MOL2NET, International Conference Series on Multidisciplinary Sciences

Machine learning techniques and the identification of new potentially active compounds against *Leishmania infantum*.

***Naivi Flores-Balmaseda*[a,*], *Susana Rojas-Socarrás*[a], *Juan Alberto Castillo-Garit*[a,b]**

[a]*Unit of Computer-Aided Molecular "Biosilico" Discovery and Bioinformatic Research (CAMD-BIR Unit), Faculty of Chemistry-Pharmacy. Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba.* ***nflores@uclv.edu.cu***

[b]*Unidad de Toxicología Experimental, Universidad de Ciencias Médicas de Villa Clara, Santa Clara, Villa Clara, Cuba, Cuba*

.

Graphical Abstract

Abstract.

Leishmaniasis is defined as a set of diseases of very varied clinical presentation produced by obligate intracellular parasites belonging to the genus *Leishmania*. They have been classified by the World Health Organization in category I of infectious diseases and are part of neglected tropical pathologies. *Leishmania infantum* mainly affects children under five years of age and has been associated with an increase in the appearance of cutaneous and visceral leishmaniasis. The search for new therapeutic alternatives remains a challenge and *in silico* studies are alternative tools to solve this problem. With the main objective of identify potentially effective compounds against *Leishmania infantum* through *in silico* studies, artificial Intelligence techniques implemented in the WEKA program and molecular descriptors 0D-2D of DRAGON software are used in this research. A new database was created and the clusters analysis (AC) *k-means* was used to design the training and prediction series. Four models were obtained with the following techniques: IBk, J48, MLP and SMO that reached percentages of classification higher than 80% for training and prediction series, whose predictive power was confirmed through external and internal validation procedures. The use of the models obtained in the virtual screening of the international database DrugBank and synthesis compounds allowed the optimal identification of 120 new potentially active compounds against *Leishmania infantum* amastigote form.

Keywords

Leishmaniasis; machine learning techniques; protozoo; WEKA software; *Leishmania infantum*; amastigote.

Introduction

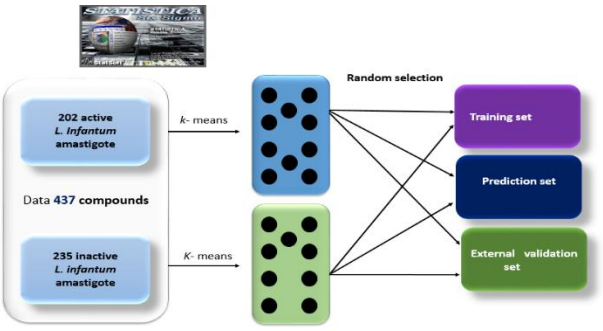
Leishmaniasis incidence has increased from the years 80, and it has won a relevant position among the causes of death for infectious illnesses worldwide[1]. They have been classified by the World Health Organization in category I of infectious diseases and are part of neglected tropical pathologies. *Leishmania infantum* mainly affects children under five years of age and has been associated with an increase in the appearance of cutaneous and visceral leishmaniasis[2]. The search for new therapeutic alternatives remains a challenge and *in silico* studies are alternative tools to solve this problem[3].

Materials and Methods

In this work, 437 PubChem bioassays tested compounds against the amastigote form of *L. infantum* parasite were selected to construct a new database with a high degree of structural variability; they have been tested experimentally through trials with very similar procedures. To classify them into active or inactive against this stage of parasite life the IC_{50} was used. Different families of 0-2D molecular descriptors were calculated using DRAGON software[4].

Conglomerate analysis (AC) implemented in the STATISTICA 8.0 processing package was carried out, as a way of evaluating the existing structural diversity and distribution within the groups of active and inactive observations respectively, figure 1. The active and inactive compounds were in turn divided into different subsets by means of two conglomerate analyzes of the k-MCA type[5]. From each conglomerate, the compounds for the conformation of the training, prediction and external validation series were randomly selected; the used procedure is shown in Figure 2. WEKA's selection procedures were used to obtain a subset of variables for models development[6].

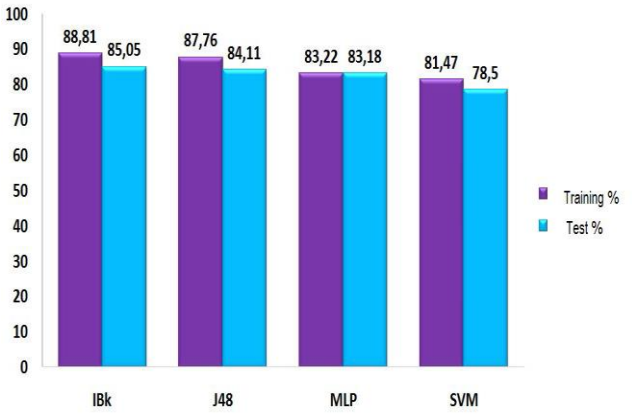
Figure 1. Conglomerate analysis and series conformation.



Results and Discussion

Four models were obtained with the following techniques: *k-Nearest Neighbors* (IBK), *Classification Trees* (J48), *Artificial Neural Network* (MLP for its acronym *MultiLayer Perceptron*) and *Support Vector Machine* (SMO for *Sequential Minimal Optimization*). For training and prediction series, they reached percentages of classification higher than 80% whose predictive power was confirmed through external and internal validation procedures (sensitivity, specificity, Matthews's correlation coefficient, false positive relationship and accuracy for the training and prediction series were determined for each model). Classification percentages for training and prediction series in the final models obtained in this work results higher in IBk model followed by J48, figure 2. The external validation of 44 previously bioassayed compounds (PubChem) yielded positive results for the four models used for amastigotes demonstrating its high degree of predictability, robustness and reproducibility.

Figure 2. Classification percentages (Accuracy) for TS and PS in final models.



A total of 5 128 compounds of different origin (DrugBank international database, and new synthesis compounds) that had not been tested experimentally against L. infantum were virtually screened, this allowed the identification of new potentially active agents against the amastigote form of this parasite, resulting in the identification of a wide structural variety compounds for each of the four models. Virtual screening allowed the optimal identification of 120 new potentially active compounds against Leishmania infantum amastigote form, which will be evaluated experimentally in subsequent studies to corroborate their activity.

References

1. Organización Panamericana de la Salud: Leishmaniasis: Informe Epidemiológico en las Américas: Washington: Organización Panamericana de la Salud, 2016 Available in http://www.paho.org/ accessed on 12 sept 2016

2. Ramos, JM; Segovia, M; Estado actual del tratamiento farmacológico de la leishmaniasis. [cited 2016]; Available from:

http://www.seq.es/seq/html/revista_seq/0197/rev1.html.

3. Timothy G. Geary; Judy A. Sakanari, and Conor R. Caffrey Anthelmintic Drug Discovery:
Into the Future. *Journal of Parasitology* 2015, 101, No. 2, 125-133.

4. Todeschini R, Consonni V, Mauri A, Pavan M. DRAGON-Software for the calculation of molecular descriptors, version 5.5 for Windows; Talete SRL: Milan, Italy. 2007

5. Xu, J; Hagler A. Chemoinformatics and drug discovery. Molecules. 2002;7(8):566-600.Hall M,

6. Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter. 2009 Nov 16;11(1):10 -8.