An evolutionary approach for the identification of a cocktail of drugs for the treatment of Chagas disease

Name: Víctor Andrés Yubero Rodríguez
Supervisor: Alberto Paccanaro **Co-supervisor:** Luca Cernuzzi

Contents

1 Introduction

Chagas disease is a life-threatening illness caused by the protozoan parasite called Trypanosoma cruzi (*T. cruzi*) [1]. According to World Health Organization, about 6 to 7 million

people worldwide are estimated to be infected with the parasite. Chagas disease is endemic mainly in areas of 21 Latin American countries, and it has spread to other continents over the last century due to increased travel and global population [2]. In Paraguay there is an estimate of 150.000 people affected with this disease [3]

The main mean of transmission is through an insect vector known as triatomine bug, which gets the parasite by biting and sucking blood from an infected host. When the contaminated bug feeds of human blood, it defecates on the victim, depositing *T. cruzi* together with its feces. The person rubs the injury introducing the stool in it, therefore, the parasite [1]. Otherwise, people also can become infected with Chagas through consumption of bad-uncooked food contaminated with the parasite, blood transmission from infected donors, congenital transmission from mother infected to newborn [4].

Chagas disease consists of two phases: the acute, and the chronic phase. The acute phase lasts for about two months after infection [2]. During the chronic phase, the parasites are hidden mainly in the heart and digestive muscles. Over the years, the infection can lead to affect the heart causing cardiomyopathy and dysrhythmias, and producing a sudden death, or the gut causing mega-oesophagus or mega-colon [5]. In this last phase, the infection with *T. cruzi* in human is lifelong.[6]

Chagas disease is currently treated using only two drugs in the acute phase: Nifurtimox and Benznidazole. Nonetheless, this phase is often asymptomatic or symptoms can be mild as a self-limiting febrile illness, swelling of lymph nodes among others manifestations that are not unique to Chagas disease. Due to above, it is hard to diagnose in this phase. Most of the times, the disease is detected once it has entered onto its chronic phase [7]. Nifurtimox and Benznidazole are low-effective for treating the chronic phase of infection [8].

In order to decrease spread of Chagas disease, control strategies are focused on preventing non-vector transmission [9][10]. In addition, vector-control programmes are important in areas where triatomine vectors are present indoors or peridomestic infestation still exists. Continuous application of modern pyrethroid insecticides has been necessary for control in endemic areas [9][6]. However, Chagas disease needs more treatment options. The two drugs used in the acute phase have failed to control the disease as both appear to have limited efficacy for treating the chronic phase of infection [1][8].

There is a noticeable lack of effective drugs against the *T. cruzi* in the chronic phase due to limited biological knowledge about it. However, the growing interest in this neglected disease has opened new horizons for this research [11].

This research will focus on trying to solve the problem of a lack of effective treatments for Chagas disease once entering in its chronic phase. We will develop a computational comparative genomics approach to identify cocktails of FDA (Food and Drugs Administration) approved drugs that could potentially work for the treatment of Chagas disease.

In the following section we will introduce biological concepts as well in order to understand appropriately the aim of our work.

2 Background

The aim in this work is to obtain sets of putative drug combinations to treat Chagas disease in the chronic phase through the development of a comparative genomics approach. The following section will introduce an overview of concepts needed for the necessary understating of our approach.

2.1 Sequencing

In genetic, sequencing refers to techniques used to determinate the primary structure of a large molecule produced by living organisms. It results in a symbolic linear of chemical structures description known as a sequence. Deoxyribonucleic acid (DNA) molecule is composed of four types called nucleotides. DNA sequencing means determining the order of these four nucleotides: Adenine (A), Thymine (T), Cytosine (C), and Guanine (G). Besides, ribonucleic acid (RNA) molecules are also a linear of four nucleotides but, unlike to DNA molecules, they are formed by Uracil (U) instead of Thymine (T) and the other three nucleotides [12].

The role of DNA is genetic instructions storage. In all living cells, these instructions are stored as a code made up of the four nucleotides and are transmitted generation to generation. The sequence determines the information available for building and maintaining an organism, similar to the way in which letters of the alphabet appear in a certain order to form words and sentences. The information required to represent whichever complex organism is stored on a number of DNA molecules. This collection of molecules is called the organism's genome. The latter, in turn, includes genes which are coding regions of DNA.

Protein molecules are built using the information encoded in DNA and also play a central role in biological processes on which life depends. They are composed of a chain of smaller molecule called amino acids. Thousands of different proteins are present in a cell and the synthesis of each type of protein being directed by a different gene. Proteins make up of the cellular structure (hair, skin, and fingernails consist largely of protein) [13].

There is a key relationship between DNA, RNA and proteins also known as central dogma of biology. There is a single direction of flow of genetic information from DNA, which acts as the information store, through RNA molecules that fulfills the function of working copy of DNA transferring the necessary genetic code for the creation for proteins [14].

Sequence analysis programs require particular format in which the sequence file must be represented. The FASTA format is the one most often used. The FASTA sequence format begins with a single-line description. The definition line is identified by a "$>$" symbol followed by the name and origin of the sequence. From the second line the sequence is defined. Sequences are expected to be represented in standard one-letter symbols (nucleotides or amino acids). It is recommended that all text lines have a length less than 80 characters.

As we can see, all this process can be associated as a flow of information which can be dealt them through computational tools.

2.2 Comparative Genomics

Once a nucleotide or amino acid sequence has been assembled, comparative genomics can be used to determine if the sequence is similar to that of a known gene.

Comparative genomics is a field of biological research focused on the comparison of the genome sequence of one organism against that of another to gain a better understanding of how species have evolved and to determine the function of gene regions in genome. One can extract a great deal of information from such analyses that is a great value in evolutionary biology. However, not only evolutionary relationship between organisms can trace out using comparative genomics, but also internal differences and similarities between species [15].

