



Plasmod-PPI: A web-server predicting complex biopolymer targets in plasmodium with entropy measures of protein–protein interactions

Yamilet Rodríguez-Soca^a, Cristian R. Munteanu^b, Julian Dorado^b, Juan Rabuñal^b, Alejandro Pazos^b, Humberto González-Díaz^{a,*}

^a Department of Microbiology & Parasitology, Faculty of Pharmacy, USC, 15782, Santiago de Compostela, Spain

^b Department of Information and Communication Technologies, Computer Science Faculty, University of A Coruña, Campus de Elviña, 15071, A Coruña, Spain

ARTICLE INFO

Article history:

Received 18 October 2009

Received in revised form

7 November 2009

Accepted 12 November 2009

Available online 26 November 2009

Keywords:

Protein–Protein interactions

Plasmodium proteome

Protein 3D-Electrostatic interactions

ABSTRACT

We can define structural indices of polymer or biopolymer complex structures and use them in the prediction of new drug targets in parasites. For instance, *Plasmodium falciparum* causes the most severe form of Malaria and kills up to 2.7 million people annually whereas *Plasmodium vivax* is geographically the most widely distributed cause with more than 80 million clinical cases. Due to drug resistance and toxicity, discovering novel drug targets is mandatory; such as Protein–Protein Complexes unique in this pathogen and not present in human host (pPPCs). Additionally, the 3D structure of an increasing number of Plasmodium proteins is being reported in public databases making easier the development of bio-informatics models to predict pPPCs. In addition, some PPCs expressed both in parasite and human, such as DHFR synthase, play a significant role in drug resistance in both Malaria and Human Cancer. However, there are no general models to predict pPPCs using indices of PPC biopolymer structure. Therefore, we introduced herein new Markov Chain numerical descriptors of protein–protein Interactions (PPIs) based on electrostatic entropy measures and calculated these parameters for 5257 pairs of proteins (774 pPPCs and 4483 non-pPPCs) from more than 20 organisms, including parasite and human hosts. We found a simple Classification Tree with high Accuracy, Sensitivity, and Specificity (90.2–98.5%) both in training and independent test sub-sets and implemented this predictor in the user-friendly web server PlasmodPPI freely available at <http://miaja.tic.udc.es/Bio-AIMS/PlasmodPPI.php>.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction

Plasmodium falciparum (*P. falciparum*) represents one of the strongest selective forces on the human genome. This stable and perennial pressure has contributed to the progressive accumulation in the exposed populations of genetic adaptations to malaria. Descriptive genetic epidemiology provides the initial step of a logical procedure of consequential phases spanning from the identification of genes involved in the resistance/susceptibility to diseases, to the determination of the underlying mechanisms and finally to the possible translation of the acquired knowledge in new control tools [1]. In addition, *Plasmodium vivax* (*P. vivax*) is geographically the most widely distributed cause of malaria in people, with up to 2.5 billion people at risk and an estimated 80 million to 300 million clinical cases every year, including severe

disease and death. Despite this large burden of disease, *P. vivax* is overlooked and left in the shadow of the enormous problem caused by *P. falciparum* in Sub-Saharan Africa. Both technological advances enabling the sequencing of the *P. vivax* genome and a recent call for worldwide malaria eradication have placed a new emphasis on the importance of addressing *P. vivax* as a major public health problem. However, because of this parasite's biology, it is especially difficult to interrupt the transmission of *P. vivax*, and experts agree that the available methods for preventing and treating both infections with *P. vivax* and *P. falciparum* are inadequate [2]. Malaria, perhaps one of the most serious and widespread diseases encountered by mankind, continues to be a major threat to about 40% of the world's population, especially in the developing world. As malaria vaccines remain problematic, chemotherapy still is the most important weapon in the fight against the disease. However, almost all available drugs have been compromised by the highly adaptable parasite, and the increasing drug resistance of *P. falciparum* continues to be the main problem. Therefore, the limited clinical repertoire of effective drugs and the emergence of multi-resistant strains substantiate the need for new proteins, or the discovery of

* Corresponding author. Tel.: +34 981 563100; fax: +34 981 594912.

E-mail addresses: humberto.gonzalez@usc.es, gonzalezdiaz@yahoo.es (H. González-Díaz).

new functions for known proteins, that may become targets of new anti-malarial compounds or the discovery of proteins involved in multi-drug resistance [3–8]. It is thus imperative that the development of new methods and strategies becomes a priority [2]. In this regard, stable protein–protein complexes formed by Protein–Protein Interactions (PPIs) may become interesting targets for new drugs and other treatment methods or strategies. For instance, there are high-molecular-weight rhoptry proteins of *P. falciparum* in a multi-protein complex consisting of proteins of 140, 130, and 110 kDa. The complex of rhoptry proteins binds to human and mouse erythrocyte membranes in association with a 120 kDa SERA protein. These proteins are believed to participate in the process of erythrocyte invasion. Sam-Yellowed have used six different antibodies (polyclonal and monoclonal) known to precipitate the high-molecular-weight rhoptry protein complex to analyze the structural relationship of proteins within the complex. The results provided insights concerning the mechanism of protein–protein interaction within the complex [9].

These types of results indicate that physically stable protein–protein biopolymer complexes (pPPC) made up of unique PPIs of *Plasmodium* sp. parasites (pPPIs) and not present in humans or other hosts may be promising targets for the development of safe drugs with low toxicity. On the contrary, the prediction of non-pPPC (non-unique *Plasmodium* sp. parasites but also present in humans) may become a source for the discovery of targets related to drug resistance not only for the treatment of malaria but also of human cancer. For instance, Human Dihydrofolate Reductase (DHFR) constitutes a primary target for antifolate drugs in cancer treatment, whereas DHFRs from *P. falciparum* and *P. vivax* are primary targets in the treatment of malaria. A recent review [10] has discussed the structural and functional impact of active-site mutations with respect to enzyme activity and antifolate resistance of DHFRs from mammals, protozoa and bacteria. DHFR is a monomeric protein with only one chain in structures deposited in PDB. However, DHFR synthase is a non-pPPC polymeric protein, which is directly involved in DHFR synthesis and consequently in drug resistance. For instance, the structure of DHFR synthase reported in the file with PDB-ID 3HBB is a PPC with four different protein chains. In this regard, a computational model able to predict non-pPPC such as DHFRs may be interesting for the prediction of protein targets involved in drug resistance in both parasite and mammalian, which may be useful in the design of chemo-protective agents.

In any case, the high number of possible genes/proteins discovered in genome/proteome of *Plasmodium* sp. determines a higher number of possible pPPC/non-pPPC structures derived from different PPIs in parasite and human hosts, which makes difficult the exhaustive experimental investigation in terms of time and resources [11,12]. In fact, many researchers in the field of Molecular and Biochemical Parasitology have recognized the high importance of different computational tools (statistical models, servers, databases) to study the proteome and/or genome of *P. falciparum* and *P. vivax* [13–18]. This fact determines that the development of predictive models for pPPIs/non-PPIs discrimination becomes a very useful tool aimed at discovering new drug targets. There are many theoretical methods for the prediction of PPIs in humans and other organisms. Many of them are based on the same approaches used for the study of protein structure–function relationships but extended to PPIs such as: sequence alignment techniques, phylogenic techniques, or alignment-free parameters besides other methods, like molecular modeling, incorporate knowledge about the 3D structure of the proteins involved in the PPIs. These methods often make use of complex trees representations (as input or output of the analysis) to represent these interactions as PPIs trees. Sequence-only methods are

often faster than 3D ones and need less structural information. On the contrary, 3D methods give a more clear idea on the structure of the protein and may be used to predict proteins with known spatial structure but unknown function [19–27]. The importance of these latter methods is that these functionally non-annotated structures become common in the Protein Data Bank (PDB) with the development of powerful characterization techniques [28]. Another role of the computational methods is the possibility to study not only the wild-type proteins but also the computational analysis of mutations [29–33]. Specifically, in this work, we are interested in computational methods to predict pPPIs that determine the formation of non-covalent but physically stable PPCs between two proteins that can be isolated and the 3D structure, chemically characterized as a potential drug target. Protein complexes are fundamental for understanding principles of cellular organizations. As the sizes of PPI trees are increasing, accurate and fast protein complex prediction from these PPI trees can be useful as a guide for biological experiments to discover novel protein complexes [34]. Otherwise, it is the direct prediction of complexes by protein–protein docking but it may become computationally expensive if we aim at performing the screening of large databases [35]. It is also of major importance to recall that nowadays it is not enough to develop a predictive model; we should also implement it into public servers, preferably of free access, for the use of the scientific community. The server packages developed by Chou and Shen [36–39], which predict the function of proteins from structural parameters or explore protein structures, are good examples in this regard. In any case, to the best of our knowledge, there is no web server available in the literature or at least a theoretical method to predict unique pPPC in *Plasmodium* and not present in humans or other parasites or hosts, based on the 3D structure of the two proteins involved in pPPIs or non-PPIs interactions.

