

Cite this: *Mol. BioSyst.*, 2011, **7**, 1938–1955www.rsc.org/molecularbiosystems

PAPER

MISS-Prot: web server for self/non-self discrimination of protein residue networks in parasites; theory and experiments in *Fasciola* peptides and *Anisakis* allergens^{†‡}**Humberto González-Díaz,^{*a} Laura Muíño,^a Ana M. Anadón,^a Fernanda Romaris,^a Francisco J. Prado-Prado,^b Cristian R. Munteanu,^c Julián Dorado,^c Alejandro Pazos Sierra,^c Mercedes Mezo,^d Marta González-Warleta,^d Teresa Gárate^e and Florencio M. Ubeira^a***Received 21st February 2011, Accepted 7th March 2011*

DOI: 10.1039/c1mb05069a

Infections caused by human parasites (HPs) affect the poorest 500 million people worldwide but chemotherapy has become expensive, toxic, and/or less effective due to drug resistance. On the other hand, many 3D structures in Protein Data Bank (PDB) remain without function annotation. We need theoretical models to quickly predict biologically relevant Parasite Self Proteins (PSP), which are expressed differentially in a given parasite and are dissimilar to proteins expressed in other parasites and have a high probability to become new vaccines (unique sequence) or drug targets (unique 3D structure). We present herein a model for PSPs in eight different HPs (*Ascaris*, *Entamoeba*, *Fasciola*, *Giardia*, *Leishmania*, *Plasmodium*, *Trypanosoma*, and *Toxoplasma*) with 90% accuracy for 15 341 training and validation cases. The model combines protein residue networks, Markov Chain Models (MCM) and Artificial Neural Networks (ANN). The input parameters are the spectral moments of the Markov transition matrix for electrostatic interactions associated with the protein residue complex network calculated with the MARCH-INSIDE software. We implemented this model in a new web-server called MISS-Prot (MARCH-INSIDE Scores for Self-Proteins). MISS-Prot was programmed using PHP/HTML/Python and MARCH-INSIDE routines and is freely available at: <http://miaja.tic.udc.es/Bio-AIMS/MISS-Prot.php>. This server is easy to use by non-experts in Bioinformatics who can carry out automatic online upload and prediction with 3D structures deposited at PDB (mode 1). We can also study outcomes of Peptide Mass Fingerprinting (PMFs) and MS/MS for query proteins with unknown 3D structures (mode 2). We illustrated the use of MISS-Prot in experimental and/or theoretical studies of peptides from *Fasciola hepatica* cathepsin proteases or present on 10 *Anisakis simplex* allergens (Ani s 1 to Ani s 10). In doing so, we combined electrophoresis (1DE), MALDI-TOF Mass Spectroscopy, and MASCOT to seek sequences, Molecular Mechanics + Molecular Dynamics (MM/MD) to generate 3D structures and MISS-Prot to predict PSP scores. MISS-Prot also allows the prediction of PSP proteins in 16 additional species including parasite hosts, fungi pathogens, disease transmission vectors, and biotechnologically relevant organisms.

^a Department of Microbiology & Parasitology, Faculty of Pharmacy, University of Santiago de Compostela, 15782, Santiago de Compostela, Spain. E-mail: gonzalezdiazh@yahoo.es^b Department of Organic Chemistry, Faculty of Pharmacy, University of Santiago de Compostela, 15782, Santiago de Compostela, Spain^c Department of Information and Communication Technologies, Computer Science Faculty, University of A Coruña, Campus de Elviña, 15071, A Coruña, Spain^d Laboratorio de Parasitología, Centro de Investigaciones Agrarias de Mabegondo - Xunta de Galicia, 15318, Abegondo, A Coruña, Spain^e Servicio de Parasitología, Centro Nacional de Microbiología, Instituto de Salud Carlos III, Majadahonda, Madrid, 28220, Spain[†] Authors contributions: HGD and FMU: design, protein 3D-QSAR study and statistical analysis, web server design, MM/MD experiment, experimental-theoretic study of *F. hepatica* and *A. simplex* peptides, paper redaction, and critical revision. LMP, AMA, FF, MM, MGW, and TG: experimental-theoretic study of *F. hepatica* peptides. AMA and FF: theoretical study of *A. simplex* peptides. FJPP: MM/MD studies, protein 3D-QSAR study and statistical analysis. CRM, JD, and APS: web server implementation.[‡] Electronic supplementary information (ESI) available. See DOI: 10.1039/c1mb05069a

Introduction

There are many different species of Human Parasites (HPs), whose members cause important parasitic diseases in humans and animals, but the number of available drugs to treat them is very limited. In order to increase the available therapeutic arsenal, a goal of major importance may be the development of computational software that can predict the therapeutic potential of new drugs or of drug-target proteins in different parasites.¹ Predicting proteins differentially expressed in only one parasitic organism (self proteins) that are not present in the corresponding host may become a gateway to seek targets with unique drug-binding pockets. These proteins are reasonably expected to be good candidates for finding new drug targets, so they may be considered as proteins with high targetability. A possible strategy to discriminate between self and non-self protein sequences is to use the methodology we previously used for drug target/non-drug target discrimination that may be complemented with the development of a free web-tool to aid researchers to perform the necessary calculus. With this tool one can apply sequence alignment procedures to predict the function and target scores (in sequence terms) of these proteins or peptides such as the very successful BLAST-like procedures.^{2–8} However, taking into account that the techniques relying on alignment may fail because there are no similar candidates in the databases or the most similar templates have not been annotated,⁹ we propose the use of alignment-free Machine Learning methods for predicting the protein functional class based on sequence (1D) parameters, but regardless of the sequence–sequence similarity.^{10–13} For instance, the so-called pseudo amino acid composition indices and other predictive methods have been introduced by Chou *et al.*, Cai and other authors, to predict the function of proteins irrespective of sequence similarity.^{14–22} Moreover, in terms of drug targetability it is very important to consider not only the primary sequence but also the 3D structure of the protein. Consequently, some groups have used structural parameters with spatial information and the Support Vector Machine (SVM) to predict protein function from 3D structures.^{23,24} This methodology is very similar to Quantitative Structure–Activity/Property (QSAR/QSPR) methods used for drugs but applied to large protein and/or drug–protein structures. Generally speaking, a QSAR/QSPR study uses more or less detailed structural indices (composition, sequence, 2D, and/or 3D) of molecular systems (drugs, RNAs, proteins, protein–protein complexes) to predict the biological activity or desired properties, which are usually expensive to be measured experimentally, especially for large databases. With this aim, QSAR/QSPR make use of Statistical, Artificial Intelligence (AI) including Artificial Neural Networks (ANNs) and/or Machine Learning (ML) methods to correlate the structural indices with the property of the systems, obtaining predictor models with different degrees of complexity. Regardless of the type of system described, the structural descriptor used, or the learning method selected, there are a series of common steps for almost all QSAR/QSPR studies, see recent reviews on this topic.^{25,26} A versatile QSAR/QSPR method was introduced by González-Díaz *et al.* as Markovian Chemicals *In silico* Design (MARCH-INSIDE 1.0) for the

computational design of small-sized drugs. In successive studies, we have extended this method to perform fast calculation of 2D and 3D alignment-free numeric parameters (such as entropy, average potentials, and spectral moments) to describe more complex systems. Currently, the method has been renamed as Markov Chains Invariants for Networks Simulation & Design (MARCH-INSIDE 2.0). The approach uses a Markov Chain Model (MCM) to calculate parameters of small-sized and complex chemical structures.^{27–29} To this end, MARCH-INSIDE describes the system as a stochastic matrix of interactions and/or transitions between the parts of the system and associates this matrix with a graph or complex network representation of this system, at the same time. This describes more adequately the broad use of the method to numerically characterize the structure of drugs,³⁰ RNA,³¹ and proteins,^{32–34} as well as drug–drug networks,³⁵ drug–protein interactions,³⁶ PPIs, and other systems such as a MCM associated to a graph. MARCH-INSIDE uses networks similar to others, known in Proteomics, Molecular Biology and Molecular Microbiology, where the nodes (connected by links) are atoms (bonds), amino acids (electrostatic interactions), proteins (PPIs), genes (co-expression), organisms and microorganisms (parasite–host interactions).^{37–44} In Fig. 1 we depict: (A) the 3D structure model for a section of a complex made up of human cathepsin L (PDB-ID 3HHA) and (B) the respective complex network graph representation of the protein structure. At this structural level, the nodes are amino acids and we link two nodes with an edge if the distance between them is less than 7 Å⁴⁵ (an optimal cut-off value, recently determined by a scan study). This type of network is also known as contact map or protein residue network.^{46–53} In a very recent review, we have discussed the details and many applications of the MARCH-INSIDE method based on a Markov transition matrix of a complex network to discover both drugs and targets in Molecular Microbiology and Parasitology.⁵⁴ The last upgrade of MARCH-INSIDE (carried out by Munteanu and González-Díaz) was the implementation of the Internet portal Bio-AIMS (<http://miaja.tic.udc.es/Bio-AIMS/>) with different web-server packages that may be used to predict different functions of proteins from PDB files. These servers are inspired by the same philosophy of online free access and are used by all the international research communities, as mentioned in the previous paragraph. In particular, the package TargetPred from the server offers two new protein-QSAR servers. The first one, ATCUNPred (<http://miaja.tic.udc.es/Bio-AIMS/ATCUNPred.php>), is available for prediction of ATCUN-mediated DNA-cleavage anticancer proteins.⁵⁵ The second server, EnzClassPred, is available at <http://miaja.tic.udc.es/Bio-AIMS/EnzClassPred.php> and may be used for prediction of enzyme classes from PDB files without function annotation.⁵⁶

In this work, we approached the QSAR/QSPR problem of 3D-protein selfness in parasites using the MARCH-INSIDE method. Our aim is to seek a QSAR/QSPR model discriminating parasite self proteins or peptides (PSPs) that differentially express in one out of nine different HPs genera from non-self proteins (nPSP) based on a protein 3D structure. The model may be used also to predict the contribution of a peptide to increase/decrease the probability of a self protein

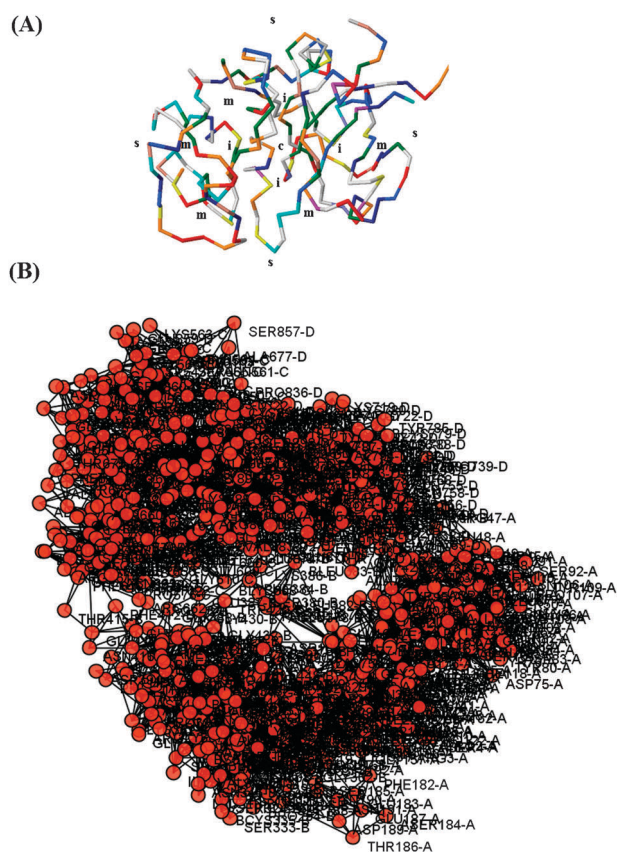


Fig. 1 Representations of a cathepsin L (PDB-ID 3HHA): (A) 3D structure and (B) complex network.

(3D structure) to be a unique target not present in other organisms (targetability). We implemented the predictor in a new web server named MISS-Prot. The expansion of this acronym is MISS-Prot: MARCH-INSIDE Scores for Self-Proteins. MISS-Prot is freely available to public at <http://miaja.tic.udc.es/Bio-AIMS/MISS-Prot.php>. Last but not least, we illustrated the use of this model with a real case study. In this regard, we reported the 1DE experimental isolation, MALDI-TOF MS characterization, MASCOT, BLAST, and QSAR prediction of the peptides identified by MS/MS analysis of a L1-cathepsin found in *F. hepatica*. We also report the first combined study MM/MD + MISS-Prot of peptides present in known allergens of *Anisakis simplex*. In Fig. 2 we depicted a flowchart within all the steps that we took in this work to construct the new classifiers and server.

Materials and methods

Theoretic methods

MARCH-INSIDE parameters of protein structure. In previous works we have predicted the protein function based on different protein structural parameters derived from a Markov matrix that accounts for electrostatic interactions between amino acid pairs in the 3D structure of the protein. One of the classes of parameters used was called the Spectral Moments $\pi_k(R)$ of the matrix. These values are used herein as inputs to construct the protein QSAR model for PPs.

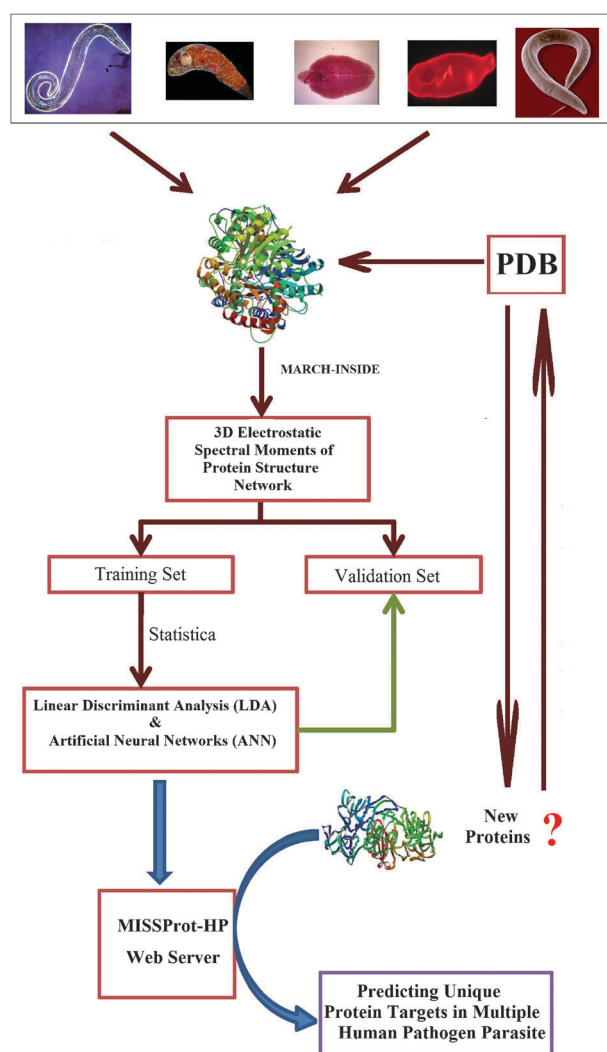


Fig. 2 Flowchart for all the steps given in the construction of the classifiers and server.

