ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΣΤΗΝ ΙΑΤΡΙΚΗ ΚΑΙ ΤΗ ΒΙΟΛΟΓΙΑ

Κατεύθυνση: Βιοπληροφορική

Αλγόριθμοι στη Μοριακή Βιολογία

Αναφορά - IS200 transposases HMM και Φυλλογενετικά Δέντρα

Επιβλέπων: **Βερνίκος Γεώργιος**, Ερευνητής - Λέκτορας, ΕΚΠΑ

ΑΘΗΝΑ

ΦΕΒΡΟΥΑΡΙΟΣ 2013

Αλγόριθμοι στη Μοριακή Βιολογία

Αναφορά - IS200 transposases HMM και Φυλλογενετικά Δέντρα

Μπεγέτης Ι. Νικόλαος

ΑΜ: ΠΙΒ0111

ΕΠΙΒΛΕΠΩΝ :

Βερνίκος Γεώργιος, Ερευνητής - Λέκτορας, ΕΚΠΑ

Λέξεις-Κλειδιά :

IS200 transposase, hmm, S. Typhi CT18, S. Typhi TY2, blast, taxa, phylogenetic trees, supertree

Algorithms in Molecular Biology

Report - IS200 transposases HMM and Phylogenetic Trees

Begetis I. Nikolaos

A.N.: PIV0111

SUPERVISOR:

Vernikos George, Investigator - Lecturer, NKUA

Keywords:

IS200 transposase, hmm, S. Typhi CT18, S. Typhi TY2, blast, taxa, phylogenetic trees, supertree

Table Of Contents

List Of Figures

CHAPTER 1

Initially, in this chapter it is provided some information about our assignment's subject so that the reader of the report can easily follow the text flow. Subsequently, there will be provided the answers to the assignment's demands step by step.

1.1 Introduction

This section of the report is intended to give a few definitions relevant to the assignment's prerequisites. To start with, molecular biology is the discipline of biology that deals with the molecular basis of biological activity and studies the complex interactions among biological molecules. This field overlaps with other areas of biology and chemistry, and in particular with genetics and biochemistry.

A key component of genetics and molecular biology is genome. The genome is the alpha and the omega of an organism's hereditary information, and it is encoded either in DNA or, in cases, in RNA. The genome sequence includes both the genes (i.e., sequences of DNA/RNA that encode protein) and the non-coding sequences of the DNA/RNA. In addition to this, all the genome's genes have a specific location in the DNA sequence on a chromosome, called locus [1].

Sometimes, though, a DNA sequence may change its position within the genome, i.e. chromosomal segment is transferred to a new position on the same or another chromosome, with result in creating mutations and altering the cell's genome size (Figure 1.1). This DNA sequence is called Transposable element (abbrf. TE).

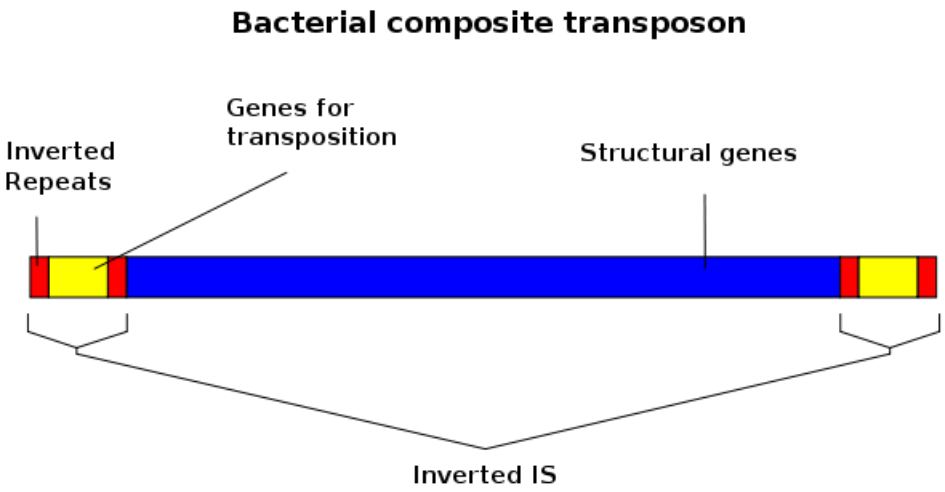


Figure 1.1: *Bacterial Composite Transposon*[1].

1.1.1 Insertion Sequence

In reference to all the above, an insertion sequence [2] (abbr. IS) is a short DNA sequence that acts as a simple transposable element (Figure 2.1). Insertion sequences can be discriminated due to the small size relatively to other transposable elements (generally around 700 to 2500 bp in length), and also because only code for proteins is implicated in the transposition activity. These proteins are usually the transposase which catalyses the enzymatic reaction allowing the IS to move, and also one regulatory protein which either stimulates or inhibits the transposition activity. It is also worth mentioning that the coding region in an insertion sequence is usually flanked by inverted repeats.

A particular IS may be named according to the form ISn, where n is a number (e.g. IS1, IS2, IS3, IS10, IS50, IS911, IS26 etc.). Although insertion sequences are usually discussed in the context of prokaryotic genomes (Figure 2.1), certain eukaryotic DNA sequences belonging to the family of Tc1/mariner transposable elements may also be considered to be insertion sequences.

1.1.2 IS200 transposases - Assignment's Requirements

In this case study we are assigned to make a report for IS200[2] (Figure 2.1, 2.1) transposases and especially we have to find all the genes that encode the IS200 transposase and exist in Salmonella Typhi CT18 and Salmonella Typhi TY2 by finding statistically significant

[1] http://en.wikipedia.org/wiki/File:Composite_transposon.svg
[2] http://pfam.sanger.ac.uk/family/Y1_Tnp

predictions respectively. In addition, we have to make comparisons in the results and create phylogenetic trees, a consensus tree and a supertree.

Before moving on to the next section and having given all the above informational material it would be recommended to give some further definitions about the bacteria in general, the Salmonella enterobacteria and the Salmonella Typhi in particular.

1.1.2.1 Bacteria

Bacteria constitute a large domain of prokaryotic microorganisms. Despite, typically a few micrometres in length, bacteria have a wide range of shapes, ranging from spheres to rods and spirals. Bacteria were among the first life forms to appear on Earth, and are present almost everywhere; growing in soil, water, acidic hot springs, radioactive waste, deep in the Earth's crust, as well as in organic matter and the live bodies of plants and animals, providing outstanding examples of mutualism in the digestive tracts of humans, termites and cockroaches.

1.1.2.2 Salmonella

Salmonella is a genre of enterobacteria with diameters around 0.7 to 1.5 µm, lengths from 2 to 5 µm, and flagella that grade in all directions (i.e., peritrichous). They are chemoorganotrophs, obtaining their energy from oxidation and reduction reactions using organic sources, and are facultative anaerobes. Moreover, Salmonella is closely related to the Escherichia genus and are found worldwide in cold- and warm-blooded animals (including humans), and in the environment. They cause illnesses such as typhoid fever, paratyphoid fever, and food-borne illness. It is worth to mention also that IS200 element is found in Escherichia Coli[3] too (Figure 1.1.2.2), as these both have a 60-70% similarity in their DNA. Yet, the IS200 element is more identical to that of Salmonella Typhimurium [3], than Salmonella Typhi. A list of Salmonella serovars that encode transposase IS200 is provided in Figure 1.1.2.2.

Figure 1.2: *tnp-transposase for IS200 element Taxonomy. Inside the parentheses is the number of annotated features per bacterial groups*[5].

[3]http://www.ncbi.nlm.nih.gov/gene/12887025

[5]http://www.ncbi.nlm.nih.gov/gene/?term=tnp+transposase+for+insertion+sequence+element+IS200

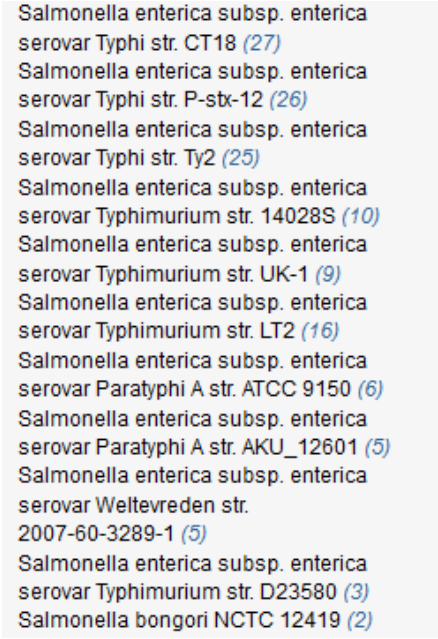


Figure 1.3: *tnp-transposase for IS200 element in Salmonella. Inside the parentheses is the number of annotated features per serovars*[7].

1.1.2.3 Salmonella Typhi

Salmonella Typhi[8] (abbr. S. Typhi) [4][7], is the causative agent of typhoid fever (Figure 1.4). Although typhoid fever is not widespread in the United States, it is very common in underdeveloped countries, and causes a serious, often fatal disease. The symptoms of typhoid fever include nausea, vomiting, fever and death. Unlike the other Salmonella serovars (Typhirium and Enteritidis), S. Typhi can only infect humans, and no other host has been identified. The main source of S. Typhi infection is from swallowing infected water. Food may also be contaminated with S. Typhi, if it is washed or irrigated with contaminated water[9].

A more detailed schema that includes all the organisms that are relevant to IS200 element is shown in Figure 2.1.

[7]http://www.ncbi.nlm.nih.gov/gene/?term=(tnp+transposase+for+insertion+sequence+element+IS200)+AND+"Salmonella"[porgn:__txid590]
[8]http://microbewiki.kenyon.edu/index.php/Salmonella_typhi
[9]http://www.salmonella.org/info.html

Figure 1.4: *Salmonella Typhi*[10].

1.2 Salmonella Typhi CT18

In this section one can find, step by step, the whole procedure followed so as to find all the genes of Salmonella Typhi CT18[11] that encode the IS200 transposase element (using the Artemis[12] Genome Browser and Annotation tool), and then use these DNA sequences with a multiple sequence alignment in order to build the hidden markov model (abbr. hmm) profile for the CT18 IS200 family sequences.

Subsequently, the produced hmm is searched in S.Typhi's CT18 genome and the output with all predictions is written in a .tab file that can be executed by Artemis keeping all the information about the predicted positions in the genome and the relevant score. The genome score is converted in a color scale of white-red for small and big scores, respectively.

Having done all the aforementioned the .tab file is loaded in Artemis and the results are inspected to find if there exists any overlap with the existing annotation for transposases IS200. Then, dynamically, the optimal score threshold is designated in order to determine the statistically important predictions. An arbitrary cutoff is selected and finally there are presented all the statistically significant predictions for the S.Typhi's CT18 genome.

[10] http://phys.org/news83593133.html
[11] http://www.ncbi.nlm.nih.gov/bioproject/57793
[12] http://www.sanger.ac.uk/resources/software/artemis/
[13] www.genome.jp/kegg-bin/show_organism?org=sty

Genome information

T number	T00064
Org code	sty
Aliases	SALTI, 220341
Full name	Salmonella enterica subsp. enterica serovar Typhi CT18
Definition	Salmonella enterica subsp. enterica serovar Typhi CT18 (Salmonella typhi CT18)
Annotation	manual
Taxonomy	TAX: 220341
Lineage	Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Salmonella
Data source	RefSeq (Project:57793)
Original DB	GeneDB, Sanger
Keywords	Human pathogen
Disease	H00111 Typhoid fever
Comment	Isolated in December 1993 at the Mekong Delta region of Vietnam, from blood of a 9-year-old girl who was suffering from typhoid.
Chromosome	Circular
Sequence	RS: NC_003198
Length	4809037
Plasmid	pHCM1; Circular
Sequence	RS: NC_003384
Length	218160
Plasmid	pHCM2; Circular
Sequence	RS: NC_003385
Length	106516
Statistics	Number of nucleotides: 5133713 Number of protein genes: 4769 Number of RNA genes: 112
Reference	PMID: 11677608
Authors	Parkhill J, et al.
Title	Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18.
Journal	Nature 413:848-52 (2001)

Figure 1.5: *S. Typhi Genome CT18 Information*[13].

