Integration and Data Mining in Drug Target Detecting for Schistossoma mansoni

Francimary Procopio Garcia[1], Kele Teixeira Belloze[1],

1 CEFET/RJ

Abstract

Classified as a neglected disease, in spite of it's acting in underdeveloped countries, schistosomiasis, caused by Schistossoma Mansoni, is considered one of the most important endemic diseases in the world, having an estimated number of around 240 million infected and 700 million people living in an area with a high risk of transmission. There's currently one sole drug recommended by the World Health Organization for schistosomiasis treatment which, although being effective in the elimination of the vermin, it shows collateral effects and can only act upon its mature form. Therefore, researching for new alternative drug targets in combating schistosomiasis is required. This work has as main objective the identification and classification of possible new drug targets for S. mansoni. The proposed methodology for the development of this work is described as follows. At first the identification of ortholog proteins between S. mansoni and three eukaryotic models organisms will be done: Caenorhabditis elegans (nematode), Saccharomyces cerevisiae (yeast) and Mus musculus (mouse), based on the concept of gene essentiality. Subsequently, the process of identifying homologous proteins between the S. mansoni proteins raised in the previous step and druggable proteins (targets for drugs), available at Drugbank and Therapeutic Target Database (TTD) databases, will be conducted. These two steps will result in an intermediate database composed of essential and druggable candidate proteins of the organism under study, represented by primary sequences integrated with the aggregate annotations during the accomplishment of these two steps of the methodology. From the candidate proteins raised, the research will proceed on identifying information of these protein's secondary structures, in order to enrich the database conceived. For this step a homology based approach will be adopted using the secondary protein structures available at Protein Data Bank (PDB). Data from this integrated database will be categorized using frequent patterns models such as Apriori to identify consistent behaviors among candidate proteins and provide them with an druggability index. A decision tree model will be used to identify the candidates with the highest combination weight and validated through cross validation functions, where the available data will be divided into two mutually exclusive subsets, one for training (parameter estimation) and another for testing (validation). The percentage of candidates prediction with the highest druggability will be calculated and their druggability indexes will be validated and discussed according to data obtained in the literature. As a result of this work, it is expected to obtain a list of S. mansoni proteins which may be indicated as drug targets, and thus contribute to the initial step of the drug development process. This work is in an initial phase of studies in which a literature review is being carried out.

Funding: CEFET/RJ