The detailed explanation has been published before^{28,29,32,57–62} and reviewed in detail recently.⁶³ As follows, we give the formula for $\pi_k(R)$ values and some general explanations:

$$\pi_k(R) = \sum_{i=j \in R}^k p_{ij}(R) \quad (1)$$

It is remarkable that the spectral moments are the direct sum of the probabilities $p_{ij}(R)$ with which the effect of the electrostatic interaction propagates from the amino acid i th to other amino acids j th next to it and returns to i th after k -steps. These probabilities refer to amino acids considered isolated in the space ($k = 0$), direct interactions between amino acids and others in contact with them in the residue network ($k = 1$) and spatial ($k > 1$) indirect interactions between amino acids placed at a distance equal to k -times the cut-off distance ($r_{ij} = k \times r_{\text{cut-off}}$). The method uses a Markov Chain Model (MCM) to calculate these probabilities. However, for the sake of simplicity, a truncation or cut-off function α_{ij} is applied in such a way that a short-term interaction takes place in a first approximation only between neighbouring amino acids ($\alpha_{ij} = 1$ if $r_{ij} < r_{\text{cut-off}}$). Otherwise, the interaction is vanished

($\alpha_{ij} = 0$). The relationship α_{ij} may be visualised in the form of a protein structure complex network. In this network the nodes are the C_α atoms of the amino acids and the edges connect pairs of amino acids with $\alpha_{ij} = 1$. Euclidean 3D space $r_3 = (x, y, z)$ coordinates of the C_α atoms of amino acids listed in protein PDB files. For the calculation, all water molecules and metal ions were removed.⁵⁴ All calculations were carried out with our in-house software MARCH-INSIDE 2.0.⁵⁴ For the calculation, the MARCH-INSIDE software always uses the full matrix, never a sub-matrix, but the last summation term may run either for all amino acids or only for some specific protein regions (R). These regions or orbits are often defined in geometric terms and called core, inner, middle or surface region. In Fig. 1 we denote the regions of the protein (c corresponds to core, i to inner, m to middle, and s to surface orbits, respectively). The diameters of the orbits, as a percentage of the longest distance $r_{\max}$ with respect to the centre of charge, are the following: $0 \leq c \leq 25$, $25 < i \leq 50$, $50 < m \leq 75$, $75 < s \leq 100$. Additionally, we consider the total orbit (t) that contains all the amino acids in the protein (orbit diameter 0 to 100% of $r_{\max}$). Consequently, we can calculate different $\pi_k(R)$ for the amino acids contained in the regions (c, i, m, s , or t) and placed at a topological distance k from each other within this orbit (k is the order).^{27–29,33,64} In this work, we have calculated altogether $5(\text{types of regions}) \times 6(\text{orders considered}) = 30$ $\pi_k(R)$ indices for each PPp or nPPp.

Dataset. The protein structures were downloaded from PDB⁶⁵ using the following schemes for PDB-database search: (i) introducing as input parameter the name of the different parasites (*Plasmodium*, *Fasciola* spp., *Ascaris* spp., *Giardia* spp., *Trypanosoma* spp., *Toxoplasma*, *Entamoeba* spp.) in the search item called source organism (for positive cases). The second alternative (ii) was to introduce the PDB IDs for all the proteins contained in the list reported in the article by Dobson and Doig,⁶⁶ which are considered as negative cases after curation to confirm the source organism. The dataset consists of 15 341 cases (protein/organism pairs) including 1733 positive cases referred to PSPs and 13 608 cases referred to nPSPs. The list covers more than 20 organisms, including bacteria, virus, fungi, and parasites, as well as their human or cattle hosts. In total 1300 PSPs and 10 212 nPSPs were used to train the model. In addition, 433 PSPs and 3396 nPSPs were used in the external validation set. Detailed information about the PDB ID, the values of the electrostatic spectral moment indices, the corresponding observed classification, and the predicted classification for each PSP or nPSP protein/organism pair are given in Table S1 of the ESI†.

Alignment-free 3D-QSAR analysis of PSPs. Linear Discriminant Analysis (LDA) and Artificial Neural Networks (ANN) were used to construct the classifiers. One of the most important steps in this work was the organisation of the spreadsheet containing the raw data used as inputs for the LDA and for the ANN because this is not a classic classifier. Herein, the schematisation of the paper is peculiar. Our expectation is to use a two-group discriminant function to classify proteins into two possible groups: proteins that belong to a particular group and proteins that do not belong to this

group. To this end, we have to indicate somehow which group we pretend to predict in each case. In this regard, we took the following steps that have been explained in detail in recent works:^{56,67}

(1) First of all, the PSPs (proteins expressed in at least one HP) were divided according to their genus membership but without specifying the species. The genus of the HP that differentially expresses one PSP is the first term in the genus species taxonomy classification of the source organism; this information was obtained from the PDB file of the protein.⁶⁵

(2) We created raw data representing each protein item as a vector made up of 90 3D-structural variables (inputs), one dependent variable (PSPg), and two auxiliary variables Genus observed (Go) and Genus Query (Gq); as auxiliary variables, Go and Gq were never used in the analysis.

(3) Using $G_o = o$ we express that a given protein is a PSP, differentially expressed by at least one parasite organism that belongs to the observed genus o .

(4) Using $G_q = q$ we express that we are using the model to predict whether a given protein is a PSP or not (is differentially expressed by) at least one parasitic organism that belongs to the query genus q .

(5) The dependent variable PSPg is a dummy variable (Boolean); $PSPg = \text{yes}$ if $G_q = G_o = g$ (the PSP protein is differentially expressed by a parasite with the observed genus equal to the query genus; otherwise $PSPg = \text{no}$). It means that we can repeat each protein more than once in the raw data. In fact, we could repeat each protein 9 times corresponding to 9 different G_q or $G_o = g = 1, 2, 3 \dots 9$ values found in these data. In an alphabetical order, we have the following possible G_o and G_q values: 1-*Ascaris*, 2-*Entamoeba*, 3-*Fasciola*, 4-*Giardia*, 5-*Leishmania*, 6-*Tolypocladium*, 7-*Toxoplasma*, 8-*Trypanosoma*, and 9-*Plasmodium*.

(6) In order to sort data, the first time that a protein appeared we used the $G_q = G_o$, and consequently $PSPg = \text{yes}$. It means that we used in G_q the real genus (G_o) of the organism that expresses this protein. In this case, the LDA model had to give the highest probability to the group $PSPg = \text{yes}$ because it had to predict the real organism that expresses the protein.

(7) The other times we used $G_q \neq G_o$ (query genus different from the real or observed genus), and the LDA model had to predict the highest probability for the group $PSPg = \text{no}$, meaning that this is not a PSP for the parasite of this genus. Conversely, non-PSP proteins (or simply nPSPs) have more than one line entry for each protein (selected at random) with different G_q values in a proportion of above 10 nPSPs (decoys) by each PSP.

(8) The problem in this type of organisation of raw data is that 30 $\pi_k(R)$ values are protein constants. Consequently, if LDA is based only on these values, the model will necessarily fail when we change G_q values. An inconvenience in this regard occurs if we pretend to use the model for a real PSP, since we have only one unspecific prediction and we need 9 specific probabilities (9 possible G_q values and only one real $G_q = G_o$). It means that, for a PSP, a perfect classifier has to predict a probability equal to 1 (or approximately 1) confirming the real G_o of the organism that expresses this protein and

8 probability values approximately equal to 0 for the other possible values of Gq.

(9) We can solve this problem by introducing characteristic variables of each Gq but without giving any information in the input about the Go for the organism that expresses a given protein. To this end, we used the average values $\pi_k(R)_g$ of each $\pi_k(R)$ for all PSPs expressed in the organisms of the same genus. We also calculated the deviation ($\Delta\pi_k(R)_g = \pi_k(R) - \pi_k(R)_g$) of the $\pi_k(R)$ from the $\pi_k(R)_g$ for the respective genus indicated in Gq. Altogether, we have then $(30 \pi_k(R)$ values of the PSP) + $(30 \pi_k(R)$ average values for each Gq) + $(30 \Delta\pi_k(R)_g$ values of deviations for each PSP from the average of the respective Gq) = 90 input variables.

(10) It is of major importance to understand that we never used as input Go or Gq, so the model only includes as input the $\pi_k(R)$ values for the protein entry and the average and deviations of these values from the Gq, which is not necessarily the real Go. The general formula for this class of the LDA model is shown below, where $S_g(r_3)$ (the model output) is not the dependent variable PSPg and neither is the output probability. $S_g(r_3)$ is a real valued score that predicts the propensity of a protein to act as a PSP differentially expressed in one HP organism that belongs to genus Gq = g:

$$\begin{aligned}
 S_g(r_3) &= a_0 + \sum_{k=0, o=0}^{k=5, o=4} b_k \cdot \pi_k(R) + \sum_{k=0, o=0}^{k=5, o=4} c_k \cdot \pi_k(R)_g \\
 &\quad + \sum_{k=0, o=0}^{k=5, o=4} d_k \cdot (\pi_k(R) - \pi_k(R)_g) \\
 &= a_0 + \sum_{k=0, o=0}^{k=5, o=4} b_k \cdot \pi_k(R) + \sum_{k=0, o=0}^{k=5, o=4} c_k \cdot \pi_k(R)_g \\
 &\quad + \sum_{k=0, o=0}^{k=5, o=4} d_k \cdot \Delta\pi_k(R)_g
 \end{aligned} \quad (2)$$

In fact, the variable $S_g(r_3)$ is the output of the LDA model and may be characterised as a real value score proportional to the probability with which the organisms of genus Gq with average values $\pi_k(R)_g$ differentially expresses the PSP protein p with structural parameters $\pi_k(R)$ and deviation $\Delta\pi_k(R)_g$ from the training sub-set of proteins expressed in this genus. LDA forward stepwise analysis was carried out for variable selection to build up the models.³⁰ All the variables included in the model were standardized in order to bring them onto the same scale.⁶⁸ The square of the canonical regression coefficient (Rc) and Wilk's statistics (U) were examined in order to assess the discriminatory power of the model ($U = 0$ perfect discrimination, being $0 < U < 1$); the separation of the two groups of proteins was statistically verified by the Fisher ratio (F) test with an error level $p < 0.05$.

BLAST alignment analysis of PSPs. The most relevant peptide fragments of the new protein were submitted to BLASTp to determine the similarity of the sequence with other known proteins.⁶⁹ The BLASTp procedure was carried out with the non-redundant NCI database such as the query database, allowing BLASTp to search for conserved domains through the CD search tool.⁷⁰ We calculated

$S_g(r_1) = (\text{BLAST score in bits})/100$ for the protein in the database expressed in one parasite of genus g, with the highest similarity to the peptide, and a known biological function.

Calculation of PSP probabilities. The score values $S_g(r_1)$ and $S_g(r_3)$ calculated with different alignment procedures and/or QSAR classifiers have different scales. This situation determines that different scores are not directly comparable. To avoid this problem, we did not use directly the scoring values $S_g(r_1)$ and $S_g(r_3)$ but transformed them into a probability scale. In the particular case of 3D-QSAR derived by LDA, we can use Mahalanobis distances to calculate the probabilities $p_g(r_3)$ with which an organism of genus g differentially expresses different PSPs or peptides with respect to organisms of the other genus. Unfortunately, we cannot directly obtain Mahalanobis distances to calculate $p_g(r_1)$ or $p_g(r_3)$ probability values from outcomes obtained by BLAST alignment or 3D-QSAR models based on ANN models.^{67,71} For these practical reasons we suggest using the same probability scale for all classes of scores. In this work, we suggest transforming the score values into probability values as follows:

$$p_g(r_n) = \frac{S_g(r_n) + |S_g(r_n)_{\min}| + 1}{\sum_{q=1}^{q=8} S_g(r_n) + |S_g(r_n)_{\min}| + 1} \quad (3)$$

LOMETS study. LOMETS is a locally installed meta-server method for protein structure prediction. It generates protein structure predictions by ranking and selecting models from 8 state-of-the-art threading programs. LOMETS implementation, validation, and use have been published before.⁷² In this work we used LOMETS to generate the 3D structure of peptides of *F. hepatica* confirmed by MS/MS (see experiment 1), which are necessary as inputs of MISS-Prot to predict the selfness score with mode 2U. First, we introduced the sequence of the peptides in the LOMETS input box. In the setup we kept the options implemented by default. The model ranking scheme of LOMETS is as follows. For a given target, 160 models are generated by 8 component servers where each server generates 20 models as sorted by their Z-scores in each algorithm. The best 10 models are finally selected from the 160 models based on the following scoring function:

$$\text{score}(i,j) = (Z(i,j)/Z_0(i)) \cdot \text{conf}(i) + \text{seq-id}(i,j) \quad (4)$$

where $Z(i, j)$ is the Z-score of the j th model of the i th server, $Z_0(i)$ is the cutoff of the i th server, $\text{conf}(i)$ is the confidence of the i th server, which is defined as the average TM-score to the native of all predictions in a large-scale benchmark test. $\text{Seq-id}(i,j)$ is the sequence identity to query of the j th model of the i th server. We list as follows the parameters $Z_0(i)$ ($\text{conf}(i)$) for each server (i): 1-FUGUE 7.0 (0.383),⁷³ 2-HHsearch 9.5 (0.439),⁷⁴ 3-MUSTER 5.8 (0.454),⁷⁵ 4-PROSPECT 23.2 (0.446),⁷⁶ 5-PPA-I 9.3 (0.447),⁷² 6-SAM 9.5 (0.377),⁷⁷ 7-SP 37.0 (0.437),⁷⁸ 8-SPARKS 7.0 (0.429).⁷⁹ Finally, the models are ranked based on the TM-score, *i.e.* the models with higher average TM-score to other models are

