UPTEC X 03 001
MAR 2003

ISSN 1401-2138

JOHAN GEIJER

Databases for the *Trypanosoma cruzi* genome project

Master's degree project

Molecular Biotechnology Programme
Uppsala University School of Engineering

UPTEC X 03 001	**Date of issue 2003-11**

Author

Johan Geijer

Title (English)

Databases for the *Trypanosoma cruzi* genome project

Title (Swedish)

Abstract

Three different types of T*rypanosoma cruzi* sequences (EST, GSS and BACend) have been assembled using Phrap. The resulting contigs have been annotated using open reading frame prediction, gene prediction and homology search. All the results have been stored in newly constructed databases and user-friendly webinterfaces have been developed to access the data. The databases were implemented to facilitate the analysis process of biological data in the *Trypanosoma cruzi* genome project.

Keywords

Trypansoma cruzi, EST, GSS, BACend, PHRAP, assembly, annotation, databases

Supervisors

Björn Andersson & Daniel Nilsson
Center for Genomics and Bioinformatics (CGB), Karolinska Institute

Examiner

Lena Åslund
Rudbeck Laboratory, Uppsala University

Project name	Sponsors
Language **English**	Security
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **27**

Biology Education Centre Biomedical Center Husargatan 3 Uppsala
Box 592 S-75124 Uppsala Tel +46 (0)18 4710000 Fax +46 (0)18 555217

Databases for the *Trypanosoma cruzi* genome project

Johan Geijer

Sammanfattning

Den tropiska Chagas sjukdomen orsakas av parasiten *Trypanosoma cruzi*. Som ett led i WHOs parasitprojekt har ett arbete för att avläsa parasitens arvsmassa påbörjats. Förhoppningen är att avläsningsprojektet ska leda till att verkningsfulla läkemedel mot parasiten utvecklas, samt att kunskapen om parasiter ökar.

Genomisk forskning är idag en snabbt växande disciplin som använder sig av den senaste tekniken inom it- och datavetenskap. I genforskningen genereras stora mängder av ny-upptäckt data i en accelererande takt. Genomprojektet *Trypanosoma cruzi* består av 5 olika laboratorier i olika delar av världen som samarbetar för att läsa av parasitens arvsmassa. Dessa institutioner producerar stora mängder av korta sekvensbitar från det ursprungliga genomet, sk reads.

Detta examensarbete har syftat till att samla ihop det sekvenserade materialet, dvs alla reads, från de olika laboratorierna och organisera denna datamängd i en databas. Fortsättningsvis har syftet varit att försöka pussla ihop alla korta sekvensbitar till längre sammanhängande bitar och därefter använda olika datorprogram för att hitta intressanta regioner i dessa längre bitar. Alla resultat har sedan samlats i en databas och ett grafiskt användargränssnitt som kan användas via Internet har skapats. Dessa databaser underlättar för forskare i deras fortsatta arbete med parasiten.

Examensarbete 20 p i Molekylär bioteknikprogrammet

Uppsala universitet November 2003

Contents

1.Introduction

The advent of high-throughput sequencing, genomics and proteomics has generated a wealth of data. Bioinformatics is a rapidly growing recently developed field, which gives us the opportunity to gain new insights into the biology of organisms by the analysis of this huge amount of data. It combines the fields of computer science, mathematics and biology, and involves the study of the methods and problems associated with the task of bringing information to the researcher, sorting this mass of information in a meaningful way and then analyzing it. It involves the construction of bioinformatics pipelines which may include many different programs designed for a specific purpose, and which can then be used by the researcher to answer diverse questions, which arise in their fields.

Trypanosoma cruzi, a protozoan parasite from the order Kinetoplastida is the causative agent of the Chagas disease in Central- and South America. Today almost 20 million people are affected and the mortality is 10 % [1]. The WHO launched a Genome Initiative in 1994 that covered five different parasites, among them *T.cruzi*. The goal is to use genome-scale analysis to rapidly gain further insight into the biology of the organism and thereby make it possible to find effective therapies against the parasite.

In the genome project thousands of short sequences (reads) ~500 bp are produced from the original genome. One of the most difficult and critical steps in DNA sequence analysis is to align reads that have overlapping segments and to produce the best copy of the original genome sequence as possible. To consolidate all of the read sequences, an assembly program, Phrap [2] is used. The output from the assembly program is a number of contigs[1].
An important step in gene discovery procedure is the computational annotation of these contigs. There are numerous ways to annotate DNA-sequences. Most frequently used are homology search programs combined with gene finding programs. Usually one also runs an open reading frame finding program. In addition to these tools there are various ways to search for repeats and protein motifs and add the obtained information to the list of annotations.

Presently there is no optimal gene finding program for *T.cruzi*. Instead it is necessary to use the most advantageous that exists, e.g GLIMMER [3,4] and Testcode [5]. These programs make gene predictions and mark sections on the contig sequences as protein coding or protein non-coding. One of the most widely used bioinformatics tools for annotation used today involves searching a database for sequences similar to a query sequence. If two sequences are similar, it is likely that they also have a similar structure, and therefore also a similar function. Thus, if you have an as-yet-uncharacterised sequence, finding homologs in the databases can give you an idea of what its identity might be. The most commonly used program is the BLAST [6] family of programs, which gives biologically relevant matches in a reasonable amount of time.

To store all of the data and information that accumulates from an assembly and annotation project in an organized way, a database is often used. In this format it is possible to efficiently structure and link all of the collected data. For researchers to benefit from all this assembly and annotation information stored in the database, however, two additional things are required: ready access to the collected pool of information and a way to extract from this pool only the information of interest to a given researcher. A solution is to construct a user-friendly webinterface that is connected to the database.

The purpose of this project was to build a bioinformatic pipeline in the *T.cruzi* genome project (Figure 1). The aim was to cluster more than 36 000 different sequences developed in

[1] A set of overlapping sequences that are assembled into a contigous segment.

the *T.cruzi* genome project, and to annotate the resulting contigs. The results were stored in an easy editable database and a www-based interface was constructed to access the database. The webinterface would also provide options to perform to homology search against the database and/or to extract information from it.

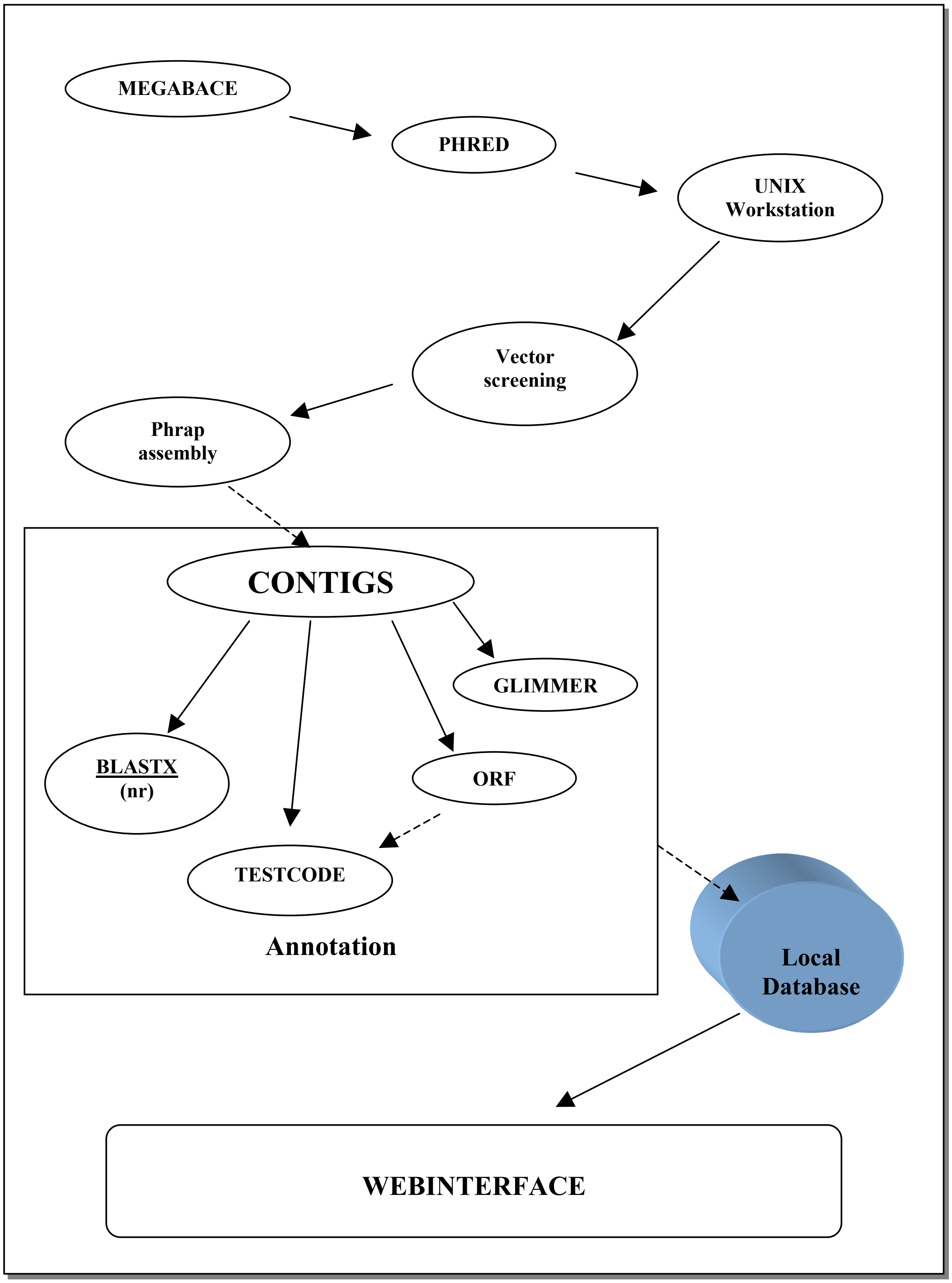


Figure 1: A simple model of the bioinformatic pipeline constructed in this project. The MegaBace produces chromatograms from the sequenced DNA. The chromatograms are interpreted by Phred and stored on a UNIX workstation. Before Phrap assembly, the sequences are trimmed. The assembled sequences are annotated using Glimmer, Orf, Testcode and BlastX. The assemblies together with the annotations are stored in a local database. A user-friendly webinterface is constructed to access the database.

In the following section a background to the *T.cruzi* genome project and some theory behind this project is given. In section 3 the methods used in this work are described and the corresponding results are presented in section 4 together with some discussion.

2. Background and theory

2.1 The *Trypanosoma cruzi* genome project

Parasite diseases are a major health problem in developing countries. One of them is Chagas' disease caused by the parasite *T.cruzi*. The disease is affecting 16-18 million people [1] in South- and Latin America (Figure 2) and the death rate is 10 %. In addition more than 100 million people remain at risk. According to WHO [1], in a global perspective, Chagas represents the third largest parasite disease burden after malaria and schistosomiasis. The severity and symptoms of the disease vary in different geographical regions, maybe as an effect of the heterogeneity among isolates of *T.cruzi* parasites or differences of the immune system in the human population. At present there exist no potent and safe drug or vaccine against the disease. In 1994 WHO introduced parasite genome programs, in which many laboratories from South America, US and Europe are involved [7]. Among others, *T.cruzi* is one of the parasites included in this program. Assistant professor Björn Andersson and his research crew at the Centre for Genomics and Bioinformatics, KI, started to sequence Chromosome 3 of *T. cruzi* in 1996. The final goals in the *T.cruzi* project are to sequence the whole genome and to map and identify the genes, with the aim to develop a more efficient cure.

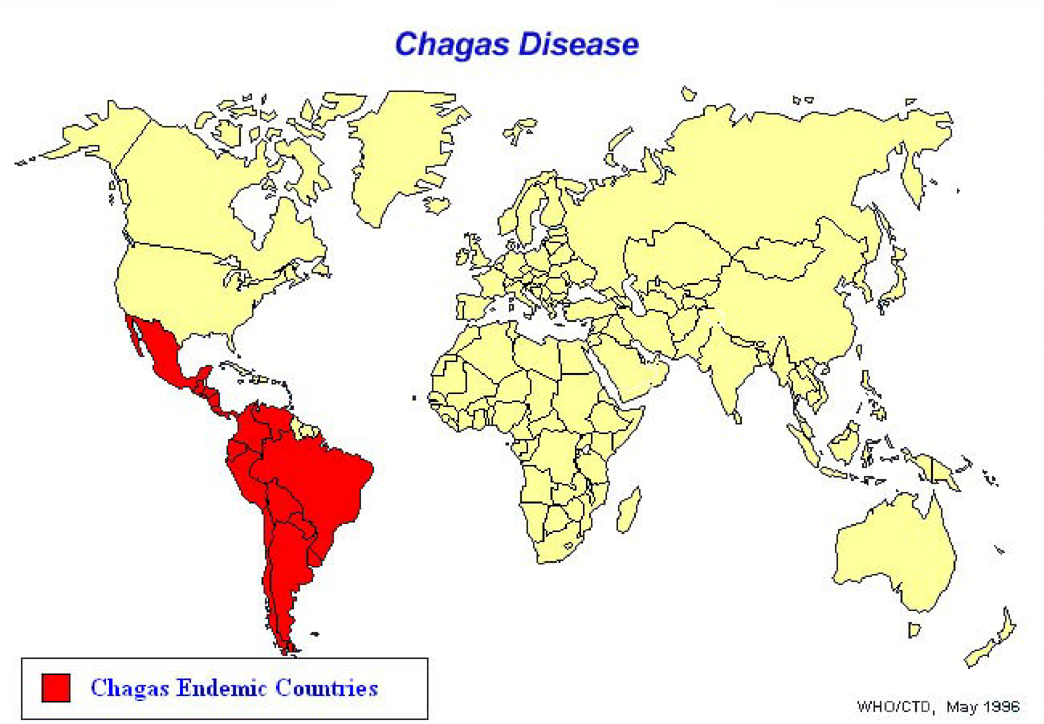


Figure 2: The geographical distribution of Chagas' disease [1]. Only Central and South America are subjected to the disease. Illustration used with permission from WHO.

