Combinatorial Chemistry & High Throughput Screening. 2015;18: 735-750

Bio-AIMS collection of chemoinformatics web tools based on molecular graph information and artificial intelligence models

Cristian R. Munteanu[1], Humberto González-Díaz[2,3], Rafael García[1], Mabel Loza[4] and Alejandro Pazos[1]

[1] *Informatton and Communication Technologies Department, Faculty of Computer Science, University of A Coruna, 15071 A Coruña, Spain*
[2] *Department of Organic Chemistry* 11, *Faculty of Science and Technology, University of the Basque Country UPV/EHU, 48940 Leioa, Vizcaya, Spain*
[3] *IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Vizcaya, Spain*
[4] *Grupo BioFarma-USEF, Departamento de Farmacologia, Facultad de Farmacia, Campus Universitario Sur s/n*, 15782 *Santiago de Compostela, Spain*

Abstract
The molecular information encoding into molecular descriptors is the first step into *in silico* Chemoinformatics methods in Drug Design. The Machine Leaning methods are a complex solution to find prediction models for specific biological properties of molecules. These models connect the molecular structure information such as atom connectivity (molecular graphs) or physical-chemical properties of an atom/group of atoms to the molecular activity (Quantitative Structure - Activity Relationship, QSAR). Due to the complexity of the proteins, the prediction of their activity is a complicated task and the interpretation of the models is more difficult. The current review presents a series of 11 prediction models for proteins, implemented as free Web tools on an Artificial Intelligence Model Server in Biosciences, Bio-AIMS (http://bio-aims.udc.es/TargetPred.php). Six tools predict protein activity, two models evaluate drug - protein target interactions and the other three calculate protein - protein interactions. The input information is based on the protein 3D structure for nine models, ID peptide amino acid sequence for three tools and drug SMILES formulas for two servers. The molecular graph descriptor-based Machine Learning models could be useful tools for *in silico* screening of new peptides/proteins as future drug targets for specific treatments.

Keywords
Molecular information, machine Learning, protein graphs, python scripts, QSAR models, Web tools.

1. INTRODUCTION

The *in silico* screening methods are the first step in Drug Development or protein function analysis. The advantages of this theoretical screening are the low cost, fast prediction, flexibility of the models for specific tasks, and high precision of predictive models. Therefore, Chemoinformatic methods are used by any pharma study in order to propose a small number of possible active drugs or optimal protein targets [1- 3].

The models are based on mathematical relationships between the structure and the properties of the molecules and their biological activity such as Qualitative Structure- Activity/Property Relationships (QSAR/QPDR). These models have been intensively used in Medical Chemistry and other Bio Sciences. In the past, the application of the QSAR was limited to small molecules or systems. Currently, these methods have been extended to larger systems. Thus, it is possible to predict the function of a protein based on its three-dimensional (3D) structure, the function of a DNA secondary structure, the interaction of drugs with multiple molecular targets [4, 5]. Several applications have been published in Medicinal Chemistry [6-12], Proteomics [13- 18], Drug Metabolism [19-23], Pharmaceutical Design [24- 28], Bioinfonnatics [29-34], Nanotoxity [35,36].

These results show that the graph/network theory can be extended to different systems such as genome networks, interaction networks of proteins, host-parasite networks, linguistics networks, social networks [37-42] and Internet [43]. Thus, a network/graph is an interconnected system which shares information and it is made up of nodes (elements of the systems) linked by relationships, The nodes can be atoms, molecules, organisms or other systems.

2. IN SILICO PROTEIN ACTIVITY PREDICTION

There are an important number of publications with prediction models for specific drug biological activity, drug toxicity, protein target interaction with drugs or protein - protein interactions, but the majority of these models are not free online tools. This makes difficult the prediction of new drugs or protein targets by all scientists.

Several programming tools give the possibility to calculate molecular descriptors and to find QSAR models: Chemical Development Kit (CDK) [44], Bioclipse [45], R packages such as QSARdata (Quantitative Structure Activity Relationship (QSAR) Data Sets, http://cran.r-project.org/web/packages/QSARdata/index.html), rcdk (http://cran.r-project.org/web/packages /rcdk/index.html), ChemmineR [46], RRegrs (Regressions in R, https://github.com/muntisa/ RRegrs) or Python tools such as pyWeka (Python Script for Weka Classifications, https://github.com/muntisa/pyWeka) and MathChem [47].

A reduced number of drug QSAR models are presented on online servers: Toxtree [48], OpenTox [49], OCHEM tools [50]. The following sections will describe some of the online protein QSAR tools and the collection of QSAR protein prediction servers from the Artificial Intelligence Model Server in Biosciences, Bio-AlMS.

2.1. Proteln Activity Web Tools

Different types of molecular information have been used to implement protein QSAR-like models as online tools. One of them is Cell-PLoc for the localization of proteins [51] in different organisms such as eukaryotes (Euk-mPLoc), plants (Plant-PLoc), Gram-negative bacteria (Gneg-PLoc), Gram-positive bacteria (Gpos-PLoc), viruses (Virus-PLoc), humans (Hum-mPLoc). The subcellular localization of proteins is an important issue in molecular cell biology, proteomics, system biology and drug discovery. Cell-PLoc is a package of Web servers obtained by hybridizing the 'higher level' approach with the *ab initio* approach. Using these Web servers, the scientists are able to predict the protein localization with a high expected accuracy, as demonstrated by a series of cross-validation tests on the benchmark data sets that covered up to 22

subcellular location sites. The maximum protein sequence identity is 25% for the protein in the same subcellular-location subset. The servers could also work with proteins that have different subcellular locations. These proteins with multiple locations are very interesting both in basic research and drug discovery, because they may play important biological functions. The computational time for each prediction is generally less than 5 seconds. The package is freely accessible at: http://www.csbio.sjtu.edu.cnl/bioinf/Cell-PLoc/.

Another Web tool for the prediction of the protein subnuclear localization is Nuc-PLoc. This predictor is based on pseudo amino acid composition (PseAA) descriptors [52] and pseudo position-specific scoring matrix (PsePSSM). This model can identify nuclear proteins among the following nine subnuclear locations: chromatin, heterochromatin, nuclear envelope, nuclear matrix, nuclear pore complex, nuclear speckle, nucleolus, nucleoplasm and nuclear promyelocytic leukaemia (PML) body. Nuc-PLoc is based on an ensemble classifier formed by fusing the evolution information of a protein and its pseudo amino acid composition. The user-friendly web-server is publicly available at: http://www.csbio.sjtu.edu.cnlbioinflNuc-PLoc/.

Other Web server for proteins is Signal-3L, Signal proteins with a 3-layer approach [53]. Signal peptides have a crucial role in the cells for directing the nascent proteins to their cellular and extracellular locations; thus, they are used to find new drugs or to reprogram the cells for gene therapy. The avalanche of the new protein sequences generated in the post-genomic era creates a critical need for fast, simple and cheap screening methods for protein property evaluation. This tool implements a new predictor based on a novel method for predicting signal peptide sequences and their cleavage sites in human, plant, animal, eukaryotic, Gram-positive, and Gram-negative protein sequences, respectively. The predictor works in three steps: (1) identification of a query protein as secretory or non-secretory by an ensemble classifier formed by fusing OET-KNN (optimized evidence- theoretic K-nearest neighbor) classifiers based on pseudo amino acid composition indices; (2) selection of a candidate set for the possible signal peptide cleavage sites of a query secretory protein by a subsite-coupled discrimination algorithm; (3) determination of the final cleavage site by fusing the global sequence alignment outcome for each of the aforementioned candidates through a voting system. Being a very fast tool, Signal-3L could be useful for the analysis of large-scale datasets. The free tool is available at: http://www.csbio.sjtu.edu.cnlbioinf/Signal-3U.

A similar Web tool is Signal-CF, which deals with the prediction of signal peptide sequences and their cleavage sites in eukaryotic and bacterial protein sequences using an automatic 2-layer predictor [54]: the 1st layer can identify a query protein as secretory or non-secretory; if it is secretory, the 2nd layer will identify the cleavage site of its signal peptide. The name of the server contains C for "coupling" and F for "fusion": the tool is created by incorporating the subsite coupling effects along a protein sequence and by fusing the results derived from many width-different scaled windows through a voting system. Signal-CF is available as a free web-server at: http://www.csbio.sjtu.edu.cnlbioinf/Signal-CF/.

Other servers are linked to proteases such as HJVcleave [55], which predicts HIV protease cleavage sites in proteins or Protldent, which can identify the proteases and their types by fusing functional domain and sequential evolution information [56]. In addition, the enzyme functional classes and sub-classes can be predicted with EzyPred [57]. The last tool is a 3-layer predictor with the overall success rates higher than 90%: the 1st layer identifies a query protein as enzyme or non-enzyme, the 2nd layer predicts the main functional enzyme class, and the 3rd layer describes the sub- functional class. The maximum protein identity is 40% in the same class or subclass. EzyPred is freely accessible at: http://www.csbio.sjtu.edu.cnlbioinf/EzyPred/.

The following Cheminformatics Web tools predict protein properties using Markov molecular graph descriptors.

2.2. Bio-AIMS Server Collection

The main purpose of Bio-AIMS (http://bio-aims.udc.es) is to extend the prediction power of the Complex Networks/Graphs, molecular information encoding and Artificial Intelligence techniques for macromolecules such as proteins/peptides. The second aim is to make available these prediction models as free online tools that could be used by scientists with little knowledge of Chemoinformatics or Machine Learning.

Bio-AIMS offers theoretical models based on Artificial Intelligence, Computational Biology and ChemIBio- informatics for the prediction of protein biological activity, drug-protein / protein-protein interactions, hotspots for protein-protein / DNA-protein interactions and personalized diagnostics. The models are constructed using different inputs such as drug / protein descriptors, molecular interaction networks, genetic mutations, brain activity records, and statistical / Machine Learning methods.

Bio-AlMS is divided in three parts (Fig. 1):

- TargetPred - Target Prediction presents 12 models for predicting the function of proteins, interaction of proteins in parasites or humans and drug-protein interactions by using data such as protein sequences or 3D structures and drug chemical structures

- (SMILES). The current review IS focused on II models from this section (Fig. 2).