Besides, González-Díaz et al. introduced the method called 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 to describe RNA secondary structures based on molecular vibration information [40], and 3D structure of proteins based on Van der Waals [41] or electrostatic interactions [42]. Recently, 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 also complex chemical structures [43–45]. 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 to a graph or complex network representation of this system, at the same time. This describes more adequately the broad uses of the method to numerically characterize the structure of drugs [46], RNA [40], and proteins [41,47,48], as well as drug–drug networks [49], drug–protein interactions [50], PPIs, and other systems such as an MCM associated to a graph. In this regard, MARCH-INSIDE uses networks similar to other 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) [51–58]. In Fig. 1 we depict the 3D structure and the Van der Waals surface for Thioredoxin (PDB-ID SYRC) a pPPC present in *P. falciparum* clone 3d7 (A) and the respective protein structure complex network graph for one of the proteins of the pPPC (B). At this structural level, the nodes are amino acids and we link two nodes with an edge if the distance between them is lower than 15 Å (this type of network is also known as contact map or protein residue networks) [59–66]. In a very recent review, we have discussed the details and many

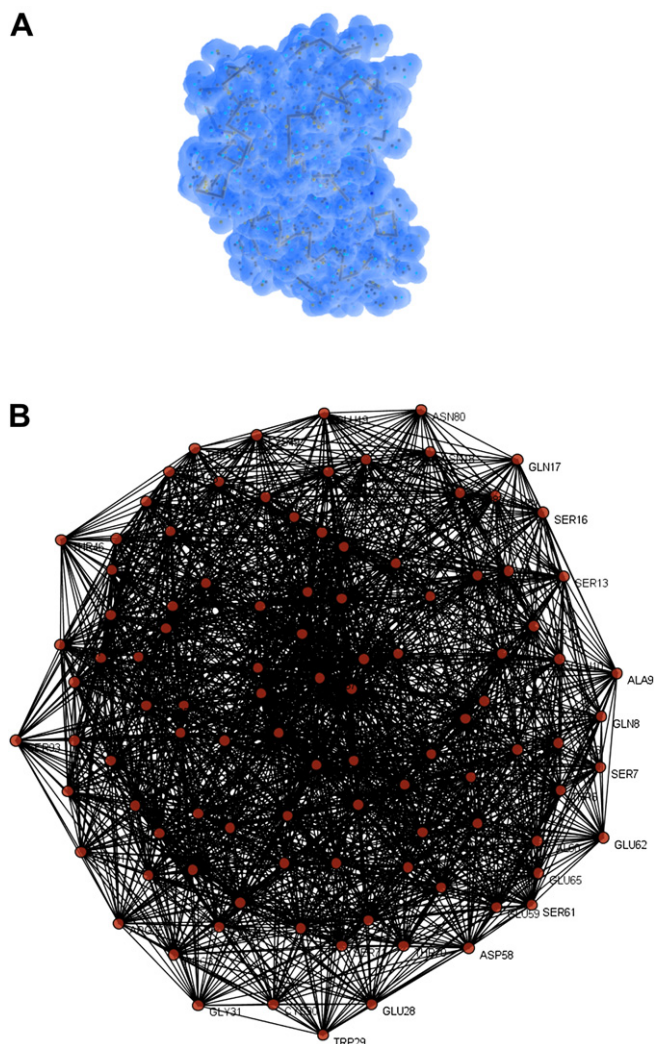


Fig. 1. 3D structure and Van der Waals surface for a *P. falciparum* protein (A) and complex network (B).

applications of the MARCH-INSIDE method to Molecular Microbiology [67].

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 on the same philosophy of online free access and use by all the international research community, as mentioned in the previous paragraph. In particular, the server called TargetPred package offers two new Protein-QSAR servers. The first, ATCUNPred (<http://miaja.tic.udc.es/Bio-AIMS/ATCUNPred.php>) is available for prediction of ATCUN-mediated DNA-cleavage anticancer proteins [68]. The second server, EnzClassPred is available at <http://miaja.tic.udc.es/Bio-AIMS/EnzClassPred.php> and can be used to predict enzyme classes from PDB files without function annotation [69]. For all these reasons, in this work we use the MARCH-INSIDE approach for the first time to solve the problem of predicting specific pPPCs from the 3D structure of two proteins that may undergo pPPCs or not. Last but not least, we implemented the predictor in a new web server named PlasmodPPI freely available to public at <http://miaja.tic.udc.es/Bio-AIMS/PlasmodPPI.php>. In Fig. 2 we depict a flowchart for all the steps taken in this work to generate the new classifiers and server.

2. Materials and methods

2.1. Electrostatic entropy measures for PPIs

In previous works we have used different entropy invariants derived from an MCM to describe the 3D structure of one protein backbone in structure–property relationship studies. The $\theta_k(R)$ parameters used represent the average electrostatic entropy (θ) due to the interactions between all pairs of amino acids allocated inside a specific protein region (R) and placed at a distance k from each other. In this work we want to use $\theta_k(R)$ values of two proteins, $\theta_k(^1R)$ for protein 1 and $\theta_k(^2R)$ for protein 2, in order to generate structural parameters describing PPI between these proteins. To this end, we introduced herein for the first time a new type of PPI invariants in the sense that they do not depend on the interchange of proteins so that we do not need to label and distinguish them for calculation. We introduce, with this aim, three types of invariants (ti) ${}^{ti}\theta_k(R)$: PPI Average Entropy Invariant ($ti = a$), PPI Entropy Difference Invariant ($ti = d$), and PPI Entropy Product Invariant ($ti = p$):

$${}^a\theta_k(R) = {}^a\theta_k(^1R_1, ^2R_1) = \frac{1}{2}[\theta_k(^1R_1) + \theta_k(^2R_1)] \quad (1)$$

$${}^d\theta_k(R) = {}^d\theta_k(^1R_1, ^2R_1) = |\theta_k(^1R_1) - \theta_k(^2R_1)| \quad (2)$$

$${}^p\theta_k(R) = {}^p\theta_k(^1R_1, ^2R_1) = \theta_k(^1R_1) \cdot \theta_k(^2R_1) \quad (3)$$

Notably, in order to guarantee that these parameters are invariant to protein labeling as 1 or 2, we have to always use the same ${}^1R = {}^2R = R$ and $k_1 = k_2 = k$ values. In order to calculate the $\theta_k(R)$ values for each protein the method uses as a source of protein macromolecular descriptors the stochastic matrices ${}^1\Pi_e$ built up as squared matrices ($n \times n$), where n is the number of amino acids (aa) in the protein. The subscript e points to the electrostatic type of molecular force field. In previous works we have predicted the protein function based on $\theta_k(R)$ values for different types of interactions or molecular fields. The main types of molecular fields used are the following: Electrostatic, vdW, and HINT entropies. In this paper, we calculated $\theta_k(R)$ values only for Electrostatic entropies. These values have been used herein to calculate PPIs invariants and next as inputs to generate the QSAR model (see description of PPI invariants above). However, the detailed explanation for the calculation of $\theta_k(R)$ values has been published before. As follows, we give the formula for ${}^k\theta(R)$ values and some general explanations [41,67,70]:

$$\theta_k(R) = - \sum_{j=1}^n {}^kP_j(R) \cdot \log[{}^kP_j(R)] \quad (4)$$

It is remarkable that the average entropy measures depend on the absolute probabilities ${}^kP_j(R)$ according to which the amino acid j th has an electrostatic interaction with the rest of amino acids that lie within the same protein region R . These probabilities refer to amino acids placed at a distance equal to k -times the cut-off distance ($r_{ij} = k \cdot r_{\text{cut-off}}$). The method uses a Markov Chain Model (MCM) to calculate these probabilities, which also depend on the 3D interactions between all pairs of amino acids placed at distance r_{ij} in r_3 in the protein structure. 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 neighboring aa ($\alpha_{ij} = 1$ if $r_{ij} < r_{\text{cut-off}}$). Otherwise, the interaction is banished ($\alpha_{ij} = 0$). The relationship α_{ij} may be displayed as a protein structure complex network. In this network the nodes are the C_α atoms of the amino acids and the edges connect