2.2 *Trypanosoma cruzi* biology and genetics

The *T.cruzi* parasite is a flagellated protozoan, a lower eukaryote that belongs to the order Kinetoplastida. The parasite passes through a complex life cycle, exhibiting different morphological stages in their invertebrate and mammalian hosts. The parasites, in form of metacyclic trypomastigotes, are transmitted to man via blood-sucking bugs (Figure 3).

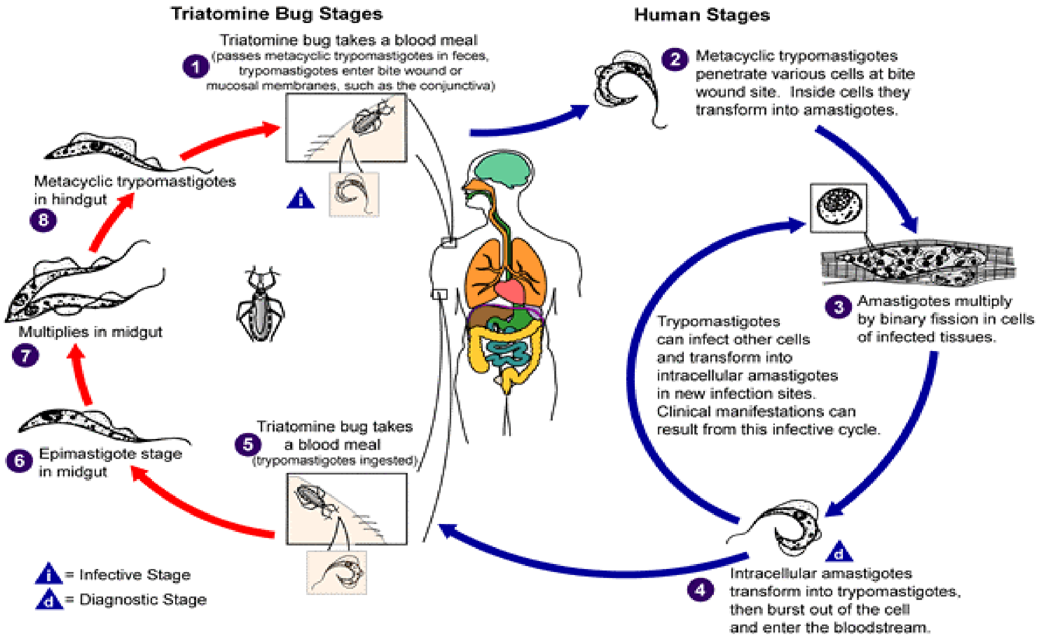

Triatomine Bug Stages
Human Stages
1
Triatomine bug takes a blood meal (passes metacyclic trypomastigotes in feces, trypomastigotes enter bite wound or mucosal membranes, such as the conjunctiva)
i
2
Metacyclic trypomastigotes penetrate various cells at bite wound site. Inside cells they transform into amastigotes.
3
Amastigotes multiply by binary fission in cells of infected tissues.
Trypomastigotes can infect other cells and transform into intracellular amastigotes in new infection sites. Clinical manifestations can result from this infective cycle.
4
Intracellular amastigotes transform into trypomastigotes, then burst out of the cell and enter the bloodstream.
d
5
Triatomine bug takes a blood meal (trypomastigotes ingested)
6
Epimastigote stage in midgut
7
Multiplies in midgut
8
Metacyclic trypomastigotes in hindgut
i = Infective Stage
d = Diagnostic Stage

Figure 3: The life cycle of *Trypanosoma cruzi* [8]. Illustration used with permission from Centers for Disease Control & Prevention.

Inside the human body, they invade a variety of mammalian cells and differentiate into the amastigote stage within these cells. After multiplying intracellularly, the parasites can re-enter the blood in the form of trypomastigotes and infect new cells or be ingested by the bug. In the midgut of the bug, the parasites differentiate into small epimastigotes, some of which transform into the infectious metacyclic trypomastigotes.

T.cruzi contains many organelles [9], the flagella is used by the trypomastigote as a means to move. The kinetoplast is the energy manufacturer of the cell. Peroxisomes contain the glycolysis enzymes and acidocalcisomes are mainly a mineral storage.

The genetics of *T.cruzi* is more similar to prokaryotes than to eukaryotes. The genome seems to be almost intronless with a few exceptions. There are also some strong indications that the genes are expressed polycistronically[1]. This is further supported by a similar organization of the related kinetoplastid *Leishmania major* chromosome 1 [10]. Individual mRNAs seems to be created in a process involving coupled trans-splicing of a 5’ spliced leader and poly-adenylation to the 3’ tail. Control of gene expression seems to be achieved mainly by post-transcriptional mechanisms [11, 12, 13], involving changes to mRNA stability rather than conventional changes of primary RNA transcript production rates.

[1] Several proteins are encoded on one expressed primary RNA transcript

T.cruzi has been estimated to have ~12,000 genes from the gene density of about 1 gene per 3.5-4 kb, as revealed by genomic sequencing [14] and the haploid genome size of ~45 Mb in the *T.cruzi* reference clone CL Brener [14].

An interesting feature that the *T.cruzi* share with other eukaryotes is extensive repetitive elements. In the *T.cruzi* genome it seems to represent a substantial portion of the genome [15]. Their functional implications still remain unknown. There exist many different types of repeats. The main groups are repeats organized in tandem (clustered) and interspersed repeat elements. Some repeats can be repeated 200-300 times [15]. Another interesting feature of the genome is that the sizes of the different chromosomes show large variations between different strains [16] and also between chromosome homologues in the genome.

2.3 Sequencing & Basecalling

2.3.1 Sequences

In this project three different types (EST, GSS & BACends) of DNA sequence were used. All of the sequences were derived from the *T.cruzi* CL Brener strain. In the *T.cruzi* genome project a MEGABACE 1000 DNA sequencing system was used for the production of reads[2]. The output from the MEGABACE 1000 is a chromatogram file (Figure 6). These chromatograms are then used as inputs into the basecalling program Phred [2,17,18] that produces the actual reads.

2.3.1.1 EST

An EST, Expressed Sequenced Tag, is a tiny portion of an entire gene, a fragment of a cDNA clone that has been sequenced. The process by which ESTs are manufactured requires the construction of a cDNA library. Normalization [19] and library subtraction is often used to increase relative representation of less abundantly transcribed mRNAs. cDNA-clones are sequenced a single time (single pass sequencing), from one or both ends of the DNA insert (Figure 4), using universal primers which are complementary to the vector at the multiple cloning site. The forward primer may be located near the 5' or the 3' end of the cloned insert, depending on how the inserts were directionally cloned. Only 300-500 readable bases are produced from each sequenced read, and yet a full gene transcript may be several thousands of bases long. ESTs thus provide a "tag level" association with an expressed gene sequence, trading quality and total sequence length for the high quantity of genes that can be tagged in a given amount of time.

Generation of EST data results in “low quality” sequence information. A single read is generated for each EST, and as such will contain errors from its generation at each step. These can include clone orientation, associated clone ID chimeras and missing 3' and 5' reads. Because the data are single-pass unedited sequences, they are also subject to errors caused by compressions and basecalling (see below) problems resulting in frameshifts.
The ESTs used in this project were produced from a normalized cDNA library [19] from the CL Brener strain.

[2] An estimate of the original piece of sequence.

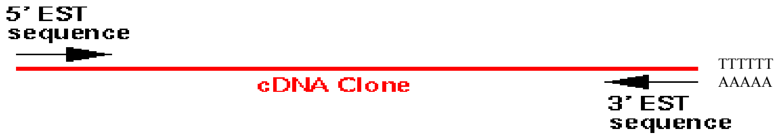


Figure 4: EST manufacturing using single pass sequencing. Sequencing from the 5' end of the cDNA forms 5' ESTs and sequencing from the reverse direction forms 3' ETSs.

2.3.1.2 GSS

Different from an EST, a GSS (genome survey sequence) is genomic in origin, rather than cDNA (mRNA). A GSS is a random "single pass read" genome survey sequence. Its quality is just as low as the quality of an EST. The GSS used in this project were produced from a random genomic *T.cruzi* library, derived from the CL Brener strain.

2.3.1.3 BACend

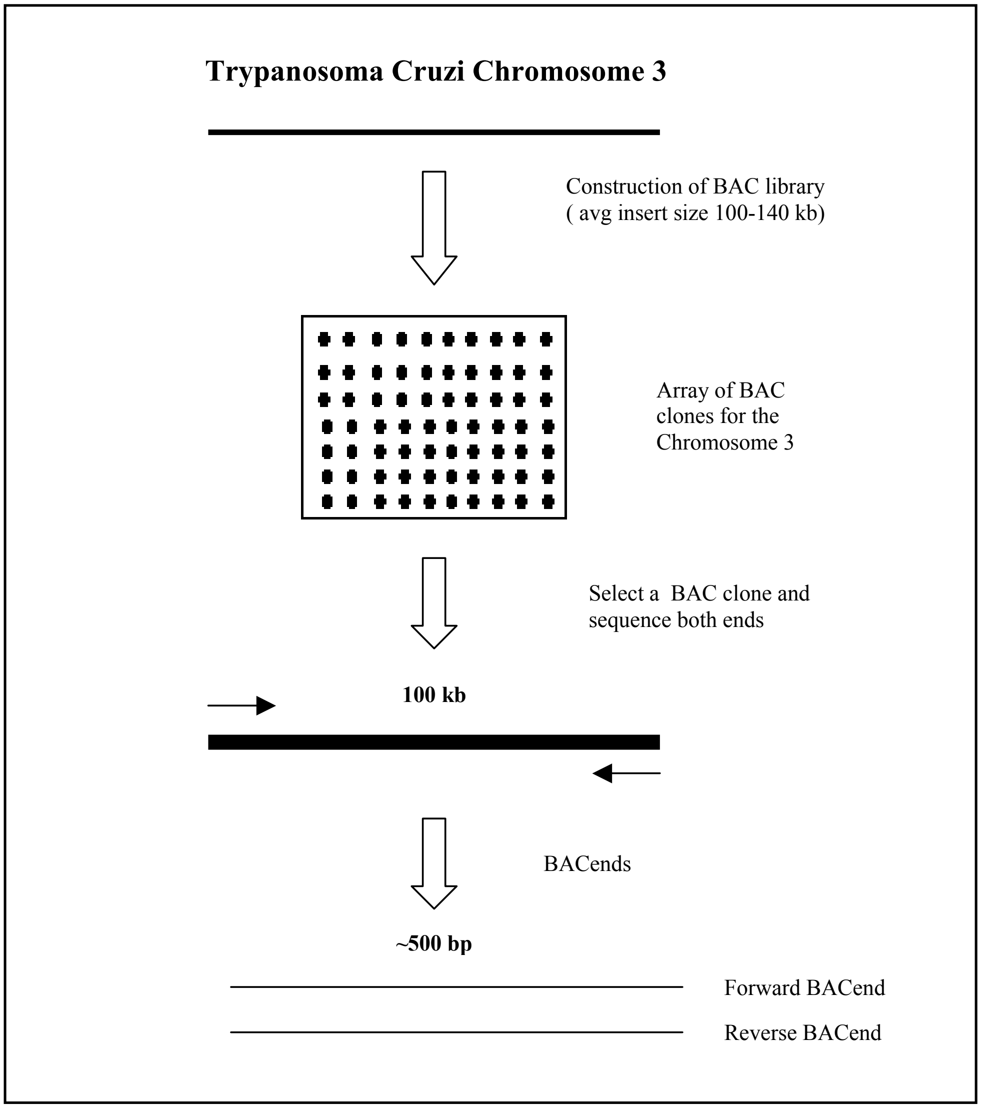


Figure 5: Construction of BACend sequences.

BACend sequences are generated from BAC (Bacterial Artificial Chromosome) clones (Figure 5). Since restriction enzymes are used when constructing BAC clones from the chromosome, a sequence marker is given where the BAC clone is cut. A BACend sequence is therefore often called a "semi-random" DNA sequence.

2.3.2 PHRED - Basecalling

The process of assigning a sequence from raw data (chromatogram files from the MEGABACE) is called *basecalling* (Figure 6). This is a process that can give rise to errors. If this step doesn't produce a correct DNA sequence, any subsequent analysis of the sequence is affected. Expressed sequence tags (EST's) and GSS's have the highest error rate (1/10 – 1/100 errors per base) [20].

