Molecular Informatics. 2014; 33(4): 276-285

LECTINPred: web Server that uses complex networks of protein structure for prediction of lectins with potential use as cancer biomarkers or in parasite vaccine design

Cristian R. Munteanu,[a] Nieves Pedreira,[a] Julián Dorado,[a] Alejandro Pazos,[a] Lázaro G. Pérez-Montoto,[b] Florencio M. Ubeira,[b] and Humberto González-Díaz[c,d]

[a] *Department of Information and Communication Technologies, Computer Science Faculty, University of A Coruña, Campus de Elviña s/n, 15071 A Coruña, Spain.*
[b] *Department of Microbiology & Parasitology, Faculty of Pharmacy, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain*
[c] *Department of Organic Chemistry II, Faculty of Science and Technology, University of the Basque Country, UPV/EHU, 48940, Leioa, Bizkaia, Spain*
[d] *IKERBASQUE, Basque Foundation for Science, 48011, Bilbao, Spain*

Abstract

Lectins (Ls) play an important role in many diseases such as different types of cancer, parasitic infections and other diseases. Interestingly, the Protein Data Bank (PDB) contains +3000 protein 3D structures with unknown function. Thus, we can in principle, discover new Ls mining non-annotated structures from PDB or other sources. However, there are no general models to predict new biologically relevant Ls based on 3D chemical structures. We used the MARCH-INSIDE software to calculate the Markov-Shannon 3D electrostatic entropy parameters for the complex networks of protein structure of 2200 different protein 3D structures, including 1200 Ls. We have performed a Linear Discriminant Analysis (LDA) using these parameters as inputs in order to seek a new Quantitative Structure-Activity Relationship (QSAR) model, which is able to discriminate 3D structure of Ls from other proteins. We implemented this predictor in the web server named LECTINPred, freely available at http://bio-aims.udc.es/LECTINPred.php. This web server showed the following goodness-of-fit statistics: Sensitivity=96.7 % (for Ls), Specificity=87.6 % (non-active proteins), and Accuracy=92.5 % (for all proteins), considering altogether both the training and external prediction series. In mode 2, users can carry out an automatic retrieval of protein structures from PDB. We illustrated the use of this server, in operation mode 1, performing a data mining of PDB. We predicted Ls scores for +2000 proteins with unknown function and selected the top-scored ones as possible lectins. In operation mode 2, LECTINPred can also upload 3D structural models generated with structure-prediction tools like LOMETS or PHYRE2. The new Ls are expected to be of relevance as cancer biomarkers or useful in parasite vaccine design.

Keywords

Lectins; Cancer biomarkers; Parasite vaccine design; Complex networks; QSAR models; Web server; Markov chains; Entropy

1. Introduction

Lectins (Ls) are proteins involved in very important biological processes. Innate immunity is the earliest response to invading microbes/parasites and it acts to contain infection in the first minutes to hours of challenge. Unlike adaptive immunity that relies upon clonal expansion of cells that emerge days after antigenic challenge, the innate immune response is immediate. In fact, C-type Ls receptors (CLRs) have long been known as pattern-recognition receptors involved in the recognition of pathogens by the innate immune system.1 However, evidence is accumulating that many CLRs are also able to recognize endogenous self-ligands and that this recognition event often plays an important role in immune homeostasis.2 Geijtenbeek and Gringhuis3 reviewed the expression of CLRs by dendritic cells as a crucial step for tailoring immune responses to pathogens. Recently, Osanya et al.[4] have shown that *Leishmania sp*. lipophosphoglycan (LPG), a pathogen glycolipid, modulated essential interactions with host phagocytic cells, promoting immune-modulation. LPG altered the production of pro-inflammatory cytokines via a toll-like receptor (TLR2)-mediated mechanism in vitro and in vivo, in mice infected with *Leishmania major.* The type of inflammation and rate of bead entry into macrophages and dendritic cells were different compared with control, indicating selective Ls receptor/oligosaccharide interactions, mediating cell entry and cytokine production. In the author's own words: "these novel findings may prompt the development of targeted oligosaccharide adjuvants against chronic infections". On the other hand, the Ls called Galectins have been recognized as important mediators of immune homeostasis and disease regulation, but comparatively little is known about their role in parasite infection. In this sense, Young et al.5 have published this year the following work: *Galectin secretion and binding to adult Fasciola hepatica during chronic liver fluke infection of sheep.* This work showed that both galectin-14 and galectin-11 directly interact with the parasite in the bile ducts after being released in liver tissue during chronic liver fluke infection. Galectin-11 may also be involved in epithelial cell turnover and cancerogenesis. In addition, fluorescently-labeled Ls have been used to uncover hidden antigens in parasites like Fasciola in vaccine design preliminary efforts.6 Soluble mediators, including complement components and the Mannose-Binding Lectin (MBL) make an important contribution to innate immune protection and work along with epithelial barriers, cellular defenses such as phagocytosis, and pattern-recognition receptors that trigger pro-inflammatory signaling cascades. Ip and Takahashi et al.7 discussed the protection provided by this complex defense system and the particular contribution of soluble mediators such as MBL. Several prediction models of the tertiary structure of the lectins have been presented by Chou and Geng's teams.8

Epidemiological studies have suggested that the genetically determined variations in MBL serum concentrations influence the susceptibility to and the course of different types of diseases.9 For instance, liver cancer is the predominant cause of cancer mortality in males of Southern China and Taiwan. The current therapy is not satisfactory, and more effective treatments are needed. In search for new therapies for liver tumor, Lei and Chang10 have found that Concanavalin A (Con-A), a lectin from Jack bean seeds, can have a potent anti-hepatoma effect. The prototype of Con-A with an anti-hepatoma activity supports the search for other natural Ls as anti-cancer compounds. However, other important uses may be expected for new lectin-like proteins. For instance, lectin binding as a biological test in vitro for the prediction of functional activity of human spermatozoa11 or a study of pathologic processes using phytoLs,12 Winn et al.13 have demonstrated that sialic acid binding Ig-like lectin 6 (Siglec-6) together with other proteins could have clinical utility for predicting and/or diagnosing Preeclampsia (PE). PE affects 4–8 % of human pregnancies causing significant maternal and neonatal morbidity and mortality. Kamoto and Satomura et al.14 assessed the usefulness of lectin-reactive alpha-fetoprotein (AFP-L3 %) in hepatocellular carcinomas. Measurement of soluble Lectin-like oxidized LDL receptor (LOX)-1 in vivo may provide a novel diagnostic tool for the evaluation and prediction of atherosclerosis and vascular disease.15 Noguchi and Thomas et al.16 carried out a further analysis of predictive value of *Helix pomatia* lectin binding to primary breast cancer for axillary and internal mammary lymph node metastases.