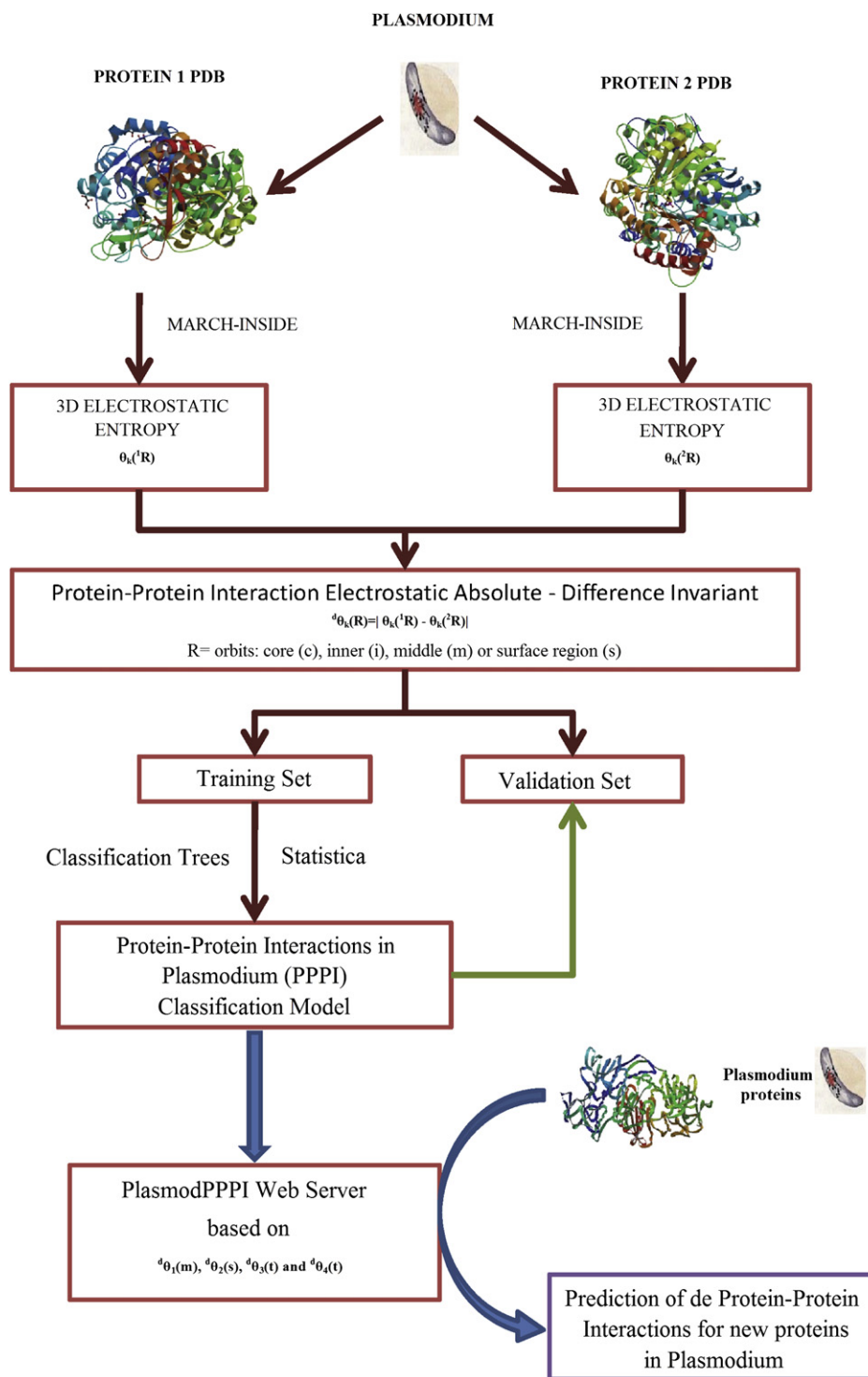


Fig. 2. Example of spatial distribution of core, inner, middle, and surface amino acids.

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 on protein PDB files. For the calculation, all water molecules and metal ions were removed [67]. All calculations were carried out with our in-house software MARCH-INSIDE 2.0 [71].

For the calculation, the MARCH-INSIDE software always uses the full matrix, never a sub-matrix, but may run the last summation term either for all amino acids or only for some specific groups, called Orbits or Regions (R). These regions are often defined in

geometric terms and called core, inner, middle or surface region. In Fig. 3 we represented the orbits of protein (*c* corresponds to core, *i* to inner, *m* to middle, and *s* to surface orbits, respectively). The diameters of the orbits, are: $0 \leq \text{orbit } c < 25$, $25 \leq \text{orbit } i < 50$, $50 \leq \text{orbit } m < 75$, and $76 \leq \text{orbit } s \leq 100$; expressed in terms of percentage of the longest distance r_{\max} with respect to the center of charge. Additionally, we take into consideration the total orbit (*t*) that contains all the amino acids in the protein (orbit diameter 0–100% of r_{\max}). Consequently, we can calculate different $\theta_k(R)$ for the

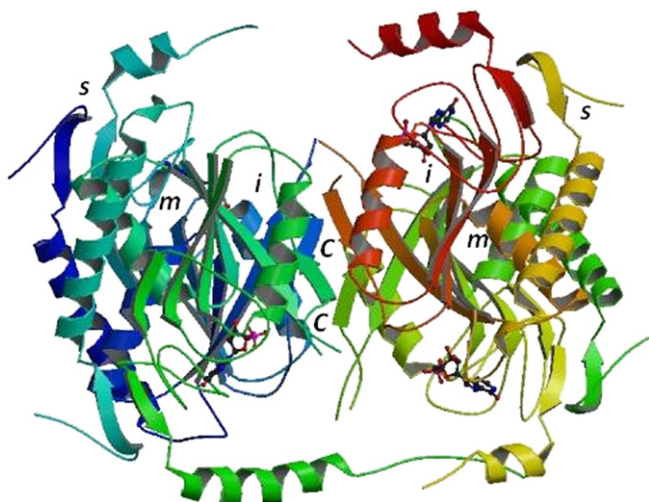


Fig. 3. Flowchart for all the steps given in the construction of the classifiers and server.

amino acids contained in an orbit ($c, i, m, s,$ or t) and placed at a topological distance k within this orbit (k is the order named) [72–75]. In this work, we calculated altogether $5(\text{types of regions}) \times 6(\text{orders considered}) = 30 \theta_k(R)$ indices for each protein.

In order to carry out the calculations referred to in equation (1) for any kind of entropy and detailed in the previous equations, for electrostatic entropy, the elements (${}^1p_{ij}$) of ${}^1\Pi_e$ and the absolute initial probabilities ${}^A p_0(j)$ were calculated as follows [67]:

$${}^1p_{ij} = \frac{\alpha_{ij} \cdot E_{ij}}{\sum_{m=1}^{\delta+1} \alpha_{im} \cdot E_{im}} = \frac{\alpha_{ij} \cdot \frac{q_i \cdot q_j}{(d_{ij})^2}}{\sum_{m=1}^{\delta+1} \alpha_{im} \cdot \frac{q_i \cdot q_m}{(d_{im})^2}} \quad (5)$$

$${}^A p_0(j) = \frac{\frac{q_j}{d_{0j}}}{\sum_{m=1}^n \frac{q_m}{(d_{0m})^2}} \quad (6)$$

where q_i and q_j are the electronic charges for amino acids i th-aa and the j th-aa and the neighborhood relationship (truncation function $\alpha_{ij} = 1$) is turned on if these amino acids participate in a peptidic hydrogen bond or $d_{ij} < d_{\text{cut-off}} = 5 \text{ \AA}$ [67]. In this regard, the truncation of the molecular field is usually applied to simplify all the calculations in large biological systems. The distance d_{ij} is the Euclidean distance between the C_α atoms of the two amino acids and d_{0j} the distance between the amino acid and the center of charge of the protein. Both kinds of distances were derived from the x, y and z coordinates of the amino acids collected from the protein PDB files. All calculations were carried out with our in-house software MARCH-INSIDE. All water molecules and metal ions were removed for the calculation [67].