Common features of two organisms will often be encoded within the DNA that is conserved between the species. More precisely, the DNA sequences encoding the proteins and RNAs responsible for functions that were conserved from the last common ancestor should be preserved in contemporary genome sequences. Understanding the ancestry of the functional features compared is essential to our understanding and applications of genome comparison [16].

The next two sections will explain the central computational methods that we will use to carry out comparative genomics on *T. cruzi*. The aim throughout is to seek opportunities

arising from the increasing availability of whole-genome data for related organisms [17].

2.3 Sequence Alignment

Comparing sequences have never been a simply task. The difficulties arise because DNA and protein sequences can change during evolution. Nucleotides at originally corresponding positions, and the amino acids they encode, can change as a result of point mutation. Also, even the sequence lengths can be quite different as result of evolution. Such changes, in gene sequence and length can effectively mas any underlying sequence similarity. In order to reveal similarities between sequences, these must be aligned each other. The sequence alignment consists of finding the best way to pair (match) two sequences, so that there is a maximum correspondence between the nucleotides or amino acids. To do this, one of the sequences needs to be run with relation to the other to find the position where the maximum matches are given. It seeks to achieve when comparing sequences is to line them up in such a way that, when they do derive from a common ancestor, nucleotides or amino acids derived from the same ancestral are aligned [17][18].

For a better understating of how sequence alignment works, take the two hypothetical amino acids sequences THISSEQUENCE and THATSEQUENCE. If we align them so that as many identical letter as possible pair up we get

```
T H I S S E Q U E N C E
| |       | | | | | | |
T H A T S E Q U E N C E
```

Where the letters in red type are identical. As we can see with short and similar sequence, this alignment clearly identifies their strong similarity to each other.

Now, when sequences become more different from each other, they become more difficult to compare. For instance, how do we compare the two sequences THATSEQUENCE and THISISASEQUENCE, in which a mutation has led to the insertion of the tree amino acids, I, S and A into one of the original sequences?. Lining them up from the beginning loses much of the similarity we can see exists. Because of the difference in length, it also creates false matches between non-corresponding positions. To treat this problem, gaps are introduced into one or

```
T H A T S E Q U E N C E
| |           |       |
T H I S I S A S E Q U E N C E
```

both of the sequences so that maximum similarity is preserved. There is never just one possible

```
T H I S I S A - S E Q U E N C E
| |         | | | | | | | | |
T H - - - - A T S E Q U E N C E
```

alignment between any two sequences, and the best one is not always obvious.

The fundamental question is identifying whether the similarities observed between two sequences are due to chance or whether they are due to the derivation of the sequences from a common ancestral sequence, and are thus homologous.

Firstly, it must be mentioned the differences between the terms "homology" and "similarity". Similarity is simply a descriptive term telling you that the sequences in question show some degree of match. Homology, in contrast, has distinct evolutionary and biological implications. It is generally defined as referring specifically to similarity in sequence or structure due to descent from a common ancestor. Homologues genes are therefore genes derived from the same ancestral gene [19].

Before aligning two or more sequences, a choice about their most likely evolutionary relationship has to be made. Whether this relationship has preserved homology across the entire length of the sequences, we have to use a global alignment algorithm. However, whether only sub-regions of the sequences are homologues, a local alignment algorithm should be used.

Global algorithm such as Needleman-Wunsch algorithm and local algorithm such as Smith-Waterman are optimal in the sense that they are guaranteed to return the best possible solution. However, in keeping with a trend in computational biology to consider evolutionary events relevant on the scale of long sequences, these algorithms are computationally costly.

With the rapid growth of sequence databases, interest in faster or heuristic alignment procedures has increased. The most popular heuristic local pairwise alignment algorithm is BLAST (Basic Local Alignment Search Tool) [20] and BLAST e-value is used for determining cut-offs. The e-value is a statistical parameter reflecting the probability of finding a similar sequence in the database. The e-value takes into consideration the size of proteins (or nucleotides), as well as databases being searched. This statistical parameter is very powerful, and also the most used [17].

2.4 Phylogenetic Tree and 18S rRNA

Phylogenetic tree is a representation of the hierarchical relationships among organisms arising through evolution. It is comprised by nodes linked together by branches. The nodes represent taxonomic units, more specifically, the terminal nodes represent known sequences from organisms, whereas the internal nodes represent hypothetical ancestors. These are classified and divided into rank-based taxonomic categories. Each internal node is attached to one branch representing evolution from its ancestor, and two or more branches representing its descendants. The tree topology gives us the information on the order of relationships [21][22].

The use of molecular data in phylogenetic brought with it a great revolution. The access to DNA sequences increased the number of homologous characters that could be compared. A few genes became reference markers. In particular, owing to its considerable degree of conservation across all organisms, the gene that encodes small subunit ribosomal RNA (SSU rRNA) was extensively used for the classification of microorganisms.

In the eukaryotic kingdom, the analysis is based on the small subunit 18S rRNA, as well as on the construction of the phylogenetic tree the said kingdom. It is reasonable to believe that two organisms that have similarities between their 18S rRNA, will have a very similar cellular machinery [23][24].

2.5 Metabolic Pathways

Here we will also introduce two essential concepts to follow up our work. One of the important characteristics that we compare and study in this approach are the metabolic pathways and also the enzymes.

A metabolic pathway is a coordinated sequence of chemical reactions by which a living organism transforms an initial source molecules or substrates, such as sugar, into different, more readily usable materials. These reactions occur inside of a cell.

On the other hand, enzymes are proteins catalysts in charge of the chemical reactions occurring within the metabolic pathway. In each step of the pathway, the enzymes convert molecules into products by attaching or breaking off chemical groups from molecules. In the simplest sense, enzymes are similar to traffic lights in that they can slow, speed up, and stop metabolic processes [25].