- DiseasePred - Disease Prediction contains 2 online personalized diagnostic tools: SNPSchizo for schizophrenia based on patient Single Nucleotide Polyrnorphism (SNP) [58] and AlzPred = SVV for Alzheimer using Spectroscopy Voxel Volum.

MolStructPred - Molecular Structural Prediction of protein and nucleic acid structures and macromolecular interactions; it contains 2 servers: SASA-HS-PNA = SASA-based hotspot prediction for protein - nucleic acid interactions, and SASA-HS-PP = SASA-basd hotspot prediction for protein – protein intercations.

The server represents a free online solution to evaluate biological properties, drug interactions and diseases based on bio-chemical information. The user can send through a simple Web interface the information (chemical structure formula, protein sequences, genetic mutations, etc.) and the server evaluates a property or disease with its own resources. From February 2010 to February 2015 there have been over 6400 visitors of Bio-AIMS, from 112 countries.

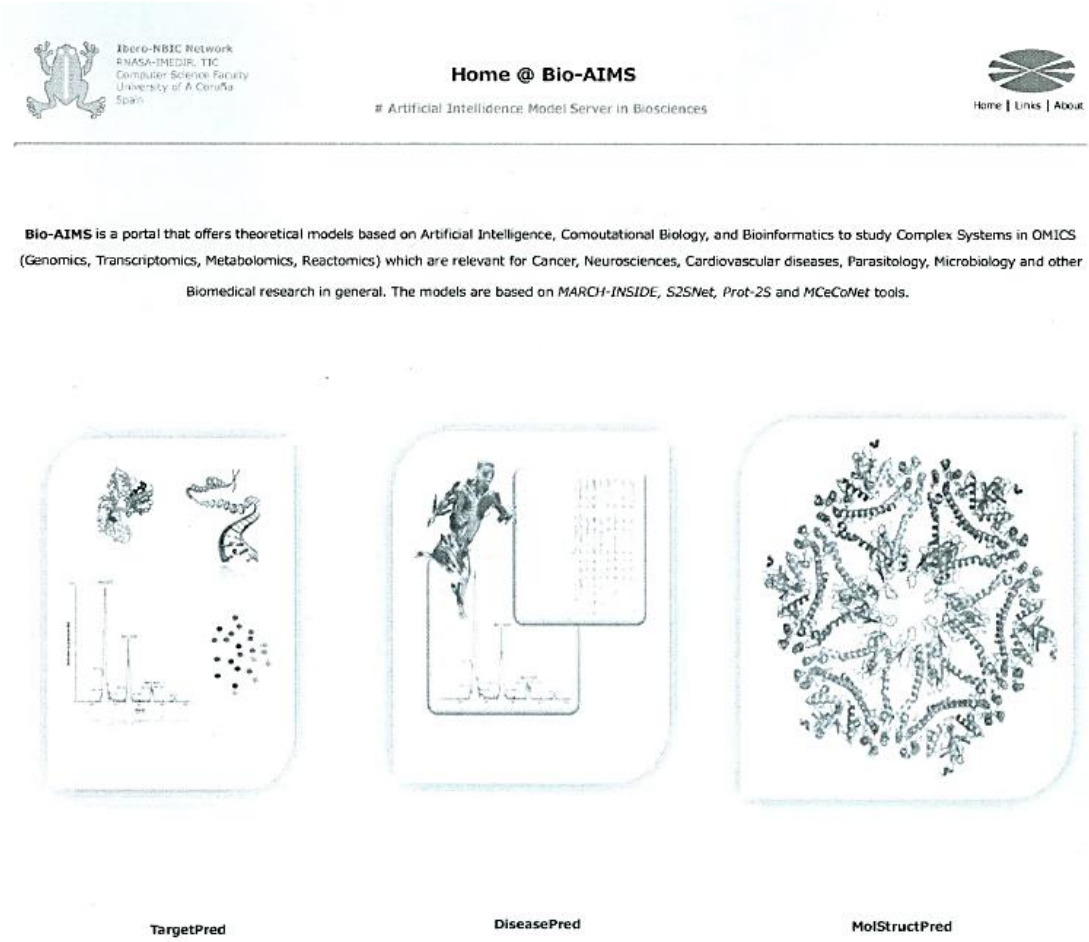

Fig. (1). Bio-AIMS server.

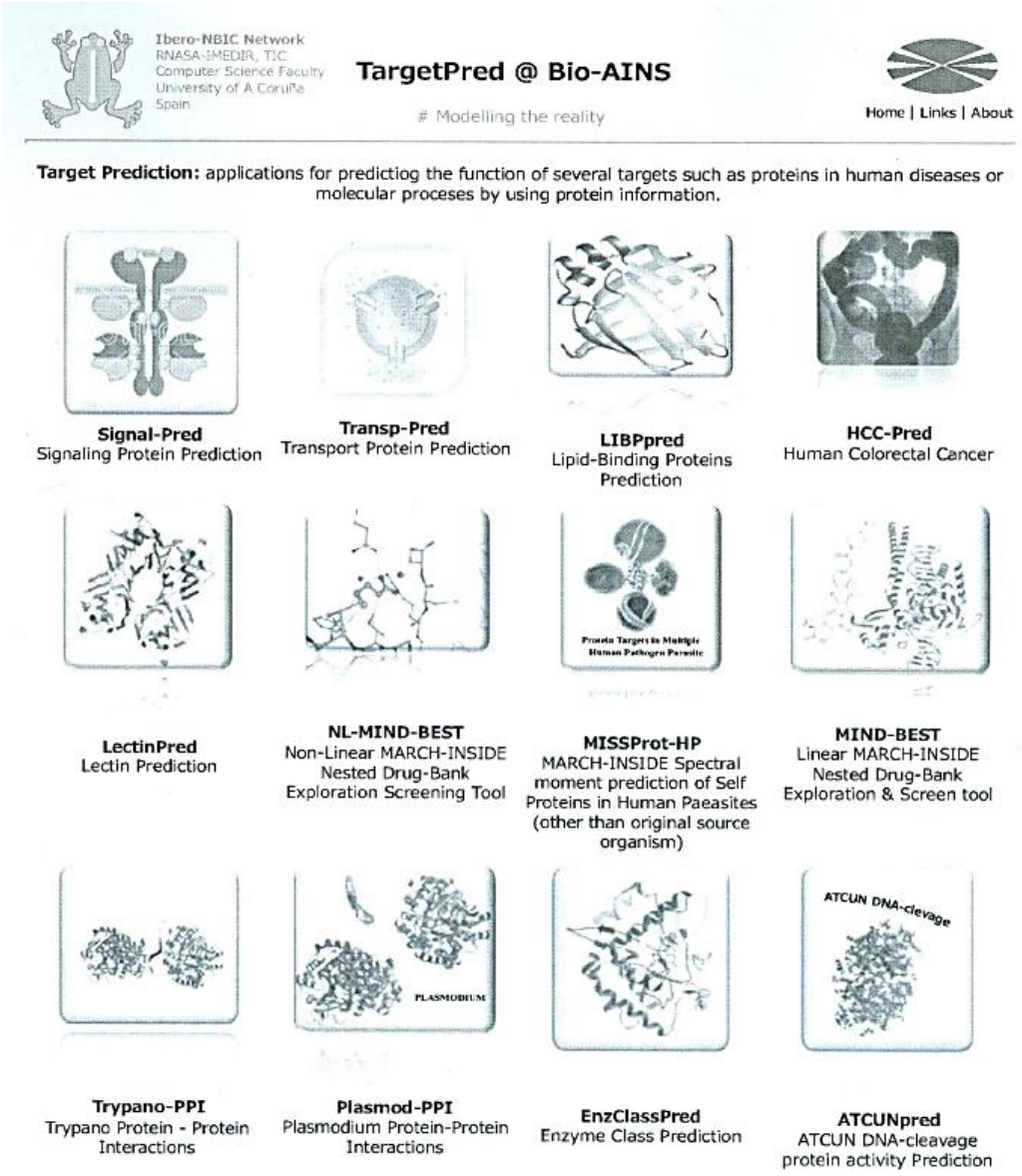

Fig. (2). Target Prediction servers frOID Bio-AIMS.

Target Prediction section contains 12 models obtained with the methodology presented in Fig. (3):

- Inputs: The inputs for the models can be protein POB name, SMILE chemical formulas for drugs or peptide sequences. These inputs are transformed into molecular indices using two in-house applications.

- Descriptor calculation tools: MARCH-INSIOE (Python version) [59] and S2SNet - Sequence to Star Network [60, 61]. Both software were programmed in Python [62] and BioPython [63]. The proteins to be evaluated are automatically downloaded from the POB databank [64, 65]. These descriptors are used to find linear and non-linear models using statistical and Machine Learning methods [66] from dedicated software.

- Model software: STA TlST1CA [67], Weka [68-70] and R [71]. The models establish quantitative relations between the structure of a system (protein molecule, protein-proteinlprotein-drug interaction networks) and an activity / property of the system (ex: biological activity of a protein). These models have been implemented into online Web servers.

- Web servers were created with XHTML [72], PHP [73], Python [62], and R [71].

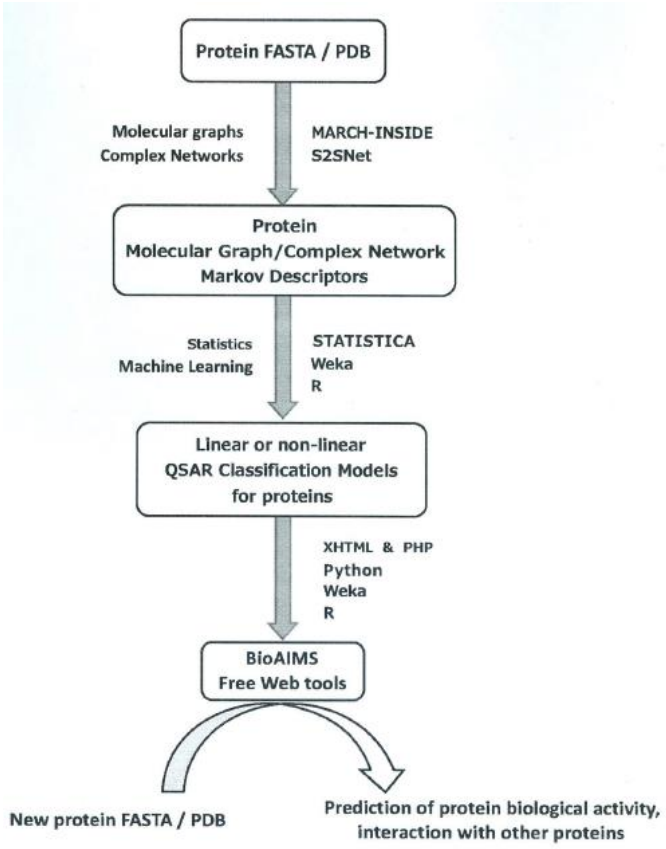

Fig. (3). General metbodology flow for Bio-AIMS Targel Prediction servers.