2.2. LDA models

LDA is frequently used for classification/prediction problems in physical anthropology, but it is unusual to find examples where researchers consider the statistical limitations and assumptions required for this technique. In this work, all LDA models have been trained with the software STATISTICA 6.0[®], for which our laboratory holds rights of use [76]. In LDA we use several variable selection techniques to seek the model: i) *All Effects* (include all parameters), ii) *Forward-stepwise*, iii) *Forward-entry*, iv) *Backward-stepwise*, v) *Backward-removal*, and vi) *Best subsets*. Unless we specify a different value, we always set a prior probability of

$p(\text{pPPI}) = p(\text{npPPI}) = 0.5$. The LDA discriminant equation was obtained using as input the three types of PPI invariants ${}^t i \theta_k(R)$. The general form of the equation obtained by LDA is:

$$S(\text{pPPC}) = \sum_{R,k,t}^{5,5,3} a_{R,k,t} \cdot {}^t i \theta_k(R) + a_0 \quad (7)$$

$S(\text{pPPC})$, the output of this model, is a real value variable that scores the propensity of a protein pair to undergo a pPPI interaction and not npPPIs forming a physically stable PPCs only in *Plasmodium* sp. The χ^2 and p -level value were examined in order to test the statistical significance of the model. The Accuracy, Specificity, Sensitivity were used to quantify the goodness-of-fit and the discriminatory power of the model. Different authors like have applied this type of LDA model using different classes of input variables to construct QSAR models for proteins or nucleic acids [77–80].

2.3. CT models

CTs have been used to test a non-linear model which is not based on assumptions of parametric distribution of data as well as non-linear models [81]. We used as Ordered Predictors the variables obtained in the Forward stepwise of the LDA. Starting from now on, several split methods were carried out: i) CT Discriminant-based Linear Combinations (CT-LC), ii) Discriminant-based univariate splits (CT-US), and CRT-style exhaustive search from univariate splits (CRT). In CRT we used three different measures of Goodness-of-fit Gini Measure, Chi-Square, and G-Square. Like in LDA we always set a prior probability of $p(\text{pPPI}) = p(\text{npPPI}) = 0.5$, unless we specify a different value. Last, we used a FACT-style direct stopping rule with a value of 0.01 to control the length of the CT. All the CTs have been trained with the software STATISTICA 6.0[®], for which our laboratory holds rights of use [76].

2.4. Dataset

The protein structures were downloaded from PDB [82] using the following schemes for PDB-database search: (i) introducing the name of the parasite species (*Plasmodium*) as input parameter in the search item called source organism (for positive cases) or (ii) introducing the PDB-IDs for all the proteins contained in the list reported in the article of Dobson and Doig [83]. The positive cases (pPPI) are those protein–protein pairs that make up a stable complex that has been structurally characterized (3D structure) in *Plasmodium* species (*Plasmodium* sp). The list of negative cases (npPPI), search scheme (b), contain enzymes and other proteins present in humans and many other organisms including other parasites that are not present in *Plasmodium* sp. The dataset consisted of 5257 pairs of proteins (774 pPPIs and 4483 npPPIs) from more than 20 organisms, including parasites and human or cattle hosts. Altogether, 581 pPPIs and 3395 npPPIs were used in training and 193 pPPIs and 1088 npPPIs were used in validation. Detailed information about the PDB-ID, the values of the electrostatic entropy indices, the corresponding observed classification, and the predicted classification for each pPPI or npPPI pair are given in the Supporting information.

3. Results and discussion

Several researchers have demonstrated the high performance of different types of computational classifiers in protein or PPI structure–function relationship studies based on different algorithms as is the case, for instance, of the works carried out by Chou

et al. [84–90], Fernandez and Caballero [91–93]. In particular, the LDA algorithm, a simpler type of the classifier used herein, was employed to train linear models based on different combinations of parameters [94].

3.1. Linear discriminant analysis (LDA) models

A simple Linear Discriminant Analysis (LDA), with only four variables, was developed to assign each protein pair as pPPI or npPPI. The best equation found was:

$$S(\text{pPPC}) = -0.09506 \cdot {}^d\theta_3(m) - 0.02219 \cdot {}^d\theta_4(s) \\ - 0.62697 \cdot {}^d\theta_5(t) + 0.51126 \cdot {}^d\theta_4(t) - 0.30646 \\ N = 3976 \quad \chi^2 = 947.95 \quad p < 0.00 \quad (8)$$

The statistical parameters for the above equation are: Number of protein entries in training (N), Chi-square statistic (χ^2), and error level (p -level), which have to be <0.05 [95]. All the statistical data of this model are summed up in Table 1. The discriminant function reported in the results section presented statistically significant results of goodness-of-fit for both training and validation series, carried out with an external series of pPPI and npPPI that were never used to train the model. Interestingly four variables, ${}^d\theta_3(m)$, ${}^d\theta_4(s)$, ${}^d\theta_4(t)$ and ${}^d\theta_5(t)$, out of more than 30 parameters calculated appear in many models. These parameters have the general formula ${}^d\theta_k(R) = |\theta_k(R)_{\text{prot1}} - \theta_k(R)_{\text{prot2}}|$, which are the absolute difference between the electrostatic entropy values $\theta_k(R)$ for amino

acids on the surface of the two proteins forming the PPI pairs. This fact indicates that the difference between the surface electrostatic entropy is very important not only for PPI interactions in general but also to discriminate the unique complex present in *Plasmodium* sp. (pPPIs) and not in other organisms. The model presents a good overall classification of pPPI and npPPI. This level of accuracy is generally accepted by other researchers that have applied LDA to find QSAR models useful in molecular parasitology and related areas; e.g., the works of García-Domenech, Marrero-Ponce, Bruno-Blanch, Galvez, Gozalbes and others predicting active compounds against *Trypanosoma cruzi*, *Mycobacterium avium*, *Toxoplasma gondii*, *P. falciparum*, *Trichomonas vaginalis*, *Fasciola hepatica*, and other parasites [96–100]; see also the works of Marrero-Ponce on protein and DNA/RNA QSAR studies [101–103].

3.2. Artificial neural network (ANN) models

The comparison of linear and non-linear models is essential to test how directly our parameters are correlated to the biological property [104]. The automatic selection of variables (features) was activated for all models. In particular, the Linear Neural Network (LNN) algorithm and other types of Artificial Neural Network (ANN), were used herein to train different linear and non-linear models based on different combinations of parameters. Table 1 also depicts the results for the best models found. The profile of the ANN model was specified with a simple notation as follows: ANN type N_{iv} : N_{in} - N_{H1} - N_{H2} - N_{on} : N_{ov} . The ANN types presented, besides LNN, are Multi-Layer Perceptron (MLP), Probabilistic Neural Network (PNN), and Radial Basis Function (RBF) [105]. The parameter N_{iv} is the number of input variables, N_{in} is the number of input neurons (one per input variable), N_{H1} is the number of neurons in the first Hidden layer (H1), N_{H2} is the number of neurons in the second Hidden layer (H1), N_{on} is the number of output neurons, and N_{ov} is the number of output variables.

Table 1
Summary of results for LDA, CT, and ANN analysis.