In order to give a better understanding, in the figure bellow we can see the metabolic pathway responsible for the production of Xylitol, also known as sugar alcohol. Xylitol is

usually added in chewing gum because its fermentation does not produce acids, therefore it is not cariogenic. Inside the cell is the metabolic pathway. Glucose enters and after a few reactions we see that Xylitol comes out. Enzymes carry out the chemical reactions (represented by ellipses) [26].

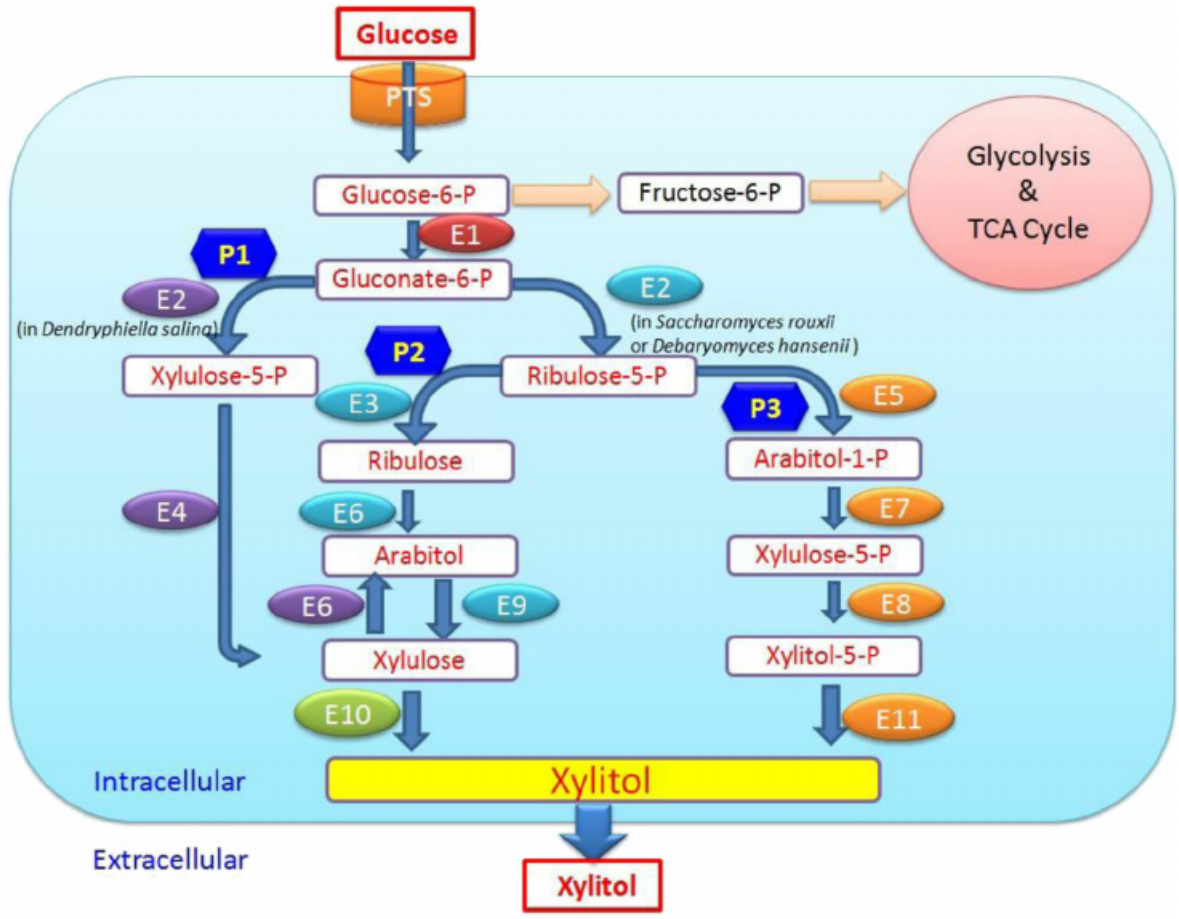

Figure 1

2.6 Databases

Databases in general are classified into primary, secondary and composite databases. A primary database contains information of the sequence or structure. They are Genbank, maintained by the National Center for Biotechnology Information (NCBI), USA, The EMBL. The secondary database contains derived information from primary database. Composite database amalgamates a variety of different primary data sources, which obviates the need to search multiple searches [27][28].

We will address databases that will be of our concern. These would be sequence, phylogenetic, and metabolic pathways databases, as well as drug information databases which helps us in the search of putative drugs for the treatment.

2.6.1 Sequence Databases

Nucleotide and amino acid sequence databases represent the most widely used. There are three main databases that store and available raw nucleotide and amino acid sequences. Genbank, maintained by the National Center for Biotechnology Information (NCBI) located in USA. European Molecular Biology Laboratory (EMBL) which is in UK and the DNA databank

of Japan (DDJ), in Japan. They have uniform data formats and exchange data on daily basis. See table 1 and table 2.

Database	URL	Feature
GenBank	http://www.ncbi.nlm.nih.gov/	NIH's archival genetic sequence database
EMBL	http://www.ebi.ac.uk/embl/	EBI's archival genetic sequence database
DDBJ	http://www.ddbj.nig.ac.jp/	NIG's archival genetic sequence database
SGD	http://www.yeastgenome.org/	A repository for baker's yeast genome and biological data
EBI genomes	http://www.ebi.ac.uk/genomes/	It provides access and statistics for the completed genomes
Ensembl	http://www.ensembl.org/	Database that maintains automatic annotation on selected eukaryotic genomes
UniGene	http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene	Each UniGene cluster contains sequences that represent a unique gene, as well as related information
dbEST	http://www.ncbi.nlm.nih.gov/dbEST/	Division of GenBank that contains expression tag sequence data

Table 1: Summary of Nucleotide Sequence Databases.

Database	URL	Feature
Swiss-Prot/TrEMBL	http://www.expasy.org/sprot/	Description of the function of a protein, its domains structure, post-translational modi cations etc,
UniProt	http://www.pir.uniprot.org/	Central repository for PIR, Swiss-Prot, and TrEMBL

Table 2: Summary of Protein Sequence Databases.

