Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
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
Article history:
Received 10 February 2009
Received in revised form 6 May 2009
Accepted 14 May 2009
Available online 3 June 2009

Keywords:
Graph theory
Parasite proteomics
Leishmania

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 $\xi(L)$ 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 $\xi(L)$ values are invariants of a Markov matrix $(P)$, 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 $\xi(L)$ 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.

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.
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 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 been 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 automation 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 physicochemical 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 microbial 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 Proteone 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 TFP 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 Leishmanioses 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* [100]). If a template protein in the database has a high *M*, 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 we 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* that is not sufficiently high to accurately annotate the query protein (−40). In an excellent work have been reported an alternative to *M* 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 relies 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 Randić [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* × *n* non-negative real symmetric matrix A with elements *a* 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 $\xi(\Omega)$ values.

\[
\text{Inv}(A) = \frac{1}{n-1} \sum_{i=1}^{n} \left( \sum_{j=1}^{n} a_{ij} \right)
\]

\[
\text{Zinv}(A) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left( \sum_{j=1}^{n} a_{ij} \right)
\]

\[
D_{\text{inv}}(\text{Inv}) = \frac{1}{m} \sum_{k=1}^{m} \left| \text{Inv}(p) - \text{Inv}(q) \right|^2
\]

\[
D_{\text{zinv}}(\text{Zinv}) = \frac{1}{m} \sum_{k=1}^{m} \left| \text{Zinv}(p) - \text{Zinv}(q) \right|^2
\]

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 (electron-repulsions, 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 Ts 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 Ts 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 monoadducts with the DNA. The common 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 monoadducts with the DNA. 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 monoadducts with the DNA.
and the vector of node energies $\mathbf{E} = \{E_1, E_2, E_j, \ldots E_n\}$, the calculation of $\xi(L)$ values is straightforward to realize by means of Chapman–Kolgomorov equations; these indices can be interpreted as Mean-Energy values for a 2D Euclidean space representation for MDTs:

$$
\xi_k(L) = \pi_0^k \cdot \mathbf{H} \cdot \mathbf{e} = \pi_0^k \cdot \mathbf{H} \cdot \mathbf{e} = \sum_{i-j}^n \pi_j d_i E_j \tag{6}
$$

### 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 0°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, the 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.
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: ±0.0 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 distance-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 $\Delta E_j (j = 1, 2, 3,\ldots100)$ 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, $\Delta r$, is the maximum allowed atomic displacement used in the generation of trial configurations. The default value of $\Delta r$ 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. 3. DNA–Drug complex and some views of drug intercalation.

Fig. 4. Details of DNA–Drug intercalation.
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 \( dE_j \) (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.

\[
d_{E_j} (\text{rnd}) = d_{E_j} + \text{random}(j, \text{max}(d_{E_j}), \text{min}(d_{E_j}))
\]  

These random MDTs contain 100 energy values \( dE_j(\text{rnd}) \) 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 \( \xi = \{d_{E_1}, d_{E_2}, d_{E_3}, ... d_{E_n}\} \). 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:

\[
D_{pq}(E) = \sqrt{\sum_{j=0}^{100} (pE_j - qE_j)^2}
\]  

\[
D_{pq}(\xi) = \sqrt{\sum_{j=0}^{5} (p\xi_j(L) - q\xi_j(L))^2}
\]

| Drug | ID \text{ID}_{50}^a | Ret. \text{Ret.}^b |
|------|--------------------|-------------------|
| 1    | C Me Me H Me Me H | 0.34 [35]        |
| 2    | C H H OMe H H H   | 0.66 [30]        |
| 3    | C H CH_2OH H Me H | 0.84 [32]        |
| 4    | C Me H H Me Me H  | 0.89 [76]        |
| 5    | C H H H Me H H    | 1.00 [34]        |
| 6    | C Me H H Me H H   | 1.01 [77]        |
| 7    | C H H Me Me Me    | 1.26 [35]        |
| 8    | C Me H H Me Me H  | 1.34 [77]        |
| 9    | C H H H Me H H    | 1.52 [78]        |
| 10   | C Me H H Me Me H  | 1.79 [35]        |
| 11   | C H CH_2OH H H H  | 2.32 [32]        |
| 12   | C H Me Me H Me Me | 27.6 [30]        |
| 13   | N H H H Me Me H   | 0.13 [34]        |
| 14   | N H Me H H H H    | 0.14 [34]        |
| 15   | N H H Me Me Me Me | 0.18 [79]        |
| 16   | N H H Me Me Me Me | 0.25 [34]        |
| 17   | N Me H H Me Me H  | 0.67 [34]        |
| 18   | N H Me H Me Me H  | 0.68 [79]        |
| 19   | N H Me Me Me Me H | 0.97 [34]        |
| 20   | N Me H Me Me Me H | 1.83 [79]        |
| 21   | N H Me H Me Me H  | 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 interaction 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 calculations 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 demonstrated 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 photoreaction, 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 monoaducts 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 monoaduct 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

