SciForum
MOL2NET

A Combined Approach of Ligand-based and Structure-based Virtual Screening to Select Structures with Potential Antichagasic Activity from SISTEMATX Sesquiterpene Lactones Database.

Chonny Herrera-Acevedo[1], Luciana Scotti [1] and Marcus Tullius Scotti [1*]

[1] Post-Graduate Program in Natural and Synthetic Bioactive Products, Federal University of Paraíba, 58051-900 João Pessoa, PB, Brazil; caherrera@ltf.ufpb.br (C.H.A); luciana.scotti@gmail.com (L.S); mtscotti@gmail.com (M.S)
***** Correspondence: mtscotti@gmail.com; Tel.: +55-83-99869-0415

Received: / Accepted: / Published:

Abstract. Chagas disease is an endemic disease caused by *Trypanosoma cruzi*, which affects more than eight million people, mostly in the Americas. A search for new treatments is necessary to control and eliminate this disease. Sesquiterpene lactones (SLs) are an interesting group of secondary metabolites characteristic of Asteraceae that have presented a wide range of biological activities. From the ChEMBL database, we selected a diverse set of 4,452, 1,635 and 1,322 structures with tested activity against the three T. cruzi parasitic forms, amastigote, trypomastigotes and epimastigote, respectively, to create random forest (RF) models with an accuracy of greater than 74 % for cross-validation and test sets. Afterwards, a ligand-based virtual screen of the entire SLs of Asteraceae database stored in SistematX (1,306 structures) was performed. In addition, a structure-based virtual screen was also performed for the same set of SLs using molecular docking for *T. cruzi* cruzain. Finally, using an approach combining ligand-based and structure-based virtual screening along with the equations proposed in this study to normalize the probability scores, we verified potentially active compounds and established a possible mechanism of action.

Keywords: Asteraceae; Chagas' disease; Ligand-based virtual screening; Structure-based virtual screening; Sesquiterpene lactones; Machine learning.

Mol2Net YouTube channel*: http://bit.do/mol2net-tube*
YouTube link: *please, paste here the link to your personal YouTube video, if any.*

1. Introduction

Chagas' disease is an endemic disease caused by Trypanosoma cruzi, which affects more than seven million people, mostly in the Americas [1]. The search for new treatments is necessary for the control and elimination of this disease. Natural products have been an invaluable source of inspiration for the development of therapeutic agents [2,3]. Sesquiterpene lactones (SLs) are one of those interesting small molecules for the search of new chemotherapies against infectious diseases [4,5].

Using a combined approach of ligand-based and structure-based virtual screening (VS) with the entire SLs databank stored in SistematX (http://sistematx.ufpb.br), we verified potentially active compounds against Trypanosoma cruzi and established a possible mechanism of action.

2. Results and Discussion

Ligand-based VS

The training set hit-rate values for the three RF models are quite close to or exactly 100%; nevertheless, for cross-validation and test sets values range from 64.6% to 91.1%, with epimastigote and trypomastigote models serving as better predictors of inactive molecules than the amastigote model. The specificity of the epimastigote model is better than the other two models, as the percentage of true negative compounds predicted in the test set (91.1%) was higher than the cross-validation set (85.6%). The amastigote model is the most sensitive of the three, presenting a true positive prediction rate of 76.7% and 79.1% for the cross-validation and test sets, respectively. In turn, the models for the two other parasitic forms were approximately 10% less sensitive to the values reached in the amastigote model.

Using this machine learning algorithm, a virtual screen was performed on a set comprising 1,306 molecules obtained from SistematX. For amastigotes, 34 SLs were predicted to be antichagasic compounds, with probability values ranging from 0.50 to 0.58. Some common structural features are observed among the structures with higher probability values, SLs **1–2** (Figure 1). are acetylated molecules germacranolides contained an epoxide moiety in their structures.