All models from Target Prediction section predict protein properties (biological activity, molecular interaction, drug target) and 11 models are presented in the following sections. Bio-AlMS generates a unique folder with a random name for each calculation. AII the input and output files could be found in this folder and all files are kept on the server for future access, using a unique URL for each calculation.

2.2.1. EnzClassPred - Enzyme Class Prediction

One of the protein biological activities is the enzymatic role. Different publications proposed methods to predict enzyme proteins and the enzyme family c1ass [74-78]. In the previous sections, the EzyPred server has been presented. The new Bio-AIMS Web tool, EnzClassPred represents a different molecular information codification using a Markov chain model (MCM) and protein molecular graphs for the enzyme c1assification (EC) number predictor. Thus, the protein 3D structures are turned into specific molecular descriptors: 3D entropy and moments of the molecular alpha-carbon contact network. These indices are the input for Statistics / Machine Learning techniques such as Linear Discriminant Analysis (LDA) and/or artificial neural networks (ANN) from STAT1STICA The best QSAR c1assifier to predict the first EC number has been implemented into EnzClassPred and is represented by a non- linear ANN, a Multi-Layer Perceptron (MLP) 4:4-9-8-1: 1 (4 input descriptors, two hidden layers of 9 and 8 neurons), with an overall accuracy of 98.85% [79]. The database for the model contained 4755 proteins (859 enzymes and 3896 non-enzymes) divided into both training and external validation series. The tool could be used to identify and predict peptides of prokaryote and eukaryote parasites and their hosts, as well as other superior organisms, which may be of interest in drug development or target identification. EnzClassPred uses PHP/HTMLlPython and MARCH- JNSIDE routines and is available for free at: http://bio-aims.udc.eslEnzClassPred.php.

EnzClassPred Web interface is a mini mal one, where the user should input only the standard PDB name (Fig. 4). An example of output result for three proteins is presented below:

```
Process ID = 742854ff1f20a0ccd
PDB List = 1XS0 1A0M 1EP9
... please wait ....
PDB Update/Verification …
1XS0 1A0M 1EP9 Done!
Processing PDBs ...
1XS0 1A0M 1EP9
Result file = Results/742854ff1f20a0ccd/ECP.calc.txt
EnzClassPred @ Bio-AIMS
Enzyme Class Prediction
by using MARCH-INSIDE and MLP 4:4-9-8-1:1 (98.57% accuracy)
Results  =  http://bio-aims.udc.es/Results/742854ff1f20a0ccd/ECP.calc.txt
2015-03-10 17:43:13


Enzyme classes (EC)
....................
1 - OXIDOREDUCTASE 2 - TRANSFERASE
3 - HVDROLASE
4 - LVASE
5 - ISOMERASE 6 - LIGASE
Eval = evaluated
(YES = if we are using our model)
(NO = if the function can be found inside the PDB)
PDB EvalEC.1 EC.2 EC.3 EC.4 EC.5 EC.6
====================================
1XS0 ND NO ND VES NO NO NO
1A0M VES NO VES VES VES NO VES
1EP9 NO NO VES NO NO NO NO
```

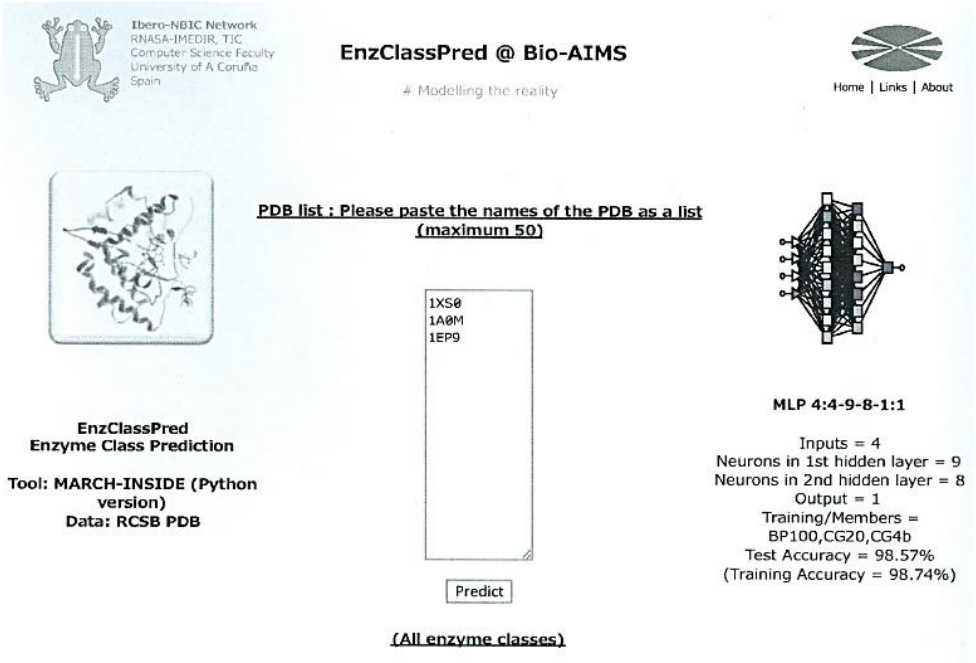

Fig. (4). EnzClassPred Web grapbical interface.

2.2.2. ATCUNpred - ATCUN DNA-Cleavage Protein Activity Prediction

The amino terminal Cu(II)- and Ni(Il)-binding (A TCUN) motif is a small metal-binding site. Discovered in serum albumin, it was demonstrated that it participated to the DNA cleavage with the NH2-aal-aa2-His3 [80] sequence, to the central nervous system function, cancer growth [81], Alzheimer's disease [82], and to several other biochemical reactions. These motifs become therapeutic agents for the chemical nuclease design [83, 84].

ATCUNpred was developed to predict the metal- mediated biological activity based only on the 3D structure of metal-unbound proteins and it is focused on the amino terminal Cu(II)- and Ni(II)-binding (ATCUN) motifs that participate in the DNA cleavage and have antitumor activity [85]. It is the first A TCUN classification model based on the 3D electrostatic spectral moments for 415 different proteins, including 133 potential ATCUN antitumor proteins. ATCUNpred is a linear model obtained with the Linear Discriminant Analysis and it discriminates between ATCUN-DNA cleavage proteins and non-active proteins with 91.32% accuracy (379 out of 415 of proteins including both training and external validation series). For the first time, the model predicted the DNA cleavage function of proteins from the pathogen parasites. Possible ATCUN-like proteins have been predicted with a probability over 99% in nine parasite families such as *Trypanosoma, Plasmodium, Leishmania,* or *Toxoplasma.* The results showed possible A TCUN proteins such as oxidoreductases, signaling proteins, lyases, membrane proteins, ligases, hydrolases, transferases, cell adhesion proteins, metal binders, translation proteins, transporters, structural proteins, and isomerases.

ATCUNpred Web interface is similar to the EnzClassPred with only a list of standard PDB name inputs and it is freely available at: http://bio-aims.udc.es/ATCUNPred.php.

An example of output is presented below:

```
ATCUNpred @ Bio-AIMS
ATCUN ONA-cleavage protein activity Prediction
by using MARCH- INSIDE and LOA based on electrostatic spectral moments
(Accuracy of 91.32%)
Results = http://bio-aims.udc.es/Results/1768754ff1e51100cd/ATCUNpred.
calc2. txt
2015-03-10 17:39:46
PDB    ATCUN Prediction
====================
1AZP   0.28 %
1I4M   66.01 %
1B0U   80.97 %
```

2.2.3. Trypano-PPI - Trypanosome Protein-Protein Interactions (TPPI)

Trypanosoma brucei has an important role in the African diseases, for causing the African trypanosomiasis in humans. Therefore, over 60 million people from the countries of sub- Saharan Africa are threatened, with consequences for the human health and the economy. *Trypanosoma cruzi* is another pathogen responsible for Chagas disease in South America. This disease causes acute illness and death, especially in young children. Thus, the discovery of novel drug targets in *Trypanosome* proteome represents an urgent task for the scientists.

Trypano-PPI implements for the first time a model that can predict unique protein-protein interactions (PPIs) in Trypanosome (TPPIs) [86]. The model classifier had the inputs as new protein-protein complex invariants based on the Markov average electrostatic potential for amino acids located in different regions of i-*th* protein and placed at a distance *k* one from each other. More than 30 different types of parameters for 7866 pairs of proteins (l023 TPPIs and 6823 non-TPPIs) from more than 20 organisms have been calculated, including parasites and human or cattle hosts. The model implemented into Trypano-PPI is represented by a simple linear model with only two parameters that predict over 90% of TPPIs and non-TPPIs (training and test sets). Trypano-PPI is able to predict how unique a protein-protein complex in Trypanosome proteome is with respect to other parasites and hosts. Thus, this tool could be important for the antitrypanosome drug target discovery.

Trypano-PPI interface contains two lists of inputs as protein chains (with standard PDB names) which will be tested for interaction (Fig. 5) and it is freely available at: http://bio-aims.udc .es/TrypanoPPI. php.