2.6.2 Pathway Databases

These databases are derived from the comparative study of metabolic pathways. Metabolic pathway databases may draw on enzyme, compound and gene databases, but usually go beyond them to show how enzyme reactions link to one another. In addition, these may contain information on pathways shared by many organisms. See table 4

Database	URL	Feature
KEGG	http://www.genome.jp/kegg/	Protein structure repository that provides tools for analyzing these structures
BioCyc	http://www.biocyc.org/	Classi cation of protein 3D structures in a hierarchical scheme of structural classes
BRENDA	http://www.brenda-enzymes.org/	Hierarchical classi cation of protein domain structure
EMP	http://emp.mcs.anl.gov/	Database of Enzymes and Metabolic pathways public server
BRITE	http://www.genome.jp/kegg/brite.html	Biomolecular Relations in Information, Transmission and Expression

Table 3: Summary of Pathway Databases.

2.6.3 Drug Interaction Databases

These databases focus directly on known drugs or drug metabolites and attempt to link the genomic or proteomic information being gathered about the relevant genes or proteins with the drug compounds themselves.

Database	URL	Feature
DrugBank	https://www.drugbank.ca/	It is a bioinformatics/cheminformatics resource that combines detailed drug data with comprehensive drug target information
TTD	http://bidd.nus.edu.sg/group/cjttd/	Therapeutic Target Database
Thomson Pharma	https://www.thomson-pharma.com/	

Table 4: Summary of Drug Databases.

3 State of the Art

In our present work, we are studying the literature in order to analyse comparative genomic techniques employed to approach our problematic. Literature searches were made in PubMed (https://www.ncbi.nlm.nih.gov/pubmed/), Google Scholar (https://scholar.google.com.py) and

Nature (https://www.nature.com/). Besides, it is worth to emphasize constant collaboration and knowledge share on the subject by members of PaccanaroLab from the RHUL (Royal Holloway University of London) and a team of biologists from the CEDIC (Centro para el Desarrollo de la Investigación Científica).

3.1 Comparative Genomics in Drug Discovery

3.1.1 At Organism Level

The search for appropriate vaccine candidates and drug targets against some disease-causing parasites has so far been confronted with several limitations due to the unavailability of biological material, appropriate molecular resources, and knowledge of the parasite biology.

Many researchers have resorted to use of comparative genomics methods in order to obtain more biological knowledge about an organism. Jane M. Carlton et al has described a comparative genomic method to show that the parasite resembles other malaria parasite in gene content and metabolic potential. Such a way to shed light on its distinctive biological features, and as a means to find potential drugs. This because of the lack of study that exists about parasite Plasmodium vivax, the human malaria parasite [29].

Also, Michelle Lizotte-Waniewski et al has used this methods for the purpose of establishing a solid foundation for a better understanding of the biology of O. volvulus as well as for the identification of novel targets or vaccines against onchocerciasis based on immunological and rational hypothsis-driven research [30].

3.1.2 At Metabolic Pathway Level

Many papers have been published to address the comparative genomics and metabolic pathways analysis with a predefined computational systemic workflow. This workflow consists of identifying of novel therapeutic candidates against most specifically bacteria [31][32][33]. At first, obtaining the whole genome and proteome of the organism and identifying metabolic pathways unique to pathogen and common to pathogen and host. A manual comparison was then made to determine which metabolic pathways did not appear in the host but were present in the pathogen, according to the KEGG database annotations. Proteins from common and unique pathways were identified and the respective amino acid sequences were obtained from the Swiss-Prot database. After all that, it comes a very important aspect to select a potential drug candidate. At first, only proteins from the pathogen-specific metabolic pathways were subjected to BLAST analysis. Then, proteins from common metabolic pathways were also compared by BLAST analysis. In each scenario, searching was restricted to proteins from humans only through an option available under BLAST parameters. Hits were filtered on the basis of expectation value inclusion threshold being set to 0.005. Proteins, that did not have hits below the e-value inclusion threshold of 0.005, were picked out as non-homologous proteins. The BLAST search aligning the non-homologous essential proteins to the list of drug targeted proteins downloaded from DrugBank was used to examine the druggability of each of the non-homologous essential proteins.

3.2 Comparative Genomics in T. Cruzi and Drug Discovery

There are currently a few applications of comparative genomics on t. cruzi. In 2005, drafts sequences of the genomes of Trypanosoma brucei, Trypanosoma cruzi and Leishmania major, also known as the Tri-Tryp genomes, were published. As result of this, extensive analysis of these genomic sequences has greatly increased our understanding of the biology of these parasites and their host-parasite interactions.

Teixeira et al [34] provides us the recent advances in the comparative genomics of these three species. The identification of more than 8000 new protein-coding genes, many of which are shared between the Leishmania and Trypanosoma genera, vastly expands the potential drug targets available for investigation. Additional data about sequences derived from T. cruzi CL Brener, which was the strain chosen for the initial T. cruzi genome project, as well as functional genomics.

Additional data about sequences derived from T. cruzi CL Brener, which was the strain chosen for the initial T. cruzi genome project, as well as functional genomics. In addition to facilitating the identification of key parasite molecules that may provide a better understanding of these complex disease, genome studies offer a rich source of new information that can be used to define potential new drug targets for controlling these parasitic infections.

Hughes et al [24] show us a phylogenetic analysis of 18S rRNA sequences from the families Trypanosomatidae and Bodonidae. They aligned 18S rRNA sequences using the ClustalW program and phylogenetic trees were constructed by four methods: the minimum evolution methods; the maximum parsimony method; the quartet maximum likelihood method; and the Bayesian method.

Gregory J. Crowther et al [35] provides strategies to prioritize pathogen proteins based on whether their properties meet criteria considered desirable in a drug target. These criteria are based upon both sequence-derived information and functional data on expression, essentially, phenotypes and metabolic pathways.

