from the related species *L. major*, namely ftp://ftp.sanger.ac.uk/pub/databases/L.major_sequences/LEISHPEP/, using MASCOT software program (www.matrixscience.com). The MASCOT search parameters were adjusted according to the MS experiment carried out and to 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; Instrument type: MALDI-TOF-TOF. We introduced the MS signals corresponding to one of the unidentified 2D electrophoresis spots (protein) into the MASCOT analysis system. The sample was recorded on this web page with the search title: Sample Set ID: 1122, AnalysisID: 1466, Maldi WellID: 17500, Spectrum ID: 7971, Path=\040519\Leishmania\New Analysis 2. The database used was Leishmania 290 703 (with 7467 sequences; and 4 469 604 residues).

2.8. MC based MD study of PMFs and DNA–Drug complex

The MDTs or energetic profiles of all the starting structure of peptides were also obtained by means of the MC method, using the HyperChem package [159,160]. In this sense, the force field AMBER94 [161] of molecular mechanics 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 Monte Carlo simulation the geometry of all the structures of peptides where optimized with this same force field. Finally, the simulation was executed in the vacuum at 300 K and 100 optimization steps obtaining MDTs with 100 potential energy $^{d}E_{j}$ ($j = 1, 2, 3,…100$) values each one. We obtained 35 MDTs for 35 peptides. In order to obtain realistic MDTs there is an additional parameter 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 acceptance ratio is a running average of the ratio of the number of accepted moves to attempted moves. Optimal values are close to 0.5. Varying the step size can have a large effect on the acceptance ratio. The step size, Δ**r**, is the maximum allowed atomic displacement used in the generation of trial configurations. The default value of **r** in HyperChem is 0.05 Å [159]. For most organic molecules, this will result in an acceptance ratio 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
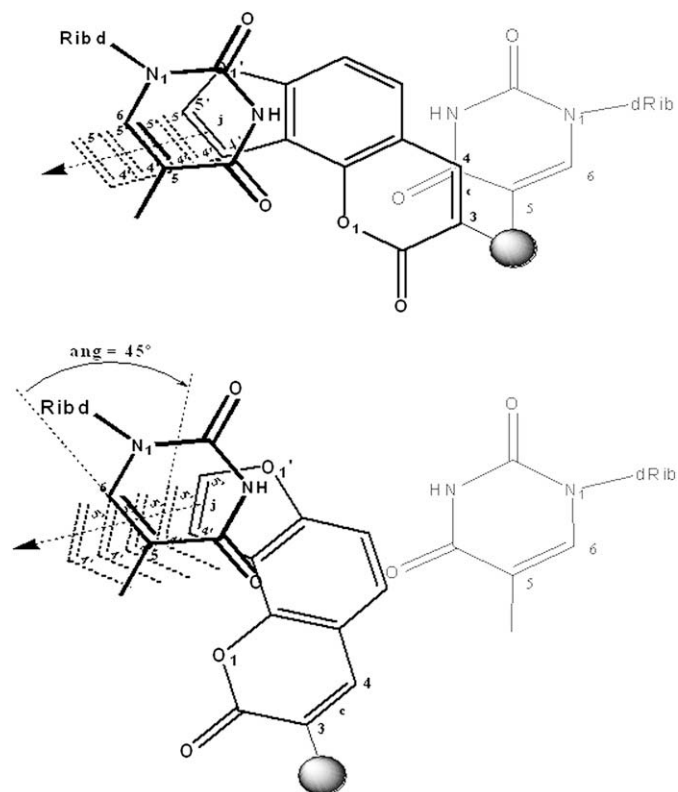
Fig. 4. Details of DNA–Drug intercalation.

Table 1
Representations for starting conformations used.

Dist.[a]	1	1	0.5	0.5	0
Ang.[a]	0°	45°	0°	45°	0°

Dist.[a]	−0.5	−0.5	−1	−1	1
Ang.[a]	0°	45°	0°	45°	0°

[a] *Dist.*: discrete distance between the geometric centers of the two double bonds (in the plane projection). The possible modular values are 0; 0.5 and 1. Positive value if the compound was moved inwards DNA pocket and negative if it was moved outwards DNA. *Ang.*: magnitude of clockwise rotation of compound (0 or 45°).

sampling. There has been little research to date on what the optimum value of the acceptance ratio should be. Most researchers tend to try for an average value around 0.5; smaller values may be appropriate when longer runs are acceptable and more extensive sampling is necessary [159].

The DNA–Drug Docking MDTs or energetic profiles of all the starting intercalation complexes were obtained also by means of the MC using the same parameters than for PMFs. We obtained 21 MDTs for psoralens and 154 MDTs for the 36 different angelicins. We also analyzed 36 averaged MDTs for each angelicin taking the average energy $^{d}E_j$ (avg) for all the initial positions of one compounds at each one of the 100 steps. All these MDTs form a total of $21 + 154 + 36 = 211$ MDTs. In addition, we analyzed other 211 MDTs (decoy trajectories) obtained as a random deviation from each one of the previous 211 MDTs.

$$^{d}E_j(\text{rnd}) = {}^{d}E_j + \text{random}(j, \max({}^{d}E_j), \min({}^{d}E_j)) \tag{7}$$

These random MDTs contain 100 energy values $^{d}E_j(\text{rnd})$ obtained with the random generator of Excel by adding a random deviation term to each $^{d}E_j$ within the max–min limits of $^{d}E_j$ for all the previous MDTs. The utility of these decoy trajectories is to test the robustness of the method to deviations of the MDTs selected. In total we studied 422 MDTs. The information about all these 422 MDTs including $\xi_k(L)$ values relevant to this work was recorded on the online Supplementary material.

2.9. Phylogenetic analysis of MDTs from PMFs

Using the vector of initial the vector of $\xi_k(L)$ values for a peptide found in the PMF of a new protein we can calculate peptide–peptide distance $D(\xi)_{pq}$ between peptides p and q. This distance may be used as alternative to the distance $D(E)_{pq}$ based directly on the energy values of the MDTs $\varepsilon = [{}^{d}E_1, {}^{d}E_2, {}^{d}E_j, \dots {}^{d}E_n]$. In principle, we can use different distance functions; here we select the Euclidean distance because of the Euclidean nature of the Cartesian space used to derive the LNs. Using the Tree Joining Cluster (TJC) analysis algorithm implemented on the software Statistica we were able to construct, visualize, and compare the phylogenetic trees based on both distances for 35 peptides found

on the PMF of the new protein. The equations to calculate both distances are:

$$D_{pq}(E) = \sqrt{\sum_{j=0}^{100}\left({}^{p}E_j - {}^{q}E_j\right)^2} \tag{8}$$

$$D_{pq}(\xi) = \sqrt{\sum_{j=0}^{5}\left({}^{p}\xi_j(L) - {}^{p}\xi_j(L)\right)^2} \tag{9}$$

Table 2
Lineal furocoumarins (psoralens) and their aza-analogues used.

Drug	Z	R_3	R_4	R_5	$R_{4'}$	$R_{5'}$	R_8	ID_{50}[a]	Ref.[b]
1	C	Me	Me	H	Me	H	H	0.34	[35]
2	C	H	H	OMe	H	H	H	0.66	[30]
3	C	H	CH_2OH	H	Me	H	OMe	0.84	[32]
4	C	Me	H	H	Me	H	OMe	0.89	[76]
5	C	H	H	H	H	H	OMe	1.00	[34]
6	C	Me	H	H	Me	H	H	1.01	[77]
7	C	H	H	H	Me	Me	H	1.26	[35]
8	C	Me	H	H	Me	H	Me	1.34	[77]
9	C	H	H	H	H	H	H	1.52	[78]
10	C	Me	H	H	Me	Me	H	1.79	[35]
11	C	H	CH_2OH	H	Me	H	H	2.32	[32]
12	C	H	Me	H	H	Me	Me	27.6	[30]
13	N	H	H	H	Me	H	–	0.13	[34]
14	N	H	Me	H	H	H	–	0.14	[34]
15	N	H	H	H	Me	Me	–	0.18	[79]
16	N	H	H	Me	Me	Me	–	0.25	[34]
17	N	Me	Me	H	Me	H	–	0.67	[34]
18	N	H	Me	H	Me	H	–	0.68	[79]
19	N	H	H	Me	Me	H	–	0.97	[34]
20	N	Me	Me	H	Me	Me	–	1.83	[79]
21	N	H	Me	H	Me	Me	–	3.66	[79]

[a] The experimental antiproliferative activity in Ehrlich Ascites tumor cells expressed as ID_{50} relative to 8-MOP.
[b] References in which the activity of compounds was reported.