1.2.1 Genes of S. Typhi CT18 encoding IS200 transposase element

In order to find all the genes of Salmonella Typhi CT18 that encode the IS200 transposase element there was used the Artemis Genome Browser and Annotation tool. At first, following some valid manual notes[14,15] we searched the Sanger Institute[16] and found the data[17] needed to begin our research. As an entry in Artemis we used the feature file named St.art, included in data. This file contains feature table[18] lines (the FT lines) of the S. Typhi genome at the head and follows the genome's sequence with its features (Figure 2.2), i.e. regions of DNA that have been annotated with a key or type, and zero or more qualifiers. Figure 1.2.1 shows the Artemis window with the St.art entry.

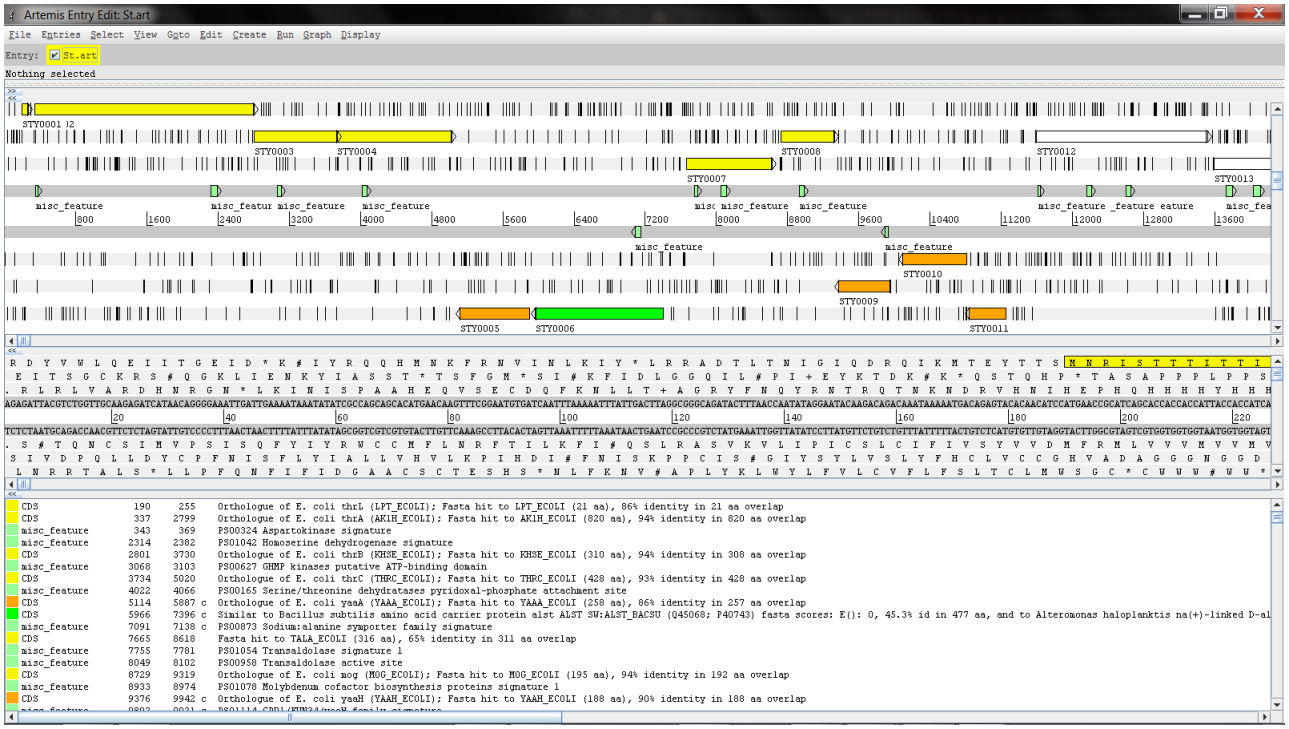


Figure 1.6: *Artemis window - St.art Entry*

Then, in order to find all the genes of S. Typhi's CT18 genome that encode the IS200 transposase element we selected the coding sequences that encode and produce the IS200 transosase, using the "Select" menu tab of Artemis and in the "Feature Selector" picking out the first two checkboxes: "Select by Key" giving 'CDS' and "Select by Qualifier" giving 'Product'

[14] http://bioinfo.hr/bioinfo2011/materials/Artemis_Manual_Cambridge_June2011.pdf
[15] ftp://ftp.sanger.ac.uk/pub4/resources/software/artemis/artemis.pdf
[16] http://www.sanger.ac.uk/
[17] ftp://ftp.sanger.ac.uk/pub/pathogens/Salmonella/typhi/
[18] ftp://ftp.ebi.ac.uk/pub/databases/embl/doc/FT_current.html

that contains the text 'IS200' (Figure 1.2.1). Figure 2.2 shows in detail the 25 Coding DNA Sequences (abbr. CDS) and the associated information (features).

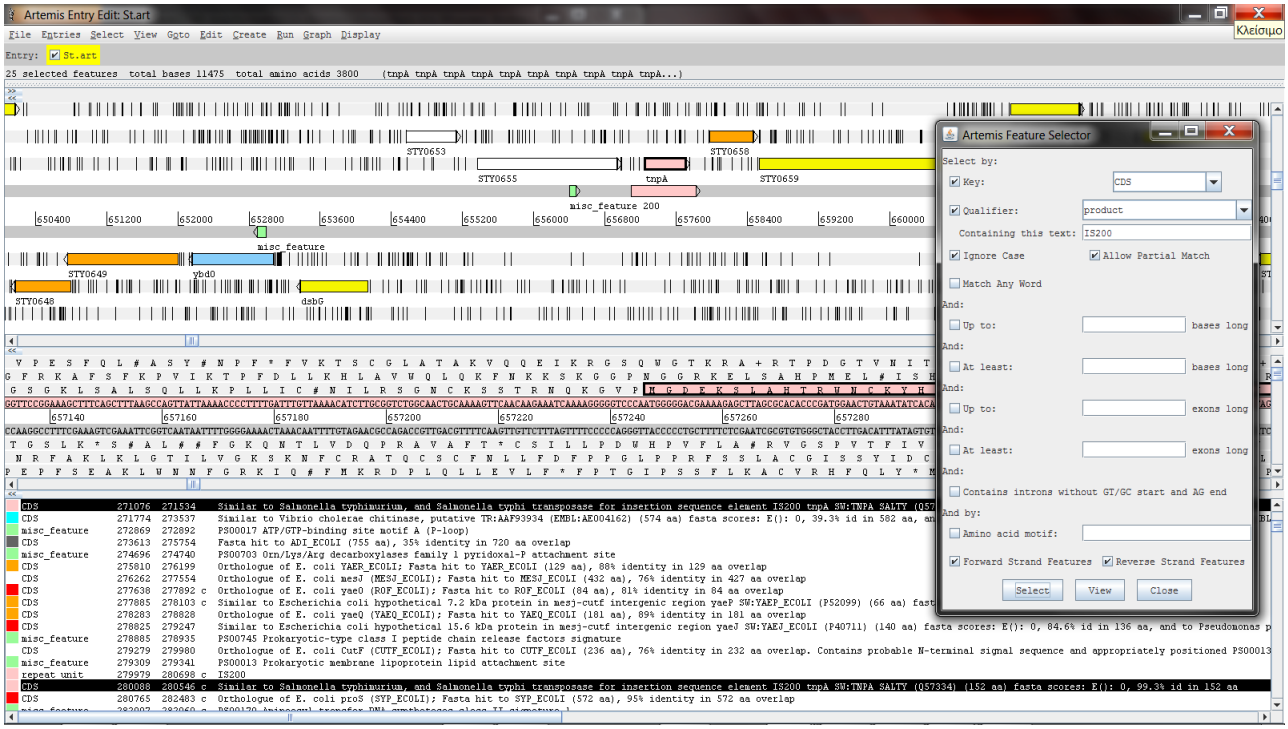


Figure 1.7: *Artemis window - CDSs producing IS200 transposase. At the top left of the window it is shown that there exist 25 CDSs that produce IS200 transposase and at the main window view and the bottom window view there are highlighted both the coding sequences in S. Typhi CT18 genome and the features' annotations for IS200 transposase*

Now that the coding DNA sequences are found we need to export their DNA bases in a .fasta file which we can use later to make a multiple sequence alignment. To do that, we select all the featured CDSs found from above and right-click on them to select "Write", then "Bases of Selection" and, after that, "FASTA Format" (Figure 1.2.1). In our case study the filename we gave to the exported file is *CT18_IS200.fasta* (Figure 2.2).

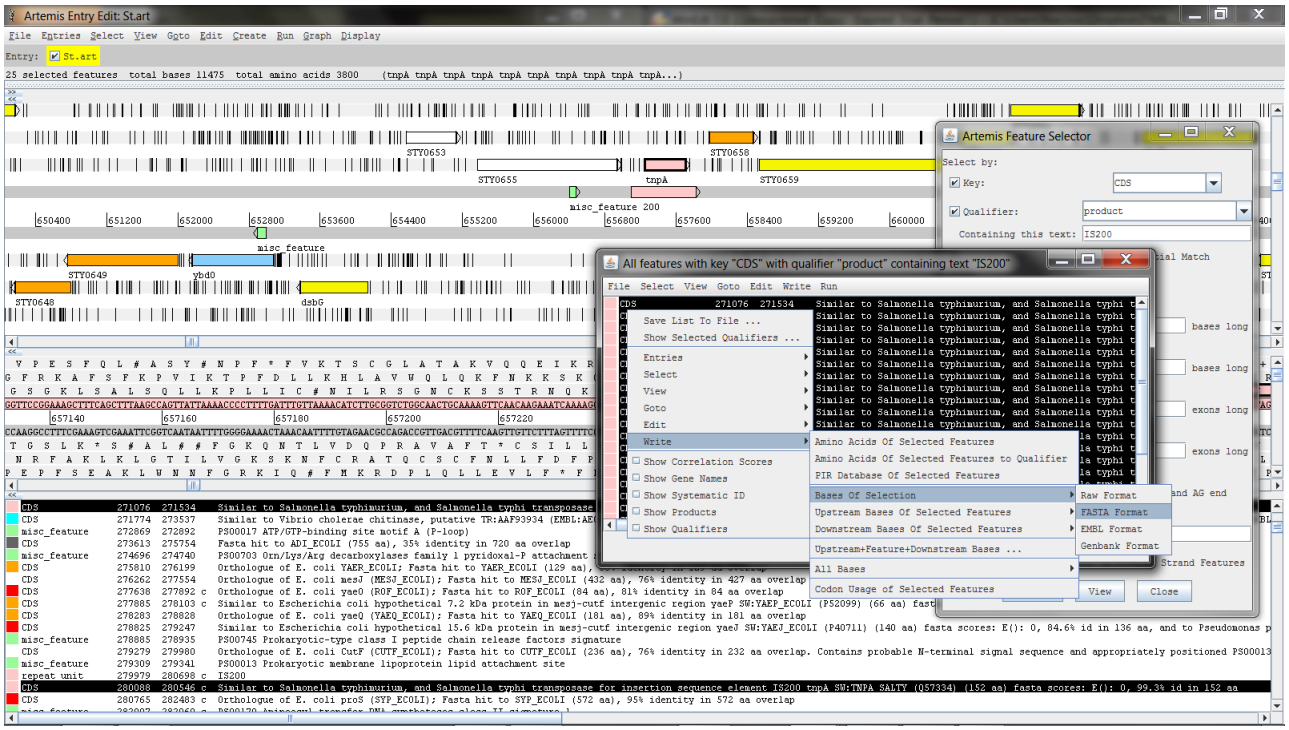


Figure 1.8: *Artemis window - CDSs Fasta Export*

1.2.2 Multiple sequence alignment for IS200 CDSs of S. Typhi CT18

The fasta file is ready and the next task to be achieved is to make a multiple sequence alignment among the coding DNA sequences found in *CT18_IS200.fasta*. To do this we search the tools of the European Bioinformatics Institute[19] (abbr. EBI) and selected the tool ClustalW2[20] for the multiple sequence alignment (abbr. msa) of our S. Typhi CT18 genes that produce IS200 transposase. ClustalW2 is suitable for this procedure in contrast to some of the rest tools, because it supports not only multiple sequence alignment for protein but also for DNA. ClustalW2 attempts to calculate the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen.