The software used for basecalling was Phred [2,17,18]. It uses Fourier analysis to resolve fluorescence traces to predict an evenly spaced set of peak locations, and then uses dynamic programming to match the actual peak locations with the predicted result. It then annotates the output from basecalling with the probability that the call is an error. Phred scores (quality values) represents scaled negative log probability that a base call is an error (Eq 1); hence, the higher the Phred score, the lower the probability that an error has been made. So a Phred score of 20 means that the peak is incorrectly called one time in 100 (10^2). In general a Phred score of 20 (Q-20 in Figure 6) means that the sequence is reliable. These values can be helpful in determining if a region needs to be re-sequenced. The sequence assembly program Phrap [2] also needs the quality values as input.

Phred reads in the chromatogram file and outputs a "phd" file that has one base and its corresponding quality score per line. Ideally, all peaks in a chromatogram trace file would be spaced an even distance apart. In reality, sequence composition and the dye primer chemistry can alter the distance between peaks as shown below (Figure 6), making it difficult to identify bases accurately. Phred uses information from the region surrounding each peak to determine the probability that a base has been identified correctly. Phred also identifies loop/stem sequence motifs based on dye primer data and splits peaks if there are indications that CC or GG peaks have merged.

$$\text{Phred quality value, } Q = -10*\log_{10}(P) \quad \text{(eq 1)}$$

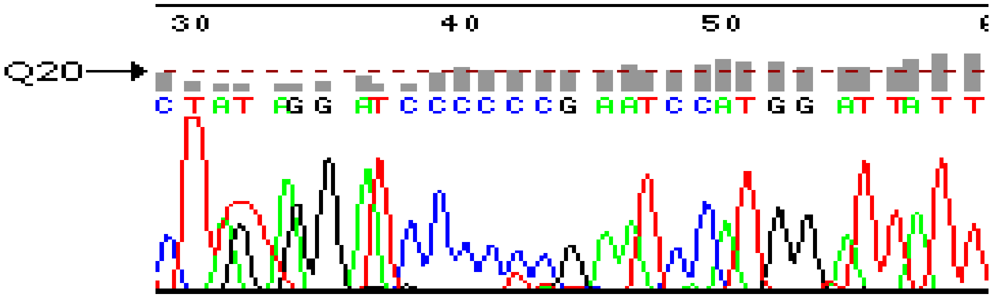


Figure 6: A MEGABACE chromatogram interpreted by the PHRED basecalling program. The Q-20 line shown is the PHRED score 20. This value is often used as a threshold for a reliable sequence.

2.4 Assembly

2.4.1 The assembly concept

The goal of fragment assembly is to order overlapping reads to recreate the original sequence (Figure 7). Sequence assembly is the process of constructing the "best guess" contig/clone – sequence from a set of overlapping reads of the clone [22]. The problem is complicated by the fact that there could be errors in the read sequences and the fact that there could be many repeats in the clone sequence. The tools used for sequence assembly are **Crossmatch** [2] (used for vector screening), **Phrap** [2] (Phragment assembly program) and **Consed** [2,21](A graphical tool for sequence finishing, described below).

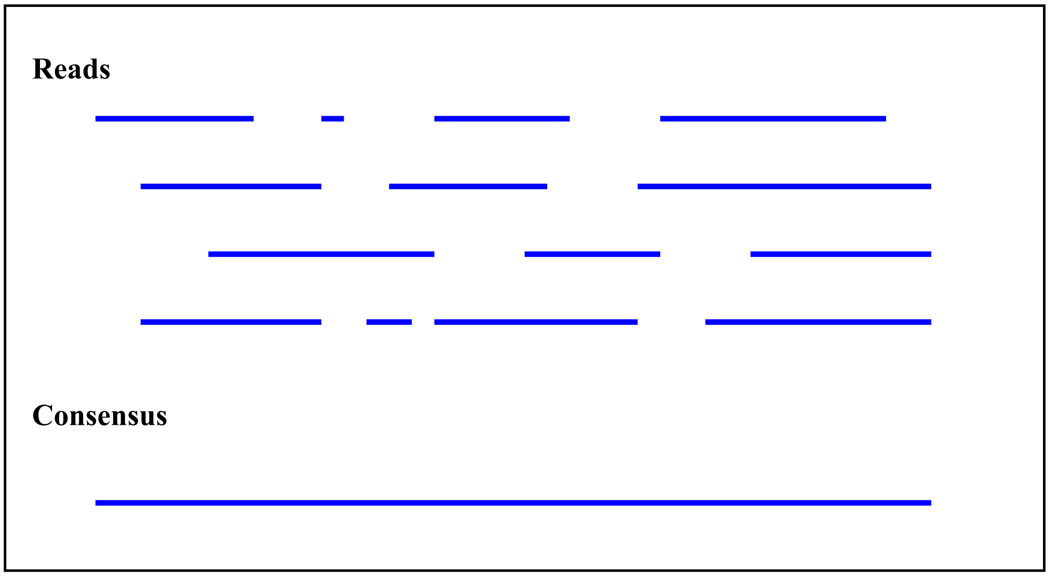


Figure 7: Reconstruction of the sequence of a clone.

2.4.2 Crossmatch

One of the most effective methods to remove contaminants in raw data reads is to compare each read against vector sequences and other potential "false" sequences using an algorithm that is reasonably fast and accurate. Crossmatch (a local alignment program), an implementation of the Smith-Waterman algorithm developed by Phil Green [2] is often used for this type of screening. The program takes quality and fasta files as input and produces screened versions of these files as an output (Figure 9).

2.4.3 PHRAP

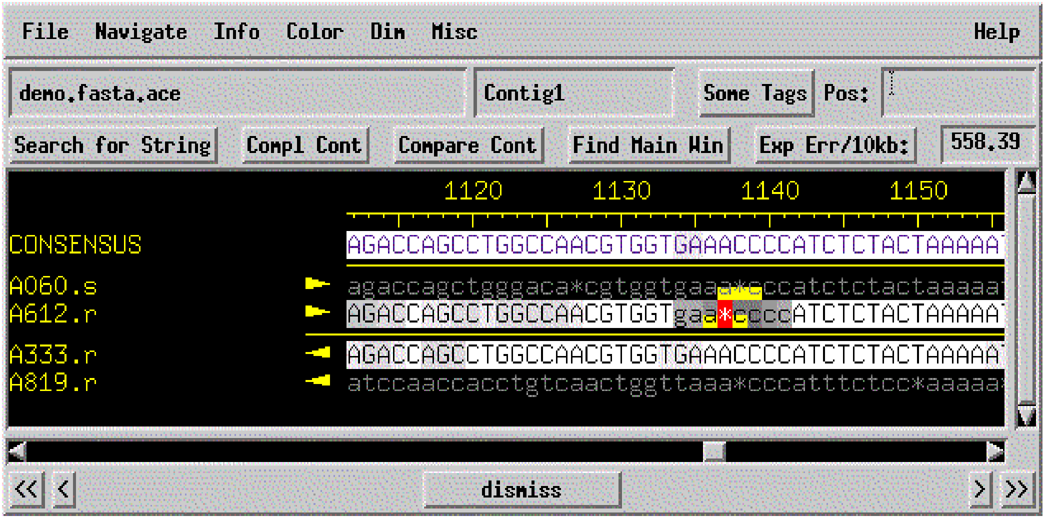
Phrap [2] is an assembly program for shotgun DNA sequence data. Input files to Phrap are of two types: sequence fasta files and quality files (Figure 9). It uses the quality information provided by Phred to discriminate repeats and sequencing errors in the assembly process and to construct contig sequences as a mosaic of the highest quality parts of the reads. It provides extensive information about the assemblies contained in phrap.out-, *.ace- and *contigs.qual files.

The program can handle very large datasets; hundreds of thousands of reads are easily manipulated [2]. There are several different parameters that can be set to govern the Phrap result. There are mainly two parameters (minscore and minmatch) that are used to alter the conditions for the assembly, while the others are set as default. Minmatch is minimum length of matching word (bases) to nucleate SWAT (Smith-Waterman) comparison. If minmatch = 0, a full (non-banded) comparison is done. Increasing minmatch can dramatically decrease the time required for the pairwise sequence comparisons; it also tends to have the effect of increasing assembly stringency. However it may cause some significant matches to be missed, and it may increase the risk of incorrect joins in phrap in certain situations (by causing implied overlaps between reads with high-quality discrepancies to be missed). The minscore is simply the minimum alignment score. The alignment score has to be higher than this value for a join to take place.

ESTs and GSS's by their nature have a degree of erroneous sequence data, complicated by short length and some mis-annotation in the ESTs. Stringent one-pass assembly methods tend to result in fewer, shorter consenus sequences. Looser systems for clustering result in larger, more 'sloppy' clusters, with various expressed forms being represented within each cluster. Each approach has its advantages and disadvantages. Stringent clustering provides greater initial fidelity, at a cost of lower coverage of expressed gene data and a lower inclusion rate of expressed gene forms. Loose clustering provides greater coverage, at a cost of possible inclusion of paralogous expressed genes, lower fidelity data, but at a gain of greater inclusion of alternate expressed forms.

2.4.4 Consed

To validate the success of a Phrap-assembly when using certain parameters it is important to be able to visualise the assembly-results. This is achieved by using Consed (Figure 8). It is a graphical program for viewing and editing Phrap assemblies [21].



demo.fasta.ace

Contig1

558.39

```
                       1120      1130      1140      1150
CONSENSUS        AGACCAGCCTGGCCAACGTGGTGAAACCCCATCTCTACTAAAAA
A060.s           agaccagctgggaca*cgtggtgaaa*cccatctctactaaaaa
A612.r           AGACCAGCCTGGCCAACGTGGTgaa*ccccATCTCTACTAAAAA
A333.r           AGACCAGCCTGGCCAACGTGGTGAAACCCCATCTCTACTAAAAA
A819.r           atccaaccacctgtcaactggttaaa*cccatttctcc*aaaaa
```

Figure 8: Consed, image showing the aligned reads and the consensus sequence.

The base background color indicates the quality of a base call as determined by Phred and Phrap. Light background corresponds to high quality, dark background to low quality. Black color means that the sequence was of too low quality to be aligned by Phrap. Phrap takes base qualities into account when calculating the contig ("consensus") sequence. In the region shown in (figure 8), only two sequences (A333.r and A612.r) are high quality. Phred quality scores enable Phrap to identify the best sequence (A333.r in this case) in a region with discrepancies, and to base the contig sequence on this sequence.

In this project Consed was used only as a tool to view and navigate through the consensus sequence and identify regions of low quality. It was not used for editing the consensus sequences but to find the best minscore parameter in the Phrap assembly, i.e find the minscore that gave the best assembly.

2.4.5 Phred/Crossmatch/Phrap/Consed Pipeline

The Phred, Phrap, Crossmatch and Consed procedures can be summarized as a pipeline. The pipeline (figure 9) starts with chromatogram inputs from the MegaBace sequencer. Phred then interprets the chromatograms and produces phd-files. These files contain bases and the corresponding quality values. Vector sequence and other disturbing sequence are removed using the Crossmatch program. The assembly program Phrap puts the sequences together and Consed is used to view the result.