2.10. Model building of DNA–Drug intercalation complexes
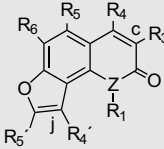
For our study we used the decanucleotide of sequence d(CCGCTAGCGG) and the software application HyperChem [160,162], a fragment of DNA with double Helix in B form and sugars in 2′ endo form. This decanucleotide sequence has been used in different studies concerning psoralens intercalation [160,163]. The structure of all the compounds selected for DNA–Drug inter-action studies were optimized using the interactive model building package of HyperChem [162]. The optimization of their geometries was carried out by the Semi-empirical Quantum Mechanics calcu-lations with method PM3 [164] using the Polak–Ribiere algorithm and the options implemented by default in the mentioned package. Thus, the minimized molecular structures were intercalated by hand approach in the DNA fragment, using the HyperChem package and taking into account the following experimentally demon-strated statements:

1. In the dark, the poly[dA–dT] poly[dA–dT] sequence in DNA is the most favorable site for intercalation since the further photoreaction takes place mainly on the 5,6 double bond of the thymine [165]. So, the optimized molecules were inserted among the thymine units in a parallel plane to the bases and, according to our decision, in a halfway position (Fig. 3, left).

2. The furocoumarins have two reactive sites, but after photore-action, different types of cycloadducts can be formed: mono (furan-side or pyrone-side) and di-adducts (the cross-link) [166]. Although psoralens are able to form all the cycloadduct types, angelicins forms only monoadducts owing to their angular molecular structure. Keeping this in mind, for each lineal molecule we modeled only one starting conformation, for which the cycloadduct formation by either one or other reactive site (furan or pyrone-side) is equally feasible from a geometric point of view. For each angular molecule we decided to model two starting conformations, one for each monoadduct formation (for the furan-side that we named as j-conformation and for the pyrone-side that we named as c-conformation).

3. The stereochemistry of the furocoumarins adducts is cis–syn [167,168]. Consequently, the molecules were oriented in such a way that the intercalation complex favors mainly the formation of cycloadducts with this stereochemistry. In the case of the furan-side, the stereochemistry syn means that

Table 3
Angular furocoumarins (angelicins) and their aza-analogues used.

Compound	Z	R_1	R_3	R_4	R_5	R_6	$R_{4'}$	$R_{5'}$	ID_{50}[a]	Ref.[b]
22	O	–	COMe	H	H	H	H	H	<0.01	[7]
23	O	–	COPh	H	H	H	H	H	<0.01	[7]
24	O	–	$CON(Et)_2$	H	H	H	H	H	<0.01	[7]
25	O	–	$CONH(CH_2)_2OH$	H	H	H	H	H	<0.01	[7]
26	O	–	$CONH(CH_2)_2OEt$	H	H	H	H	H	<0.01	[7]
27	O	–	$CONH(CH_2)_2NMe_2$	H	H	H	H	H	<0.01	[7]
28	O	–	$CON[(CH_2)_2OH]_2$	H	H	H	H	H	<0.01	[7]
29	O	–	$CON(CH_2)_2NMe$	H	H	H	H	H	<0.01	[7]
30	O	–	$CONH_2$	H	H	H	H	H	0.05	[7]
31	O	–	$CON(CH_2)_2O$	H	H	H	H	H	0.06	[7]
32	O	–	CO_2H	H	H	H	H	H	0.07	[7]
33	O	–	$CON(Me)_2$	H	H	H	H	H	0.07	[7]
34	O	–	CO_2Me	H	H	H	H	H	0.20	[7]
35	O	–	Me	H	H	H	H	H	0.20	[81]
36	O	–	Me	Me	H	H	Me	H	0.03	[35]
37	O	–	Me	Me	H	H	H	H	0.35	[81]
38	O	–	CO_2Et	H	H	H	H	H	0.40	[81]
39	O	–	H	H	H	H	H	H	0.55	[82]
40	O	–	H	Me	H	H	H	H	0.55	[35]
41	O	–	H	Me	H	H	CH_2OMe	Me	0.60	[81]
42	O	–	H	H	H	H	H	Me	0.80	[82]
43	O	–	H	H	H	H	Me	H	0.81	[82]
44	O	–	H	Me	H	H	H	Me	1.27	[81]
45	O	–	H	H	H	H	Me	Me	1.47	[81]
46	O	–	H	H	Me	H	Me	H	5.30	[82]
47	O	–	H	Me	H	H	Me	H	5.75	[82]
48	O	–	H	Me	Me	H	Me	H	5.78	[81]
49	N	H	H	Me	H	H	Me	H	0.48	[83]
50	N	H	H	CH_2OH	H	Me	H	Me	0.66	[84]
51	N	H	H	Me	H	Me	Me	CH_2OH	1.07	[83]
52	N	H	H	Me	H	Me	H	Me	1.36	[83]
53	N	Me	H	CH_2OMe	H	Me	H	Me	2.09	[84]
54	N	H	H	Me	H	H	Me	Me	2.59	[83]
55	N	H	H	Me	H	Me	Me	H	4.62	[83]
56	N	Me	H	CH_2OH	H	Me	H	Me	5.60	[84]
57	N	H	H	Me	H	Me	Me	Me	9.25	[83]

[a] The experimental antiproliferative activity in Ehrlich Ascites tumor cells expressed as ID_{50} relative to 8-MOP.
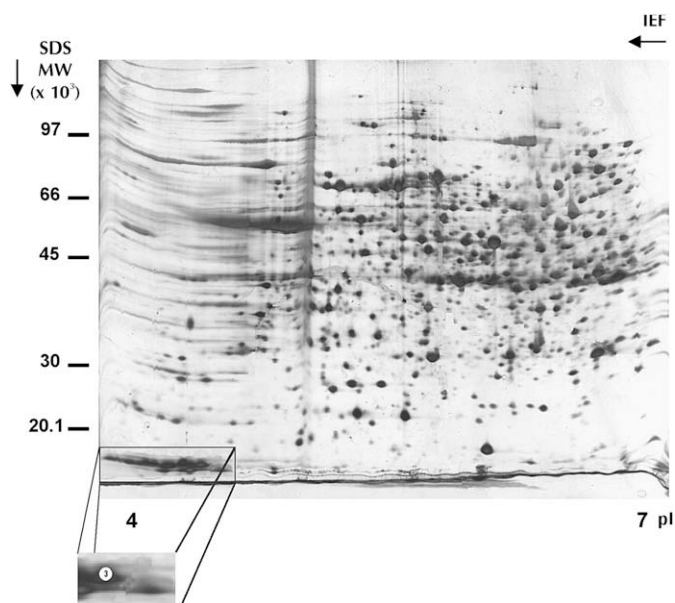[b] References in which the activity of compounds was reported.

Fig. 5. 2DE map for *L. infantum* promastigote homogenate (spot 3 is the new protein).

the furan $O_{1'}$ and the pyrimidine N_1 are going to be on the adjacent corners of the future cyclobutane ring. For the pyrone-side, the stereochemistry *syn* is defined as having the carbonyl-carbon of the pyrone ring and the N_1 of the pyrimidine on the adjacent corners of the future cyclobutane ring (Fig. 3, right).

On the other hand, some of the studied angular molecules present ramifications in the C3 carbon that hindered us to model appropriately their j-conformation, due to steric problems with the thymine ring. We also found steric impediments in the backbone of the DNA when these ramifications are much bigger. In all these cases we decided to model several alternative starting conformations for which the steric effects were eliminated. For the majority of the cases we just varied the insertion degree of molecule in the DNA; in the most critical cases we also had to rotate the molecule clockwise, see Figs. 3 and 4, and Table 1 for details.

