ABSTRACT

UTR PREDICTION PROGRAMS FOR TRYPANOSOMA BRUCEI

by

Maria Moutafis

In the past few years, the field of bioinformatics has seen a rapid increase in the need for use of various sequence analysis tools. As we advance in the fields of science and technology, new programs and software are constantly being developed in this field. Rapidly expanding gene sequence databases and rapidly evolving sequence analysis tools are providing researchers with ways to search for highly similar query sequences whether they are nucleotide, protein, or gene databases. This thesis will focus on sequence alignment tools, specifically concentrating on tools that help determine/predict non-coding regions of sequences also known as untranslated regions of sequence or UTRs. These regions tend to exhibit less cross-species conservation than do coding sequences.

There are many software packages available today for use in determining UTRs, most of which are only available for use by not-for-profit organizations such as universities. Such programs include BLAST and ORBIT. However, not all tools are the same, requiring the researcher to modify each query accordingly. In addition, not all tools give the same results back to a specific query, some lack flexibility, and others lack user-friendliness. So how does a scientist know which one to choose? Which one is more accurate and how often is it accurate? These are the questions this thesis will answer, by running sequences on both platforms and checking the results against PatSearch.

UTR PREDICTION PROGRAMS FOR TRYPANOSOMA BRUCEI

by

Maria Moutafis

A Thesis
Submitted to the Faculty of
New Jersey Institute of Technology
In Partial Fulfillment of the Requirements for the Degree of
Master of Science in Computational Biology

Department of Computer Science

January 2008

APPROVAL PAGE

UTR PREDICTION PROGRAMS FOR TRYPANOSOMA BRUCEI

Maria Moutafis

Jason T. L. Wang, Ph.D., Thesis Advisor — Date
Professor, Department of Computer Science, NJIT

Vincent Oria, Ph.D., Committee Member — Date
Associate Professor, Department of Computer Science, NJIT

Dimitri Theodoratos, Ph.D., Committee Member — Date
Associate Professor, Department of Computer Science, NJIT

BIOGRAPHICAL SKETCH

Author: Maria Moutafis

Degree: Master of Science

Date: January 2008

Undergraduate and Graduate Education:

- Master of Science in Computational Biology,
  New Jersey Institute of Technology, Newark, NJ, 2008

- Bachelor of Arts in Biology,
  Rutgers University, New Brunswick, NJ, 2001

Major: Computational Biology

Presentations and Publications:

Wilson, G. M., Sutphen, K., Moutafis, M., Sinha, S., & Brewer, G. (2001). Structural remodeling of an A+U-rich RNA element by cation or AUF1 binding. *Journal of Biological Chemistry, 276,* 38400-38409.

To my beloved husband, parents, and brother, thank you for standing by me and supporting me through everything, I love you all very much.

Σά βγείς στόν πηγαιμό γιά την 'Ιθάκη,
νά εύχεσαι να'ναι μακρύς ο δρόμος,
γεμάτος περιπέτειες, γεμάτος γνώσεις....

Πάντα στόν νου σου νά'χεις την 'Ιθάκη.
Τό φθάσιμον εκεί ειν' ο προορισμός σου.
'Αλλά μη βιάζεις το ταξείδι διόλου.
Καλλίτερα χρόνια πολλά νά διαρκέσει
και γέρος πιά ν'αράξεις στο νησί,
πλούσιος μέ όσα κέρδισες στον δρόμο,
μη προσδοκώντας πλούτη νά σέ δώσει η 'Ιθάκη.

-Κ. Π. Καβάφης, ΙΘΑΚΗ
-C. P. Cavafy, Ithaca

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my advisor Dr. Jason T.L. Wang for his guidance and support throughout this process; thank you for believing in me. I would also like to thank Dr. Vincent Oria and Dr. Dimitri Theodoratos for taking the time to participate on my thesis committee. A special thank you is also given to my fellow student, Vandana Patel in the Computer Science Department at NJIT, for her assistance and insight on UTR prediction software, computer, and statistical aspects of this project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES
(Continued)

Figure **Page**

CHAPTER 1

INTRODUCTION

1.1 Basics of Cell Biology

The building block of all organisms is the cell, hence the reason it is widely known in the world of science as the fundamental unit of life. It's hard to fathom that something so small and invisible to the naked eye can hold so much information and be responsible for so many functions, yet it is true. Cells utilize genetic information to guide the production of most of the cell's components. This genetic information is stored in molecules of DNA and is duplicated prior to cell division. By doing so, the cell ensures that each newly formed "daughter" cell inherits a complete set of genetic commands.

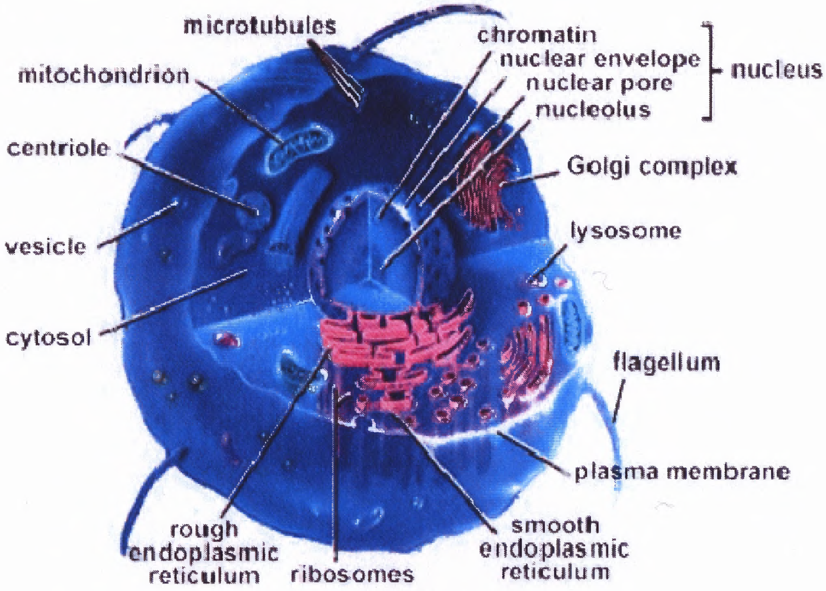


Figure 1.1 Eukaryotic cell - Illustrates a typical eukaryotic cell and its major organelles.
http://www.modares.ac.ir/elearning/mnaderi/Genetic%20Engineering%20course%20II/images/wpe16.gif

Humans are eukaryotic, meaning that their genetic material is organized into a membrane bound nucleus in the cell. The cell's genetic information is stored in DNA (deoxyribonucleic acid) molecules which are contained in an intertwined mass of fibers in the nucleus known as chromatin. When the cell is ready to divide, the chromatin fibers condense into compact structure known as chromosomes. This makes it easier for the cell to provide its daughter cells with a copy of the genetic information. To better understand how genetic information is read and processed, it is necessary to first become aware of what it is made of.

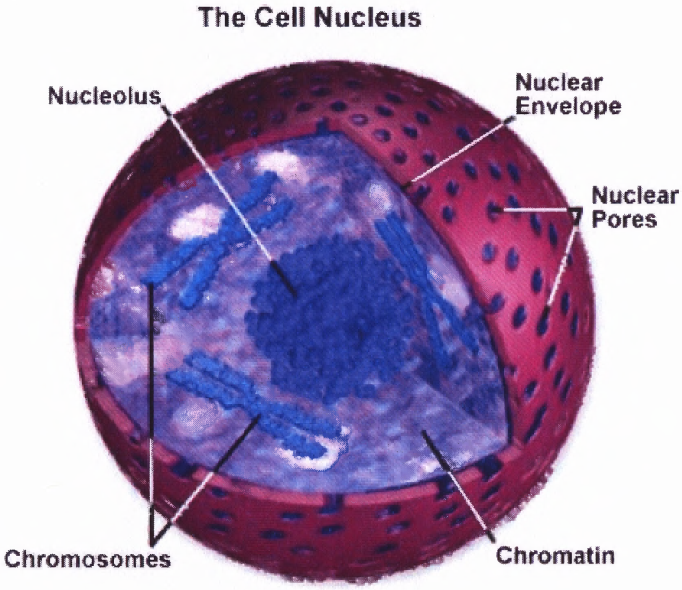


Figure 1.2 Nucleus of the cell - Three dimensional model of a nucleus filled with genetic information.
http://micro.magnet.fsu.edu/cells/plants/images/cellnucleus.jpg

DNA is a double helix twisted chain that is comprised of sugar-phosphate backbones located on the outside of the helix and nitrogenous bases projecting toward the center that are held together by hydrogen bonds. It consists of four nitrogenous bases: adenine (A), thymine (T), guanine (G), and cytosine (C). In normal DNA, adenine pairs with thymine and guanine pairs with cytosine (Kleinsmith & Kish, 1995). The two chains that comprise the helix run in opposite directions, meaning that one chain runs from 5′ → 3′ and the other runs from 3′ → 5′ and therefore are complementary to each other. They key factor of DNA is its nitrogenous bases. It is the sequence of these bases that is used to encode the genetic information.

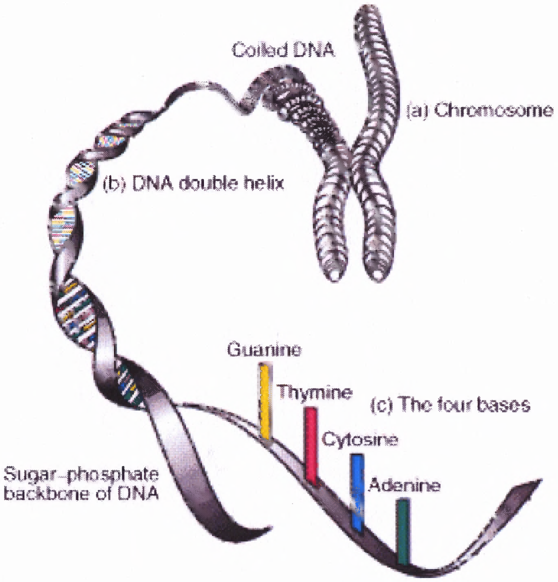


Figure 1.3 The structure of a chromosome – condensed chromatin fibers form the chromosome which is made of coiled DNA.
http://www.llnl.gov/str/June03/gifs/Stubbs1.gif

RNA is the second type of nucleic acid found in cells and is very similar to DNA, differing structurally only in two ways. It too is comprised of sugar-phosphate backbone however; instead of the sugar being deoxyribose the sugar found in RNA is ribose. The other difference pertains to the nitrogenous bases. Adenine, guanine, and cytosine are present in both DNA and RNA, but they differ in the last base. For RNA, the fourth base is uracil (U) and not thymine. Another important difference between DNA and RNA is the size of the chains. Typically, RNA chains have lengths hundreds or thousands of bases long where as DNA chains can be millions of bases long.

1.2 What are Proteins?

Proteins are very diverse macromolecules that are essential to all living cells in order to function properly. Each protein is made by the joining together of a unique combination amino acids. There are twenty amino acids needed for the formation of proteins, which can be seen listed in the table below. The process of protein formation known as translation and is an extremely complex process that takes place in the ribosome of the cell. In order to better understand the concept, a review of the basic process is needed.

Table 1.1 List of 20 Amino Acids

Glycine (Gly) G	Alanine (Ala) A	Valine (Val) V	Leucine (Leu) L	Isoleucine (Ile) I
Phenylalanine (Phe) F	Tryptophan (Trp) W	Methionine (Met) M	Proline (Pro) P	Serine (Ser) S
Threonine (Thr) T	Cysteine (Cys) C	Tyrosine (Tyr) Y	Asparagine (Asn) N	Glutamine (Gln) Q
Lysine (Lys) K	Arginine (Arg) R	Histidine (His) H	Aspartic Acid (Asp) D	Glutamic Acid (Glu) E

As described in the previous section, DNA contains genetic material that needs to be encoded. Each protein-coding gene is transcribed into a short template known as messenger RNA or mRNA for short. The mRNA is what enters the ribosome and is "translated" into an amino acid chain also known as a polypeptide. The process also needs the assistance of another RNA, tRNA or transfer RNAs that are specific for individual amino acids. tRNAs have anticodons complementary to the codons in mRNA and can be "charged" covalently with amino acids at their 3′ terminal ends (Kleinsmith & Kish, 1995). There are 64 different codon combinations possible with a triplet codon of three RNA nucleotides. This can be seen visually in Figure 1.4.

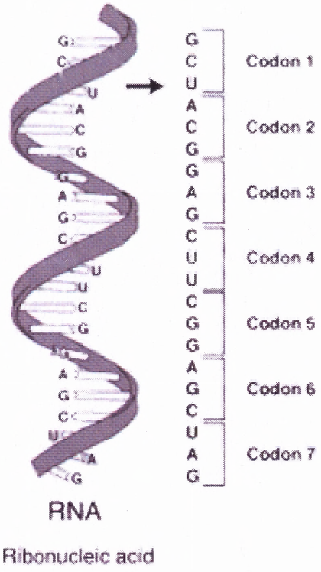


Figure 1.4 RNA codons – depicts how a RNA strand would be read by the ribosome. http://encyc.connectonline.com/encyclopedia/images/thumb/c/cd/RNA-codon.png/180px-RNA-codon.png

All 64 codons of the standard genetic code are assigned for either amino acids or stop signals. For example, when an RNA fragment sequence of ACGUCAGCC is read beginning with the A (always reading from the 5′ end to the 3′ end), it is noted that there are three codons; ACG, UCA and GCC. Each of these specifies one amino acid, in this case, threonine, serine, and alanine. There are four codons that stand apart from the rest because they serve a different role than the other codons. The codon AUG in addition to coding for methionine, also serves as the most notable "start" codon. It along with other initiation factors triggers the beginning of translation. The other three codons are UGA, UAG, and UAA which all code for a "stop" codon that initiates termination of translation. Table 1.2 lists all the possible codons and the amino acid they translate into. The proteins form a linear sequence that will fold and take shape to make the structural protein.

Table 1.2 List of RNA Codons

Ala	A	GCU, GCC, GCA, GCG	**Leu**	L	UUA, UUG, CUU, CUC, CUA, CUG
Arg	R	CGU, CGC, CGA, CGG, AGA, AGG	**Lys**	K	AAA, AAG
Asn	N	AAU, AAC	**Met**	M	AUG
Asp	D	GAU, GAC	**Phe**	F	UUU, UUC
Cys	C	UGU, UGC	**Pro**	P	CCU, CCC, CCA, CCG
Gln	Q	CAA, CAG	**Ser**	S	UCU, UCC, UCA, UCG, AGU,AGC
Glu	E	GAA, GAG	**Thr**	T	ACU, ACC, ACA, ACG
Gly	G	GGU, GGC, GGA, GGG	**Trp**	W	UGG
His	H	CAU, CAC	**Tyr**	Y	UAU, UAC
Ile	I	AUU, AUC, AUA	**Val**	V	GUU, GUC, GUA, GUG
Start		AUG, GUG	***Stop***		UAG, UGA, UAA

CHAPTER 2

SIGNIFICANCE OF T. BRUCEI

2.1 Why the need to study parasites?

Parasites are organisms that need to live on other organisms (hosts) in order to survive. The study of parasites, their hosts and the relationship between them is called parasitology. Most of the research regarding parasites falls into three fields: Medical, Veterinary, and Ecology. The largest of the three fields is medical parasitology which deals with parasites that infect humans. These parasites include single-celled protozoa and multicellular worms which range in size from those that are microscopic to those that are visible by the human eye. The medical field of parasitology is the field that interests us and is the basis for this project. But why is there a need to study parasites and their hosts?

Parasitic infections plague thousands of people, cattle and other domesticated animals yearly, especially those living in Africa and South America. Of importance to us is the disease caused by Trypanosoma brucei, called African trypanosomiasis or African sleeping sickness. The disease is debilitating and lethal and yet currently there are no vaccines for it and no modern drugs to treat it. The few drugs that do exist are all very toxic to humans. The big pharmaceutical companies lack the motivation to help in this field mostly due to the sheer location of the disease and hence, the inability of people in those areas to pay for their drugs and vaccines. Those in the academic area have stepped up to help. They are putting their resources to work to try and understand how the species replicates and survives.

2.2 Why T. brucei?

Trypanosomes have been around for millions of years and have evolved as their hosts have in order to guarantee their survival. This parasite undergoes complex changes in order to ensure its survival in both its insect host and its mammalian host. The insect vector for T. brucei is the tsetse fly. It lives in the gut of the fly until it is ready to migrate to the fly's salivary glands, from where it will be injected into the mammalian host. In the mammal, it will live in the bloodstream where it can reinfect the fly after biting and also be transferred to other humans through bodily fluid exchange. Of its cell structure, the most notable feature is its variable surface glycoprotein (VSG) coat which it uses to avoid detection by the host's immune system.