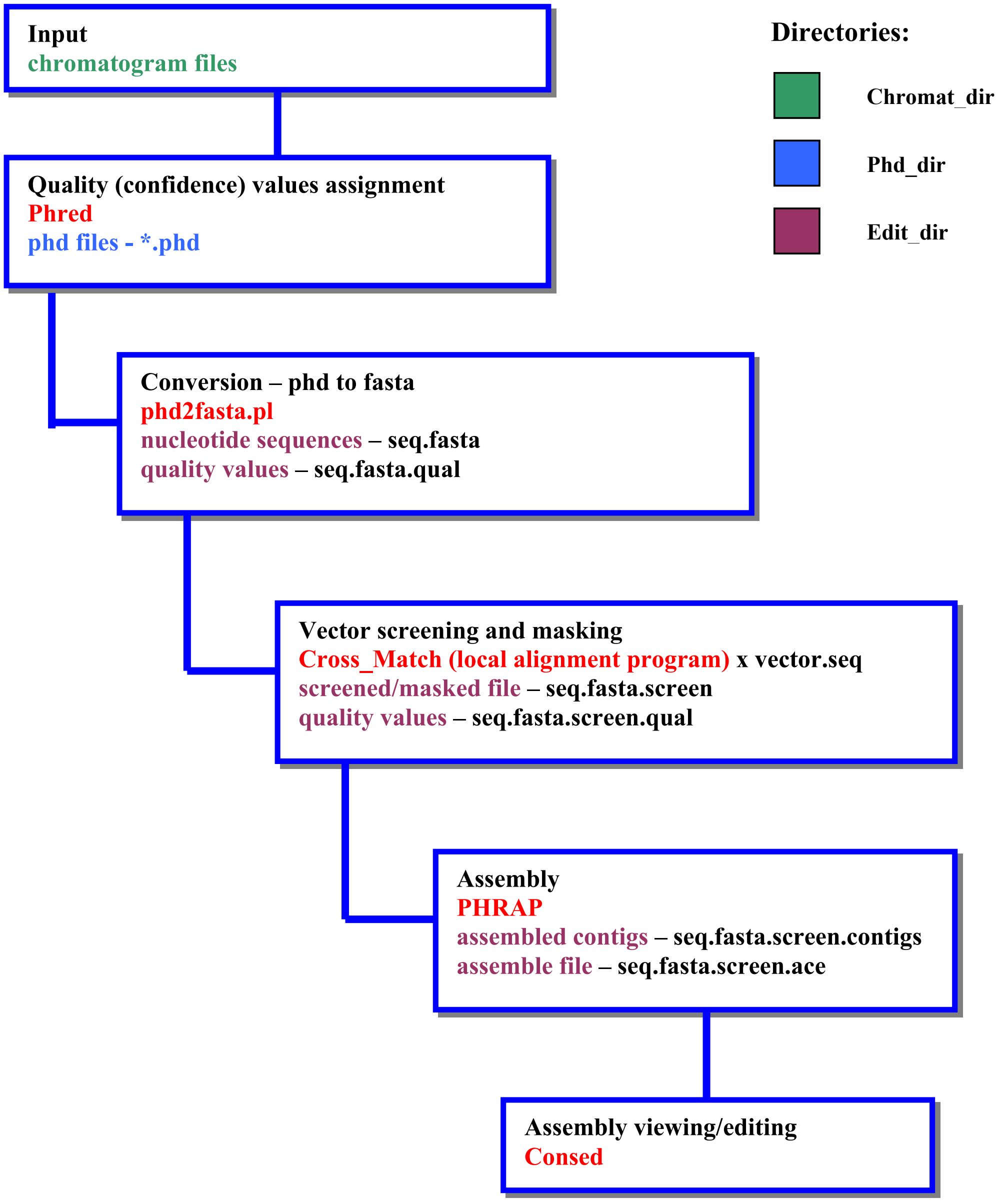

Input
chromatogram files
Quality (confidence) values assignment
Phred
phd files - *.phd
Conversion – phd to fasta
phd2fasta.pl
nucleotide sequences – seq.fasta
quality values – seq.fasta.qual
Vector screening and masking
Cross_Match (local alignment program) x vector.seq
screened/masked file – seq.fasta.screen
quality values – seq.fasta.screen.qual
Assembly
PHRAP
assembled contigs – seq.fasta.screen.contigs
assemble file – seq.fasta.screen.ace
Assembly viewing/editing
Consed
Directories:
Chromat_dir
Phd_dir
Edit_dir

Figure 9: Flowchart showing the Phred/Phrap/Consed pipeline.

2.5 Contig/Gene Annotation

Contig/gene annotations can be defined as: features on the genome derived through the transformation of raw genomic sequences into information by integrating computational tools, auxiliary biological data and biological knowledge (Figure 10). Auxiliary data can for example be genetic or physical maps. It can also be expression data like ESTs from a specific tissue or stage in an organism's life cycle. Some typical computational annotation tools that have been used in this project are Testcode [5] and Glimmer [3,4] for gene finding, BlastX [6] for homology search and an ORF program to find potential open reading frames in the different contigs. Some database resources that often are used are curated sequence feature data sets like non-redundant protein databases, repeat element databases etc. Genome sequence from related species like *Trypanosoma brucei* and genome sequence from more distant species are also often used.

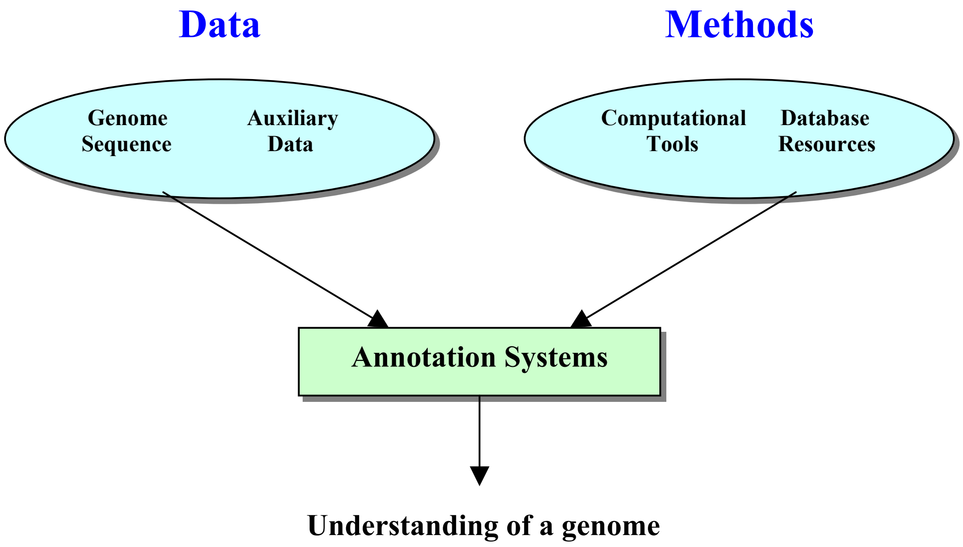


Figure 10: Annotation process overview.

The features that are outputs from these programs are intervals or ordered sets of intervals on a sequence that describe some biological attribute. The collection (set) of all of these features for a sequence is the annotation of that sequence.

2.5.1 Open reading frame prediction

2.5.1.1 ORF

An Open Reading Frame Finder is a tool for locating open reading frames in a DNA sequence. It is a good course of action to apply such a program and to get a first indication of putative coding regions. The output from the ORF finder is also often used as an input in a gene finder program. There almost exist as many ORF programs as there are genome projects. For this project a new ORF finder was constructed using C++ (see section 3.3.1).

2.5.2 Gene finding

2.5.2.1 GLIMMER2

Glimmer2 (Gene Locator and Interpolated Markov Modeller 2) uses interpolated context models (ICMs), which is a further extension of Interpolated Markov Models (IMM's) to identify the coding regions and distinguish them from noncoding DNA [3,4]. It is a system designed for finding genes in microbial DNA, especially in the genomes of bacteria and archea. The ICM approach, uses a combination of Markov models from 1st through 8th-order, weighting each model according to its predictive power. Glimmer 2.02, which was used in this project, uses 3-periodic nonhomogenous Markov models in their ICMs.

The Glimmer system consists of two main programs. The first of these is the training program, ***build-icm***. This program takes an input set of sequences, builds and outputs the ICM for them. These input sequences can be complete genes or just partial ORFs. For a new genome, this training data can consist of those genes with strong database hits as well as very long open reading frames that are statistically almost certain to be genes. The second program is ***glimmer***, which uses this ICM to identify putative genes in an entire genome. Glimmer automatically resolves conflicts between most overlapping genes by choosing one of them. It also identifies genes that are suspected to truly overlap, and flags those for further inspection by the user [3,4].

2.5.2.2 Testcode

Several methods for sophisticated gene prediction have been proposed. One approach is to analyse the sequence in search of regularities of the nucleotide pattern. It has been shown that, in a protein encoding reading frame, certain patterns will occur in a periodic fashion [5]. The detection of such patterns in a relative large range (window of 200 bp or more) is the operational hypothesis of the **Testcode** program.

Testcode predicts coding regions by looking for non-randomness [5]. An algorithm developed by Fickett measures the A, C, G and T – content of a sequence, and the degree to which each nucleotide is favoured in each codon position. The Fickett algorithm weights these eight values (4 content + 4 position-value) to form a single score (Testcode indicator) indicating the probability that the sequence encodes a protein. Testcode is a formal implementation of Fickett's method. The statistic is high when measures of compositional bias with a periodicity of three are high. The key measures of bias are simply the four measures:

$$\text{N-position} = \text{Maximum}(n_{(1)}, n_{(2)}, n_{(3)}) / \text{Minimum}(n_{(1)}, n_{(2)}, n_{(3)}) \quad \text{(eq 2)}$$

where $n_{(1)}$, $n_{(2)}$ and $n_{(3)}$ are the composition of each nucleotide at positions (1,4,7,...), (2,5,8,...) and (3,6,9,...) and N = A, C, G and T. The composition is simply the number of observations of n in the window.

2.5.3 Homology search

2.5.3.1. BLASTX

One way to identify a putative gene in a new genome is to find a close homolog from the same or another organism. By far, the most popular tool for searching sequence databases is the program BLAST(Basic Local Alignment Sequence Tool). It uses the method of Altschul [6] to search for similarities between a query sequence and all the sequences in a database. The query sequence and the database you want to search can be either protein or nucleic acid in any combination. BlastX, compares a nucleotide query sequence translated in all reading frames against a protein sequence database. This tool can be used to find potential translation products of an unknown nucleotide sequence. The query is translated, and each of the six products is compared to each database sequence in a separate protein-protein pairwise comparison.

In the output the best hits are presented first. A score value and an E-value are presented. The score value shows (among other things) how well the two sequences match each other. The value should be high. The E-value shows the probability that the two sequences are matched at random. This value should be low and depends on the sequence length and how large the accessed databases are.

3. Material and methods

3.1 Collection of sequence data

The sequence data that have been used in this project were collected from 5 different laboratories (CGB, Uppsala, Granada, FioCruz and San Martin) involved in the *Trypanosoma cruzi* project. Totally 36 370 different reads were collected (see Table 1). Chromatogram, sequence and quality files as well as clone names and clone numbers were collected together with the accession numbers and annotations from Genbank. All of this data were organised and then inserted into a raw database table (one table for each type of raw data) in the database that had been constructed for the project.

EST (Expressed sequenced tags):	**10 268 reads**
GSS (Genome survey sequence):	**11 459 reads**
BACend:	**14 643 reads**
Totally:	**36 370 reads**

Table 1: Summary of collected sequences.

3.2 Assembly

The sequences were first pre-processed using **Crossmatch**. The program was used to screen each of the sequences for vector, polyA and spliced leader contamination. It was used to compare the whole set of reads to a set of vector sequences (used in the construction of the sequences), polyA-sequence and also to the spliced leader. If any potential vector sequence, polyA or spliced leader was found somewhere in the sequence (mostly at either end of the sequence), the crossmatch program converted that region to X's. A masked version of the reads was produced. These bases were then automatically ignored by subsequent programs and not included in the Phrap assembly. The processed reads were assembled with Phrap. The strategy in this project was to assemble different combinations of the sequences (Table 2) and the main goal was to assemble all of the sequences at once (EST +GSS +BACend). Since the procedure to assemble all of the reads was quite resource demanding (needs a lot of memory) it was executed during nighttime. The hardware used was a PIII dual processors and a 1.5 Gb memory.

Consed was used to view the Phrap results by navigation through the contig and make sure that it looked all right. Some effort was put into analysing the assemblies using Consed but no editing was performed. The assembly process was initiated using the same minscore parameter as applied in a previous EST assembly [23], minscore = 40. The results were analysed and additional assemblies were tested using different minscores (Table 2).

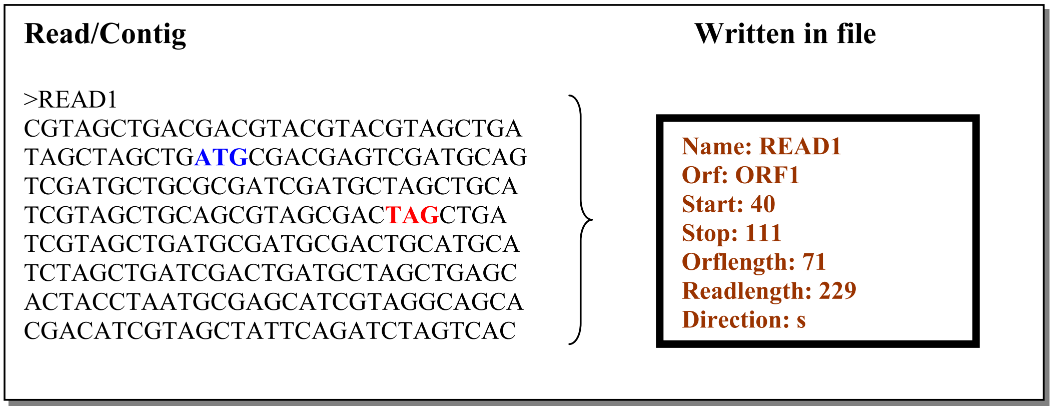
Sequences assembled	*Assembly name*	*Parameters*
EST	EST1	Minscore:39
	EST2	Minscore:40
	EST3	Minscore:45
GSS	GSS1	Minscore:39
	GSS2	Minscore:40
BACend	BAC1	Minscore:40
EST + GSS	ESTGSS1	Minscore:40
GSS + BACend	GSSBAC1	Minscore:40
EST + BACend	ESTBAC1	Minscore:40
EST + GSS+ BACend	ESTGSSBAC1	Minscore:40
	ESTGSSBAC2	Minscore:45

Table 2: The different assemblies and their corresponding parameters. The minmatch parameter were set to default = 14.