In the following article, Ekins et al [36] has constructed a Pathway Genome Database for T. cruzi. This means that they have inferred metabolic pathways present in T. cruzi and also, they have identified which enzymes are in each metabolic pathway.

With the study of the literature, we believe that it much remains to be discovered in the area of comparative genomics in T. cruzi in a way to discover an effective treatment for Chagas disease. Retrieving all data mentioned from previous studies, give us broad landscape to implement our proposal looking for a effective treatment of Chagas disease.

4 Solution Proposal

We will calculate evolutionary similarity measures that will allow us to transfer existing knowledge from well-studied organisms to *T. cruzi*. Evolutionary similarities will be calculated at two levels:

4.1 At Organism Level

Our idea is that if some drugs affect organisms similar to the *T. cruzi*, then they could also be reused for Chagas Disease. The reasoning is that the drug will continue to be effective if the drug's mechanism of action can be preserved, therefore, similar organisms should possess similar binding targets and comparable metabolic pathways, that when disrupted by the drug, will render the organism to be harmless for the host.

The first step is to define which organisms are similar and close to the *T. cruzi* following a comparative genomic-approach. We decided upon two ways to collect organisms. The first one is to compare 18S rRNA sequences of eukaryotic organisms against the *T. cruzi*'s sequence, and the second one is to select organisms closely related to the *T. cruzi* using the Eukaryota phylogenetic tree [37].

An 18S rRNA is used to establish evolutionary relationship between species. Ribosomal RNA is presumed to date back to the earliest forms of life showing very little change throughout time. To obtain a list of similar organisms, by comparing their 18S rRNA sequences, we will

use BLAST. It will allow us to rank the resulting organisms by quantifying their similarity to the *T. cruzi*.

In addition, we will explore the Eukaryota phylogenetic tree. The nearest organisms to the *T. cruzi* found in the tree have a very close evolutionary relationship. To obtain a list, we have to find all the children of a common ancestor located a few levels up the tree. The resulting organisms can only be considered useful as long as the number of levels we climb the tree is very small. This is because the higher we go to find a common ancestor, we will be including organisms that could be very far from the *T. cruzi* and thus lowering our chances of finding similarities at organism level. Finding the closest organisms in the tree does not give us any measure of how similar they are to the *T. cruzi*, so the list we obtain will not be ranked.

Once the list of organisms is extracted, the next step in our approach is to find the drugs. Drugbank database containing information about drugs, their mechanisms, their groups and their targets [38]. Almost every drug in the database has specific organisms which it affects.

Finally, in order to obtain a list of drugs, we perform a mapping between every drug's affected organisms and all similar and close relative to the *T. cruzi* we extracted.

4.2 At Metabolic Pathway Level

Based on available biological data, we will begin by inferring which metabolic pathways occur in *T. cruzi*. In addition, we will collect drug target information of FDA approved drugs. Furthermore, we will identify similar proteins between drug-targeted proteins and proteins of *T. cruzi* using a homology approach. The idea in this level is to find a set of drugs that could show antitrypanosomal activity by potentially disrupting metabolic pathways of *T. cruzi*.

Metabolic pathways explain the biological processes occurring within organisms. If we can find drugs that bind to pathway-specific proteins as their mechanism of action, then we could be disrupting biological processes and potentially kill the organism.

5 Advances

5.1 Mining Organisms

5.1.1 18S rRNA Sequences

Our primary source was the SILVA rRNA Database Project [37], from which we extracted a quality checked dataset of 18S ribosomal RNA sequences in FASTA format from Eukaryotic organisms. As the reference sequence for BLASTN, we used the *Trypanosoma Cruzi Clone CL Brener*, which is the reference organism used in the *Trypanosoma Cruzi Genome Porject* [39]. Using the 18S rRNA dataset from SILVA, we created the BLASTN database that contains 363.948 sequences in total. After finding the score for each alignment between the reference sequence and the sequences in the BLASTN database, the results are sorted by e-value, the most significant hits appearing at the top. The results obtained from the search contains 4715 different organisms that are highly similar to the *T. cruzi* ranked by their corresponding e-value. Table 5 shows some example results from the BLASTN search. The e-values for all of these organisms are very close to zero, that makes these results very significant and promising.

5.1.2 Eukaryota Phylogenetic Tree

The other technique to mine the organisms is to extract the nearest organisms to the *T. cruzi* in the Eukaryota phylogenetic tree. We used the SILVA's Eukaryotic Taxonomy by the Eukaryotic Taxonomy Working Group (ERWG) [37], which is an improved and unified taxonomy for Eukaryota. It provides a taxonomic hierarchy for eukaryotic organisms. The

e-value	Organism	e-value	Organism
0.00	Koruga bonita	8.00E-84	Candida
0.00	Trichomonas vaginalis	3.00E-58	Escherichia coli
0.00	Trichonympha magna	5.00E-55	Apicomplexa
0.00	Pseudotrichonympha hertwigi	5.00E-55	Ochroconis sexualis
0.00	Pseudotrichomonas keilini	5.00E-55	Venturia inaequalis
0.00	Trichomitopsis minor	5.00E-55	Ochroconis lascauxensis
0.00	Tetratrichomonas limacis	2.00E-54	Sorosphaerula viticola
0.00	Hypotrichomonas mariae	6.00E-54	Helicosporidium
0.00	Simplicimonas moskowitzi	6.00E-54	Curvularia
0.00	Pseudotrichonympha	6.00E-54	Mniaecia jungermanniae
0.00	Snyderella yamini	7.00E-53	Operculina
0.00	Pseudotrichomonas	7.00E-53	Korotnevella
0.00	Histomonas meleagridis	4.00E-50	Pluteus granulatus
0.00	Hypotrichomonas	4.00E-50	Korotnevella hemistylolepis
0.00	Trichomonas equibuccalis	4.00E-50	Bankera fuligineoalba
0.00	Joenia	4.00E-50	Ganoderma sinense

Table 5: Examples of BLASTN Results of Similar Organisms to the *Trypanosoma Cruzi*

used dataset was the taxonmy mapping (taxmap) for eukaryotes from SILVA's release 123. The way to extract the organisms we want, is to navigate the tree using the taxonomy mapping and specifically find all the interior nodes that match with the taxonomic unit: Trypanosoma. Afterwards we climb k levels up the tree and find all the children of the found node (common ancestor), including the ones of the Trypanosoma. The number k of levels to climb the tree is arbitrary but needs to be small to keep a close relationship between the resulting organisms. For these results, we used k = 2. The resulting list has 371 organisms close to the *T. cruzi* in the tree, with k = 2. It is important to remember that we only care about the topology of the tree to state a relationship. We are not quantifying the distance between each pair of organisms. Consequently, the list is unranked. See Table 6 for some examples.