Both, the displacement outwards DNA and the molecule rotation were carried out in the halfway and parallel plane to the nitrogen bases. In this sense, the geometric criterion used was the relative distance (in the plane projection) between the geometric centers of the double bonds (j or c bond for furocoumarins and 5,6 bond for the thymine) that will take part in the photoaddition and the relative angle between them. In both Table 1 and Fig. 2, the variations of these geometric parameters used to model the j-conformations are represented in a simplified way. Taking this aspects into consideration the notation of a MD trajectory is given here as follows. We used the notation: m-[Bond/Dist./Ang.]; where: m is the number of the compound in Table 2 or Table 3, Bond = j, c, or j&c are the chemical bonds susceptible of photoaddition in this position; whereas Dist. and Ang. are the distance and angular intercalation parameters, respectively (see also Table 1).

2.11. DNA–Drug dataset and statistical analysis

In this study we selected different furocoumarins and some of their aza-analogous, whose antiproliferative activities in Ehrlich Ascites tumor cells have been determined (Tables 2 and 3). We obtained in total 422 MDTs for these compounds. We constructed 422 LNs (one for each MD trajectory) transformed them in a vector

of 11 $\pi_k(d)$ values for the compound and 11 $\xi_k(L)$ values for the MDt (see previous sections). Were grouped all these 422 MDTs on two sets composed by MDTs of complexes between DNA and active compounds and other composed by trajectories of active compounds with no-optimal MDTs and/or trajectories of non-active compounds. In general, compounds as 4′-MAP and the 4-MBAP, with activities (ID_{50} relative to 8-MOP) of 0.13 and 0.14 are considered as poorly active [169,170]. The biological activity of these compounds is normally studied by evaluating of their capacity of forming an intercalated complex with DNA and their ability of photo-binding through mono- or bi-functional addition to the same macromolecule [171]. A traditional procedure to determine the photobiological and antiproliferative activity of furocoumarins is based on ID_{50}, the UVA dose that reduces to 50% of the DNA synthesis in Ehrlich Ascites tumor cells (EATC) in presence of tested compound at certain concentration (18–20 μM). The protocols used in the activity determination are heterogeneous, however it is very common the use of the 8-MOP as reference to express the activity [169,170,172].

Keeping in mind all the above-mentioned aspects, we classified the 57 compounds, compiled for our dataset in two observed activity groups: 0 for the inactive compounds ($LD_{50} \leq 0.1$) and 1 for the active ones ($LD_{50} > 0.1$). QSBR studies were carried out to obtain models that allow us to classify the furocoumarins derivatives in one of these two activity groups. We selected Linear Discriminant Analysis (LDA) [173,174] to fit the discriminant function as implemented in the LDA module of the STATISTICA 6.0 software package [175]. Forward-stepwise algorithm was used for variable selection [176–178]. The statistical significance of the LDA model was determined with Fisher ratio (F) and the respective p-level (p). All the variables included in the model were standardized in order to bring it into the same scale. Subsequently, a standardized linear discriminant equation that allows to compare their coefficients is obtained [179]. We also inspected the Accuracy, Sensitivity, and Specificity of the model for both training and external validation series. Last, cases/adjustable parameters ratios (ρ), and number of variables to be explored to avoid over-fitting or chance correlation [176,177]. The general form of this model is the following, where MD score is the real valued variable (output of the model) that scores the goodness of fit to guide DNA–Drug Docking.

Table 4
Top-20 MASCOT scored protein template candidates.

Protein	Accession	Mass	Score	Description
1	LmjF36.6010	73 697	61	Hypothetical
2	LmjF31.2850c	22 350	42	Ribosomal
3	CHR32_tmp.120	24 347	39	Cyclophilin
4	CHR7-11_tmp.271	16 228	35	Possible ubiquinone biosynthesis protein
5	CHR27_tmp.124c	191 035	34	Cytoskeleton
6	CHR7-11_tmp.32c	18 913	34	Hypothetical
7	CHR16-22_tmp.83	17 424	34	Hypothetical
8	CHR27_tmp.35c	17 424	34	Hypothetical
9	L344.4	52 863	34	Hypothetical
10	CHR28_tmp.181	65 936	31	Hypothetical
11	CHR30_tmp.180c	24 128	30	Ribosomal
12	CHR26_tmp.127	460 680	30	Hypothetical
13	LmjF25.0160c	24 004	29	Hypothetical
14	CHR32_tmp.448c	11 907	29	Brain Sjogren's syndrome nuclear antigen
15	CHR34_tmp.181c	11 907	29	Brain Sjogren's syndrome nuclear antigen
16	LmjF36.3570	66 043	28	Signal recognition
17	LmjF31.1700c	75 452	28	Hypothetical
18	L1648.05	72 486	28	Putative tubulin-tyrosine ligase
19	CHR34_tmp.158	47 641	28	Ribosomal
20	CHR34_tmp.157	47 743	28	Ribosomal

$$\text{MD score} = \sum_{k=0}^{10} a_k \cdot \xi_k(L) + \sum_{k=0}^{10} b_k \cdot \chi_k(D) + c_0 \qquad (10)$$

The equation do not only consider the MDT of DNA–Drug Docking but incorporates also the Mean Atomic Electronegativity values for atoms placed at distance k on the chemical structure of the drug $\chi_k(D)$. The $\chi_k(D)$ contains only information about atom-atom connectivity patterns and electronic distribution on the drug but do not incorporate MDT information. In consequence, $\chi_k(D)$ are unable to account for MDTs. The approach used to calculate $\chi_k(D)$ with MARCH-INISDE is essentially the same used to calculate $\xi_k(D)$ values but use the Markov matrix of a molecular graph instead of an LN. The details for calculating $\chi_k(D)$ values are well known and have been published; thus we refer to these references by reasons of space [180,181].

3. Results and discussion

3.1. Experiment 1

3.1.1. 2-DE isolation of a novel sequence

In this section we presented an example of the practical use of the $\xi_k(L)$ values to construct phylogenetic trees for comparison of peptides found in the PMF of a new query protein. In Fig. 5 we illustrate an overall view of the 2DE 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 p*I*. 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. To start investigation on the nature of these proteins initially we the spot marked with an arrow and encircled in the zoom image for this area, see Fig. 5.

3.1.2. 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 peptides was obtained from MALDI-TOF MS analysis. However, we focus our attention in this study on the protein corresponding to spot #3. Once we have obtained the data from MALDI-TOF MS analysis of spot #3, the more relevant MS signals were introduced into the MASCOT search engine [182,183]. Due to the fact that the MASCOT collection of annotated databases does not contained data about *L. infantum* proteome, we chose the *L. major* database of annotated proteins with MS recorded because of its similarity to *L. Infantum* [184]. Even being a protein fragment of low MW, the MASCOT search of MS signals found one hit with an M_s higher than 51 ($p < 0.05$) for spot #3 (see Table 4).

The top Mowse score found was 61, correspondent to the protein LmjF36.6010 of *L. major* with mass 73 697 but without know function annotation. The second highest Mowse score of 42 correspond to the protein LmjF31.2850c assigned as one ribosomal protein of *L. major* specie with mass 22 350. The other proteins to complete a total of 20 with similar Mowse scores were summarized in Table 4. All these proteins have Mowse values lower that the threshold value of 51 used to identify proteins with significant similarity. In any case, many of them have been also recorded with unknown function or as hypothetical proteins. Taking into

Table 5
A set of 35 peptides found on the PMFs of the new protein with and MASCOT templates.