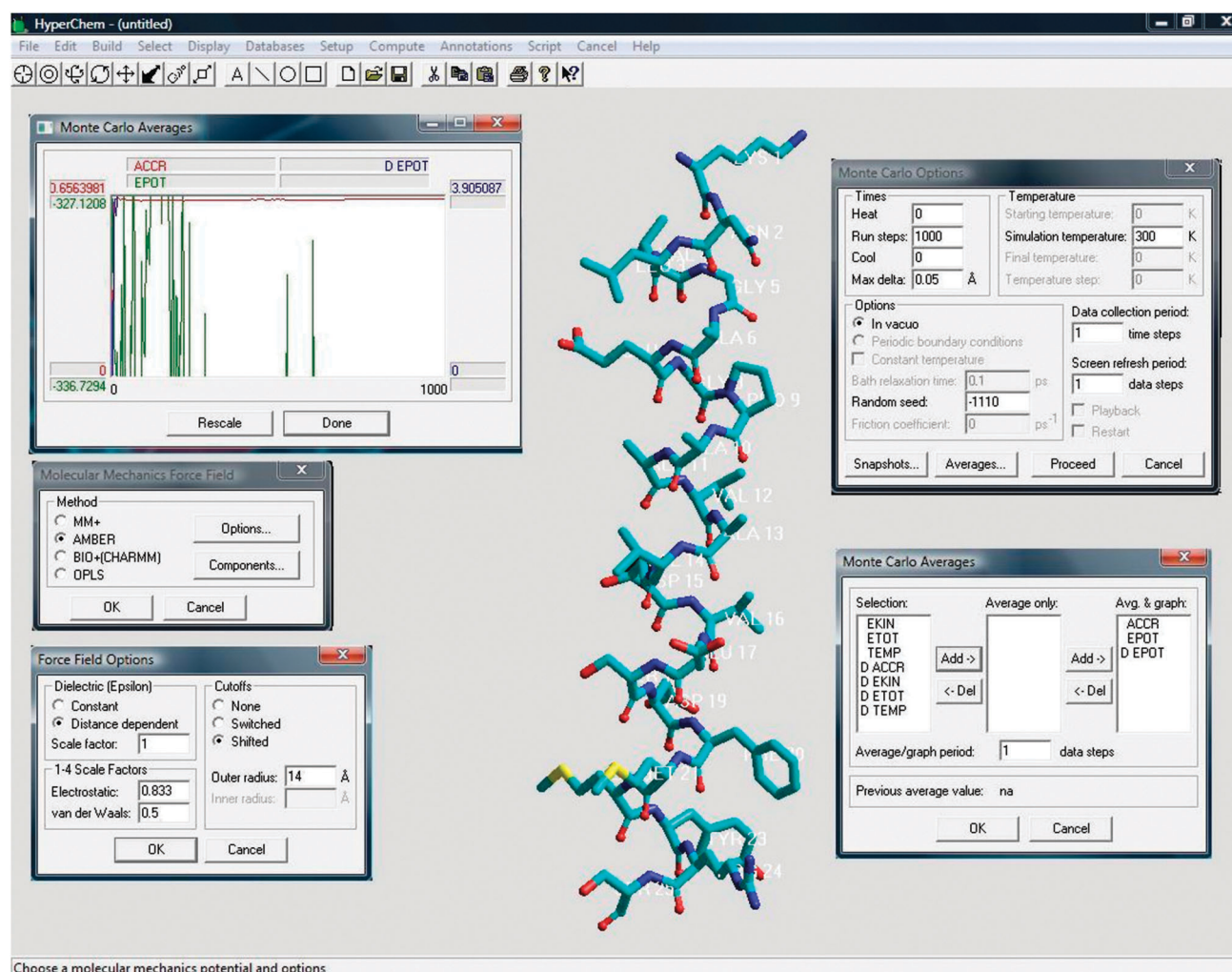


Fig. 3 Snapshot of the software HyperChem showing MM/MD specifications used.

ranked higher.⁷² LOMETS also consider the following format of spatial restraints:

1. Side-chain contacts: describes the predicted side chain contacts between pairs of residues. The first line indicates the total number of pairs of predicted side chain contacts. Columns 1 and 2 are residue orders of predicted pairs. Column 3 is the confidence of the contact, which is defined as the relative frequency of contacts appearing in the top 20 templates.

2. C_{α} atom contacts: describes the predicted C_{α} atom contacts between pairs of residues. The first line describes the total number of pairs of predicted C_{α} contacts. Columns 1 and 2 are residue orders of predicted pairs. Column 3 is the confidence of contact predictions.

3. Short-range C_{α} distances: describes the predicted C_{α} distances in short-range, i.e., $|i - j| < 7$. The first line is the total number of short-range C_{α} atom pairs. Columns 1 and 2 are residue orders of predicted pairs. Column 4 is the predicted distance between the C_{α} atom pairs. Column 5 is the standard deviation of the distance predictions. All distances are in units of Angstrom.

4. Long-range C_{α} distances: describes the predicted C_{α} distances in a long-range, i.e., $|i - j| > 6$. The first line describes the total number of long-range C_{α} atom pairs.

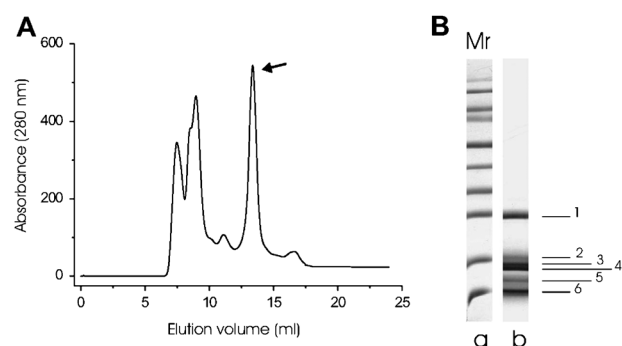


Fig. 4 ESAs of *F. hepatica*: (A) size-exclusion chromatogram obtained by FPLC separation of ESAs from *F. hepatica*. The arrow indicates peak number IV of the chromatogram according to Mezo *et al.*,⁸⁴ (B) SDS-PAGE analysis of proteins contained in peak number IV (lane b). The proteins present in band 1 were analysed by MS/MS spectrometry. The relative molecular masses of standard markers (lane a) from top to bottom were: 170, 130, 95, 72, 55, 43, 34, 26, 17 and 10 kDa.

Columns 1 and 2 are residue orders of predicted pairs. Column 3 is the predicted distance between C_{α} atom pairs. There are up to 4 values of distances for each pair, which are taken from the top 4 template alignments.

HyperChem study. This is a Molecular Mechanics/Molecular Dynamics (MM/MD) study using the program HyperChem. HyperChem is a program for general use in molecular modeling that can be used for protein structure prediction. In this work we used HyperChem to generate the 3D structures of peptides of allergens present in *A. simplex* (see experiment 2), which were necessary as inputs of MISS-Prot to predict the selfness score with mode 2. First, we introduced the sequence of the 30 peptides in the HyperChem sequence editor⁸⁰ and carried out a geometry optimisation with the AMBER Molecular Mechanics (MM) force field.⁸¹ In the setup we kept the options implemented by default but allowing a shifted cut-off truncation with $r_{\text{out}} = 14 \text{ \AA}$ (outer radius). Next, Molecular Dynamics (MD) simulation was carried out using a Monte Carlo (MC) algorithm *in vacuo* at 300 K and 1000 optimisation stages (run steps) obtaining 30 Molecular Dynamics Trajectories (MDTs) or energetic profiles, one for each peptide, with 1000 potential energy values, each one corresponding to 1000 different conformations of the same peptide: ${}^dE_j(r_3)$ with $j = 1, 2, 3, \dots, 1000$.^{82,83} To this end, we also used the force field AMBER⁸¹ with the distant-dependent dielectric constant (scale factor 1), electrostatic and Van der Waals values by default and the same cut-off used for optimisation. All the other specifications for the force field were set equal to HyperChem default values (see Fig. 3) and the atom type was recalculated keeping their current charges. In order to obtain realistic geometries, there is an additional parameter that we monitored 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 dialogue box of HyperChem (see also Fig. 3). The ACCR is a running average of the number ratio of accepted moves to attempted moves. Varying the step size can have a large effect on the ACCR value. The step size, Δr , is the maximum allowed atomic displacement used in the generation of trial configurations. The default value of r in HyperChem is $0.05 \text{ \AA}$.⁸² For most organic molecules, this will result in an ACCR of about 0.5, which means that about 50% of all the moves are accepted.

Experimental methods

Obtention of excretory–secretory antigens (ESAs) of *F. hepatica*. ESAs of *F. hepatica* were obtained as previously described by Mezo *et al.*⁸⁴ Briefly, liver adult flukes collected from bile ducts of naturally infected cows were washed twice, first in sterile saline solution containing antibiotics (100 IU ml^{-1} penicillin and $100 \text{ }\mu\text{g ml}^{-1}$ streptomycin) and glucose (2 mg ml^{-1}) at 38°C , and then in Roswell Park Memorial Institute cell culture medium (RPMI 1640 supplemented with 20 mM HEPES, 0.3 g L^{-1} L-glutamine, 2 g L^{-1} sodium bicarbonate and antibiotics) at 38°C under $5\% \text{ CO}_2$ in air. Flukes were then transferred to a 75 cm^2 tissue culture flask (Iwaki, Sciteck Div. Asahi Techno Glass, Funabashi City, Chiba, Japan) and maintained in a culture medium (3 ml per fluke) at 38°C under $5\% \text{ CO}_2$ in air. After 24 h of incubation, the medium containing ESAs was removed, the protease activity inhibited with a protease inhibitor cocktail (SigmaFAST™ Protease Cocktail Tablet, Sigma-Aldrich, Madrid, Spain), and centrifuged at $10000 \times g$ for 20 min at 4°C . Then,

the supernatant was passed through a $0.45 \text{ }\mu\text{m}$ pore filter disc, and submitted to a process of washing/concentration with distilled water using Microsep Microconcentrators (3000 molecular weight cut-off; Filtron Technology Corporation, Northborough, Massachusetts) and the concentrate dried using a vacuum concentrator (SpeedVac, Thermo Scientific, Barcelona, Spain) and stored at 4°C until use.

Size-exclusion chromatography (SEC). Samples of ESAs of *F. hepatica* were dissolved in PBS at a final protein concentration of 3.5 mg ml^{-1} and fractionated by size-exclusion chromatography using an FPLC system (ÄKTA Basic 10, Amersham Biosciences Europe BmbH, Barcelona, Spain) equipped with a column of Superdex 75 HR 10/30 column (Amersham Biosciences) as described previously.⁸⁴ Briefly, 0.5 ml samples were eluted at a flow rate of 0.3 ml min^{-1} and the protein concentration was monitored at 280 nm . The molecular weight of the eluted proteins was estimated using a mixture of proteins of known molecular weight (Gel Filtration LMW Calibration Kit, Amersham Biosciences). Finally, the protein concentration of each peak obtained was measured by the Pyrogallol Red Method (Sigma) and stored in aliquots at -80°C until further analysis.

Polyacrylamide gel electrophoresis (SDS-PAGE). The proteins from an aliquot of purified *F. hepatica* ESAs corresponding to peak IV (see Fig. 4A), obtained as described above by exclusion chromatography, were resolved by SDS-PAGE under reducing conditions ($2.5 \text{ }\mu\text{g}$ of protein per lane) using a mini-vertical gel electrophoresis unit (Mighty Small II, Hoefer Inc., Holliston, MA). One-dimensional electrophoresis (1DE) was performed at 200 V constant current at 20°C for 1 h using the Laemmli (1970) buffer system with discontinuous slab gels made up of a 5% polyacrylamide stacking gel and a $10\text{--}20\%$ linear-gradient polyacrylamide resolving gel of 1 mm thickness. 1D-SDS-PAGE molecular weight standards (BioRad) were used to calibrate gels.

Mass spectrometry (MS). Proteins in the polyacrylamide gels were stained with Imperial Protein Stain (Pierce) overnight and the excess dye removed with distilled water. Then the selected bands (band #1 in this study; see Fig. 4B) were excised using a scalpel, washed twice with water, shrunk 15 min with 100% acetonitrile and dried in a Savant SpeedVac for 30 min. Then, the samples were reduced with 10 mM dithioerythritol in 25 mM ammonium bicarbonate for 30 min at 56°C and subsequently alkylated with 55 mM iodoacetamide in 25 mM ammonium bicarbonate for 15 min in the dark. Finally, samples were digested with $12.5 \text{ ng }\mu\text{l}^{-1}$ sequencing grade trypsin (Roche Molecular Biochemicals) in 25 mM ammonium bicarbonate ($\text{pH } 8.5$) overnight at 37°C . After digestion, the supernatant was collected and $1 \text{ }\mu\text{l}$ was spotted onto a MALDI target plate and allowed to air-dry at room temperature. Then, $0.4 \text{ }\mu\text{l}$ of 3 mg ml^{-1} of the α -cyano-4-hydroxy-cinnamic acid matrix (Sigma-Aldrich) in 50% acetonitrile were added to the dried peptide digest spots and allowed again to air-dry at room temperature.

MALDI-TOF MS analyses were performed in a 4800 Proteomics Analyzer MALDI-TOF/TOF mass spectrometer (Applied Biosystems, Framingham, MA) at the Genomics and

Proteomics Center, Universidad Complutense de Madrid, operated in positive reflector mode, with an accelerating voltage of 20 000 V. All mass spectra were calibrated internally using peptides from the auto digestion of trypsin. The MALDI-TOF/TOF mass spectrometry analysis produces peptide mass fingerprints. Those observed with a signal-to-noise greater than 10 were collated and represented as a list of monoisotopic molecular weights. To confirm protein identification, peptides corresponding to some selected PMF peaks were subjected to MS/MS sequencing analyses using the 4800 Proteomics Analyzer (Applied Biosystems, Framingham, MA). The suitable MS spectra precursors for MS/MS sequencing analyses were selected, and fragmentation was carried out using the acquisition method in the 1 kV ion reflector mode, collision induced dissociation (CID), and the precursor mass window ± 5 Da. The plate model and default calibration were optimized for the MS-MS spectra processing. For protein identification, the non-redundant NCBI database or SwissProt was searched using MASCOT 2.1 (<http://www.matrixscience.com>) through the Global Protein Server v3.5 from Applied Biosystems. The Mascot Search Parameters were (i) Taxonomy: metazoa (animals); (ii) Database: SwissProt and NCBIInr; (iii) Enzyme: Trypsin, allow upto 1 missed trypsin cleavage site; (iv) Modifications: carbamidomethyl cysteine as fixed modification and oxidized methionine as variable modification; (v) Peptide mass tolerance: 50–100 ppm (PMF) (MS-MS or combined search); (vi) Peptide charge state: +1 MS-MS; (vii) Fragments tolerance: 0.3 Da. The parameters for the combined search (peptide mass fingerprint and MS-MS spectra) were the same as described above. In all protein identifications, the probability scores were greater than the score fixed by MASCOT as significant with a p -value < 0.05 . Protein identification by *de novo* sequencing from fragmentation spectra of peptides was performed using the DeNovo tool software (Applied Biosystems). Tentative sequences were manually checked and validated, and the homology search for these sequences was obtained using the BLAST tool at <http://www.ncbi.nlm.nih.gov/BLAST>.