Initially, as in most tools, in step 1 we set the sequence type as 'DNA' and the input by uploading the '*CT18_IS200.fasta*' file, which as indicated above contains 25 CDSs.

In step 2 we preferred setting the alignment type as 'slow' because it more accurate. We set the DNA weight matrix for the slow pairwise alignment nucleotide sequence comparison matrix used to score alignment as 'IUB' (default), the GAP Open for the first residue in a gap with a '10' value (default) and the Gap Extension for each additional residue in a gap with a '0.1' value (default).

[19] http://www.ebi.ac.uk/Tools/

[20] http://www.ebi.ac.uk/Tools/msa/clustalw2/

In step 3 the DNA Weight Matrix for the multiple alignment nucleotide sequence comparison matrix used to score the alignment was set as 'IUB' again, the Gap Open for the first residue in a gap was set with a '10' value, the Gap Extension for each additional residue in a gap was set with a '0.20' value, the penalty for Gap Distances that are closer together was set in the distance of '5' gaps, the No End Gaps that disable the gap separation penalty when scoring gaps reach the ends of the alignment was set as 'No', the Iteration was set with the iteration type of 'tree', so that the iteration is done at each step of alignment process, the Num Iter that indicates the maximum number of iterations to perform was set as '5', the Clustering type was set as 'NJ' indicating the Neighbour-joining clustering (Saitou and Nei 1987) and finally the output format for generated multiple sequence alignment was set as 'GCG MSF', because it is the only one supported by the tool we will use for building the hmm profile (HMMER). The output, GCG Multiple Sequence File alignment format, was also set in an 'aligned' order of the final alignment determined by the guide tree.

Figure 1.2.2 shows the submitted form of ClustalW2 and Figure 2.2 shows the resulted CT18_IS200.msf file after the submission of the form done in step 4 of ClustalW2 Multiple Sequence Alignment Tool, while Figure 1.2.2 shows the MSA results visualized by JalView that is hyperlinked from ClustalW2.

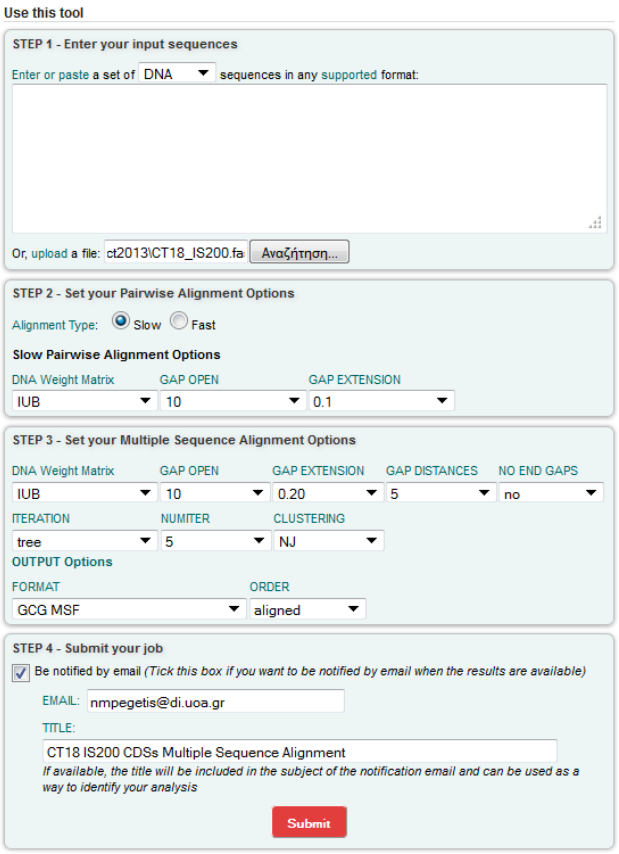


Figure 1.9: *ClustalW2 Multiple Sequence Alignment*

Figure 1.10: *ClustalW2 Multiple Sequence Alignment - Jalview Visualization*

We also loaded the output CT18_IS200.msf file in Unipro UGENE[21] visualization tool which is one of the best public licensed tools for alignment visualization[22]. Figure 1.2.2 shows the visualized result. A further set of screenshots is presented in the miscellaneous chapter (Figure 2.2, 2.2)

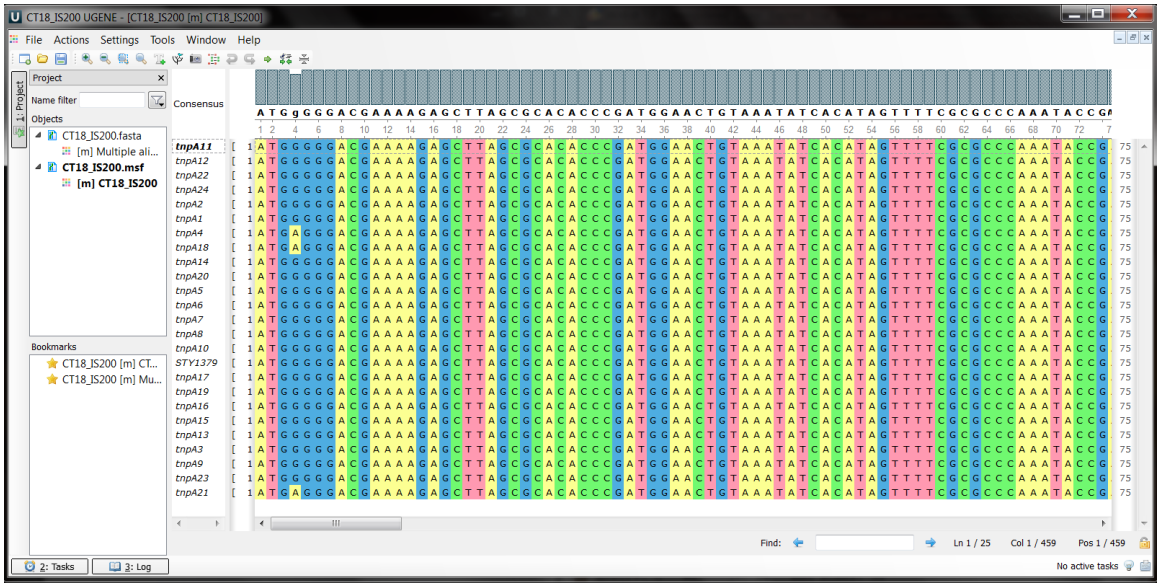


Figure 1.11: *UGENE MSA of CT18_IS200.msf. The main central window represents a part of the msa from the 1st-75th base of each CDS. At the top of the main window it is shown the consensus sequence after the msa. Below that, the msa of all CDSs is represented with each of the AGTC bases colored respectively.*

[21] http://ugene.unipro.ru/

[22] http://en.wikipedia.org/wiki/List_of_alignment_visualization_software

1.2.3 Hidden Markov Model profile for IS200 CDSs of S. Typhi CT18

Having completed the MSA task for the IS200 transposase family of S. Typhi CT18, next comes the building of the Hidden Markov Model profile (abbr. hmm). In Bioinformatics (ITMB, MD, 2012-13) class we were taught to use the HMMER v1.8 tool for building and searching the HMM profile for sequence families. HMMER v1.8 tool is the only one which is so specialized for nucleosomes and so we preferred using it. In this report, though, we are also going to use the latest version of HMMER (v3.0), because it is featured enough (for what we are assigned to do) and is provided not only with the console executables, but also in GUI environment (as a plugin of UGENE tool), so both of the HMMER will be used, the v1.8, via the console window and the v3.0 via the GUI.

In this report both methods of building HMMs and searching these methods will be tried out. Beginning from the console environment having downloaded HMMER v1.8 software[23] for Windows, and having read the HMMER v1.8 and v3.0 release notes[24] and their User Guides[25] we confirmed that they both are fully compatible for the kind of DNA sequencing we are assigned to do. So, subsequently, we switched the console to the directory where we downloaded and extracted the HMMER software and tried out the command:

hmmb -h

As indicated from the build help results, the usage of hmmb can take a lot of parameters and options as shown in Figure 2.2 (2.3 HMM3). The main, though, necessary parameters of this command is to be followed by the hmm output file, and after that, the msa input file (in an GCG MSF format). So our executed command should be:

hmmb CT18_IS200_cmd.hmm CT18_IS200.msf

Then, after getting the CT18_IS200_cmd.hmm file (Figure 1.2.3) the task of building the hidden markov model is completed and the next step to be followed is to search for the builded hmm by querying in the S. Typhi CT18 genome sequence.

Figure 2.2 (2.3 HMM3) shows the results of hmm build command and the help screenshot of the:

hmmfs -h

console command that provides us with the usage of this command and its options. It also shows at the bottom the next command that we run:

hmmfs -c CT18_IS200_cmd.hmm St.dna CT18_IS200.out

where the result of the execution (searched also the complement strand) is printed in file CT18_IS200.out. With this command we search the S. Typhi CT18 genome (St.dna) with the previous output hidden markov model of IS200, CT18_IS200_cmd.hmm. If we do not redirect

[23] http://hmmer.janelia.org/software
[24] ftp://selab.janelia.org/pub/software/hmmer3/3.0/RELEASE-NOTES
[25] ftp://selab.janelia.org/pub/software/hmmer3/3.0/Userguide.pdf

the output in a file and the output results are printed in the stdout - the screen (Figure 2.2 (2.49 HMM3)), and as indicated because we didn't set any options there were used all the default options.

Now, it is time to see all the above executions in console, with a GUI. In the previous section that we talked about multiple sequence alignment we mentioned that we decided to use the UGENE tool because of the plugins and features that are installed with it.

To begin with, following the guiding lines introduced from the UGENE User Manual[26] (Chapter 11.14) the first thing to be done is to load S. Typhi's CT18 genome on our project (Figure 2.2). Then, we select from the menu the HMMER3 plugin following the path: "Tools", "HMMER tools", "HMMER3 tools", "Build HMM3 profile". On the HMM3 window that popups the Input alignment file should be the GCG MSF file, CT18_IS200.msf, and the builded profile should be the Hidden Markov Model CT18_IS200.hmm, that we want for output (Figure 2.2). When building the hmm of S. Typhi CT18 for IS200 transposase, we can select some other options too, from the build window tabs. The same options had been indicated in the console window before (Figure 2.2).

A part of the resulted CT18_IS200.hmm is shown in Figure 1.2.3.

St.art | CT18_IS200.fasta | CT18_IS200.msf | CT18_IS200.hmm | CT18_IS200_annotations.gb

```
HMMER3/b [3.0 | March 2010]
NAME  CT18_IS200
LENG  459
ALPH  DNA
RF    no
CS    no
MAP   yes
DATE  Sat Feb 16 13:01:01 2013
NSEQ  25
EFFN  1.217651
CKSUM 4134559237
STATS LOCAL MSV      -11.0166  0.69846
STATS LOCAL VITERBI  -12.5702  0.69846
STATS LOCAL FORWARD   -5.1174  0.69846
HMM          A        C        G        T
             m->m     m->i     m->d     i->m     i->i     d->m     d->d
  COMPO   1.24650  1.59397  1.28103  1.46269
          1.38629  1.38629  1.38629  1.38629
          0.06030  3.53154  3.53154  1.46634  0.26236  0.00000        *
      1   0.38320  2.42022  2.08065  2.25786      1 - -
          1.38629  1.38629  1.38629  1.38629
          0.06030  3.53154  3.53154  1.46634  0.26236  1.09861  0.40547
      2   2.19287  1.97304  2.26408  0.43782      2 - -
          1.38629  1.38629  1.38629  1.38629
          0.06030  3.53154  3.53154  1.46634  0.26236  1.09861  0.40547
      3   2.12660  2.54317  0.34336  2.37774      3 - -
          1.38629  1.38629  1.38629  1.38629
          0.06030  3.53154  3.53154  1.46634  0.26236  1.09861  0.40547
```

Figure 1.12: *HMM build using HMMER v3.0 - CT18_IS200.hmm*

On the next step, having built CT18_IS200.hmm, what is needed is to search for these hidden markov models in S. Typhi's CT18 genome for the CDSs including IS200 transposase. Figure 2.2 shows how we used the hmm file to search these models in the S. Typhi's genome, and Figure 2.2 shows the GUI result. What is more, it is worth to mention that in this step too, the options shown before in the console's hmmsearch, now are provided by the user through selection lists and checkboxes. To conclude with this subsection, the output file

[26] http://ugene.unipro.ru/downloads/UniproUGENE_UserManual.pdf

format that we selected to export the result of the HMM search is "genbank", because it is pretty alike that one of FT .tab used in Artemis, as we want in the next step (Figure 1.2.3).