Template	Peptide	Observed	M_r (expt)	M_r (calc)	Delta	Sequence
LmjF36.6010	P01	833.44	832.43	832.41	0.02	AGWTVDGK
	P02	877.46	876.46	876.44	0.02	LEMLESR
	P03	999.53	998.52	998.56	−0.04	RDALQLQR
	P04	1317.76	1316.75	1316.67	0.08	DEAIQSLTRER
	P05	1405.84	1404.83	1404.73	0.1	LMLTDSVSPALSR
	P06	1581.93	1580.92	1580.79	0.13	EKIMLAQEVTTMR
	P07	1626	1625	1624.86	0.14	MLQHASLKSDPLAAK
	P08	1802.07	1801.06	1800.89	0.17	IMLAQEVTTMRAMYK
LmjF31.2850c	P09	823.44	822.43	822.39	0.04	MFPAETK
	P10	861.45	860.44	860.43	0.01	LGAEVELM
	P11	867.47	866.46	866.4	0.06	AMTEMLR
	P12	1405.84	1404.83	1404.7	0.13	DAMVKLGAEVELM
	P13	1449.84	1448.83	1448.72	0.11	MRQSVLACDVVR
CHR32_tmp.120	P14	823.44	822.43	822.42	0.01	TFLSAER
	P15	841.43	840.42	840.41	0.01	GYDVIMK
	P16	911.49	910.48	910.45	0.03	VEMELFK
	P17	1449.84	1448.83	1448.76	0.07	VEMELFKDVVPK
	P18	1581.93	1580.92	1580.8	0.12	MHSEALVISYFLR
CHR7-11_tmp.271	P19	823.44	822.43	822.41	0.02	EYEALAK
	P20	839.44	838.43	838.39	0.04	MPVDYSK
	P21	877.46	876.46	876.45	0.01	ETVMGKGR
CHR27_tmp.124c	P22	823.44	822.43	822.44	−0.01	FMKLER
	P23	833.44	832.43	832.45	−0.02	KAENMLK
	P24	839.44	838.43	838.44	−0.01	RLEHER
	P25	841.43	840.42	840.39	0.03	AACTPGHK
	P26	921.5	920.49	920.5	0	SKSFDIPK
	P27	999.53	998.52	998.48	0.04	YLAAEYGGR
	P28	1043.6	1042.6	1042.54	0.06	DVQEALNVR
	P29	1405.84	1404.83	1404.71	0.12	GSALNDRAFEVAR
	P30	1449.84	1448.83	1448.71	0.12	DRACQLAELVMK
	P31	1493.86	1492.85	1492.73	0.13	EQLPEGHSADLAAR
CHR7-11_tmp.32c	P32	823.44	822.43	822.4	0.03	AEERYR
	P33	833.44	832.43	832.4	0.03	MHNLYR
	P34	1043.6	1042.6	1042.55	0.04	IGVNRAEER
	P35	2201	2199.99	2200.13	−0.14	ELTTVDATAQQTPWWRVAK

consideration we can consider this protein as low-Mowse score case (because no protein in MASCOT search with known function has a high score). As we referred in introduction precisely the PMF of this type of protein may be of high interest. In Table 5 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 [157] and possible *Leishmania* dynein proteins [83]. In this table we have shown only the 35 more interesting peptides matching with the MS of other proteins on the MASCOT search. Considering the high importance of phylogenetic analysis in the next section we propose a new algorithm for phylogenetic tree construction based MDTs and using this set of peptides as case study.

3.1.3. MC exploration of peptides found on PMF of the new protein

After MS characterization of the PMF of the new protein, we decided to use the $\xi_k(L)$ values as inputs to construct a new type of phylogenetic tree. In so doing, we obtained firstly the MDTs for the more interesting 35 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 [159,160]. We also reported the (E_1) and energy gradient (δ_1) obtained after optimization of the structure with AMBER force field as well as the last energy value (E_{100}) obtained on the MC exploration of the MDT. Last, we report in Table 6 the ACCR values for the MDT of the 35 peptides; which are all lower than 0.5. In consequence, we can accept the MD results and use them to construct a phylogenetic tree.

3.1.4. MDT phylogenetic for PMF of new query protein

Using information about the distribution of monomers (amino acids or nucleotides) throughout the biopolymer chain have been the major tendency on phylogenetic analysis [185]. In the Introduction, we discussed the importance of new molecular phylogenetic approaches for polypeptide chains based on other sources of information such as MDTs. In Materials and method we outlined the possibility of construction of a phylogenetic tree for the PMFs of the new protein described above using Equation (9). For it, we have calculated first the $\xi_k(L)$ for the 35 more relevant peptides (see Table 7) and later the peptide–peptide distance using Equation (9). In Fig. 6 we illustrate that there are notable differences on the grouping of the 35 peptides if we use the traditional sequence similarity method or alternatively the present approach. This results show that in principle the distance $D_{pq}(\xi)$ between a peptide p and other q based on $\xi_k(L)$ values of MDTs codify information essentially different to sequence similarity. In this sense, the present molecular phylogenetic algorithm may become an alternative to traditional methods.

3.2. Experiment 2

3.2.1. DNA–Drug docking

Using the new $\xi_k(L)$ values as inputs we can obtain a classifier to discriminate DNA–Drug complexes of the two classes defined in materials and methods. The best model we found was:

$$MD\ score = -2.24 \times \xi_0(L) - 1.59 \times \chi_2(D) - 2.11 \quad n = 316$$

$$F = 120.99 \ p < 0.01 \ \rho = 70.33 \qquad (11)$$

Table 6
Some MDT energy values for peptides found on the PMF of the new protein.

Pept.	Sequence	E_0	δ_0	E_1	δ_1	E_{100}	ACCR$_{100}$
P01	AGWTVDGK	7347.62	5054.22	28.73	0.10	111.48	0.47
P02	LEMLESR	11.17	10.48	−41.42	0.09	65.72	0.49
P03	RDALQLQR	−77.59	11.29	−163.12	0.99	−48.02	0.47
P04	DEAIQSLTRER	619.67	296.52	−149.80	0.10	−17.93	0.47
P05	LMLTDSVSPALSR	15036.18	7987.95	−55.85	0.10	91.25	0.48
P06	EKIMLAQEVTTMR	577.02	197.41	−79.93	0.09	91.45	0.48
P07	MLQHASLKSDPLAAK	14695.52	7374.20	64.10	0.10	263.42	0.47
P08	IMLAQEVTTMRAMYK	899.72	220.08	−87.59	0.10	115.66	0.48
P09	MFPAETK	310.54	61.75	25.45	0.10	110.59	0.47
P10	LGAEVELM	546.32	269.08	6.70	0.09	105.72	0.48
P11	AMTEMLR	0.35	10.45	−77.19	0.10	11.64	0.47
P12	DAMVKLGAEVELM	1049.69	295.86	−30.37	0.10	132.30	0.48
P13	MRQSVLACDVVR	1404.96	364.28	−143.87	0.10	4.06	0.48
P14	TFLSAER	83.16	41.35	−61.16	0.10	23.92	0.47
P15	GYDVIMK	1519.70	505.90	−9.82	0.10	78.21	0.48
P16	VEMELFK	168.78	39.88	−9.63	0.10	83.26	0.49
P17	VEMELFKDVVPK	594555.63	582651.63	−42.04	0.10	112.65	0.48
P18	MHSEALVISYFLR	1598.97	358.15	−20.42	0.10	145.47	0.47
P19	EYEALAK	390.84	181.36	2.04	0.10	81.61	0.46
P20	MPVDYSK	453.14	181.59	3.69	0.09	90.22	0.47
P21	ETVMGKGR	29.55	12.55	−59.94	0.10	34.41	0.47
P22	FMKLER	110.27	40.55	−22.21	0.10	64.97	0.48
P23	KAENMLK	70.97	11.13	−26.46	0.10	64.15	0.48
P24	RLEHER	−14.75	10.81	−128.88	0.10	−40.43	0.46
P25	AACTPGHK	17181.88	11045.33	84.06	0.09	169.13	0.48
P26	SKSFDIPK	1325711.25	1771949.63	40.49	0.10	141.63	0.48
P27	YLAAEYGGR	647.60	234.33	−87.52	0.10	27.46	0.47
P28	DVQEALNVR	486.06	244.65	−88.87	0.10	20.24	0.48
P29	GSALNDRAFEVAR	525.10	213.31	−162.52	0.10	−17.81	0.47
P30	DRACQLAELVMK	506.82	211.08	−97.64	0.10	39.80	0.48
P31	EQLPEGHSADLAAR	164.26	26.06	−59.73	0.10	80.00	0.47
P32	AEERYR	270.57	183.70	−114.27	0.10	−32.83	0.46
P33	MHNLYR	363.75	180.53	−26.20	0.10	55.96	0.47
P34	IGVNRAEER	3.65	14.52	−139.72	0.10	−13.62	0.48
P35	ELTTVDATAQQTPWWRVAK	35088.39	7955.64	−108.83	0.10	130.40	0.47

Table 7
Some $\xi_k(L)$ values for 35 peptides found on the PMF of the new protein.