Results and discussion

LDA model

The search for approaches that complement or improve classical alignment tools for protein function annotation, like BLAST, constitutes a goal of major importance, see for instance the works of Ivanciuc.^{85–89} In this work we accomplish this goal with a 3D-QSAR study using the 3D parameters $\pi_k(R)$, $\pi_k(R)_g$, and $\Delta\pi_k(R)$ as inputs. A simple LDA, with only five variables was the best LDA model found, able to assign each protein to 1 out of 9 different PSP classes corresponding to 9 different genera of HPs (HPGs) or to predict it as a nPSP. The best equation found was the following:

$$\begin{aligned} S_g(r_3) &= 0.007 \cdot \Delta\pi_0(t) - 0.027 \cdot \Delta\pi_0(m) + 0.002 \cdot \pi_0(i) \\ &+ 0.025 \cdot \pi_0(m) - 0.007 \cdot \pi_0(t) - 2.114 \\ &= 0.007 \cdot (\pi_0(t) - \pi_0(t)_g) - 0.027 \cdot (\pi_0(m) - \pi_0(m)_g) \\ &+ 0.002 \cdot \pi_0(m) - 0.007 \cdot \pi_0(t) - 2.114 \quad N \\ &= 11\,512 \quad \chi^2 = 263.9 \quad p < 0.001 \end{aligned} \quad (5)$$

The statistical parameters for the above equation are as follows: number of protein entries in training (N), Chi-square statistic (χ^2), and error level (p -level) which has to be < 0.05 .⁹⁰ This discriminant function presented good sensitivity results both in training (94.8%) and validation sub-sets (94.4%), see Table 1. However, the specificity of the model is unacceptable (lower than 60%) in all sub-sets. It determines that the final accuracy $< 60\%$ of the model is certainly very low with respect to the standards reported in the literature for applications of LDA to QSAR models both for low-weight drugs or proteins that are useful in molecular parasitology and related areas: *e.g.*, the works carried out by García-Domenech, Marrero-Ponce, Bruno-Blanch, Galvez, Gozalbes and others who predicted compounds active against *Trypanosoma cruzi*, *Mycobacterium avium*, *Toxoplasma gondii*, *Plasmodium falciparum*, *Trichomonas vaginalis*, *Fasciola hepatica*, and other parasites;^{91–95} see also the works performed by Marrero-Ponce on protein and DNA/RNA QSAR studies.^{96–98} It was impossible to increase the goodness-of-fit for this model by adding more variables. On the other hand, all the variables in the models are $\pi_0(R)$ variables and this fact strongly reduces the ability of the model to codify structural information in order to distinguish between different proteins. In fact, $\pi_0(R)$ values are zero-order spectral moments ($k = 0$) that in this case are equal to the number of amino acids in a given region and do not codify any information about the nature of the amino acids or their geometrical distribution.^{28,55} Actually, a zero-order spectral moment is equal to the number of nodes in the molecular graph (atoms, amino acids, nucleotides, RNAs, proteins, or organisms) depending on the structural level described by the network.^{54,57} Taking into consideration all these facts, we decided to reject the LDA model and search for more accurate QSAR models using other methods.

ANN models

We tried to find a better model by processing our data with different ANNs.^{99,100} We carried out four different types of networks, namely, Probabilistic Neural Network (PNN), Radial Basis Function (RBF), Multi Layer Perceptron (MLP), and Linear Neural Network (LNN). We report a summary of the ANN that we have trained and tested in Table 1. In Fig. 5A we present the topology of some of the networks trained. All these QSAR models were more complex than the linear LDA model, but the accuracy, sensitivity, and/or specificity obtained in some cases were higher than 90%. We selected the model with higher accuracy but keeping a balance between specificity and sensitivity. Taking into consideration that simpler LDA failed and the number of cases in training is very high, we decided to look for better models without checking the model simplicity. The high number of cases in training ($N = 11\,512$) excludes the possibility of chance correlation due to an excessive number of parameters. The best model found was an LNN with high sensitivity = 91.2, specificity = 91.1, and accuracy = 91.2 in the training sub-set. This QSAR model also presented high accuracy, specificity, and sensitivity values equal to 91.0% in an external validation sub-set (see Table 1). We also validated the model by means of a ROC curve¹⁰¹ analysis (see Fig. 5B). The values of the area under the ROC curve for this model are

Table 1 Comparative study of linear vs. non-linear models for input mode 1 (default)

Model	Parameter	Set	Train			Test (CV)		
			%	nPSP	PSP	%	nPSP	PSP
LDA-FS 5:5-1:1	Specificity	nPSP	52.3	5345	4867	52.0	1765	1631
	Sensitivity	PSP	94.8	68	1232	94.9	22	411
	Accuracy	Total	57.1			56.8		
MLP 21:21-30-1:1	Specificity	nPSP	49.2	5021	5191	49.7	1687	1709
	Sensitivity	PSP	49.4	658	642	49.2	220	213
	Accuracy	Total	49.2			49.6		
PNN 90:90-11512-2-2:1	Specificity	nPSP	100.0	10 212	0	100.0	3396	0
	Sensitivity	PSP	43.5	734	566	43.6	244	189
	Accuracy	Total	93.6			93.6		
PNN 89:89-11512-2-2:1	Specificity	nPSP	100.0	10 212	0	100.0	3396	0
	Sensitivity	PSP	43.5	734	566	43.6	244	189
	Accuracy	Total	93.6			93.6		
MLP 11:11-29-1:1	Specificity	nPSP	87.3	8917	1295	87.2	2962	434
	Sensitivity	PSP	86.5	176	1124	86.6	58	375
	Accuracy	Total	87.2			87.2		
MLP 9:9-26-28-1:1	Specificity	nPSP	77.7	7938	2274	77.7	2638	758
	Sensitivity	PSP	77.3	295	1005	77.4	98	335
	Accuracy	Total	77.7			77.6		
MLP 5:5-19-22-1:1	Specificity	nPSP	67.8	6922	3290	67.7	2300	1096
	Sensitivity	PSP	59.8	523	777	59.6	175	258
	Accuracy	Total	66.9			66.8		
RBF 17:17-1-1:1	Specificity	nPSP	100.0	10 212	0	100.0	3396	0
	Sensitivity	PSP	80.7	251	1049	80.6	84	349
	Accuracy	Total	97.8			97.8		
LNN 90:90-1:1	Specificity	nPSP	91.1	9303	909	91.0	3089	307
	Sensitivity	PSP	91.2	115	1185	91.0	39	394
	Accuracy	Total	91.1			91.0		

0.96 (in training and validation) both very close to 1 (the highest possible value for a perfect classifier) and notably different from 0.5 (the typical value of a random classifier).

MISS-Prot server (input mode 1). On the other hand, as the final objective is the implementation of the model in a web server that performs all the operations in a hidden mode keeping a user-friendly interface for the scientist, the complexity of the model behind this interface does not presuppose any problem for the user. Certainly, we have to consider that with the advent of internet it is important not only to develop new predictive models for proteome research but also to carry out the implementation of these models in public web servers available to other research groups.^{102–109} In this sense, we implemented the LNN model in a server at the portal Bio-AIMS (<http://miaja.tic.udc.es/Bio-AIMS/MISS-Prot.php>). This free online public tool is based on PHP/HTML/Python programming languages and MARCH-INSIDE routines for the calculation of the protein parameters (see Materials and methods section). This is the first web server for PSPs/nPSPs discrimination (in 3D structural terms) predicting PSP scores in 9 different HP genera with respect to other parasites, virus, bacteria, fungi, or human and animal hosts. In Fig. 6 we illustrate the input (A) and output (B) web pages of this server. The user only has to paste the PDB codes for all the query proteins and the server calculates $S_3(r_3)$ scores for these proteins for the main nine parasite organisms studied in this work. We call this use of the server the default or input mode 1.

Experimental study of peptides found in cathepsin proteins. L-cathepsins are secreted proteases that have functions in *F. hepatica* parasite virulence including tissue invasion and

suppression of host immune responses. In fact, Robinson and Tort *et al.*¹¹⁰ used proteomics methods alongside phylogenetic studies and characterized the profile of cathepsin L proteases secreted by adult *F. hepatica* and hence identified those involved in host–pathogen interactions. Phylogenetic analyses showed that the *Fasciola* cathepsin L gene family expanded by a series of gene duplications followed by divergence that gave rise to three clades associated with mature adult worms (Clades 1, 2, and 5) and two clades specific to infective juvenile stages (Clades 3 and 4). The proteomics studies of these authors identified representatives from Clades 1, 2, and 5 but not from Clades 3 and 4 in adult *F. hepatica* secretory products. Clades 1 and 2 account for 67.39 and 27.63% of total secreted cathepsin Ls, respectively, suggesting that their expansion was positively driven and that these proteases are most critical for parasite survival and adaptation. Secreted *F. hepatica* cathepsin Ls are also of interest as targets for serodiagnostic of fascioliasis in humans¹¹¹ and ruminants^{112–114} and as vaccine candidates.¹¹⁵

In this work, we used SEC, IDE, and MS/MS analysis to perform a study of some cathepsin L peptides from *F. hepatica*. Considered a secondary zoonotic disease until the mid-1990s, currently, human fasciolosis is at present emerging or re-emerging in many countries, including increases of prevalence and intensity and geographical expansion. Research in recent years has justified the inclusion of fasciolosis in the list of important human parasitic diseases.¹¹⁶ Proteomic techniques may become very useful tools to identify new proteins by homology sequences. For instance, some authors use a combination of electrophoresis techniques and MS to isolate and characterise new proteins.¹¹⁷ Carrying out MS/MS identification of protein peptides is a very useful

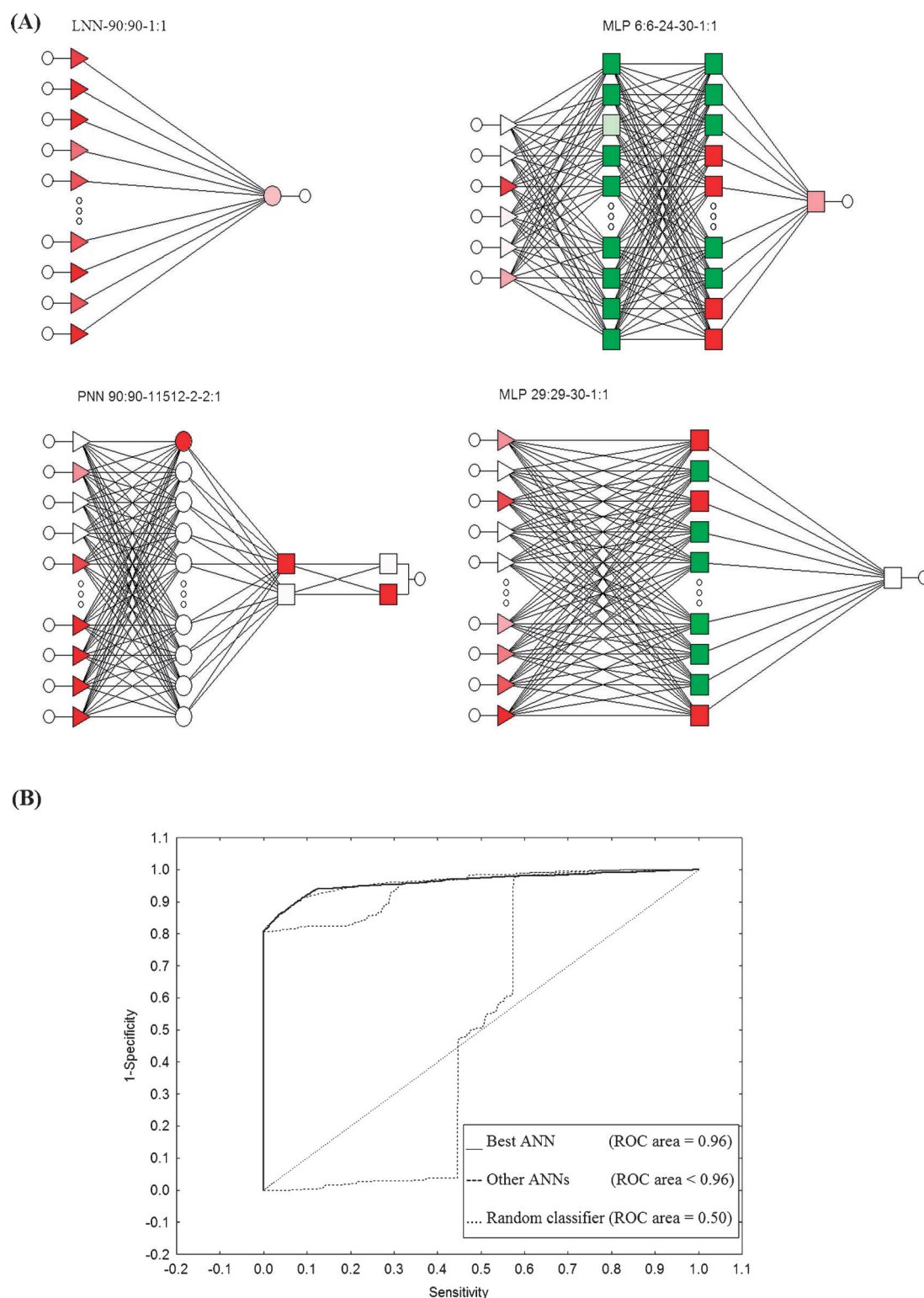


Fig. 5 ANN study: (A) topology of some models trained and (B) the ROC curve for the LNN predictor.

procedure in this sense.¹¹⁸ In MS/MS analysis we can use computational search engines, such as MASCOT, to match MS outcomes for the new peptides in the most similar proteins.^{119,120} It means that MASCOT seeks collection of MS signals and the corresponding sequence of peptides present in known proteins that match up with our MS input.

MASCOT uses the Mowse score (M_s) to rank and select the best protein/peptide candidates.¹²¹ If a template protein in the database has a high $M_s > 52$, in this way we can detect a high sequence homology and perform function annotation. However, the final function annotation of a large protein does not imply that all their peptides found by MS/MS analysis

(A)



RGB Groups
RNASeq, TIC
University of A Coruña
Microbiology & Parasitology
University of Santiago de Compostela
Spain

MISSProt-HP @ Bio-AIMS

Modelling the reality



Home | Links | About



Protein Targets in Multiple
Human Pathogen Parasites

genome: Leishmania, Plasmodium, Trypanosoma

MISSProt-HP

MARCH-INSIDE Spectral moment
prediction of Self Proteins in Human
Parasites
(other than original source organism)

Tool: MARCH-INSIDE (Py)
Data: RCSB PDB

Mode 1: PDBs for 9 Parasites

PDB list: Please paste the names of the PDBchains as a list
(maximum 50)

13PKA
1JLRA
1DQFB
1QNGA
1EEXD
1Y9AC
1A5CB

Predict

(all types of parasites)

Mode 1U: PDBs for User-Defined Organisms

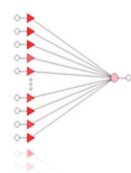
PDB list: Please paste the names of the PDBchains as a list
(maximum 50)

13PKA
1JLRA
1DQFB

Source organism: Homo sapiens

Predict

** the model wasn't trained for this type of organisms



LNN 90:90-1:1

Programmed by Prado-Prado F.
& González-Díaz H.