Technique			Training sub-set			Validation sub-set		
Profile	Parameters	Group	%	npPPI	pPPI	%	npPPI	pPPI
LDA	Specificity	npPPI	85.0	2886	509	82.4	897	191
	Sensitivity	pPPI	94.8	30	551	92.7	14	179
	Accuracy	Total	86.4	–	–	84.0	–	–
CT	Specificity	npPPI	98.5	3343	52	98.0	1066	22
	Sensitivity	pPPI	91.2	51	530	90.2	19	174
	Accuracy	Total	97.4	–	–	96.8	–	–
US	Specificity	npPPI	95.6	3247	148	96.5	1050	38
	Sensitivity	pPPI	83.8	94	487	84.5	30	163
	Accuracy	Total	93.9	–	–	94.7	–	–
CRT	Specificity	npPPI	97.6	3315	80	97.8	1064	24
	Sensitivity	pPPI	84.7	89	492	83.4	32	161
	Accuracy	Total	95.7	–	–	95.6	–	–
Chi-square	Specificity	npPPI	97.6	3315	80	97.8	1064	24
	Sensitivity	pPPI	84.7	89	492	83.4	32	161
	Accuracy	Total	95.7	–	–	95.6	–	–
G-square	Specificity	npPPI	98.6	3348	47	98.4	1071	17
	Sensitivity	pPPI	81.8	106	475	80.3	38	155
	Accuracy	Total	96.2	–	–	95.7	–	–
MLP	Sensitivity	pPPI	83.3	484	97	82.9	160	33
	Specificity	npPPI	84.0	544	2851	82.9	186	902
	Accuracy	Total	83.9	–	–	82.9	–	–
MLP	Sensitivity	pPPI	83.1	483	98	81.9	158	35
	Specificity	npPPI	83.0	577	2818	81.6	200	888
	Accuracy	Total	83.0	–	–	81.7	–	–
RBF	Sensitivity	pPPI	18.9	110	471	20.2	39	154
	Specificity	npPPI	17.3	2807	588	15.5	919	169
	Accuracy	Total	17.6	–	–	16.2	–	–
LNN	Sensitivity	pPPI	92.6	538	43	90.2	174	19
	Specificity	npPPI	92.2	264	3131	90.4	104	984
	Accuracy	Total	92.3	–	–	90.4	–	–

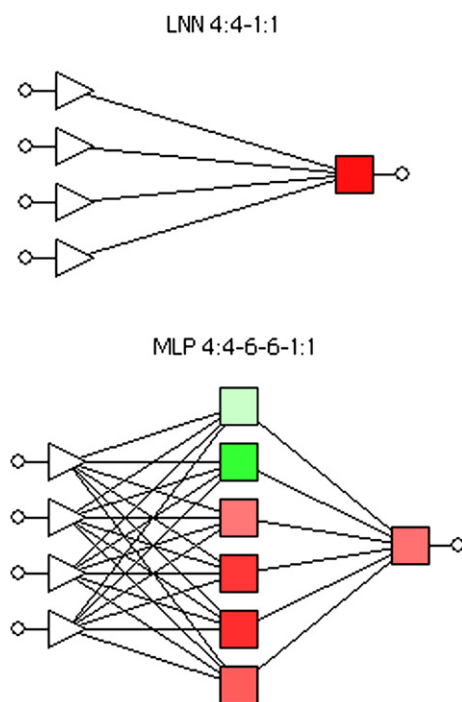


Fig. 4. Illustrative example of the topology used for different ANNs trained in this work.

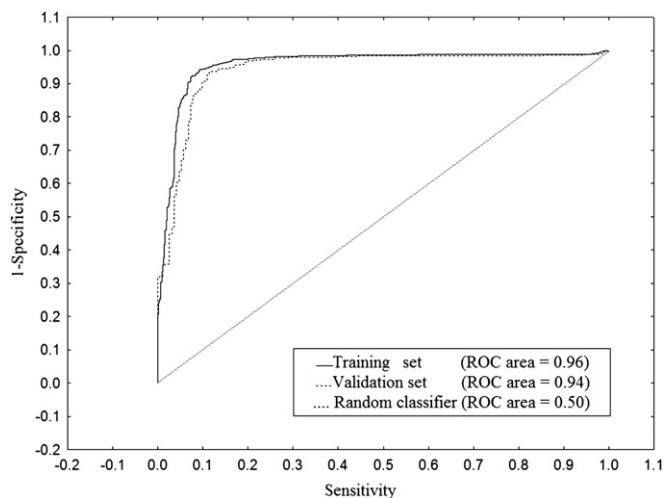


Fig. 5. ROC curve for pPPC predictor.

In particular, the model LNN 4:4–1:1 is the simplest model found with the highest levels of Sensitivity = 92.6, Specificity = 92.2 and Accuracy = 92.3 in the training set. These values are excellent considering that this predictor uses only two molecular descriptors of the PPI pair, which is a very complex structure in chemical terms, to fit a large data set of 581 pPPIs and 3395 npPPIs. The profile 4:4–1:1 indicates that this model assign the values of four input variables to four input neurons that perform a weighed sum and assigns the result to one output neuron; which gives the final result of the case classification according to the threshold value that has been optimized. In addition, the model LNN 4:4–1:1 also presented a higher levels of Sensitivity = 90.2, Specificity = 90.4 and Accuracy = 90.4 in external validation (test) set (see Table 1). In Fig. 4 we illustrate the topology of this LNN network compared with a non-linear ANN. Interestingly, four variables

$d\theta_3(m)$, $d\theta_4(s)$, $d\theta_4(t)$ and $d\theta_5(t)$, out of more than 30 parameters calculated, appear in many models. This fact indicates that the difference between the electrostatic entropy is very important not only for PPI interactions in general but also to discriminate a unique complex present in *Plasmodium* (pPPIs). On the other hand, the product and average invariant types (${}^a\theta_k(R)$ and ${}^p\theta_k(R)$) do not appear to be relevant.

We also validated the linear model by means of a ROC curve analysis (see Fig. 5) to demonstrate that there is a linear and not an indirect non-linear relationship between our indices and the classification of pPPCs [106]. The values of the area under the ROC curve for this model are 0.95 and 0.96 very close to 1 (the highest possible value) and notably different from 0.5 (the typical value of a random classifier). This kind of analysis is an accepted tool in Bioinformatics to demonstrate which classification methods outperform the other methods, e.g. the study carried out by Xu and Du related to PPIs [107] or the work of Mahdavi and Lin [108]. This first search points to a linear instead of non-linear relationship between pPPI prediction and $d\theta_k(R)$ values, giving additional proofs of the validity of our methodology. For instance, in Table 1 we can see that more complicated models with non-linear profiles do not improve the linear model and give even worse results sometimes.

3.3. Classification Tree (CT) models

Last, considering that non-linear ANN did not notably improved LDA, we used the variables pre-selected by LDA as inputs for a Classification Tree (CT) analysis. With complete data sets, LDA may be a simpler and sometimes better choice. However, the testing of data prior to analysis is necessary, and CTs are recommended either as a replacement for LDA or as a supplement whenever data do not meet relevant assumptions [109]. Table 1 also depicts the results for the best CT models found. The automatically selection of variables (features) was activated for all models if available. In Fig. 6 we illustrate the graph representation

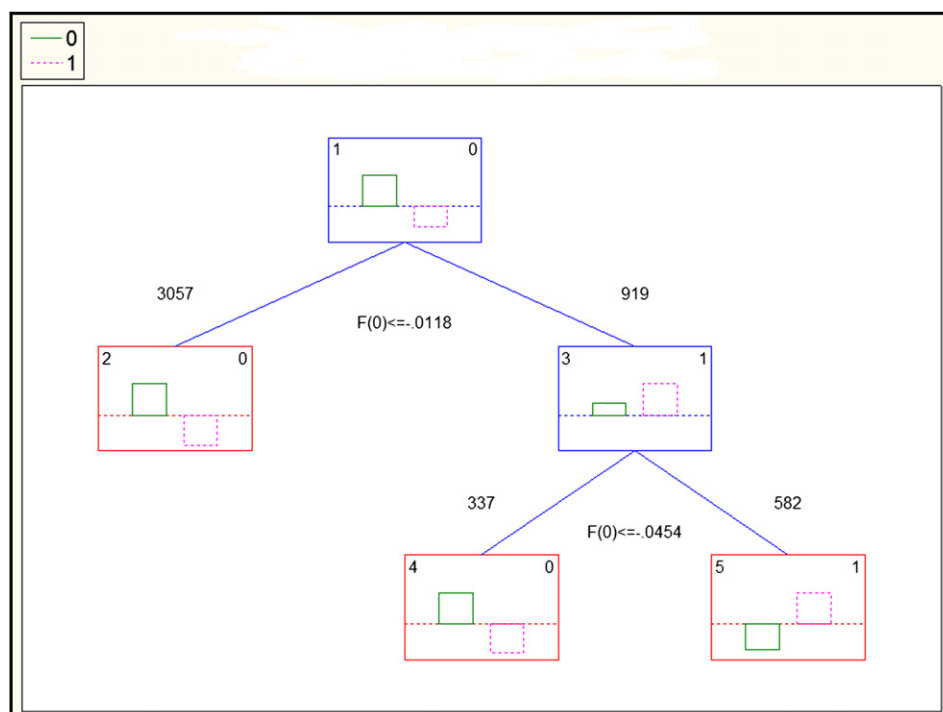


Fig. 6. Structure of the CT model found.