| Compound | Z | R₁ | R₂ | R₃ | R₄ | R₅ | R₆ | ID₅₀ᵃ | Ref.ᵇ |
|----------|---|----|----|----|----|----|----|-------|-------|
| 22       | O | –  | CO₂Me| H | H | H | H | <0.01 | [7]   |
| 23       | O | –  | CO₂Ph| H | H | H | H | <0.01 | [7]   |
| 24       | O | –  | CON(T₂₂₂₂)| H | H | H | H | <0.01 | [7]   |
| 25       | O | –  | CONH(CH₂)₂OH| H | H | H | H | <0.01 | [7]   |
| 26       | O | –  | CONH(CH₂)₂OH₂| H | H | H | H | <0.01 | [7]   |
| 27       | O | –  | CONH(CH₂)₂O| H | H | H | H | <0.01 | [7]   |
| 28       | O | –  | CON[(CH₂)₂OH]₂| H | H | H | H | <0.01 | [7]   |
| 29       | O | –  | CON(CH₂)₃NMe| H | H | H | H | <0.01 | [7]   |
| 30       | O | –  | CONH₂| H | H | H | H | 0.05  | [7]   |
| 31       | O | –  | CON(CH₂)₃O| H | H | H | H | 0.06  | [7]   |
| 32       | O | –  | CO₂H| H | H | H | H | 0.07  | [7]   |
| 33       | O | –  | CON(Me₂)| H | H | H | H | 0.07  | [7]   |
| 34       | O | –  | CO₂Me| H | H | H | H | 0.20  | [7]   |
| 35       | O | –  | Me| H | H | H | H | 0.20  | [81]  |
| 36       | O | –  | Me| H | H | H | Me | 0.03  | [35]  |
| 37       | O | –  | Me| H | H | H | Me | 0.35  | [81]  |
| 38       | O | –  | CO₂Et| H | H | H | H | 0.40  | [81]  |
| 39       | O | –  | H| H | H | H | H | 0.55  | [82]  |
| 40       | O | –  | H| Me| H | H | H | 0.55  | [35]  |
| 41       | O | –  | H| Me| H | H | CH₂OMe| Me | 0.60  | [81]  |
| 42       | O | –  | H| H | H | H | H | 0.80  | [82]  |
| 43       | O | –  | H| H | H | Me | H | 0.81  | [82]  |
| 44       | O | –  | H| Me| H | H | H | Me | 1.27  | [81]  |
| 45       | O | –  | H| H | H | Me | H | Me | 1.47  | [81]  |
| 46       | O | –  | H| Me| H | Me | H | Me | 5.30  | [82]  |
| 47       | O | –  | H| Me| H | Me | H | Me | 5.75  | [82]  |
| 48       | O | –  | H| Me| H | Me | H | Me | 5.78  | [81]  |
| 49       | N | H | H | Me| H | H | Me | 0.48  | [83]  |
| 50       | N | H | H | CH₂OH| H | Me| H | Me | 0.66  | [84]  |
| 51       | N | H | H | Me| H | Me| Me | Me | CH₂OH| 1.07  | [83]  |
| 52       | N | H | H | Me| H | Me| Me | CH₂OH| 1.36  | [83]  |
| 53       | N | Me| H | CH₂OMe| H | Me| H | Me | 2.09  | [84]  |
| 54       | N | H | H | Me| H | Me| H | Me | 2.59  | [83]  |
| 55       | N | H | H | Me| H | Me| H | Me | 4.62  | [83]  |
| 56       | N | Me| H | CH₂OH| 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 ID₅₀ relative to 8-MOP.
ᵇ References in which the activity of compounds was reported.
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 $C_3$ 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 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 $11 \tau(d)$ values for the compound and $11 \xi(L)$ 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-MBA, 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 $\mu$M). 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$_{50} < 0.1$) and 1 for the active ones (LD$_{50} > 0.1$). QSRR 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 ($r$), 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