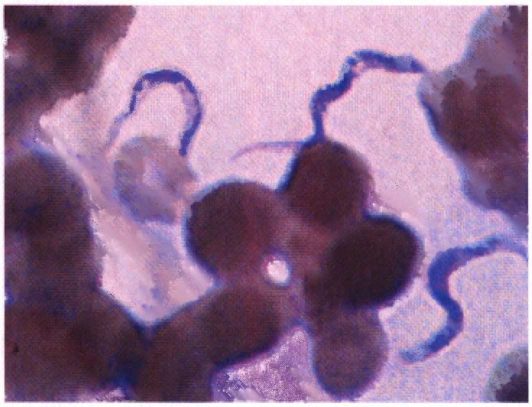


Figure 2.1 Image of Trypanosoma brucei, bloodstream form – blood smear from a patient with African trypanosomiasis.
http://www.dpd.cdc.gov/dpdx/HTML/TrypanosomiasisAfrican.htm

However, the parasite's complex life cycle and cell structure are not the only things that have captured the interest of scientists. The parasite also has a unique genetic trait as well. Unlike eukaryotic cells which generate mRNA from pre-mRNA by cis-splicing, T. brucei uses trans-splicing. This splicing fuses a short capped RNA (called a spliced leader) to each protein coding pre-mRNA. The 5′ end of the spliced leader is then added to the pre-mRNA. This means that all the mature mRNA in T. brucei will begin with the same nontranslated leader sequence (Benz et al., 2005). By researching the uniqueness of the parasite's traits, researchers are hoping they will be able to come up with ways to diminish this disease.

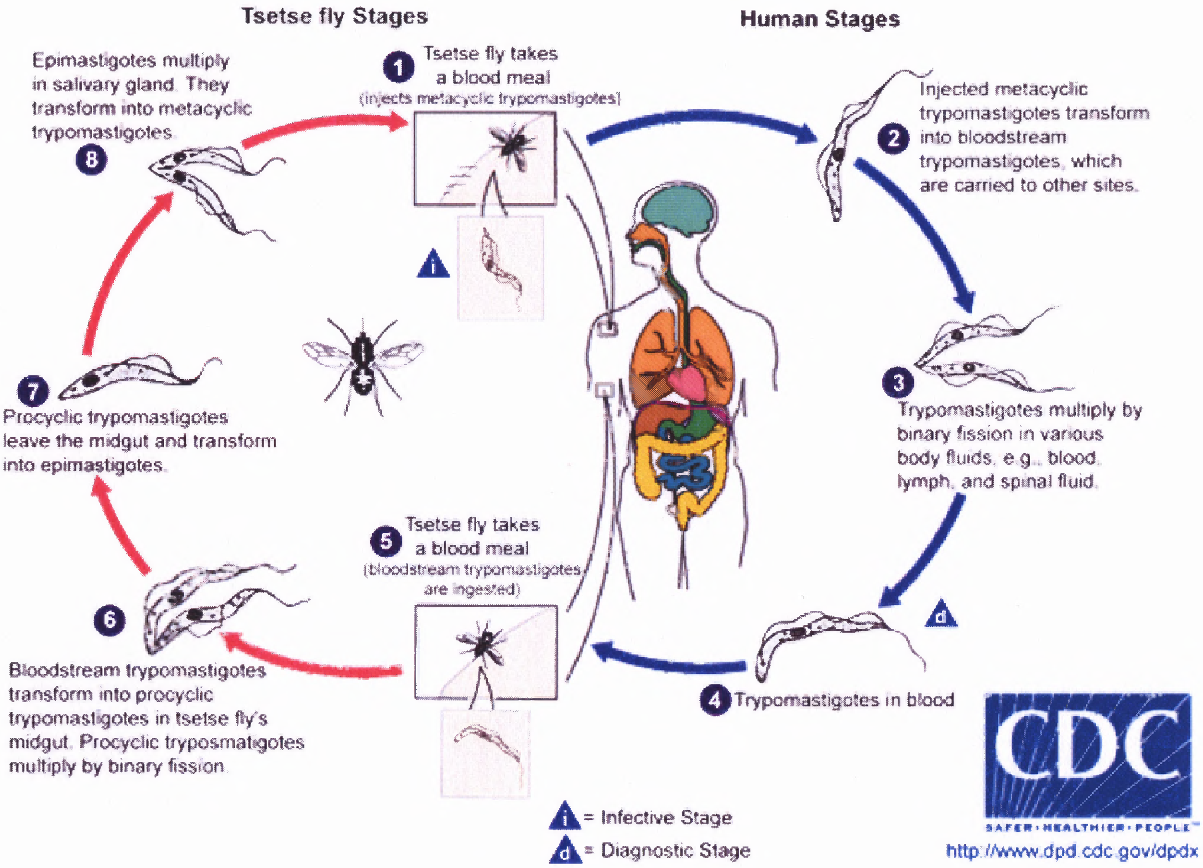

Tsetse fly Stages
Human Stages
1 Tsetse fly takes a blood meal (injects metacyclic trypomastigotes)
2 Injected metacyclic trypomastigotes transform into bloodstream trypomastigotes, which are carried to other sites.
3 Trypomastigotes multiply by binary fission in various body fluids, e.g., blood, lymph, and spinal fluid.
4 Trypomastigotes in blood
5 Tsetse fly takes a blood meal (bloodstream trypomastigotes are ingested)
6 Bloodstream trypomastigotes transform into procyclic trypomastigotes in tsetse fly's midgut. Procyclic tryposmatigotes multiply by binary fission.
7 Procyclic trypomastigotes leave the midgut and transform into epimastigotes.
8 Epimastigotes multiply in salivary gland. They transform into metacyclic trypomastigotes.
i
d
i = Infective Stage
d = Diagnostic Stage
CDC
SAFER·HEALTHIER·PEOPLE™
http://www.dpd.cdc.gov/dpdx

Figure 2.2 Life cycle of Trypanosoma brucei – goes through the infectious and disease stages of the parasite.
http://www.dpd.cdc.gov/dpdx/HTML/TrypanosomiasisAfrican.htm

CHAPTER 3

DATA COLLECTION

3.1 GeneDB and Data Collection

The first part of this project entailed finding the mRNA sequences of T. brucei. This data was obtained through the use of GeneDB (http://www.genedb.org/). Initially, GeneDB began as a way to develop and maintain a database resource for three organisms: *Schizosaccharomyces pombe*, *Leishmania major* and *Trypanosoma brucei*. Currently, GeneDB holds data for both eukaryotic and prokaryotic organisms. The data within GeneDB are manually annotated, frequently updated and easy to precisely query. Under the site's homepage, there are three columns: searches, sequence searches, and datasets.

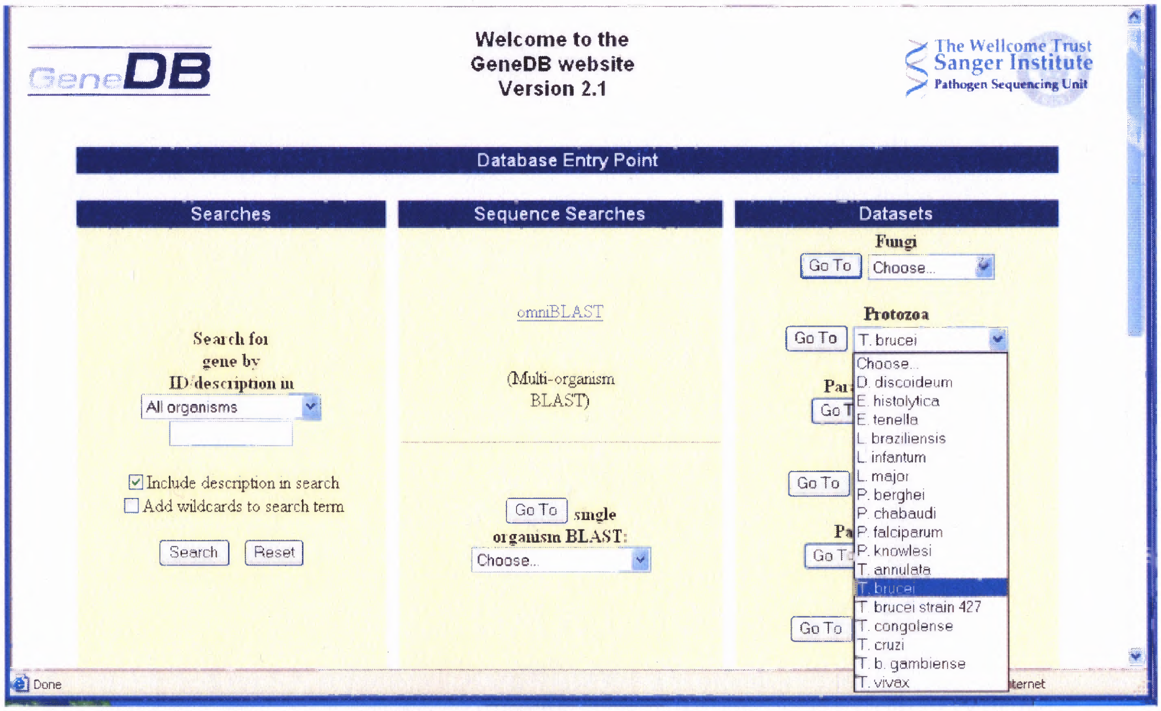
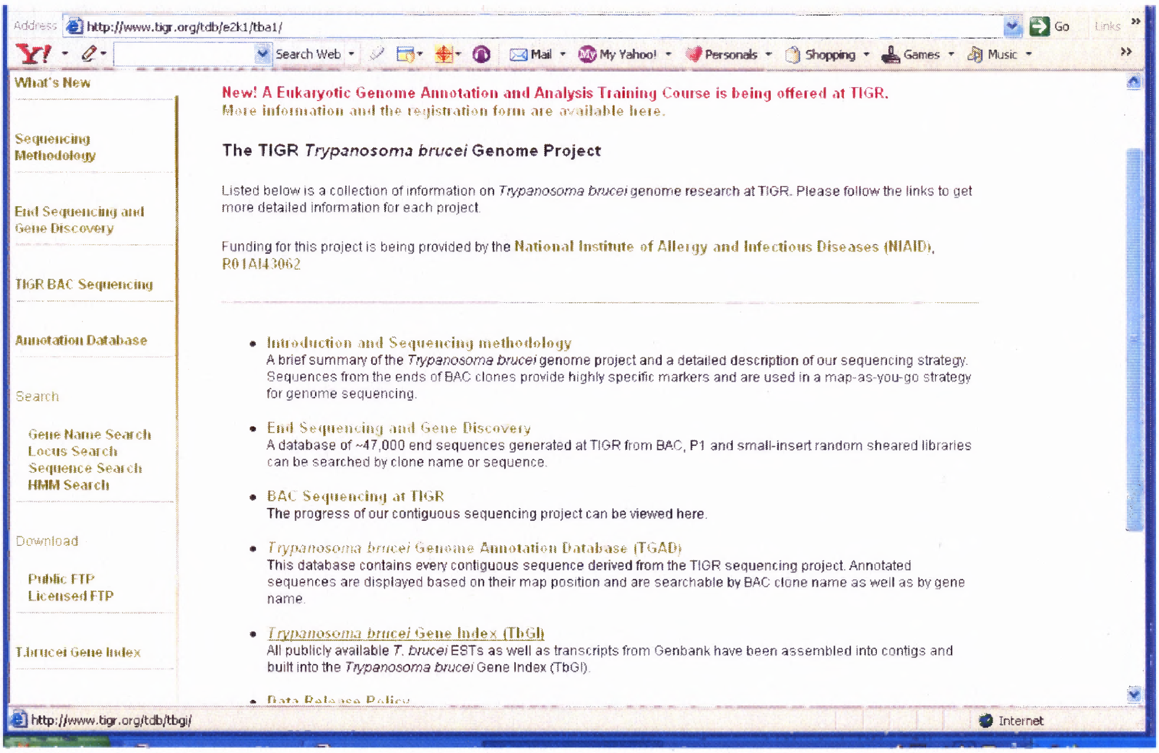


Figure 3.1 Homepage of GeneDB.
http://www.genedb.org/

Choosing T. brucei from the Protozoa Section's pull-down menu found under "Datasets" brings you to the T. brucei specific site, http://www.genedb.org/genedb/tryp/. Once at the home page of T. brucei, scroll to the bottom of the page to see a section titled "Links" and click on "TIGR T. brucei project". This will direct you to the Trypanosoma brucei Genome Project page of the TIGR database, http://www.tigr.org/tdb/e2k1/tba1/. TIGR (The Institute for Genomic Research) is a not-for-profit research center that is committed to the interpreting and analyzing of genomes. Once on this page, select the Heading "Trypanosoma brucei Gene Index (TbGI)".



New! A Eukaryotic Genome Annotation and Analysis Training Course is being offered at TIGR.
More information and the registration form are available here.

The TIGR *Trypanosoma brucei* Genome Project

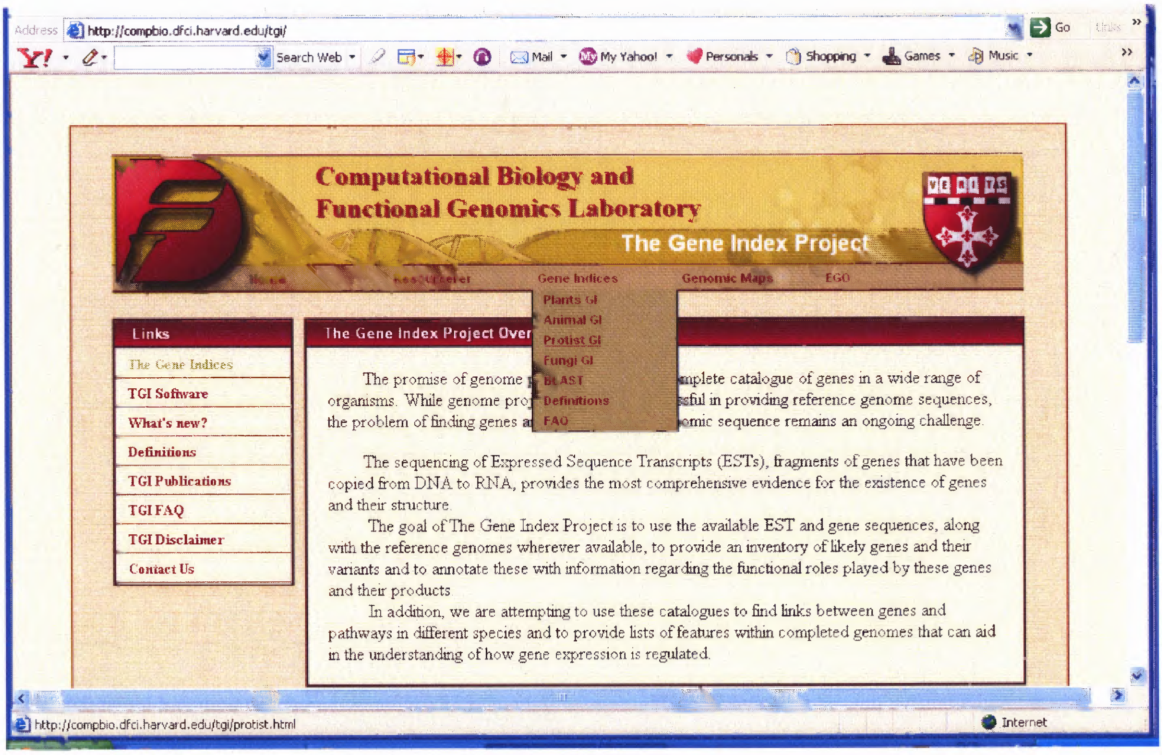
Listed below is a collection of information on *Trypanosoma brucei* genome research at TIGR. Please follow the links to get more detailed information for each project.

Funding for this project is being provided by the National Institute of Allergy and Infectious Diseases (NIAID), R01AI43062

- Introduction and Sequencing methodology
  A brief summary of the *Trypanosoma brucei* genome project and a detailed description of our sequencing strategy. Sequences from the ends of BAC clones provide highly specific markers and are used in a map-as-you-go strategy for genome sequencing.
- End Sequencing and Gene Discovery
  A database of ~47,000 end sequences generated at TIGR from BAC, P1 and small-insert random sheared libraries can be searched by clone name or sequence.
- BAC Sequencing at TIGR
  The progress of our contiguous sequencing project can be viewed here.
- *Trypanosoma brucei* Genome Annotation Database (TGAD)
  This database contains every contiguous sequence derived from the TIGR sequencing project. Annotated sequences are displayed based on their map position and are searchable by BAC clone name as well as by gene name.
- *Trypanosoma brucei* Gene Index (TbGI)
  All publicly available *T. brucei* ESTs as well as transcripts from Genbank have been assembled into contigs and built into the *Trypanosoma brucei* Gene Index (TbGI).
- Data Release Policy

Figure 3.2 Homepage of TIGR T. brucei project.
http://www.tigr.org/tdb/e2k1/tba1/

At this point, TIGR will automatically redirect you to a new page for the Gene Index Project at Harvard's Computational Biology and Functional Genomics Laboratory at http://compbio.dfci.harvard.edu/tgi/. Under the header Gene Indices, there is a drop down menu from where you will choose Protist GI. Selecting T. brucei from the list of protists will redirect the user to a site that is specific to T. brucei, http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=t_brucei. From the category titled Sequence Reports, the user should select CAT# Download which finally directs the site to the TbGI sequence library. From this point, the user may obtain the ESTs for T. brucei. ESTs are expressed sequence transcripts that are fragments of genes that have been copied from DNA to RNA. Check off the EST box under Sequence Type and one by one choose from the five catalog numbers. Once selected, click the download sequence button and you will be given the choice to either open or save the sequences to file.