Mode 2: ENT for 1 of 9 Parasites

Upload & evaluate one ENT file created with HyperChem MM/MD
algorithm (max file size 2MB)

Source organism: Trypanosoma

ENT file

Examine...

Evaluate

* the model was trained for this type of parasites

Mode 2U: ENT for User-Defined Organisms

Upload & evaluate one ENT file created with HyperChem MM/MD
algorithm (max file size 2MB)

Source organism: Homo sapiens

ENT file

Examine...

Evaluate

* the model wasn't trained for this type of organisms

(B)



RGB Groups
RNASeq, TIC
University of A Coruña
Microbiology & Parasitology
University of Santiago de Compostela
Spain

MISSProt-HP @ Bio-AIMS

Modelling the reality



Home | Links | About

Process ID = 105624afdd28ad5230

PDB List = 1ML3D 2GHID

... please wait

PDB Update/Verification ...

1ML3D 2GHID Done!

Processing PDBs ...

1ML3D 2GHID

Result file = [Results/105624afdd28ad5230/ProtSpect-HPP.calc.txt](http://miaja.tic.udc.es/Bio-AIMS/Results/105624afdd28ad5230/ProtSpect-HPP.calc.txt)

ProtSpect-HPP @ Bio-AIMS

Prediction of 3D-Selfness of Protein targets in Human Pathogen Parasites by using Spectral Momen
by using MARCH-INSIDE and LNN 90:90-1:1 (91.0% accuracy)

Results = <http://miaja.tic.udc.es/Bio-AIMS/Results/105624afdd28ad5230/ProtSpect-HPP.calc.txt>
2009-11-13 22:41:31

Parasite type (P)

.....
P1-Trypanosoma P2-Toxoplasma P3-Tolypocladium
P4-Plasmodium P5-Leishmania P6-Giardia
P7-Fasciola P8-Entamoeba P9-Ascaris

PDB	P1	P2	P3	P4	P5	P6	P7	P8	P9
1ML3D	YES	YES	NO	YES	YES	YES	NO	NO	YES
2GHID	YES	YES	YES	YES	YES	YES	NO	YES	YES

Fig. 6 MISS-Prot web server interface (<http://miaja.tic.udc.es/Bio-AIMS/MISSProt-HP.php>).

have a positive contribution to the protein activity by themselves and also does not exclude that these peptides, if present, may have other functions in proteins from other organisms.

In addition to protein sequence homology, we can take into account that proteins and their peptides are 3D structures that could be compared with a target 3D genome to investigate

the selfness of query peptides. In the following example we used 1DE followed by MS/MS analysis, and subsequent evaluation by MASCOT, BLAST, and 3D-QSAR analysis to predict both 1D and 3D ($S_{10}(r_1)$ and $S_{10}(r_3)$) values for self/non-self scores of *F. hepatica* peptides with respect to the human proteome. In Fig. 4 we illustrate the results obtained after SEC separation (A) and an overall view of the 1DE bands (B) obtained from *F. hepatica* samples. After excision, the band was subjected to in-gel trypsin digestion and three of the most abundant resulting peptides subjected to MS/MS analysis. Then, the MS signals were introduced into the MASCOT search engine.^{122,123} We selected MASCOT collection of annotated databases because it contains data on the *F. hepatica* proteome of annotated proteins with MS recorded. A similar combination was successfully used in the past to study *Trichinella* sp. antigens¹²⁴ and *Leishmania infantum* dynein proteins.⁵⁹ With the MASCOT search we could identify three peptide sequences AVPDKIDWR ($M_s = 48$), NSWGLSWGER ($M_s = 74$) and QFGLETSSYPYTA-VEGQCR ($M_s = 127$), all of which matched *Fasciola* L-cathepsins. Considering only *F. hepatica* proteins we obtained 7 hits (see Table 2) with maximal M_s scores = 249. Among them, 4 hits corresponded to *F. hepatica* L-cathepsins, without specification of clade, while 2 hits were classified as a secreted L1-cathepsin (gi|116488416) and procathepsin L1 (gi|163310848), respectively. This later,¹²⁵ having mass $M = 35\,390$ Da, was used as a template for matching the set of identified *F. hepatica* peptides (Fig. 7). It should be noted that the masses of *F. hepatica* L cathepsins showed in Table 2 ($M = 35\text{--}37$ kDa) are higher than the mass observed in 1DE (Fig. 4, band 1) because the secreted L-cathepsins lacked the signal peptide and the inhibitor propeptide.^{126,127}

Theoretical studies of peptides found in cathepsin proteins. To carry out the theoretical study of the peptide found we have to consider that: (i) we introduced a new server for predicting unique targets in different parasites including *Fasciola* sp., (ii) unique peptides (in terms of 3D structure) may be present in specific drug-target pocket sites, and (iii) unique peptides may also become important in vaccine development. Then, we decided to predict both 1D and 3D self/non-self scores, $S_{10}(r_1)$ and $S_{10}(r_3)$, for the selected peptides derived from the *F. hepatica* L1-cathepsin.

BLASTp study of peptides found in *F. hepatica* L1-cathepsin. We use the Basic Local Alignment Search Tool (BLAST), which is the most frequently used tool for calculating sequence

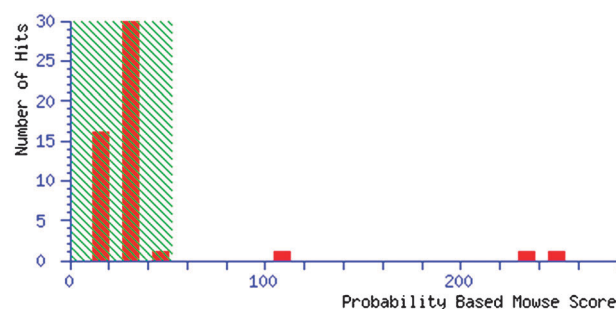


Fig. 7 MASCOT search sequence coverage for cathepsin L1 (gi|163310848) from *F. hepatica*. The sequence corresponding to the signal peptide is underlined and the sequence of the propeptide is highlighted in blue. Matching positions of identified peptides are highlighted in red.

similarity, for calculation of 1D self/non-self scores $S_{10}(r_1)$. The way most people use BLAST is to input a nucleotide or protein sequence as a query against all (or a subset of) the public sequence databases, pasting the sequence into the textbox on one of the BLAST web pages. There are many different variations of BLAST available to use for different sequence comparisons, e.g., a DNA query to a DNA database, a protein query to a protein database, and a DNA query, translated in all six reading frames, to a protein sequence database. In this work, we use BLASTp, a variation for the comparison of a protein query to a protein database. From BLASTp results, we obtained similarity identities = 9/14 (65%) for the *F. hepatica* peptide with sequence qfgletessypytavegqcr, and peptide gldseesypyvavd as the one most similar to it in the human proteome. This peptide is present in the sequence of cliticocypin, a Cathepsin V complex chain A, with a PDB ID 3H6S. Another BLAST result was a similarity identity = 7/8 (88%) for the *F. hepatica* peptide with sequence avpdkidwr and peptide vpdksdwr as the most similar in the human proteome. This peptide is present in the sequence of the CWF19L1 protein. The last BLAST result was a similarity identity = 8/10 (80%) for *F. hepatica* nswglswger and peptide nswgpawger as the most similar in the human proteome; this peptide was predicted to be present in the sequence of the tubulointerstitial nephritis antigen-like 1 protein.

LOMETS study of peptides found in cathepsin proteins. Unfortunately, MS/MS provide information only about the sequences of peptides but our QSAR model uses the 3D structure to carry out the predictive study. Consequently, we decided to run structure prediction jobs at the LOMETS web server in order to obtain different models of the 3D structure of these peptides. Following the same line of the previous section, our main interest in the study of the peptides of the new unknown proteins is to know which of them have a positive contribution to the differential expression of proteins in the different HPg. It may allow us to select peptides for drug

Table 2 Best protein hits found with MASCOT for *F. hepatica* proteins

Protein	Accession ^a	Mass	Score	Description
1	gi 163310848	35 390	249	Procathepsin L1, chain A
2	gi 116488416	37 149	249	Secreted cathepsin L1
3	gi 41152540	26 600	249	Cathepsin L protein
4	gi 13774082	35 611	249	Cathepsin L-like
5	gi 41152538	37 090	249	Cathepsin L protein
6	gi 50403821	36 978	249	Cathepsin L1 proteinase
7	gi 31558997	36 978	249	Cathepsin L

^a Database: NCBIInr 080227 (6189142 sequences); 10 Apr 2008 at 16:45:09 GMT.

Table 3 Self/non-self study of peptides of one protein of *F. hepatica* (mode 2U)

3D structure models ^a								Selfness scores ^b	
Rank	Template	AL	C	Z-score	SID	C-score	Soft.	$S_{10}(r_1)$	$S_{10}(r_3)$
QFGLTESSYPYTAVEGQCR									
1	1IWDA	20	1	18.621	0.45	High	SAM	30.8	14.6
2	2O6XA2	20	1	56.987	1	High	SAM		14.6
3	2O6XA	20	1	11.213	1	High	MUSTER		14.6
4	2O6XA	19	0.95	10.461	1	High	HHsearch		14.6
5	1XKGA	20	1	10.477	0.45	High	SP3		14.6
AVPDKIDWR									
1	2O6XA	9	1	5.864	1	High	MUSTER	25.2	14.6
2	2O6XA2	9	1	31.705	1	High	SAM		14.6
3	2GS0B	9	1	16.283	0.444	High	SPARKS2		14.6
4	3IECE	9	1	15.078	0.333	High	SP3		14.6
5	2C0YA2	9	1	5.989	0.556	High	MUSTER		14.6
NSWGLSWGER									
1	1XKGA2	10	1	7.768	0.5	High	MUSTER	29.1	14.6
2	3IAXB	10	1	20.85	0.4	High	SPARKS2		14.6
3	2O6XA2	10	1	39.485	0	High	SAM		14.6
4	3E1ZB	10	1	7.836	0.7	High	MUSTER		14.6
5	3IAXB	8	0.8	16.348	0.25	High	SP3		14.6

^a Models generated using the web server LOMETS (<http://zhanglab.cmb.med.umich.edu/LOMETS/>), template = PDB ID of the template protein, AL = peptide-template alignment length, C = Coverage, Z-score = Z-score of the *j*th model of the *i*th program, SID = sequence ID is the sequence identity to the query of the *j*th model of the *i*th program, C-score = coverage score. ^b Selfness scores $S_g(r_n)$ in humans ($g = 10$) for *F. hepatica* peptides: $S_{10}(r_3)$ = score calculated with MISS-Prot for 3D structures ($n = 3$), $S_{10}(r_1)$ = scores calculated with BLAST using only sequence information ($n = 1$).

Table 4 Average values for different organisms (continued in Table 5)

μ_k^a avg.	Organism code or label (g values) ^b												
	Pathogen parasites									Parasite hosts			
	1	2	3	4	5	6	7	8	9	10	11	12	13
0;0	19.2	6.2	3.1	8.2	10.8	0.5	6.8	10.5	6.3	9.8	9.4	22.1	6.6
1;0	0.8	0.3	0.1	0.3	0.5	0.0	0.3	0.6	0.3	0.4	0.4	1.0	0.3
2;0	1.1	0.4	0.2	0.4	0.7	0.0	0.4	0.8	0.4	0.7	0.7	1.4	0.5
3;0	1.4	0.5	0.2	0.5	0.9	0.0	0.6	1.1	0.5	0.8	0.8	1.7	0.7
4;0	1.6	0.6	0.2	0.5	1.0	0.0	0.7	1.2	0.6	1.0	1.0	2.0	0.8
5;0	1.8	0.6	0.3	0.6	1.2	0.0	0.8	1.4	0.6	1.1	1.1	2.2	0.9
0;1	69.1	53.7	23.1	53.3	71.2	5.8	43.2	54.6	40.6	55.5	47.6	134.3	40.6
1;1	3.5	3.0	1.2	2.0	3.5	0.2	2.3	3.1	1.9	2.8	2.4	6.5	2.1
2;1	5.3	4.1	1.6	2.9	5.1	0.3	3.3	4.3	2.8	4.1	3.7	9.5	3.2
3;1	6.8	5.4	2.2	3.7	6.6	0.4	4.3	5.8	3.7	5.3	4.7	12.0	4.2
4;1	7.9	6.2	2.5	4.2	7.7	0.5	4.9	6.6	4.3	6.2	5.6	14.0	5.0
5;1	8.9	7.1	2.9	4.7	8.7	0.6	5.7	7.6	4.9	7.1	6.3	15.6	5.7
0;2	43.8	109.3	43.3	90.3	104.7	22.0	66.3	85.4	68.4	81.3	87.1	240.9	56.1
1;2	2.6	6.2	2.5	4.6	6.2	1.0	3.8	5.6	3.8	4.5	4.6	13.0	3.4
2;2	3.9	8.7	3.6	6.6	9.0	1.5	5.5	7.8	5.6	6.8	7.0	19.4	5.2
3;2	5.0	11.7	4.9	8.6	11.9	2.0	7.3	10.7	7.4	8.7	9.0	24.8	6.8
4;2	6.0	13.3	5.7	9.9	13.7	2.4	8.4	12.1	8.5	10.2	10.7	29.0	8.1
5;2	6.8	15.4	6.7	11.2	15.7	2.7	9.6	14.2	9.8	11.6	12.1	32.5	9.2
0;3	81.2	177.9	84.0	266.4	113.6	57.5	139.4	137.5	136.6	136.8	156.5	365.2	76.2
1;3	5.8	8.9	7.3	15.8	8.2	4.3	10.6	9.4	8.6	0.0	10.7	0.0	5.5
2;3	8.4	12.7	9.7	23.1	11.8	5.8	15.0	13.4	12.3	0.0	15.8	0.0	8.3
3;3	10.8	16.5	13.8	29.8	15.5	8.1	20.3	17.8	16.4	0.0	20.5	0.0	10.9
4;3	12.6	19.0	15.3	34.4	17.9	9.1	23.5	20.4	18.9	0.0	24.0	0.0	12.9
5;3	14.1	21.6	18.5	38.9	20.5	11.0	27.4	23.4	21.7	0.0	27.2	0.0	14.7
0;4	213.3	346.9	153.6	418.2	300.4	85.8	255.8	288.0	251.9	283.4	300.7	762.5	179.6
1;4	12.6	18.3	11.1	22.7	18.4	5.5	17.0	19.0	14.8	0.0	18.2	0.0	11.3
2;4	18.7	25.9	15.1	33.1	26.6	7.7	24.3	26.7	21.4	0.0	27.2	0.0	17.3
3;4	24.0	34.1	21.0	42.5	34.9	10.6	32.5	35.7	28.3	0.0	35.1	0.0	22.5
4;4	28.1	39.0	23.7	49.1	40.3	12.1	37.5	40.8	32.6	0.0	41.2	0.0	26.7
5;4	31.7	44.8	28.3	55.4	46.1	14.3	43.5	47.3	37.6	0.0	46.7	0.0	30.5

^a $\mu_k(O)_{avg}$ average values for different organisms; the first term is *k* and the second O. ^b Organisms for web-server modes 1 and 2: *Ascaris spp.* ($g = 1$), *Entamoeba spp.* ($g = 2$), *Fasciola spp.* ($g = 3$), *Giardia spp.* ($g = 4$), *Leishmania spp.* ($g = 5$), *Plasmodium spp.* ($g = 6$), *Toxoplasma spp.* ($g = 7$), *Trypanosoma spp.* ($g = 8$), *Tolypocladium spp.* ($g = 9$). Organisms for web-server modes 3 and 4: *Homo sapiens* ($g = 10$), *Bos taurus* ($g = 11$), *Sus scrofa* ($g = 12$), *G. gallus* ($g = 13$).

design and/or obtain information for drug target discovery. In this regard, we have to calculate the $\pi_k(R)$ values for all peptides and substitute these values in the QSAR model to predict $S_{10}(r_3)$ scores. To this end, we need first the 3D structures of the peptides in order to calculate the $\pi_k(R)$ values. For this study we used the 3 *F. hepatica* peptides identified by MS/MS. We firstly obtained the optimal 3D folded structures using the web server LOMETS. In Table 3 we have also summarised the results of the LOMETS simulation of the 3D structure of these peptides.