| Protein Accession | Mass | Score | Description |
|-------------------|------|-------|-------------|
| LmjF36.6010 | 73 697 | 61 | Hypothetical |
| LmjF31.2850c | 22 350 | 42 | Ribosomal |
| CHR32_tmp.120 | 24 347 | 39 | Cyclophilin |
| CHR7-11_tmp.271 | 16 228 | 35 | Possible ubiquinone biosynthesis protein |
| CHR27_tmp.124c | 191 035 | 34 | Cytoskeleton |
| CHR7-11_tmp.32c | 18 913 | 34 | Hypothetical |
| CHR16-22_tmp.83 | 17 424 | 34 | Hypothetical |
| CHR27_tmp.35c | 17 424 | 34 | Hypothetical |
| L344.452 | 52 863 | 34 | Hypothetical |
| CHR28_tmp.181 | 65 936 | 31 | Hypothetical |
| CHR30_tmp.180c | 24 128 | 30 | Ribosomal |
| CHR26_tmp.127 | 460 680 | 30 | Hypothetical |
| LmjF25.0160c | 24 004 | 29 | Hypothetical |
| CHR32_tmp.448c | 11 907 | 29 | Brain Sjögren’s syndrome nuclear antigen |
| CHR34_tmp.181c | 11 907 | 29 | Brain Sjögren’s syndrome nuclear antigen |
| LmjF34.3570 | 66 043 | 28 | Signal recognition |
| LmjF31.1700c | 75 452 | 28 | Hypothetical |
| LmjF36.6010 | 73 697 | 28 | Putative tubulin-tyrosine ligase |
| CHR34_tmp.158 | 47 641 | 28 | Ribosomal |
| CHR34_tmp.157 | 47 743 | 28 | Ribosomal |


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

(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-INSIDE is essentially the same used to calculate \( \varepsilon(L) \) 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 \( \varepsilon(L) \) values to construct phylogenetic trees for comparison of proteins found in the PMF of a new query protein. In Fig. 5 we illustrate an overall view of the 2DE map obtained from the 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 contains 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 \( M_x \) 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 than 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 | \( M_x \) (exp) | \( M_x \) (calc) | Delta | Sequence |
|----------|---------|----------|----------------|----------------|-------|----------|
| LmjF36.6010 | P01 | 833.44 | 832.43 | 832.41 | 0.02 | AGWTVDGK |
| P02 | 877.46 | 876.46 | 876.44 | 0.02 | LEMLSR |
| P03 | 999.53 | 998.52 | 998.56 | 0.04 | RDALQLQR |
| P04 | 1317.76 | 1316.75 | 1316.67 | 0.08 | DEAQLSTRRER |
| P05 | 1405.84 | 1404.83 | 1404.73 | 0.1 | LMLTDSVPSLAR |
| P06 | 1581.93 | 1580.92 | 1580.79 | 0.13 | EKMLAQJEVTMR |
| P07 | 1626 | 1625 | 1624.86 | 0.14 | MLQHASLSDPLAARK |
| P08 | 1802.07 | 1801.06 | 1800.89 | 0.17 | IMLQAEQTVTMRAMYK |
| LmjF31.2850c | P09 | 823.44 | 822.43 | 822.39 | 0.04 | MFPAETK |
| P10 | 861.45 | 860.44 | 860.43 | 0.01 | LGAELVM |
| P11 | 867.47 | 866.46 | 866.4 | 0.06 | AMTEMRL |
| P12 | 1405.84 | 1404.83 | 1404.7 | 0.13 | DAMVVLKAEVLM |
| P13 | 1449.84 | 1448.83 | 1448.72 | 0.11 | MRQVYLCADVVR |
| CHR32_tmp.120 | P14 | 823.44 | 822.43 | 822.42 | 0.01 | TFLSAER |
| P15 | 841.43 | 840.42 | 840.41 | 0.01 | GYDVLMK |
| P16 | 910.48 | 910.45 | 910.45 | 0.03 | VEMELPK |
| P17 | 1449.84 | 1448.83 | 1448.76 | 0.07 | VEMELFIVDVPK |
| P18 | 1581.93 | 1580.92 | 1580.8 | 0.12 | MHSEALVISVFLR |
| CHR7-11_tmp.271 | P19 | 823.44 | 822.43 | 822.41 | 0.02 | EYEALAK |
| P20 | 839.44 | 838.43 | 838.39 | 0.04 | MVPYDYSK |
| P21 | 877.46 | 876.46 | 876.45 | 0.01 | ETVMGKGR |
| CHR7-12_tmp.124 | 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 | SIKSKDPIK |
| P27 | 999.53 | 998.52 | 998.48 | 0.04 | YLAAEYCGG |
| P28 | 1043.6 | 1042.6 | 1042.54 | 0.06 | DVQEA'LNR |
| P29 | 1405.84 | 1404.83 | 1404.71 | 0.12 | GSALNDRAFEVAR |
| P30 | 1449.84 | 1448.83 | 1448.71 | 0.12 | DRACQELAEVMK |
| P31 | 1493.86 | 1492.85 | 1492.73 | 0.13 | EQLPEGHSAALIAAR |
| CHR7-11_tmp.32c | P32 | 823.44 | 822.43 | 822.4 | 0.03 | AEEBYR |
| P33 | 833.44 | 832.43 | 832.4 | 0.03 | MHNLYR |
| P34 | 1043.6 | 1042.6 | 1042.55 | 0.04 | ICSVNRAEER |
| P35 | 2201 | 2199.99 | 2200.13 | 0.14 | ELTTVDAQQTJWTIPWWRVAK |
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 $\xi_{L}(L)$ 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 have shown only the 35 more interesting peptides matching with the MS of other proteins on the PMF of the new protein described above using Equation (9). For it, we have calculated first the $\xi_0(L)$ 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 $d_{pq}(L)$ between a peptide p and other q based on $\xi_{L}(L)$ 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.