3.3 Contig Annotation

3.3.1 ORF

A short program written in C++ called **Orf** has been created in this project. It is an analysis tool, which locates all open reading frames of a selectable minimum size in a read/ DNA sequence. This tool identifies all open reading frames using the standard genetic codes for start (ATG) and stop (TAA, TAG, TGA). It is constructed as a sliding window that sweeps the read/contig and scans for the codons mentioned. In each read/contig being analysed the first nucleotide is always registered as a start position and the last nucleotide as a stop position. That eliminates the risk of reading through a read/contig that only has one stop or start codon or neither of them in the sequence. As a default, to be registered as an open reading frame (orf) the distance between a pair of consecutive start and stop codons must be at least 50 nucleotides. The user can change this cutoff value. Shorter orf's are discarded. If an orf longer than 50 bp has been found, the length of the orf and read as well as the start and stop position are written to a file (Figure 11). The orf is named according to which reading frame (1, 2 or 3) it belongs to and the frame direction (s or r). This program was applied to all the contigs from all of the assemblies. The results were then stored in the Orf table in the database constructed for this project.



```
>READ1
CGTAGCTGACGACGTACGTACGTAGCTGA
TAGCTAGCTGATGCGACGAGTCGATGCAG
TCGATGCTGCGCGATCGATGCTAGCTGCA
TCGTAGCTGCAGCGTAGCGACTAGCTGA
TCGTAGCTGATGCGATGCGACTGCATGCA
TCTAGCTGATCGACTGATGCTAGCTGAGC
ACTACCTAATGCGAGCATCGTAGGCAGCA
CGACATCGTAGCTATTCAGATCTAGTCAC
```

```
Name: READ1
Orf: ORF1
Start: 40
Stop: 111
Orflength: 71
Readlength: 229
Direction: s
```

Figure 11: The ORF program. It extracts open reading frame information from a read or contig. ORF1 indicate that the orf is in the first reading frame. The direction s means that the orf is in the forward direction.

3.3.2 Glimmer

The Interpolated context models [4] (ICM's) used in this project were trained using a set of well-studied *Trypanosoma cruzi* genes from Genbank. Each contig in an assembly was used as an input sequence into Glimmer 2.02. The minimum gene length was set to 50. It is the length of the smallest fragment considered to be a gene. The length is measured from the first base of the start codon to the last base before the stop codon. The minimum overlap length was set to 10. It is the lower bound limit of the number of bases overlap between two genes that is considered a problem, and overlaps shorter than this are ignored. The outputs from Glimmer 2.02 were parsed and the most important information, like starts and stops of putative genes were collected and inserted into the database table glimmer, included in the database (Figure 12).

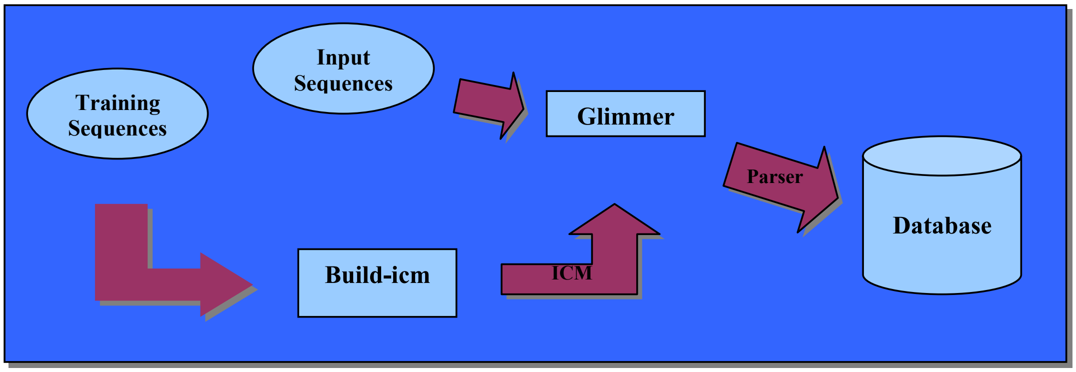


Figure 12: Glimmer flowchart.

3.3.3 Testcode

In the Testcode [5] approach the results stored in the orf table were used together with the contig sequences as inputs into the testcode program. The start and stop positions in each contig stored in the orf table were extracted from the database. Since the testcode algorithm is independent of the reading frame, only the longest orf was used. These positions were then used to choose a region in the contigs that would be used as an input into the testcode program. The window size was set to 200 [5], because the Testcode statistics is not applicable for windows of less than 200 bases and a sliding window approach was executed.

The testcode program calculated the testcode indicator for each position in the input sequence. A C++-program was written in which the range of the indicator was divided into 10 equal intervals. Two different cut-offs values were calculated. The first cut-off (Cutoff1) was after the first four intervals and the second cutoff (Cutoff2) was after the sixth interval. To be able to predict the sequence in question an average value for the testcode indicators were calculated. If the average number <cutoff1 the sequence region was assigned non-coding region. If the average was >cutoff1 but <cutoff2 the sequence was assigned middle region and if the average was >cutoff2 the sequence was assigned coding region. A parser for the output from the C++-script was build and the extracted information was stored in the database table, testcode.

3.3.4 BLASTX

The assembled sequences were searched against a non-redundant (nr) database by using BlastX [6]. This nr database included all non-redundant GenBank CDS translations and all proteins from PDB, SwissProt, PIR and PRF [24]. The results were parsed by using a script written in C++. The matches for whom the E-value was below 10^{-2} were stored in the table, blast_x, in the database.

3.4 Design and construction of biological database

All the information collected and generated in this project was stored in a database. The Relational Database Management System (RDBMS) used in the project was PostgreSQL [25]. It is a full-featured object-relational DBMS that supports user-defined data types and functions in addition to a broad set of SQL functions and types. For a database with a size as the one constructed in this project it is a good choice of DBMS.

Data in a relational database table is organised in rows, with each row representing one record in the database. A row may contain several separate pieces of information (fields). Each field in the database must contain one distinct piece of information. The job of the RDBMS is to make connections between related tables by rapidly finding the common elements that establish those relationships.

An ER-model of the database was constructed (Figure 13). It describes what's stored in the different tables in the database and the relation between the tables. The database contains one table for each type of cluster (estcluster, estgsscluster etc.) and also one table for each type of raw data (rawdataest, rawdatagss and rawdatabac). In addition one table was constructed for each type of annotation. Each row in the cluster tables was labelled with the name of a read from the rawdata tables. In that way each read is connected to a cluster and hence connected to each of the annotation tables (Figure 13).

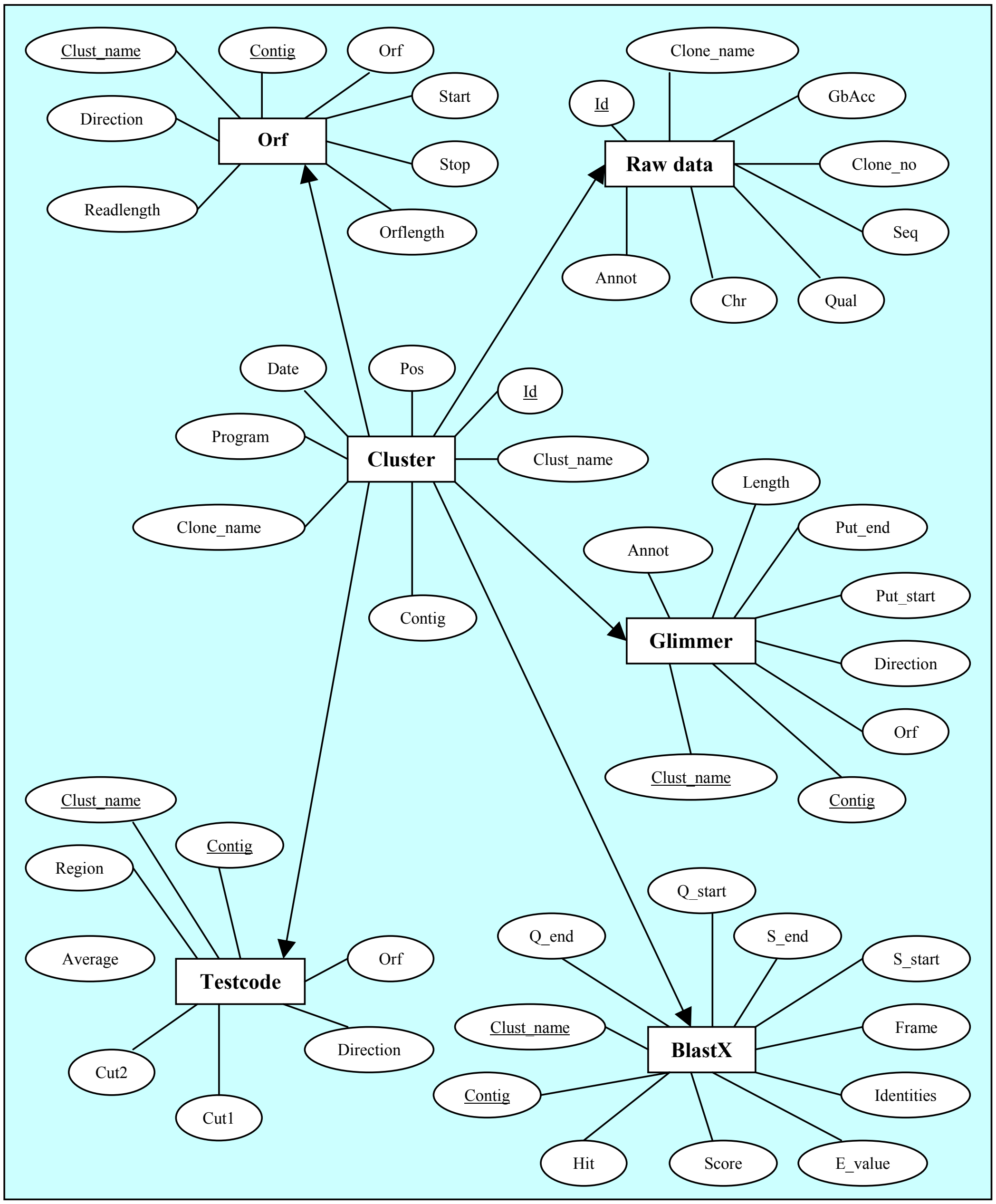


Figure 13: ER-model for the database. The primary keys are underlined.

3.5 Webinterface

A www-based webinterface was made for the visualisation of the information in the databases. The interface was implemented using Html [26], with Php [27] and Perl [28,29] as scripting languages. The Php scripts are executed by the Php preprocessor, which extracts data from the database. The results are then together with the rest of the Php-script interpreted into Html-code. The final Html-page gets sent back to the user's browser (Figure 14).

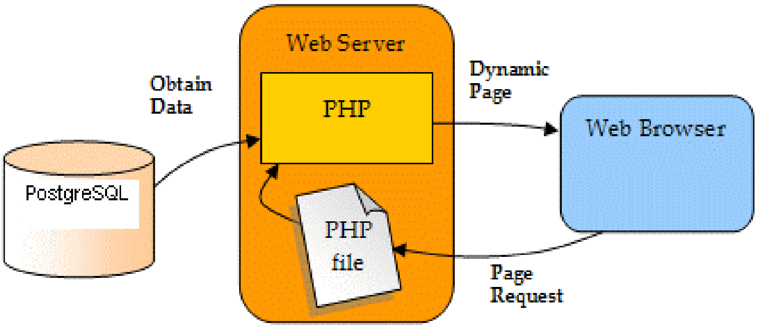


Figure 14: Simplification of how PHP works.

4. Results and discussion

Using Phrap, 11 different assemblies where made using 7 different combinations of reads (Table 3). A comparison between the EST2 – cluster (minscore 40) and the ESTGSSBAC1 (minscore 40) cluster has been made, evaluating the number of contigs, number of orf and glimmer hits and finally some results from the BlastX.

4.1 PHRAP Assembly

Assembly [1)]	*Reads* [2)]	*Minscore* [3)]	*Contigs* [4)]	*Singeltons* [5)]	*Contigs with est* [6)]
EST1	10 268[a]	39	1714	4360	1714
EST2	"	40	1928	3331	1928
EST3	"	45	1930	3297	1930
GSS1	11 459	39	1571	3154	-
GSS2	"	40	1570	3371	-
BAC1	14 643	40	1589	1685	-
ESTGSS1	21 727	40	4641	5634	2456
GSSBAC1	26 102	40	2842	4499	-
ESTBAC1	24 911	40	3785	4532	2151
ESTGSSBAC1	36 370	40	4776	6608(2602)	2594
ESTGSSBAC2	"	45	4784	7113(2666)	2577

Table 3: A summary of all the results from the different Phrap assemblies. 1) Shows the name of the assembly, 2) The number of reads put into Phrap. 3) The minscore parameter. 4) The number of contigs produced in the assembly. 5) Represents the number of singletons. The number in parenthesis is the number of singletons that is an EST. 6) The number of contigs with at least one est.

The resulting 1928 clusters from the EST2 assembly contained 6937 reads. Most of the clusters (71%) in this assembly contained 2 or 3 sequences and the largest cluster comprised 50 sequences. The redundancy of the cDNA library, estimated from the fraction of sequences that assembled into clusters was ~68%, a relatively high value considering the normalization step. In total 5257 "transcripts[b]" were identified in the EST2 assembly.

[a] The total number of ESTs included in the assembly is 10 268, out of which 5607 are annotated as 5' – ESTs , 2327 are annotated as 3' – ESTs and 2334 have no annotation.