MISS-Prot scoring of peptide outcomes (input mode 2). Input mode 2 refers to the prediction with the new server of $S_g(r_3)$ scores for the selected peptides. We used MISS-Prot to predict selfness scores for humans $S_{10}(r_3)$ of the peptides studied here. To this end, after LOMETS calculation of different models for the plausible 3D structures we upload the best structural models to the server (PDB files). MISS-Prot uses essentially the same algorithm as MARCH-INSIDE 2.0 to calculate the $\pi_k(R)$ values for the 15 structural models for the 3 peptides studied. In Table 3 we illustrate the 3D self/non-self scores obtained for these peptides. Notably, all peptides of the *F. hepatica* protein present positive self/non-self scores in *H. sapiens* ($g = 10$), which means a high probability to be

expressed by human cells. Consequently, both sequence-based scores $S_{10}(r_1)$ and 3D structure-based scores $S_{10}(r_1)$ indicate high probability of expression of these peptides both in the parasite and in the human host. This is important because the alignment procedure does not consider the 3D structure and may fail when the database does not contain templates similar to the concerned query or if those available are also proteins with an unknown function, as in this case.^{66,128}

Notes on input mode 1U and mode 2U. The present server was trained and validated with the explicit objective of predicting PSPs from the 3D structure released to PDB only. The server predicts PSPs for eight of the major classes of HPs: $g = 1$ -Ascaris, 2-Entamoeba, 3-Fasciola, 4-Giardia, 5-Leishmania, 6-Plasmodium, 7-Toxoplasma, and 8-Trypanosoma. For these organisms the user only has to introduce the PDB code for the protein and the server has as input an internal list with the average values of the parameters (see Table 3). We called this case the input mode 1. Although the server was created to calculate $S_g(r_3)$ values only for the above parasites, we should not forget that the model depends both on protein parameters and average parameters of the proteins expressed on a given organism. Consequently, it is

Table 5 Contains average values for other organisms (continued from Table 4)

$\mu_k^a(O)$ avg.	Organism code or label (g values) ^b										
	Vectors		Other pathogens								Biotechnology
	14	15	16	17	18	19	20	21	22	23	24
0:0	6.7	7.8	5.1	8.4	15.8	8.9	22.7	7.3	10.4	8.2	13.8
1:0	0.3	0.3	0.2	0.3	0.0	0.4	1.0	0.3	0.4	0.4	0.0
2:0	0.4	0.4	0.4	0.5	0.0	0.6	1.4	0.4	0.7	0.6	0.0
3:0	0.5	0.6	0.5	0.6	0.0	0.8	1.8	0.5	0.8	0.7	0.0
4:0	0.6	0.6	0.5	0.8	0.0	0.9	2.1	0.6	1.0	0.8	0.0
5:0	0.7	0.7	0.6	0.9	0.0	1.0	2.4	0.7	1.1	0.9	0.0
0:1	46.5	56.5	33.0	62.3	100.5	55.1	110.7	46.7	57.6	51.8	79.2
1:1	2.2	2.5	1.6	2.7	4.8	3.3	5.2	2.1	2.7	2.5	3.7
2:1	3.2	3.8	2.4	4.1	7.2	4.7	7.9	3.1	4.0	3.6	5.5
3:1	4.1	4.8	3.1	5.3	9.2	6.4	10.1	4.0	5.1	4.6	7.1
4:1	4.8	5.6	3.7	6.2	10.7	7.3	11.9	4.7	6.0	5.4	8.3
5:1	5.4	6.3	4.2	7.0	12.1	8.7	13.4	5.3	6.7	6.1	9.3
0:2	84.7	105.9	72.2	133.7	176.3	126.5	141.6	91.3	84.4	96.5	132.5
1:2	4.3	5.4	4.0	6.7	9.4	8.0	8.4	4.6	4.4	5.4	7.0
2:2	6.5	8.1	6.1	10.2	14.1	11.5	12.6	7.0	6.6	8.0	10.6
3:2	8.3	10.4	7.9	13.0	18.1	15.4	16.3	9.0	8.5	10.2	13.6
4:2	9.7	12.2	9.4	15.3	21.2	17.7	19.1	10.6	9.9	11.9	16.0
5:2	10.9	13.7	10.6	17.3	23.8	20.6	21.6	11.9	11.2	13.4	18.1
0:3	152.3	199.7	122.4	209.2	288.1	215.8	333.4	104.4	221.3	122.3	206.7
1:3	9.5	12.5	7.7	13.0	18.7	16.0	22.3	6.6	0.0	8.7	0.0
2:3	14.0	18.7	11.7	19.4	28.0	23.1	33.5	9.8	0.0	12.8	0.0
3:3	17.9	24.0	15.1	24.8	36.0	30.6	43.4	12.6	0.0	16.3	0.0
4:3	20.8	28.1	17.8	28.9	42.3	35.5	51.2	14.7	0.0	18.9	0.0
5:3	23.4	31.7	20.2	32.4	47.8	40.9	58.1	16.5	0.0	21.2	0.0
0:4	290.2	369.9	232.7	413.6	580.7	406.2	608.4	249.7	373.6	278.7	432.2
1:4	16.5	20.7	13.7	22.8	0.0	27.6	36.8	13.6	0.0	16.8	0.0
2:4	24.5	31.0	20.6	34.1	0.0	39.8	55.5	20.4	0.0	24.8	0.0
3:4	31.3	39.7	26.6	43.7	0.0	53.0	71.5	26.1	0.0	31.6	0.0
4:4	36.5	46.6	31.4	51.1	0.0	61.3	84.3	30.6	0.0	36.9	0.0
5:4	41.1	52.5	35.6	57.5	0.0	71.0	95.4	34.4	0.0	41.4	0.0

^a $\mu_k(O)_{avg}$ average values for different organisms; the first term is k and the second O . ^b Organisms for web-server modes 3 and 4: *M. musculus* ($g = 14$), *R. norvegicus* ($g = 15$), *B. stearothermophilus* ($g = 16$), *B. subtilis* ($g = 17$), *E. coli* ($g = 18$), *Pseudomonas sp.* ($g = 19$), *S. typhimurium* ($g = 20$), *Staphylococcus sp.* ($g = 21$), *Streptococcus sp.* ($g = 22$), *Streptomyces sp.* ($g = 23$) and *Saccharomyces sp.* ($g = 24$). Vectors = infectious disease transmission vectors; Biotech. = organisms interesting for pharmaceutical or food industrial biotechnology.

technically possible to estimate $S_g(r_3)$ values for proteins of other organisms if we introduce as inputs for the classifier new average values for proteins of a different organism. The reader should remember at this point that the classifier was trained with proteins of many different organisms as control cases. For these situations, we incorporated to the server the input modes: mode 1U and mode 2U. These input modes are essentially the same as mode 1 and mode 2 but the server prompts you to introduce the name of the new organisms, parasites or not (e.g., host organisms and/or transmission vectors), as well as the average values of these organisms (see Table 4 and 5). Using these new average values as inputs and the values for the query proteins the server will calculate the $S_g(r_3)$ values for the query proteins on the new organism. In mode 2U the server allows you to type the PDB codes of the query proteins.

Lastly, in mode 4 the server allows you to browse your own files and upload 3D structure models (.ent files) with the 3D structure of the protein generated by MM/MD studies. We have to be aware that by using the input mode 3 and mode 4 we predict self scores $S_g(r_3)$ for proteins of a new parasite, host, transmission vector or other organism that was not present on the original dataset and used to train and validate the model (at least as a positive case). Consequently, predictions derived with input mode 1U and mode 2U have to be used with higher caution than predictions with input mode 1 and mode 2. In Tables 4 and 5 we list the values of the average parameters for the default HPs of the server as well as

for other organisms such as important fungi pathogens like *Tolypocladium* spp. (g = 9), parasite hosts like *Homo sapiens* (g = 10), *Bos Taurus* (g = 11), *Sus scrofa* (g = 12), *Gallus gallus* (g = 13), and/or pathogen transmission vectors like *Mus musculus* (g = 14), *Rattus norvegicus* (g = 15). We also included other human pathogen organisms like *Bacillus stearothermophilus* (g = 16), *Bacillus subtilis* (g = 17), *Escherichia coli* (g = 18), *Pseudomonas* sp. (g = 19), *Salmonella typhimurium* (g = 20), *Staphylococcus* sp. (g = 21), *Streptococcus* sp. (g = 22), or biotechnologically relevant organisms like *Streptomyces* sp. (g = 23) and *Saccharomyces* sp. (g = 24). Remember that g is only a code or label for identifying the organism.

MISS-Prot study of Ani s allergens (mode 2U). We illustrate the use of input mode 2 for MIProtS-HP server with 10 known allergen proteins (called Ani s 1 to Ani s 10)^{129–149} of the parasite *A. simplex*. This HP organism was not present on the original list of HPs g = 1 to 8. Details of these proteins have been deposited at the database ALLERGEN of the I.U.I.S. Allergen Nomenclature Sub-committee (<http://www.allergen.org/Allergen.aspx>), which operates under the auspices of the International Union of Immunological Societies (I.U.I.S.) and the World Health Organization (W.H.O.). We have withdrawn from ALLERGEN the GeneBank accession codes for these sequences (<http://www.ncbi.nlm.nih.gov/pubmed>). However, at present there are no 3D structures released to PDB for these proteins. We used the software HyperChem to

Table 6 3D-Selfness study of peptides of *A. simplex* antigens using HyperChem to predict 3D structures + MISS-Prot to predict selfness scores (input mode 2U)^a

Peptide	Sequence	s	Accession	sbp	Biochem.	Ref.	E_{pot}	ΔE_{pot}	ACCR	$S_{10}(r_3)^b$
A1	MASMQHFSLAALLLA	1	AB100095	1	—	131	−22.3	27.5	0.62	0.031
A2	MHLITALVLLLQLIH			16			28.5	29.6	0.62	0.037
A3	MSDTLYRSPSMAIRS			32			−440.9	26.6	0.62	0.033
A4	MFQSTFFLVLMVCVA	2	AF173004	1	Paramyosin	135	−32.4	27.9	0.63	0.021
A5	AKYGSQFCKNLLANC			16			−58.5	27.2	0.65	0.055
A6	MKVLFFVAAVVAVSL			32			28.3	28.6	0.65	0.020
A7	MKLCILAVAVFVVAV	3	Y19221	1	Tropomyosin	135	42.1	28.2	0.63	0.014
A8	ASICLGADRTCEQL			16			−176.5	25.2	0.65	0.035
A9	FITSSPIQEVVGGP			32			13.6	24.2	0.63	0.043
A10	STADMGALTSMSVAD	4	AM279414	1	Cysteine	136	−11.8	25.8	0.62	0.043
A11	DNALDRADAAEEKVR			16	Protease		−223.1	26.7	0.62	0.041
A12	ILMQTEAGMLGSSD			32	Inhibitor		−41.3	22.7	0.61	0.014
A13	ALADDTPPPPFLAG	5	AB274998	1	SXP	143	45.4	24.9	0.62	0.045
A14	TARFANKDHCPPNEE			16	RAL-2		−35.0	26.3	0.64	0.049
A15	LSSTGATLPMQSPWQ			32	Protein		−76.0	24.7	0.65	0.049
A16	AQDQGPPLPKFLDG	6	AB274999	1	Serine	144	96.2	24.8	0.63	0.037
A17	SAQGPPPLPPFVANA			16	Protease		38.1	28.5	0.63	0.029
A18	PLDKGTPTQEGGVK			32	Inhibitor		40.4	26.5	0.65	0.059
A19	GPVVGGSGIGNVWEK	7	EF158010	1	—	129	14.4	24.6	0.62	0.107
A20	LGSLTRLEDKIRLLQ			16			−218.0	31.6	0.63	0.053
A21	QMTDKLERIEEELRD			32			−320.7	22.8	0.58	0.037
A22	VDVNDPEIKELAGKS	8	AB300625	1	SXP/	145	44.6	26.3	0.63	0.025
A23	APQDVVKAFPELLKK			16	RAL-2		5.9	26.6	0.64	0.098
A24	YNECGNPCQEKCDNG			32	Protein		−109.1	26.0	0.60	0.027
A25	IPPVVSSCITSGMAK	9	EU074790	1	SXP/	142	48.6	23.8	0.67	0.025
A26	ADQATKDFAAALAQ			16	RAL-2		−80.3	26.3	0.62	0.023
A27	PPAVQAEFRQLANGA			32	Protein		−98.2	29.0	0.65	0.049
A28	MDAIKKKMQAMKIEK	10	GU187358	1	Unknown	—	62.8	32.1	0.64	0.039
A29	MQSRIVVAVVVFCCI			16			−101.4	28.0	0.61	0.031
A30	MKTLIVAALFCTIGM			32			12.3	57.5	0.64	0.014

^a ACCR = acceptance ratio, E_{pot} = potential energy, ΔE_{pot} = potential energy difference for the 3D structure model obtained by the MM/MD study using HyperChem. ^b $S_{10}(r_3)$ = selfness scores in humans (g = 10) for 3D structures ($n = 3$) of peptides present in *A. simplex* antigens.

generate 3D structure models of peptides found on the structure of the 10 allergens from *A. simplex*. We introduced the sequences as the input in the sequence editor of HyperChem and MM/MD experiments described in the Materials and methods section were carried out. We used only three 15-aa-long peptides from each allergen starting at positions 1, 16, and 32 of the allergen sequence. Using MISS-HP we uploaded the 3D structure (.ent files) of 30 peptides (A1 to A30) and predicted their selfness scores for humans $S_{10}(r_3)$. In Table 6 we illustrate the results obtained for these peptides: code, sequence, positions, energetic details, and predicted scores. Notably, all peptides present very low selfness scores in *H. sapiens* ($g = 10$), it means low probability to be expressed by human cells. The result is consistent with self/non-self discrimination theories postulating that the human immune systems recognize as allergens peptides with low sequence similarity.^{150–152} However, this prediction confirms low selfness for sequences and also in terms of 3D structure information.