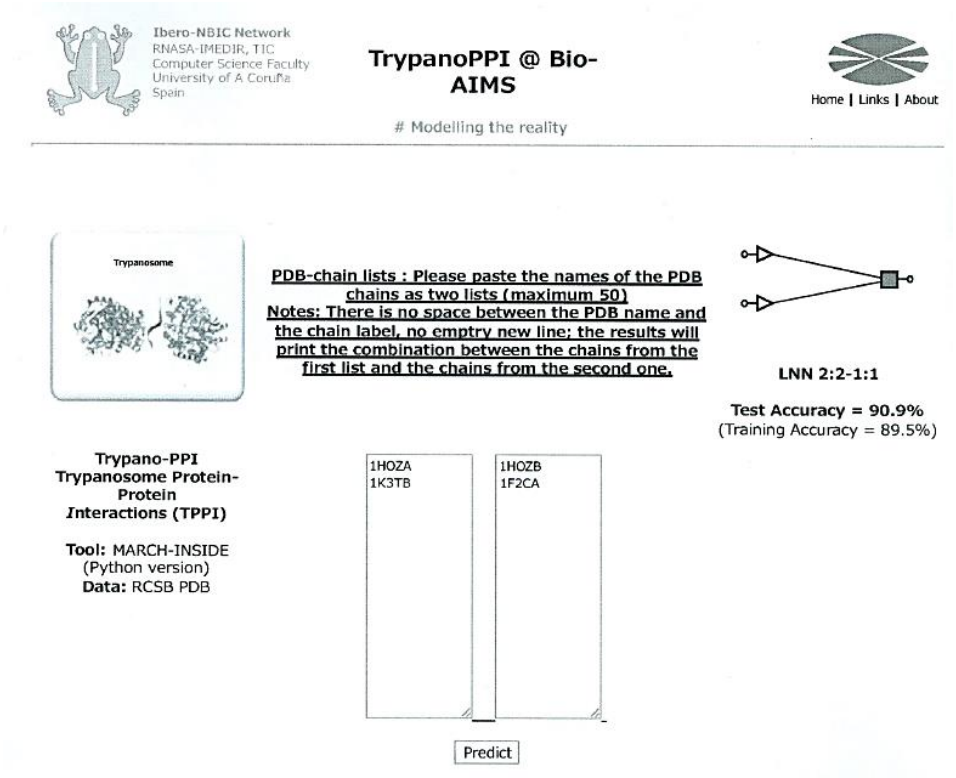

Fig. (5). Trypano-PPI Web grapbical interface

The output for four protein chains is presented below:

```
Process ID = 307254ff21a03dge4
PDB List 1= 1HOZA 1K3TB
POB List 2 = 1HOZB 1F2CA
… please wait ....
PDB Update/verification [List 1] …
1HOZA 1K3TB
PDB Update/Verification [List 2] …
1HOZB 1F2CA
Processing POB-chain List 1 …
1HOZA 1K3TB
Processing POB-chain List 2 …
1HOZB 1F2CA
Result file =
Results/307254ff21a03dge4/TrypanoPPI.calc.txt
TrypanoPPI @ Bio-AIMS
Biopython server to predict if a pair of proteins form a physically stable complex unique of Trypanosoma (not present in human or other parasites) based on electrostatic potential indices of Protein-Protein Interactions (PPIs) by using MARCH-INSIDE (Python version) and LNN 2:2-1:1 (90.9% accuracy).
These complexes may be interesting candidates for specific anti-Trypanosoma drug targets.
```

```
Results = http://bio-aims.udc.es/Results/307254ff21a03dge4/TrypanoPPI.
calc.txt
Calculated at 2015-03-10 17:53:52
Chain1 Chain2 Complex
=====================================
1HOZA  1HOZB  YES
1HOZA  1F2CA  NO
1K3TB  1HOZB  NO
1K3TB  1F2CA  NO
```

2.2.4. Plasmod-PPI- Plasmodium Protein-Protein Integrations (PPPI)

Plasmodium falciparum causes the most severe form of malaria. It kills up to 2.7 million people annually. In addition, *Plasmodium vivax* is geographically the most widely distributed and it produced over 80 million clinical cases. The drug resistance and toxicity are two major problems for these diseases. Therefore, there is a real need for new drug target methods. One possibility involves the Protein-Protein Complexes, unique in this pathogen and not present in human hosts (PPPCs). In addition, some PPCs expressed both in parasites and humans, such as DHFR synthase, play a significant role in drug resistance in both malaria and human cancer.

Considering there is no general model to predict pPPCs using indices of PPC biopolymer structure, Plasmod-PPI Web server implemented a unique classifier to predict these interactions [87]. The inputs for the model are the Markov Chain numerical descriptors of protein-protein interactions (PPIs) based on electrostatic entropy measures. These parameters have been calculated for 5257 pairs of proteins (774 pPPCs and 4483 non-pPPCs), from more than 20 organisms (including parasite and human hosts). The implemented prediction model is a simple Classification Tree with accuracy, sensitivity, and specificity between 90.2 and 98.5% (training and test sets).

Plasmod-PPI interface contains two lists of inputs as protein chains (with standard PDB names) that will be tested (in a similar way as Trypano-PPl). It is available for free at: http://bio-aims. udc.eslPlasmodPPI .php.

The output for six protein chains is presented below:

```
Process ID = 10 10S6654ff227d22ea5
PDB List 1 = 3C5IA 2F6IE 15YRC
PDB List 2 = 3C5IE 2GHUA 15YRF
… please wait …
PDB Update/Verification [List 1] …
3C5IA 2F6IE 1SYRC
PDB Update/Verification [List 2] …
3C5IE 2GHUA lSYRF
Processing PDB-chain List 1 …
3C5IA 2F6IE 1SYRC
Processing PDB-chain List 2 …
3C5IE 2GHUA 1SYRF
Result file =
Results/1086654ff227d22ea5/PlasmodPPI.calc.txt
Plasmod-PPI @ Bio-AIMS
Biopython server to predict if a pair of proteins form a physically
stable complex unique of Plasmodium (not present in human or other
```

```
parasites) based on electrostatic entropy indices of Protein-Protein
Interactions (PPIs)
by using MARCH-INSIDE (Python version) and CT (96.8% accuracy) .
These complexes may be interesting candidates for specific anti-
Plasmodium / anti-cancer drug targets.
Results = http://bio-aims.udc.es/Results/1086654ff227d22ea5/PlasmodPPI.
calc.txt
Calculated at 2015-03-10 17:57:33
Chain1 Chain2 Complex
=====================================
3C5IA  3C5IE  YES
2F6IE  2GHUA  NO
15YRC  1SYRF  YES
```

2.2.5. NL-MIND-BEST - Non-Linear MARCH-INSIDE Nested Drug-Bank Exploration & Screening Tool

The increased importance of the drug-protein interactions in the drug design for specific targets in diseases led to the need for this tool. Ligands or Drug-target pairs (DTPs/nDTPs) of drugs with high affinity/non-affinity for different targets have been selected and QSAR models have been constructed. Genera11y, most QSAR models predict activity against one protein target only and/or they have not been implemented in the form of public web server. To solve this problem, NL-MIND-BEST server implemented a multi- target QSAR (mt-QSAR) classifier using MARCH-INSIDE for the calculation of the structural parameters for drugs and target proteins (approved by the U.S. Food and Drug Administration, FDA). The best model has been obtained with the Artificial Neuronal Network (ANN) technique and it is represented by a Multi-Layer Perceptron (MLP) such as MLP 20:20-15-1:1 with a sensitivity of90.12%, a specificity of 90.46%) and an accuracy of 90.41% for the training dataset [88]. The test results are characterized by a sensitivity of 91.72%, a specificity of 91.22% and a total accuracy of 91.30%.

The Web application is based on PHP/HTML/Python and MARCH-INSIDE routines. NL-MIND-BEST Web interface contains two types of ca1culations for drug - protein interaction prediction:

- Mode 1: drugs as SMILES and proteins as standard PDBs (from PDB Databank);
- Mode 2: drugs as SMILES and proteins as uploaded ENT files created with HyperChem MMlMD algorithm; thus, theoretical PDBs could be tested as possible targets for a list oftargets.

This server is free on demand in order to be used only by academic institutes and it is available at: http://bio-aims.udc.esINL-MIND-BEST.php.

The following rows present the predictions for two drugs against two proteins (all possible combinations drug - protein):

```
Process ID = 1713354ff3620322cc
… please wait …
PDB Update/Verification …
1A36 1A8M Done!
Calculating …
Result file = .. /Results/1713354ff3620322cc/NL-MIND-BEST.Modelo txt
```

```
Class  PDB    Similar PDB  Drug   Name
       Similar Drug Drug-PDB     Similarity
=========================================================
YES    1A36   1A36: 100.0% Irinotecan
       Irinotecan: 100.0%  100.0%
No     1A36   1A36: 100.0% Simvastatin
       Simvastatin : 100.0%      100.0%
No     1A8M   1A8M: 100.0% Irinotecan
       Irinotecan: 100.0%  100.0%
YES    1A8M   1A8M: 100.0% Simvastatin
       Simvastatin: 100.0% 100.0%
```

