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

Cristian R. Munteanu¹, Humberto González-Díaz²³, Rafael García¹, Mabel Loza⁴ and Alejandro Pazos¹

¹ Information and Communication Technologies Department, Faculty of Computer Science, University of A Coruña, 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 Farmacología, 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 Learning 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], Bioinformatics [29-34], Nanotoxicity [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. Protein 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.cn/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.cn/bioinf/Nuc-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.cn/bioinf/Signal-3L.

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.cn/bioinf/Signal-CF/.

Other servers are linked to proteases such as HVcleave [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.cn/bioinf/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 ChemIBioinformatics 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-AIMS 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.

- **DiseasePred - Disease Prediction** contains 2 online personalized diagnostic tools: SNPSchizo for schizophrenia based on patient Single Nucleotide Polymorphism (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-based 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.
Bio-AIMS is a portal that offers theoretical models based on Artificial Intelligence, Computational Biology, and Bioinformatics to study Complex Systems in OMICS (Genomics, Transcriptomics, Metabolomics, Proteomics) which are relevant for Cancer, Neurosciences, Cardiovascular diseases, Parasitology, Microbiology and other biomedical research in general. The models are based on MARROW, S2Stab, Prot-2S and HolosNet tools.

**Fig. (1). Bio-AIMS server.**
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.

---

Fig. (2). Target Prediction servers frOID Bio-AIMS.
- Model software: STATISTICA [67], Weka [68-70] and R [71]. The models establish quantitative relations between the structure of a system (protein molecule, protein-protein-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].

![Diagram of Bio-AIMS Target Prediction](Fig. (3). General methodology flow for Bio-AIMS Target 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-AIMS generates a unique folder with a random name for each calculation. All 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 class [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 classification (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 STATISTICA. The best QSAR classifier 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/HTML/Python and MARCH-JINSIDE routines and is available for free at: http://bio-aims.udc.es/EnzClassPred.php.

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

<table>
<thead>
<tr>
<th>Process ID</th>
<th>PDB List</th>
<th>Result file</th>
</tr>
</thead>
<tbody>
<tr>
<td>74285ff1f20a0cccd</td>
<td>1XS0 1A0M 1EP9</td>
<td>Results/74285ff1f20a0cccd/ECP.calc.txt</td>
</tr>
</tbody>
</table>

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/74285ff1f20a0cccd/ECP.calc.txt
2015-03-10 17:43:13

Enzyme classes (EC)

<table>
<thead>
<tr>
<th>No.</th>
<th>Enzyme Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>OXIDOREDUCTASE</td>
</tr>
<tr>
<td>2</td>
<td>TRANSFERASE</td>
</tr>
<tr>
<td>3</td>
<td>HYDROLASE</td>
</tr>
<tr>
<td>4</td>
<td>LVASE</td>
</tr>
<tr>
<td>5</td>
<td>ISOMERASE</td>
</tr>
<tr>
<td>6</td>
<td>LIGASE</td>
</tr>
</tbody>
</table>

Eval = evaluated
(YES = if we are using our model)
(NO = if the function can be found inside the PDB)
PDB Eval	EC.1	EC.2	EC.3	EC.4	EC.5	EC.6

<table>
<thead>
<tr>
<th>PDB</th>
<th>Eval</th>
<th>EC.1</th>
<th>EC.2</th>
<th>EC.3</th>
<th>EC.4</th>
<th>EC.5</th>
<th>EC.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1XS0</td>
<td>ND</td>
<td>NO</td>
<td>NO</td>
<td>VES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>1A0M</td>
<td>VES</td>
<td>NO</td>
<td>VES</td>
<td>VES</td>
<td>VES</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>1EP9</td>
<td>NO</td>
<td>VES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
</tr>
</tbody>
</table>
2.2.2. **ATCUNpred - ATCUN DNA-Cleavage Protein Activity Prediction**

The amino terminal Cu(II)- and Ni(II)-binding (ATCUN) 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-aal2-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 ATCUN 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 ATCUN 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.calk2. 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 (1023 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 antitypansomose 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.
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.
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 PPCs 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 PPCs and 4483 non-PPCs), 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-PPI). 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 1SYRC
PDB List 2 = 3C5IE 2GHUA 1SYRF
... please wait ...
PDB Update/Verification [List 1] ...
  3C5IA 2F6IE 1SYRC
PDB Update/Verification [List 2] ...
  3C5IE 2GHUA 1SYRF
Processing PDB-chain List 1 ...
  3C5IA 2F6IE 1SYRC
Processing PDB-chain List 2 ...
  3C5IE 2GHUA 1SYRF
Result file = Results/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
1SYRC 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. Generally, 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 of 90.12%, 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 calculations 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 MMiMD algorithm; thus, theoretical PDBs could be tested as possible targets for a list of targets.

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

<table>
<thead>
<tr>
<th>YES 1A36 1A36: 100.0% Irinotecan</th>
<th>Irinotecan: 100.0% 100.0%</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO 1A36 1A36: 100.0% Simvastatin</td>
<td>Simvastatin: 100.0% 100.0%</td>
</tr>
<tr>
<td>YES 1A8M 1A8M: 100.0% Irinotecan</td>
<td>Irinotecan: 100.0% 100.0%</td>
</tr>
<tr>
<td>YES 1A8M 1A8M: 100.0% Simvastatin</td>
<td>Simvastatin: 100.0% 100.0%</td>
</tr>
</tbody>
</table>

---

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.

MISSProt-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 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 = 2456154ff39605dba3
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.Model1U.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/MISSProt- HP.Model1U.txt
2015-03-10 19:35:13

Parasite type (P)

Parasite type (P)

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

PDB P1 P2 P3 P4 P5 P6
UO P7 P8 P9 Source

Organism (SO)

================================================================================================
13PKA 0.0% 12.9% 4.2% 0.7% 7.8% 13.9%
34.0% SO 10.7% 15.8% Trypanosoma
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 - Lipid-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.es/LIBPpred.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.Model1.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%
180U* 45.34%

* the input contains no chain => UBpred 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 (FASTA 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 [PeptideName1] [PeptideName2]
[Sequence]. The output predicting if one peptide sequence is related to HCC: http://bio-aims.udc.es/lHCCPred.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 model 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 classification, and 2090 are non-transporters. These indices were used as input in Weka to find the best QSAR Machine Learning classification model 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 classification 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 freely available at: http://bio-aims.udc.es/lTranspPred.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 classification models to complex molecules such as proteins. All the tools are based on topological indices of molecular protein graphs and 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 article 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
REFERENCES


[34] Dave, K.; Banerjee, A. Bioinformatics analysis of functional relations between CNPs regions. *Curro Bioinf.*, 2011, 6 (1), 122-128.