Organism	Organism
Trypanosoma corvi	Trypanosoma copemani
Trypanosoma conorhini	Trypanosoma congolense
Trypanosoma cobitis	Trypanosoma chelodinae
Trypanosoma chattoni	Trypanosoma cascavelli
Trypanosoma brucei	Trypanosoma boissoni
Phytomonas	Phanerobia pelophila
Phacus warszewiczii	Phacus triqueter
Phacus trimarginatus	Lepocinclis acus
Leishmania turanica	Leishmania tropica
Leishmania major	Leishmania infantum
Leishmania guyanensis	Leishmania gerbilli
Euglena viridis	Euglena velata
Euglena tristella	Euglena stellata

Table 6: Example Organisms close to the *Trypanosoma Cruzi* in the *Eukaryota* Phylogenetic Tree

5.2 Mining Drugs

The final stage of this methodology is to find drugs as candidates for the treatment of Chagas Disease. The main source of data for this stage was DrugBank. The database we used was the 4.3 version, released in February 2016. It contains information about 8203 drugs, out of wich 2027 drugs have information about the affected organism. An extensive search was performed to select the drugs. We have searched whether any of the organisms from the lists we generated, were found among any of the drug's affected organisms. Every time there is a match, the drug name is extracted, along with the information about the affected organisms and the drug's group or groups The result is a filtered list of drug that affect certain organisms. These affected organisms are very similar or relative close to *T. cruzi*. These organisms could also have comparable cellular machinery with the *T. cruzi*. Therefore, all of these drugs are potential candidates for the treatment of Chagas Disease.

5.3 Preliminary Results

The results so far consist in a list of 24 different drugs obtained from 4 similar organisms to the *T. cruzi*. See Table 7. Almost every result we obtained, 3 out of the 4 organisms and 23 out of the 24 drugs, was from the correspondence with the BLAST results, whereas only 1 organism and 1 drug, Trypanosoma brucei and Eflornithine (bolded in in Table3), with the Eukaryota phylogenetic tree results. Figure 2 shows a schematic diagram of our approach.

The evaluation of these drugs is complex, requiring experiments to obtain significant conclusions. In vitro test are scheduled, to be performed by the CEDIC (Centro para el Desarrollo de la Investigación Científica), as part of the project which encloses this work as well. However, the list we obtain is encouraging, as it contains a drug which has been suggested in the past as a potential cure for Chagas disease. Even if the drug was later discarded because it failed in late stage trials, it is extremely encouraging to see that our comparative genomics approach has been able to identify it.

The drug in question is Posaconazole. Posaconazole has shown desructive activity against the *T. cruzi* in murine models. In the acute phase of Chagas disease, Posaconazole cured up to 90% of animals infected with *T. cruzi*. In the model of chronic Chagas disease, the differences were even greater: Posaconazole was associated with cure rates of up to 60% in animals infected with *T. cruzi* [40][41].

This results are presented as a pre-selection of drugs, they still need a more thorough evaluation. More information about the drugs should be further examined: toxicity, side effects, mechanisms of action.

The fact that Posaconazole, a drug that has been a very promising candidate for many years, was found among our results, demonstrates that the simple methodology we are presenting could indeed be useful in finding reusable drugs and also save researchers a lot of time and resources. It also gives more validity to the other found drugs.We also presented a database of highly similar and closely related organisms to the *T. cruzi*.

Organism	Drug Name
Trichomonas vaginalis	Tinidazole
Candida	Posaconazole
Candida	Efinaconazole
Candida	Isavuconazonium
Candida	Caspofungin
Candida	Silver sulfadiazine
Candida	Sulfanilamide
Candida	Anidulafungin
Candida	Micafungin
Escherichia coli	Arbekacin
Escherichia coli	Carbenicillin
Escherichia coli	Avibactam
Escherichia coli	Sulfamethoxazole
Escherichia coli	Ceftriaxone
Escherichia coli	Cefotaxime
Escherichia coli	Loracarbef
Escherichia coli	Pentostatin
Escherichia coli	Amdinocillin
Escherichia coli	Tetracycline
Escherichia coli	Doripenem
Escherichia coli	Neomycin
Escherichia coli	Gramicidin D
Escherichia coli	Trimethoprim
Trypanosoma brucei	**Eflornithine**

Table 7: Preselection of Drugs to be Reused for the Treatment of Chagas Disease

References

[1] J. R. Coura and P. A. Viñas, "Chagas disease: a new worldwide challenge," *Nature*, vol. 465, no. n7301_supp, pp. S6–S7, 2010.

[2] W. H. Organization, "Chagas disease fact sheet no. 340." Available at: http://www.who.int/mediacentre/factsheets/fs340/en/, Marc 2017. Accessed: 20.03.2017.

[3] O. P. de la Salud, "Artículo sobre la enfermedad de chagas en paraguay." Available at: http://www.paho.org/par/index.php?option=com_content&view=article&id=677:enfermedad-chagas-calcula-50-000-nuevos-casos-ano-america-150-000-personas-infectadas-paraguay&Itemid=258. Accessed: 15.05.2017.