All these facts have determined that the discovery of new Ls or Ls-like proteins and/or Ls targeted by drug lead compounds has become a goal of major interest. In any case, the number of potential protein candidates to be investigated with experimental procedures in order to possibly annotate them as Ls is very high. The same happens with a very large number of potential compounds to be studied as inhibitors of different Ls. In this sense, the development of new web servers may be very useful to predict different types of Ls based on different Machine Learning techniques. For instance, a web server for C-type Ls prediction based on Support Vector Machine (SVM) algorithms has been published on the web (http://software.iiar.res.in/svm/ctypelectin index.html). Other web server has been developed to predict Galectins, also based on SVM (http://software.iiar.res.in/svm/galectin index.html).

Remarkably, Kumar et al. have developed another SVM-based web server called CancerPred for the prediction of Cancer Ls (http://www.imtech.res.in/raghava/cancer_pred/index.html).17 In our opinion, Quantitative Structure-Activity Relationship (QSAR)18 and other computational methods may be very useful for predicting new Ls proteins and/or compounds having Ls as protein targets. Sirois, Giguere, and Roy19 presented the first QSAR model for Galectin-3 glycomimetic inhibitors based on docked structures to the carbohydrate recognition domain (CRD). We can use many physicochemical parameters to characterize proteins in these studies such as charges or hydrophilicity parameters20 or numerical parameters derived from a graph or network representation of the molecular systems (including but not limited to protein structure, as in this case).21 However, there is no report of a general QSAR model for the prediction of Ls-like proteins. As demonstrated by a series of recent publications,22 to establish a really useful statistical predictor for a protein system, we need to consider the following procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the protein or biological samples with an effective mathematical expression or model that can truly reflect their intrinsic correlation with the attribute to be predicted; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (v) establish a user-friendly web-server for the predictor that is accessible to the public. Below, let us describe how to deal with these steps one by one.

González-Díaz et al. introduced a new QSAR method that could be useful to solve the Ls prediction problem. The method was called **MAR**kovian **CH**emicals **IN SI**lico **DE**sign (**MARCH-INSIDE 1.0**) for the computational design of small-sized drugs. The approach uses a Markov Chain model (MCM) of the intra-molecular movement of electrons to calculate structural parameters of drugs. In successive studies, we have extended this method to perform fast calculation of 2D and 3D alignment—free structural parameters based on molecular vibrations in RNA secondary structures, or electrostatic potential, and Van der Waals interactions in proteins. Recently, the method has been renamed as **MAR**kov **CH**ains **I**nvariants for **N**etworks **SI**mulation & **DE**sign (MARCH-INSIDE 2.0). This explores more adequately the broad uses of the method that describes the structure of drugs,23 RNA,24 and proteins,25,26 as well as drug-drug networks,27 and drug-protein interactions.28 The MARCH-INSIDE method may also be used to study PPIs, bacteria-bacteria co-aggregation, parasite-host interactions and other systems with an MCM associated to a network. In very recent reviews, we have discussed the last applications of this method.29–31 We should also make reference to the recent implementation carried out by Munteanu and González-Díaz of the Internet portal called Bio-AIMS, freely available for the use of the international research community. This portal includes the web-server packages TargetPred (http://bio-aims.udc.es/TargetPred.php) with new Protein-QSAR servers based on MARCH-INSIDE. One of the servers is ATCUNPred,32 useful for predicting ATCUN-mediated DNA-cleavage anticancer proteins. The second server is EnzClassPred,33 which implements one of the MARCH-INSIDE-based QSAR models for the prediction of enzyme function.34 Two additional servers based on MARCH-INSIDE are: Trypano-PPI35 and Plasmod-PPI.36 These are the first servers that predict self protein-protein complexes in *Trypanosome* sp. or *Plasmodium sp.* Proteomes, opening new opportunities for anti-trypanosome or anti-malarial drug target discovery.

For all these reasons, we decided to develop herein the first 3D-QSAR method useful to discriminate between Ls and non-Ls proteins (nLs) using MARCH-INSIDE 2.0. To this end, firstly we have calculated different local and global parameters of the 3D residue contact networks of large series of Ls and nLs proteins (see Figure 1). The 3D network parameters calculated are of three different classes: average electrostatic potentials ξ_k(R), together with spectral moments of π_k(R) and entropy measures θ_k(R) of the electrostatic field of amino acids placed at distance *k* from each other within different regions R of the protein 3D structure. Next, we have carried out a statistical analysis in order to seek a linear equation (3D-QSAR model) that links the 3D electrostatic parameters of the protein structural network with S(Ls) values. The S(Ls) output is a real-valued variable that scores the propensity of a protein to act as a Ls. In addition, we have implemented the model in a public web server for the prediction of these proteins called LECTINPred. Last, we have illustrated the use of this web server to carry out online data mining of the PDB. We have predicted S(Ls) values for +1000 proteins. This type of study may help us to discover new Ls useful as human cancer biomarkers of drug targets.

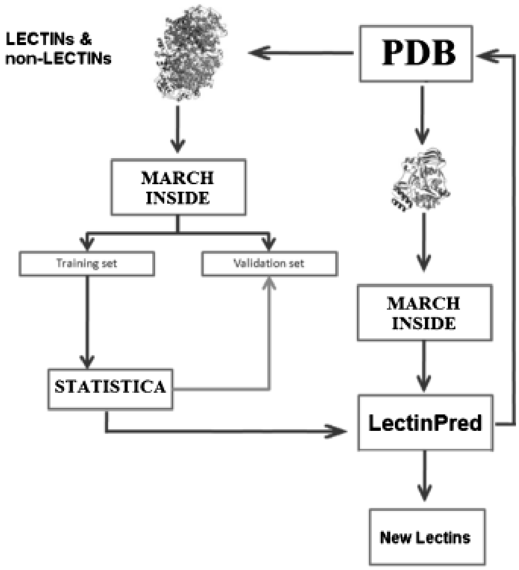


Figure 1. Flowchart for all the steps necessary to construct the classifiers and server.