Table 2
Structure of the CT-LC model.

Parameters	Parent nodes				
	1	2	3	4	5
Child nodes					
Left branch	2		4		
Right branch	3		5		
npPPI	3395	3018	377	325	52
pPPI	581	39	542	12	530
Predicted class	npPPI	npPPI	pPPI	npPPI	pPPI
Split conditions ($LC_i \leq$ Split constant)					
LC_i	LC_1	LC_2	LC_3	LC_4	LC_5
Split constant	0.011758	0	0.045360	0	0
$d\theta_3(m)$	-0.000827	0	-0.004075	0	0
$d\theta_4(s)$	-0.000193	0	-0.001044	0	0
$d\theta_4(t)$	0.005454	0	0.018150	0	0
$d\theta_5(t)$	-0.004447	0	-0.014544	0	0

of the CT-LC trained in this work and in Table 2 we give details about the structure of this CT and the split rules derived. In particular, the model CT-LC is the simplest CT model found with the highest levels of Sensitivity = 91.2% Specificity = 98.5% and Accuracy = 97.4% in the training set. These values are excellent considering that this predictor uses only two molecular descriptors of the PPI pair; which is a very complex structure in chemical terms, to fit a large data set of 582 TPPIs and 3394 non-TPPIs (see Table 1). In fact, the CT analysis yielded the best model found in this work.

3.4. PlasmodPPI, a server for PPC plasmodium targets


Last, 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 [36–39,110–113]. In this regard, we implemented this predictor into a web server freely available to public at <http://miaja.tic.udc.es/Bio-AIMS/PlasmodPPI.php>. This is the first model and web server that



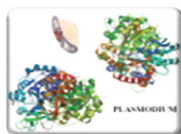
RGB Groups
RNASA, TIC
University of A Coruña
Microbiology & Parasitology
University of Santiago de Compostela
Spain

PlasmodPPI @ Bio-AIMS

Modelling the reality



Home | Theory | About



Plasmod-PPI
Plasmodium Protein-Protein Interactions (PPPI)


Tool: MARCH-INSIDE (Python version)
Data: RCSB PDB

PDB-chain lists : Please paste the names of the PDB chains as two lists (max. 50)

Notes: There is no space between the PDB name and the chain label, no empty new line; the results will print the pairs between the chain from list 1 with the chain from list 2 (not the combination of the list items).

3C5IA
2F6IE
1SYRC


3C5IE
2GHUA
1SYRF



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

PlasmodPPI.calc @ Bio-AIMS

Modelling the reality



Home | Theory | About

```

Process ID = 137494af1487ed0412
PDB List 1 = 3C5IA 2F6IE 1SYRC
PDB List 2 = 3C5IE 2GHUA 1SYRF

... please wait ....

PDB Update/Verification [List 1] ...
3C5IA 2F6IE 1SYRC

PDB Update/Verification [List 2] ...
3C5IE 2GHUA 1SYRF

Processing PDB-chain List 1 ...
3C5IA 2F6IE 1SYRC
Processing PDB-chain List 2 ...
3C5IE 2GHUA 1SYRF

Result file = Results/137494af1487ed0412/PlasmodPPI.calc.txt

Calculated at 2009-11-04 10:25:19

Chain1 Chain2 Complex
=====
3C5IA 3C5IE YES
2F6IE 2GHUA NO
1SYRC 1SYRF YES

```

Fig. 7. Example of use of PlasmodPPI web tool: (A) Input and (B) Output pages.

predicts how unique is a protein–protein complex in Plasmodium proteome with respect to other parasites and hosts breaking new ground for anti-plasmodium drug target discovery.

In order to demonstrate the practical utility of this Web server, three examples of protein chain pairs have been used to evaluate the possibility to make up unique complexes in Plasmodium, a human pathogen parasite: 3C5IA–3C5IE, 2F6IE–2GHUA and 1SYRC–1SYRF. Fig. 7 presents the input (A) and output (B) web pages of the PlasmodPPI tool. The first pair contains the first chain A of the *Plasmodium knowlesi* choline kinase (a transferase, 3C5I) and the cleaved fragment of N-terminal expression tag (chain E), all expressed in *Escherichia coli*. Choline kinase is the first enzyme in the Kennedy pathway (CDP-choline pathway) for the biosynthesis of the most essential phospholipid, phosphatidylcholine, in Plasmodium. In addition, choline kinase also plays a pivotal role in trapping essential polar head group choline inside the malaria parasite. The inhibition of choline kinase will lead to a decrease in phosphocholine, which in turn causes a decrease in phosphatidylcholine biosynthesis, resulting in death of the parasite. This pair of protein chains is evaluated to make up the unique complex in Plasmodium that can be a target for new anti-parasite drugs. The second pair example is formed by the chain E of the 2F6I hydrolase [114], a ATP-dependent CLP protease (serine-type endopeptidase) from *Plasmodium falciparum* (expressed in *E. coli*) and the chain A of the 2GHU hydrolase, Falcipain-2 (FP-2) of *P. falciparum* [115]. FP-2 is a papain-family (C1A) cysteine protease that plays an important role in the parasite life cycle by degrading erythrocyte proteins, most notably hemoglobin. Inhibition of FP-2 and its paralogues prevents parasite maturation. These two chains of hydrolases are not evaluated by our tool to form a unique complex. This can be explained by the different targets of these hydrolases and different cellular localizations (2F6I in cytoplasm and 2GHU in food vacuole for hemoglobin degradation and cleavage of cytoskeletal elements). The last example is formed by the chains C and F of the 1SYR protein, a *Plasmodium falciparum* thioredoxin in the genetic structure with an unknown function [116]. These chains are evaluated to form a unique complex according to the localization of both chains in the same protein. PlasmodPPI tool can become important for the discovery of new anti-plasmodium drug targets and can be useful as model for building similar models for other types of parasites or other organisms.

4. Conclusions

The overall findings suggest that the new type of parameters introduced herein is useful to numerically characterize the structure of PPCs, formed after PPIs, in protein structure–function studies. We also demonstrate that it is possible to distinguish between PPCs (pPPCs cases) formed according to unique PPIs in *Plasmodium* sp. (pPPIs) and not present in other parasites or host organisms using these parameters. We generate and compare linear and non-linear classifiers. We show that it is possible to predict PPIs that undergo pPPC formation with a simple linear classifier based on the absolute difference between 3D protein surface electrostatic entropies of the pair proteins. The model was implemented in a public web server, available for free-of-charge use to the scientific community.

Acknowledgments

We thank the kind and professional attention of Prof. J.E. Mark (Computational & Theoretical Polymer Science editor for Polymer) as well as the opinion of the reviewers. Gonzalez–Díaz H. and Munteanu C.R. acknowledge research contract financed by the Contract/grant sponsor: Isidro Parga Pondal Program, Xunta de

Galicia. The authors thank for the partial financial support from the grants 2007/127 and 2007/144 from the General Directorate of Scientific and Technological Promotion of the Galician University System of the Xunta de Galicia and from grant (Ref. PIO52048 and RD07/0067/0005) funded by the Carlos III Health Institute.