[4] J. A. Urbina, "Specific chemotherapy of chagas disease: relevance, current limitations and new approaches," *Acta tropica*, vol. 115, no. 1, pp. 55–68, 2010.

[5] A. R. Teixeira, M. M. Hecht, M. C. Guimaro, A. O. Sousa, and N. Nitz, "Pathogenesis of chagas' disease: parasite persistence and autoimmunity," *Clinical microbiology reviews*, vol. 24, no. 3, pp. 592–630, 2011.

[6] A. Prata, "Clinical and epidemiological aspects of chagas disease," *The Lancet infectious diseases*, vol. 1, no. 2, pp. 92–100, 2001.

[7] A. Rassi and J. A. Marin-Neto, "Chagas disease," *The Lancet*, vol. 375, no. 9723, pp. 1388–1402, 2010.

[8] J. Clayton, "Chagas disease: pushing through the pipeline," *Nature*, vol. 465, no. n7301_supp, pp. S12–S15, 2010.

[9] A. Rassi, J. P. Dias, and J. A. Marin-Neto, "Challenges and opportunities for primary, secondary, and tertiary prevention of chagas' disease," *Heart*, vol. 95, no. 7, pp. 524–534, 2009.

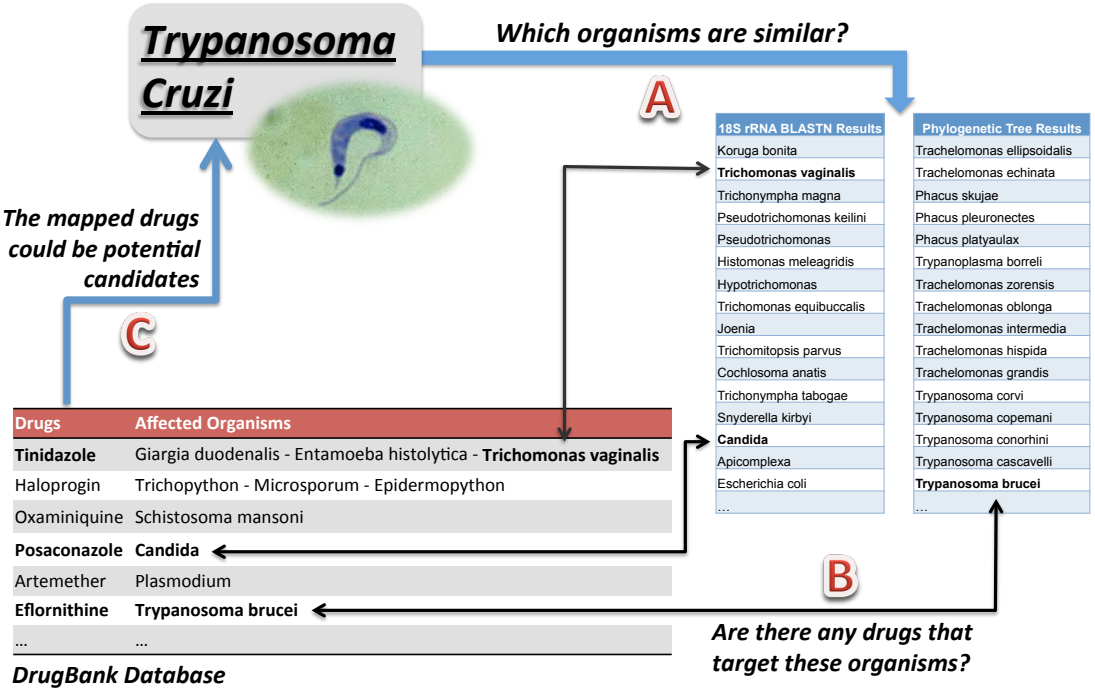

Figure 2: **A)** Get a list of similar organisms using BLASTN results of 18S rRNA sequences, and another list of close organisms in the *Eukaryota* phylogenetic tree. **B)** Look at the DrugBank database and find drugs that target the organisms we found. **C)** The mapped drugs are preselected as potential candidates to be reused against the *Trypanosoma Cruzi*.

[10] J. R. Coura and J. C. P. Dias, "Epidemiology, control and surveillance of chagas disease: 100 years after its discovery," *Memórias do Instituto Oswaldo Cruz*, vol. 104, pp. 31–40, 2009.

[11] E. Chatelain, "Chagas disease drug discovery: toward a new era," *Journal of biomolecular screening*, vol. 20, no. 1, pp. 22–35, 2015.

[12] J. O. B. Marketa J. Zvelebil, *Understanding Bioinformatics*, ch. The Nucleic Acid World, pp. 3–24.

[13] *Genomes. 2nd edition*, ch. Understanding a Genome Sequence. 2002.

[14] F. Crick, "Central dogma of molecular biology," *Nature*, vol. 227, no. 5258, pp. 561–563, 1970.

[15] S. Sivashankari and P. Shanmughavel, "Comparative genomics-a perspective," *Bioinformation*, vol. 1, no. 9, p. 376, 2007.

[16] R. C. Hardison, "Comparative genomics," *PLoS Biol*, vol. 1, no. 2, p. e58, 2003.

[17] B. Haubold and T. Wiehe, "Comparative genomics: methods and applications," *Naturwissenschaften*, vol. 91, no. 9, pp. 405–421, 2004.

[18] W. R. Pearson, "An introduction to sequence similarity ("homology") searching," *Current protocols in bioinformatics*, pp. 3–1, 2013.

[19] J. O. B. Marketa J. Zvelebil, *Understanding Bioinformatics*, ch. Producing and Analyzing Sequence Alignments, pp. 71–114.

[20] F. Altschul, Stephen, W. Gish, *et al.*, "Basic local alignment search tool," *J. Mol. Biol.*, 1990.

[21] F. Delsuc, H. Brinkmann, and H. Philippe, "Phylogenomics and the reconstruction of the tree of life," *Nature Reviews Genetics*, vol. 6, no. 5, pp. 361–375, 2005.