2. Computational Methods

2.1. MARCH-INSIDE Method

In this work, the information about the molecular structure of the proteins is codified by using the MCM method with the $^{1}\mathbf{\Pi}$ matrix (the short-term electrostatic interaction matrix). The matrix $^{1}\mathbf{\Pi}$ is constructed as a squared matrix ($n\times n$), where n is the number of amino acids (*aa*) in the protein.37,38 In previous works we have predicted protein function based on 3D-Potentials for different type of interactions or molecular fields derived from $^{1}\mathbf{\Pi}$. The main types of the used molecular fields are: E (electronegativity), vdW (Van der Waals), and HINT (hydrophobic interaction field) potentials.26,38,39 In this paper, we have calculated π_k(R) and θ_k(R) values only for E and HINT potentials. We have omitted the vdW term due to a simple reason; the HINT potential includes a vdW component. The values have been used here as inputs to construct the QSAR model. The detailed explanation has been published before. As follows, we give the formula for π_k(R), θ_k(R) and ξ_k(R) and some general explanations:(1), (2), (3)

$$\xi_k(\mathrm{R}) = -\sum_{j\in \mathrm{R}} {}^{k}p_j(\mathrm{R})\cdot \xi_0(j) \qquad (1)$$

$$\theta_k(\mathrm{R}) = -\sum_{j\in \mathrm{R}}^{n} {}^{k}p_j(\mathrm{R})\cdot \log\left[{}^{k}p_j(\mathrm{R})\right] \qquad (2)$$

$$\pi_k(\mathrm{R}) = \sum_{i=j\in \mathrm{R}}^{n} {}^{k}p_{ij}(\mathrm{R}) \qquad (3)$$

It is remarkable that the spectral moments depend on the probability ${}^{k}p_{ij}(\mathrm{R})$ with which the effect of the interaction f propagates from amino acid i^{th} to other neighboring amino acids j^{th} and returns to i^{th} after k-steps. On the other hand, both the average electrostatic potential and the entropy measures depend on the absolute probabilities ${}^{k}p_{j}(\mathrm{R})$ with which the amino acid j^{th} has an interaction of type f with the rest of amino acids. In any case, both probabilities refer to a first (k=1) direct interaction of type f between 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 a 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* (α_{ij}=1 if $r_{ij}<r_{\text{cut-off}}$). Otherwise, the interaction is banished (α_{ij}=0). The relationship α_{ij} may be visualized in the form of a protein structure complex network (see Figure 2). In this network, the nodes are the C_α atoms of the amino acids and the edges connect pairs of amino acids with α_{ij}=1. Euclidean 3D space $\mathbf{r}_3=(x, \mathrm{y}, \mathrm{z})$ coordinates of the C_α atoms of amino acids listed on protein PDB files. For calculation all water molecules and metal ions were removed.30 All calculations were carried out with our in-house MARCH-INSIDE 2.0 software.30 Thus, the study used the default value of 7 Å for the $\mathrm{r}_{\text{cut-off}}$. The MARCH-INSIDE software uses the full matrix to calculate the natural powers of the matrix but it can sum the values of all nodes (amino acids) to calculate total indices or only for some nodes in order to calculate local protein descriptors. It means that MARCH-INSIDE can calculate total values or local values that correspond to the selected amino acids placed in specific spatial regions or orbits (R). These regions are often defined in geometric terms and are called core, inner, middle or surface region. The protein is virtually divided into the following regions: *c* corresponds to core, *i* to inner, *m* to middle, and *s* to surface regions, respectively. The diameters of the regions are 0 to 25 for region *c*, 25 to 50 for region *i*, 50 to 75 for region *m*, and 75 to 100 for region *s*. These values are given in terms of percentage of the longest distance $r_{\max}$ with respect to the centre of charge. Additionally, we consider the total region (*t*) that contains all the amino acids in the protein (region diameter 0 to 100 % of $r_{\max}$). Consequently, we can calculate different parameters ($\pi_k(\mathrm{R})$, $\xi_k(\mathrm{R})$, and $\theta_k(\mathrm{R})$) for the amino acids contained in a region (*c*, *i*, *m*, *s*, or *t*) and placed at a topological distance *k* within this region (*k* is the name of the order).26,40 In this work, we calculated a total of 90 indices (3 types of indices×5 types of regions×6 higher order considered) for each protein.

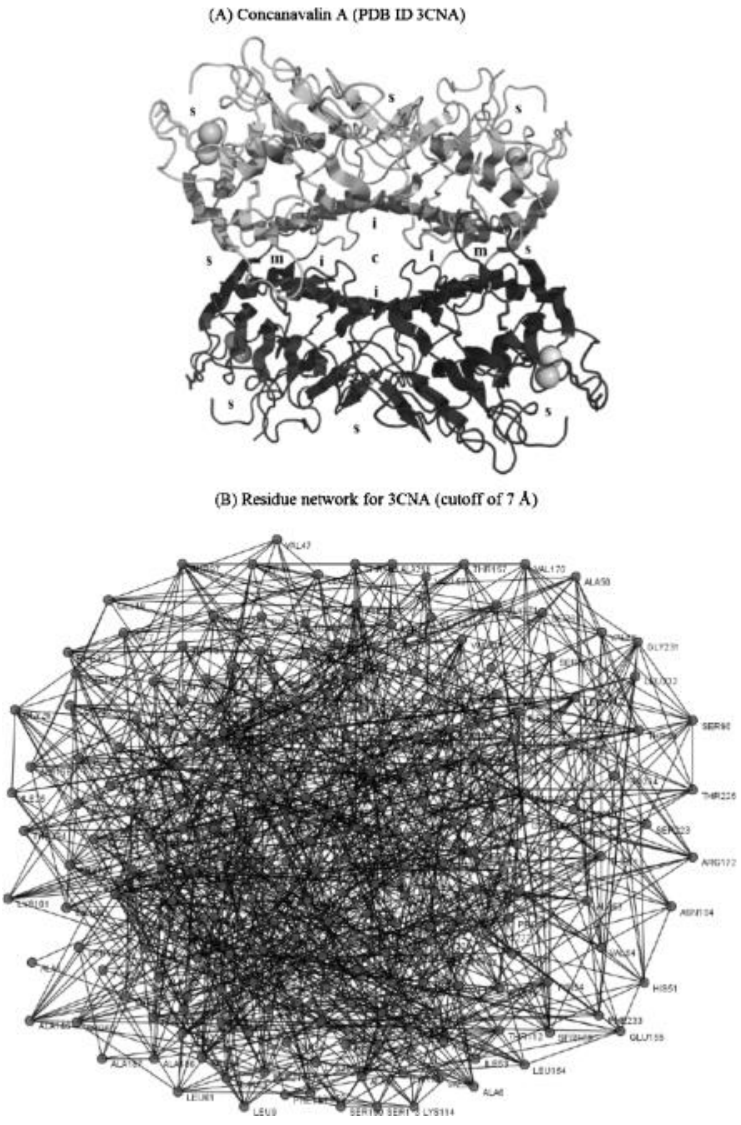


Figure 2. Representations of a lectin with PDBID 3CNA (concanavalin): (A) 3D structure model for a full complex and (B) complex network graph.