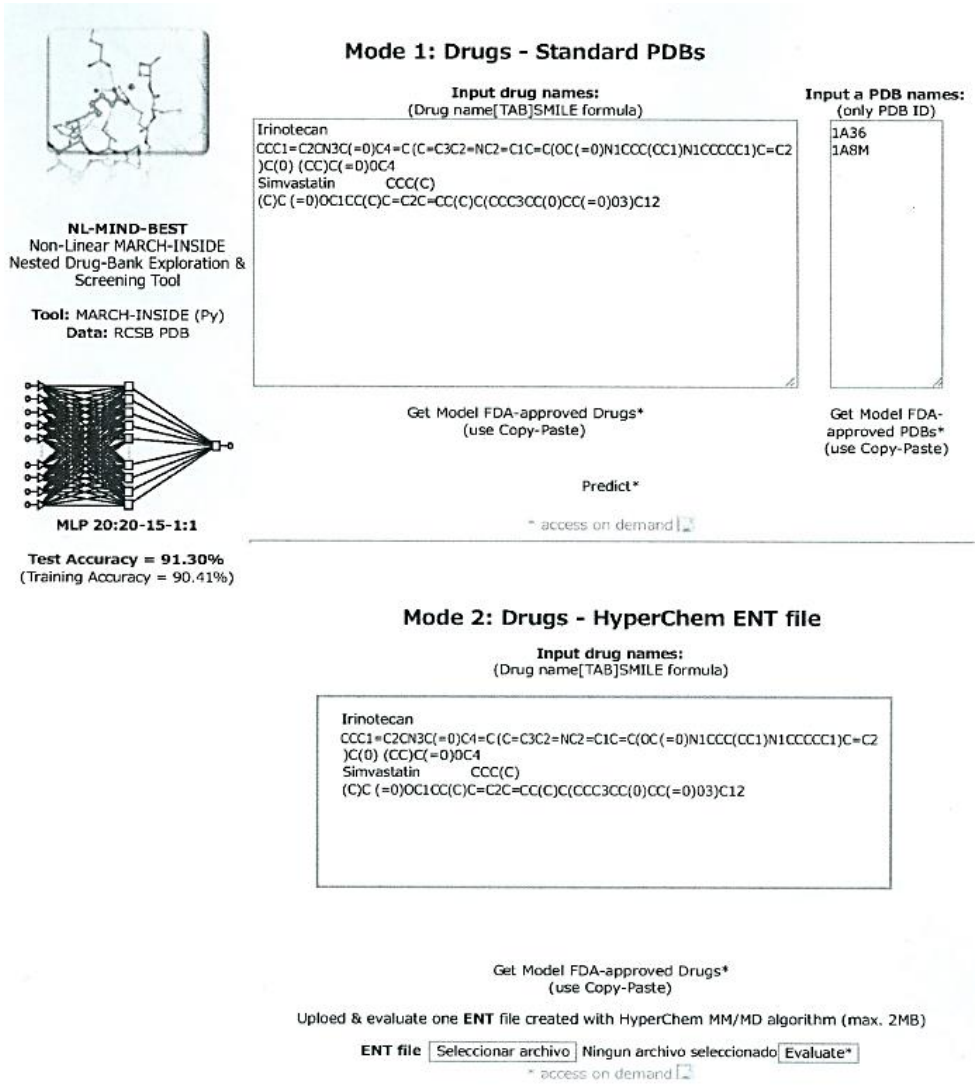

Fig. (6). NL-MIND-BEST Web graphical interface.

2.2.6. MIND-BEST - Linear MARCH-INSIDE Nested Drug-Bank Exploration & Screening Tool

This tool is the linear solution similar to the NL-MIND- BEST tool. The accuracy of the best LDA model was 94.4% (3859/4086 cases) for training and 94.9% (190912012 cases) for the external validation series [89]. It is available at: http://bio-aims.udc.esIMIND-BEST.php.

2.2.7. MISSProt-HP - MARCH-INSIDE Spectral moment prediction of Self Proteins in Human Parasites (Other than the Original Source Organism)

500 million people worldwide are affected by infections caused by human parasites (HPs). On the other hand, the chemotherapy is expensive, toxic, and sometimes less effective because of the drug resistance. Thus, there is a need for methods to predict biologically relevant Parasite Self Proteins (PSP), which are expressed differentially in a given parasite and are dissimilar to proteins expressed in other parasites. These peptide unique sequences have a high probability to become new vaccines or unique 3D structure drug targets.

MISSProt-HP implements a classification model for PSPs in eight different HPs such as *Ascaris, Entamoeba, Fasciola, Giardia, Leishmania, Plasmodium, Trypanosoma,* and *Toxoplasma* [90]. The accuracy of the model is 90% in the case of 15,341 training and validation cases. This model combines several methods: protein residue networks, Markov Chains and ANN. The input parameters are calculated with MARCH-INSIDE as the spectral moments of the Markov transition matrix for electrostatic interactions associated with the protein residue complex network.

MISS-Prot-HP was implemented into PHP/HTML/Python and it is easy to use by non-experts in Bioinformatics. It is freely available at: http://bio-aims.udc.esIMISSProt-HP.php.

MISSProt-HP Web interface is the most complex from Bio-AJMS and is divided into four types of calculations:

- Mode 1: prediction of Self Proteins in Human Parasites for Ascaris spp., Entamoeba spp., Fasciola spp., Giardia spp., Leishmania spp., Tolypocladium spp., Toxoplasma spp., Trypanosoma spp., Plasmodium spp.;

- Mode 1U: the same prediction, but in user-defined organism, such as Homo sapiens, Bos taurus, Sus scrofa, G. gallus, M. musculus, R. norvegicus, B. stearothermophilus, B. subtilis, E. coli, Pseudomonas spp., S. typhimurium, Staphylococcus spp., Streptococcus spp., Streptomyces spp. and Saccharomyces spp.;

- Mode 2: evaluation against 9 parasites of the uploaded PDB files obtained with LOMETS or ENT files from HyperChem;

- Mode 2U: evaluation of the uploaded files as Mode 3, but using a user-defined organism.

The output for Mode 1U for 3 protein chains using Homo sapiens as organism is presented below:

The probability for each organism (p_org), other than the original source organism, is calculated as follows:

```
Process ID = 2456154ff3960Sdba3
PDB List = 13PKA 1JLRA 1DQPB
... please wait ...
PDB Update/Verification ...
13PKA 1JLRA 1DQPB Done!
Processing PDBs ...
13PKA 1JLRA 1DQPB
```

```
Result file = Results/2456154ff3960Sdba3/MISSProt-HP.Mode1U.txt
MISSProt-HP @ Bio-AIMS
MARCH-INSIDE Spectral moment prediction of Self Proteins in Human
Parasites
by using LNN 90:90-1:1 (91.0% accuracy)
Results = http://bio-aims.udc.es/Results/2456154ff3960Sdba3/MI5SProt-
HP.Mode1U.txt
2015-03-10 19:35:13
Parasite type (P)
……..
P1-Ascaris          P2-Entamoeba        P3-Fasciola
P4-Giardia          P5-Leishmania       P6-Tolypocladium
P7-ToxoplasmaP8-Trypanosoma             P9-Plasmodium
UO-Homo sapiens

PDB          P1     P2     P3     P4     P5     P6
             P7     P8     P9     UO     Source
Organism (SO)
================================================
13PKA  0.0%   12.9%  4.2%   0.7%   7.8%   13.9%
             34.0%  SO     10.7%  15.8%  Trypanosoma
1JLRA  0.0%   17.4%  5.7%   1.0%   10.4%  18.7%
             SO     11.3%  14.4%  21.2%  Toxoplasma
1DQPB  0.0%   12.0%  3.9%   SO     7.2%   12.9%
             31.6%  7.8%   10.0%  14.7%  Giarda
```

2.2.8. LIBPpred - Llpid-Binding Proteins Prediction

Lipid-Binding Proteins (LIBPs) or Fatty Acid-Binding Proteins (F ABPs) play an important role in many diseases such as different types of cancer, kidney injury, atherosclerosis, diabetes, intestinal ischemia and parasitic infections. This tool proposed for the first time a model to predict new LIBPs based on protein 3D structures. The QSAR model was built on 3D electrostatic parameters of 1801 different proteins, including 801 LIBPs, calculated with the MARCH-INSIDE software [91]. The model is a linear classifier that can predict with an accuracy of 89.11% if a new protein can bind to lipids. The tool is available at: http://bio-aims.udc.esILIBPpred.php.

LIBPpred Web interface is divided into two modes:

- Mode 1: prediction of protein chains as lipid-binding peptides using as input the standard PDB names;
- Mode 2: the same prediction, but using as input an uploaded LOMET PDB file.

The output for the Mode 1 is described below:

```
Process ID = 1578754ff3c0eda057
... please wait ...
PDB Update/Verification ...
1QGHK 114M 2QZTB 1B0U Done!
Calculating ...
Result file = Results/1578754ff3c0eda057/LIBPpred.Mode1.txt
LIBPpred @ Bio-AIMS
```

Mode 1: Standard PDB input
Lipid-Binding Proteins Prediction
by using MARCH- INSIDE and LDA based on electrostatic spectral moments (Accuracy of 89.11%)
Results = http://bio-aims/Results/1578754ff3c0eda057/ 2015-03-10 19:46:40
PDBChain LIBP Prediction
==========
1QGHK 0.00%
114M* 29.66%
2QZTB 100.00%
1B0U* 45.34%
==========
* the input contains no chain => UBPpred used the entire protein (all the chains)

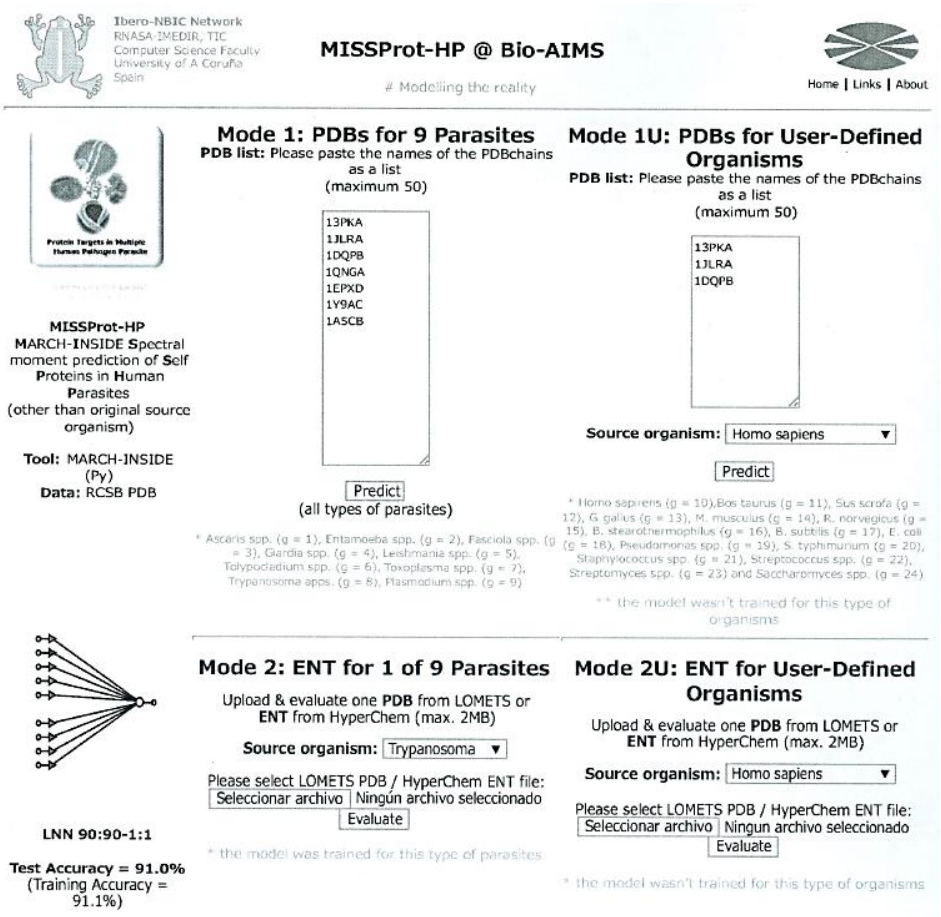

Fig. (7). MISSProt-HP Web graphical interface.