[b] Means the contigs with one or more ESTs included plus the number of singletons that are ESTs. There are no evidence that these transcripts represent unique genes. There's a big possibility that some fraction of them could be nonoverlapping sequences derivied from the same transcript.

The 4776 clusters from the ESTGSSBAC1 assembly contained 29762 reads. 63 % of the clusters contained 2 or 3 sequences and the largest cluster had 731 sequences. The total amount of “transcripts” was 5196. This is roughly the same amount as in the EST2 assembly. The difference is that in the large assembly there are genomic[c] DNAs flanking the ESTs in most of the 2594 contigs that contains one or more ESTs. This means that depending on where in the contig the est/ests is/are placed there can be additional genomic sequence upstream or downstream of a potential “transcript”. These kinds of contigs (Figure 15) are of special interest when searching for 5’- and 3’ - UTR’s. They are also obviously good targets for gene finding.

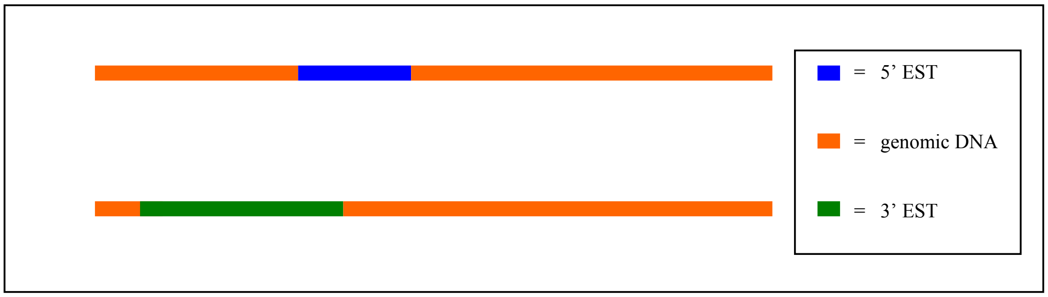


Figure 15: Showing some of the potential assembly results from the large assembly.

4.2 ORF

Some query searches in the ORF database table were made to find out how many contigs in the different assemblies that have orf’s longer than 500 bp (Figure 16). The search was made in all of the six reading frames. As one would expect there are a lot more potential open reading frames in the assemblies where the EST’s are included. These results in some way show the importance of having EST’s in the assembly

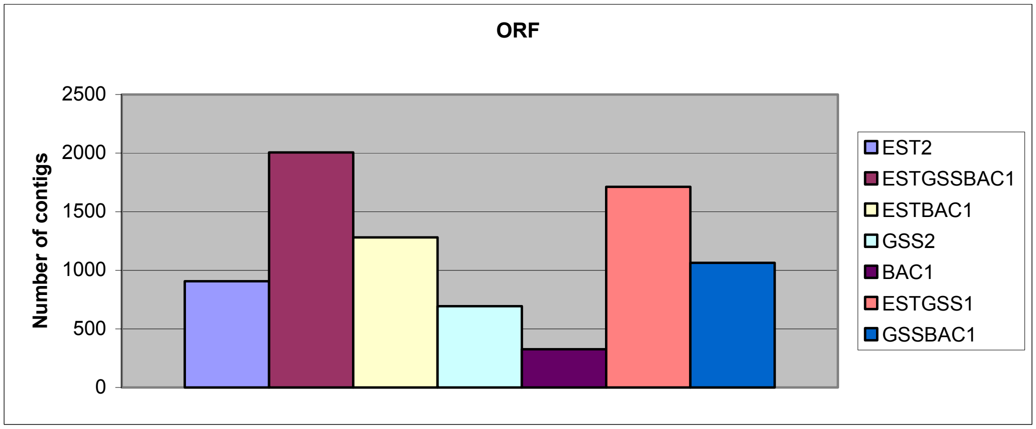


Figure 16: Summary of the number of contigs in the different assemblies that have an orf longer than 500 bp. The figure shows the assemblies made using minscore 40.

4.3 GLIMMER

The database was screened to identify the number of “long-orf contigs” that also have Glimmer hits. Only the results from the EST2- and ESTGSSBAC1-queries are shown. Figure 17 shows that the number of coding contigs according to Glimmer with long orfs has increased from ~300 to ~800 between the EST - and the large ESTGSSBAC - assembly. Also

[c] DNA (BACend or GSS) that can be located anywhere in the genome, not just from the stretches that produce mRNA.

the number of coding contigs including EST's has increased from ~300 to ~500. This indicates that a sequence material that has been added to the EST sequences in the largest assembly increased the amount of coding information in the contigs quite a bit.

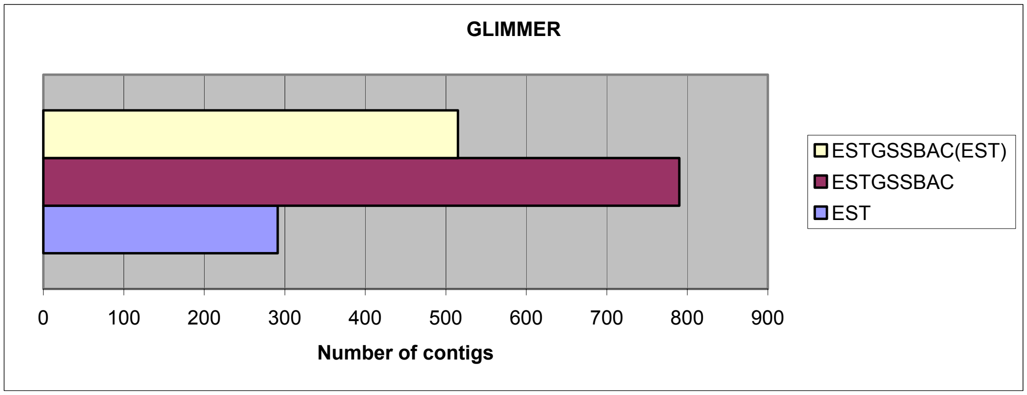


Figure 17: Results showing the number of contigs from EST2 and ESTGSSBAC1 assembly that have a orf > 500 bp and also a GLIMMER hit. The top bar shows the number of ESTGSSBAC1 contigs with at least one EST included.

4.4 BLASTX

The database has been used to compare the BlastX hits for the EST2- and ESTGSSBAC1-assemblies. A cut-off of $p < 10^{-10}$ was used in the database queries [30]. The results from the EST query showed that 736 contigs had a putative identity. The same kind of result from the ESTGSSBAC query showed that 995 contigs could be assigned a putative identity. In the larger ESTGSSBAC1 assembly only the 2594 contigs containing an EST were investigated. A number of queries were also made to find out the distribution of the hits. In the EST2 assembly 41 % of the top hits[1] (Figure 18) were Trypanosoma hits and 27% Leishmania[2] hits.

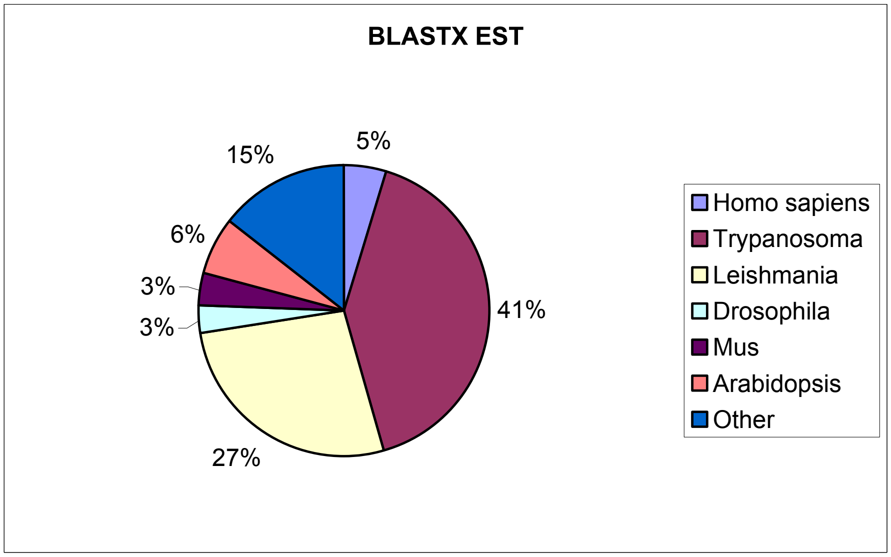


Figure 18: Summary of BlastX top hit results for EST2 assembly. E-value threshold was $1e^{-10}$.

[1] The hit with the lowest E_value.

[2] A parasite similar to Trypanosoma parasites.

The same kind of search for the ESTGSSBAC1 assembly showed that 42 % of the top hits (Figure 19) were Trypanosoma hits and 29 % Leishmania hits. It appears that the larger assembly has a somewhat larger number of contigs with Trypanosoma top hits. Interestingly, the average Trypanosoma top hit E_value is ten times (1e-13 vs 1e-12) lower in the larger assembly. That indicates a slightly greater Trypansoma gene resemblance in the contigs from the ESTGSSBAC1 assembly containing at least one EST.

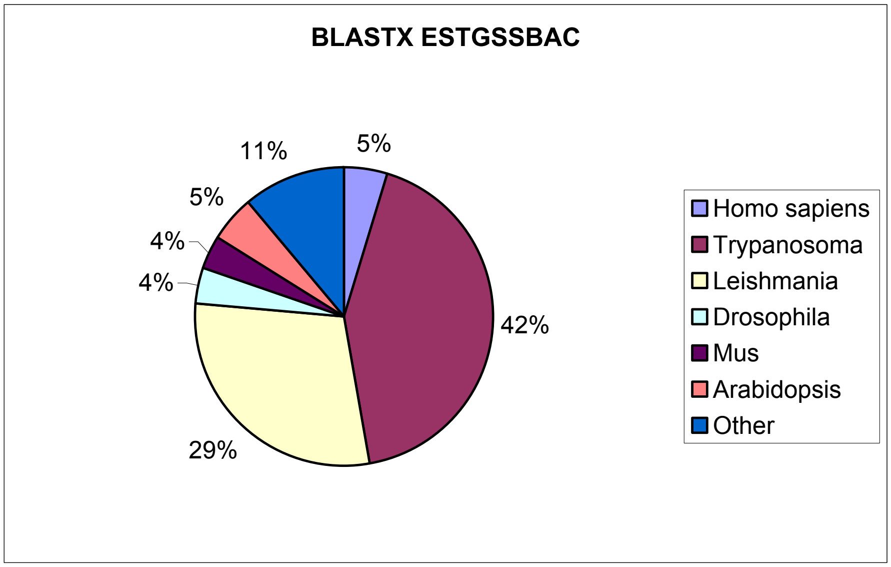


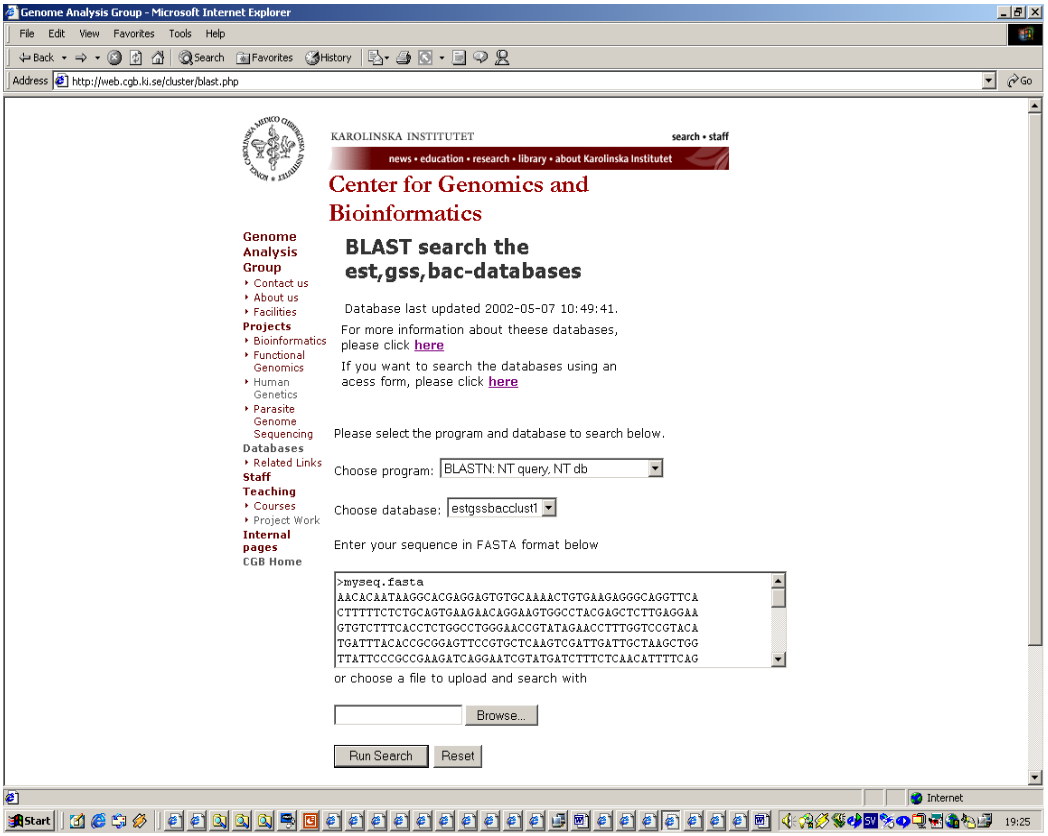
Figure 19: Summary of BlastX top hit results for ESTGSSBAC1 assembly E-value threshold was 1e^{-10}.