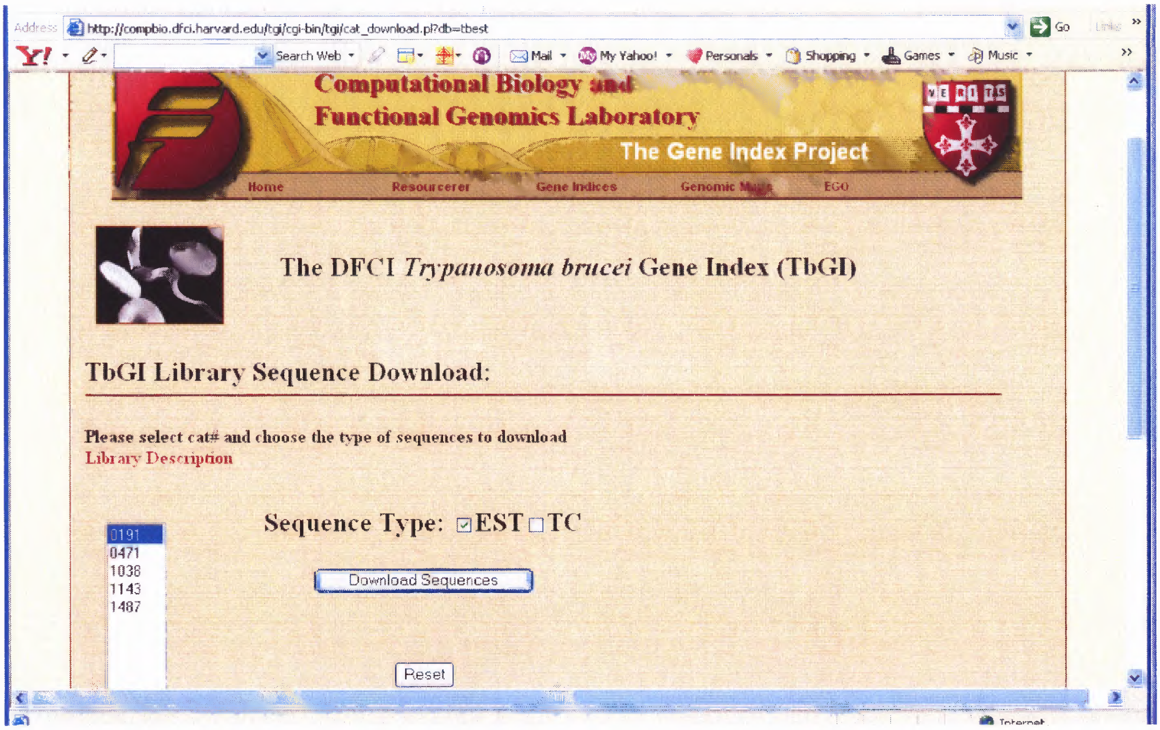
The promise of genome p… …mplete catalogue of genes in a wide range of organisms. While genome pro… …ssful in providing reference genome sequences, the problem of finding genes a… …omic sequence remains an ongoing challenge.

The sequencing of Expressed Sequence Transcripts (ESTs), fragments of genes that have been copied from DNA to RNA, provides the most comprehensive evidence for the existence of genes and their structure.

The goal of The Gene Index Project is to use the available EST and gene sequences, along with the reference genomes wherever available, to provide an inventory of likely genes and their variants and to annotate these with information regarding the functional roles played by these genes and their products.

In addition, we are attempting to use these catalogues to find links between genes and pathways in different species and to provide lists of features within completed genomes that can aid in the understanding of how gene expression is regulated.

Figure 3.3 Homepage of The Gene Index Project.
http://compbio.dfci.harvard.edu/tgi/



http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/cat_download.pl?db=tbest

The DFCI *Trypanosoma brucei* Gene Index (TbGI)

TbGI Library Sequence Download:

Please select cat# and choose the type of sequences to download
Library Description

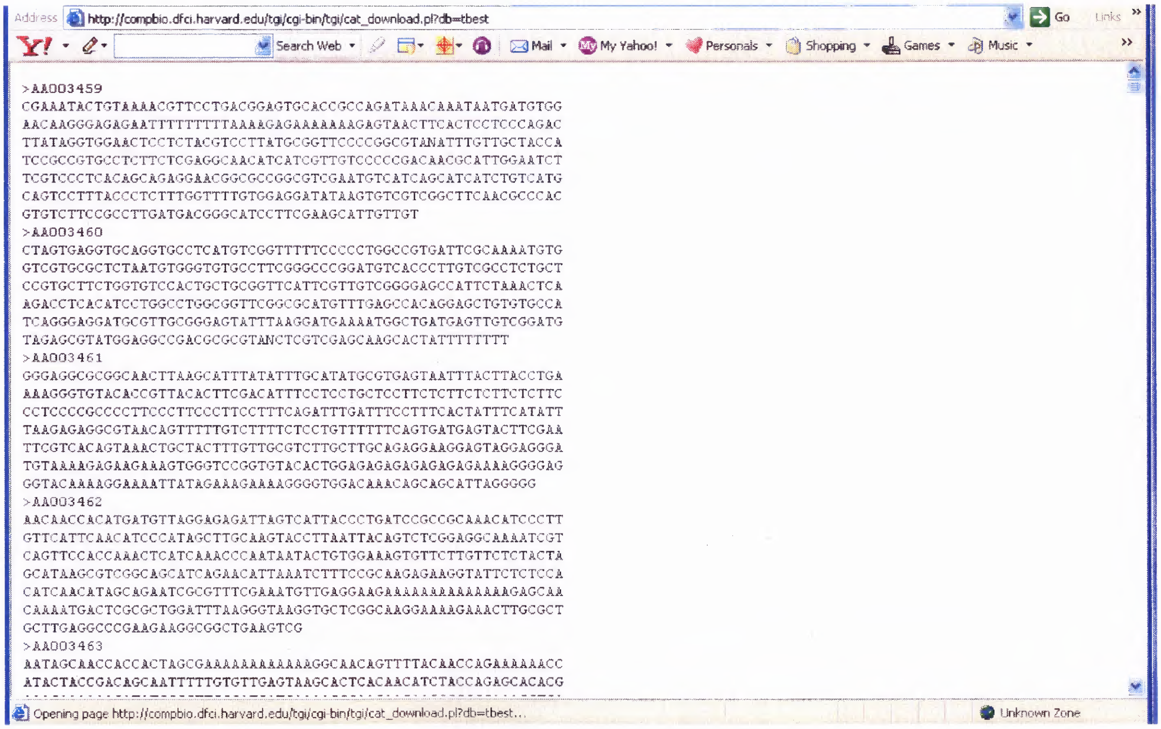
0191
0471
1038
1143
1487

Sequence Type: ☑EST □TC

Download Sequences

Reset

Figure 3.4 TbGI Library Sequence download page.
http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/cat_download.pl?db=tbest



http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/cat_download.pl?db=tbest

```
>AA003459
CGAAATACTGTAAAACGTTCCTGACGGAGTGCACCGCCAGATAAACAAATAATGATGTGG
AACAAGGGAGAGAATTTTTTTTTAAAAGAGAAAAAAAAGAGTAACTTCACTCCTCCCAGAC
TTATAGGTGGAACTCCTCTACGTCCTTATGCGGTTCCCCGGCGTANATTTGTTGCTACCA
TCCGCCGTGCCTCTTCTCGAGGCAACATCATCGTTGTCCCCCGACAACGCATTGGAATCT
TCGTCCCTCACAGCAGAGGAACGGCGCCGGCGTCGAATGTCATCAGCATCATCTGTCATG
CAGTCCTTTACCCTCTTTGGTTTTGTGGAGGATATAAGTGTCGTCGGCTTCAACGCCCAC
GTGTCTTCCGCCTTGATGACGGGCATCCTTCGAAGCATTGTTGT
>AA003460
CTAGTGAGGTGCAGGTGCCTCATGTCGGTTTTTCCCCCTGGCCGTGATTCGCAAAATGTG
GTCGTGCGCTCTAATGTGGGTGTGCCTTCGGGCCCGGATGTCACCCTTGTCGCCTCTGCT
CCGTGCTTCTGGTGTCCACTGCTGCGGTTCATTCGTTGTCGGGGAGCCATTCTAAACTCA
AGACCTCACATCCTGGCCTGGCGGTTCGGCGCATGTTTGAGCCACAGGAGCTGTGTGCCA
TCAGGGAGGATGCGTTGCGGGAGTATTTAAGGATGAAAATGGCTGATGAGTTGTCGGATG
TAGAGCGTATGGAGGCCGACGCGCGTANCTCGTCGAGCAAGCACTATTTTTTTTT
>AA003461
GGGAGGCGCGGCAACTTAAGCATTTATATTTGCATATGCGTGAGTAATTTACTTACCTGA
AAAGGGTGTACACCGTTACACTTCGACATTTCCTCCTGCTCCTTCTCTTCTCTTCTCTTC
CCTCCCCGCCCCTTCCCTTCCCTTCCTTTCAGATTTGATTTCCTTTCACTATTTCATATT
TAAGAGAGGCGTAACAGTTTTTGTCTTTTCTCCTGTTTTTTCAGTGATGAGTACTTCGAA
TTCGTCACAGTAAACTGCTACTTTGTTGCGTCTTGCTTGCAGAGGAAGGAGTAGGAGGGA
TGTAAAAGAGAAGAAAGTGGGTCCGGTGTACACTGGAGAGAGAGAGAGAGAAAAGGGGAG
GGTACAAAAGGAAAATTATAGAAAGAAAAGGGGTGGACAAACAGCAGCATTAGGGGG
>AA003462
AACAACCACATGATGTTAGGAGAGATTAGTCATTACCCTGATCCGCCGCAAACATCCCTT
GTTCATTCAACATCCCATAGCTTGCAAGTACCTTAATTACAGTCTCGGAGGCAAAATCGT
CAGTTCCACCAAACTCATCAAACCCAATAATACTGTGGAAAGTGTTCTTGTTCTCTACTA
GCATAAGCGTCGGCAGCATCAGAACATTAAATCTTTCCGCAAGAGAAGGTATTCTCTCCA
CATCAACATAGCAGAATCGCGTTTCGAAATGTTGAGGAAGAAAAAAAAAAAAAAAGAGCAA
CAAAATGACTCGCGCTGGATTTAAGGGTAAGGTGCTCGGCAAGGAAAAGAAACTTGCGCT
GCTTGAGGCCCGAAGAAGGCGGCTGAAGTCG
>AA003463
AATAGCAACCACCACTAGCGAAAAAAAAAAAAAGGCAACAGTTTTACAACCAGAAAAAACC
ATACTACCGACAGCAATTTTTGTGTTGAGTAAGCACTCACAACATCTACCAGAGCACACG
```

Opening page http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/cat_download.pl?db=tbest...

Figure 3.5 Example of sequence download.

3.2 ESTs

As described in the introduction chapter, each DNA must be converted into messenger RNA (mRNA) because RNA functions as a template for protein synthesis. Then, through a process called translation, the mRNA is decoded and guides the synthesis of a protein. Therefore, isolating mRNA is vital to the whole process. However, it is not so easy to do because it requires extensive processing and transporting. In addition, mRNAs vary greatly in stability, ranging from several minutes to even days. To get around this problem, scientists found a way to convert mRNA into complementary DNA (cDNA), which is much more stable and designates only expressed DNA sequence.

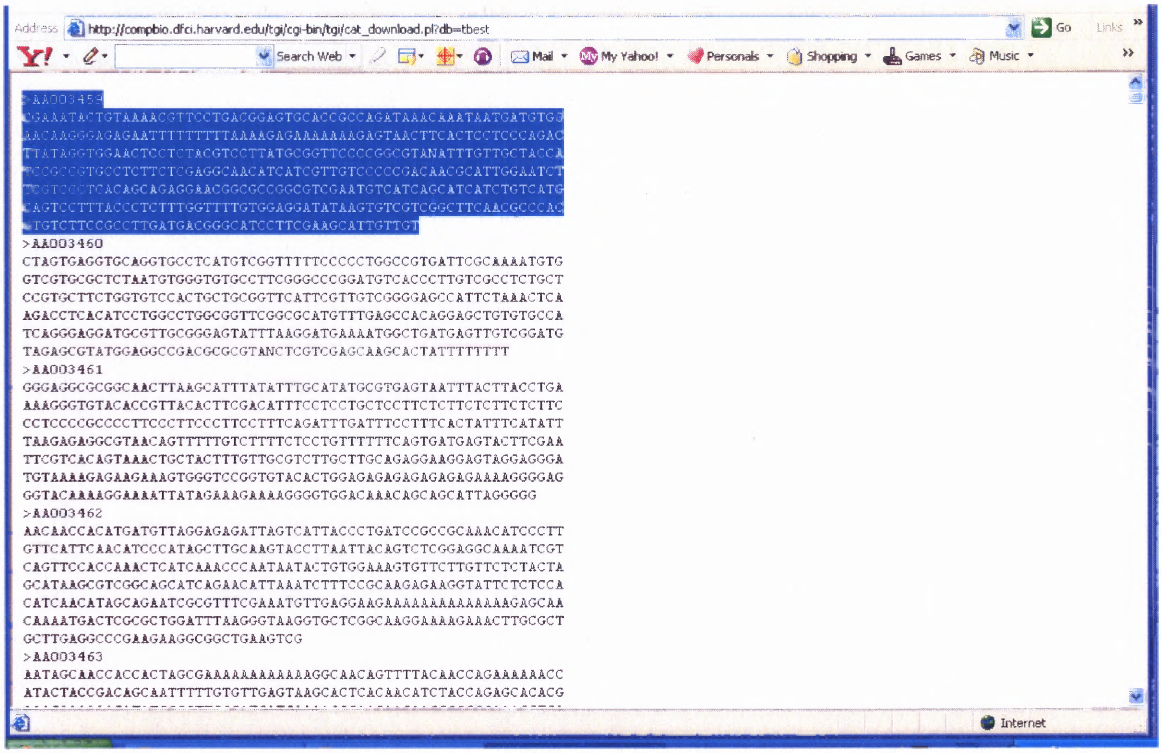
It is from this cDNA that ESTs are determined. There are two types of ESTs, each of which gets its name from the portion of the cDNA it is sequenced from: 5′ EST and a 3′ EST. A 5′ EST is obtained from the beginning segment of the cDNA and usually is the portion that codes for a protein. These regions tend to be conserved across species and do not change much within a gene family. A 3′ EST is obtained from the ending segment of the cDNA and are more likely to fall within regions that are non-coding or untranslated regions (UTRs). These regions tend to exhibit less cross-species conservation than do coding sequences (Quackenbush et al. 2000). It is these UTRs that were are interested in for this project and will be discussed again in Chapters 6-8 of this project.

CHAPTER 4

BLAST

BLAST (Basic Local Alignment Search Tool) is the best known and most frequently used tool for calculating sequence similarity. It contains more search modes than other alignment tools eliminating both the need to run multiple times and the need for use of multiple programs. Contributing factors that make it so appealing to users is its speed and ease of use. There are many variations of BLAST, each used for alignment of different query sequences against different databases, i.e. BLASTn, BLASTp, BLASTx, etc. All the variations of BLAST, information and applications on it are free and readily available on the NCBI website at http://www.ncbi.nlm.nih.gov/BLAST/. There is no need to go into details of how BLAST works because thousands of projects already have done so. This chapter will therefore, only be an explanation of the process of blasting downloaded sequences.