2.2. LDA Model

The Linear Discriminant Analysis **(**LDA) is frequently used for classification/prediction problems in physical anthropology, but it is unusual to find examples in which researchers consider the statistical limitations and assumptions required for this technique. In this work, all LDA models have been trained with the STATISTICA 6.0 software, for which our laboratory holds rights of use.41 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(LIBP)=p(nLIBP)=0.5. The LDA discriminant equation was obtained using as input the three types of Markov chain invariants θ_k(R). The general form of the equation obtained by LDA is:(4)

$$\mathrm{S(Ls)} = \sum_{\mathrm{R},k,ti}^{5,5,3} a_{\mathrm{R},k} \cdot \xi_k(\mathrm{R}) + b_{\mathrm{R},k} \cdot \theta_k(\mathrm{R}) + c_{\mathrm{R},k} \cdot \pi_k(\mathrm{R}) \ + d \qquad (4)$$

S(Ls) is the above-mentioned output of the model. It is a real-valued variable that scores the propensity of a protein to act as Ls. The χ^2 and p-level value were examined in order to test the statistical significance of the model. The Accuracy, Specificity and Sensitivity were used to quantify the goodness-of-fit and the discriminatory power of the model. Different authors have applied this type of LDA model using different classes of input variables to construct QSAR models for drugs,42,43 proteins or nucleic acids.43,44 In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling or K-fold (such as 5-fold and 10-fold) cross-validation test, and jackknife test.45 However, as elucidated in Chou46 and demonstrated by Equations 28–32 therein, among the three cross-validation methods, the jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark dataset, and hence has been increasingly used and widely recognized by investigators to examine the accuracy of various predictors.47 However, in this study we used the independent dataset to test our model in order to reduce the computational time as many investigators did with SVM as prediction engine.

2.3. Dataset

The protein structures were downloaded from PDB48 using the following schemes for PDB-database search: (i) introducing as input parameter the text sugar-binding protein in the search item, called function for positive cases. Scheme (ii) was used to get negative cases introducing the PDB IDs for all the proteins contained in the list reported in the article published by Dobson and Doig.49 The positive cases are those proteins with function annotation as Ls in the PDB. The list of negative cases (nLs) from the search scheme (ii) contains enzymes and other proteins present in humans and many other organisms, including other parasites (see Supporting Information 1). The nLs proteins have known functions different from Ls. The dataset was made up of 2200 proteins (1200 Ls and 1000 nLs) from more than 20 organisms, including parasites and human or cattle hosts. Detailed information about the PDB ID, the values of the Markov-Shannon electrostatic entropy indices, the corresponding observed classification, and the predicted classification for each protein are given in the Supporting Information 2. To avoid homology bias and remove the redundant sequences from the benchmark dataset, a cut-off threshold of 25 % is normally recommended46 to exclude those proteins from the benchmark datasets that have equal to or greater than 25 % sequence identity compared to any other as done in the literature.50 However, in this study we have not used such a stringent criterion because the currently available data do not allow us to do so. Otherwise, the number of proteins for some subsets would be too low to have statistical significance.

3. Results and Discussion

3.1. Alignment-Free LDA Model for Ls

One of our previous works has reported the applicability of the LDA in protein and/or drug QSAR studies.51 The best QSAR LDA model in this study is described by Equation 5 and was obtained with the *Forward stepwise* method from STATISTICA:41(5)

$$\begin{aligned} S(\mathrm{Ls}) = {} & 107.367\theta_1(m) - 111.637\theta_5(m) + 345.891\theta_1(s) \\ & - 745.918\theta_3(s) + 399.047\theta_5(s) + 3.603 \qquad (5) \\ \mathrm{N} = 1650 \quad & R_c = 0.81 \quad \chi^2 = 1811.152 \quad p < 0.001 \end{aligned}$$

The statistical parameters of the model are: Canonical Regression Coefficient (R_c), Chi-square (χ^2) and model significance level (p-level).52 N represents only the number of proteins used to train the model. We split the dataset at random in a training series (75 %), used for model construction; and a prediction one (25 %) used for model validation. The high $R_c > 0.8$ indicates a strong linear correlation between input and output. The value of p-level <0.05 for the Chi-square test indicates a statistically significant discrimination between the two groups of proteins. In addition, the model has shown good Accuracy, Specificity, and Sensitivity values in both training series and external validation series. The classification matrices for the training, validation and both (training + validation together) series are presented in Table 1 (full dataset). The PDB ID, θ_k, and S(Ls) values for all proteins used to train or validate (cv) the model are given in the Supporting Information 2 (available online). This result confirms a statistically significant relationship between MARCH-INSIDE parameters and Ls activity. Taking into consideration that this classifier is a simpler linear equation with only nine input parameters we can conclude that this may become a very useful model.

In order to test the possible biases generated by the unbalanced full dataset composition (1000 nLs vs. 1200 Ls), two additional datasets have been tested: 1 : 1 (1000 nLs vs. 1000 Ls) and 5 : 1 (1000 nLs vs. 200 Ls). These two datasets have been used to find the same type of classification (LDA), with the same five topological indices (see Equation 5) and the same training/validation split (75 % vs. 25 %). The dataset 1 : 1 has been used to test the variation of the prediction when the dataset is balanced. The dataset 5 : 1 represents an extreme test where the number of Ls cases is five times lower than the number of nLs cases (closer to the reality). Therefore, these two new datasets have been obtained by randomly removing Ls cases. Table 1 demonstrates that the total accuracy prediction (Both, Total Accuracy) of Ls proteins with an LDA model based on five TIs is lower with only 0.6 % for the balanced dataset (1 : 1) and it is lower with 2.6 % for the close-to-reality composition dataset (5 : 1, nLs : Ls).

Table 1. Results of the 3D-QSAR study of LIBPs with LDA.