Pept.	Sequence	$\xi_0(L)$	$\xi_1(L)$	$\xi_2(L)$	$\xi_3(L)$	$\xi_4(L)$	$\xi_5(L)$
P01	AGWTVDGK	2.0	69.4	54.3	49.2	44.1	40.7
P02	LEMLESR	2.1	71.2	55.7	50.5	45.3	41.8
P03	RDALQLQR	2.1	72.9	57.0	51.6	46.3	42.8
P04	DEAIQSLTRER	2.1	71.6	56.0	50.7	45.5	42.0
P05	LMLTDSVSPALSR	2.0	69.4	54.3	49.2	44.1	40.7
P06	EKIMLAQEVTTMR	2.0	68.5	53.5	48.5	43.5	40.2
P07	MLQHASLKSDPLAAK	2.0	69.4	54.2	49.1	44.1	40.7
P08	IMLAQEVTTMRAMYK	2.0	69.2	54.0	49.0	43.9	40.5
P09	MFPAETK	2.0	69.4	54.2	49.1	44.1	40.7
P10	LGAEVELM	2.0	69.3	54.2	49.1	44.0	40.6
P11	AMTEMLR	2.1	72.1	56.4	51.1	45.9	42.4
P12	DAMVKLGAEVELM	2.0	69.3	54.2	49.1	44.0	40.6
P13	MRQSVLACDVVR	2.2	73.9	57.8	52.3	46.9	43.3
P14	TFLSAER	2.1	71.8	56.2	50.9	45.6	42.1
P15	GYDVIMK	2.2	69.2	54.3	49.3	44.3	41.0
P16	VEMELFK	2.1	69.7	54.5	49.3	44.2	40.8
P17	VEMELFKDVVPK	2.0	69.5	54.4	49.3	44.2	40.8
P18	MHSEALVISYFLR	2.0	69.3	54.2	49.1	44.0	40.7
P19	EYEALAK	2.0	69.3	54.2	49.1	44.0	40.6
P20	MPVDYSK	2.0	69.5	54.3	49.2	44.1	40.7
P21	ETVMGKGR	2.1	71.9	56.2	51.0	45.7	42.2
P22	FMKLER	2.1	69.8	54.9	49.9	44.9	41.6
P23	KAENMLK	2.2	70.6	55.0	50.0	44.9	41.6
P24	RLEHER	2.1	72.6	56.8	51.4	46.1	42.6
P25	AACTPGHK	2.0	69.3	54.2	49.1	44.0	40.6
P26	SKSFDIPK	2.0	69.5	54.4	49.3	44.2	40.8
P27	YLAAEYGGR	2.1	71.4	55.8	50.6	45.4	41.9
P28	DVQEALNVR	2.1	72.5	56.7	51.4	46.1	42.5
P29	GSALNDRAFEVAR	2.1	72.9	57.0	51.6	46.3	42.7
P30	DRACQLAELVMK	2.1	70.4	55.1	49.9	44.8	41.3
P31	EQLPEGHSADLAAR	2.0	69.5	54.3	49.2	44.1	40.7
P32	AEERYR	2.1	72.3	56.6	51.3	46.0	42.5
P33	MHNLYR	2.1	72.2	56.5	51.3	46.1	42.6
P34	IGVNRAEER	2.1	71.4	55.8	50.6	45.4	41.9
P35	ELTTVDATAQQTPWWRVAK	2.0	69.1	54.0	49.0	43.9	40.5

The output of the model, MD score, is a real value variable that scores the predicted goodness of fit for one MD trajectory. After the forward-stepwise selection of $\xi_k(L)$ and $\chi_k(L)$ values the model retained only two parameters $\xi_0(L)$ and $\chi_2(L)$. These values can be interpreted as the average or mean value of energy for states on the 2D space (nodes on the LN) and mean electronegativity value for atoms placed at distance $k = 2$ on the drug, respectively. The model was trained with a training series and later validated with and external validation series. In training series the model correctly classifies 78 out of 80 (specificity = 97.50%) optimal and 192 out of 236 (sensitivity = 81.36%) no-optimal MDTs.

In statistical prediction, the following three cross-validation methods are often used to examine a classifier for its effectiveness in practical application: independent dataset test, sub-sampling test, and jackknife test [186]. However, as elucidated by [8] and demonstrated in [188], 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 dataset, and hence has been increasingly used by investigators to examine the accuracy of various predictors [189–197]. In the current study, for simplifying the demonstration, we just used the independent dada for validation. In external validation series the model correctly classifies 25 out of 26 (specificity = 96.15%) optimal and 65 out of 80 (sensitivity = 81.25%) no-optimal MDTs. This results represent total accuracy = 85.44% and 84.91% in training and validation respectively. This result indicates significant goodness of fit for this linear classifier based on the results reported before for LDA-QSAR classifiers. See for instance the LDA models used to predict anti-Leishmania, and in general other anti-parasitic or anti-microbial drugs or other classes of activities by Galvez, García-Domenech, Marrero-Ponce, Castillo-Garit, Casañola-Martin, and other authors [91,198–208].
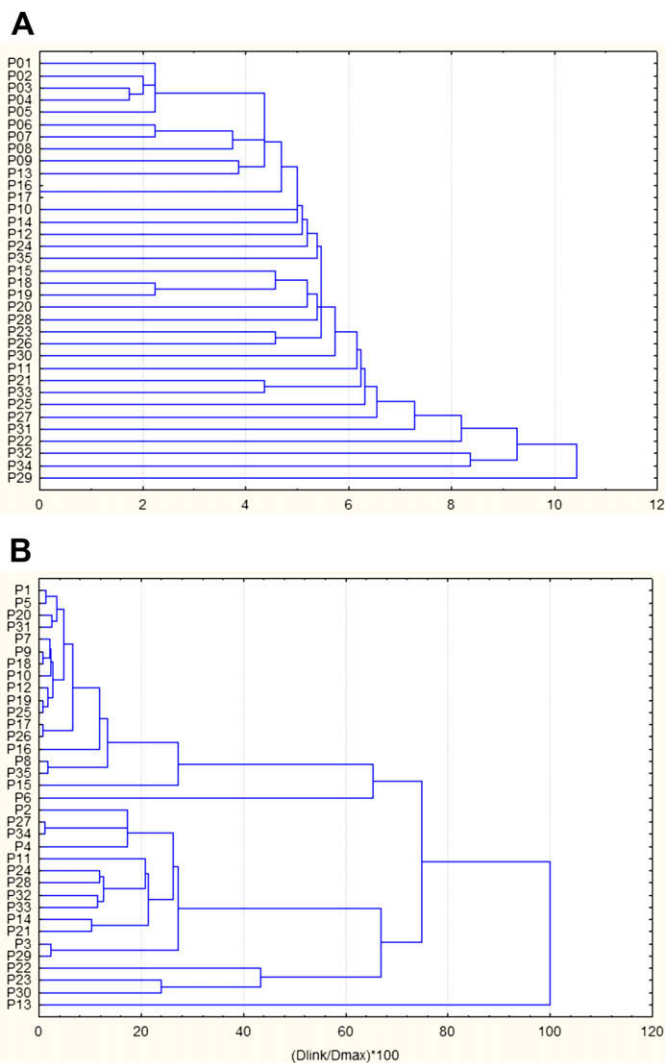
Fig. 6. Different phylogenetic trees: Sequence similarity (A) and $\xi_k(L)$ values of MDTs (B).

4. Conclusions

MDTs of biopolymers can be numerically described with a new class of invariants $\xi_k(L)$ representing spatial distribution of Mean-Energy values on a 2D Euclidean space. The procedure forces one MD trajectory to fold into a 2D Cartesian coordinates system using a step-by-step procedure driven by simple rules. The graphical representation of this space has the form of a lattice network symbolized as LN. We can use $\xi_k(L)$ values of LN to develop new algorithms to perform molecular Phylogenetic analysis of peptides based on MDTs instead of the sequence of the polypeptide. The new procedure combined with 2D Electrophoresis and MALDI-TOF MS can be applied to analyze Peptide Mass Fingerprints of new proteins. We can use the same idea to seek scoring functions for DNA–Drug Docking simulations. The work opens new perspective on the analysis and applications MD on biopolymers sciences.

H. acknowledges financial support of Program Isidro Parga Pondal funded by Xunta de Galicia and European Social Fund (E.S.F.).

Appendix. Supplementary data

Supplementary data associated with this article can be found in the online version, at doi:10.1016/j.polymer.2009.05.055.