2.2.9. LectinPred - Lectin Prediction

Lectins are sugar-binding proteins that are highly specific for their sugar moieties and they play a significant role in biological recognition phenomena involving cells and proteins, such as different types of cancer, parasitic infections and other diseases. For example, some viruses use lectins to attach themselves to the cells of the host organism during infection. This tool presents a linear classifier based on Markov Shannon entropies of proteins and it can evaluate if a new protein can bind to sugars (this protein could be a lectin) with an accuracy over 90.00% [92]. Because there are +2000 proteins with 3D structure, but without a known function, LectinPred has been used to predict possible lectin proteins. The descriptors of the model have been calculated with MARCH-INSIDE (Python version) and they encode 3D electrostatic entropy of the protein molecule complex networks. A series of 2200 PDBs have been used, including 1200 lectins. The QSAR model has been obtained with the Linear Discriminant Analysis (LDA) and it is able to discriminate between lectin and non-lectin protein 3D structures.

The implemented model as a free online server could be accessed at http://bio-aims.udc.es/LECTINPred.php. The model is characterized by sensitivity of 96.7% (for lectins), specificity of 87.6% (non-lectins), and accuracy of 92.5% (for all proteins), considering altogether both the training and external prediction series.

Both the interface of LectinPred and the results are similar to LIBPpred. This server has two modes:

- Mode 1 - the input is the name of the PDB or the PDB+chain letter. All the PDBs will be automatically retrieved from the PDB databank.

- Mode 2 - In case there is no standard PDB, the user could upload PDBs generated with LOMETS or PHYRE2.

Because there is no direct relation between the lectin property of proteins and any molecular property, this model represents a unique tool that established a complex relationship between the 3D protein structure and the electronegativity of the amino acids and the lectin property.

2.2.10. HCC-Pred - Human Colorectal Cancer

Cancer is a complex disease and, therefore, it is difficult to identify specific biomarkers or the false positives and false negative predictions. Consequently, a good option is the use of the complex networks / graphs theory. This method allows describing any real system, from the small molecules to the complex genetic, neural or social networks by turning real properties into topological indices.

This tool is based on a QSAR-like model that is able to predict peptides related to colorectal cancer [93]. In the first step, the protein primary structure data (F ASTA peptide amino acid sequences) have been converted into specific Randic's star networks topological indices using S2SNet software for a set of 1054 proteins related or not to two types of cancer: human breast cancer (HBC) and human colon cancer (HCC). These descriptors were the inputs for the general discriminant analysis method from ST ATISTICA, in order to find a unique input-coded multi-target classification model with accuracies of the training/predicting set of90.0% for the forward stepwise model type.

From this model, the equation which corresponds to the HCC has been extracted and implemented as a free online Web tool. The simple inputs are peptide amino acid sequences with two labels. This tool could be useful in Clinical Proteomics, where a large volume of peptide sequences are obtained without any known biological activity.

The Web interface of HCCPred contains only one input list as peptide amino acid sequences (one letter codification). The format of each line input is [PeptideNamel] [PeptideName2]

[Sequence). The output predicting if one peptide sequence is related to HCC: http://bio-aims.udc.esIHCCPred.php.

2.2.11. Transp-Pred - Transport Protein Prediction

Another important topic in Drug Metabolism is the transport of molecules within cells. Because of the high costs of the experimental testing for these types of molecules, theoretical models are considered for this screening task, as a cheap and fast solution.

Thus, Transp-Pred was implemented using a QSAR mode1 that is able to predict whether a new peptide amino acid sequence (FASTA format) could be a transporter [94]. In the first step, the primary structure of a protein was represented as a molecular Star graph and the corresponding topological indices have been calculated. The dataset consisted of 2503 protein chains, out of which 413 are transporters, according to the PDB database c1assification, and 2090 are non-transporters. These indices were used as input in Weka to find the best QSAR Machine Learning c1assification mode1 that can evaluate the transporter function of a new protein chain. The best method found is the Support Vector Machine Recursive Feature Elimination, which produced a c1assification model based on only 20 attributes with a true positive rate of 83% and a false positive rate of 16.7%.

The interface of Transp-Pred is similar to the HCCPred: there is only an input list with peptide sequences that will be tested for the transport biological activity. The tool is free1y available at: http://bio-aims.udc.eslTranspPred.php.

CONCLUSION

This review describes a collection of free online protein QSAR tools known as Bio-AIMS. These unique servers allow the researchers to evaluate the biological activity of proteins, as well as their interaction with drugs or other proteins. These QSAR models are extended from the classic drug QSAR c1assification models to complex molecules such as proteins. All the tools are based on topological indices of molecular protein graphs and protein-protein/protein-drug interaction networks. The Machine Learning methods employed to find the models are different, producing linear or non-linear models, from the simple Linear Discriminant Analysis method to the very abstract Support Vector Machine Recursive Feature Elimination.

Bio-AIMS contributes to the open science concept with free user-friendly tools for scientists from different fields of science. The future models will be based on 100% open software such as Weka, R and Python.

CONFLICT OF INTEREST

The authors confirm that this artic1e content has no conflict of interest.

ACKNOWLEDGEMENTS

The authors acknowledge the support provided by the Galician Network of Drugs R+D REGID (Xunta de Galicia R2014/025) and by the "Collaborative Project on Medical Informatics (CIMED)" PI 13/00280 funded by the Carlos III Health Institute from the Spanish National plan for Scientific and Technical Research and Innovation 2013-2016 and the European Regional Development Fund / GAIN (FEDER - CONECTAPEME - INTERCONECTA). This work was partially supported by the Galician Network for Colorectal Cancer Research (Red Gallega de Cáncer Colorrectal - REGICC, Ref.: CN 20121217*)*, Institute for Biomedical Informatics of A

Coruña (INIBIC), and Center for Research of Information and Communication Technologies (CITIC).

REFERENCES

[1] Roy, K. Application of chemometrics and cheminfonnatics in antimalarial drug research. *Combo Chem. High Throughput Screen.,* 2015, 18 (2), 89-90.
[2] Tang, C. Y.; Hung, C. L.; Hsu, C. H.; Zheng, H.; Lin, C. Y. Novel computing technologies for bioinformatics and cheminformatics. *Biomed. Res. Int.,* 2014, 2014, 392150.
[3] Akella, L. B.; DeCaprio, D. Cheminformatics approaches 10 analyze diversity in compound screening libraries. *Curro Opin. Chem. Biol.,* 2010, 14 (3), 325-330.
[4] Randic, M.; Pompe, M. The variable molecular descriptors based on distance related matrices. J. *Chem. Inf Comput. Sci.,* 2001. 41 (3),575-581.
[5] Gonzalez-Diaz, H. Quantitative studies on Structure-Activity and Structure-Property Relationships (QSAR/QSPR). *Curro Top. Med. Chem.,* 2008, 8 (1 S), 1554.
[6] Wang, J. F.; Chou, K. C. Molecular modeling of cytochrome P450 and drug metabolism. *Curro Drug Metab.,* 2010, 11 (4),342-346.
[7] Mrabet, Y.; Semmar, N. Mathematical methods to analysis of topology, functional variability and evolution of metabolic systems based on different decomposition concepts. *Curro Drug Metab.,* 2010,11 (4), 315-341.
[8] Li, Z.; Huang, C.; Bai, S.; Pan, X.; Zhou, R.; Wei, Y.; Zhao, X. Prognostic evaluation of epidennal fatty acid-binding protein and ca1cyphosine, two proteins implicated in endometrial cáncer using a proteomic approach. *Int. J. Cancer,* 2008, 123 (10),2377-2383.
[9] Martinez-Romero, M.; Vazquez-Naya, J. M.; Rabunal, J. R.; Pita- Fernandez, S.; Macenlle, R.; Castro-Alvarino, J.; Lopez-Roses, L.; Ulla, J. L.; Martinez-Calvo, A. V.; Vazquez, S.; Pereira, J.; Porto- Pazos, A. B.; Dorado, J.; Pazos, A.; Munteanu, C. R. Artificial intelligence techniques for colorectal cancer drug metabolism: ontology and complex network. *Curro Drug Metab.,* 2010, 11 (4), 347-36S.
[10] Khan, M. T. Predictions of the ADMET properties of candidate drug molecules utilizing different QSAR/QSPR modelling approaches. *Curro Drug Metab.,* 2010, 11 (4), 285-295.
[11] Junior, F. J.; Scotti, L.; Botelho, S. P.; Da Silva, M. S.; Scotti, M. T. Benzo- and Thienobenzo-Diazepines: Multi-target Drugs for CNS Disorders, *Mini Rev. Med. Chem.,* 2015,15(8),630-647.
[12] Scotti, L.; Scotti, M. T..; Silva, V. B.; Santos, S. R.; Cavalcanti, S. *C.;* Mendonca, F. J., Jr. Chemometric studies on potemiallarvicidal compounds against Aedes aegypti. *Med. Chem.,* 2014,10 (2),201- 210.
[13] Chen, J.; Shen, 8. Computational Analysis of Amino Acid Mutation: a Proteome Wide Perspective. *Curr. Proteomics,* 2009, 6 (4),228-234.
[14] Chou, K. C. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Curro Proteomics.* 2009,6 (4),262-274.
[15] Ivanciuc, O. Machine learning Quantitative Structure-Activity Relationships (QSAR) for peptides binding to Human Amphiphysin-I SH3 domain. *Curro Proteomics,* 2009, 6 (4),289- 302.
[16] Pérez-Montoto, L. G.; Prado-Prado, F.; Ubeira, F. M.; González- Díaz, H. Study of Parasitie lnfections, Cancer, and other Diseases with Mass-Spectrometry and Quantitative Proteome-Disease Relationships. *Curro Proteomics,* 2009, 6 (4),246-261.
[17] Torrens, F.; Castellano, G. Topologieal Charge-Transfer Indiees: From Small Molecules to Proteins. *Curro Proteomics,* 2009, 6 (4), 204-213.
[18] Vázquez, J. M.; Aguiar, V.; Seoane, J. A.; Freire, A.; Serantes, J. A.; Dorado, J.; Pazos, A.; Munteanu, C. R. Star Graphs of Protein Sequences and Proteome Mass Spectra in Cancer Prediction. *Curr Proteomics,* 2009, 6 (4), 275-288.
[19] Chou, K. C. Graphic rule for drug metabolism systems. *Curro Drug Metab.,* 2010,11 (4), 369-78.
[20] Garcia, I..; Diop, Y. F.; Gomez, G. QSAR & complex network study of the HMGR inhibitors structural diversity. *Curro Drug Metab.,* 2010,11 (4), 307-314.
[21] Gonzalez-Diaz, H. Network topological indices, drug metabolism, and distribution. *Curro Drug Metab.,* 2010, 11 (4), 283-284.
[22] Gonzalez-Diaz, H.; Duardo-Sanchez, A.; Ubeira, F. M.; Prado- Prado, F.; Perez-Montoto, L. G.; Coneu, R.; Podda, G.; Shen, S. Review of MARCH-INSIDE & complex networks predietion of drugs: ADMET, ami-parasite activity, metabolizing enzymes and cardiotoxicity proteome biomarkers. *Curro Drug Metab.,* 2010, 11 (4),379-406.
[23] Zhong, W. Z.; Zhan, J.; Kang, P.; Yamazaki, S. Gender specific drug metabolism of PF-02341 066 in rats-role of sulfoconjugation. *Curro Drug Metab.,* 2010, 11 (4), 296-306.
[24] Gonzalez-Diaz, H.; Romaris, F.; Duardo-Sanchez, A; Perez- Montoto, L. G.; Prado-Prado, F.; Patlewiez, G.; Ubeira, F. M. Predieting drugs and proteins in parasite infections with topological indices of complex networks: theoretical background s, applications, and legal issues. *Curr. Pharm. Des.,* 2010, 16 (24), 2737-64.