Appendix. Supplementary data

Supplementary data associated with this article can be found in online version, at doi:10.1016/j.polymer.2009.11.029.

References

- [1] Verra F, Mangano VD, Modiano D. Parasite Immunol 2009;31(5):234–53.
- [2] Mueller I, Galinski MR, Baird JK, Carlton JM, Kochar DK, Alonso PL, et al. Lancet Infect Dis 2009;9(9):555–66.
- [3] Bonilla JA, Bonilla TD, Yowell CA, Fujioka H, Dame JB. Mol Microbiol 2007;65(1):64–75.
- [4] Turschner S, Efferth T. Mini Rev Med Chem 2009;9(2):206–2124.
- [5] Sanchez CP, Rotmann A, Stein WD, Lanzer M. Mol Microbiol 2008;70(4):786–98.
- [6] Sanchez CP, Rohrbach P, McLean JE, Fidock DA, Stein WD, Lanzer M. Mol Microbiol 2007;64(2):407–20.
- [7] Nunes MC, Goldring JP, Doerig C, Scherf A. Mol Microbiol 2007;63(2):391–403.
- [8] Siden-Kiamos I, Ecker A, Nyback S, Louis C, Sinden RE, Billker O. Mol Microbiol 2006;60(6):1355–63.
- [9] Sam-Yellowe TY. Exp Parasitol 1993;77(2):179–94.
- [10] Volpato JP, Pelletier JN. Drug Resist Updat 2009;12(1–2):28–41.
- [11] Carucci DJ, Yates 3rd JR, Florens L. Int J Parasitol 2002;32(13):1539–42.
- [12] Coppel RL, Black CG. Int J Parasitol 2005;35(5):465–79.
- [13] Bender A, van Dooren GG, Ralph SA, McFadden GI, Schneider G. Mol Biochem Parasitol 2003;132(2):59–66.
- [14] Carlton JM, Muller R, Yowell CA, Fluegge MR, Sturrock KA, Pritt JR, et al. Mol Biochem Parasitol 2001;118(2):201–10.
- [15] Coppel RL. Mol Biochem Parasitol 2001;118(2):139–45.
- [16] Cui L, Fan Q, Hu Y, Karamycheva SA, Quackenbush J, Khuntirat B, et al. Mol Biochem Parasitol 2005;144(1):1–9.
- [17] Gunasekera AM, Patankar S, Schug J, Eisen G, Kissinger J, Roos D, et al. Mol Biochem Parasitol 2004;136(1):35–42.
- [18] Huestis R, Fischer K. Mol Biochem Parasitol 2001;118(2):187–99.
- [19] Sharon I, Davis JV, Yona G. Methods Mol Biol 2009;541:61–88.
- [20] Liu L, Cai Y, Lu W, Feng K, Peng C, Niu B. Biochem Biophys Res Commun 2009;380(2):318–22.
- [21] Skrabanek L, Saini HK, Bader GD, Enright AJ. Mol Biotechnol 2008;38(1):1–17.
- [22] Najafabadi HS, Salavati R. Genome Biol 2008;9(5):R87.
- [23] Kim S, Shin SY, Lee IH, Kim SJ, Srimam R, Zhang BT. Nucleic Acids Res 2008;36(Web Server issue):W411–5.
- [24] Jaeger S, Gaudan S, Leser U, Rebholz-Schuhmann D. BMC Bioinformatics 2008;8(9 Suppl):S2.
- [25] Burger L, van Nimwegen E. Mol Syst Biol 2008;4:165.
- [26] Scott MS, Barton GJ. BMC Bioinformatics 2007;8:239.
- [27] Zvebil MJ, Tang L, Cookson E, Selkirk ME, Thornton JM. Mol Biochem Parasitol 1993;58(1):145–53.
- [28] von Grotthuss M, Plewczynski D, Ginalski K, Rychlewski L, Shakhnovich EI. BMC Bioinformatics 2006;7:53.
- [29] Lappalainen I, Thusberg J, Shen B, Vihinen M. Proteins 2008;72(2):779–92.
- [30] Shen B, Bai J, Vihinen M. Protein Eng Des Sel 2008;21(1):37–44.
- [31] Shen B, Vihinen M. Protein Eng Des Sel 2004;17(3):267–76.
- [32] Liu ML, Shen BW, Nakaya S, Pratt KP, Fujikawa K, Davie EW, et al. Blood 2000;96(3):979–87.
- [33] Shen B, Nolan JP, Sklar LA, Park MS. Nucleic Acids Res 1997;25(16):3332–8.
- [34] Chua HN, Ning K, Sung WK, Leong HW, Wong L. J Bioinform Comput Biol 2008;6(3):435–66.
- [35] Smith GR, Sternberg MJ. Curr Opin Struct Biol 2002;12(1):28–35.
- [36] Shen HB, Chou KC. Anal Biochem 2008;373(2):386–8.
- [37] Shen HB, Chou KC. Protein Eng Des Sel 2007;20(11):561–7.
- [38] Chou KC, Shen HB. Biochem Biophys Res Commun 2007; doi:10.1016/j.bbrc.2007.10.06.1027.
- [39] Chou KC, Shen HB. Nat Protoc 2008;3(2):153–62.
- [40] González-Díaz H, de Armas RR, Molina R. Bioinformatics 2003;19(16):2079–87.
- [41] González-Díaz H, Saiz-Urra L, Molina R, Santana L, Uriarte E. J Proteome Res 2007;6(2):904–8.
- [42] Gonzalez-Diaz H, Molina R, Uriarte E. FEBS Lett 2005;579(20):4297–301.
- [43] Concu R, Podda G, Uriarte E, Gonzalez-Diaz H. J Comput Chem 2009;30:1510–20.
- [44] Gonzalez-Diaz H, Saiz-Urra L, Molina R, Gonzalez-Diaz Y, Sanchez-Gonzalez A. J Comput Chem 2007;28(6):1042–8.
- [45] González-Díaz H, Pérez-Castillo Y, Podda G, Uriarte E. J Comput Chem 2007;28:1990–5.