References

[1] Chou KC. Curr Med Chem 2004;11(16):2105–34.
[2] Zheng H, Wei DQ, Zhang R, Wang C, Wei H, Chou KC. Med Chem 2007;3(5):488–93.
[3] Chou KC, Maggiora GM, Nemethy G, Scheraga HA. Proc Natl Acad Sci U S A 1988;85:4295–9.
[4] Chou KC, Wei DQ, Du QS, Sirois S, Zhong WZ. Curr Med Chem 2006;13(27): 3263–70.
[5] Chou KC. J Mol Biol 1992;223(2):509–17.
[6] Du QS, Huang RB, Chou KC. Curr Protein Pept Sci 2008;9:248–59.
[7] Gonzalez-Diaz H, Gonzalez-Diaz Y, Santana L, Ubeira FM, Uriarte E. Proteomics 2008;8(4):750–78.
[8] Chou KC, Shen HB. Nat Protoc 2008;3:153–62.
[9] Chou KC, Shen HB. J Proteome Res 2007;6(5):1728–34.
[10] Chou KC, Shen HB. J Proteome Res 2006;5(8):1888–97.
[11] Chou KC. Curr Protein Pept Sci 2005;6(5):423–36.
[12] Chou KC, Shen HB. Biochem Biophys Res Commun 2007;360(2):339–45.
[13] Chou KC, Shen HB. Biochem Biophys Res Commun 2008;376:321–5.
[14] Shen HB, Chou KC. Biochem Biophys Res Commun 2007.
[15] Xiao X, Wang P, Chou KC. J Comput Chem 2008, doi:10.1002/jcc.21163.
[16] Shen HB, Chou KC. Anal Biochem 2008;375:388–90.
[17] Shen HB, Chou KC. Biochem Biophys Res Commun 2007;363(2):297–303.
[18] Chou KC, Chen NY. Sci Sin 1977;20:447–57.
[19] Chou KC. Biopolymers 1987;26(2):285–95.
[20] Chou KC. Biophys Chem 1986;25(2):105–16.
[21] Chou KC, Kiang YS. Biophys Chem 1985;22(3):219–35.
[22] Chou KC. Biophys J 1985;48(2):289–97.
[23] Chou KC. Biophys Chem 1984;20(1–2):61–71.
[24] Chou KC. Biochem J 1984;221(1):27–31.
[25] Chou KC. Biophys J 1984;45(5):881–9.
[26] Chou KC. Biochem J 1983;215(3):465–9.
[27] Chou KC. Biochem J 1983;209(3):573–80.
[28] Martel P. Prog Biophys Mol Biol 1992;57:129–79.
[29] Sinkala Z. J Theor Biol 2006;241:919–27.
[30] Chou KC, Chen NY, Forsen S. Chem Scripta 1981;18:126–32.
[31] Chou KC, Zhang CT, Maggiora GM. Biopolymers 1994;34:143–53.
[32] Chou KC, Mao B. Biopolymers (Biospectroscopy) 1988;27:1795–815.
[33] Chou KC. Biophys Chem 1988;30(1):3–48.
[34] Chou JJ, Li S, Klee CB, Bax A. Nat Struct Biol 2001;8:990–7.
[35] Gordon G. J Cell Physiol 2007;212:579–82.
[36] Gordon G. J Biomed Sci Eng 2008;1:152–6.
[37] McCammon JA, Gelin BR, Karplus M. Nature 1977;267(5612):585–90.
[38] Karplus M, McCammon JA. Nat Struct Biol 2002;9(9):646–52.
[39] McCammon JA, Karplus M. Nature 1977;268(5622):765–6.
[40] Navarro E, Tejero R, Fenude E, Celda B. Biopolymers 2001;59(2):110–9.
[41] Navarro E, Fenude E, Celda B. Biopolymers 2004;73(2):229–41.
[42] Navarro E, Fenude E, Celda B. Biopolymers 2002;64(4):198–209.
[43] Gia O, Marciani Magno S, Gonzalez-Diaz H, Quezada E, Santana L, Uriarte E, et al. Bioorg Med Chem 2005;13(3):809–17.
[44] Monleon D, Esteve V, Celda B. Biochem Biophys Res Commun 2003; 303(1):81–90.
[45] Chou KC. J Proteome Res 2005;4(5):1657–60.
[46] Hamacher K. J Comput Chem 2007;28(16):2576–80.
[47] King EL, Altman C. J Phys Chem 1956;60:1375–8.
[48] Chou KC, Jiang SP, Liu WM, Fee CH. Sci Sin 1979;22:341–58.
[49] Andraos J. Can J Chem 2008;86:342–57.
[50] Chou KC. Biophys Chem 1990;35(1):1–24.
[51] Althaus IW, Chou KC, Lemay RJ, Franks KM, Deibel MR, Kezdy FJ, et al. Biochem Pharmacol 1996;51(6):743–50.
[52] Zhang CT, Chou KC. J Mol Biol 1994;238(1):1–8.
[53] Chou KC, Zhang CT, Elrod DW. J Protein Chem 1996;15(1):59–61.
[54] Qi XQ, Wen J, Qi ZH. J Theor Biol 2007;249:681–90.
[55] Prado-Prado FJ, Gonzalez-Diaz H, de la Vega OM, Ubeira FM, Chou KC. Bioorg Med Chem 2008;16(11):5871–80.
[56] Diao Y, Li M, Feng Z, Yin J, Pan Y. J Theor Biol 2007;247:608–15.
[57] Gonzalez-Diaz H, Vilar S, Santana L, Uriarte E. Curr Top Med Chem 2007;7(10):1015–29.
[58] Wolfram S. A new kind of science. In: Media W, editor. Champaign, IL; 2002.
[59] Xiao X, Shao SH, Chou KC. Biochem Biophys Res Commun 2006;342(2): 605–10.
[60] Xiao X, Shao S, Ding Y, Huang Z, Chen X, Chou KC. J Theor Biol 2005;235(4):555–65.
[61] Gao L, Ding YS, Dai H, Shao SH, Huang ZD, Chou KC. J Pharm Biomed Anal 2006;41(1):246–50.
[62] Xiao X, Shao S, Ding Y, Huang Z, Chen X, Chou KC. Amino Acids 2005;28(1):29–35.
[63] Xiao X, Chou KC. Protein Pept Lett 2007;14:871–5.
[64] Chen M, Huang WQ. Genomics Proteomics Bioinformatics 2005;3(4):225–30.
[65] Thachuk C, Shmygelska A, Hoos HH. BMC Bioinformatics 2007;8(1):342.
[66] Zhang XS, Wang Y, Zhan ZW, Wu LY, Chen L. J Bioinformatics Comput Biol 2005;3(2):385–400.
[67] Jiang M, Zhu B. J Bioinformatics Comput Biol 2005;3(1):19–34.