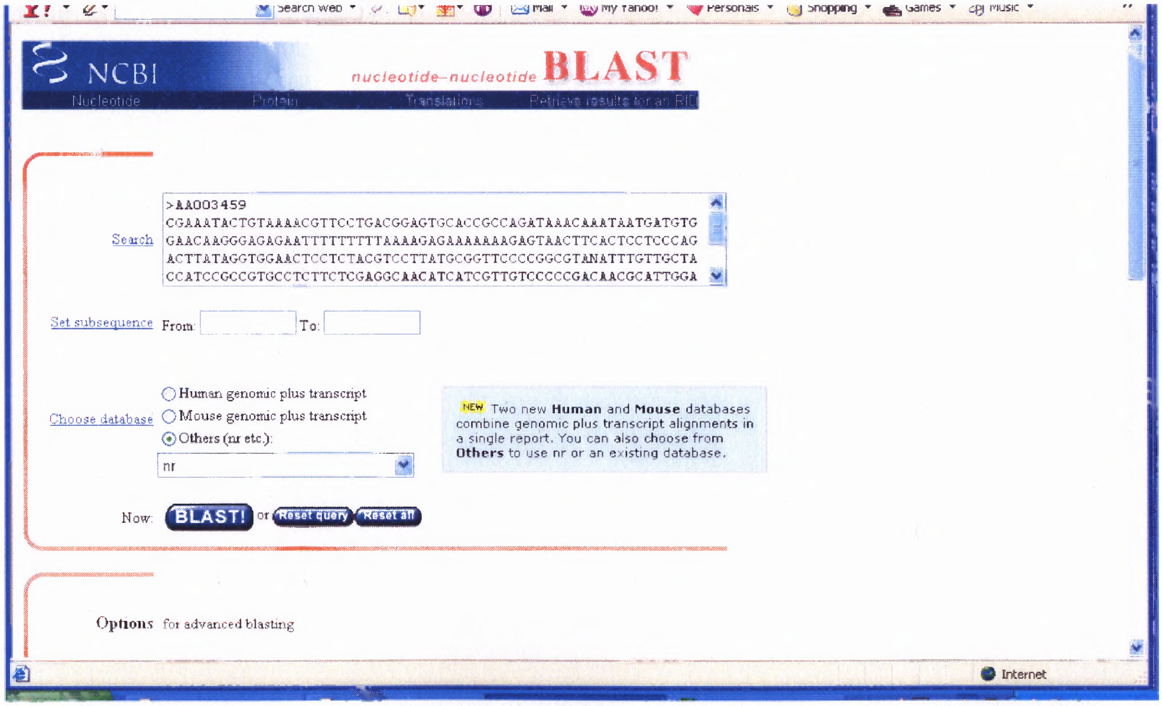
From NCBI's main BLAST page, http://www.ncbi.nlm.nih.gov/BLAST/, look for the Nucleotide sub-heading and select nucleotide-nucleotide BLAST (BLASTn). Now we are directed to the main page where you will be entering all our sequences. From the list of downloaded sequences obtained in the previous chapter, one sequence is selected and copied into the search box. It is important to remember that only one sequence can be aligned at a time. Also make note, that the database chosen must be "Others" which automatically places nr in the selection box. If this is not chosen, all queries will be made against the human database. Once the sequence is placed in the search box, click on the "BLAST!" button. Upon doing so, a message appears stating the query has been placed into the BLAST queue and a request ID is assigned to this specific query.



Address http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/cat_download.pl?db=tbest

```
>AA003459
CGAAATACTGTAAAACGTTCCTGACGGAGTGCACCGCCAGATAAACAAATAATGATGTGG
AACAAGGGAGAGAATTTTTTTTTAAAAGAGAAAAAAAGAGTAACTTCACTCCTCCCAGAC
TTATAGGTGGAACTCCTCTACGTCCTTATGCGGTTCCCCGGCGTANATTTGTTGCTACCA
TCCGCCGTGCCTCTTCTCGAGGCAACATCATCGTTGTCCCCCGACAACGCATTGGAATCT
TCGTCCCTCACAGCAGAGGAACGGCGCCGGCGTCGAATGTCATCAGCATCATCTGTCATG
CAGTCCTTTACCCTCTTTGGTTTTGTGGAGGATATAAGTGTCGTCGGCTTCAACGCCCAC
TGTCTTCCGCCTTGATGACGGGCATCCTTCGAAGCATTGTTGT
>AA003460
CTAGTGAGGTGCAGGTGCCTCATGTCGGTTTTTCCCCCTGGCCGTGATTCGCAAAATGTG
GTCGTGCGCTCTAATGTGGGTGTGCCTTCGGGCCCGGATGTCACCCTTGTCGCCTCTGCT
CCGTGCTTCTGGTGTCCACTGCTGCGGTTCATTCGTTGTCGGGGAGCCATTCTAAACTCA
AGACCTCACATCCTGGCCTGGCGGTTCGGCGCATGTTTGAGCCACAGGAGCTGTGTGCCA
TCAGGGAGGATGCGTTGCGGGAGTATTTAAGGATGAAAATGGCTGATGAGTTGTCGGATG
TAGAGCGTATGGAGGCCGACGCGCGTANCTCGTCGAGCAAGCACTATTTTTTTT
>AA003461
GGGAGGCGCGGCAACTTAAGCATTTATATTTGCATATGCGTGAGTAATTTACTTACCTGA
AAAGGGTGTACACCGTTACACTTCGACATTTCCTCCTGCTCCTTCTCTTCTCTTCTCTTC
CCTCCCCGCCCCTTCCCTTCCCTTCCTTTCAGATTTGATTTCCTTTCACTATTTCATATT
TAAGAGAGGCGTAACAGTTTTTGTCTTTTCTCCTGTTTTTTCAGTGATGAGTACTTCGAA
TTCGTCACAGTAAACTGCTACTTTGTTGCGTCTTGCTTGCAGAGGAAGGAGTAGGAGGGA
TGTAAAAGAGAAGAAAGTGGGTCCGGTGTACACTGGAGAGAGAGAGAGAGAAAAGGGGAG
GGTACAAAAGGAAAATTATAGAAAGAAAAGGGGTGGACAAACAGCAGCATTAGGGGG
>AA003462
AACAACCACATGATGTTAGGAGAGATTAGTCATTACCCTGATCCGCCGCAAACATCCCTT
GTTCATTCAACATCCCATAGCTTGCAAGTACCTTAATTACAGTCTCGGAGGCAAAATCGT
CAGTTCCACCAAACTCATCAAACCCAATAATACTGTGGAAAGTGTTCTTGTTCTCTACTA
GCATAAGCGTCGGCAGCATCAGAACATTAAATCTTTCCGCAAGAGAAGGTATTCTCTCCA
CATCAACATAGCAGAATCGCGTTTCGAAATGTTGAGGAAGAAAAAAAAAAAAAAAGAGCAA
CAAAATGACTCGCGCTGGATTTAAGGGTAAGGTGCTCGGCAAGGAAAAGAAACTTGCGCT
GCTTGAGGCCCGAAGAAGGCGGCTGAAGTCG
>AA003463
AATAGCAACCACCACTAGCGAAAAAAAAAAAAAGGCAACAGTTTTACAACCAGAAAAAACC
ATACTACCGACAGCAATTTTTGTGTTGAGTAAGCACTCACAACATCTACCAGAGCACACG
```

Figure 4.1 Downloaded sequences - example of selection of a sequence.



Search
```
>AA003459
CGAAATACTGTAAAACGTTCCTGACGGAGTGCACCGCCAGATAAACAAATAATGATGTG
GAACAAGGGAGAGAATTTTTTTTTAAAAGAGAAAAAAAGAGTAACTTCACTCCTCCCAG
ACTTATAGGTGGAACTCCTCTACGTCCTTATGCGGTTCCCCGGCGTANATTTGTTGCTA
CCATCCGCCGTGCCTCTTCTCGAGGCAACATCATCGTTGTCCCCCGACAACGCATTGGA
```

Set subsequence From: To:

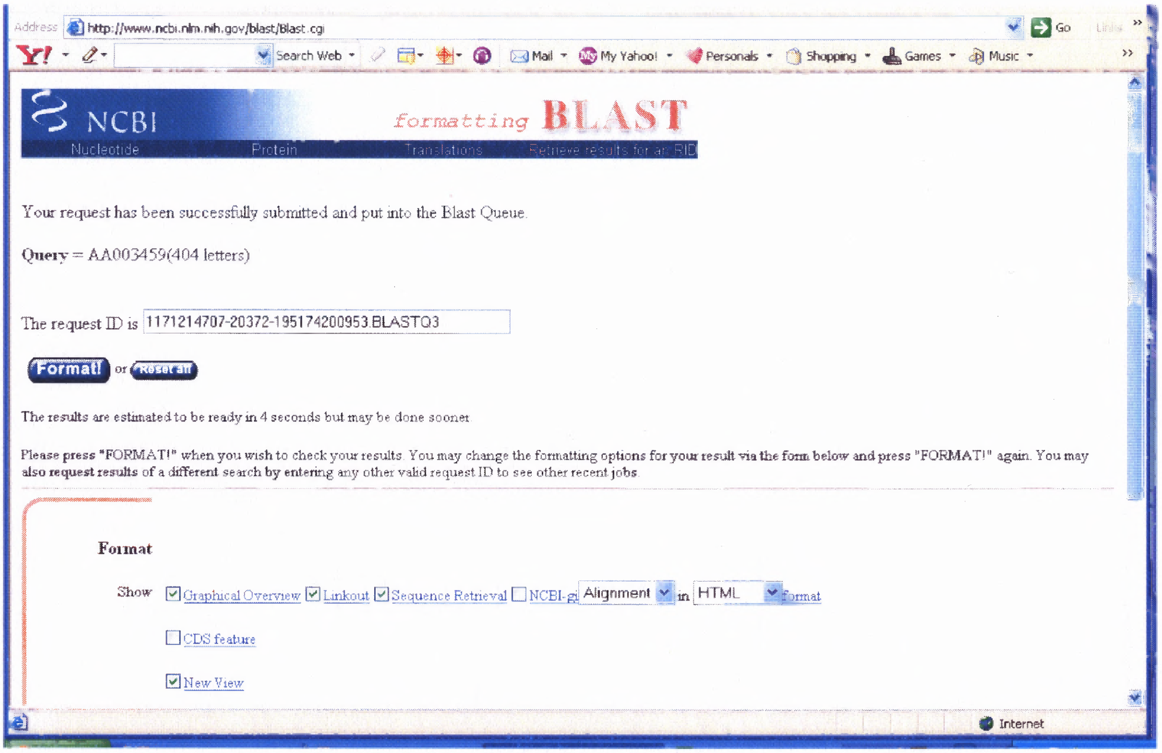
Choose database
○ Human genomic plus transcript
○ Mouse genomic plus transcript
● Others (nr etc.):
nr

NEW Two new **Human** and **Mouse** databases combine genomic plus transcript alignments in a single report. You can also choose from **Others** to use nr or an existing database.

Now: BLAST! or Reset query Reset all

Options for advanced blasting

Figure 4.2 Example of BLAST search.



Your request has been successfully submitted and put into the Blast Queue.

Query = AA003459(404 letters)

The request ID is 1171214707-20372-195174200953.BLASTQ3

Format! or Reset all

The results are estimated to be ready in 4 seconds but may be done sooner.

Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.

Format

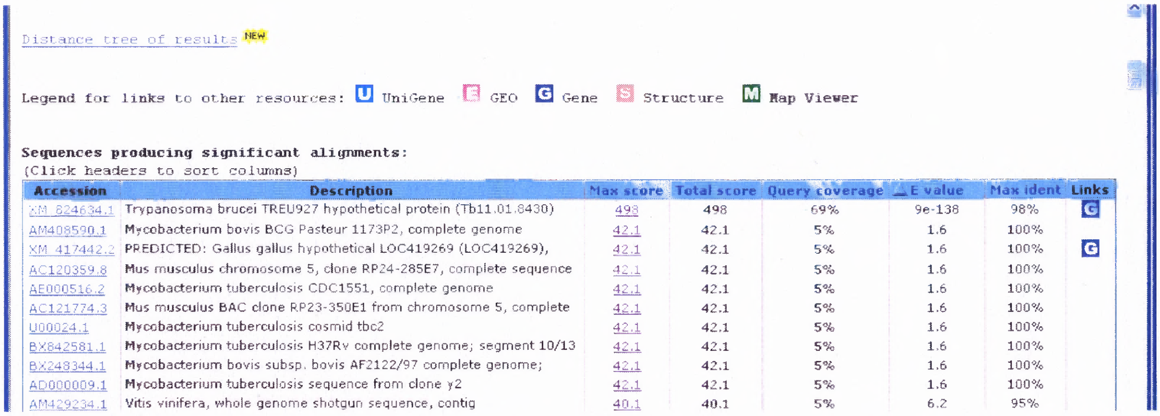
Show ☑ Graphical Overview ☑ Linkout ☑ Sequence Retrieval ☐ NCBI-gi Alignment in HTML Format

☐ CDS feature

☑ New View

Figure 4.3 Query assigned request ID.

From this page, "Format!" is clicked to redirect the user to the results page where the report is obtained. A traditional BLAST report will consist of three sections: 1. the header, containing all the information about the query and the database searched; 2. online-database descriptions of each database sequence; 3. the pair-wise sequence alignments. BLAST uses a statistical method to determine a bit Max Score and an expect value (E-value) for each alignment. As a general rule, the higher the bit score the better the alignment and the lower the E-value (the closer to zero it is) the more significant the hit. However, one must be aware of these values. They do not tell the biological significance of the alignment, so reading of the descriptions is also necessary. The significant alignment of the highlighted sequence queried is shown in Figure 4.4.



Distance tree of results

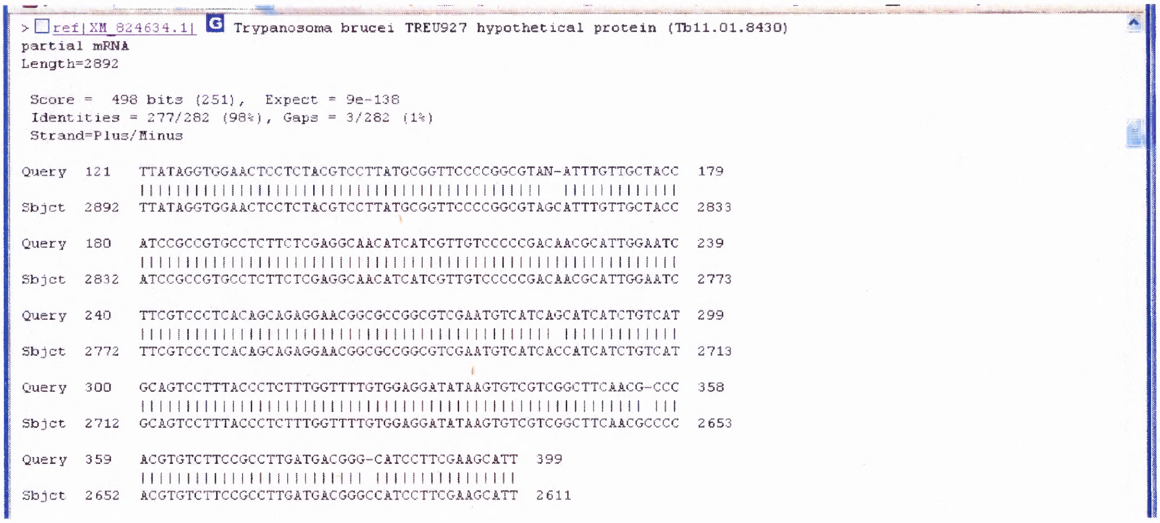
Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
XM_824634.1	Trypanosoma brucei TREU927 hypothetical protein (Tb11.01.8430)	498	498	69%	9e-138	98%	G
AM408590.1	Mycobacterium bovis BCG Pasteur 1173P2, complete genome	42.1	42.1	5%	1.6	100%	
XM_417442.2	PREDICTED: Gallus gallus hypothetical LOC419269 (LOC419269),	42.1	42.1	5%	1.6	100%	G
AC120359.8	Mus musculus chromosome 5, clone RP24-285E7, complete sequence	42.1	42.1	5%	1.6	100%	
AE000516.2	Mycobacterium tuberculosis CDC1551, complete genome	42.1	42.1	5%	1.6	100%	
AC121774.3	Mus musculus BAC clone RP23-350E1 from chromosome 5, complete	42.1	42.1	5%	1.6	100%	
U00024.1	Mycobacterium tuberculosis cosmid tbc2	42.1	42.1	5%	1.6	100%	
BX842581.1	Mycobacterium tuberculosis H37Rv complete genome; segment 10/13	42.1	42.1	5%	1.6	100%	
BX248344.1	Mycobacterium bovis subsp. bovis AF2122/97 complete genome;	42.1	42.1	5%	1.6	100%	
AD000009.1	Mycobacterium tuberculosis sequence from clone y2	42.1	42.1	5%	1.6	100%	
AM429234.1	Vitis vinifera, whole genome shotgun sequence, contig	40.1	40.1	5%	6.2	95%	

Figure 4.4 Partial screenshot of possibly significant alignments – displays bit scores and E-values of results.

The first alignment listed has an E-value of 9e-138 which is an extremely low value. This tells the user that the hit is significant. The next item to take note of is the Max Score of 498. This score is extremely higher than the other scores listed, which shows that this alignment is a good one. After reading the description of the alignment, it is evident that the species is also correct because it lists it as Trypanosoma brucei.



```
> □ref|XM_824634.1| G Trypanosoma brucei TREU927 hypothetical protein (Tb11.01.8430)
partial mRNA
Length=2892

 Score =  498 bits (251),  Expect = 9e-138
 Identities = 277/282 (98%), Gaps = 3/282 (1%)
 Strand=Plus/Minus

Query  121   TTATAGGTGGAACTCCTCTACGTCCTTATGCGGTTCCCCGGCGTAN-ATTTGTTGCTACC  179
             |||||||||||||||||||||||||||||||||||||||||||||  |||||||||||||
Sbjct  2892  TTATAGGTGGAACTCCTCTACGTCCTTATGCGGTTCCCCGGCGTAGCATTTGTTGCTACC  2833

Query  180   ATCCGCCGTGCCTCTTCTCGAGGCAACATCATCGTTGTCCCCCGACAACGCATTGGAATC  239
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2832  ATCCGCCGTGCCTCTTCTCGAGGCAACATCATCGTTGTCCCCCGACAACGCATTGGAATC  2773

Query  240   TTCGTCCCTCACAGCAGAGGAACGGCGCCGGCGTCGAATGTCATCAGCATCATCTGTCAT  299
             |||||||||||||||||||||||||||||||||||||||||||||| |||||||||||||
Sbjct  2772  TTCGTCCCTCACAGCAGAGGAACGGCGCCGGCGTCGAATGTCATCACCATCATCTGTCAT  2713

Query  300   GCAGTCCTTTACCCTCTTTGGTTTTGTGGAGGATATAAGTGTCGTCGGCTTCAACG-CCC  358
             |||||||||||||||||||||||||||||||||||||||||||||||||||||||| |||
Sbjct  2712  GCAGTCCTTTACCCTCTTTGGTTTTGTGGAGGATATAAGTGTCGTCGGCTTCAACGCCCC  2653

Query  359   ACGTGTCTTCCGCCTTGATGACGGG-CATCCTTCGAAGCATT  399
             ||||||||||||||||||||||||| ||||||||||||||||
Sbjct  2652  ACGTGTCTTCCGCCTTGATGACGGGCCATCCTTCGAAGCATT  2611
```

Figure 4.5 Partial screenshot of significant alignment – displays alignment information for T. brucei TREU927 hypothetical protein (Tb11.01.8430), partial mRNA.