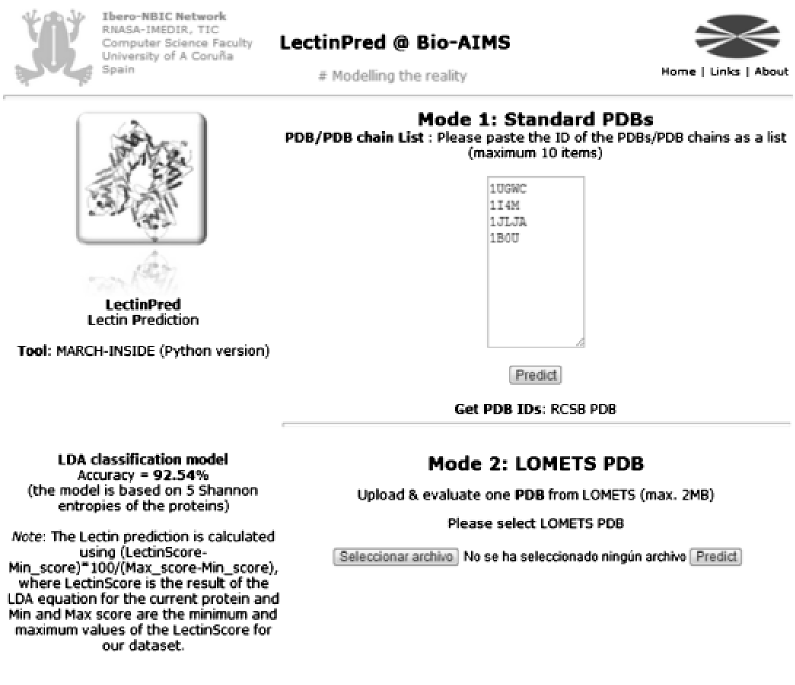
Data Sub-set	Group	Parameter	%	nLs	Ls
1000 nLs vs. 1200 Ls (Full dataset)					
Training	nLs	Specificity	88.3	662	88
	Ls	Sensitivity	96.9	28	872
	Total	Accuracy	93.0		
Validation	nLs	Specificity	85.6	214	36
	Ls	Sensitivity	96.0	12	288
	Total	Accuracy	91.3		
Both	nLs	Specificity	87.6	876	124
	Ls	Sensitivity	96.7	40	1160
	Total	Accuracy	92.5		
1000 nLs vs. 1000 Ls (1 : 1 dataset)					
Training	nLs	Specificity	88.1	661	89
	Ls	Sensitivity	96.3	28	722
	Total	Accuracy	92.2		
Validation	nLs	Specificity	86.0	215	35
	Ls	Sensitivity	96.0	10	240
	Total	Accuracy	91.0		
Both	nLs	Specificity	87.6	876	124
	Ls	Sensitivity	96.2	38	962
	Total	Accuracy	91.9		
1000 nLs vs. 200 Ls (5 : 1 dataset)					
Training	nLs	Specificity	89.9	674	76
	Ls	Sensitivity	90.0	15	135
	Total	Accuracy	89.9		
Validation	nLs	Specificity	89.2	223	27
	Ls	Sensitivity	94.0	3	47
	Total	Accuracy	90.0		
Both	nLs	Specificity	89.7	897	103
	Ls	Sensitivity	91.0	18	182
	Total	Accuracy	89.9		

3.2. *LECTINPred web Server*

At the present time, looking for a fast and accurate predictive model is not enough; it should also be implemented into public servers available online to the scientific community. The server packages developed by Chou and Shen that predict the function of proteins from structural parameters or explore protein structures53 are good examples in this sense. These may be used by proteome research scientists by interacting with user-friendly interfaces. It means that the user does not need to be an expert on the theoretical details behind this kind of models, including the vast literature published by Chou et al. on the development of models with pseudo-amino acid composition parameters or the use of Machine Learning classification techniques (Artificial Neural Networks, Hidden Markov Models, and Support Vector Machines.54 PseAA composition of a protein represents a set of discrete numbers that is derived from its amino acid sequence and that is different from the classical amino acid composition. Thus, PseAA is able to encode sequence order or pattern information. However, to the best of our knowledge, there is no QSAR-based server for the prediction of Ls. In this sense, we have implemented the best LDA model found here at the web portal Bio-AIMS as an online server called LECTINPred. The acronym **LECTINPred** comes from the words **LECTIN**s and **P**redictor. LECTINPred is located at: http://bio-aims.udc.es/LECTINPred.php. This online tool is based on PHP/HTML and Python routines coupled to nested MARCH-INSIDE classic algorithm to calculate input molecular structure parameters.29

3.2.1. *LECTINPred Mode 1*

Figure 3 depicts the user interface for LECTINPred including mode 1 (top of the web page). The user only has to paste the PDB ID of the query proteins with unknown functions. With these PDB ID codes, LECTINPred automatically connects to the PDB database, uploads the PDB files with the 3D structure of the protein, constructs the Markov matrix of electrostatic interactions and calculates the total and regional (R) average electrostatic entropy values θ_k(R) for each query protein.



Ibero-NBIC Network
RNASA-IMEDIR, TIC
Computer Science Faculty
University of A Coruña
Spain

LectinPred @ Bio-AIMS

Modelling the reality

Mode 1: Standard PDBs

PDB/PDB chain List : Please paste the ID of the PDBs/PDB chains as a list (maximum 10 items)

1UGWC
1I4M
1JLJA
1B0U

Get PDB IDs: RCSB PDB

LectinPred
Lectin Prediction

Tool: MARCH-INSIDE (Python version)

LDA classification model
Accuracy = 92.54%
(the model is based on 5 Shannon entropies of the proteins)

Note: The Lectin prediction is calculated using (LectinScore-Min_score)*100/(Max_score-Min_score), where LectinScore is the result of the LDA equation for the current protein and Min and Max score are the minimum and maximum values of the LectinScore for our dataset.

Mode 2: LOMETS PDB

Upload & evaluate one PDB from LOMETS (max. 2MB)

Please select LOMETS PDB

No se ha seleccionado ningún archivo

Figure 3. Web-user interface of LECTINPred online tool (http://bio-aims.udc.es/LECTINPred.php).

3.2.2. Mining PDB with LECTINPred (Example of use of Mode 1)

Both the existence in PDB of +3000 proteins with unknown function and the interest on the discovery of new Ls (potential drug targets in parasite infections or cancer biomarkers) prompt us to carry out a data-mining search of new Ls candidates in PDB. For this study, we have implemented the key function PDB mining in the new server LECTINPred. By clicking this key the server performs automatic search of all PDB files with unknown function at a reference date. After that, LECTINPred extracts all C_α coordinates from these files and calculates the necessary π_k(R) values for all these proteins. Last, the server uses these values as inputs of the best model found and predicts the S(Ls) values for all these proteins. The proteins with the highest scores may be selected as candidates for experimental assays in order to confirm the Ls function. Each time we use the PDB mining key, the server updates the prediction for all new PDB files present in the last version of the PDB synchronized with LECTINPred. We have predicted S(Ls) values for a total of 2693 proteins selected to have an unknown function (or only a hypothetical function predicted) and low sequence homology in current PDB release. Only 59 out of a total of 1368 proteins studied (released to PDB with unknown function) were predicted as possible Ls with S(Ls) >70 %, from those only 8 presented as S(Ls) >80 %. It means that LECTINPred has a low propensity to predict unknown proteins as Ls (only 0.58 % of the studied proteins have S(Ls) >80 %). In our opinion, this result is coherent with the relative low abundance of Ls in nature with respect to other functions. We depict the higher S(Ls) values found in the Supporting Information 2. These possible new Ls, if confirmed, may be useful as cancer biomarkers or in parasite vaccine design.