[25] Marrero-Ponce, Y.; Casanola-Martin, G. M.; Khan, M. *T.;* Torrens, F.; Reseigno, A.; Abad, C. Ligand-Based Computer-Aided Discovery of Tyrosinase Inhibitors. Applications of the TOMOCOMD-CARDD Method to the Elucidation of New Compounds. *Curro Pharm. Des.,* 2010, 16 (24),2601-2624.
[26] Munteanu, C. R.; Fernandez-Blanco, E.; Seoane, J. A.; Izquierdo- Novo, P.; Rodriguez-Femandez, J. A.; Prieto-Gonzalez, J. M.; Rabunal, J. R.; Pazos, A. Drug discovery and design for complex diseases through QSAR computational methods. *Curro Pharm. Des.,* 2010, 16 (24),2640-2655.
[27] Speck-Planche, A.; Cordeiro, M. N. Multitasking models for quantitative structure-biological effeet relationships: eurrent status and future perspectives to speed up drug discovery. *Expert Opino Drug Discov,* 2015, 10 (3).245-256.
[28] de Araujo, R. S.; Guerra, F. Q.; de, O. L. E.; de Simone, C. A.; Tavares, J. F.; Sconi, L.; Scotti, M. *T.;* de Aquino, T. M.; de Moura, R. O.; Mendonca, F. J.; Barbosa-Filho, J. M. Synthesis, Structure-Activity Relationships (SAR) and *in Silico* Studies of Coumarin Derivatives with Antifungal Activity. *Int. J. Mol. Sci.,* 2013,14 (1),1293-1309.
[29] Garcia, L; Fall, Y.; Gómez, G. Trends in Bioinfonnatics and Chemoinformatics of Vitamin D analogues and their protein targets. *Curro Bioinf.,* 2011, 6 (1), 16-24.
[30] lvanciuc, T.; Ivaneiuc, O.; Klein, D. J. Network-QSAR with Reaction Poset Quantitative Superstructure-Activity Relationships (QSSAR) for PCS Chromatographie Properties, *Curro Bioinf.,* 2011,6 (l), 25-34.
[31] Bhauacharjee, B.; Jayadeepa, R. M.; Banerjee, S.; Joshi, J.; Middha, S. K.; Mole, J. P.; Samuel, J. Review of Complex Network and Gene Ontology in pharmacology approaches: Mapping natural compounds on potential drug target Colon Cancer network. *Curr·. Bioinf.,* 2011, 6 (1),44-52.
[32] Wan, S. S.; Hu, L. L.; Niu, S.; Wang, K.; Cai, Y. D.; Lu, W. C.; Chou, K. C. Identification of multiple subcellular locations for proteins in budding yeast. *Curr. Bioinf.,* 2011, 6 (1),71-80.
[33] Speck-Planche, A.; Cordeiro, M. N. D. S. Application of Bioinforrnatics for the search of novel anti-viral therapies: Rational design of anti-herpes agents. *Curro Bioinf.,* 2011, 6 (1), 81-93.
[34] Dave, K.; Banerjee, A. Bioinformatics analysis of functional relations between CNPs regions. *Curro Bioinf.,* 2011, 6 (1), 122- 128.
[35] Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N. Computational modeling in nanomedicine: prediction of multiple antibacterial profiles of nanopanicles using a quantitative structure-activity relationship penurbation model. *Nanomedicine (Lond),* 2015, 10 (2), 193-204.
[36] Kleandrova, V. V.; Luan, F.; Gonzalez-Diaz, H.; Ruso, J. M.; Melo, A; Speck-Planche, A.; Cordeiro, M. N. Computational ecotoxicology: simultaneous prediction of ecotoxic effects of nanoparticles under different experimental conditions. *Environ. Int.,* 2014,73,288-294.
[37] Breiger, R. The Analysis of Social Networks. In: *Handbook of Data Analysis,* Hardy, M.; Bryman, A., Eds. Sage Publications: London, 2004; pp 505-526.
[38] Abercrornbie, N.; HiII, S.; Turner, B. S. Social structure. In: *The Penguin Dictionary of Sociology,* 4th ed.; Penguin: London, 2000.
[39] Craig, C. Social Structure. In: *Dictionary of the Social Sciences,* Oxford University Press: Oxford, 2002.
[40] White, H., Scott Boorman and Ronald Breiger .. "." Social Structure from Multiple Networks: I Blockmodels of Roles and Positions. *American Journal of Sociology;* 1976,81,730-780.
[41] Wellman, S.; Berkowitz, S. D. *Social Structures: A Network Approach,* Cambridge University Press: Cambridge, 1988.
[42] Newman, M. E. J. The structure and function of complex networks. *SIAM Review,* 2003, 45, 167-256.
[43] Bonchev, D. On the complexity of directed biological networks. *SAR QSAR Environ. Res.,* 2003, 14 (3),199-214.
[44] Steinbeek, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): an open- source Java library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.,* 2003, 43 (2), 493-500.
[45] Spjuth, O.; Helmus, T.*;* Willighagen, E. L.; Kuhn, S.; Eklund, M.; Wagener, J.; Murray-Rust, P.; Steinbeck, C.; Wikberg, J. E. Bioclipse: an open source workbench for chemo- and bioinformatics. *BMC Bioinformatics,* 2007, 8, 59.
[46] Cao, Y.; Charisi, A.; Cheng, L. C.; Jiang, *T.;* Girke, T. ChemmineR: a compound mining framework for R. *Bioinformatics,* 2008, 24 (15), 1733-1734.
[47] Alexander, V.; Dragan, S. MathChem: A Python Package For Calculating Topological Indices. *MATCH Commun. Math. Comput. Chem.,* 2014,71,657-680.
[48] Patlewicz, G.; Jeliazkova, N.; Safford, R. J.; Worth, A. P.; Aleksiev, B. An evaluation of the imp1ementation of the Cramer c1assification scheme in the Toxtree software. *SAR QSAR Environ. Res.,* 2008, 19 (5-6),495-524.
[49] Willighagen. E. L.; Jeliazkova, N.; Hardy, B.; Grafstrom, R. C.; Spjuth, O. Computational toxicology using the OpenTox application programming interface and Bioclipse. *BMC Res. Notes,* 2011,4,487.
[50] Sushko, l.; Novotarskyi, S.; Komer, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; Todesehini, R.; Varnek, A; Marcou, G.; Enl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, c.; Baskin, ll; Palyulin, V. A.; Radchenko, E. V.; Welsh, W. J.; Kholodovych, V.; Chekmarev, D.; Cherkasov, A.; Aires-de-Sousa, J.; Zhang, Q. Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkaehenko, V.; Tetko, L V. Online chemical modeling environment