[68] Gupta A, Manuch J, Stacho L. J Comput Biol 2005;12(10):1328–45.
[69] Gupta A, Manuch J, Stacho L. Proc IEEE Comput Syst Bioinformatics Conf 2004:311–8.
[70] Berger B, Leighton T. J Comput Biol 1998;5(1):27–40.
[71] Agarwala R, Batzoglou S, Dancik V, Decatur SE, Hannenhalli S, Farach M, et al. J Comput Biol 1997;4(3):275–96.
[72] Gates MA. J Theor Biol 1986;119:319–28.
[73] Nandy A. Comput Appl Biosci 1996;12(1):55–62.
[74] Leong PM, Morgenthaler S. Comput Appl Biosci 1995;11:503–7.
[75] Randic M, Guo X, Basak SC. J Chem Inf Comput Sci 2001;41(3):619–26.
[76] González-Díaz H, Saíz-Urra L, Molina R, Uriarte E. Polymer 2005;46:2791–8.
[77] González-Díaz H, Pérez-Bello A, Uriarte E. Polymer 2005;46:6461–73.
[78] González-Díaz H, Molina RR, Uriarte E. Polymer 2004;45:3845–53.
[79] Cruz-Monteagudo M, Munteanu CR, Borges F, Cordeiro MNDS, Uriarte E, Chou K-C, et al. Polymer 2008;49(25):5575–87.
[81] González-Díaz H, Vilar S, Santana L, Uriarte E. Curr Top Med Chem 2007;7(10): 1025–39.
[82] Aguero-Chapin G, Gonzalez-Diaz H, de la Riva G, Rodriguez E, Sanchez-Rodríguez A, Podda G, et al. J Chem Inf Model 2008;48(2):434–48.
[83] Dea-Ayuela MA, Perez-Castillo Y, Meneses-Marcel A, Ubeira FM, Bolas-Fernandez F, Chou KC, et al. Bioorg Med Chem 2008;16(16):7770–6.
[84] Perez-Bello A, Munteanu CR, Ubeira FM, Lopes De Magalhaes A, Uriarte E, Gonzalez-Díaz H. J Theor Biol 2008.
[85] Gonzalez-Diaz H, Prado-Prado F, Ubeira FM. Curr Top Med Chem 2008;8(18):1676–90.
[86] Cruz-Monteagudo M, González-Díaz H, Borges F, Dominguez ER, Cordeiro MN. Chem Res Toxicol 2008;21:619–32.
[87] Cruz-Monteagudo M, Munteanu CR, Borges F, Cordeiro MN, Uriarte E, Gonzalez-Diaz H. Bioorg Med Chem 2008;16(22):9684–93.
[88] Ferino G, Gonzalez-Diaz H, Delogu G, Podda G, Uriarte E. Biochem Biophys Res Commun 2008;372(2):320–5.
[89] Rama-Iñiguez S, Dea-Ayuela MA, Sanchez-Brunete JA, Torrado JJ, Alunda JM, Bolas-Fernández F. Antimicrob Agents Chemother 2006;50(4):1195–201.
[90] Sánchez-Brunete JA, Dea-Ayuela MA, Rama S, Bolás F, Alunda JM, Raposo R, et al. Antimicrob Agents Chemother 2004;48(9):3246–52.
[91] Roldos V, Nakayama H, Rolon M, Montero-Torres A, Trucco F, Torres S, et al. Eur J Med Chem 2008;43(9):1797–807.
[92] Chenik M, Chaabouni N, Achour-Chenik YB, Ouakad M, Lakhal-Naouar I, Louzir H, et al. Biochem Biophys Res Commun 2006;341:541–8.
[93] Sarciron ME, Terreux R, Prieto Y, Cortes M, Cuellar MA, Tapia RA, et al. Parasite 2005;12(3):251–8.
[94] Opperdoes FR, Szikora JP. Mol Biochem Parasitol 2006.
[95] Zick A, Onn I, Bezalel R, Margalit H, Shlomai J. Nucleic Acids Res 2005;33(13):4235–42.
[96] Aksu S, Scheler C, Focks N, Leenders F, Theuring F, Salnikow J, et al. Proteomics 2002;2(10):1452–63.
[97] Tebbe A, Klein C, Bisle B, Siedler F, Scheffer B, Garcia-Rizo C, et al. Proteomics 2005;5(1):168–79.
[98] Hirosawa M, Hoshida M, Ishikawa M, Toya T. Comput Appl Biosci 1993;9(2):161–7.
[99] Resing KA, Meyer-Arendt K, Mendoza AM, Aveline-Wolf LD, Jonscher KR, Pierce KG, et al. Anal Chem 2004;76(13):3556–68.
[100] Savitski MM, Nielsen ML, Kjeldsen F, Zubarev RA. J Proteome Res 2005;4(6): 2348–54.
[101] Savitski MM, Nielsen ML, Zubarev RA. Mol Cell Proteomics 2005;4(8): 1180–8.
[102] Han L, Cui J, Lin H, Ji Z, Cao Z, Li Y, et al. Proteomics 2006;6(14):4023–37.
[103] Sanchez R, Morgado E, Grau R. J Math Biol 2005;51(4):431–57.
[104] Sanchez R, Morgado E, Grau R. Bull Math Biol 2005;67(1):1–14.
[105] Sanchez R, Grau R. Bull Math Biol 2005;67(5):1017–29.
[106] Liao B, Ding K. J Comput Chem 2005;26(14):1519–23.
[107] Liao B, Wang TM. J Chem Inf Comput Sci 2004;44(5):1666–70.
[108] Liao B, Wang TM. J Comput Chem 2004;25(11):1364–8.
[109] Liao B, Xiang X, Zhu W. J Comput Chem 2006;27(11):1196–202.
[110] Yu-Hua Y, Liao B, Tian-Ming W. J Mol Struct THEOCHEM 2005;755:131–6.
[111] Liao B, Wang T. J Biomol Struct Dyn 2004;21:827–32.
[112] Liao B, Ding K, Wang T. J Biomol Struct Dyn 2005;22:455–64.
[113] Liao B, Wang T, Ding K. Mol Simul 2005;31(14):1063–71.
[114] Liao B, Luo J, Li R, Zhu W. Int J Quant Chem 2006;106(8):1749–55.
[115] Zhu W, Liao B, Ding K. J Mol Struct THEOCHEM 2005;757:193–8.
[116] Randic M, Vracko M. J Chem Inf Comput Sci 2000;40(3):599–606.
[117] Aguero-Chapin G, González-Díaz H, Molina R, Varona-Santos J, Uriarte E, Gonzalez-Diaz Y. FEBS Lett 2006;580(3):723–30.
[118] Randič M, Vračko M, Nandy A, Basak SC. J Chem Inf Comput Sci 2000;40:1235–44.