[22] S. Whelan, P. Liò, and N. Goldman, "Molecular phylogenetics: state-of-the-art methods for looking into the past," *TRENDS in Genetics*, vol. 17, no. 5, pp. 262–272, 2001.

[23] A. K. Bachhawat, "Comparative genomics-a powerful new tool in biology,"

[24] A. L. Hughes and H. Piontkivska, "Phylogeny of trypanosomatidae and bodonidae (kinetoplastida) based on 18s rrna: evidence for paraphyly of trypanosoma and six other genera," *Molecular biology and evolution*, vol. 20, no. 4, pp. 644–652, 2003.

[25] K. Faust, D. Croes, and J. van Helden, "Prediction of metabolic pathways from genome-scale metabolic networks," *Biosystems*, vol. 105, no. 2, pp. 109–121, 2011.

[26] H. Cheng, J. Lv, H. Wang, B. Wang, Z. Li, and Z. Deng, "Genetically engineered pichia pastoris yeast for conversion of glucose to xylitol by a single-fermentation process," *Applied microbiology and biotechnology*, vol. 98, no. 8, pp. 3539–3552, 2014.

[27] H. Mewes, R. Doelz, and D. George, "Sequence databases: an indispensible source for biotechnological research," *Journal of biotechnology*, vol. 35, no. 2-3, pp. 239–256, 1994.

[28] N. Toomula, A. Kumar, D. Kumar, and V. Bheemidi, "Biological databases-integration of life science data," *J. Comput. Sci. Syst. Biol*, vol. 4, pp. 87–92, 2012.

[29] J. M. Carlton, J. H. Adams, J. C. Silva, S. L. Bidwell, H. Lorenzi, E. Caler, J. Crabtree, S. V. Angiuoli, E. F. Merino, P. Amedeo, *et al.*, "Comparative genomics of the neglected human malaria parasite plasmodium vivax," *Nature*, vol. 455, no. 7214, pp. 757–763, 2008.

[30] M. Lizotte-Waniewski, W. Tawe, D. B. Guiliano, W. Lu, J. Liu, S. A. Williams, and S. Lustigman, "Identification of potential vaccine and drug target candidates by expressed sequence tag analysis and immunoscreening of onchocerca volvulus larval cdna libraries," *Infection and immunity*, vol. 68, no. 6, pp. 3491–3501, 2000.

[31] S. Ghosh, J. Prava, H. B. Samal, M. Suar, and R. K. Mahapatra, "Comparative genomics study for the identification of drug and vaccine targets in staphylococcus aureus: Mura ligase enzyme as a proposed candidate," *Journal of microbiological methods*, vol. 101, pp. 1–8, 2014.

[32] N. Batool, M. Waqar, and S. Batool, "Comparative genomics study for identification of putative drug targets in salmonella typhi ty2," *Gene*, vol. 576, no. 1, pp. 544–559, 2016.

[33] P. Chawley, H. B. Samal, J. Prava, M. Suar, and R. K. Mahapatra, "Comparative genomics study for identification of drug and vaccine targets in vibrio cholerae: Mura ligase as a case study," *Genomics*, vol. 103, no. 1, pp. 83–93, 2014.

[34] S. M. Teixeira, R. M. C. d. Paiva, M. M. Kangussu-Marcolino, and W. D. DaRocha, "Trypanosomatid comparative genomics: contributions to the study of parasite biology and different parasitic diseases," *Genetics and molecular biology*, vol. 35, no. 1, pp. 1–17, 2012.

[35] G. J. Crowther, D. Shanmugam, S. J. Carmona, M. A. Doyle, C. Hertz-Fowler, M. Berriman, S. Nwaka, S. A. Ralph, D. S. Roos, W. C. Van Voorhis, *et al.*, "Identification of attractive drug targets in neglected-disease pathogens using an in silico approach," *PLoS neglected tropical diseases*, vol. 4, no. 8, p. e804, 2010.

[36] S. Ekins, J. L. de Siqueira-Neto, L.-I. McCall, M. Sarker, M. Yadav, E. L. Ponder, E. A. Kallel, D. Kellar, S. Chen, M. Arkin, *et al.*, "Machine learning models and pathway genome data base for trypanosoma cruzi drug discovery," *PLoS neglected tropical diseases*, vol. 9, no. 6, p. e0003878, 2015.

[37] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner, "The silva ribosomal rna gene database project: improved data processing and web-based tools," *Nucleic acids research*, vol. 41, no. D1, pp. D590–D596, 2012.

[38] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey, "Drugbank: a comprehensive resource for in silico drug discovery and exploration," *Nucleic acids research*, vol. 34, no. suppl_1, pp. D668–D672, 2006.

[39] B. Zingales, M. E. S. Pereira, R. P. Oliveira, K. A. Almeida, E. S. Umezawa, R. P. Souto, N. Vargas, M. I. Cano, J. F. da Silveira, N. S. Nehme, *et al.*, "Trypanosoma cruzi genome project: biological characteristics and molecular typing of clone cl brener," *Acta tropica*, vol. 68, no. 2, pp. 159–173, 1997.

[40] J. Molina, O. Martins-Filho, Z. Brener, *et al.*, "Activities of the triazole derivative sch 56592 (posaconazole) against drug-resistant strains of the protozoan parasitetrypanosoma (schizotrypanum) cruzi in immuno-competent and immunosuppressed murine hosts," *Antimicrobial agents and chemotherapy*, vol. 44, no. 1, pp. 150–155, 2000.

[41] W. J. Hoekstra, T. Y. Hargrove, Z. Wawrzak, *et al.*, "Antiparasitic effect in vitro, activity in a murine model of chagas disease, and structural characterization in complex with the target enzyme cyp51 from trypanosoma cruzi of the potent clinical candidate vt-1161," *Antimicrobial agents and chemotherapy*, pp. AAC–02287, 2015.