Otherwise, 17 SLs were predicted to be anti-T. cruzi compounds for the trypomastigote parasitic form, with probability values ranging from 0.50 to 0.64. Desacetyl-isotenulin (**3**, Figure 1) was the structure with the highest probability value. The structures of the active molecules are similar (guaianolides). Finally, the epimastigote model was less selective than the other two models, as 420 active molecules were predicted, with probability values ranging from 0.50 to 0.82. As in the amastigote model, structural similarity was observed between SLs with higher probability values (**5-6**).

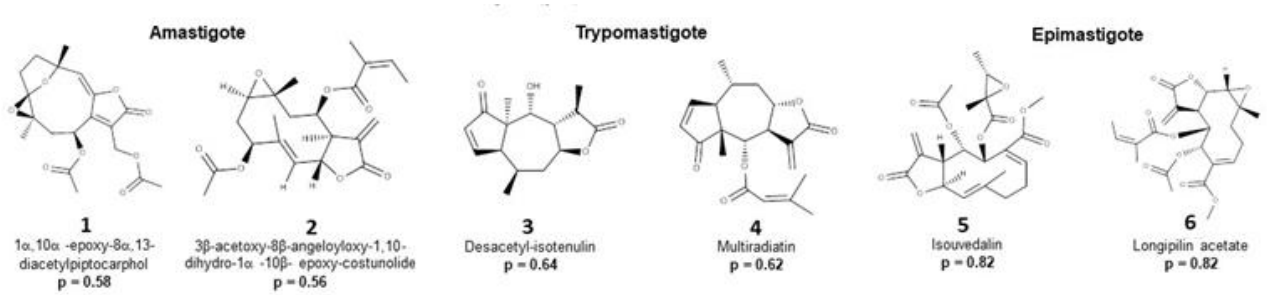

Figure 1. Potentially active sesquiterpene lactones identified using a ligand-based virtual screening; *p*= active probability value.

Structure-based VS

Initially, molecular docking was validated by redocking of the original ligand for *T. cruzi* cruzain. This score is listed in Table 1 with their respective RMSD value.

Table 1. The docking energy (kJ/mol) of two of the best-ranked SLs from the structure-based approach for cruzain. Ligand = energy (kJ/mol) for the PDB ligand and the RMSD values obtained from the redocking procedure.

***T. cruzi* protein**	**SL (KJ/mol)**	**Ligand (KJ/mol)**	**Redocking RMSD**
Cruzain	**7** (-91.4) **8** (-84.2)	(-80.0)	0.79

After, a virtual screen of 1,306 SLs was performed. Based on the binding energy values, all tested molecules were ranked using the following probability calculation (p_s, Equation 1):

Equation 1:

$$p_s = \frac{E_i}{E_{min}} \quad \text{if} \quad E_i < E_{ligand}$$

where p_s = structure-based probability; E_i = docking energy of compound *i*, and *i* ranges from 1 to 1306 (SLs dataset); E_{min} = the lowest energy value of the dataset; E_{ligand} = the ligand energy from protein crystallography.

For 753 SLs, values greater than 0.5 and binding energy values less than the ligand were observed. The structures **7** and **8** (Figure 2A), two guaianolide SLs extracted from *Lactuca georgica*, presented the highest active probability values in structure-based VS. Figure 2B shows the conformations of both SLs in the active site of Cruzain, as well as the hydrogen-bonding (H-bond) interactions of compound **7** (Figure 5B) with residues Cys 25, Trp 26 and Trp 184. Molecule **8** also participated in H-bond interactions with Cys 25 and Trp 184. In both SLs, an H-bond was observed between the carbonyl moiety of carbon-2 with Trp 184.