4.5 Web interface

The construction of web interfaces resulted in two versions of interfaces. The Blast – search page (Figure 20) was used as a start page to search the database with an input sequence in fasta format and one access form page (Figure 23) to easily extract information from the database.

The web interfaces have been built in a way to make it possible for the user to extract any kind of information from the different assemblies or from the separate reads. In the start page you can choose which Blast – program to use, depending on if you enter a nucleotide or protein fasta input sequence, and what assembly to blast against. On this page there is also a link to an information page. Here you can read more about the data stored in the database. The start page also includes a link to the access form page (Figure 23).

When a Blast search is finished a new result page appears (Figure 21). It shows all the Blast hits (contigs) and it is possible to click on one of the contig hits to receive additional information about the contig. The contig page contains information about which reads that is included in the contig and the start position of every read. The clone name is linked to a page where all the information about that specific read is presented and also a link to Genbank through the Genbank accession number. The contig link (Figure 22) opens a page where all the results from the orffinder program (Orf), the genefinder programs Glimmer2, Testcode and finally results from the BlastX program for that contig are shown in table form. It is possible to do a new limited search at this stage and submit a new query with different search parameters.



Genome Analysis Group - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://web.cgb.ki.se/cluster/blast.php
Go
KAROLINSKA INSTITUTET
search • staff
news • education • research • library • about Karolinska Institutet
Genome Analysis Group
Contact us
About us
Facilities
Projects
Bioinformatics
Functional Genomics
Human Genetics
Parasite Genome Sequencing
Databases
Related Links
Staff
Teaching
Courses
Project Work
Internal pages
CGB Home
Internet
Start
19:25

Center for Genomics and Bioinformatics

BLAST search the est,gss,bac-databases

Database last updated 2002-05-07 10:49:41.

For more information about theese databases, please click here

If you want to search the databases using an acess form, please click here

Please select the program and database to search below.

Choose program: BLASTN: NT query, NT db

Choose database: estgssbacclust1

Enter your sequence in FASTA format below

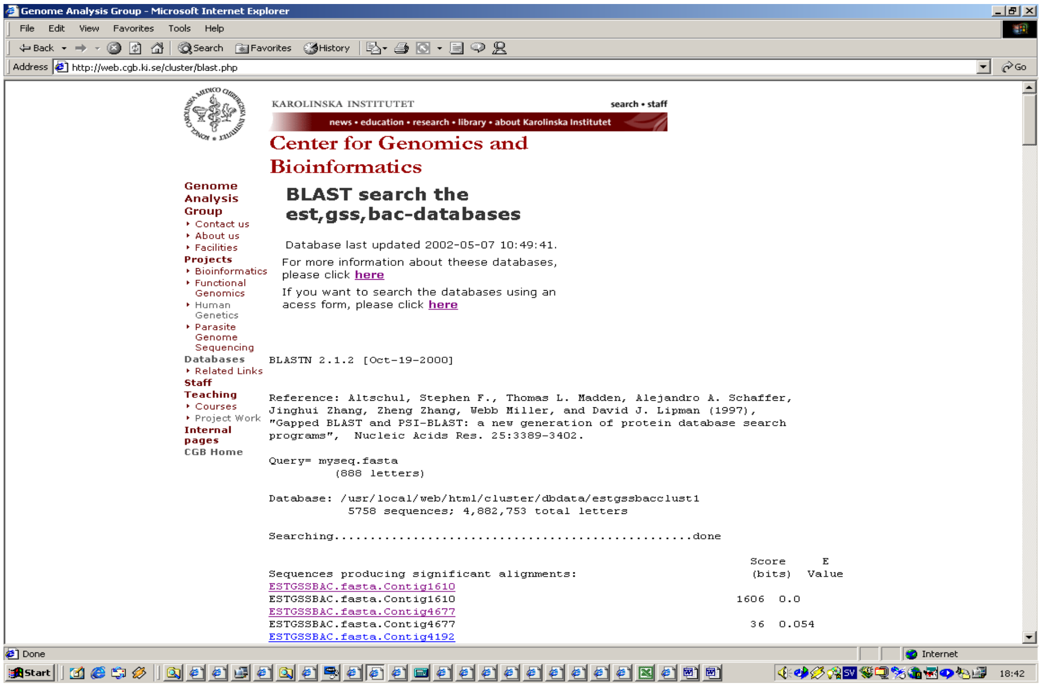
```
>myseq.fasta
AACACAATAAGGCACGAGGAGTGTGCAAAACTGTGAAGAGGGCAGGTTCA
CTTTTTCTCTGCAGTGAAGAACAGGAAGTGGCCTACGAGCTCTTGAGGAA
GTGTCTTTCACCTCTGGCCTGGGAACCGTATAGAACCTTTGGTCCGTACA
TGATTTACACCGCGGAGTTCCGTGCTCAAGTCGATTGATTGCTAAGCTGG
TTATTCCCGCCGAAGATCAGGAATCGTATGATCTTTCTCAACATTTTCAG
```

or choose a file to upload and search with

Browse...

Run Search Reset

Figure 20: The BLAST search interface. Here it is possible to choose the type of Blast search, which kind of assembly to search, and to enter the sequence to use as query sequence in the Blast search.



Genome Analysis Group - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://web.cgb.ki.se/cluster/blast.php
Go
KAROLINSKA INSTITUTET
search • staff
news • education • research • library • about Karolinska Institutet
Genome Analysis Group
Contact us
About us
Facilities
Projects
Bioinformatics
Functional Genomics
Human Genetics
Parasite Genome Sequencing
Databases
Related Links
Staff
Teaching
Courses
Project Work
Internal pages
CGB Home
Done
Internet
Start
18:42

Center for Genomics and Bioinformatics

BLAST search the est,gss,bac-databases

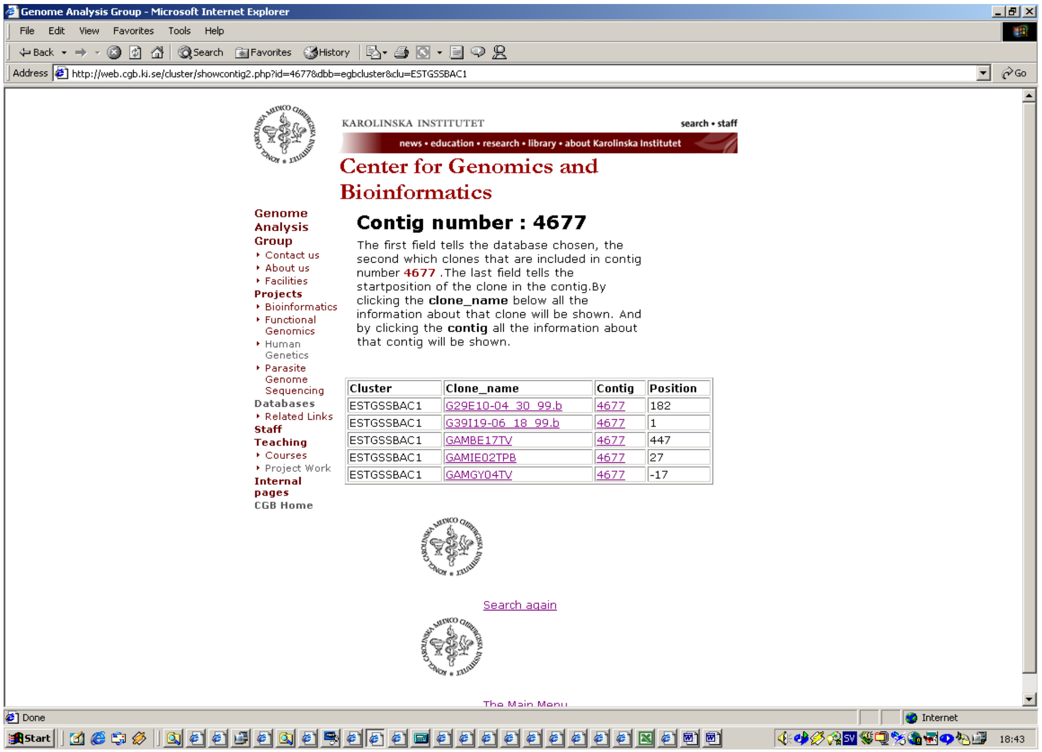
Database last updated 2002-05-07 10:49:41.

For more information about theese databases, please click here

If you want to search the databases using an acess form, please click here

```
BLASTN 2.1.2 [Oct-19-2000]


Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Query= myseq.fasta
         (888 letters)

Database: /usr/local/web/html/cluster/dbdata/estgssbacclust1
           5758 sequences; 4,882,753 total letters

Searching..................................................done


                                                                   Score     E
Sequences producing significant alignments:                        (bits)  Value
ESTGSSBAC.fasta.Contig1610
ESTGSSBAC.fasta.Contig1610                                        1606   0.0
ESTGSSBAC.fasta.Contig4677
ESTGSSBAC.fasta.Contig4677                                          36   0.054
ESTGSSBAC.fasta.Contig4192
```

Figure 21: The results from a blast search using a nucleotide input sequence.


Genome Analysis Group - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://web.cgb.ki.se/cluster/showcontig2.php?id=4677&dbb=egbcluster&clu=ESTGSSBAC1
KAROLINSKA INSTITUTET
search • staff
news • education • research • library • about Karolinska Institutet
Center for Genomics and Bioinformatics
Genome Analysis Group
Contact us
About us
Facilities
Projects
Bioinformatics
Functional Genomics
Human Genetics
Parasite Genome Sequencing
Databases
Related Links
Staff
Teaching
Courses
Project Work
Internal pages
CGB Home
Search again
The Main Menu
Done
Internet
Start
18:43

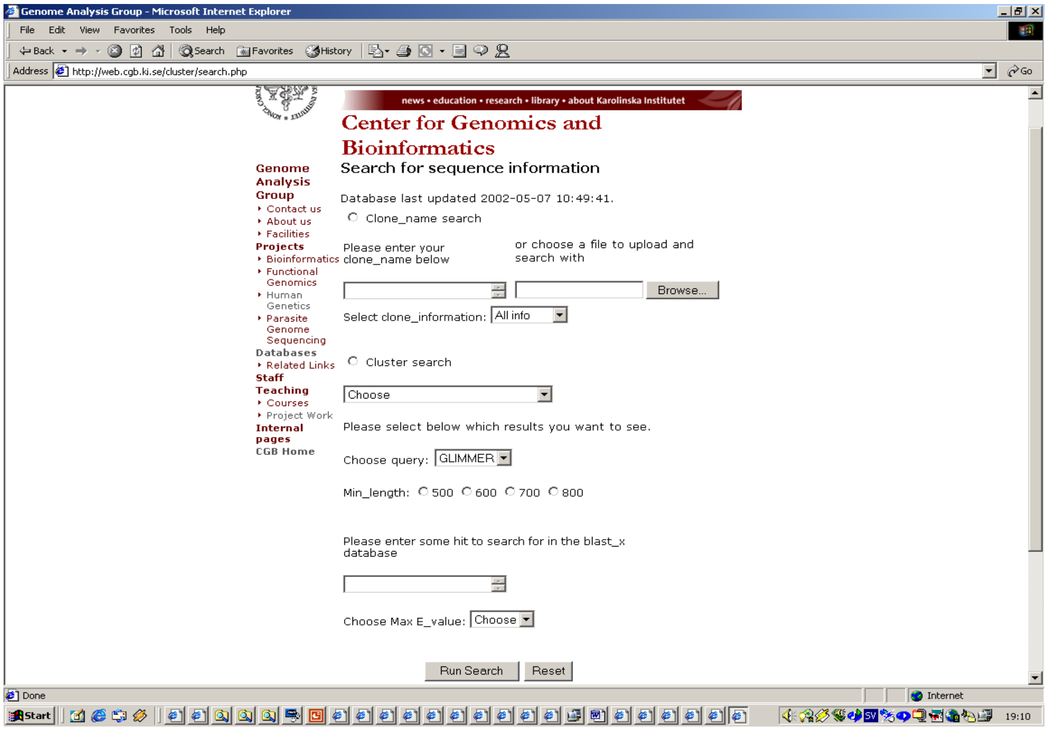
Contig number : 4677

The first field tells the database chosen, the second which clones that are included in contig number **4677** .The last field tells the startposition of the clone in the contig.By clicking the **clone_name** below all the information about that clone will be shown. And by clicking the **contig** all the information about that contig will be shown.

Cluster	Clone_name	Contig	Position
ESTGSSBAC1	G29E10-04_30_99.b	4677	182
ESTGSSBAC1	G39I19-06_18_99.b	4677	1
ESTGSSBAC1	GAMBE17TV	4677	447
ESTGSSBAC1	GAMIE02TPB	4677	27
ESTGSSBAC1	GAMGY04TV	4677	-17

Figure 22: The result page when clicking on one of the Blast hits. The table shows more information about the specific contig chosen.