3.2.3. LECTINPred Mode 2

There are other potential uses of this server. How should one predict S(Ls) values for proteins with known sequence but unknown 3D structure and function that have not been released to PDB? Mode 2 is essentially the same as mode 1, but the server prompts the users to upload .ent and .pdb files with 3D structures of proteins generated by using LOMETS web server55 developed by Prof. Zhang et al. at Michigan University. Figure 3 depicts the user interface for LECTINPred mode 2 (bottom of the web page). LOMETS is a local threading meta-server, for quick and automated predictions of protein tertiary structures and spatial constraints. The LOMETS server is freely available to the academic community at http://zhang.bioinformatics.ku.edu/LOMETS. The mode 2 can also upload pdb files of protein 3D structural models generated with other structure prediction web servers like PHYRE256 (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index). After generating PDB files with LOMETS, we can upload them to LECTINPred. This is the same strategy used to develop the mode 2 of the web server MIND-BEST, in order to predict drug-target interactions between drugs and proteins with unknown 3D structure.57 Anyhow, we have to be aware that using this input mode 2 we can predict S(Ls) values using 3D structural models generated only by modeling. Consequently, predictions derived with input mode 2 have to be used with more caution than predictions obtained with input mode 1.

4. Conclusions

The discovery of new Ls is a goal of great importance and several authors have presented interesting results. The present work has demonstrated that there is a strong linear relationship between 3D electrostatic field entropies calculated with the MARCH-INSIDE approach and the propensity of proteins to act as Ls. Consequently, using these parameters we can seek a linear QSAR useful to predict Ls. In addition, as user-friendly and publicly accessible web-servers represent the future direction of developing practically more useful predictors,58 we have introduced the web-server LECTINPred at http://bio-aims.udc.es/LECTINPred.php. The online implementation of this model in the web server LECTINPred allows public researchers around the world to predict online new Ls at no cost. LECTINPred may be used to mine the PDB. We have demonstrated the PDB mining option performing a predictive study of +1000 proteins with unknown function looking for new Ls. The user can also predict Ls scores for new protein sequences if uploads to LECTINPred the 3D models of proteins with unknown structure generated with well-known servers as in the case of LOMETS. The new Ls are expected to be of relevance as cancer biomarkers or useful in parasite vaccine design.

Acknowledgements

This work was financially supported by the Research Projects AGL2010-22290-C03-01 and AGL2011-30563-C03-01 from the *Ministry of Science and Innovation*, Spain. Munteanu CR thanks the Isidro Parga Pondal Research Program funded by *Xunta de Galicia* and the *European Social Funds* (*ESF*) for partial financial support.

References

1. L. M. van den Berg, S. I. Gringhuis, T. B. Geijtenbeek, Ann. N. Y. Acad. Sci. 2012, 1253, 149–158.
2. J. J. Garcia-Vallejo, Y. van Kooyk, Immunol. Rev. 2009, 230, 22–37.
3. T. B. Geijtenbeek, S. I. Gringhuis, Nat. Rev. Immunol. 2009, 9, 465–479.
4. A. Osanya, E. H. Song, K. Metz, R. M. Shimak, P. M. Boggiatto, E. Huffman, C. Johnson, J. M. Hostetter, N. L. Pohl, C. A. Petersen, in Am. J. Pathol., Vol. 179, American Society for Investigative Pathology, Elsevier, New York, 2011, pp. 1329–1337.
5. A. R. Young, G. J. Barcham, H. E. McWilliam, D. M. Piedrafita, E. N. Meeusen, in Vet. Immunol. Immunopathol., Elsevier, Amsterdam, 2012, pp. 362–367.
6. H. C. McAllister, A. J. Nisbet, P. J. Skuce, D. P. Knox, in J. Helminthol., 2011, pp. 1–7.
7. W. K. Ip, K. Takahashi, R. A. Ezekowitz, L. M. Stuart, Immunol. Rev. 2009, 230, 9–21.
8. a) J. G. Geng, M. Chen, K. C. Chou, Curr. Med. Chem. 2004, 11, 2153–2160; b) K. C. Chou, FEBS Lett. 1995, 363, 123–126; c) K. C. Chou, J. Protein Chem. 1996, 15, 161–168; d) K. C. Chou, R. L. Heinrikson, J. Protein Chem. 1997, 16, 765–773.
9. P. Garred, C. Honore, Y. J. Ma, L. Munthe-Fog, T. Hummelshoj, Mol. Immunol. 2009, 46, 2737–2744.
10. H. Y. Lei, C. P. Chang, J. Biomed. Sci. 2009, 16, 10.
11. I. Mladenovic, L. Hajdukovic, O. Genbacev, M. Cuperlovic, M. Movsesijan, Hum. Reprod. 1993, 8, 258–265.
12. M. Vasil'chuk Iu, N. F. Pogorelaia, N. S. Tkachenko, Klin. Khir. 1992, 9–11.
13. V. D. Winn, M. Gormley, A. C. Paquet, K. Kjaer-Sorensen, A. Kramer, K. K. Rumer, R. Haimov-Kochman, R. F. Yeh, M. T. Overgaard, A. Varki, C. Oxvig, S. J. Fisher, Endocrinology 2009, 150, 452–462.
14. T. Kamoto, S. Satomura, T. Yoshiki, Y. Okada, F. Henmi, H. Nishiyama, T. Kobayashi, A. Terai, T. Habuchi, O. Ogawa, Jpn. J. Clin. Oncol. 2002, 32, 472–476.
15. N. Kume, T. Kita, Curr. Opin. Lipidol. 2001, 12, 419–423.
16. M. Noguchi, M. Thomas, H. Kitagawa, K. Kinoshita, N. Ohta, M. Nagamori, I. Miyazaki, Br. J. Cancer 1993, 67, 1368–1371.
17. R. Kumar, B. Panwar, J. S. Chauhan, G. P. Raghava, in BMC Res. Notes, Vol. 4, England, 2011, p. 237.
18. a) J. Devillers, A. T. Balaban, Topological Indices and Related Descriptors in QSAR and QSPR, Gordon and Breach, The Netherlands, 1999; b) Q. S. Du, R. B. Huang, K. C. Chou, Curr. Protein Pept. Sci. 2008, 9, 248–260; c) Q. S. Du, R. B. Huang, Y. T. Wei, L. Q. Du, K. C. Chou, J. Comput. Chem. 2008, 29, 211–219; d) Q. S. Du, R. B. Huang, Y. T. Wei, Z. W. Pang, L. Q. Du, K. C. Chou, J. Comput. Chem. 2009, 30, 295–304; e) Q. Du, P. G. Mezey, K. C. Chou, J. Comput. Chem. 2005, 26, 461–470;

18f. H. Wei, C. H. Wang, Q. S. Du, J. Meng, K. C. Chou, Med. Chem. 2009, 5, 305–317.