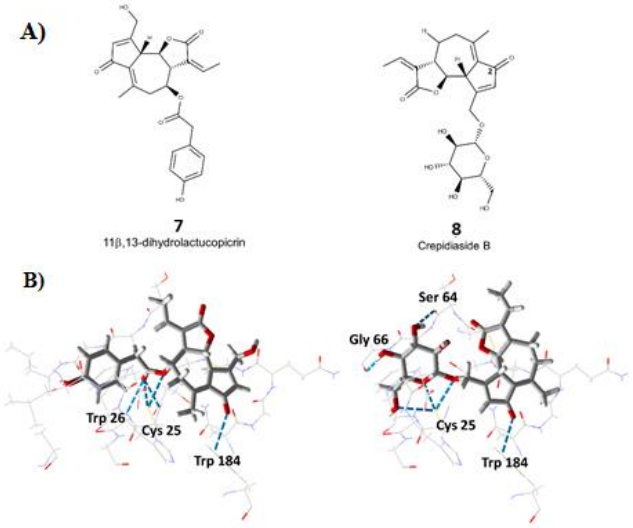

Figure 2. A) Structure of crepidiaside B (**7**) and 11β, 13-dihydrolactucopicrin (**8**). B) Docking conformations SL **7** and **8** in the pocket of *T. cruzi* Cruzain (PDB ID: 4XUI). The blue dotted line represents H-bond interactions between SLs **7** and **8** with Cruzain residues (black labels).

Ligand - based and Structure based VS combined approach.

Using the equation 3, an approach combining structure-based and ligand-based virtual screening was performed to verify potentially active molecules as well as their possible mechanism of action, facilitating the identification of potential multitarget compounds.

Equation 3:

$$p_c = \frac{p_s + (1 + TN)\ x\ p}{2 + TN}$$

where pc = combined probability p_s = structure based probability; TN = true negative rate; p = ligand-based probability

Table 2 summarizes the results for the best-ranked SLs obtained using the combined approach. Some structures that previously displayed a high active probability value in the ligand-based virtual screen appear to be interesting potential structures for each *T. cruzi* parasitic form.

Table 2. The best-ranked structures for each parasitic form obtained using an approach combining ligand-based and structure-based virtual screening.; p = active probability value in ligand-based VS; p_s = active probability value in structure-based VS. p_c = combined probability value

Cruzain				
Parasitic form	**Structure**	***p***	p_s	p_c
Amastigote	**1**	0.58	0.83	0.67
Trypomastigote	**4**	0.62	0.64	0.63
Epimastigote	**9**	0.73	0.91	0.79

Structure **1** and **4,** have the highest p_c values for amastigote and trypomastigote parasitic form, these two compounds also presented high probability scores in Ligand-based VS. Structure **9** (Figure 3), emerges as an interesting structure that acts in cruzain of epimastigotes, since that have good results in the two VS methodologies as well as in the combined-approach.

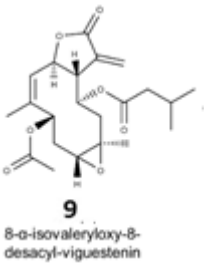

Figure 3. Structure of 8-α-isovaleryloxy-8-desacyl-viguestenin

3. Materials and Methods

From the ChemBL database were obtained 4,452, 1,635 and 1,322 structures with activity against the three parasitic forms of *T. cruzi*, amastigotes, trypomastigotes and epimastigotes, respectively (https://www.ebi.ac.uk/chembl/). The compounds were classified using values of pIC_{50} (-log IC_{50}), which led us to divide them into active ($pIC_{50} \geq 5$) and inactive ($pIC_{50} < 5$) structures.

For all structures including 1,306 SLs obtained from SistematX database, SMILES codes were used as input data in Marvin; ChemAxon (version 16.11.28 (2016), a calculation module developed by ChemAxon, http://www.chemaxon.com/). We used Standardizer software (Jchem, version 16.11.28 (2016), a calculation module developed by ChemAxon, http://www.chemaxon.com/); ChemAxon to canonize structures, add hydrogens, perform aromatic form conversions, and clean the molecular graph in three dimensions. After were calculated 128 3D-molecular descriptors in Volsurf+ software. Obtained results were imported to Knime 3.1.0 software (www.knime.org). All variables were submitted to autoscaling and after were partitioned to generate two groups, a training group composed by the 80% of the whole molecules set and a test group composed by the remaining 20%. Using a Random Forest algorithm, three models were performed. Models were evaluated through cross validation and a test set (20%). After a Ligand-based VS of the 1,306 SLs were performed in these models (Figure 4).