The next step is to access the specific alignment by clicking on the max score of the alignment of interest, in this case that of Trypanosoma brucei TREU927, Score 498. The first item listed is the BLAST identifier. The BLAST identifier should be saved because it is always linked to the sequence noted so that new queries do not need to be made every time you would like to refer back to this sequence. Listed after the identifier is the species and what this sequence codes for in that species, in this case T. brucei and TREU927 hypothetical protein respectively. Directly after that, it tells the user that this particular sequence is a partial mRNA sequence.

By clicking on the BLAST identifier, it brings the user to the reference page of the sequence. On this page, the nucleotide sequence length may be obtained under the section labeled CDS (coding sequence). All the following information: downloaded sequence, gene identifier, coding sequence, and nucleotide numbers are all kept in an Excel file. This proves extremely useful, when trying to compare this tool to the next tool, ORBIT.

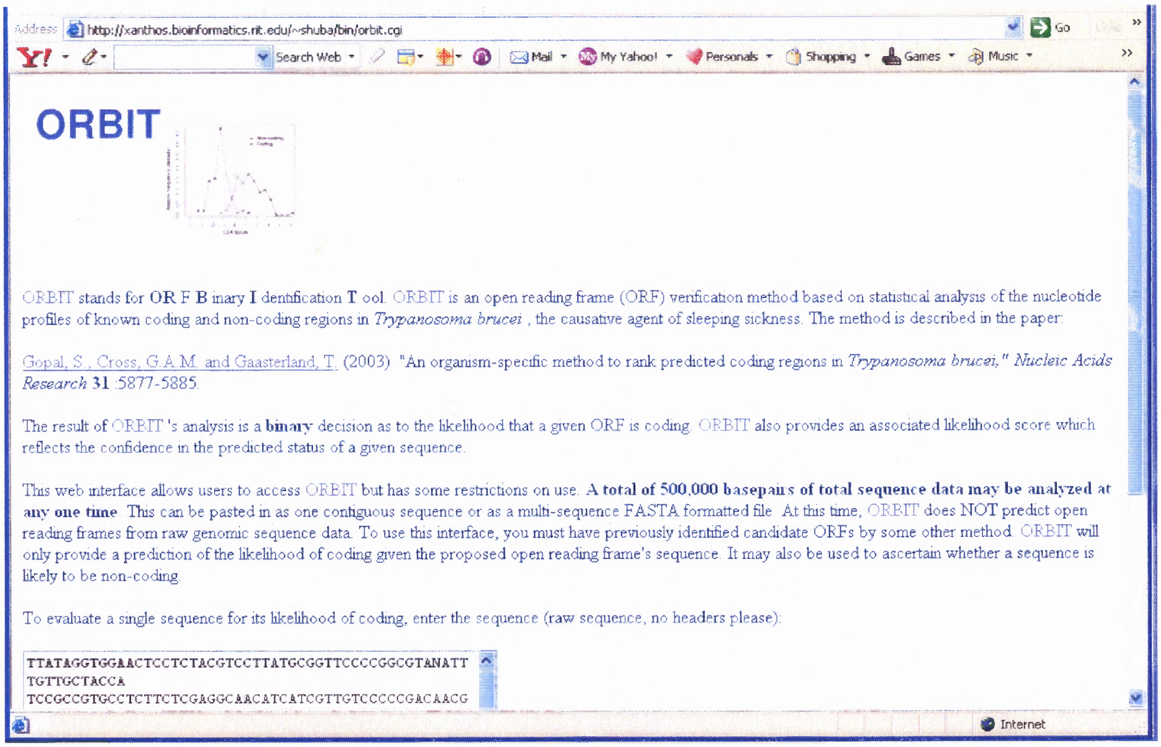
CHAPTER 5

ORBIT

Initial sequence alignment was done using the well known alignment tool BLAST. However, another tool is needed to verify and/or compare the results obtained from BLAST. This other tool is called ORBIT (Open Reading Frame Binary Identification Tool). It uses a statistical analysis of nucleotide profiles of known coding and non-coding regions of T. brucei to create an open reading frame verification method (Gopal et al. 2003). The sequences entered must be non-genomic sequences and only be open reading frames.

The output of ORBIT is not a sequence alignment like in BLAST but a prediction of the likelihood of coding and whether or not the sequence queried is likely to be coding or non-coding. The statistical process of how ORBIT functions will not be mentioned in this thesis. For the purpose of this thesis, only the physical task of entering a sequence, getting results, and understanding what the results mean, are needed. For a detailed look into the statistical basis of ORBIT, please refer to the paper listed on ORBIT's homepage, http://xanthos.bioinformatics.rit.edu/~shuba/bin/orbit.cgi.

To begin the verification process, an open reading frame sequence must be entered in ORBIT's query box. To obtain this sequence, make sure to copy into the query box the aligned sequence found from the BLAST results (do not enter the whole sequence that was initially downloaded). Once the sequence has been entered into the box, the button on the bottom of the screen labeled "Evaluate" is clicked. The resulting page is seen in Figure 5.2.



ORBIT

ORBIT stands for OR F B inary I dentification T ool. ORBIT is an open reading frame (ORF) verification method based on statistical analysis of the nucleotide profiles of known coding and non-coding regions in *Trypanosoma brucei* , the causative agent of sleeping sickness. The method is described in the paper:

Gopal, S., Cross, G.A.M. and Gaasterland, T. (2003). "An organism-specific method to rank predicted coding regions in *Trypanosoma brucei*," *Nucleic Acids Research* **31**:5877-5885.

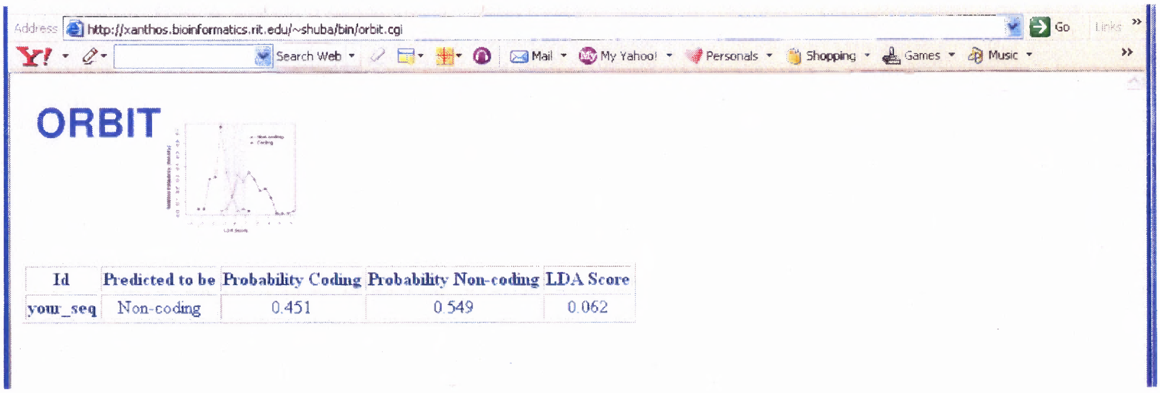
The result of ORBIT 's analysis is a **binary** decision as to the likelihood that a given ORF is coding. ORBIT also provides an associated likelihood score which reflects the confidence in the predicted status of a given sequence.

This web interface allows users to access ORBIT but has some restrictions on use. **A total of 500,000 basepairs of total sequence data may be analyzed at any one time**. This can be pasted in as one contiguous sequence or as a multi-sequence FASTA formatted file. At this time, ORBIT does NOT predict open reading frames from raw genomic sequence data. To use this interface, you must have previously identified candidate ORFs by some other method. ORBIT will only provide a prediction of the likelihood of coding given the proposed open reading frame's sequence. It may also be used to ascertain whether a sequence is likely to be non-coding.

To evaluate a single sequence for its likelihood of coding, enter the sequence (raw sequence, no headers please):

```
TTATAGGTGGAACTCCTCTACGTCCTTATGCGGTTCCCCGGCGTANATT
TGTTGCTACCA
TCCGCCGTGCCTCTTCTCGAGGCAACATCATCGTTGTCCCCCGACAACG
```

Figure 5.1 Homepage of ORBIT – screenshot of the homepage with sequence entered in query box.
http://xanthos.bioinformatics.rit.edu/~shuba/bin/orbit.cgi

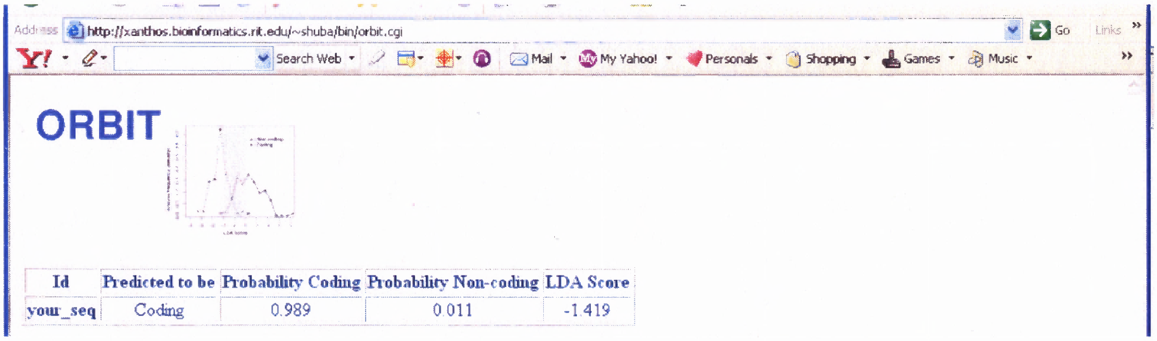


ORBIT

Id	Predicted to be	Probability Coding	Probability Non-coding	LDA Score
your_seq	Non-coding	0.451	0.549	0.062

Figure 5.2 One of two possible results of ORBIT – partial screenshot of the results of this query.

The information on this page that is of most interest to us is the "Predicted to be" and the "LDA Score" columns. According to ORBIT, the sequence that was found by BLAST is a non-coding region and the LDA Score is a reflection of that result. A LDA score of zero or less classifies a sequence as coding while a LDA score greater than zero classifies a sequence as non-coding. The LDA score in this example is 0.062. Hence, it is classified as a non-coding sequence. Figure 5.3 shows the results of a query that was found to be positive for a coding region and which also was found to be a partial mRNA sequence from BLAST results. This second example, a sequence that agrees between both alignment tools as being a coding region, is what is chosen to continue on with UTR generation.



Address: http://xanthos.bioinformatics.rit.edu/~shuba/bin/orbit.cgi

Id	Predicted to be	Probability Coding	Probability Non-coding	LDA Score
your_seq	Coding	0.989	0.011	-1.419

Figure 5.3 Other possible result of ORBIT – partial screenshot of the results of a positive query.

CHAPTER 6

UTR GENERATION

6.1 UTRs

A UTR (untranslated region) is part of a gene that does not translate into a protein. There are two types of UTRs: 5′ UTR and 3′ UTR. The 5′ UTR starts from the 5′ end of the mRNA up to the first codon used in translation. The 3′ UTR starts from the last codon used in translation up to the 3′ end of the mRNA. UTRs are believed to have several roles in gene expression, including mRNA stability, mRNA localization, and translational efficiency, (Pesole et al. 2000) and regulation by miRNAs.

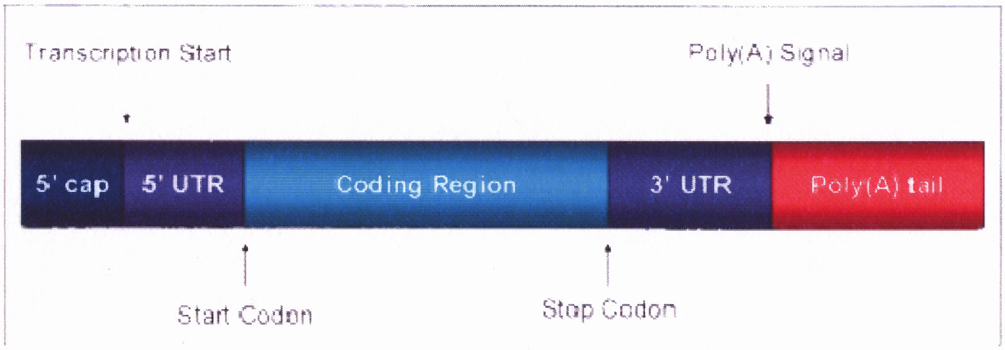
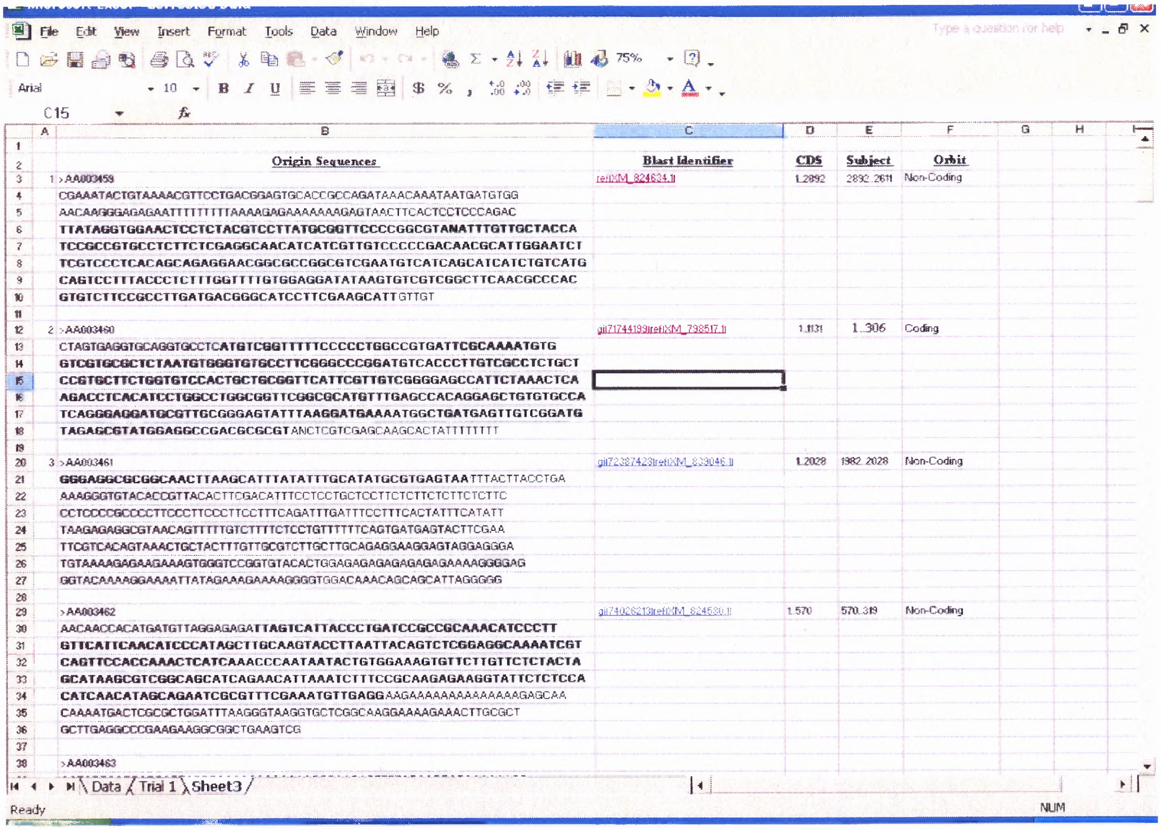


Figure 6.1 Structure of mature mRNA - mRNA includes a 5' cap, 5' UTR, coding region, 3' UTR, and poly (A) tail.
http://www.answers.com/topic/messenger-rna

6.2 UTR Generation and Prediction

As sequences were downloaded, they were saved into an Excel file (see Figure 6.2), under a column titled "Origin Sequences". Those sequences were run one at a time on NCBI's BLAST website, http://www.ncbi.nlm.nih.gov/BLAST/. The BLAST identifier was placed in a separate column to be easily accessible if further review is needed and thereby avoiding the repeating of BLAST. The section of sequence that was found to be similar with the origin sequence is represented by the nucleotides that are in bold font. From the report page accessed through the BLAST identifier the CDS (coding sequence) region was obtained, which is the information listed in the third column. The fourth column contains the nucleotide base positions of the points the aligned sequences pair up. The final column lists the result of the sequence that was run on ORBIT, whether it was predicted to be a coding or non-coding sequence. The sequence entered here was not the full origin sequence that had been downloaded, but the bolded sequence found to be similar by BLAST.