19. S. Sirois, D. Giguere, R. Roy, Med. Chem. (Shariqah, United Arab Emirates) 2006, 2, 481–489.
20. a) J. P. Zbilut, A. Giuliani, A. Colosimo, J. C. Mitchell, M. Colafranceschi, N. Marwan, C. L. Webber, Jr., V. N. Uversky, J. proteome Res. 2004, 3, 1243–1253; b) B. Shen, J. Bai, M. Vihinen, Prot. Eng. Des. Sel. 2008, 21, 37–44.
21. F. Torrens, G. Castellano, Curr. Proteomics 2009, 204–213.
22. a) W. Chen, P. M. Feng, H. Lin, K. C. Chou, Nucleic Acids Res. 2013, 41, e68; b) W. Chen, H. Lin, P. M. Feng, C. Ding, Y. C. Zuo, K. C. Chou, PLoS ONE 2012, 7, e47843; c) K. C. Chou, Z. C. Wu, X. Xiao, Mol. BioSyst. 2012, 8, 629–641; d) Y. Xu, J. Ding, L. Y. Wu, K. C. Chou, PLoS ONE 2013, 8, e55844; e) X. Xiao, P. Wang, W. Z. Lin, J. H. Jia, K. C. Chou, Anal. Biochem. 2013, 436, 168–177; f) K. C. Chou, Z. C. Wu, X. Xiao, PLoS ONE 2011, 6, e18258;
23. L. Santana, E. Uriarte, H. González-Díaz, G. Zagotto, R. Soto-Otero, E. Mendez-Alvarez, J. Med. Chem. 2006, 49, 1149–1156.
24. H. González-Díaz, R. R. de Armas, R. Molina, Bioinformatics 2003, 19, 2079–2087.
25. a) G. Aguero-Chapin, J. Varona-Santos, G. A. de la Riva, A. Antunes, T. Gonzalez-Villa, E. Uriarte, H. Gonzalez-Diaz, J. Proteome Res. 2009, 8, 2122–2128; b) R. Concu, M. A. Dea-Ayuela, L. G. Perez-Montoto, F. Bolas-Fernandez, F. J. Prado-Prado, G. Podda, E. Uriarte, F. M. Ubeira, H. Gonzalez-Diaz, J. Proteome Res. 2009, 8, 4372–4382.
26. H. González-Díaz, L. Saiz-Urra, R. Molina, L. Santana, E. Uriarte, J. Proteome Res. 2007, 6, 904–908.
27. L. Santana, H. Gonzalez-Diaz, E. Quezada, E. Uriarte, M. Yanez, D. Vina, F. Orallo, J. Med. Chem. 2008, 51, 6740–6751.
28. D. Vina, E. Uriarte, F. Orallo, H. Gonzalez-Diaz, Mol. Pharm. 2009, 6, 825–835.
29. H. Gonzalez-Diaz, F. Prado-Prado, F. M. Ubeira, Curr. Top. Med. Chem. 2008, 8, 1676–1690.
30. H. González-Díaz, Y. González-Díaz, L. Santana, F. M. Ubeira, E. Uriarte, Proteomics 2008, 8, 750–778.
31. a) H. González-Díaz, S. Vilar, L. Santana, E. Uriarte, Curr. Top. Med. Chem. 2007, 7, 1025–1039; b) 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, H. Gonzalez-Diaz, Biochim. Biophys. Acta 2009, 1794, 1784–1794; c). Vilar, H. Gonzalez-Diaz, L. Santana, E. Uriarte, J. Theor. Biol. 2009, 261, 449–458.

32. C. R. Munteanu, J. M. Vazquez, J. Dorado, A. P. Sierra, A. Sanchez-Gonzalez, F. J. Prado-Prado, H. Gonzalez-Diaz, J. Proteome Res. 2009, 8, 5219–5228.
33. 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, H. Gonzalez-Diaz, Biochim. Biophys. Acta 2009, 1794, 1784–1794.
34. R. Concu, M. A. Dea-Ayuela, L. G. Perez-Montoto, F. Bolas-Fernandez, F. J. Prado-Prado, G. Podda, E. Uriarte, F. M. Ubeira, H. Gonzalez-Diaz, J. Proteome Res. 2009, 8, 4372–4382.
35. Y. Rodriguez-Soca, C. R. Munteanu, J. Dorado, A. Pazos, F. J. Prado-Prado, H. Gonzalez-Diaz, J. Proteome Res. 2010, 9, 1182–1190.
36. C. R. Munteanu, Y. Rodriguez-Soca, J. Dorado, J. Rabuñal, A. Pazos, H. González-Díaz, Polymer 2010, 51, 264–273.

37 a) H. González-Díaz, A. Pérez-Bello, E. Uriarte, Polymer 2005, 46, 6461–6473; b) H. González-Díaz, E. Uriarte, R. Ramos de Armas, Bioorg. Med. Chem. 2005, 13, 323–331.

38. L. Saiz-Urra, H. González-Díaz, E. Uriarte, Bioorg. Med. Chem. 2005, 13, 3641–3647.
39. R. Concu, G. Podda, E. Uriarte, H. Gonzalez-Diaz, J. Comput. Chem., J. Comput. Chem. 2009, 30, 1510–1520.

40 a) H. Gonzalez-Diaz, L. Saiz-Urra, R. Molina, Y. Gonzalez-Diaz, A. Sanchez-Gonzalez, J. Comput. Chem. 2007, 28, 1042–1048; b) H. Gonzalez-Diaz, R. Molina, E. Uriarte, FEBS Lett. 2005, 579, 4297–4301; c) R. Concu, G. Podda, E. Uriarte, H. Gonzalez-Diaz, J. Comput. Chem. 2009, 30, 1510–1520; d) H. González-Díaz, Y. Pérez-Castillo, G. Podda, E. Uriarte, J. Comput. Chem. 2007, 28, 1990–1995.

41. STATISTICA (data analysis software system), version 6.0, StatSoft.Inc., www.statsoft.com.Statsoft, Inc., **2002**.

42 a) A. Speck-Planche, M. T. Scotti, V. de Paulo-Emerenciano, Curr. Pharm. Des. 2010, 16, 2656–2665; b) A. Speck-Planche, M. N. D. S. Cordeiro, Current Bioinformatics 2011, 6, 81–93; c) A. Speck-Planche, M. T. Scotti, V. P. Emerenciano, A. García-López, E. Molina-Pérez, E. Uriarte, J. Comput. Chem. 2010, 31, 882–894; d) A. Speck-Planche, M. T. Scotti, A. García-López, V. P. Emerenciano, E. Molina-Pérez, E. Uriarte, Mol. Divers. 2009, 13, 445–458; e) G. M. Casanola-Martin, M. T. Khan, Y. Marrero-Ponce, A. Ather, M. N. Sultankhodzhaev, F. Torrens, Bioorg. Med. Chem. Lett. 2006, 16, 324–330; f) G. M. Casanola-Martin, Y. Marrero-Ponce, M. T. Khan, A. Ather, K. M. Khan, F. Torrens, R. Rotondo, Eur. J. Med. Chem. 2007, 42, 1370–1381; g) G. M. Casanola-Martin, Y. Marrero-Ponce, M. T. Khan, A. Ather, S. Sultan, F. Torrens, R. Rotondo, Bioorg. Med. Chem. 2007, 15, 1483–1503; h) G. M. Casanola-Martin, Y. Marrero-Ponce, M. Tareq Hassan Khan, F. Torrens, F. Perez-Gimenez, A. Rescigno, J. Biomol. Screen. 2008, 13, 1014–1024.