(OCHEM): web platform for data storage, model development and publishing of chemical information. *J. Comput. Aided Mol. Des.*, 2011, 25 (6). 533-554.
[51] Chou, K. C.; Shen, H. S. Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.*, 2008, 3 (2),153-62.
[52] Chou, K. C. Prediction of protein cellular attributes using pseudo- amino acid composition. *Proteins*, 2001, 43 (3), 246-255.
[53] Shen, H. S.; Chou, K. C. Signal-3L: A 3-layer approach for predicting signal peptides. *Biochem. Biophys. Res. Commun., 2007,* 363 (2). 297-303.
[54] Chou, K. C.; Shen, H. B. Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. *Biochem. Biophys. Res. Commun.*, 2007, 357 (3),633-640.
[55] Shen, H. B.; Chou, K. C. HIVcleave: a web-server for predicting HIV protease c1eavage sites in proteins. *Anal. Biochem.*, 2008, 375, 388-390.
[56] Chou, K. C.; Shen, H. B. Protldent: a web server for identifying proteases and their types by fusing functional domain and sequential evolution information. *Biochem. Biophys. Res. Commun.*, 2008, 376 (2),321-325.
[57] Shen, H. B.; Chou, K. C. EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. *Biochem. Biophys. Res. Commun.*, 2007, 364 (1),53-59.
[58] Aguiar-Pulido, Y.; Seoane, J. A.; Rabunal, J. R; Dorado, J.; Pazos, A.; Munteanu, C. R Machine learning techniques for single nucleotide polymorphism - disease classification models in schizophrenia. *Molecules*, 2010, 15 (7), 4875-4889.
[59] González-Díaz, H.; Torres-Gomez, L. A.; Guevara, *Y.;* Almeida, M. S.; Molina, R.; Castanedo, N.; Santana, L.; Uriarte, E. Markovian chemicals *"in silico"* design (MARCH-INSIDE), a promising approach for computer-aided molecular design III: 2.5D in dices for the discovery of antibacterials. *J. Mol. Model.*, 2005, 11 (2), 116-123.
[60] Munteanu, C. R.; Gonzáles-Díaz, H. *S2SNet - Sequence to Star Network. Reg. No. 03 / 2008 / 1338, Santiago de Compostela, Spain*, Santiago de Compostela, Spain, 2008.
[61] Munteanu, C. R.; Magalhaes, A. L.; Duardo-Sanchez, A.; Pazos, A.; Gonzalez-Diaz, H. S2SNet: A Tool for Transforming Characters and Numeric Sequences into Star Network Topological Indices in Chemoinformatics, Bioinformatics, Biomedical, and Social-Legal Sciences. *Curro Bioirf.*, 2013, 8 (4),429-437.
[62] Rossum, G. V. Python Reference Manual. http://docs.python.org/ref/ref.html.
[63] Jeff Chang, B. C., Iddo Friedberg, Thomas Hamelryck, Michiel de Hoon, Peter Cock Biopython Tutorial and Cookbook. 2007.
[64] Bernstein, F. C.*;* Koetzle, T. F.; Williams, G. J.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, *T.;* Tasumi, M. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 1977, 1/2 (3), 535- 542.
[65] Berman, H. M.; Westbrook, I.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Boume, P. E. The Protein Data Bank *Nucleic Acids Res.*, 2000, 28, 235-242.
[66] Mitchell, T. *Machine Learning, 1997.*
[67] StatSoft.lnc. *STATISTlCA (data analysis software system). version 6.0, www.statsoft.com. Statsoft.Lnc* . 6.0; 2002.
[68] Witten, I. H.; Frank, E. *WEKA: Waikato Environment for Knowledge Analysis., 2000.*
[69] Frank, E.; Hall, M.; Trigg, L.; Holmes, G.; Witten, I. H. Data mining in bioinformatics using Weka. *Bioinformatics*, 2004, 20 (15),2479-2481.
[70] Ivanciuc, O. Weka machine learning for predicting the phospholipidosis inducing potential. *Curro Top. Med. Chem., 2008,* 8 (18),1691-1709.
[71] Team, R D. C. *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria*, R Foundation for Statistical Computing: Vienna, Austria, 2008.
[72] Pemberton, S.; Altheim, M.; AskJeeves, A. D.; Boumphrey, F.; Mitre, G. B.; Donoho, A. W.; Dooley, S.; Hofrichter, K..; Hoschka, P.; Ishikawa, M.; Wamer, K.; King, P.; Klante, P.; Matsui, S.; McCarron, S.; Navarro, A.; Nies, Z.; Raggen, D.; Schmitz. P.; Schnitzenbaumer, S.; Stark, P.; Wilson, C.*;* Wugofski, T.*;* Zigmond, D. XHTMUM 1.0: The Extensible HyperText Markup Language. W3C Recommendation. http://www.w3.orgITR/2000IREC-xhtmll-20000126/.
[73] Lerdorf, R. Dynantic Web Pages with PHP3. WebTechniques. http://www.php.net.
[74] Bate, P.; Warwicker, J. Enzyme/non-enzyme discrimination and prediction of enzyme active site location using charge-based methods. *J. Mol. Biol.*, 2004, *340* (2), 263-276.
[75] Cai, Y. D.; Chou, K. C. Using functional domain composition to predict enzyme family classes. *J. Proteome Res.*, 2005, 4 (1), 109- 11.
[76] Cai, Y. D.; Chou, K. C. Predicting enzyme subclass by functional domain composition and pseudo amino acid composition. *J. Proteome Res.*, 2005, 4 (3), 967-971.
[77] Cai, Y. D.; Zhou, G. P.; Chou, K. C. Predicting enzyme family classes by hybridizing gene product composition and pseudo-amino acid composition. *J. Theor. Biol.*, 2005, 234 (1), 145-149.
[78] Dobson, P. D.; Doig, A. J. Predicting enzyme c1ass from protein structure without alignments. *J. Mol. Biol.*, 2005, 345 (1),187-199.

[79] Concu, R; Dea-Ayuela, M. A.; Perez-Montoto, L. G.; Prado-Prado, F. J.; Uriane, E.; Bolas-Femandez, F.; Podda, G.; Pazos, A.; Munteanu, C. R; Ubeira, F. M.; Gonzalez-Diaz, H. 3D entropy and moments prediction of enzyme c1asses and experimental-theoretic study of peptide fingerprints in Leishmania parasites. *Biochim. Biophys. Acta,* 2009, 1794 (12), 1784-1794.
[80] Laussac, J. P.; Sarkar, B. Characterization of the copper(II)- and nickel(II)-transport site of human serum albumin. Studies of copper(II) and nickel(II) binding to peptide 1-24 of human serum albumin by 13C and IH NMR spectroscopy. *Biochemistry (Mosc.),* 1984, 23 (12), 2832-2838.
[81] Harford, C.*;* Sarkar, B. Neuromedin C binds Cu(II) and Ni(II) *vio* the A TCUN motif: implications for the CNS and cancer growth. *Biochem. Biophys. Res. Commuu.,* 1995, 209 (3),877-882.
[82] Drew, S. C, Noble, C. J.; Masters, C. L.; Hanson, G. R; Barnham, K. J. Pleomorphic copper coordination by Alzheimer's disease amyloid-beta peptide. *J. Am. Chem. Soc.,* 2009, 131 (3), 1195- 1207.
[83] Singh, R K.; Sharma, N. K.; Prasad, R.; Singh, U. P. DNA c1eavage study using copper (II)-GlyA.ibHis: a tripeptide complex based on A TCUN peptide motifs. *Protein Pepl. Lett.,* 2008, 15 (1), 13-19.
[84] Melino, S.; Gallo, M.; Trotta, E.; Mondello, F.; Paci, M.; Petruzzelli, R Metal-binding and nuclease activity of an antimicrobial peptide analogue of the salivary histatin 5. *Biochemistry (Mosc.),* 2006, 45 (51), 15373-1583.
[85] Munteanu, C. R.; Vazquez, J. M.; Dorado, *J.;* Sierra, A. P.; Sanchez-Gonzalez, A.; Prado-Prado, F. J.; Gonzalez-Diaz, H. Complex network spectral moments for A TCUN motif DNA c1eavage: first predictive study on proteins of human pathogen parasites. *J. Proteome Res.,* 2009, 8 (11), 5219-5228.
[86] Rodriguez-Soca, Y.*;* Munteanu, C. R; Dorado, J.; Pazos, A.; Prado-Prado, F. J.; Gonzalez-Diaz, H. Trypano-PPI: a web server for prediction of unique targets in trypanosome proteome by using electrostatic parameters of protein-protein interactions. *J. Proteome Res.,* 2010, 9(2),1182-1190.
[87] Yamilet Rodriguez-Soca, C. R M., Julian Dorado, Juan Rabuñal, Alejandro Pazos and Humberto González-Diaz Plasmod-PPI: a web-server predicting complex biopolymer targets in Plasmodium with entropy measures of protein-protein interactions. *Polymer,* 2010,51 (1), 264-273.
[88] Gonzalez-Diaz, H.; Prado-Prado, F.; Sobarzo-Sanchez, E.; Haddad, M.; Maurel Chevalley, S.; Valentin, A.; Quetin-Leclercq, J.; Dea- Ayuela, M. A.; Teresa Gomez-Munos, M.; Munteanu, C. R; Jose Torres-Labandeira, J.; Garcia-Mera, X.; Tapia, R. A.; Ubeira, F. M. NL MIND-BEST: A web server for ligands and proteins discovery- Theoretic-experimental study of proteins of Giardia lamblia and new compounds active against Plasmodium falciparum. *J. Theor. Biol.,* 2011, 276 (1), 229-249.
[89] Gonzalez-Diaz, H.; Prado-Prado, F.; Garcia-Mera, X.; Alonso, N.; Abeijon, P.; Caamano, O.; Yanez, M.; Munteanu, C. R.; Pazos, A.; Dea-Ayuela, M. A.; Gomez-Munoz, M. T.*;* Garijo, M. M.; Sansano, J.; Ubeira, F. M. MIND-BEST: Web server for drugs and target discovery; design, synthesis, and assay of MAO-B inhibitors and theoretical-experimental study of G3PDH protein from Trichomonas gallinae. *J. Proteome Res.,* 2011, 10 (4),1698-1718.
[90] Gonzalez-Diaz, H.; Muino, L.; Anadon, A. M; Romaris, F.; Prado- Prado, F. J.; Munteanu, C. R.; Dorado, J.; Sierra, A. P.; Mezo, M.; Gonzalez-Warleta, M.; Garate, *T.;* Ubeira, F. M. MISS-Prot: web server for self/non-self discrimination of protein residue networks in parasites; theory and experiments in Fasciola peptides and Anisakis allergens. *Mol. BioSyst.,* 2011, 7 (6),1938-1955.
[91] Gonzalez-Diaz, H.; Munteanu, C. R.; Postelnicu, L.; Prado-Prado, F.; Gestal, M.; Pazos, A. UBP-Pred: web server for lipid binding proteins using structural network parameters; PDB mining of human cancer biomarkers and drug targets in parasites and bacteria. *Mol. BioSyst.,* 2012, 8 (3),851-862.
[92] Munteanu, C. R.; Pedreira-Souto, N.; Dorado, J.; Pazos, A.; Pérez- Montoto, L. G.; Ubeira, F. M.; González-Díaz, H. LECTINPred: web server that uses complex networks of protein structure for prediction of lectins with potential use as cancer biomarkers or in parasite vaccine designo *Molecular lnformatics,* 2014, 33 (4), 276- 285.
[93] Munteanu, C. R.; Magalhaes, A. L.; Uriarte, E.; Gonzalez-Diaz, H. Multi-target QPDR classification model for human breast and colon cancer-related proteins using star graph topological indices. *J. Theor. Biol.,* 2009, 257 (2), 303-31!.
[94] Fernandez-Lozano, C.; Gestal, M.; Pedreira-Souto, N.; Postelnicu, L.; Dorado, J.; Munteanu, C. R. Kernel-based feature selection techniques for transport proteins based on star graph topological indices. *Curr. Top. Med. Chem.,* 2013, 13 (14),1681-1691.