Conclusion

This study allowed us to confirm that spectral moment parameters derived with an MCM are useful predictors for PSP/nPSP discrimination involving 3D structural and not only sequence information. We confirm by LDA a useful linear relationship. However, using ANN classifiers we may obtain notably accurate models. We also conclude that this method may help us to predict not only new PSP proteins but also to select peptide candidates found on MS/MS of new proteins that may improve PSP targetability. In particular, the new model can be used to study peptide outcomes obtained after IDE and MALDI-TOF MS analysis of HP samples. This combined strategy may be used to identify and predict proteins/peptides of prokaryotic and eukaryotic parasites which may be of interest in drug development or target identification or serve in anti-parasitic vaccine design. This type of model may be implemented in public web servers, available to the scientific community for use free-of-charge.

Acknowledgements

The authors appreciate the technical assistance and advice of Lola Gutiérrez, from the Genomics and Proteomics Center, Universidad Complutense de Madrid. H. González-Díaz and CR Munteanu are thankful for support from the research program Isidro Parga Pondal (IPP), funded by Xunta de Galicia (XdG) and European Social Fund (ESF). F. Prado-Prado thanks sponsorships of research position at USC from the research program Angeles Alvarino, funded by XdG and ESF. The present study was supported by grants AGL2010-22290-C03-01 (Ministerio de Ciencia e Innovación, Spain), the “Ibero-American Network of the Nano-Bio-Info-Cogno Convergent Technologies”, Ibero-NBIC Network (209RT0366), funded by CYTED (Ciencia y Tecnología para el Desarrollo), the COMBIOMED Network, grant (Ref. PIO52048 and RD07/0067/0005), funded by the Carlos III Health Institute, and by the European Fund for Regional Development (FEDER).

References

- 1 F. Ibarra-Velarde, Y. Vera-Montenegro, A. Huesca-Guillen, G. Canto-Alarcon, Y. Alcala-Canto and Y. Marrero-Ponce, *Ann. N. Y. Acad. Sci.*, 2008, **1149**, 183–185.
- 2 G. Zehetner, *Nucleic Acids Res.*, 2003, **31**, 3799–3803.
- 3 Y. Zhou, G. M. Huang and L. Wei, *Bioinformatics*, 2002, **18**, 1268–1269.
- 4 A. F. Neuwald and A. Poleksic, *Nucleic Acids Res.*, 2000, **28**, 3570–3580.
- 5 A. Muller, R. M. MacCallum and M. J. Sternberg, *J. Mol. Biol.*, 1999, **293**, 1257–1271.
- 6 J. Zhang and T. L. Madden, *Genome Res.*, 1997, **7**, 649–656.
- 7 P. Durand, L. Canard and J. P. Mornon, *CABIOS, Comput. Appl. Biosci.*, 1997, **13**, 407–413.
- 8 S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman, *Nucleic Acids Res.*, 1997, **25**, 3389–3402.
- 9 L. Han, J. Cui, H. Lin, Z. Ji, Z. Cao, Y. Li and Y. Chen, *Proteomics*, 2006, **6**, 4023–4037.
- 10 H. H. Lin, L. Y. Han, H. L. Zhang, C. J. Zheng, B. Xie and Y. Z. Chen, *J. Lipid Res.*, 2006, **47**, 824–831.
- 11 H. H. Lin, L. Y. Han, C. Z. Cai, Z. L. Ji and Y. Z. Chen, *Proteins: Struct., Funct., Bioinf.*, 2006, **62**, 218–231.
- 12 L. Y. Han, C. Z. Cai, Z. L. Ji, Z. W. Cao, J. Cui and Y. Z. Chen, *Nucleic Acids Res.*, 2004, **32**, 6437–6444.
- 13 L. Y. Han, C. Z. Cai, Z. L. Ji and Y. Z. Chen, *Virology*, 2005, **331**, 136–143.
- 14 Y. D. Cai, G. P. Zhou and K. C. Chou, *J. Theor. Biol.*, 2005, **234**, 145–149.
- 15 K. C. Chou, *Bioinformatics*, 2005, **21**, 10–19.
- 16 K. C. Chou, *Proteins: Struct., Funct., Bioinf.*, 2001, **43**, 246–255 (Erratum: *ibid.*, 2001, **44**, 60).
- 17 S. Q. Wang, J. Yang and K. C. Chou, *J. Theor. Biol.*, 2006, **242**, 941–946.
- 18 X. Xiao, S. Shao, Y. Ding, Z. Huang and K. C. Chou, *Amino Acids*, 2006, **30**, 49–54.
- 19 Y. D. Cai and K. C. Chou, *J. Theor. Biol.*, 2006, **238**, 395–400.
- 20 K. C. Chou and Y. D. Cai, *Proteins: Struct., Funct., Bioinf.*, 2003, **53**, 282–289.
- 21 Y. D. Cai and K. C. Chou, *J. Theor. Biol.*, 2005, **234**, 145–149.
- 22 K. C. Chou and Y. D. Cai, *J. Chem. Inf. Model.*, 2005, **45**, 407–413.
- 23 P. D. Dobson and A. J. Doig, *J. Mol. Biol.*, 2005, **345**, 187–199.
- 24 N. Bhardwaj, R. E. Langlois, G. Zhao and H. Lu, *Nucleic Acids Res.*, 2005, **33**, 6486–6493.
- 25 S. Vilar, H. Gonzalez-Diaz, L. Santana and E. Uriarte, *J. Theor. Biol.*, 2009, **261**, 449–458.
- 26 H. González-Díaz, S. Vilar, L. Santana and E. Uriarte, *Curr. Top. Med. Chem. (Sharjah, United Arab Emirates)*, 2007, **7**, 1025–1039.
- 27 R. Concu, G. Podda, E. Uriarte and H. Gonzalez-Diaz, *J. Comput. Chem.*, 2009, **30**, 1510–1520.
- 28 H. Gonzalez-Diaz, L. Saiz-Urra, R. Molina, Y. Gonzalez-Diaz and A. Sanchez-Gonzalez, *J. Comput. Chem.*, 2007, **28**, 1042–1048.
- 29 H. González-Díaz, Y. Pérez-Castillo, G. Podda and E. Uriarte, *J. Comput. Chem.*, 2007, **28**, 1990–1995.
- 30 L. Santana, E. Uriarte, H. González-Díaz, G. Zagotto, R. Soto-Otero and E. Mendez-Alvarez, *J. Med. Chem.*, 2006, **49**, 1149–1156.
- 31 H. González-Díaz, R. R. de Armas and R. Molina, *Bioinformatics*, 2003, **19**, 2079–2087.
- 32 G. Aguero-Chapin, J. Varona-Santos, G. A. de la Riva, A. Antunes, T. Gonzalez-Villa, E. Uriarte and H. Gonzalez-Diaz, *J. Proteome Res.*, 2009, **8**, 2122–2128.
- 33 H. González-Díaz, L. Saiz-Urra, R. Molina, L. Santana and E. Uriarte, *J. Proteome Res.*, 2007, **6**, 904–908.
- 34 S. H. Strogatz, *Nature*, 2001, **410**, 268.
- 35 L. Santana, H. Gonzalez-Diaz, E. Quezada, E. Uriarte, M. Yanez, D. Vina and F. Orallo, *J. Med. Chem.*, 2008, **51**, 6740–6751.
- 36 D. Vina, E. Uriarte, F. Orallo and H. Gonzalez-Diaz, *Mol. Pharmaceutics*, 2009, **6**, 825–835.
- 37 S. Bornholdt and H. G. Schuster, *Handbook of Graphs and Complex Networks: From the Genome to the Internet*, Wiley-VCH GmbH & CO. KGa., Weinheim, 2003.
- 38 J. H. Peters, G. R. Gordon, J. F. Murray, Jr., A. H. Fieldsteel and L. Levy, *Antimicrob. Agents Chemother.*, 1975, **8**, 551–557.

- 39 S. C. Kalhan and P. A. Adam, *J. Clin. Endocrinol. Metab.*, 1975, **41**, 600–610.
- 40 T. M. Witten and D. Bonchev, *Chem. Biodiversity*, 2007, **4**, 2639–2655.
- 41 D. Bonchev and G. A. Buck, *J. Chem. Inf. Model.*, 2007, **47**, 909–917.
- 42 Y. Onoda, Y. Tsunematsu and K. Kamei, *Jpn. J. Antibiot.*, 1974, **27**, 471–474.
- 43 E. Estrada, *J. Proteome Res.*, 2006, **5**, 2177–2184.
- 44 E. Estrada, *Proteomics*, 2006, **6**, 35–40.
- 45 Y. Kanazawa, *J. Antibiot.*, 1966, **19**, 175–189.
- 46 N. Gupta, N. Mangal and S. Biswas, *Proteins: Struct., Funct., Bioinf.*, 2005, **59**, 196–204.
- 47 C. L. Webber, Jr, A. Giuliani, J. P. Zbilut and A. Colosimo, *Proteins: Struct., Funct., Bioinf.*, 2001, **44**, 292–303.
- 48 U. Gobel, C. Sander, R. Schneider and A. Valencia, *Proteins: Struct., Funct., Bioinf.*, 1994, **18**, 309–317.
- 49 A. Krishnan, J. P. Zbilut, M. Tomita and A. Giuliani, *Curr. Protein Pept. Sci.*, 2008, **9**, 28–38.
- 50 A. Krishnan, A. Giuliani, J. P. Zbilut and M. Tomita, *PLoS One*, 2008, **3**, e2149.
- 51 M. C. Palumbo, A. Colosimo, A. Giuliani and L. Farina, *FEBS Lett.*, 2007, **581**, 2485–2489.
- 52 A. Krishnan, A. Giuliani, J. P. Zbilut and M. Tomita, *J. Proteome Res.*, 2007, **6**, 3924–3934.
- 53 A. Krishnan, A. Giuliani and M. Tomita, *PLoS One*, 2007, **2**, e562.
- 54 H. González-Díaz, Y. González-Díaz, L. Santana, F. M. Ubeira and E. Uriarte, *Proteomics*, 2008, **8**, 750–778.
- 55 C. R. Munteanu, J. M. Vázquez, J. Dorado, A. Pazos-Sierra, A. Sánchez-González, F. J. Prado-Prado and H. González-Díaz, *J. Proteome Res.*, 2009, **8**, 5219–5228.
- 56 H. Gonzalez-Diaz, L. G. Perez-Montoto, A. Duardo-Sanchez, E. Paniagua, S. Vazquez-Prieto, R. Vilas, M. A. Dea-Ayuela, F. Bolas-Fernandez, C. R. Munteanu, J. Dorado, J. Costas and F. M. Ubeira, *J. Theor. Biol.*, 2009, **261**, 136–147.
- 57 G. Agüero-Chapin, A. Antunes, F. M. Ubeira, K. C. Chou and H. González-Díaz, *J. Chem. Inf. Model.*, 2008, **48**, 2265–2277.
- 58 M. Cruz-Monteagudo, C. R. Munteanu, F. Borges, M. N. D. S. Cordeiro, E. Uriarte, K.-C. Chou and H. González-Díaz, *Polymer*, 2008, **49**, 5575–5587.
- 59 M. A. Dea-Ayuela, Y. Perez-Castillo, A. Meneses-Marcel, F. M. Ubeira, F. Bolas-Fernandez, K. C. Chou and H. Gonzalez-Diaz, *Bioorg. Med. Chem.*, 2008, **16**, 7770–7776.
- 60 G. Agüero-Chapin, H. Gonzalez-Díaz, G. de la Riva, E. Rodriguez, A. Sanchez-Rodriguez, G. Podda and R. I. Vazquez-Padron, *J. Chem. Inf. Model.*, 2008, **48**, 434–448.
- 61 G. Ferino, H. Gonzalez-Díaz, G. Delogu, G. Podda and E. Uriarte, *Biochem. Biophys. Res. Commun.*, 2008, **372**, 320–325.
- 62 H. Safayhi, S. E. Boden, S. Schweizer and H. P. Ammon, *Planta Med.*, 2000, **66**, 110–113.
- 63 H. Gonzalez-Díaz, F. Prado-Prado and F. M. Ubeira, *Curr. Top. Med. Chem.*, 2008, **8**, 1676–1690.
- 64 H. Gonzalez-Díaz, R. Molina and E. Uriarte, *FEBS Lett.*, 2005, **579**, 4297–4301.
- 65 V. A. Ivanisenko, S. S. Pintus, D. A. Grigorovich and N. A. Kolchanov, *Nucleic Acids Res.*, 2005, **33**, D183–D187.
- 66 P. D. Dobson and A. J. Doig, *J. Mol. Biol.*, 2003, **330**, 771–783.
- 67 C. R. Munteanu and A. L. Magalhaes, *International Journal of Bioinformatics Research and Applications*, 2009, **5**, 402–416.
- 68 M. H. Kutner, C. J. Nachtsheim, J. Neter and W. Li, *Applied Linear Statistical Models*, McGraw Hill, New York, 5th edn, 2005, ch. Multiple Regression II, pp. 271–277.
- 69 S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and L. D. J., *Nucleic Acids Res.*, 1997, **25**, 389–402.
- 70 A. Marchler-Bauer and S. H. Bryant, *Nucleic Acids Res.*, 2004, **32**, W327–W331.
- 71 S. Vilar, H. González-Díaz, L. Santana and E. Uriarte, *J. Comput. Chem.*, 2008, **29**, 2613–2622.
- 72 S. Wu and Y. Zhang, *Nucleic Acids Res.*, 2007, **35**, 3375–3382.
- 73 A. Duardo-Sanchez, in *Topological Indices for Medicinal Chemistry, Biology, Parasitology, Neurological and Social Networks*, ed. H. González-Díaz and C. R. Munteanu, Transworld Research Network, Kerala, India, 2010, ch. 12, pp. 205–212.
- 74 J. Soding, *Bioinformatics*, 2005, **21**, 951–960.
- 75 S. Wu and Y. Zhang, *Proteins: Struct., Funct., Bioinf.*, 2008, **72**, 547–556.
- 76 A. Huett and R. Xavier, *Autophagy*, 2009, **5**, 884–886.
- 77 E. Onischenko, L. H. Stanton, A. S. Madrid, T. Kieselbach and K. Weis, *J. Cell Biol.*, 2009, **185**, 475–491.
- 78 R. Redon, S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shaperro, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. Gonzalez, M. Gratacos, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodward, F. Yang, J. Zhang, T. Zerjal, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C. Lee, K. W. Jones, S. W. Scherer and M. E. Hurles, *Nature*, 2006, **444**, 444–454.
- 79 B. A. Reimel, S. Pan, D. H. May, S. A. Shaffer, D. R. Goodlett, M. W. McIntosh, L. M. Yerien, M. P. Bronner, R. Chen and T. A. Brentnall, *Curr. Proteomics*, 2009, **6**, 63–69.
- 80 : Hypercube Inc., Gainesville, FL, USA, 2002.
- 81 Y. Liu and D. L. Beveridge, *Proteins: Struct., Funct., Bioinf.*, 2002, **46**, 128–146.
- 82 M. Froimowitz, *BioTechniques*, 1993, **14**, 1010–1013.
- 83 Hypercube Inc., Gainesville, FL, USA, 7.0 edn, 2002.
- 84 M. Mezo, M. Gonzalez-Warleta and F. M. Ubeira, *J. Parasitol.*, 2003, **89**, 843–849.
- 85 O. Ivanciuc, *Curr. Top. Med. Chem.*, 2008, **8**, 1691–1709.
- 86 C. H. Schein, O. Ivanciuc and W. Braun, *Immunol. Allergy Clin. North Am.*, 2007, **27**, 1–27.
- 87 O. Ivanciuc and W. Braun, *Protein Pept. Lett.*, 2007, **14**, 903–916.
- 88 O. Ivanciuc, N. Oezguen, V. S. Mathura, C. H. Schein, Y. Xu and W. Braun, *Curr. Med. Chem.*, 2004, **11**, 583–593.
- 89 O. Ivanciuc, C. H. Schein and W. Braun, *Bioinformatics*, 2002, **18**, 1358–1364.
- 90 H. Van Waterbeemd, *Chemometric Methods in Molecular Design*, Wiley-VCH, New York, 1995.
- 91 A. Garcia-Garcia, J. Galvez, J. V. de Julian-Ortiz, R. Garcia-Domenech, C. Munoz, R. Guna and R. Borrás, *J. Biomol. Screening*, 2005, **10**, 206–214.
- 92 A. Garcia-Garcia, J. Galvez, J. V. de Julian-Ortiz, R. Garcia-Domenech, C. Munoz, R. Guna and R. Borrás, *J. Antimicrob. Chemother.*, 2004, **53**, 65–73.
- 93 R. Gozalbes, M. Brun-Pascaud, R. Garcia-Domenech, J. Galvez, G. Pierre-Marie, D. Jean-Pierre and F. Derouin, *Antimicrob. Agents Chemother.*, 2000, **44**, 2771–2776.
- 94 R. Gozalbes, J. Galvez, R. Garcia-Domenech and F. Derouin, *SAR QSAR Environ. Res.*, 1999, **10**, 47–60.
- 95 Y. Marrero-Ponce, A. Meneses-Marcel, O. M. Rivera-Borroto, R. Garcia-Domenech, J. V. De Julian-Ortiz, A. Montero, J. A. Escario, A. G. Barrio, D. M. Pereira, J. J. Nogal, R. Grau, F. Torrens, C. Vogel and V. J. Aran, *J. Comput.-Aided Mol. Des.*, 2008, **22**, 523–540.
- 96 Y. Marrero-Ponce, S. E. Ortega-Broche, Y. Echeverria Diaz, Y. J. Alvarado, N. Cubillan, G. Casas Cardoso, F. Torrens and F. Perez-Gimenez, *J. Theor. Biol.*, 2009, **259**, 229–241.
- 97 Y. Marrero-Ponce, R. Medina-Marrero, J. A. Castillo-Garrit, V. Romero-Zaldivar, F. Torrens and E. A. Castro, *Bioorg. Med. Chem.*, 2005, **13**, 3003–3015.
- 98 Y. Marrero-Ponce, D. Nodarse, H. González-Díaz, R. Ramos de Armas, V. Romero-Zaldivar, F. Torrens and E. A. Castro, *Int. J. Mol. Sci.*, 2004, **5**, 276–293.
- 99 O. Ivanciuc, in *Encyclopedia of Complexity and System Science*, ed. R. A. Meyers, Springer-Verlag, New York, 2009.
- 100 O. Ivanciuc, in *QSPR/QSAR Studies by Molecular Descriptors*, ed. M. V. Diudea, Nova Science, Huntington, NY, 2001, pp. 213–231.
- 101 T. Hill and P. Lewicki, *STATISTICS Methods and Applications. A Comprehensive Reference for Science, Industry and Data Mining*, StatSoft, Tulsa, 2006.
- 102 H. B. Shen and K. C. Chou, *Anal. Biochem.*, 2008, **373**, 386–388.
- 103 H. B. Shen and K. C. Chou, *Protein Eng., Des. Sel.*, 2007, **20**, 561–567.
- 104 K. C. Chou and H. B. Shen, *Nat. Protoc.*, 2008, **3**, 153–162.
- 105 A. Schlessinger, G. Yachdav and B. Rost, *Bioinformatics*, 2006, **22**, 891–893.