43 a) A. Speck-Planche, L. Guilarte-Montero, R. Yera-Bueno, J. A. Rojas-Vargas, A. Garcia-Lopez, E. Uriarte, E. Molina-Perez, Pest. Manag. Sci. 2011, 67, 438–445; b) A. Speck-Planche, V. V. Kleandrova, J. A. Rojas-Vargas, Mol. Divers. 2011, 15, 901–909; c) A. Speck-Planche, V. V. Kleandrova, F. Luan, M. N. Cordeiro, Bioorg. Med. Chem. 2011, 19, 6239–6244.

44 a) Y. Marrero-Ponce, R. Medina-Marrero, A. E. Castro, R. Ramos de Armas, H. González-Díaz, V. Romero-Zaldivar, F. Torrens, Molecules 2004, 9, 1124–1147; b). Ramos de Armas, H. Gonzalez Diaz, R. Molina, E. Uriarte, Proteins 2004, 56, 715–723; c) R. Ramos de Armas, H. González-Díaz, R. Molina, M. Perez Gonzalez, E. Uriarte, Bioorg. Med. Chem. 2004, 12, 4815–4822; d)R. Ramos de Armas, H. González-Díaz, R. Molina, E. Uriarte, Biopolymers 2005, 77, 247–256; e) A. Speck-Planche, M. T. Scotti, V. de Paulo-Emerenciano, Curr. Pharm. Des., 16, 2656–2665.

45. K. C. Chou, C. T. Zhang, Crit. Rev. Biochem. Mol. Biol. 1995, 30, 275–349.
46. K. C. Chou, J. Theor. Biol. 2011, 273, 236–247.
47. lit a) M. Hayat, A. Khan, J. Theor. Biol. 2011, 271, 10–17; b)K. K. Kandaswamy, K. C. Chou, T. Martinetz, S. Moller, P. N. Suganthan, S. Sridharan, G. Pugalenthi, J. Theor. Biol. 2011, 270, 56–62; c) K. C. Chou, Proteins 2001, 43, 246–255; d) P. Zakeri, B. Moshiri, M. Sadeghi, J. Theor. Biol. 2011, 269, 208–216; e) K. C. Chou, Bioinformatics 2005, 21, 10–19; f) H. Mohabatkar, Protein Pept. Lett. 2010, 17, 1207–1214; g) H. Mohabatkar, M. M. Beigi, K. Abdolahi, S. Mohsenzadeh, Med. Chem. 2013, 9, 133–137.
48. V. A. Ivanisenko, S. S. Pintus, D. A. Grigorovich, N. A. Kolchanov, Nucleic Acids Res. 2005, 33, D183–187.
49. P. D. Dobson, A. J. Doig, J. Mol. Biol. 2003, 330, 771–783.

50 a) Z. C. Wu, X. Xiao, K. C. Chou, Mol. BioSyst. 2011, 7, 3287–3297; b) X. Xiao, Z. C. Wu, K. C. Chou, J. Theor. Biol. 2011, 284, 42–51; c) K. C. Chou, Z. C. Wu, X. Xiao, Mol. BioSyst. 2012, 8, 629–641; d) X. Xiao, Z. C. Wu, K. C. Chou, PLoS ONE 2011, 6(6), e20592.

51 a) M. Perez Gonzalez, A. Morales Helguera, J. Comput. Aided Mol. Des. 2003, 17, 665–672; b) Y. Marrero-Ponce, R. Medina-Marrero, F. Torrens, Y. Martinez, V. Romero-Zaldivar, E. A. Castro, Bioorg. Med. Chem. 2005, 13, 2881–2899; c) Y. Marrero-Ponce, A. Montero-Torres, C. R. Zaldivar, M. I. Veitia, M. M. Perez, R. N. Sanchez, Bioorg. Med. Chem. 2005, 13, 1293–1304; d) H. González-Díaz, A. Sanchez-Gonzalez, Y. Gonzalez-Diaz, J. Inorg. Biochem. 2006, 100, 1290–1297.

52. H. Van Waterbeemd, in Method and Principles in Medicinal Chemistry, Vol. 2 (Eds: R. Manhnhold, P. Krogsgaard-Larsen, H. Timmerman, H. Van Waterbeemd), Wiley-VCH, New York, 1995, pp. 283–293.

53 a) H. B. Shen, K. C. Chou, Anal. Biochem. 2008, 373, 386–388; b) H. B. Shen, K. C. Chou, Protein Eng. Des. Sel. 2007, 20, 561–567; c) K. C. Chou, H. B. Shen, Biochem. Biophys. Res. Commun. 2007, 357, 633–640; d) K. C. Chou, H. B. Shen, Nat. Protoc. 2008, 3, 153–162.

54 a) K. C. Chou, J. Proteome Res. 2005, 4, 1413–1418; b) K. C. Chou, D. W. Elrod, J. Proteome Res. 2002, 1, 429–433; c) K. C. Chou, D. W. Elrod, J. Proteome Res. 2003, 2, 183–190; d) K. C. Chou, H. B. Shen, J. Proteome Res. 2006, 5, 1888–1897; e) K. C. Chou, H. B. Shen, J. Proteome Res. 2006, 5, 3420–3428.

55. S. Wu, Y. Zhang, Nucleic Acids Res. 2007, 35, 3375–3382.

56. L. A. Kelley, M. J. Sternberg, Nat. Protoc. 2009, 4, 363–371.

57. H. Gonzalez-Diaz, F. Prado-Prado, X. Garcia-Mera, N. Alonso, P. Abeijon, O. Caamano, M. Yanez, C. R. Munteanu, A. Pazos, M. A. Dea-Ayuela, M. T. Gomez-Munoz, M. M. Garijo, J. Sansano, F. M. Ubeira, J. Proteome Res. 2011, 10, 1698–1718.

58. K. C. Chou, H. B. Shen, Natural Science 2009, 2, 63–92.