St.art | CT18_IS200.fasta | CT18_IS200.msf | CT18_IS200.hmm | CT18_IS200_annotations.gb

```
LOCUS       Annotations                                    16-FEB-2013
UNIMARK     Annotations
FEATURES             Location/Qualifiers
     CDS             271076..271534
                     /HMM model="CT18_IS200"
                     /Independent e-value="4.72268e-176"
                     /Conditional e-value="4.72268e-176"
                     /Score="572.872192"
                     /Bias="5.13461e+00"
                     /Accuracy per residue="9.96909e-01"
                     /HMM region="1...459"
                     /Envelope of domain location="271076...271534"
                     /ugene_group="IS200"
     CDS             complement(280088..280546)
                     /HMM model="CT18_IS200"
                     /Independent e-value="1.04547e-175"
                     /Conditional e-value="1.04547e-175"
                     /Score="571.734436"
                     /Bias="4.98651e+00"
                     /Accuracy per residue="9.96908e-01"
                     /HMM region="1...459"
                     /Envelope of domain location="280087...280545"
                     /ugene_group="IS200"
     CDS             446265..446723
                     /HMM model="CT18_IS200"
                     /Independent e-value="2.10123e-176"
                     /Conditional e-value="2.10123e-176"
                     /Score="574.031677"
                     /Bias="4.95158e+00"
                     /Accuracy per residue="9.96908e-01"
                     /HMM region="1...459"
                     /Envelope of domain location="446265...446723"
                     /ugene_group="IS200"
     CDS             507385..507843
                     /HMM model="CT18_IS200"
                     /Independent e-value="1.31591e-174"
                     /Conditional e-value="1.31591e-174"
                     /Score="568.108398"
                     /Bias="5.55001e+00"
                     /Accuracy per residue="9.96350e-01"
                     /HMM region="1...459"
                     /Envelope of domain location="507385...507843"
                     /ugene_group="IS200"
```

Figure 1.13: *HMM search results using HMMER v3.0 - CT18_IS200.gb*

1.2.4 HMM output predictions conversion in .tab file to be used with Artemis

Moving to the next step, we have to convert the output predictions to a .tab file that can be read by the Artemis tool and keep the relative position prediction information and the hmm score. To do that, because .gb (genbank) files share a lot in common with the Artemis .tab files format, we searched for utilized converter between these two format and we came into molecular biology tools[27] implemented by University of Guelph[28], Canada. There we found a variety of converters[29] between Molecular Biology file formats, and for our case, we found a tool converting genbank files to artemis tab files[30]. Figure 1.2.4 shows the input genbank file content and the output tab content. We also found a tool[31] written in python that does the same conversion. And the console command to use it should be (being in the right directory): *gb2tab.py -f CT18_IS200_predictions.tab CT18_IS200_predictions.gb*

Converting from **tab** to **genbank** sequence file format

Sequence multi-format conversion tool

About tab	Simple two column tab separated sequence files, where each line holds a record's identifier and sequence. For example, this is used by Aligent's eArray software when saving microarray probes in a minimal tab delimited text file.
About genbank	The GenBank or GenPept flat file format.
tab content	CDS 271076..271534 /HMM model="CT18_IS200" /Independent e-value="4.72268e-176" /Conditional e-value="4.72268e-176" /Score="572.872192" /Bias="5.13461e+00" /Accuracy per residue="9.96909e-01" /HMM region="1...459" /Envelope of domain location="271076...271534" /ugene_group="IS200" CDS complement(280088..280546)
Alphabet (optional)	None DNA RNA Protein Nucleotide
genbank converted content	FT CDS 271076..271534 FT /systematic_id="pred0" FT /score=572.872192 FT CDS 356772..356783 FT /systematic_id="pred1" FT /score=172.209 FT CDS 446265..446723 FT /systematic_id="pred2" FT /score=574.031677 FT CDS 507385..507843 FT /systematic_id="pred3" FT /score=568.108398 FT CDS 657246..657704 FT /systematic_id="pred4" FT /score=573.380005 FT CDS 864041..864499 FT /systematic_id="pred5" FT /score=573.380005 FT CDS 1082125..1082583

Figure 1.14: *Conversion of CT18_IS200.gb in CT18_IS200.tab*

Now that the input CT18_IS200.tab is created with both the relative position prediction information and the hmm score, it is worth saying before moving on, that with a perl simple

[27] http://www.molbiol-tools.ca/Convert.htm
[28] http://www.molbiol-tools.ca/
[29] http://sequenceconversion.bugaco.com/converter/biology/sequences/
[30] https://app.bugaco.com//converter/biology/sequences/genbank_to_tab.php
[31] http://wiki.christophchamp.com/index.php/TAB_file_format

tool we implemented, we transformed the genbank output score in a 100% scale for the Artemis tab, we added a color table feature (abbr. FT) which ranges in a color scale of white-red depending on its score, a note FT and a product FT. The final result is shown in Figure 1.2.4.

CT18_IS200.msf | CT18_IS200.hmm | CT18_IS200_predictions.gb | CT18_IS200_predictions.tab

```
1   FT   CDS             271076..271534
2   FT                   /systematic_id="pred0"
3   FT                   /colour=255 0 0
4   FT                   /score=99.84
5   FT                   /note="putative transposase for insertion sequence
6   FT                   element IS200 with hmmer score = 901.00"
7   FT                   /product="hypothetical transposase for IS200"
8   FT   CDS             356772..356783
9   FT                   /systematic_id="pred1"
10  FT                   /colour=255 254 254
11  FT                   /score=0.30
12  FT                   /note="putative transposase for insertion sequence
13  FT                   element IS200 with hmmer score = 2.98"
14  FT                   /product="hypothetical transposase for IS200"
15  FT   CDS             446265..446723
16  FT                   /systematic_id="pred2"
17  FT                   /colour=255 0 0
18  FT                   /score=100.00
19  FT                   /note="putative transposase for insertion sequence
20  FT                   element IS200 with hmmer score = 902.47"
21  FT                   /product="hypothetical transposase for IS200"
22  FT   CDS             507385..507843
23  FT                   /systematic_id="pred3"
24  FT                   /colour=255 2 2
25  FT                   /score=99.03
26  FT                   /note="putative transposase for insertion sequence
27  FT                   element IS200 with hmmer score = 893.69"
28  FT                   /product="hypothetical transposase for IS200"
29  FT   CDS             657246..657704
30  FT                   /systematic_id="pred4"
31  FT                   /colour=255 0 0
32  FT                   /score=99.92
33  FT                   /note="putative transposase for insertion sequence
34  FT                   element IS200 with hmmer score = 901.74"
35  FT                   /product="hypothetical transposase for IS200"
```

Figure 1.15: *Final Transformed CT18_IS200.tab*

1.2.5 Artemis .tab load - Inspection of results

In this step we are going to load in Artemis tool the previously edited CT18_IS200.tab. Initially, the first thing to do is to open Artemis and load the S. Typhi CT18 genome and its annotations (St.art) as we did at the first step of this report. At the same time when the window with S. Typhi's CT18 genome appears we go to the Artemis menu and we select "File" and in File's submenu we select "Read an entry..." and we load the CT18_IS200.tab as a new entry in Artemis window. If we now un-check the entry of St.art and we only keep checked the CT18_IS200.tab entry, then in the window there will appear only the CDSs of the predictions listed in the entry file. Figure 1.2.5 shows a screenshot of this entry's CDSs view.

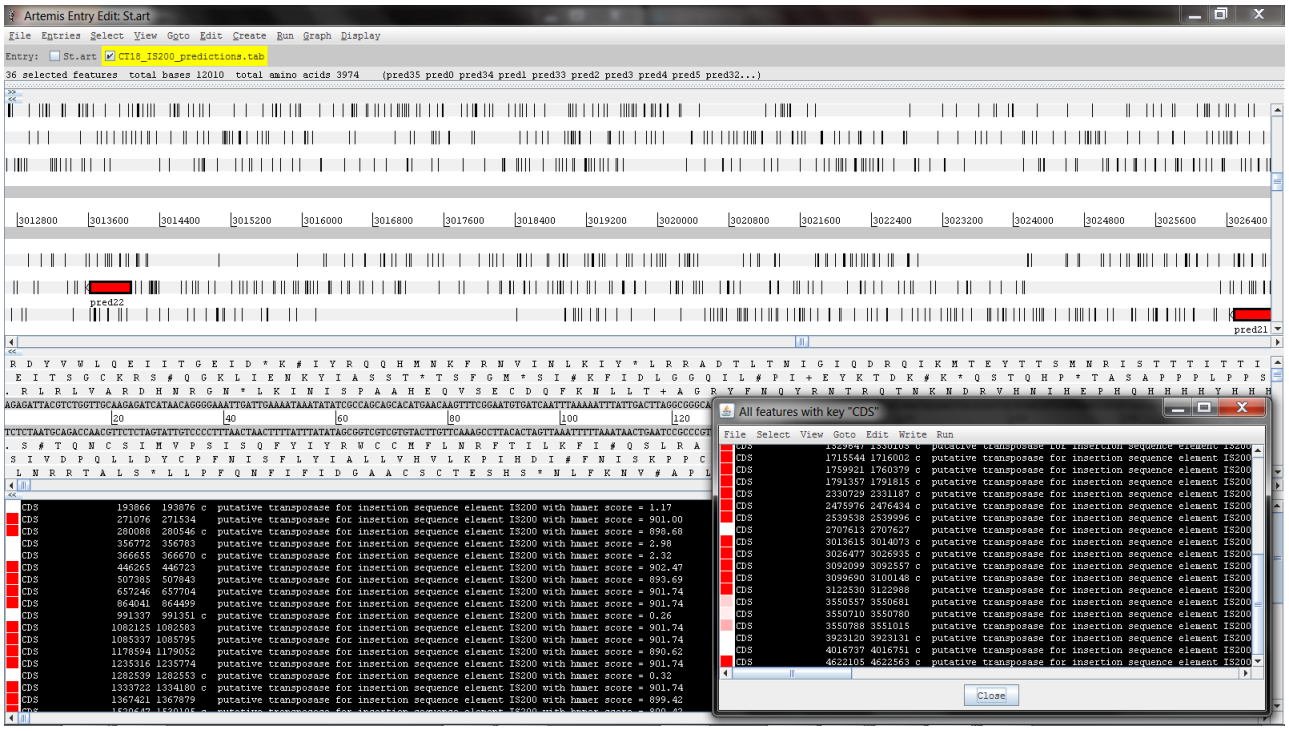


Figure 1.16: *CT18_IS200.tab entry's CDSs view*

Now that we confirmed that the entry's CDSs are shown in the proper way with a white-red color, the next thing to do for inspecting the prediction results is to check for overlaps with the already existed annotation for transposases IS200. To do that, we select (Figure 2.2) all the CDS features from the S. Typhi genome that produce IS200 transposase and we copy them in CT18_IS200.tab to check if they overlap.

As is was predicted, it is true that all the CDSs fully overlap with all the predictions. We also saved the new predictions with the part of CDSs S. Typhi's genome output in file CT18_IS200_overlap. Figure 1.2.5 shows the bottom Artemis window that corroborates our assertion and Figures 2.2 and 2.2 show the main Artemis window showing the same CDS,

one time as it is provided through S. Typhi's genome and once as it is predicted from the HMM results.