- 106 H. W. Mewes, D. Frishman, K. F. Mayer, M. Munsterkotter, O. Noubibou, P. Pagel, T. Rattei, M. Oesterheld, A. Ruepp and V. Stumpflen, *Nucleic Acids Res.*, 2006, **34**, D169–D172.
- 107 D. Xie, A. Li, M. Wang, Z. Fan and H. Feng, *Nucleic Acids Res.*, 2005, **33**, W105–W110.
- 108 J. McDermott, M. Guerquin, Z. Frazier, A. N. Chang and R. Samudrala, *Nucleic Acids Res.*, 2005, **33**, W324–W325.
- 109 R. Concu, M. A. Dea-Ayuela, L. G. Perez-Montoto, F. J. Prado-Prado, E. Uriarte, F. Bolas-Fernandez, G. Podda, A. Pazos, C. R. Munteanu, F. M. Ubeira and H. Gonzalez-Diaz, *Biochim. Biophys. Acta, Proteins Proteomics*, 2009, **1794**, 1784–1794.
- 110 M. W. Robinson, J. F. Tort, J. Lowther, S. M. Donnelly, E. Wong, W. Xu, C. M. Stack, M. Padula, B. Herbert and J. P. Dalton, *Mol. Cell. Proteomics*, 2008, **7**(6), 1111–1123.
- 111 F. M. Ubeira, L. Muino, M. A. Valero, M. V. Periago, I. Perez-Crespo, M. Mezo, M. Gonzalez-Warleta, F. Romaris, E. Paniagua, S. Cortizo, J. Llovo and S. Mas-Coma, *Am. J. Trop. Med. Hyg.*, 2009, **81**, 156–162.
- 112 M. Mezo, M. Gonzalez-Warleta, J. A. Castro-Hermida, C. Carro and F. M. Ubeira, *Vet. Parasitol.*, 2010, **168**, 36–44.
- 113 M. Mezo, M. Gonzalez-Warleta, C. Carro and F. M. Ubeira, *J. Parasitol.*, 2004, **90**, 845–852.
- 114 M. Mezo, M. Gonzalez-Warleta and F. M. Ubeira, *J. Parasitol.*, 2007, **93**, 65–72.
- 115 O. Golden, R. J. Flynn, C. Read, M. Sekiya, S. M. Donnelly, C. Stack, J. P. Dalton and G. Mulcahy, *Vaccine*, 2010, **28**, 5551–5557.
- 116 S. Mas-Coma, M. A. Valero and M. D. Bargues, *Adv. Parasitol.*, 2009, **69**, 41–146.
- 117 S. Aksu, C. Scheler, N. Focks, F. Leenders, F. Theuring, J. Salnikow and P. R. Jungblut, *Proteomics*, 2002, **2**, 1452–1463.
- 118 A. Tebbe, C. Klein, B. Bisle, F. Siedler, B. Scheffer, C. Garcia-Rizo, J. Wolfertz, V. Hickmann, F. Pfeiffer and D. Oesterheld, *Proteomics*, 2005, **5**, 168–179.
- 119 M. Hirose, M. Hoshida, M. Ishikawa and T. Toya, *CABIOS, Comput. Appl. Biosci.*, 1993, **9**, 161–167.
- 120 K. A. Resing, K. Meyer-Arendt, A. M. Mendoza, L. D. Aveline-Wolf, K. R. Jonscher, K. G. Pierce, W. M. Old, H. T. Cheung, S. Russell, J. L. Wattawa, G. R. Goehle, R. D. Knight and N. G. Ahn, *Anal. Chem.*, 2004, **76**, 3556–3568.
- 121 M. M. Savitski, M. L. Nielsen, F. Kjeldsen and R. A. Zubarev, *J. Proteome Res.*, 2005, **4**, 2348–2354.
- 122 Z. Lei, A. M. Elmer, B. S. Watson, R. A. Dixon, P. J. Mendes and L. W. Sumner, *Mol. Cell. Proteomics*, 2005, **4**, 1812–1825.
- 123 M. C. Giddings, A. A. Shah, R. Gesteland and B. Moore, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 20–25.
- 124 M. A. Dea-Ayuela and F. Bolás-Fernández, *Vet. Parasitol.*, 2005, **132**, 43–49.
- 125 M. Tamura and A. Yamazaki, *Kekkaku*, 1967, **42**, 165–172.
- 126 L. Roche, J. Tort and J. P. Dalton, *Mol. Biochem. Parasitol.*, 1999, **98**, 271–277.
- 127 C. M. Stack, C. R. Caffrey, S. M. Donnelly, A. Sessaadri, J. Lowther, J. F. Tort, P. R. Collins, M. W. Robinson, W. Xu, J. H. McKerrow, C. S. Craik, S. R. Geiger, R. Marion, L. S. Brinen and J. P. Dalton, *J. Biol. Chem.*, 2008, **283**, 9896–9908.
- 128 P. D. Dobson, Y. D. Cai, B. J. Stapley and A. J. Doig, *Curr. Med. Chem.*, 2004, **11**, 2135–2142.
- 129 D. Dickson, *Nature*, 2000, **405**, 264.
- 130 R. J. Zagursky, P. S. Conway and M. A. Kashdan, *BioTechniques*, 1991, **11**, 3638.
- 131 G. Ehren, *Nervenarzt*, 1993, **64**, 481–483.
- 132 H. F. Judson, *Gene*, 1993, **135**, 19–23.
- 133 E. Katz, *Public Hist.*, 1994, **17**, 41–50.
- 134 S. Abdulla, R. P. Alderton, R. J. Glynn, I. G. Gut, L. K. Hosking, A. Jackson, A. Kelly, W. R. Newell, E. Radley, P. Sanseau, K. L. Thorpe, J. Trowsdale and S. Beck, *DNA Sequence*, 1996, **7**, 5–7.
- 135 E. Todd, *Pennsylvania Nurse*, 1996, **51**, 21.
- 136 D. G. Knorre, D. V. Pyshnyi, E. M. Ivanova, A. A. Bondar, I. V. Morozov, N. P. Mertvetsov and V. F. Zarytova, *Dokl. Akad. Nauk*, 1996, **350**, 119–120.
- 137 T. Avis, E. K. Clark, T. L. Flack, M. Mohammadi, S. Milne, D. Niblett, S. Palmer, S. Phillips, C. Smalley, M. Tagney, K. L. Thorpe, B. Tubby, J. Westhorp and S. Beck, *DNA Sequence*, 1997, **8**, 131–135.
- 138 K. D. Pruitt, *Genome Res.*, 1998, **8**, 4–8.
- 139 S. A. Soper, D. C. Williams, Y. Xu, S. J. Lassiter, Y. Zhang, S. M. Ford and R. C. Bruch, *Anal. Chem.*, 1998, **70**, 4036–4043.
- 140 Y. Kobayashi, S. Ishizaki, Y. Nagashima and K. Shiomi, *Parasitol. Int.*, 2008, **57**, 314–319.
- 141 A. J. Theaker, G. L. Maslen, C. E. Scott, C. M. Rice, S. E. Hunt, A. King, A. J. Mungall, I. Dunham and S. Beck, *DNA Sequence*, 1997, **8**, 167–171.
- 142 R. Rodriguez-Perez, I. Moneo, A. Rodriguez-Mahillo and M. L. Caballero, *Mol. Biochem. Parasitol.*, 2008, **159**, 92–97.
- 143 W. Ansoerge, J. Zimmermann, C. Schwager, J. Stegemann, H. Erfle and H. Voss, *Nucleic Acids Res.*, 1990, **18**, 3419–3420.
- 144 W. Ansoerge, H. Voss, U. Wirkner, C. Schwager, J. Stegemann, R. Pepperkok, J. Zimmermann and H. Erfle, *J. Biochem. Biophys. Methods*, 1989, **20**, 47–52.
- 145 S. S. Ner, D. B. Goodin, G. J. Pielak and M. Smith, *BioTechniques*, 1988, **6**, 408410, 412.
- 146 A. M. Anadon, E. Rodriguez, M. T. Garate, C. Cuellar, F. Romaris, T. Chivato, M. Rodero, H. Gonzalez-Diaz and F. M. Ubeira, *Clin. Vaccine Immunol.*, 2010, **17**, 496–502.
- 147 A. M. Anadon, F. Romaris, M. Escalante, E. Rodriguez, T. Garate, C. Cuellar and F. M. Ubeira, *Clin. Exp. Immunol.*, 2009, **156**, 471–478.
- 148 E. Rodriguez, A. M. Anadon, E. Garcia-Bodas, F. Romaris, R. Iglesias, T. Garate and F. M. Ubeira, *Allergy (Oxford, U. K.)*, 2008, **63**, 219–225.
- 149 P. Puente, A. M. Anadon, M. Rodero, F. Romaris, F. M. Ubeira and C. Cuellar, *Exp. Parasitol.*, 2008, **118**, 271–274.
- 150 J. T. Marques, T. Devosse, D. Wang, M. Zamanian-Daryoush, P. Serbinowski, R. Hartmann, T. Fujita, M. A. Behlke and B. R. Williams, *Nat. Biotechnol.*, 2006, **24**, 559–565.
- 151 O. Mekenyan, D. Bonchev, N. Trinajstić and D. Peitchev, *Arzneim.-Forsch., Beih.*, 1986, **36**, 421–424.
- 152 O. Mekenyan, D. Peitchev, D. Bonchev, N. Trinajstić and I. Bangov, *Arzneim.-Forsch., Beih.*, 1986, **36**, 176–183.