The structure of *T. cruzi* protein, Cruzain (PDB ID: 4XUI) in complex with the respective inhibitor (PDB ID: 2VC), were downloaded from the Protein Data Bank—PDB. The docking procedure was performed using MOLEGRO virtual docker 6.0, using a GRID with a radius of 15 Å and a resolution of 0.30 Å to cover the ligand-binding site in the structure of cruzain (Figure 4).

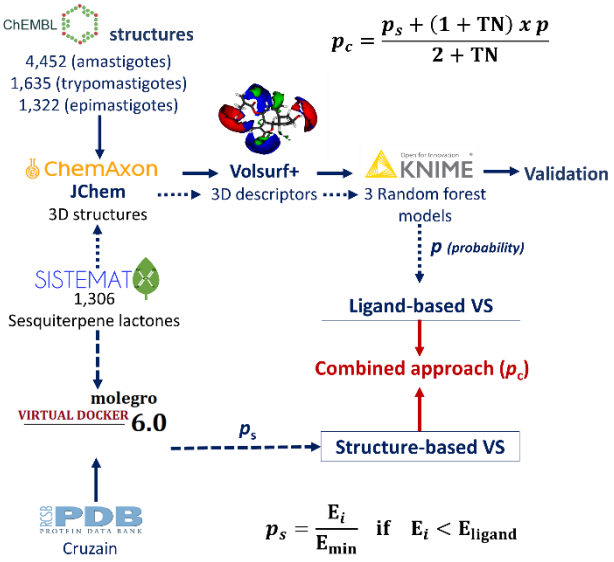

Figure 4. General scheme of the methodologies used to select potentially antichgasic compound through of a combined approach.

4. Conclusions.

In the present study, potential antichagasic SLs for the three parasitic forms and some structural features were determined from RF models of *T. cruzi*. In addition, a structure-based virtual screen using PDB structure of *T. cruzi* cruzain for the entire SL set allowed the selection of potential inhibitors of this enzyme. Finally, using a combined approach of structure-based and ligand-based VS enabled the identification of promising multitarget antichagasic SLs.

Acknowledgments
We would like to thank the Student Agreement Program of Graduate—PEC-PG of Brazilian National Research Council (CNPq)—Brazil

Author Contributions
LS built database; CHA performed all calculus; and CHA and MTS wrote the paper. All authors read and approved the final manuscript.

Conflicts of Interest
The authors declare no conflict of interest.

References and Notes

1. WHO. Available online: http://www.who.int/mediacentre/factsheets/fs340/en/, (accessed on 10 September 2017).
2. Newman, D. J.; Cragg, G. M. Natural products as sources of new drugs from 1981 to 2014. *J. Nat. Prod*, **2016**, *79*, 629-661.
3. Mishra, B. B.; Tiwari, V. K. Natural products: an evolving role in future drug discovery. *European journal of medicinal chemistry*, **2011**, *46*, 4769-4807.
4. Schmidt, T.J.; Brun, R.; Willuhn, G.; Khalid, S.A. Anti-trypanosomal activity of helenalin and some structurally related sesquiterpene lactones. *Planta Med*. **2002**, *68*, 750–751.
5. Herrera Acevedo, C.; Scotti, L.; Feitosa Alves, M.; Formiga Melo Diniz, M.D.F.; Scotti, M.T. Computer-Aided Drug Design Using Sesquiterpene Lactones as Sources of New Structures with Potential Activity against Infectious Neglected Diseases. *Molecules* **2017**, *22*, 79.