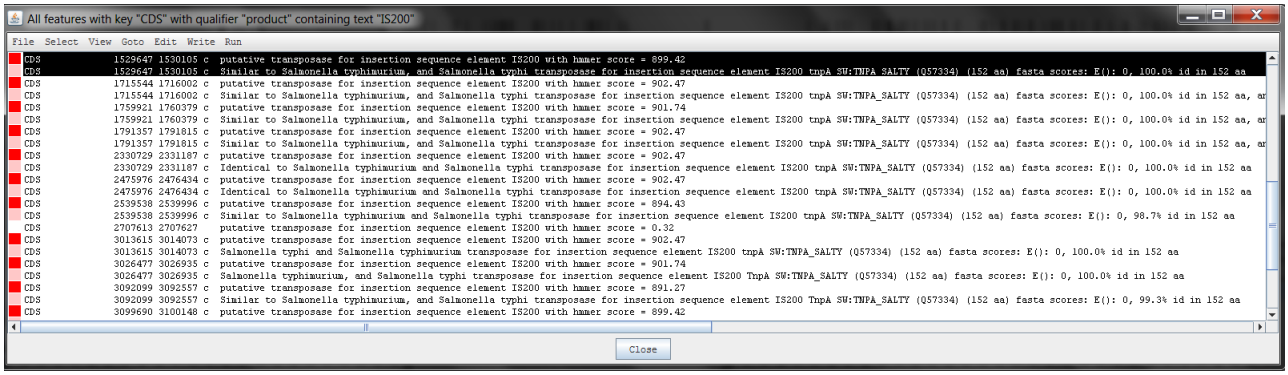


Figure 1.17: *St.dna and CT18_IS200.tab CDSs overlaping*

So, to conclude, there exists a full overlap between the CDSs of the S. Typhi CT18 that produce IS200 transposase and the predictions that we took from the Hidden Markov Models.

1.2.6 Optimal score threshold - statistically important predictions

In this new subsection we are assigned to determine with a dynamic way the optimal score threshold over which we will keep the predictions that we consider statistically important.

In order to do this task, we made a short discussion with our colleagues and we concluded that a good way of showing which predictions are statistically important is by giving a small diagram like this of Figure 1.2.5.

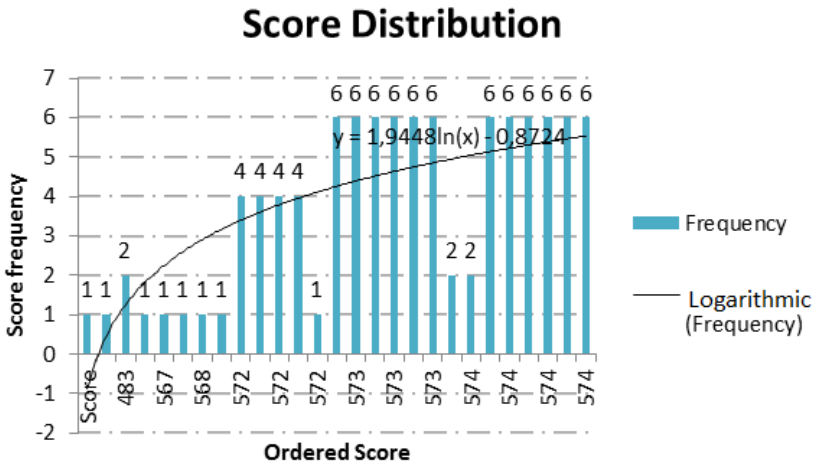


Figure 1.18: *Score Distribution of Statistically Important Predictions*

The data that we used to end up in the above figure were taken from a .csv file that we had taken as an output HMM profile from the UGENE tool (CT18_IS200_predictions.csv). We opened the .csv file in Microsoft Excel Office Tool and we configured the coma separated data to be set in columns. Then we selected the column that holds the scores for every prediction and we sorted the whole data according to this column. Having done that, we could export the frequency with which each result score was predicted, and given this information we structured a bar graph with the horizontal axis showing the score range and the vertical axis showing the frequency with which each score was predicted.

What is more, we considered giving the most close type of function, and as it is shown, a logarithmic function would be the most proper. The file that includes all the above is saved as CT18_IS200_predictions - plot.xlsx. It is worth mentioning that HMMER tool also has a logarithmic latency in finding one's matching prediction score.

Furthermore, in order to find an optimal score threshold to exclude all the unimportant predictions to get only the statistically more important predictions, as it is also shown from Figure 1.2.6, a good threshold would be to cutoff all the predictions that have a less than two times frequency and simultaneously a more than 483 score.

So, as a result, when asked from the assignment, to implement the score threshold (through Artemis tool) we can disarrear/appear predictions with a score less than the cutoff

that we wrote earlier. We start Artemis again and we load the CT18_IS200_predictions.tab entry that contains all the predictions features of S. Typhi CT18 which all include the score attribute that takes its values normalized in 0-100 range, after our perl application execution.

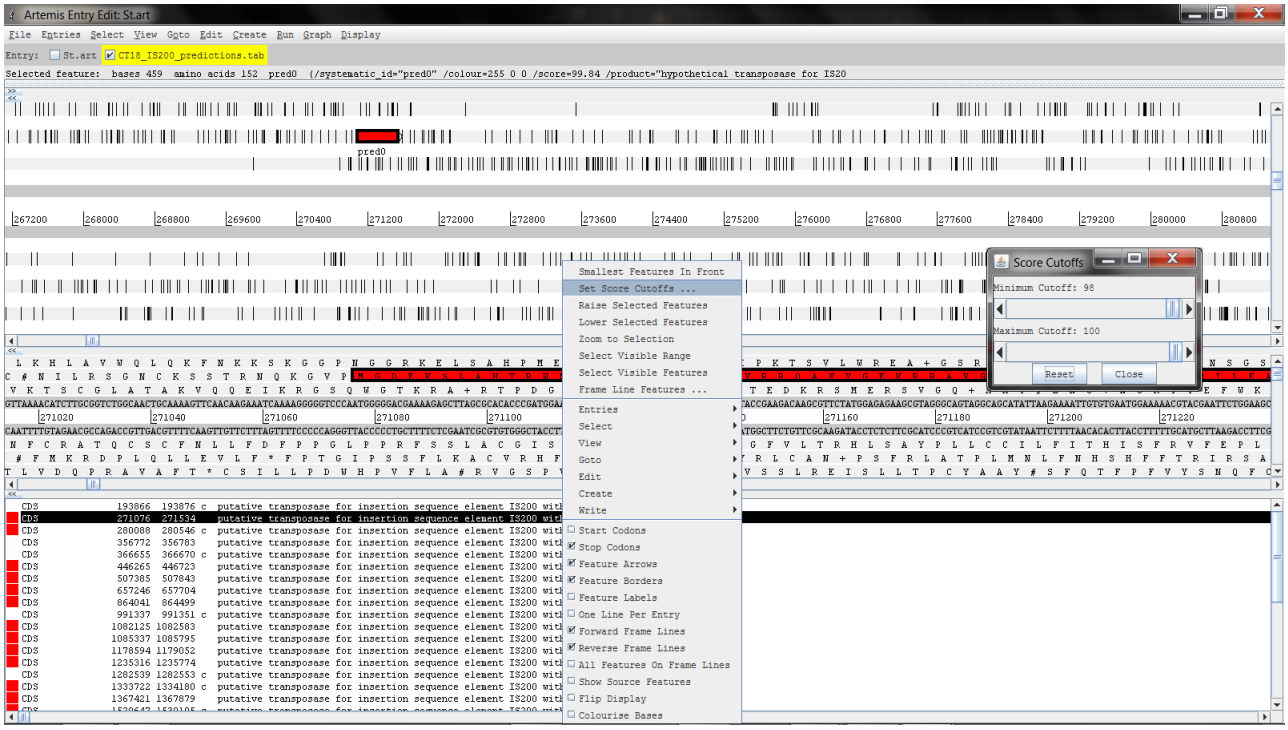


Figure 1.19: *Selection of Score Cutoff*

Figure 1.2.6 shows how to select a score cutoff with Artemis[32], and specifically for our occasion we defined a Minimum Cutoff threshold to 98 so that we keep all the 25 statistically important predictions for the CDSs that produce IS200 transposase as we examined earlier (Figure 1.2.6).

[32] ftp://ftp.sanger.ac.uk/pub4/resources/software/artemis/archive/v3/manual/x680.htm

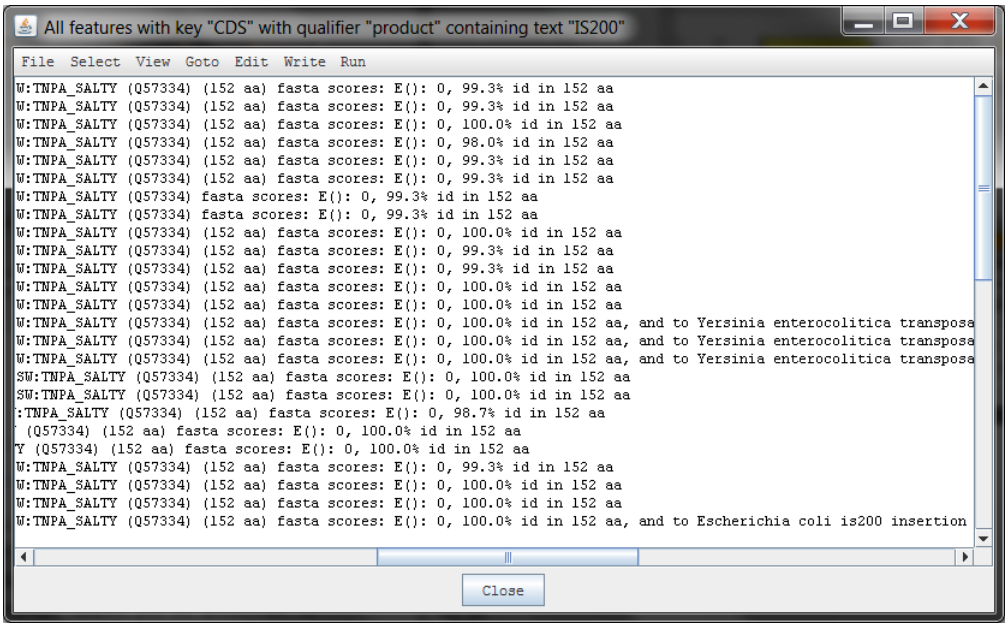


Figure 1.20: *Statistically Important Predictions given from CT18_IS200_predictions.tab. These predictions occur after having set the minimum threshold to 98, and as it is shown the CDSs predicted, still remain 25 at number.*

Finally, to conclude for this step of the assignment's task, it is worthy saying that it was expected from the beginning that HMMER search results would include at least as many statistically important sequences (25 sequences) as given in MSA tool that afterwards was used to create the HMM profile. As a more sophisticated conclusion, having come up with the above results, we can argue that because of the exact overlapping of S. Typhi's CT18 sequences with the predicted HMM sequences, all the CDS transposase IS200 that are statistically important, they all have an annotates CDS.

1.3 Salmonella Typhi TY2

In this section one can find, step by step, the whole procedure followed so as to use the produced hidden markov model (abbr. hmm) profile, which was built for the CT18 IS200 family sequences, in order to search this model in S. Typhi's TY2[33] [5] (Figure 1.3) genome and place the output with all predictions is written in a .tab file that can be executed by Artemis keeping all the information about the predicted positions in the genome and the relevant score.

KEGG **GENOME: Salmonella enterica subsp. enterica serovar Typhi Ty2** Help

Entry	T00121 Complete Genome
Name	stt, SALTI, 209261
Definition	Salmonella enterica subsp. enterica serovar Typhi Ty2
Annotation	manual Show organism
Taxonomy	TAX:209261
Lineage	Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Salmonella Taxonomy
Data source	RefSeq (Project:57973)
Original DB	Wisconsin
Keywords	Human pathogen
Disease	H00111 Typhoid fever
Comment	Isolated before the emergence of drug resistance in the 1970s.
Chromosome	Circular
Sequence	RS:NC_004631
Length	4791961
Statistics	Number of nucleotides: 4791961 Number of protein genes: 4370 Number of RNA genes: 109
Reference	PMID:12644504
Authors	Deng W, et al.
Title	Comparative genomics of Salmonella enterica serovar Typhi strains Ty2 and CT18.
Journal	J Bacteriol 185:2330-7 (2003)

Figure 1.21: *S. Typhi TY2 Genome Information*[34].

This time, though, the new TY2 .tab file results will be compared with those of CT18. Using the similarities and the differences between CT18 and TY2 genomes, the results can be visualized by Artemis Comparison Tool[35] (abbr. ACT). Phylogenetic trees will be created for all

[33] http://www.ncbi.nlm.nih.gov/bioproject/57973
[34] http://www.genome.jp/dbget-bin/www_bget?genome:T00121
[35] http://www.sanger.ac.uk/resources/software/act/

the DNA sequences that produce transposase IS200, both for CT18 and TY2 S. Typhi types and a small discussion will be provided for whether or not the phylogenetic trees created separate the IS200 transposase with a "phylogenetic separation" way.

