Poster I-19

A Trypanosoma cruzi Genome Study by Sequence-Structure HighThroughput Comparative Modelling: A First Step Towards the Identification of Potential Molecular Targets for Chagas Disease

Authors:
Priscila V. S. Z. Capriles (*LNCC/MCT*)
Shaila C. Rössle (*Ludwig - Maximilians - University of Munich*)
Ana C. R. Guimarães (*DBBM/FIOCRUZ*)
Marcos Catanho (*DBBM/FIOCRUZ*)
Paulo M. Bisch (*IBCCF/UFRJ*)
Wim Degrave (*DBBM/FIOCRUZ*)
Laurent E. Dardenne (*LNCC/MCT*)

Short Abstract: In this work, we investigated the Trypanosoma cruzi genome, by highthroughput comparative modelling, and identified/classified the subset of T. cruzi enzymes with theoretically determined 3D-structures (9.13% of the total genome) which can be investigated as potential molecular targets for structure based rational drug design researches.

Long Abstract:
The protozoa Trypanosoma cruzi is endemic in South and Central America and nearly 100 years after the discovery of the parasitic agent of Chagas disease there are no appropriate therapies that lead to cure the acute or the chronic phases of this disease.
In this work we present a highthroughput comparative modelling study of the Trypanosoma cruzi genome. The main objective is to identify the subset of Trypanosoma cruzi enzymes that can have their three dimensional structure determined by comparative modeling in order to be investigated as potential molecular targets for structure based rational drug design researches.
The following main topics were investigated and discussed: (i) which sequences can have their 3D-structures generated by comparative modelling; (ii) the quality associated to the models; (iii) the association of the modeled proteins with the different enzymatic classes and sub-classes; (iv) the use of the structural models to infer a possible biological function to hypothetical protein sequences and/or to infer a distinct function to already annotaded ones; (v) the percentage of transmembrane proteins.
To construct a list of possible molecular targets for Chagas disease, we used computational biology tools to track proteins with enzyme characteristics. To detect possible redundancies, it was performed a BLAST of 25,041 CDS (coding sequence) from T. cruzi against T. cruzi and the result was automated filtered (by the BioParser program - using a identity criterium <=95%) obtaining 20,679 sequences.
These 20,679 sequences were submitted to MHOLline, a workflow for automated large scale protein structure prediction and function annotation, using the following criteria of classification: (1) Very High Quality: models built based on templates that share a high

degree of sequence identity (>=75%) and coverage (>=90%), with the target; (2) High Quality: templates that shows high sequence indentity (>=50% and <75%) and coverage (>=90%), in relation to the target; (3) Good Quality: models built based on templates that share a relatively high degree of sequence identity (>=50%) and coverage (>=70% and <90%); (4) Medium to Good Quality: models based on medium sequence identity (>=35% and <50%) and coverage (>=70%); (5) Medium to Low Quality: models based on medium sequence identity (>=25% and <35%) and coverage (>=70%) and (6) Low Quality: models based on a low identity (>=25%) and coverage (>=50% and <70%) between the sequences.

The resulting 2,786 modeled sequences were used as entry data for a ECNumber program, that automatically find the enzyme classification number (ECN) for each sequence, using a pdb2sp.txt (from ftp://beta.rcsb.org/pub/pdb/uniformity/derived_data/ in january-2006) and enzyme.dat (from ftp://br.expasy.org/databases/enzyme/release_ with_updates/ in january-2006) data banks, getting 1,888 models with predicted enzymatic properties (Very High = 3, High = 96, Good = 49, Medium to Good = 652, Medium to Low = 674 and Low = 387 models).

These 1,888 models were classified in five categories: (I) Annotation's Matching (695 models): ECN's models matches the original CDS annotation; (II) Hypothetical (238 models): models which CDS was originally annotated as a hypothetical protein; (III) Surface Proteins (697 models): models which CDS was originally annotated as a surface protein or gp63 and (IV) Annotation's Doubt (258 models): ECN's models do not matches the original CDS annotation or produce some doubts.

Each category was analyzed according to the International Union of Biochemistry and Molecular Biology (IUBMB) enzyme nomenclature, presenting the following percentage: oxidorreductases (7.63%), transferases (24.47%), hydrolases (59.53%), lyases (2.17%), isomerases (2.86%) and ligases (3.34%). The MHOLline program was able to identify 4,719 sequences with transmembrane regions indicating that these sequences can be associated to membrane proteins.

The emerging sequences of this work will be further investigated by the Analogue Enzyme Pipeline (AnEnPi) (in development) to recognize which of these have or not similarity with human enzymes.