Every downloaded sequence had to be run in BLAST and ORBIT. If both programs agreed on their findings, whether it was determined to be coding or non-coding, they were chosen to be the predicted UTRs to be run in UTRscan. Before that can be done, the next step has to be determining whether they are true or known UTRs.



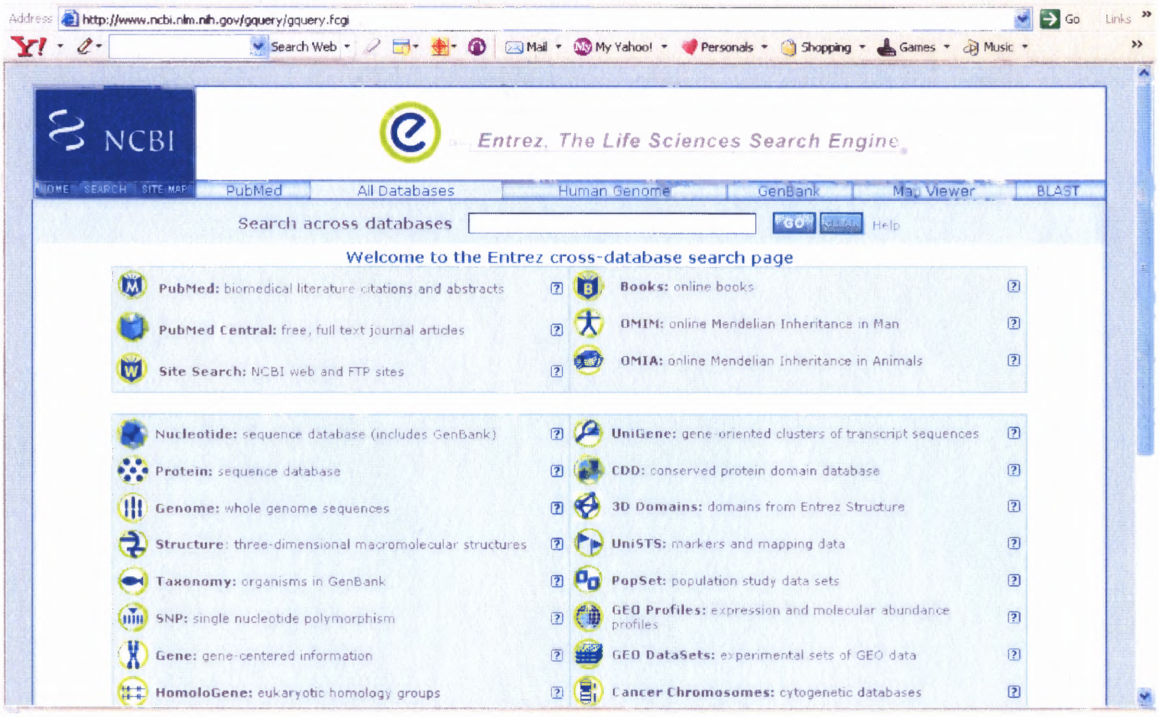
	A	B	C	D	E	F	G	H
1								
2		**Origin Sequences**	**Blast Identifier**	**CDS**	**Subject**	**Orbit**		
3	1	>AA003459	ref|XM_824634.1|	1..2892	2892..2611	Non-Coding		
4		CGAAATACTGTAAAACGTTCCTGACGGAGTGCACCGCCAGATAAACAAATAATGATGTGG						
5		AACAAGGGAGAGAATTTTTTTTAAAAGAGAAAAAAAGAGTAACTTCACTCCTCCCAGAC						
6		**TTATAGGTGGAACTCCTCTACGTCCTTATGCGGTTCCCCGGCGTANATTTGTTGCTACCA**						
7		**TCCGCCGTGCCTCTTCTCGAGGCAACATCATCGTTGTCCCCCGACAACGCATTGGAATCT**						
8		**TCGTCCCTCACAGCAGAGGAACGGCGCCGGCGTCGAATGTCATCAGCATCATCTGTCATG**						
9		**CAGTCCTTTACCCTCTTTGGTTTTGTGGAGGATATAAGTGTCGTCGGCTTCAACGCCCAC**						
10		**GTGTCTTCCGCCTTGATGACGGGCATCCTTCGAAGCATT**GTTGT						
11								
12	2	>AA003460	gi|71744199|ref|XM_798517.1|	1..1131	1..306	Coding		
13		CTAGTGAGGTGCAGGTGCCTC**ATGTCGGTTTTTCCCCCTGGCCGTGATTCGCAAAATGTG**						
14		**GTCGTGCGCTCTAATGTGGGTGTGCCTTCGGGCCCGGATGTCACCCTTGTCGCCTCTGCT**						
15		**CCGTGCTTCTGGTGTCCACTGCTGCGGTTCATTCGTTGTCGGGGAGCCATTCTAAACTCA**						
16		**AGACCTCACATCCTGGCCTGGCGGTTCGGCGCATGTTTGAGCCACAGGAGCTGTGTGCCA**						
17		**TCAGGGAGGATGCGTTGCGGGAGTATTTAAGGATGAAAATGGCTGATGAGTTGTCGGATG**						
18		**TAGAGCGTATGGAGGCCGACGCGCGT**ANCTCGTCGAGCAAGCACTATTTTTTTT						
19								
20	3	>AA003461	gi|72387423|ref|XM_833046.1|	1..2028	1982..2028	Non-Coding		
21		**GGGAGGCGCGGCAACTTAAGCATTTATATTTGCATATGCGTGAGTAA**TTTACTTACCTGA						
22		AAAGGGTGTACACCGTTACACTTCGACATTTCCTCCTGCTCCTTCTCTTCTCTTCTCTTC						
23		CCTCCCCGCCCCTTCCCTTCCCTTCCTTTCAGATTTGATTTCCTTTCACTATTTCATATT						
24		TAAGAGAGGCGTAACAGTTTTTGTCTTTTCTCCTGTTTTTTCAGTGATGAGTACTTCGAA						
25		TTCGTCACAGTAAACTGCTACTTTGTTGCGTCTTGCTTGCAGAGGAAGGAGTAGGAGGGA						
26		TGTAAAAGAGAAGAAAGTGGGTCCGGTGTACACTGGAGAGAGAGAGAGAGAAAAGGGGAG						
27		GGTACAAAAGGAAAATTATAGAAAGAAAAGGGGTGGACAAACAGCAGCATTAGGGGG						
28								
29		>AA003462	gi|74026213|ref|XM_824580.1|	1..570	570..319	Non-Coding		
30		AACAACCACATGATGTTAGGAGAGA**TTAGTCATTACCCTGATCCGCCGCAAACATCCCTT**						
31		**GTTCATTCAACATCCCATAGCTTGCAAGTACCTTAATTACAGTCTCGGAGGCAAAATCGT**						
32		**CAGTTCCACCAAACTCATCAAACCCAATAATACTGTGGAAAGTGTTCTTGTTCTCTACTA**						
33		**GCATAAGCGTCGGCAGCATCAGAACATTAAATCTTTCCGCAAGAGAAGGTATTCTCTCCA**						
34		**CATCAACATAGCAGAATCGCGTTTCGAAATGTTGAGG**AAGAAAAAAAAAAAAAAGAGCAA						
35		CAAAATGACTCGCGCTGGATTTAAGGGTAAGGTGCTCGGCAAGGAAAAGAAACTTGCGCT						
36		GCTTGAGGCCCGAAGAAGGCGGCTGAAGTCG						
37								
38		>AA003463						

Figure 6.2 Screenshot of Excel data file.

CHAPTER 7

TRUE UTRs

The process of obtaining true UTRs is quick and easy however, going through the full list of sequences and deciphering which are legitimate and which are not is a long and laborious process. To get started we have to access NCBI's homepage again, http://www.ncbi.nlm.nih.gov/. From the "Hot Spots" column to the right side of the page click on "Entrez Home". Once at the homepage of Entrez there will two columns of databases to choose from. In the case of our project we needed to click on the nucleotide database.



Welcome to the Entrez cross-database search page

PubMed: biomedical literature citations and abstracts

PubMed Central: free, full text journal articles

Site Search: NCBI web and FTP sites

Books: online books

OMIM: online Mendelian Inheritance in Man

OMIA: online Mendelian Inheritance in Animals

Nucleotide: sequence database (includes GenBank)

Protein: sequence database

Genome: whole genome sequences

Structure: three-dimensional macromolecular structures

Taxonomy: organisms in GenBank

SNP: single nucleotide polymorphism

Gene: gene-centered information

HomoloGene: eukaryotic homology groups

UniGene: gene-oriented clusters of transcript sequences

CDD: conserved protein domain database

3D Domains: domains from Entrez Structure

UniSTS: markers and mapping data

PopSet: population study data sets

GEO Profiles: expression and molecular abundance profiles

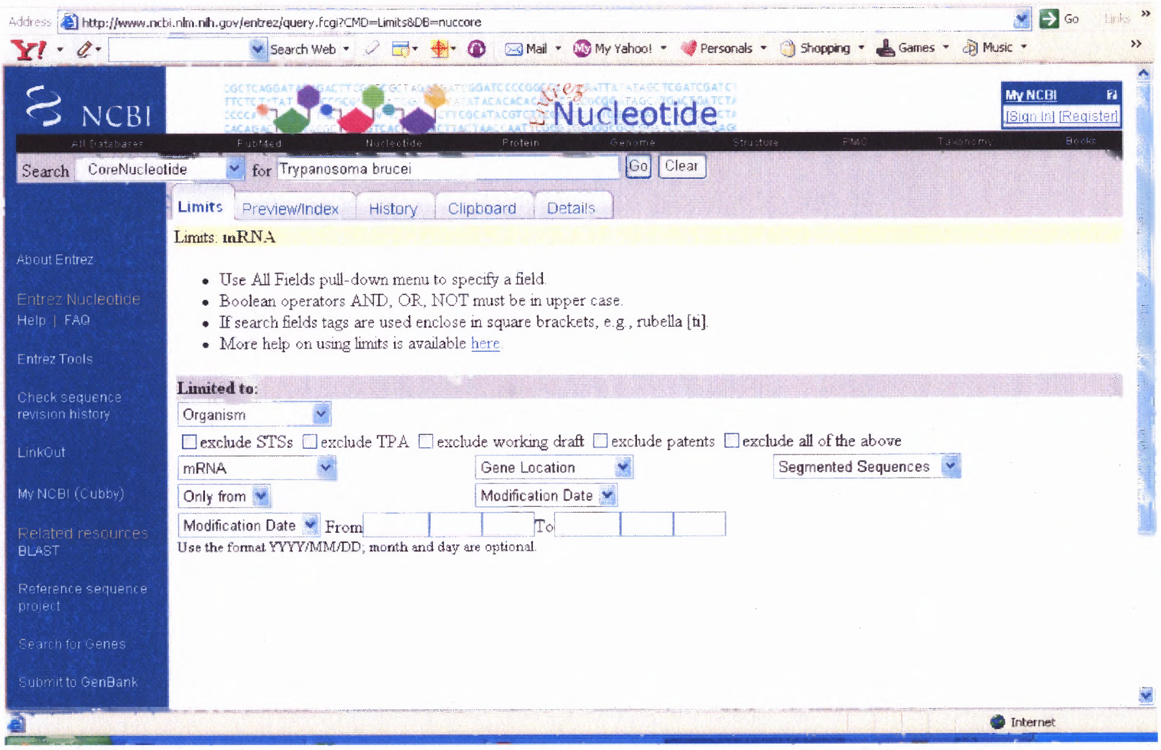
GEO DataSets: experimental sets of GEO data

Cancer Chromosomes: cytogenetic databases

Figure 7.1 Homepage of Entrez.
http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi

The next step is to run a search, which can get a little tricky because of the many available drop down menus and parameters that can be set to aid in your search.

First we begin by selecting CoreNucleotide from the Search menu and then entering Trypansoma brucei in the blank field “for”. To aid in minimizing results, click the “Limits” tab to help set up some parameters. Directly under “Limited to” is a drop down menu where the user should choose “Organism” instead of “All Fields”. In the next drop down menu after that, select “mRNA” instead of “Molecule”. Once all the above fields have been selected, click the “Go” button to run the search.



http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=Limits&DB=nuccore

Search CoreNucleotide for Trypanosoma brucei

Limits: mRNA

- Use All Fields pull-down menu to specify a field.
- Boolean operators AND, OR, NOT must be in upper case.
- If search fields tags are used enclose in square brackets, e.g., rubella [ti].
- More help on using limits is available here.

Limited to:

Organism

□ exclude STSs □ exclude TPA □ exclude working draft □ exclude patents □ exclude all of the above

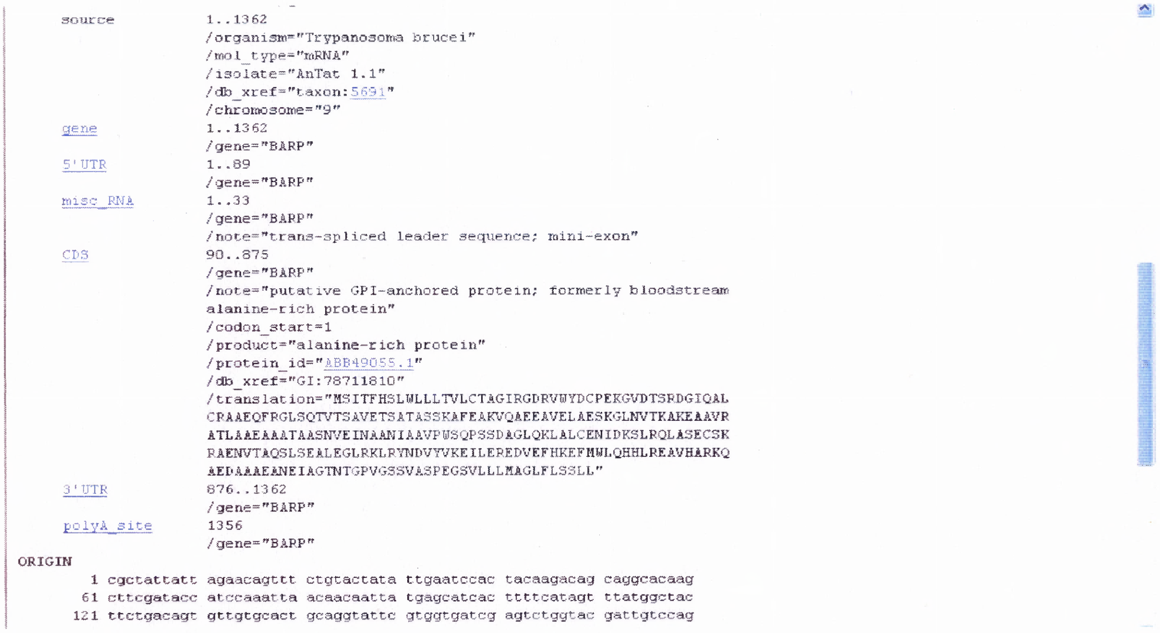
mRNA | Gene Location | Segmented Sequences

Only from | Modification Date

Modification Date From ____ ____ ____ To ____ ____ ____

Use the format YYYY/MM/DD; month and day are optional.