[120] Liu WC, Lin WH, Davis AJ, Jordan F, Yang HT, Hwang MJ. BMC Bioinformatics 2007;8:121.
[121] Wang B, Chen P, Huang DS, Li JJ, Lok TM, Lyu MR. FEBS Lett 2006;580(2):380–4.
[122] Thibert B, Bredesen DE, del Rio G. BMC Bioinformatics 2005;6:213.
[123] Zupan J, Randic M. J Chem Inf Model 2005;45(2):309–13.
[124] Liao B. Chem Phys Lett 2005;401:196–9.
[125] Liao B, Tan M, Ding K. Chem Phys Lett 2005;414(4–6):296–300.
[126] Zhang Y, Chen W. Comb Chem High Throughput Screen 2007;10(3):231–7.
[127] Zhang X, Luo J, Yang L. J Comput Chem 2007;28:2342–6.
[129] Vilar S, González-Díaz H, Santana L, Uriarte E. J Comput Chem 2008;29:2613–22.
[130] Giordanetto F, Fossa P, Menozzi G, Mosti L. J Comput Aided Mol Des 2003;17(1):53–64.
[132] Schein CH, Ivanciuc O, Braun W. Immunol Allergy Clin North Am 2007;27(1):1–27.
[133] Schein CH, Ivanciuc O, Braun W. J Agric Food Chem 2005;53(22):8752–9.
[134] Ivanciuc O, Schein CH, Braun W. Nucleic Acids Res 2003;31(1):359–62.
[135] Ivanciuc O, Mathura V, Midoro-Horiuti T, Braun W, Goldblum RM, Schein CH. J Agric Food Chem 2003;51(16):4830–7.
[136] Balaban AT, Beteringhe A, Constantinescu T, Filip PA, Ivanciuc O. J Chem Inf Model 2007;47(3):716–31.
[137] Ivanciuc O, Ivanciuc T, Klein DJ, Seitz WA, Balaban AT. J Chem Inf Comput Sci 2001;41(3):536–49.
[138] Ivanciuc O. J Chem Inf Comput Sci 2000;40(6):1412–22.
[139] Bonchev D. J Chem Inf Comput Sci 2000;40(4):934–41.
[140] Ivanciuc O, Ivanciuc T, Klein DJ. SAR QSAR Environ Res 2001;12(1–2):1–16.
[141] Bonchev D, Buck GA. J Chem Inf Model 2007;47(3):909–17.
[142] Bonchev D. SAR QSAR Environ Res 2003;14(3):199–214.
[143] Bonchev D. Chem Biodivers 2004;1(2):312–26.
[144] Kier LB, Bonchev D, Buck GA. Chem Biodivers 2005;2(2):233–43.
[145] Bornholdt S, Schuster HG. Handbook of graphs and complex networks: from the genome to the internet. Weinheim: WILEY-VCH GmbH & Co. KGa; 2003.
[146] Zhang S, Golbraikh A, Tropsha A. J Med Chem 2006;49:2713–24.
[147] Hetenyi C, Paragi G, Maran U, Timar Z, Karelson M, Penke B. J Am Chem Soc 2006;128(4):1233–9.
[148] Lill MA, Vedani A, Dobler M. J Med Chem 2004;47(25):6174–86.
[149] Smith R, Hubbard RE, Gschwend DA, Leach AR, Good AC. J Mol Graph Model 2003;22(1):41–53.
[150] Wang R, Lu Y, Wang S. J Med Chem 2003;46(12):2287–303.
[151] Ferrara P, Gohlke H, Price DJ, Klebe G, Brooks 3rd CL. J Med Chem 2004;47(12):3032–47.
[152] Santana L, Uriarte E, Roleira F, Milhazes N, Borges F. Curr Med Chem 2004;11(24):3239–61.
[153] Pathak MA, Fitzpatrick TB. J Photochem Photobiol B 1992;14(1–2):3–22.
[154] Parrish JA, Stern RS, Pathak MA, Fitzpatrick TB. In: Regan JD, Parrish JA, editors. The science of photomedicine. New York: Plenum Press; 1982. p. 595.
[155] Dall'Acqua F, Vedaldi D, Baccichetti F, Bordin F, Averbeck D. Farmaco [Sci] 1981;36(7):519–35.
[156] Pathak MA, Parrish JA, Fitzpatrick TB. Farmaco [Sci] 1981;36(7):479–91.
[157] Dea-Ayuela MA, Bolás-Fernández F. Vet Parasitol 2005;132:43–9.
[158] Gharahdaghi F, Weinberg CR, Meagher DA, Imai BS, Mische SM. Electrophoresis 1999;20:601–5.
[159] Froimowitz M. Biotechniques 1993;14(6):1010–3.
[160] Hyperchem. Gainesville, FL, USA: Hypercube Inc.; 2002.
[161] Liu Y, Beveridge DL. Proteins 2002;46(1):128–46.
[162] Hypercube Inc. Hyperchem software. Release 7.5 for windows, Molecular Modeling System. Gainesville, FL, USA: Hypercube Inc.; 2002.
[163] Eichman BF, Mooers BH, Alberti M, Hearst JE, Ho PS. J Mol Biol 2001;308(1):15–26.
[164] Clark T. A handbook of computational chemistry. New York: John Wiley & Sons; 1985.
[165] Kitamura N, Kohtani S, Nakagaki R. J Photochem Photobiol C Photochem Rev 2005;6:168–85.
[166] Tessman JW, Isaacs ST, Hearst JE. Biochemistry 1985;24(7):1669–76.
[167] Cimino GD, Gamper HB, Isaacs ST, Hearst JE. Annu Rev Biochem 1985;54:1151–93.
[168] Caffieri S, Miolo G, Dall'Acqua F, Benetollo F, Bombieri G. Photochem Photobiol 2000;72(1):23–7.
[169] Baccichetti F, Bordin F, Simonato M, Toniolo L, Marzano C, Rodighiero P, et al. Il Farmaco 1992;47(12):1529–41.
[170] Antonello C, Zagotto G, Mobilio S, Marzano C, Gia O, Uriarte E. Il Farmaco 1994;49(4):277–80.
[171] Zagotto G, Gia O, Baccichetti F, Uriarte E, Palumbo M. Photochem Photobiol 1993;58(4):486–91.
[172] Musajo L, Visentini P, Baccichetti F, Razzi MA. Experientia 1967;23:335–6.
[173] Van Waterbeemd H. Discriminant analysis for activity prediction. In: Van Waterbeemd H, editor. Chemometric methods in molecular design, vol. 2. New York: Wiley-VCH; 1995. p. 265–82.
[174] Estrada E, Molina E. J Chem Inf Comput Sci 2001;41(3):791–7.
[175] STATISTICA. Statsoft Inc.; 2001.
[176] Kowalski RB, Wold S. Pattern recognition in chemistry. In: Krishnaiah PR, Kanal LN, editors. Handbook of statistics. Amsterdam: North Holland Publishing Company; 1982. p. 673–97.
[177] Van Waterbeemd H. Chemometric methods in molecular design. New York: Wiley-VCH; 1995.
[178] Cruz-Monteagudo M, Gonzalez-Diaz H, Aguero-Chapin G, Santana L, Borges F, Dominguez ER, et al. J Comput Chem 2007;28(11):1909–23.
[179] Kutner MH, Nachtsheim CJ, Neter J, Li W. Standardized multiple regression model. Applied linear statistical models. New York: McGraw Hill; 2005. p. 271–7.
[180] Santana L, Uriarte E, González-Díaz H, Zagotto G, Soto-Otero R, Mendez-Alvarez E. J Med Chem 2006;49(3):1149–56.
[181] Santana L, Gonzalez-Diaz H, Quezada E, Uriarte E, Yanez M, Vina D, et al. J Med Chem 2008;51(21):6740–51.
[182] Lei Z, Elmer AM, Watson BS, Dixon RA, Mendes PJ, Sumner LW. Mol Cell Proteomics 2005;4(11):1812–25.
[183] Giddings MC, Shah AA, Gesteland R, Moore B. Proc Natl Acad Sci U S A 2003;100(1):20–5.
[184] Arakaki T, Le Trong I, Phizicky E, Quartley E, DeTitta G, Luft J, et al. Acta Crystallogr Sect F Struct Biol Cryst Commun 2006;62(Pt 3):175–9.
[185] Puslednik L, Serb JM. Mol Phylogenet Evol 2008;48(3):1178–88.
[186] Chou KC, Zhang CT. Crit Rev Biochem Mol Biol 1995;30(4):275–349.
[188] Chou KC, Shen HB. Anal Biochem 2007;370:1–16.
[189] Zhou XB, Chen C, Li ZC, Zou XY. J Theor Biol 2007;248(3):546–51.
[190] Lin H. J Theor Biol 2008;252(2):350–6.
[191] Zhang GY, Fang BS. J Theor Biol 2008;253(2):310–5.
[192] Zhang GY, Li HC, Fang BS. Protein Pept Lett 2008;15:1132–7.
[193] Jiang X, Wei R, Zhang TL, Gu Q. Protein Pept Lett 2008;15:392–6.
[194] Li FM, Li QZ. Protein Pept Lett 2008;15:612–6.
[195] Lin H, Ding H, Feng-Biao Guo FB, Zhang AY, Huang J. Protein Pept Lett 2008;15:739–44.
[196] Wang T, Yang J, Shen HB, Chou KC. Protein Pept Lett 2008;15:915–21.
[197] Ding YS, Zhang TL. Pattern Recognit Lett 2008;29:1887–92.
[198] Garcia-Garcia A, Galvez J, de Julian-Ortiz JV, Garcia-Domenech R, Munoz C, Guna R, et al. J Biomol Screen 2005;10(3):206–14.
[199] Garcia-Garcia A, Galvez J, de Julian-Ortiz JV, Garcia-Domenech R, Munoz C, Guna R, et al. J Antimicrob Chemother 2004;53(1):65–73.
[200] Bruno-Blanch L, Galvez J, Garcia-Domenech R. Bioorg Med Chem Lett 2003;13(16):2749–54.
[201] Gozalbes R, Brun-Pascaud M, Garcia-Domenech R, Galvez J, Pierre-Marie G, Jean-Pierre D, et al. Antimicrob Agents Chemother 2000;44(10):2771–6.
[202] Meneses-Marcel A, Rivera-Borroto OM, Marrero-Ponce Y, Montero A, Machado Tugores Y, Escario JA, et al. J Biomol Screen 2008;13(8):785–94.
[203] Marrero-Ponce Y, Meneses-Marcel A, Rivera-Borroto OM, Garcia-Domenech R, De Julian-Ortiz JV, Montero A, et al. J Comput Aided Mol Des 2008;22(8):523–40.
[204] Castillo-Garit JA, Marrero-Ponce Y, Torrens F, Garcia-Domenech R, Romero-Zaldivar V. J Comput Chem 2008;29(15):2500–12.
[205] Casanola-Martin GM, Marrero-Ponce Y, Khan MT, Ather A, Sultan S, Torrens F, et al. Bioorg Med Chem 2007;15(3):1483–503.
[206] Casanola-Martin GM, Marrero-Ponce Y, Khan MT, Ather A, Khan KM, Torrens F, et al. Eur J Med Chem 2007;42(11–12):1370–81.
[207] Marrero-Ponce Y, Montero-Torres A, Zaldivar CR, Veitia MI, Perez MM, Sanchez RN. Bioorg Med Chem 2005;13(4):1293–304.
[208] Marrero-Ponce Y, Iyarreta-Veitia M, Montero-Torres A, Romero-Zaldivar C, Brandt CA, Avila PE, et al. J Chem Inf Model 2005;45(4):1082–100.