Moreover, the results will also be analysed by using the S.Typhi's IS200 CT18 CDSs and the CT18 hmm predictions as an input for the BLAST[36] engine tool. Then, the output classification, shows the region in genome where different 'taxa' appear, which have different important blast predictions.

Subsequently, a phylogenetic tree for transposase IS200 will be created, by using all the statistically significant DNA sequences returned from BLAST. The way that phylogenetic trees are separated is caused by the different phylogenetic 'taxa' that appeared in blast predictions.

Having completed this step too, a consensus tree is created and comparisons in the different DNA topologies with all the phylogenetic trees are made. After that, a distribution of all the tree distances is created and a commentary for these results is provided.

Finally, having created all the phylogenetic trees and the consensus tree, a supertree from the reticulate network is created to add a different prospect in the phylogenetic history of the 'taxa' that where previously compared.

1.3.1 S. Typhi's CT18 HMM profile searched in S. Typhi's TY2 genome

To start with, in this subsection, we will use the produced hmm profile of S. Typhi's CT18 IS200 family sequences, in order to search this model in S. Typhi's TY2 genome and place the output result with all predictions in a new .tab file that afterwards will be executed by Artemis.

So, as we previously described in subsections 1.2.3 and 1.2.4 of this report, we will use HMMER v3.0 tool either from the command line, either as a plugin of UGENE tool, but first we have to find the S. Typhi TY2 genome sequence to search the CT18's hmm profile in it. Figure 2.3 shows where we found and downloaded the S. Typhi's TY2 genome[37]. After getting the S. Typhi TY2 DNA in a GenBank format we loaded it in UGENE (Figure 2.3) and we used the HMMER3 plugin tool to search the hmm profile using the CT18_IS200.hmm file (Figure 2.3). The output predictions results where stored to (Figure 2.3) and the output format of exporting them was GenBank (TY2_IS200_predictions.gb) so that to use the same web application to get the .tab file for Artemis and if needed to use again our perl implementation to change the exported .tab file. As we indicated in the related previous sections, the reason of exporting the predictions in a .gb output file is because it has a lot in common with .tab files. Figure 1.3.1 shows a part of the output TY2_IS200_predictions.gb file, while Figure 1.3.1 shows the final .tab file, after the editing done from the perl program we made.

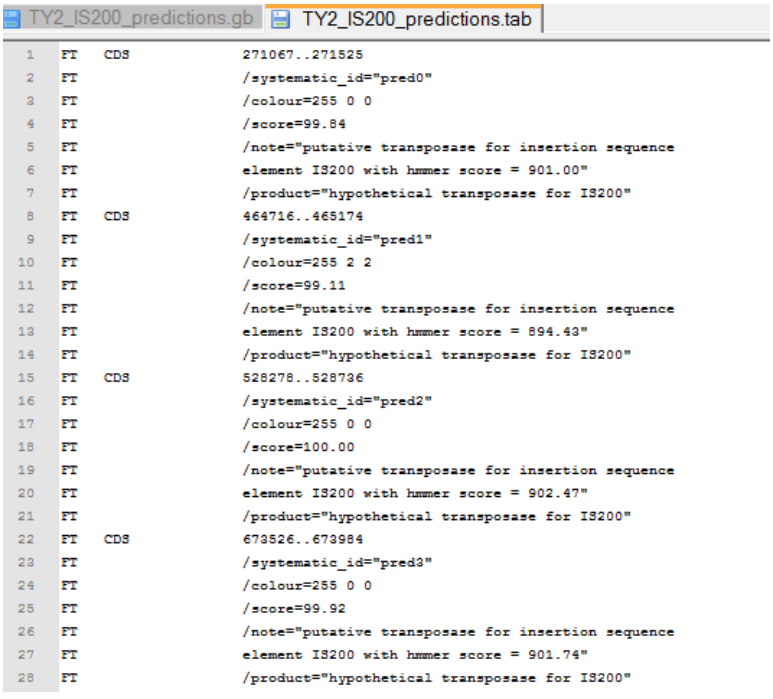
[36] http://blast.ncbi.nlm.nih.gov/

[37] http://www.ncbi.nlm.nih.gov/nuccore/NC_004631.1?report=gbwithparts&log$=seqview

TY2_IS200_predictions.gb

```
LOCUS       Annotations                                          18-FEB-2013
UNIMARK     Annotations
FEATURES             Location/Qualifiers
     CDS             186496..186954
                     /HMM model="CT18_IS200"
                     /Independent e-value="4.12808e-175"
                     /Conditional e-value="4.12808e-175"
                     /Score="569.768188"
                     /Bias="4.87234e+00"
                     /Accuracy per residue="9.96901e-01"
                     /HMM region="1...459"
                     /Envelope of domain location="186496...186954"
     CDS             complement(1255287..1255745)
                     /HMM model="CT18_IS200"
                     /Independent e-value="1.46474e-149"
                     /Conditional e-value="1.46474e-149"
                     /Score="485.538483"
                     /Bias="4.14697e+00"
                     /Accuracy per residue="9.95683e-01"
                     /HMM region="1...459"
                     /Envelope of domain location="1255286...1255744"
     CDS             complement(1683479..1683937)
                     /HMM model="CT18_IS200"
                     /Independent e-value="2.09631e-176"
                     /Conditional e-value="2.09631e-176"
                     /Score="574.035034"
                     /Bias="4.95132e+00"
                     /Accuracy per residue="9.96907e-01"
                     /HMM region="1...459"
                     /Envelope of domain location="1683478...1683936"
     CDS             1706319..1706777
                     /HMM model="CT18_IS200"
                     /Independent e-value="6.61859e-176"
                     /Conditional e-value="6.61859e-176"
                     /Score="572.388977"
                     /Bias="4.83847e+00"
                     /Accuracy per residue="9.96908e-01"
                     /HMM region="1...459"
                     /Envelope of domain location="1706319...1706777"
```

Figure 1.22: *Sample of IS200 predictions output from TY2 hmm search in GenBank format*



TY2_IS200_predictions.gb | TY2_IS200_predictions.tab

```
FT   CDS             271067..271525
FT                   /systematic_id="pred0"
FT                   /colour=255 0 0
FT                   /score=99.84
FT                   /note="putative transposase for insertion sequence
FT                   element IS200 with hmmer score = 901.00"
FT                   /product="hypothetical transposase for IS200"
FT   CDS             464716..465174
FT                   /systematic_id="pred1"
FT                   /colour=255 2 2
FT                   /score=99.11
FT                   /note="putative transposase for insertion sequence
FT                   element IS200 with hmmer score = 894.43"
FT                   /product="hypothetical transposase for IS200"
FT   CDS             528278..528736
FT                   /systematic_id="pred2"
FT                   /colour=255 0 0
FT                   /score=100.00
FT                   /note="putative transposase for insertion sequence
FT                   element IS200 with hmmer score = 902.47"
FT                   /product="hypothetical transposase for IS200"
FT   CDS             673526..673984
FT                   /systematic_id="pred3"
FT                   /colour=255 0 0
FT                   /score=99.92
FT                   /note="putative transposase for insertion sequence
FT                   element IS200 with hmmer score = 901.74"
FT                   /product="hypothetical transposase for IS200"
```

Figure 1.23: *Sample of IS200 predictions output from TY2 hmm search in Artemis .tab format*

Now that the TY2_IS200_predictions.tab is ready, we load it in Artemis tool and to investigate if there exist overlaps with the existing annotation and if there exist in which percent they overlap with the existing CDSs. So, following the same steps as in subsection 1.2.6 we select the CDSs that produce IS200 transposase and the predictions (Figure 2.3) and set, this time again, a score cutoff to 98, as shown in Figure 1.3.1, because with this threshold we can find all the statistically important predictions which it worths saying that they all overlapped with the 25 CDS annotations for IS200 transposase.

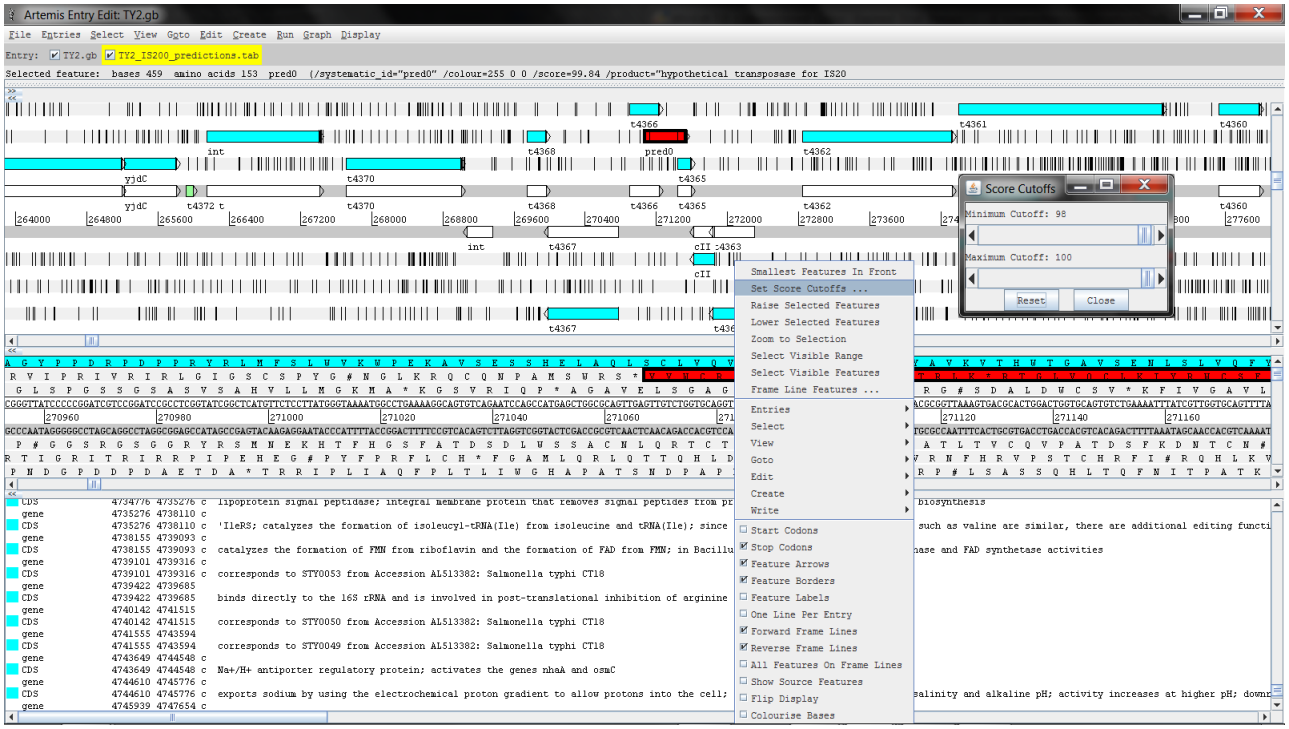


Figure 1.24: *Minimum threshold cutoff of 98 - Statistically Important Predictions of S. Typhi TY2 for IS200 transposase*

As a result to the above, we can conclude that our hidden markov model works properly in both genomes, as it was expected considering that these two kinds of Salmonella are close enough to have a full CDSs overlapping by the predictions of Typhi's CT18, as shown in Figure 1.3.1.

Figure 1.25: *S. Typhi's TY2 features overlapping with S. Typhi's CT18 predictions. On the top of this figure it is indicated the way to select the overlapping features. At the bottom of Artemis window there are shown all the overlapping features of TY2 with those of CT18.*

1.3.2 S. Typhi CT18 and TY2 Genome Comparison - WebACT

In addition to what we found in the previous subsection, in this step we are going to compare the genomes of the two S. Typhi family kinds. To visualize the similarities and the differences of TY and CT18 genomes we used the Artemis Comparison Tool[38]. Taking as a reference the [Baker & Dougan, 2007] article[39] [6], we opened the WebACT[40] tool for comparing the two bacteria genome sequences. The in the "Pre-computed" tab we selected to compare two sequences and after that we selected the names of these two sequences, as shown in Figure 1.3.2. We, then, set the same range for both sequences and that each sequence would compare its full genome to the others (Figure 2.3). Afterwards, we set a cutoff of 0.01 e-value (Figure 2.3) and we started ACT.