Figure 7.2 Entrez search query.
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=Limits&DB=nuccore



```
     source          1..1362
                     /organism="Trypanosoma brucei"
                     /mol_type="mRNA"
                     /isolate="AnTat 1.1"
                     /db_xref="taxon:5691"
                     /chromosome="9"
     gene            1..1362
                     /gene="BARP"
     5'UTR           1..89
                     /gene="BARP"
     misc_RNA        1..33
                     /gene="BARP"
                     /note="trans-spliced leader sequence; mini-exon"
     CDS             90..875
                     /gene="BARP"
                     /note="putative GPI-anchored protein; formerly bloodstream
                     alanine-rich protein"
                     /codon_start=1
                     /product="alanine-rich protein"
                     /protein_id="ABB49055.1"
                     /db_xref="GI:78711810"
                     /translation="MSITFHSLWLLLTVLCTAGIRGDRVWYDCPEKGVDTSRDGIQAL
                     CRAAEQFRGLSQTVTSAVETSATASSKAFEAKVQAEEAVELAESKGLNVTKAKEAAVR
                     ATLAAEAAATAASNVEINAANIAAVPWSQPSSDAGLQKLALCENIDKSLRQLASECSK
                     RAENVTAQSLSEALEGLRKLRYNDVYVKEILEREDVEFHKEFMWLQHHLREAVHARKQ
                     AEDAAAEANEIAGTNTGPVGSSVASPEGSVLLLMAGLFLSSLL"
     3'UTR           876..1362
                     /gene="BARP"
     polyA_site      1356
                     /gene="BARP"
ORIGIN
        1 cgctattatt agaacagttt ctgtactata ttgaatccac tacaagacag caggcacaag
       61 cttcgatacc atccaaatta acaacaatta tgagcatcac ttttcatagt ttatggctac
      121 ttctgacagt gttgtgcact gcaggtattc gtggtgatcg agtctggtac gattgtccag
```

Figure 7.3 Partial Screenshot of Entrez report page – Results for Accession # DQ246439.
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=Limits&DB=nuccore

The result of this search yields an astonishing 9284 mRNA sequences for T. brucei. The second entry in the list, Accession # DQ246439, contains both a 5′ UTR and a 3′ UTR, in addition to the rest of the information provided on the record page. Figure 7.3 illustrates this data. It is extremely time consuming and tedious to go through 9284 sequences one by one, in order to find UTRs. To help alleviate this problem it is possible to add another search to the one placed above to filter out sequences that do not have UTRs. For this to be done, the "Search" field is left the same and only change the "For" field to exactly what the inquiry is about, i.e. "3′UTR" or "5′UTR". Back at the "Limits" tab, change the "Organism" tab back to "All Fields" and make sure to keep "mRNA" selected. Finally, go over to the final tab and click "History". At this location it is possible to find all previous searches done by the user. By clicking on the Search # of the previous search run, a pop up menu becomes available. From this pop up menu, choose

"And" and then click "Go". A search of UTR sequences for Trypansoma brucei yielded the following results. A query of sequences containing only 5′ UTRs yielded 28 sequences. Oddly enough, in this case, a search of 3′ UTRs also yielded 28 sequences. Combining all three conditions, that is looking at mRNA sequences with 5′ and 3′ UTRs, resulted in only 25 sequences from the initial 9284 found.

Table 7.1 Sequences with 3′ and 5′ UTRs

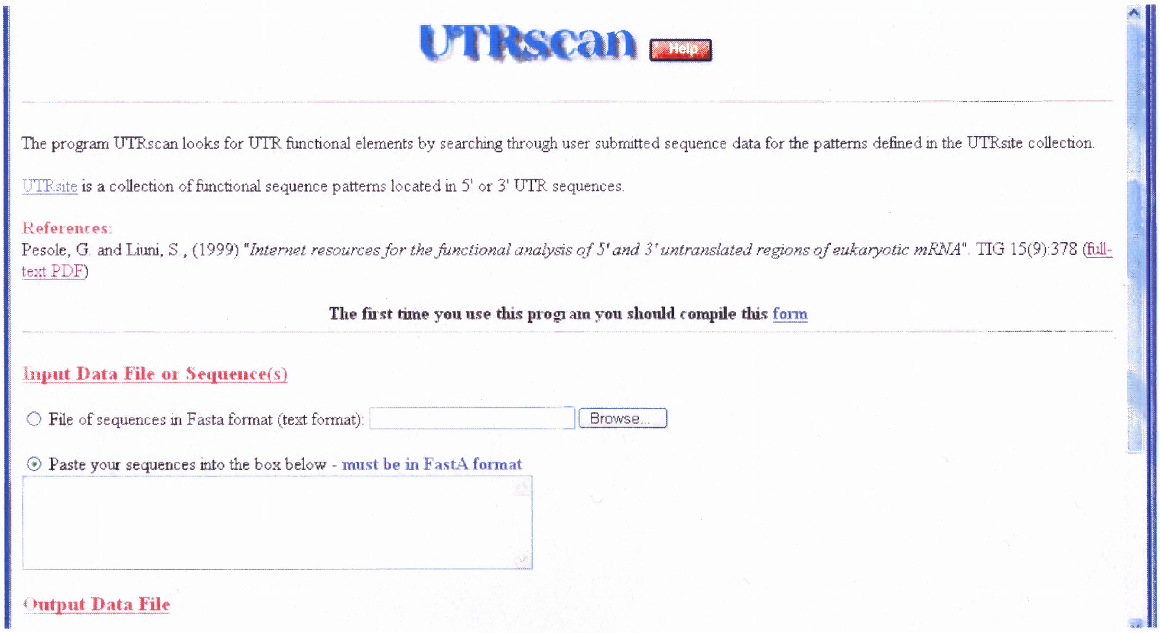
	Accession #	Description
1	DQ246439	Trypanosoma brucei isolate AnTat 1.1 alanine-rich protein (BARP) mRNA, complete cds, gi|78711809|gb|DQ246439.1|[78711809]
2	DQ841707	Trypanosoma brucei putative sialidase mRNA, complete cds gi|112031258|gb|DQ841707.1|[112031258]
3	AM295303	Trypanosoma brucei mRNA for band-VII protein (MP18 gene) gi|111378143|emb|AM295303.1|[111378143]
4	AM168497	Trypanosoma brucei mRNA for DNA-directed RNA polymerase subunit RPB12 (RPB12 gene), gi|90954544|emb|AM168497.1|[90954544]
5	AM168496	Trypanosoma brucei mRNA for DNA-directed RNA polymerase subunit 2RPB10 (2RPB10 gene), gi|90954542|emb|AM168496.1|[90954542]
6	AM167553	Trypanosoma brucei mRNA for RNA polymerase subunit 2RPB6 (2RPB6 gene), strain 427, gi|90954540|emb|AM167553.1|[90954540]
7	AM167552	Trypanosoma brucei mRNA for RNA polymerase subunit 2RPB5 (2RPB5 gene), strain 427, gi|90954538|emb|AM167552.1|[90954538]
8	AM159572	Trypanosoma brucei mRNA for RNA polymerase I subunit RPC19 (RPC19 gene) gi|88687069|emb|AM159572.1|[88687069]
9	AM159571	Trypanosoma brucei mRNA for RNA polymerase I subunit RPB8 (RPB8 gene) gi|88687067|emb|AM159571.1|[88687067]
10	AM159570	Trypanosoma brucei mRNA for RNA polymerase I subunit RPC40 (RPC40 gene) gi|88687065|emb|AM159570.1|[88687065]
11	AM159086	Trypanosoma brucei mRNA for RNA polymerase I subunit 1RPB10 (1RPB10 gene), gi|86438841|emb|AM159086.1|[86438841]
12	AM159085	Trypanosoma brucei mRNA for RNA polymerase I subunit 1RPB6 (1RPB6 gene) gi|84619253|emb|AM159085.1|[84619253]
13	AM159084	Trypanosoma brucei mRNA for RNA polymerase I subunit RPA12 (RPA12 gene) gi|84619251|emb|AM159084.1|[84619251]
14	AM159083	Trypanosoma brucei mRNA for RNA polymerase I subunit 1RPB5 (1RPB5 gene) gi|84619249|emb|AM159083.1|[84619249]
15	AM235385	Trypanosoma brucei mRNA for transcription factor TFIIB-like (TFIIB-like gene) gi|89511759|emb|AM235385.1|[89511759]
16	AY157307	Trypanosoma brucei phosphatidylinositol-phospholipase C mRNA, complete cds gi|37724009|gb|AY157307.1|[37724009]
17	AJ887988	Trypanosoma brucei mRNA for transcription factor IIA gamma subunit (TFIIA-2 gene), gi|62120370|emb|AJ887988.1|[62120370]
18	AJ887987	Trypanosoma brucei mRNA for transcription factor IIA alpha-beta subunit (TFIIA-1 gene), gi|62120368|emb|AJ887987.1|[62120368]
19	AJ879576	Trypanosoma brucei mRNA for small nuclear RNA activating protein 3 (SNAP3 gene), gi|62120366|emb|AJ879576.1|[62120366]
20	AJ879575	Trypanosoma brucei mRNA for small nuclear RNA activating protein 2 (SNAP2 gene), gi|62120364|emb|AJ879575.1|[62120364]
21	AJ437580	Trypanosoma brucei mRNA for RNA polymerase I second largest subunit (RPA2 gene), gi|19913064|emb|AJ437580.1|[19913064]
22	L30155	Trypanosoma brucei paraflagellar rod protein (PFR2) mRNA, complete cds gi|463286|gb|L30155.1|TRBP515A[463286]
23	AF074867	Trypanosoma brucei rhodesiense lysosomal/endosomal membrane protein p67 mRNA, complete cds, gi|3378149|gb|AF074867.1|AF074867[3378149]
24	M81386	Trypanosoma brucei hexose transporter mRNA, complete cds gi|162126|gb|M81386.1|TRBHTP[162126]
25	L02933	Trypanosoma brucei surface antigen mRNA, complete cds gi|162202|gb|L02933.1|TRBPCSSA[162202]

CHAPTER 8

UTRscan

UTRscan is a well known program that allows its users to search through user submitted sequences for patterns collected in the UTRsite. The UTRsite is the location where the collection of functional sequence patterns of 5′ and 3′ UTR sequences are located. It is extremely useful in the prediction of functional elements of both existing and newly produced nucleotide sequences (Pesole & Liuni, 1999). It can find motifs that fit to both 5′ and 3′ UTR sequences.

When entering sequences into the input sequence box, there are a few issues users need to be aware of. The most important one being that the sequences must be in fasta form and the other being that it cannot process batch sequences, only one sequence can be queried at a time.



The program UTRscan looks for UTR functional elements by searching through user submitted sequence data for the patterns defined in the UTRsite collection.

UTRsite is a collection of functional sequence patterns located in 5' or 3' UTR sequences.

References:
Pesole, G. and Liuni, S., (1999) "*Internet resources for the functional analysis of 5' and 3' untranslated regions of eukaryotic mRNA*". TIG 15(9):378 (full-text PDF)

The first time you use this progi am you should compile this form

Input Data File or Sequence(s)

○ File of sequences in Fasta format (text format): Browse...

● Paste your sequences into the box below - **must be in FastA format**

Output Data File

Figure 8.1 UTRscan homepage.
http://www.ba.itb.cnr.it/BIG/UTRScan/

The program also allows the user to choose the way they wish to receive their data, in text or html format. The data displayed in both formats is the same except for that the html format links the patterns to the respective UTRsite collection. An example of data output from the UTRscan website is shown in the Appendix.

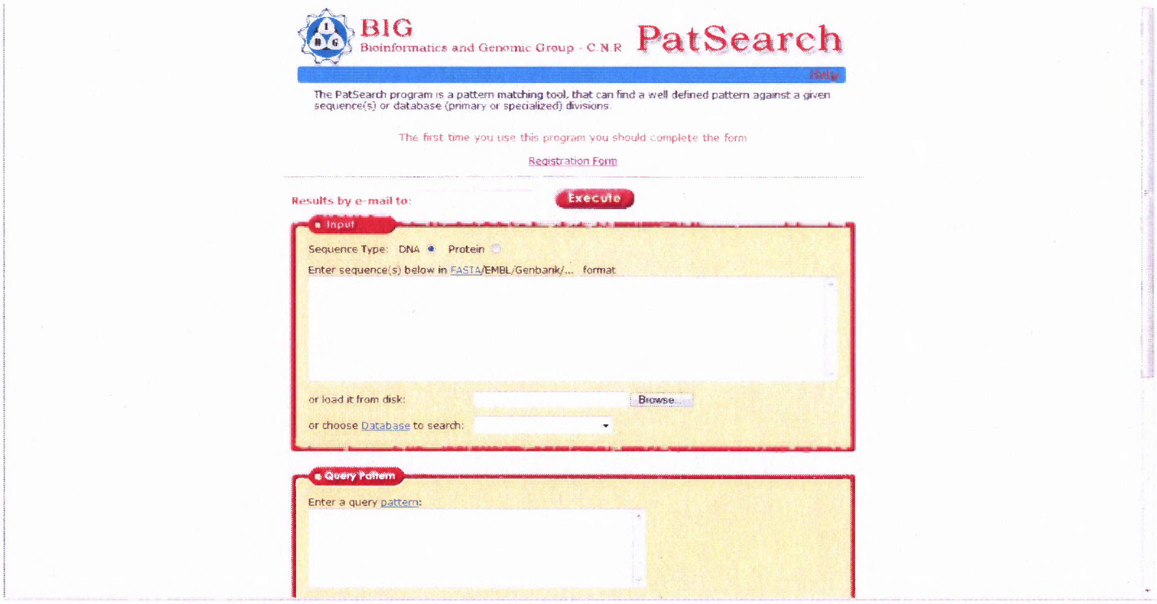
CHAPTER 9

PATSEARCH

PatSearch is software that scans sequences submitted by users looking for any sequence patterns and structural motifs while allowing for mismatches and mispairings, below a fixed threshold determined by the user. The program is available online at http://bighost.area.ba.cnr.it/BIG/PatSearch. It is available only to non-for-profit organizations and users must submit a request form for access to the program. If a user is not granted access he cannot use the program because results are emailed to the user's approved email address.

There are three typical applications of PatSearch. However, the first application is the one that is of interest in this project. It is the search for specific cis-acting elements located in 5′ UTR or 3′ UTR of eukaryotic mRNAs. The results of the subsequent search are then collected in the UTRsite database and annotated through PatSearch analysis of UTRdb. The second application involves the search for the recognition site of transcription factors of p53 homologs. The third is the search for PWMs (position weight matrices).

The program works in a slightly confusing manner. It runs by assigning "patterns" which are a combination of pattern units that can be assigned a name or left unnamed. The names that can be assigned to pattern units are p1, p2, p3, etc. always using a lowercase p. Keep in mind that patterns must be named if 1 or more reprocessing steps of already matched sequences have to be carried out. Example: p1=5…9. In this example p1 is defined as a pattern unit that means match any 5 to 9 character sequence and call them p1.



The PatSearch program is a pattern matching tool, that can find a well defined pattern against a given sequence(s) or database (primary or specialized) divisions.

The first time you use this program you should complete the form

Registration Form

Results by e-mail to:

Sequence Type: DNA ● Protein ○

Enter sequence(s) below in FASTA/EMBL/Genbank/... format

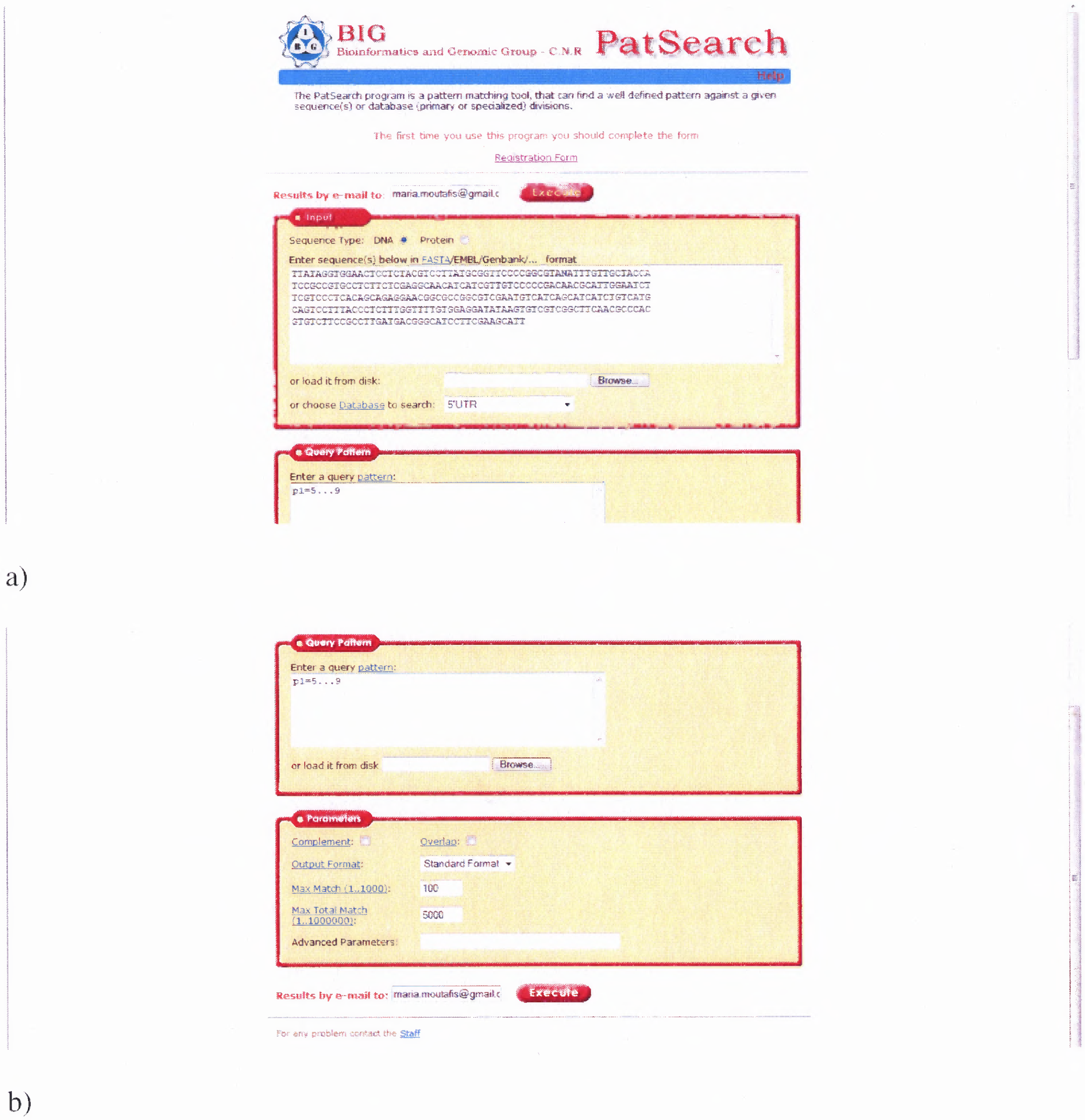
or load it from disk:

or choose Database to search:

Enter a query pattern:

Figure 9.1 PatSearch homepage.
http://www.ba.itb.cnr.it/BIG/PatSearch/

The sequence the user wishes to query is placed in the input box via copy and paste or by loading the sequence from a disk. In the same box titles Input the user then has the ability to select from the drop-down menu which database to search. The second box is titled Query Pattern. This may also be typed in the space provided or downloaded from a disk. The final box contains parameters the user wishes to set including the format of the output and maximum number of matches. When all the data is inputted, the user enters his email in the designated box and clicks the execute button. The page refreshes and gives the user a job number for their inquiry and notifies them that their request has been accepted and to check their email for the result of the query. Figure 9.1 parts a) and b), show screenshots of an example input.