Table 6

| Pept. | Sequence | $E_0$ | $\delta_0$ | $E_1$ | $\delta_1$ | $F_{100}$ | ACCR$_{100}$ |
|-------|----------|------|----------|------|----------|---------|------------|
| P01   | AGVTVDVGK | 7347.62 | 5054.22  | 28.73 | 0.10     | 111.48  | 0.47       |
| P02   | LEMLSR   | 11.73  | 10.48    | -41.42 | 0.09     | 65.72   | 0.49       |
| P03   | RMLGVQR  | -77.59 | 11.29    | -163.12 | 0.09     | 48.02   | 0.47       |
| P04   | DEAQSLTR  | 619.67 | 296.52   | -149.80 | 0.10     | -17.93  | 0.47       |
| P05   | LMLTDSVPSLR | 15036.18 | 7987.95 | -55.85 | 0.10     | 91.25   | 0.48       |
| P06   | EKIMLAEVTITMR | 577.02 | 197.41  | -79.93 | 0.09     | 91.45   | 0.48       |
| P07   | MLQHASKLSDPLAAK  | 14695.52 | 7374.20 | 64.10  | 0.10     | 263.42  | 0.47       |
| P08   | IMQTVLTDVVR | 899.72  | 220.08   | -87.59 | 0.10     | 115.66  | 0.48       |
| P09   | MFAPET  | 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   | DAMRLKGAELVM | 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 | 582631.63 | -42.04 | 0.10     | 112.65  | 0.48       |
| P18   | NHSEALISVYFLR | 1598.97  | 358.15   | -20.42 | 0.10     | 145.42  | 0.48       |
| P19   | EYEALAK | 390.84  | 181.36   | 2.04   | 0.10     | 81.61   | 0.46       |
| P20   | MVPDYDK | 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   | KAENMHLK | 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   | AACPTGCHK | 17181.88 | 11045.33 | 84.06 | 0.09     | 169.13  | 0.48       |
| P26   | SKSFDPK | 1325711.25 | 1771949.63 | 40.49 | 0.10     | 141.63  | 0.48       |
| P27   | YLAAGEYGR | 647.60  | 234.33   | -87.52 | 0.10     | 27.46   | 0.47       |
| P28   | DWQAEALVR | 486.06  | 244.65   | -88.87 | 0.10     | 20.24   | 0.48       |
| P29   | GSANDRAFEVAR | 525.10  | 213.31   | -162.52 | 0.10     | -17.81  | 0.47       |
| P30   | DRACQALEIMK | 506.82  | 211.08   | -97.64 | 0.10     | 39.80   | 0.48       |
| P31   | EQPEGHISADLAAR | 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   | MRFQVXR | 180.53  | 262.09   | 0.10   | -55.96  | 0.47       |
| P34   | IVGVRFAER | 3.65    | 14.52    | -130.72 | 0.10     | -13.62  | 0.48       |
| P35   | ELTVTDAQQTPWWRVAK | 35088.39 | 7955.64 | -108.83 | 0.10     | 130.40  | 0.47       |

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 $\xi_{L}(L)$ 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 $d_{pq}(L)$ between a peptide p and other q based on $\xi_{L}(L)$ 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 $\xi_{L}(L)$ 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:

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

$$
F = 120.99 \quad p < 0.01 \quad \rho = 0.703
$$

(11)
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 $\xi(L)$ and $\chi(L)$ values the model retained only two parameters $\xi(L)$ and $\chi(L)$. 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 data for validation. In external validation series the model correctly classifies 78 out of 80 (specificity $= 97.50\%$) optimal and 65 out of 80 (sensitivity $= 81.25\%$) no-optimal MDTs. 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 data for validation. In external validation series the model correctly classifies 78 out of 80 (specificity $= 97.50\%$) optimal and 65 out of 80 (sensitivity $= 81.25\%$) no-optimal MDTs. 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 data for validation. In external validation series the model correctly classifies 78 out of 80 (specificity $= 97.50\%$) optimal and 65 out of 80 (sensitivity $= 81.25\%$) no-optimal MDTs. 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 data for validation. In external validation series the model correctly classifies 78 out of 80 (specificity $= 97.50\%$) optimal and 65 out of 80 (sensitivity $= 81.25\%$) no-optimal MDTs. 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.