The access form (Figure 23) page offers the user to search the database for a particular read or assembly information by making specific choices in the form. For example the user can choose annotation information for a read or search for the all *T cruzi* BlastX hits in a specific assembly.


Genome Analysis Group - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://web.cgb.ki.se/cluster/search.php
news • education • research • library • about Karolinska Institutet
Center for Genomics and Bioinformatics
Genome Analysis Group
Contact us
About us
Facilities
Projects
Bioinformatics
Functional Genomics
Human Genetics
Parasite Genome Sequencing
Databases
Related Links
Staff
Teaching
Courses
Project Work
Internal pages
CGB Home
Browse...
Run Search
Reset
Done
Internet
Start
19:10

Search for sequence information

Database last updated 2002-05-07 10:49:41.

○ Clone_name search

Please enter your clone_name below

or choose a file to upload and search with

Select clone_information: All info

○ Cluster search

Choose

Please select below which results you want to see.

Choose query: GLIMMER

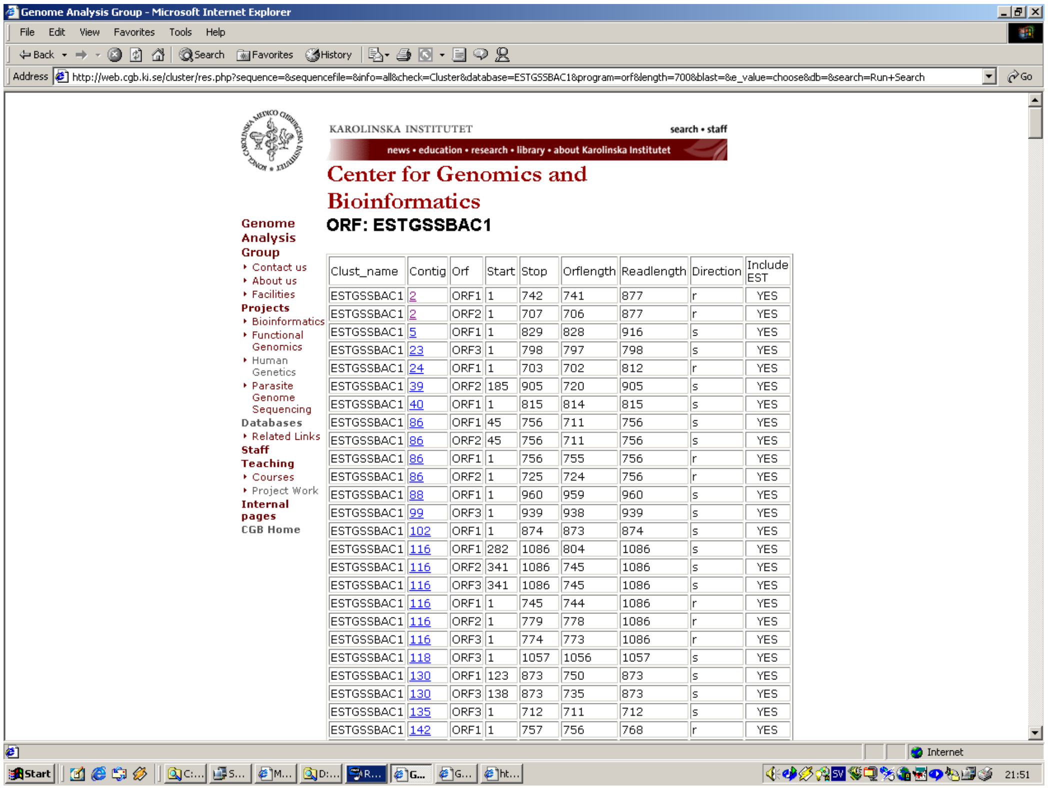
Min_length: ○ 500 ○ 600 ○ 700 ○ 800

Please enter some hit to search for in the blast_x database

Choose Max E_value: Choose

Figure 23: A simple search form to extract information from the database.

If the user for example wants to investigate all the contigs with open reading frames longer than 700 bp in a specific assembly, the results are shown on a new page (Figure 24). Here one can see information about the orf's and also be informed about if the contig has any EST included or not.


Genome Analysis Group - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://web.cgb.ki.se/cluster/res.php?sequence=&sequencefile=&info=all&check=Cluster&database=ESTGSSBAC1&program=orf&length=700&blast=&e_value=choose&db=&search=Run+Search
Go
KAROLINSKA INSTITUTET
search • staff
news • education • research • library • about Karolinska Institutet
Center for Genomics and Bioinformatics
Genome Analysis Group
Contact us
About us
Facilities
Projects
Bioinformatics
Functional Genomics
Human Genetics
Parasite Genome Sequencing
Databases
Related Links
Staff
Teaching
Courses
Project Work
Internal pages
CGB Home
Internet
Start
21:51

ORF: ESTGSSBAC1

Clust_name	Contig	Orf	Start	Stop	Orflength	Readlength	Direction	Include EST
ESTGSSBAC1	2	ORF1	1	742	741	877	r	YES
ESTGSSBAC1	2	ORF2	1	707	706	877	r	YES
ESTGSSBAC1	5	ORF1	1	829	828	916	s	YES
ESTGSSBAC1	23	ORF3	1	798	797	798	s	YES
ESTGSSBAC1	24	ORF1	1	703	702	812	r	YES
ESTGSSBAC1	39	ORF2	185	905	720	905	s	YES
ESTGSSBAC1	40	ORF1	1	815	814	815	s	YES
ESTGSSBAC1	86	ORF1	45	756	711	756	s	YES
ESTGSSBAC1	86	ORF2	45	756	711	756	s	YES
ESTGSSBAC1	86	ORF1	1	756	755	756	r	YES
ESTGSSBAC1	86	ORF2	1	725	724	756	r	YES
ESTGSSBAC1	88	ORF1	1	960	959	960	s	YES
ESTGSSBAC1	99	ORF3	1	939	938	939	s	YES
ESTGSSBAC1	102	ORF1	1	874	873	874	s	YES
ESTGSSBAC1	116	ORF1	282	1086	804	1086	s	YES
ESTGSSBAC1	116	ORF2	341	1086	745	1086	s	YES
ESTGSSBAC1	116	ORF3	341	1086	745	1086	s	YES
ESTGSSBAC1	116	ORF1	1	745	744	1086	r	YES
ESTGSSBAC1	116	ORF2	1	779	778	1086	r	YES
ESTGSSBAC1	116	ORF3	1	774	773	1086	r	YES
ESTGSSBAC1	118	ORF3	1	1057	1056	1057	s	YES
ESTGSSBAC1	130	ORF1	123	873	750	873	s	YES
ESTGSSBAC1	130	ORF3	138	873	735	873	s	YES
ESTGSSBAC1	135	ORF3	1	712	711	712	s	YES
ESTGSSBAC1	142	ORF1	1	757	756	768	r	YES

Figure 24: The result page when user has chosen to look for orf's > 700 bp in the ESTGSSBAC1 assembly.

5. Conclusions

This bioinformatic pipeline created in this project is an efficient way to do a large-scale analysis of the *T.cruzi* sequence data and to organise all the data and it's annotations in a database. The pipeline has been constructed in a way to make it easy to update the information in the database when new data has been added and another assembly has been made.

This approach to organise sequence data and its annotations is quite common in a genome project. It can be seen as a first step on the long way to a fully annotated Chromsome 3 and finally the major part of the *T.cruzi* genome.

6. Acknowledgement

I would like to thank my supervisors Daniel Nilsson, Björn Andersson and Lena Åslund for helping me and answering all my questions, and also all the people in the Björn Andersson group at the Center for Genomics and Bioinforatics for making my time here so pleasant.

7. References

[1] World Health Organisation, Burdens and trends. http://www.who.int/ctd/chagas/burdens.htm (20 Jan. 2002).

[2] The Phred/Phrap/Consed Sytem Homepage. http://www.phrap.org (12 Oct. 2001).

[3] Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER. *Nucleic Acid Research*, 23:4636-4641, 1999.

[4] Salzberg SL, Delcher AL, Kasif S, White O. Microbial gene identification using interpolated Markov models. *Nucleic Acid Research*, 2:544-548, 1998.

[5] Fickett JW. Recognition of protein coding regions in DNA sequences. *Nucleic Acids Research*, 10:5303-5318, 1982.

[6] Altschul SF, Gish W. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403-410, 1990.

[7] Zingales B, Rondinelli E, Degrave W, daSilveira JF, Levin M, Le Paslier D, Modabber F, Dobrokhotov B, Swindle J, Kelly JM. The *Trypanosoma cruzi* genome initiative. *Parasitology Today*, 13:16–22, 1997.

[8] National Center for Infectious Diseases at the Centers for Disease Control. www.dpd.cdc.gov (20 Jan. 2002).

[9] De Souza W. “From the cell biology to the development of new chemotherapeutic approaches against trypanosomatids:dreams and reality.” *Kinetoplastid Biology and Disease* 1:3 2002. http://www.kinetoplastids.com/content/1/1/3 (21 Jan. 2002).

[10] Myler PJ, Audleman L, deVos T, Hixson G, Kiser P, Lemley C, Magness C, Rickel E, Sisk E, Sunkin S, Swartzell S, Westlake T, Bastien P, Fu G, Ivens A, Stuart K. Leishmania major friedlin chromosome 1 has an unusual distribution of protein coding genes. *Genetics,* 96:2902-2906, 1999.

[11] Freitas-Junior LHG, Briones MRS, Schenkman S. Two distinct groups of mucin-like genes are differentially expressed in the developmental stages of *Trypanosoma Cruzi*. *Molecular and Biochemical Parasitology*, 93:101-114, 1998.

[12] Di Noia JM, Orso I, Sánchez DO, Frasch AC. AU-rich elements in the 3'-untranslated region of a new mucin-type gene family of *Trypanosoma cruzi* confers mRNA instability and modulates translation efficiency. *Journal of Biological Chemistry*, 275:10218-10227, 2000.

[13] Coughlin BC, Teixeira SM, Kirchhoff LV, Donelson JE. Amastin mRNA abundance in trypanosoma cruzi is controlled by a 3'-untranslated region: Position-dependent cis-element and an untranslated region-binding protein. *Journal of Biological Chemistry*, 275:12051-12060, 2000.

[14] Andersson Björn, Aslund L, Tammi M, Tran AN, Hoheisel JD, Pettersson U. Complete sequence of a 93.4-kb contig from chromosome 3 of *Trypanosoma cruzi* containing a strand-switch region. *Genome research*, 8:809-816, 1998.

[15] Requena JM, Lopez MC, Alonso C. Genomic repetitive DNA elements of *Trypanosoma cruzi. Parasitology Today,* 12:279–283, 1996.

[16] Henriksson J, Porcel B, Rydaker M, Ruiz A, Sabaj V, Galanti N, Cazzulo JJ, Frasch AC, Pettersson U. Chromosome specific markers reveal conserved linkage groups in spite of extensive chromosomal size variation in *Trypanosoma cruzi*. *Molecular and Biochemical Parasitology,* 73:63-74, 1995.

[17] Ewing B, Hillier L, Wendl MC, Green P. Base-Calling of automated sequencer traces using phred I. Accuracy assessment. *Genome Research*, 8:175-185, 1998.

[18] Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, 8:186-94, 1998.

[19] Urmenyi TP, Bonaldo MF, Soares MB, Rondinelli E. Construction of a normalized cDNA library for the Trypanosoma cruzi genome project. *Journal of Eukaryotic Microbiology,* 46:542-544, 1999.

[20] Gibas C, Jambeck P. *Developing Bioinformatic Computer Skills*, 2001, O'Reilly & Associates Inc., Sebastopol, ISBN 1-56592-664-1

[21] Gordon D, Abajian C, Green P. Consed: a graphical tool for sequence finishing. *Genome Research*, 8:195-202, 1998.

[22] Green P, Lecture slides. www.cs.technion.ac.il/Labs/cbl/teaching/uw/lect12.pdf (17 Oct. 2001).

[23] Porcel BM, Tran AN, Tammi M, Nyarady Z, Rydaker M, Urmenyi TP, Rondinelli E, Pettersson U, Andersson B, Aslund L. Gene Survey of the Pathogenic Protozoan *Trypanosoma cruzi.. Genome Research,* 10:1103-1107, 2000.

[24] National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov (28 Jan. 2002).

[25] PostgresSQL. http://www.postgresql.org (13 Oct 2001).

[26] World Wide Web Consortium. http://www.w3.org (2 Feb. 2002).

[27] Php. http://www.php.net (2 Feb. 2002).

[28] Perl. http://www.perl.org (2 Feb. 2002).

[29] Perl. http://www.perl.com (2 Feb. 2002).

[30] Andrade MA, Brown NP, Leroy C, Hoersch S, de Daruvar A, Reich C, Franchini A, Tamames J, Valencia A, Ouzounis C, Sander C. Automated genome sequence analysis and annotation. *Bioinformatics*, 15:391-412, 1999.

[31] Agüero F, Verdun RE, Frasch AC, Sanchez DO. A random sequencing approach for the analysis of the *Trypanosoma cruzi* genome: General structure, large gene and repetitive DNA families, and gene discovery. *Genome Research,* 10:1996-2005, 2000.