BIG

Bioinformatics and Genomic Group - C.N.R

PatSearch

Help

The PatSearch program is a pattern matching tool, that can find a well defined pattern against a given sequence(s) or database (primary or specialized) divisions.

The first time you use this program you should complete the form

Registration Form

Results by e-mail to: maria.moutafis@gmail.c Execute

Input

Sequence Type: DNA ● Protein ○

Enter sequence(s) below in FASTA/EMBL/Genbank/... format

```
TTATAGGTGGAACTCCTCTACGTCCTTATGCGGTTCCCCGGCGTANATTTGTTGCTACCA
TCCGCCGTGCCTCTTCTCGAGGCAACATCATCGTTGTCCCCCGACAACGCATTGGAATCT
TCGTCCCTCACAGCAGAGGAACGGCGCCGGCGTCGAATGTCATCAGCATCATCTGTCATG
CAGTCCTTTACCCTCTTTGGTTTTGTGGAGGATATAAGTGTCGTCGGCTTCAACGCCCAC
GTGTCTTCCGCCTTGATGACGGGCATCCTTCGAAGCATT
```

or load it from disk: Browse...

or choose Database to search: 5'UTR

Query Pattern

Enter a query pattern:

```
p1=5...9
```

a)

Query Pattern

Enter a query pattern:

```
p1=5...9
```

or load it from disk Browse...

Parameters

Complement: □ Overlap: □

Output Format: Standard Format

Max Match (1..1000): 100

Max Total Match (1..1000000): 5000

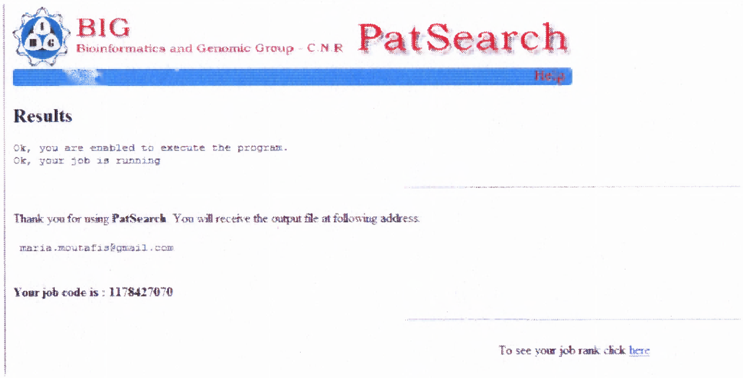
Advanced Parameters:

Results by e-mail to: maria.moutafis@gmail.c Execute

For any problem contact the Staff

b)

Figure 9.2 Input of data into PatSearch. a) Screenshot of first half of PatSearch homepage illustrating an example input. b) Screenshot of second half of PatSearch homepage illustrating an example input.



Results

```
Ok, you are enabled to execute the program.
Ok, your job is running
```

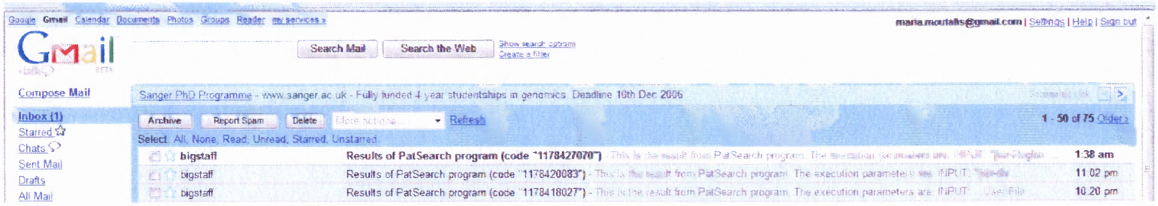
Thank you for using **PatSearch**. You will receive the output file at following address:

```
 maria.moutafis@gmail.com
```

Your job code is : 1178427070

To see your job rank click here

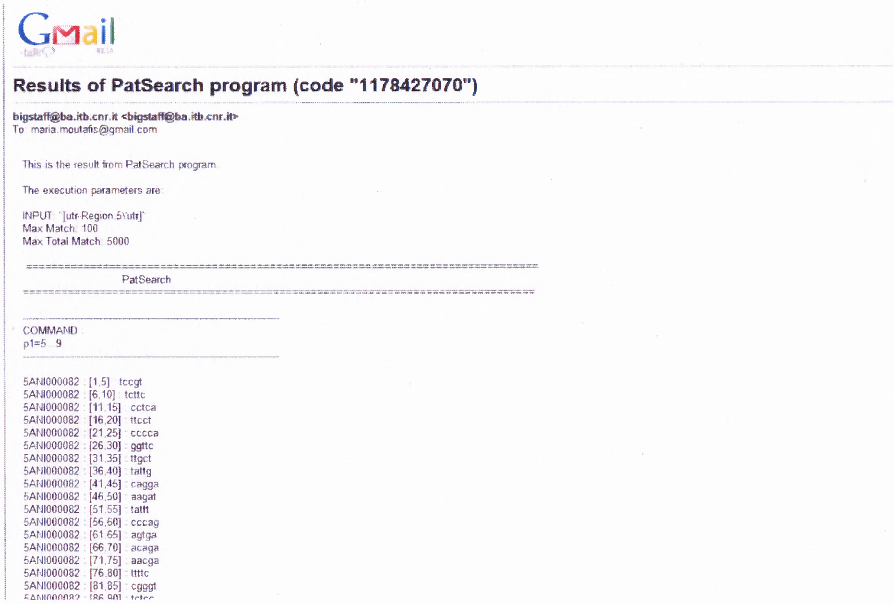
Figure 9.3 PatSearch results web page.



Sanger PhD Programme - www.sanger.ac.uk - Fully funded 4 year studentships in genomics. Deadline 10th Dec 2006

```
bigstaff   Results of PatSearch program (code "1178427070") - This is the result from PatSearch program. The execution parameters are: INPUT: "[utr-Region ...   1:38 am
bigstaff   Results of PatSearch program (code "1178420083") - This is the result from PatSearch program. The execution parameters are: INPUT: "[utr-Re ...    11:02 pm
bigstaff   Results of PatSearch program (code "1178418027") - This is the result from PatSearch program. The execution parameters are: INPUT: ... User File ... 10:20 pm
```

a)

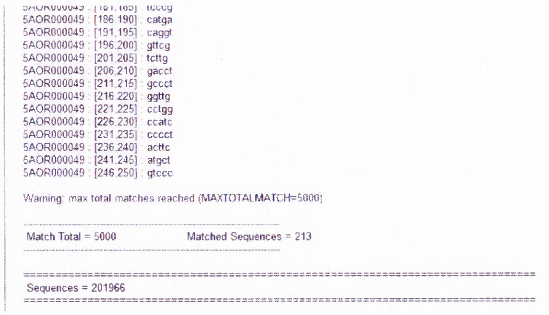


Results of PatSearch program (code "1178427070")

bigstaff@ba.itb.cnr.it <bigstaff@ba.itb.cnr.it>
To: maria.moutafis@gmail.com

```
This is the result from PatSearch program.

The execution parameters are:

INPUT: "[utr-Region:5\'utr]"
Max Match: 100
Max Total Match: 5000

 ==========================================================================
                 PatSearch
==========================================================================

-----------------------------------------------------
COMMAND :
p1=5...9
-----------------------------------------------------

5ANI000082 : [1,5] : tccgt
5ANI000082 : [6,10] : tcttc
5ANI000082 : [11,15] : cctca
5ANI000082 : [16,20] : ttcct
5ANI000082 : [21,25] : cccca
5ANI000082 : [26,30] : ggttc
5ANI000082 : [31,35] : ttgct
5ANI000082 : [36,40] : tattg
5ANI000082 : [41,45] : cagga
5ANI000082 : [46,50] : aagat
5ANI000082 : [51,55] : tattt
5ANI000082 : [56,60] : cccag
5ANI000082 : [61,65] : agtga
5ANI000082 : [66,70] : acaga
5ANI000082 : [71,75] : aacga
5ANI000082 : [76,80] : ttttc
5ANI000082 : [81,85] : cgggt
5ANI000082 : [86,90] : tctcc
```

b)



```
5AOR000049 : [181,185] : tcccg
5AOR000049 : [186,190] : catga
5AOR000049 : [191,195] : caggt
5AOR000049 : [196,200] : gttcg
5AOR000049 : [201,205] : tcttg
5AOR000049 : [206,210] : gacct
5AOR000049 : [211,215] : gccct
5AOR000049 : [216,220] : ggttg
5AOR000049 : [221,225] : cctgg
5AOR000049 : [226,230] : ccatc
5AOR000049 : [231,235] : cccct
5AOR000049 : [236,240] : acttc
5AOR000049 : [241,245] : atgct
5AOR000049 : [246,250] : gtccc

Warning: max total matches reached (MAXTOTALMATCH=5000)

-----------------------------------------------------
 Match Total = 5000            Matched Sequences = 213
-----------------------------------------------------

==========================================================================
 Sequences = 201966
==========================================================================
```

c)

Figure 9.4 PatSearch emailed results. a) View of Inbox showing email received from "bigstaff". b) Shows the top half of the emailed results. c) Shows the bottom half of the emailed results.

CHAPTER 10

CONCLUSIONS

The field of science is a fast ever-changing area of study. New and improved technologies emerge daily. They enable scientists to carry out studies with more easy and more accurate results, whether it is in the form of machines, programs/software, or techniques/protocols. However, different tools use different methods of data determination, which also results in variations in results from program to program. How does a scientist determine which is the appropriate tool to use? For the case of UTR determining programs, that was the question being answered here.

Three hundred Trypanosoma brucei sequences were downloaded and ran on BLASTn and ORBIT. Of those sequences ran on BLASTn, 203 were classified as mRNA sequences for Trypanosoma brucei. The rest of the sequences were determined to be either DNA or classified under other species. Those same sequences were run on ORBIT. ORBIT found that 121 of those sequences to be a match for mRNA sequences for Trypanosoma brucei. Of those sequences, the two programs agreed on only 89 to be coding. The difference between the two programs is astonishing. Between the two programs, it is my opinion that ORBIT is much easier to use and not as complex. However, the information obtained from BLAST was much more informative and useful. The retrieval of the above data was extremely long and tedious, especially since it was done manually, sequence by sequence.

Verifying which of the two programs gave correct predictions, was even more complicated than the data retrieval. Obtaining access/permission to use the UTRscan and PatSearch tools was a difficult task of its own. Reply time from staff for access alone took over three weeks for one of the programs. In the end, the staff of PatSearch was the only one that granted me permission to use their analysis tool. This was unfortunate, because between the two programs, it was PatSearch that was less user-friendly and more difficult with data retrieval. The 89 sequences run on PatSearch found matches for 72 of the sequences, that equals approximately 81%. These results help make BLASTn the definite tool of choice.

APPENDIX

```
--------------> UTRscan Results <-------------------
Processed sequences: 1
-----------------------------------------------------
          Pattern = HISTONE3
-----------------------------------------------------
Pattern not found
-----------------------------------------------------
           Pattern = IRE
-----------------------------------------------------
Pattern not found

-----------------------------------------------------
          Pattern = SECIS
-----------------------------------------------------
Pattern not found

-----------------------------------------------------
          Pattern = APP
-----------------------------------------------------
Pattern not found
-----------------------------------------------------
          Pattern = CPE
-----------------------------------------------------
Pattern not found

-----------------------------------------------------
          Pattern = TGE
-----------------------------------------------------
Pattern not found
-----------------------------------------------------
          Pattern = NANOS
-----------------------------------------------------
Pattern not found

-----------------------------------------------------
          Pattern = 15-LOX-DICE
-----------------------------------------------------
Found 13 matches in 1 sequences
EM_OM:M27214 :[2066,2084]    :GCCCACCCTCT CCCCC AAG
EM_OM:M27214 :[2085,2103]    :CCCTGCCCTCT TCCCC AAG
EM_OM:M27214 :[2104,2122]    :CCCTGCCCCT TTCCCC AAG
EM_OM:M27214 :[2123,2142]    :CCCCATCCTCT TTCCCC AAG
EM_OM:M27214 :[2143,2161]    :CCCCGCCCTCT TCCCC AAG
EM_OM:M27214 :[2162,2180]    :CCCCGCCCTCT TCCCC AAG
EM_OM:M27214 :[2181,2199]    :CCCCACCCTCT TCCCC AAG
EM_OM:M27214 :[2200,2220]    :CCCCGCCCTCT TTCCCCC ACG
EM_OM:M27214 :[2221,2240]    :CCCCACCCTCT TTCCCC AAG
EM_OM:M27214 :[2241,2259]    :CCCCGCCCTCT TCCCA AAG
EM_OM:M27214 :[3319,3338]    :CCCCTCCTTC TCTCTGG AGG
EM_OM:M27214 :[3345,3361]    :CCCCTTCTCC TTGG AGG
EM_OM:M27214 :[3437,3453]    :CCCCTTCTCC TTGG AGG
```

```
Checking repeats for 15-LOX-DICE    (min: 2)
        Found 1 matches for pattern

15-LOX-DICE EM_OM:M27214 :[2066,2259] :10:
GCCCACCCTCT CCCCC AAG CCCTGCCCTCT TCCCC AAG CCCTGCCCCT
TTCCCC AAG CCCCATCCTCT    TTCCCC AAG CCCCGCCCTCT TCCCC
AAG CCCCGCCCTCT TCCCC AAG CCCCACCCTCT TCCCC AAG
CCCCGCCCTCT TTCCCCC ACG CCCCACCCTCT TTCCCC AAG
CCCCGCCCTCT TCCCA AAG

-------------------------------------------------------
                 Pattern = ARE2
-------------------------------------------------------
Found 0 matches

-------------------------------------------------------
                 Pattern = POLY-PY
-------------------------------------------------------
Pattern not found

-------------------------------------------------------
                 Pattern = GLUT1
-------------------------------------------------------
Pattern not found
-------------------------------------------------------
                 Pattern = TNF
-------------------------------------------------------
Pattern not found
-------------------------------------------------------
                 Pattern = VIMENTIN
-------------------------------------------------------
Pattern not found
-------------------------------------------------------
                 Pattern = IRES
-------------------------------------------------------
Pattern not found
-------------------------------------------------------
                 Pattern = MSL2-5UTR
-------------------------------------------------------
Pattern not found

-------------------------------------------------------
                 Pattern = MSL2-3UTR
-------------------------------------------------------
```

Pattern not found

Figure A.1 Example of UTRscan output data.
http://www.ba.itb.cnr.it/BIG/UTRScan/HelpFile/Ultrscanhelp.htm#Output

REFERENCES

Benz, C., Nilsson, D., Andersson, B., Clayton, C., & Lys Guilbride, D. (2005). Messenger RNA processing sites in Trypanosoma brucei. *Molecular & Biochemical Parasitology, 143,* 125-134.

Gopal, S., Cross, G. A. M., & Gaasterland, T. (2003). An organism specific method to Rank predicted coding regions in Trypanosoma brucei. *Nucleic Acids Research, 31,* 5877-5885.

Kleinsmith, L. J., & Kish, V. M. (1995). *Principles of Cell and Molecular Biology.* New York: HarperCollins College Publishers.

Pesole, G. & Liuni, S. (1999). Internet resources for the functional analysis of 5' and 3' untranslated regions of eukaryotic mRNAs. *Trends in Genetics, 15(9),* 378.

Pesole, G., Grillo, G., Larizza, A., & Liuni, S. (2000). The untranslated regions of eukaryotic mRNAs: Structure, function, evolution and bioinformatic tools for their analysis. *Briefings in Bioinformatics, 1(3),* 236-249.

Quackenbush, J., Liang, F., Holt, I., Pertea, G., Upton, J. (2000). The TIGR Gene Indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Research, 28(1),* 141-145.