[38]http://www.webact.org/WebACT/home
[39]http://cid.oxfordjournals.org/content/45/Supplement_1/S29.full
[40]http://www.webact.org/WebACT/home

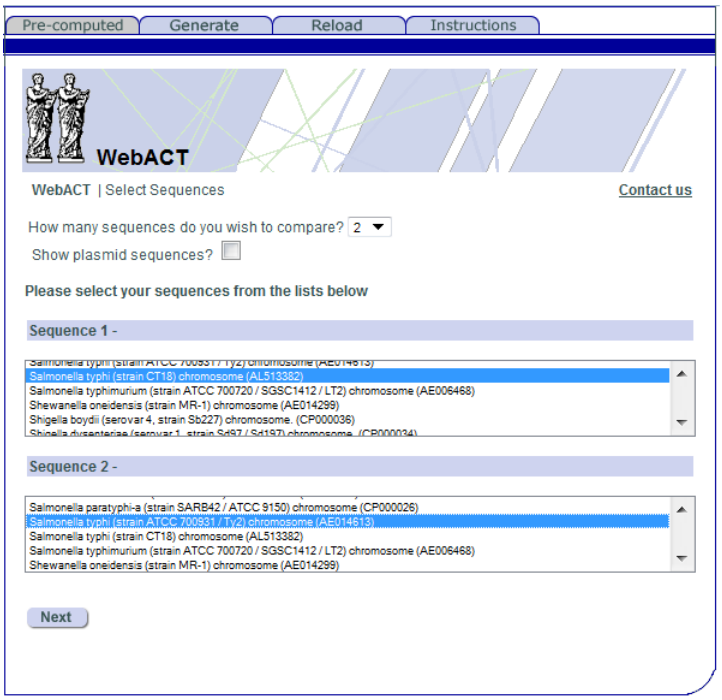


Figure 1.26: *WebACT S. Typhi's CT18 and TY2 selection*[41]

With the WebACT tool we can view the matching genome regions between these two genomes. As shown in Figure 1.3.2 we can see that both the CT18 and TY2 genome sequences match almost precisely everywhere in their DNA sequence. In this figure there are three types of colored strains, and as it can be easily presumed the red colored strains show the forward matches, the blue colored show the reverse matches, and the yellow colored show specific matches of sequences that have a specific feature, which in our case, is that yellow strains indicate the CDSs that produce IS200 transposase. In addition to that, with the WebACT tool we managed to visualize and count the sequences of both genomes, which have except for the 25 CDSs that produce IS200 transposase, the 17 CDSs, out of the 25, in common, i.e. these two S. Typhi's genomes are close enough, which means that they share in common some of the CDSs that produce IS200. Figures 2.3, 2.3, 2.3, 2.3 and 2.3.

[41] http://www.webact.org/WebACT/prebuilt

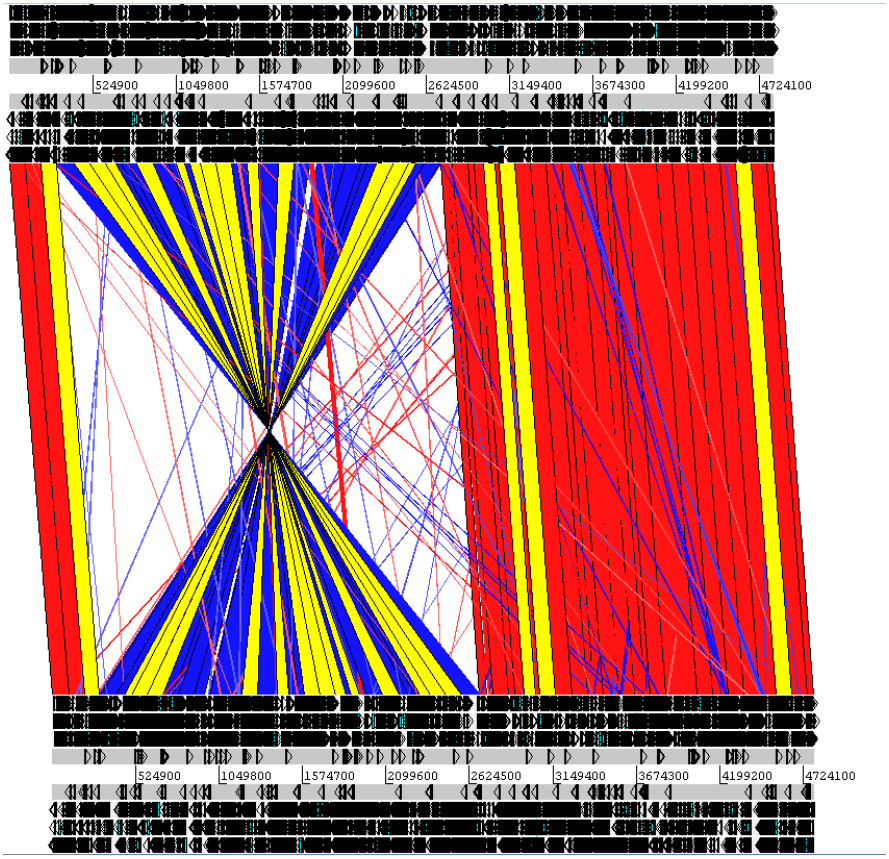


Figure 1.27: *WebACT S. Typhi's CT18 and TY2 genome comparison. The red colored strains show the forward matches, the blue colored strains show the reverse matches, and the yellow colored strains show specific matches of sequences that have a CDS feature that produce IS200 transposase.*

1.3.3 S. Typhi CT18 and TY2 IS200 Phylogenetic Tree

Having finished with the WebACT genome comparison in this step it is time to create the phylogenetic trees for all the DNA sequences that produce transposase IS200, both for S. Typhi CT18 and S. Typhi TY2. Afterwards, depending on the results that will come of the phylogenetic tree analysis, we will discuss the reason why or not, the phylogenetic trees created separate the IS200 transposase with a "phylogenetic separation" way.

So, to begin with, for the phylogenetic tree construction of both CT18 and TY2 kinds of Salmonella Typhi, we used the PHYLIP[42] tool, implemented by the University of Washington. In Bioinformatics (ITMB, MD, 2012-13) class we were taught to use the PHYLIP tool for constructing phylogenetic trees by using the console. In this subsection we will use, though, PHYLIP via a plugin version that UGENE has got. In this way we will not get confused of

[42]http://evolution.genetics.washington.edu/phylip.html

the many intermediate output files that are necessary if we use PHYLIP through the console window. Having as a reference the Unipro UGENE manual[43] [Chapters 7, 9].

Yet, some things to be done before construction the phylogenetic trees is to produce with Artemis tool a FASTA file of the S.Typhi's TY2 CDSs for IS200 transposase production, and after that to build via ClustalW2 a multiple sequence alignment, to be used by the PHYLIP tool. This time though, instead of ClustalW2 web application tool we will use it via UGENE, to test another way of doing the same thing.

Figures 2.3, 2.3 and 2.3 show how we extracted with Artemis tool the coding DNA sequences that we are interested in and stored them in a FASTA format. Figures 2.3, 2.3 show how we produced from the TY2_IS200.fasta file the TY2_IS200.msf file by using ClustalW2 plugin of UGENE, and Figure 2.3 shows the merged MSA file with both CT18 and TY2 sequence alignments, which we will use for our phylogenetic trees.

Now that we figured out how the S. Typhi's TY2 MSA came of, its time to build with PHYLIP the phylogenetic trees. Building a phylogenetic tree with UGENE's plugin tool for PHYLIP is very easy. The only needed is to have the previously referred .fasta and .msf files for both CT18 and TY2 kinds of S. Typhi, loaded already in the project. Then, to build a tree from the alignment file one can either press the "Build Tree" button on the toolbar, either select the "Tree" and "Build Tree" item in the alignment context menu or the "Actions", "Tree", "Build Tree" item in the main menu.

To make a short analysis of the phylogenetic tree options supported by PHYLIP's latest version (v3.69), there are two methods for building phylogenetic trees supported: a) the PHYLIP Neighbour-Joining method, and b)the MrBayes external tool. In our Bioinformatics class lessons we used the first one, so we will do the same now. Moreover we can parameterize our building method with some more specific options. At first we can set the distance matrix model, which computes a distance matrix for a nucleotide multiple sequence alignment with values: F84, Kimura, Jukes-Cantor or LogDet. All the four models have been taught in presentations we did in class. The Gamma distributed rates across sites option, specifies to take into account unequal rates of change at different sites. It is also assumed that the distribution of the rates follows the Gamma distribution. Moreover, the transition/transversion ratio option is the expected ratio of transitions to transversions.

Having enabled the Bootstrapping and Consensus Trees group of option, one can set the number of replicate date sets, the random number seed, which is automatically generated and if changed manually results of different runs (of a tree building) become reproducible. Furthermore, the consensus type option specifies the method to build the consensus tree. If set as "Strict" then the set of species must appear in all input trees to be included in the strict consensus tree. Else if set as "Majority Rule (extended)" then any set of species that appears in more than 50% of the trees is included. The program then considers the other sets of species in order of the frequency with which they have appeared, adding to the consensus tree any which are compatible with it until the tree is fully resolved. This option selection is the default setting and this is what will be shown later in Figure 1.3.3. If consensus type is set

[43] http://ugene.unipro.ru/downloads/UniproUGENE_UserManual.pdf

as "M1", then the tree will include in the consensus tree any sets of species that occur among the input trees more than a specified fraction of the time (see the Fraction parameter below). In this part, it worths saying that the "Strict" consensus and the "Majority Rule" consensus are extreme cases of the Ml consensus, being for fractions of 1 and 0.5 respectively. Finally if "Majority Rule" is selected, then a set of species will be included in the consensus tree if it is present in more than half of the input trees.

Having explained all the modularity that can be selected to get an output phylogenetic tree, in Figure 1.3.3 we present our selection.

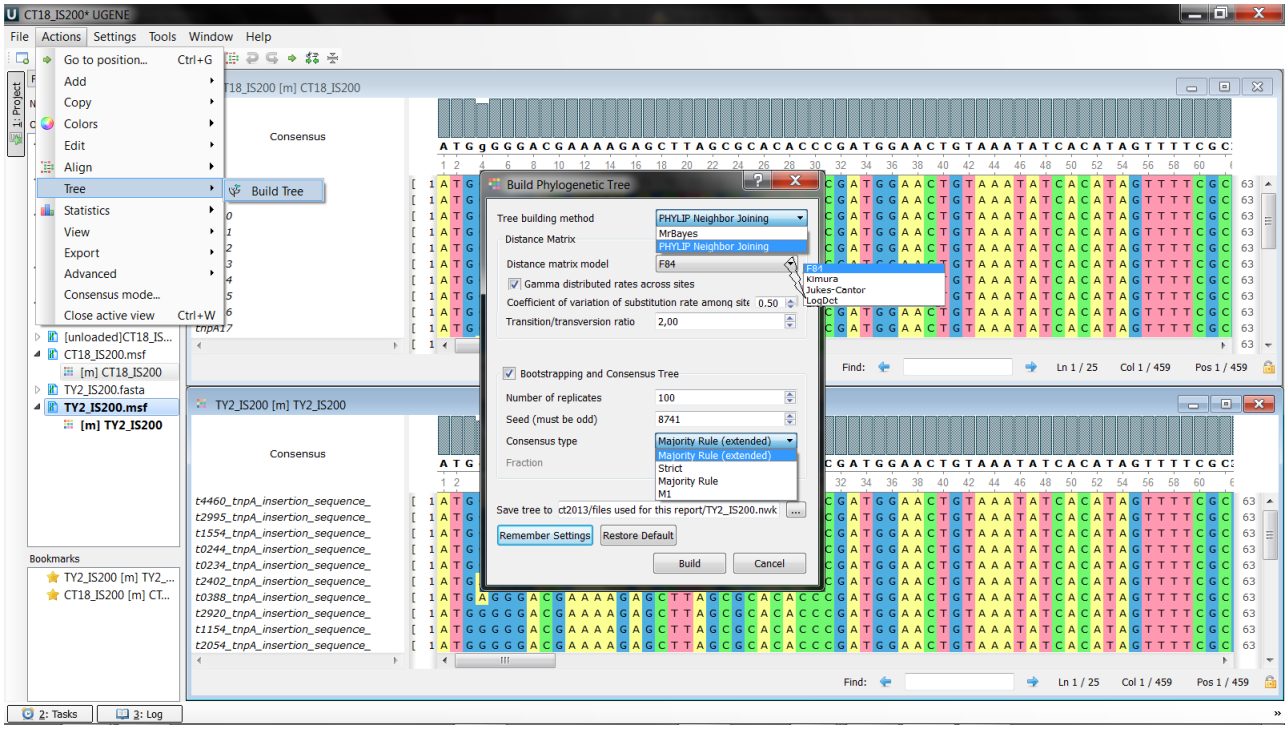


Figure 1.28: *CT18 and TY2 options selected for the Phylogenetic Tree*

The output of the above selection is shown in the Figures 1.3.3 and 1.3.3 below. Figure 2.3 and 2.3 shows also the phylogenetic tree of S. Typhi CT18 straightly and Figure 2.3 and 2.3 shows the phylogenetic tree of S. Typhi TY2. Figure 2.3 shows how to configure a tree view and Figures 2.3-2.3 show the merged phylogenetic tree again, but with some other options selected.