- [46] Santana L, Uriarte E, González-Díaz H, Zagotto G, Soto-Otero R, Mendez-Alvarez E. *J Med Chem* 2006;49(3):1149–56.
- [47] Aguero-Chapin G, Varona-Santos J, de la Riva GA, Antunes A, Gonzalez-Villa T, Uriarte E, et al. *J Proteome Res* 2009;8(4):2122–8.
- [48] Concu R, Dea-Ayuela MA, Perez-Montoto LG, Bolas-Fernandez F, Prado-Prado FJ, Podda G, et al. *J Proteome Res* 2009;8(9):4372–82.
- [49] Santana L, Gonzalez-Diaz H, Quezada E, Uriarte E, Yanez M, Vina D, et al. *J Med Chem* 2008;51(21):6740–51.
- [50] Vina D, Uriarte E, Orallo F, Gonzalez-Diaz H. *Mol Pharmacol* 2009;6(3):825–35.
- [51] Bornholdt S, Schuster HG. *Handbook of graphs and complex networks: from the genome to the internet*. Weinheim: WILEY-VCH GmbH & CO. KGa; 2003.
- [52] Mazurie A, Bonchev D, Schwikowski B, Buck GA. *Bioinformatics* 2008;24(22):2579–85.
- [53] Managbanag JR, Witten TM, Bonchev D, Fox LA, Tsuchiya M, Kennedy BK, et al. *PLoS One* 2008;3(11):e3802.
- [54] Witten TM, Bonchev D. *Chem Biodivers* 2007;4(11):2639–55.
- [55] Bonchev D, Buck GA. *J Chem Inf Model* 2007;47(3):909–17.
- [56] Bonchev D. *SAR QSAR Environ Res* 2003;14(3):199–214.
- [57] Estrada E. *J Proteome Res* 2006;5(9):2177–84.
- [58] Estrada E. *Proteomics* 2006;6(1):35–40.
- [59] Gupta N, Mangal N, Biswas S. *Proteins* 2005;59(2):196–204.
- [60] Webber Jr CL, Giuliani A, Zbilut JP, Colosimo A. *Proteins* 2001;44(3):292–303.
- [61] Gobel U, Sander C, Schneider R, Valencia A. *Proteins* 1994;18(4):309–17.
- [62] Krishnan A, Zbilut JP, Tomita M, Giuliani A. *Curr Protein Pept Sci* 2008;9(1):28–38.
- [63] Krishnan A, Giuliani A, Zbilut JP, Tomita M. *PLoS One* 2008;3(5):e2149.
- [64] Palumbo MC, Colosimo A, Giuliani A, Farina L. *FEBS Lett* 2007;581(13):2485–9.
- [65] Krishnan A, Giuliani A, Zbilut JP, Tomita M. *J Proteome Res* 2007;6(10):3924–34.
- [66] Krishnan A, Giuliani A, Tomita M. *PLoS ONE* 2007;2(6):e562.
- [67] González-Díaz H, González-Díaz Y, Santana L, Ubeira FM, Uriarte E. *Proteomics* 2008;8:750–78.
- [68] Munteanu CR, Vázquez JM, Dorado J, Pazos-Sierra A, Sánchez-González A, Prado-Prado FJ, et al. *Proteome Res* 2009; doi:10.1021/pr900556g.
- [69] Concu R, Dea-Ayuela MA, Perez-Montoto LG, Prado-Prado FJ, Uriarte E, Bolas-Fernandez F, et al. *Biochim Biophys Acta* 2009; doi:10.1016/j.bbapap.2009.1008.1020.
- [70] González-Díaz H, Molina R, Uriarte E. *Bioorg Med Chem Lett* 2004;14(18):4691–5.
- [71] Gonzalez-Diaz H, Prado-Prado F, Ubeira FM. *Curr Top Med Chem* 2008;8(18):1676–90.
- [72] González-Díaz H, Saíz-Urra L, Molina R, Uriarte E. *Polymer* 2005;46(8):2791–8.
- [73] González-Díaz H, Molina-Ruiz R, and Hernandez I. MARCH- INSIDE v3.0 (MAR kov CH ains IN variants for SI mulation & DE sign); Windows supported version under request to the main author contact email: gonzalezdiaz@yahoo.es; 2007.
- [74] Cruz-Monteagudo M, Gonzalez-Diaz H. *Eur J Med Chem* 2005;40(10):1030–41.
- [75] Gonzalez-Diaz H, Aguero-Chapin G, Varona J, Molina R, Delogu G, Santana L, et al. *J Comput Chem* 2007;28(6):1049–56.
- [76] StatSoft.Inc. STATISTICA (data analysis software system), version 6.0, www.statsoft.com. Statsoft, Inc; 2002.
- [77] Marrero-Ponce Y, Medina-Marrero R, Castro AE, Ramos de Armas R, González-Díaz H, Romero-Zaldivar V, et al. *Molecules* 2004;9:1124–47.
- [78] Ramos de Armas R, Gonzalez Diaz H, Molina R, Uriarte E. *Proteins* 2004;56(4):715–23.
- [79] Ramos de Armas R, González-Díaz H, Molina R, Perez Gonzalez M, Uriarte E. *Bioorg Med Chem* 2004;12(18):4815–22.
- [80] Ramos de Armas R, González-Díaz H, Molina R, Uriarte E. *Biopolymers* 2005;77(5):247–56.
- [81] Hill T, Lewicki P. *Statistics methods and applications. A comprehensive reference for science, industry and data mining*. Tulsa: StatSoft; 2006.
- [82] Ivanisenko VA, Pintus SS, Grigorovich DA, Kolchanov NA. *Nucleic Acids Res* 2005;33(Database issue):D183–7.
- [83] Dobson PD, Doig AJ. *J Mol Biol* 2003;330(4):771–83.
- [84] Chou KC. *J Proteome Res* 2005;4(4):1413–8.
- [85] Chou KC, Elrod DW. *J Proteome Res* 2003;2(2):183–90.
- [86] Chou KC, Shen HB. *J Proteome Res* 2006;5:1888–97.
- [87] Chou KC, Shen HB. *J Proteome Res* 2006;5:3420–8.
- [88] Chou KC, Shen HB. *J Proteome Res* 2007;6:1728–34.
- [89] Chou KC, Cai YD. *J Proteome Res* 2006;5(2):316–22.
- [90] Chou KC, Elrod DW. *J Proteome Res* 2002;1(5):429–33.
- [91] Fernández M, Caballero F, Fernández L, Abreu JI, Acosta G. *Proteins* 2008;70(1):167–75.
- [92] Caballero J, Fernandez M. *Curr Top Med Chem* 2008;8(18):1580–605.
- [93] Fernández L, Caballero J, Abreu JI, Fernández M. *Proteins* 2007;67:834–52.
- [94] Guha R, Jurs PC. *J Chem Inf Comput Sci* 2004;44(6):2179–89.
- [95] Van Waterbeemd H. Discriminant analysis for activity prediction. In: Van Waterbeemd H, editor. *Chemometric methods in molecular design*, vol. 2. New York, NY: Wiley-VCH; 1995. p. 265–82.
- [96] Garcia-Garcia A, Galvez J, de Julian-Ortiz JV, Garcia-Domenech R, Munoz C, Guna R, et al. *J Biomol Screen* 2005;10(3):206–14.
- [97] Garcia-Garcia A, Galvez J, de Julian-Ortiz JV, Garcia-Domenech R, Munoz C, Guna R, et al. *J Antimicrob Chemother* 2004;53(1):65–73.
- [98] Gozalbes R, Brun-Pascaud M, Garcia-Domenech R, Galvez J, Pierre-Marie G, Jean-Pierre D, et al. *Antimicrobial Agents Chemother* 2000;44(10):2771–6.
- [99] Gozalbes R, Galvez J, Garcia-Domenech R, Derouin F. *SAR QSAR Environ Res* 1999;10(1):47–60.
- [100] Marrero-Ponce Y, Meneses-Marcel A, Rivera-Borroto OM, Garcia-Domenech R, De Julian-Ortiz JV, Montero A, et al. *J Comput Aided Mol Des* 2008;22(8):523–40.
- [101] Marrero-Ponce Y, Ortega-Broche SE, Diaz YE, Alvarado YJ, Cubillan N, Cardoso GC, et al. *J Theor Biol* 2009;259(2):229–41.
- [102] Marrero-Ponce Y, Castillo Garit JA, Nodarse D. *Bioorg Med Chem* 2005;13(10):3397–404.
- [103] Marrero-Ponce Y. *J Chem Inf Comput Sci* 2004;44(6):2010–26.
- [104] Fernandez M, Caballero J, Tundidor-Camba A. *Bioorg Med Chem* 2006;14(12):4137–50.
- [105] Rabow AA, Scheraga HA. *J Mol Biol* 1993;232(4):1157–68.
- [106] Hill T, Lewicki P. *Statistics methods and applications*. Tulsa: StatSoft; 2006.
- [107] Xu T, Du L, Zhou Y. *BMC Bioinformatics* 2008;9:472.
- [108] Mahdavi MA, Lin YH. *Genomics Proteomics Bioinformatics* 2007;5(3–4):177–86.
- [109] Feldesman MR. *Am J Phys Anthropol* 2002;119(3):257–75.
- [110] Schlessinger A, Yachdav G, Rost B. *Bioinformatics* 2006;22(7):891–3.
- [111] Mewes HW, Frishman D, Mayer KF, Munsterkötter M, Noubibou O, Pagel P, et al. *Nucleic Acids Res* 2006;34(Database issue):D169–172.
- [112] Xie D, Li A, Wang M, Fan Z, Feng H. *Nucleic Acids Res* 2005;33(Web Server issue):W105–110.
- [113] McDermott J, Guerin M, Frazier Z, Chang AN, Samudrala R. *Nucleic Acids Res* 2005;33(Web Server issue):W324–5.
- [114] Vedadi M, Lew J, Artz J, Amani M, Zhao Y, Dong A, et al. *Mol Biochem Parasitol* 2007;151(1):100–10.
- [115] Hogg T, Nagarajan K, Herzberg S, Chen L, Shen X, Jiang H, et al. *J Biol Chem* 2006;281(35):25425–37.
- [116] Banerjee AK, Arora N, Murty US. *J Vector Borne Dis* 2009;46(3):171–83.