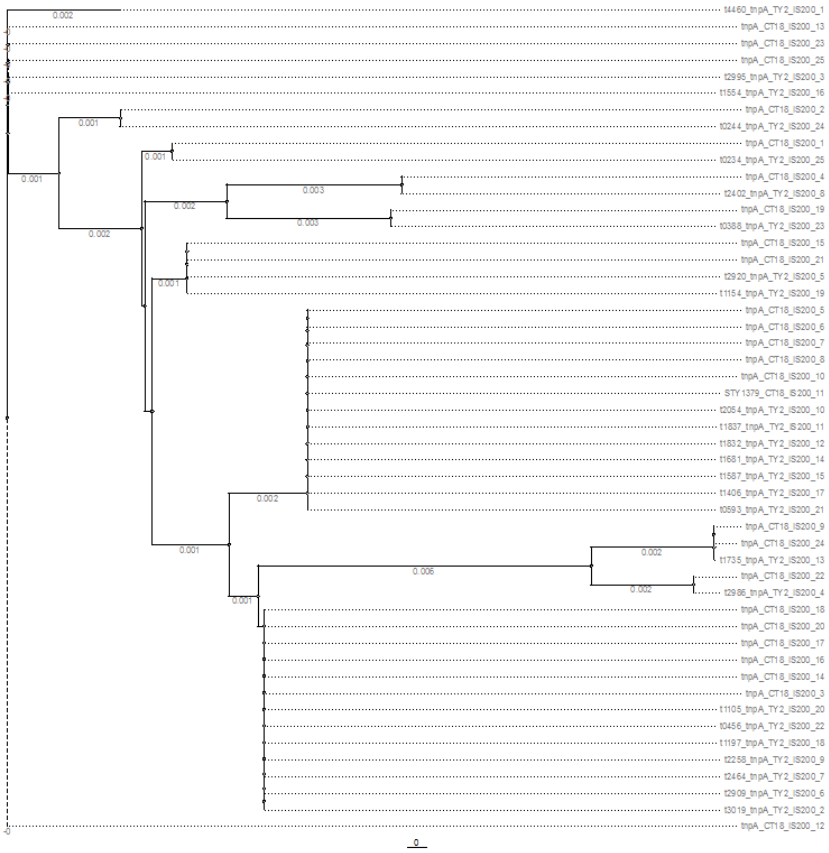


Figure 1.29: *CT18 and TY2 Phylogenetic Tree. Figure shows the phylogenetic Tree in a Rectangular and Phylogram view, with tree settings set to minimum width and height.*

To conclude with this step, as we can infer from the output results shown in figures of the phylogenetic trees, the two S. Typhi's kinds referring to IS200 transposase product do not segregate with each other in a "phylogenetic way". This conclusion can be inferred in an obvious reasoning way because except for the non-visible separation of the tree's CDSs for CT18 and TY2, also, in the previous steps we inferred that S. Typhi CT18 and S. Typhi TY2 share the 98% of their coding genome in IS200 transposase. So, we conclude and verify our initial affairs that every IS200 transposase of each of CT18 and TY2 is affiliated to one or more IS200 transposases of the other. Finally, another conclusion that we can come up with, is that because of this resemblance, is such cases, we can try thinking of checking the differentiation of IS200 transposase, regardless where the belong.

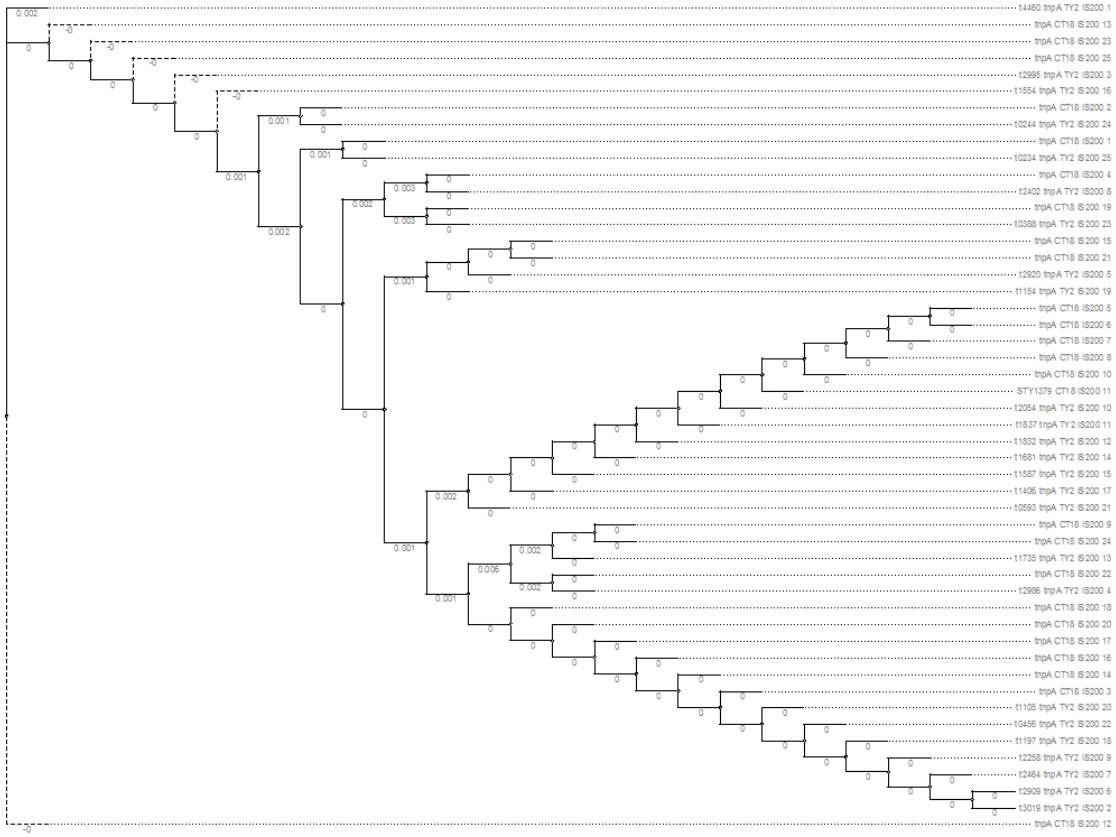


Figure 1.30: *CT18 and TY2 Phylogenetic Tree. Figure shows the phylogenetic Tree in a Rectangular and Cladogram view, with tree settings set to minimum width and height.*

1.3.4 S. Typhi CT18 IS200 BLAST Analysis for Predictions

Having found all the phylogenetic trees, in the next step the results will also be analysed by using the S.Typhi's IS200 CT18 CDSs and the CT18 hmm predictions as an input for the BLAST[44] engine tool. Then, the output classification, will show the region in genome where different 'taxa' appear, which have different important blast predictions.

So, we start by cutting off the less important predictions of S. Typhi CT18 as we previously mentioned that exist, and by keeping the 25 predictions that are identical with the S. Typhi CT18 CDSs. Then, we are going to 'feed' these sequences in a FASTA format, one by one, in Basic Local Alignment Sequence Tool (abbr. BLAST), as shown in Figure 1.3.4. Another way to do the alignment is by using Artemis Tool or UGENE tool. From now we will describe doing the local alignment through Artemis, in order to select the sequences more easily.

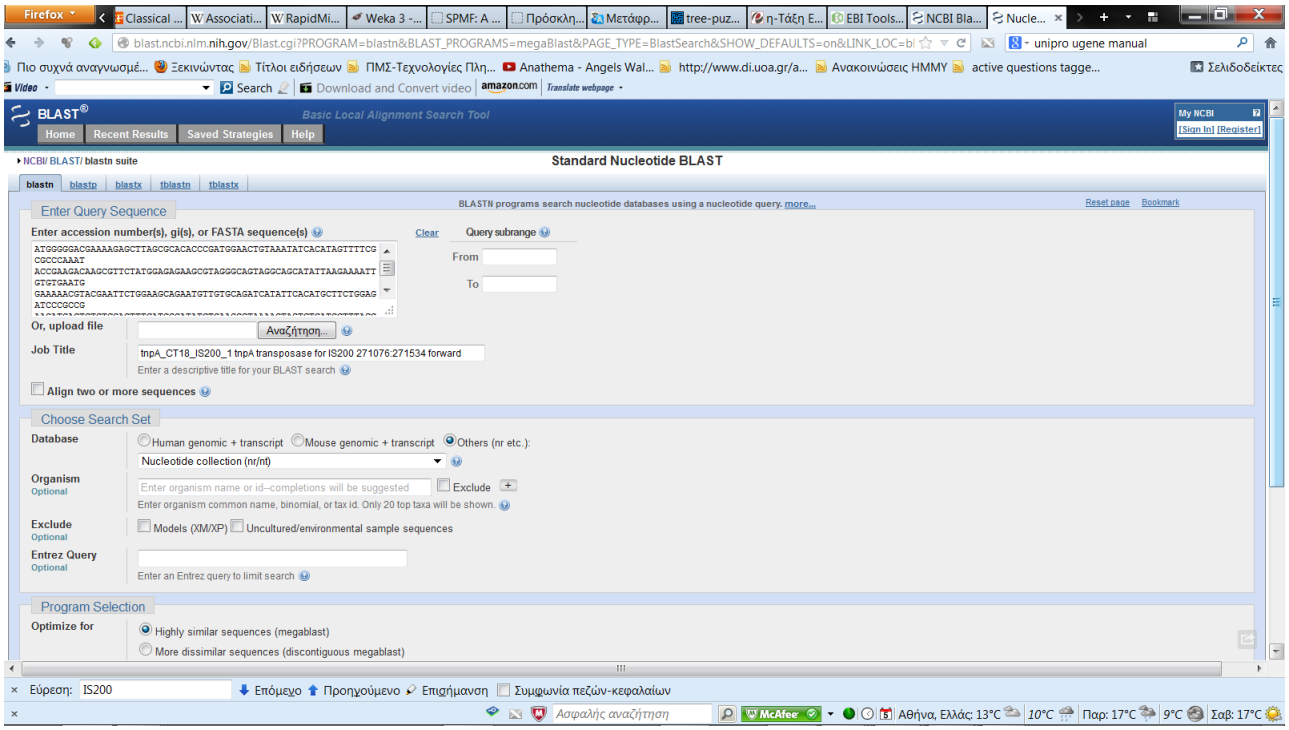


Figure 1.31: *Local Alignment using BLAST via NCBI web API configuration. The CT18 sequence fasta format that we used for querying.*

As indicated in previous sections we start Artemis Tool with the S. Typhi's CT18 annotated genome, and we select and view all CT18 features that are CDSs and produce IS200 transposase element. Then we selected from the elements' view each one element and we press in the window's menu bar the button "Run", then "NCBI Searches" and on the follow "blastn" (Figure 1.3.4). "Blastn" makes local sequence alignment for nucleonic sequences

[44] http://blast.ncbi.nlm.nih.gov/

which is what we like to do with the data that we posses. When we click on the button Artemis sends the sequence data to the BLAST API that we showed earlier and the result that appears contain all the regions of local similarity between sequences.

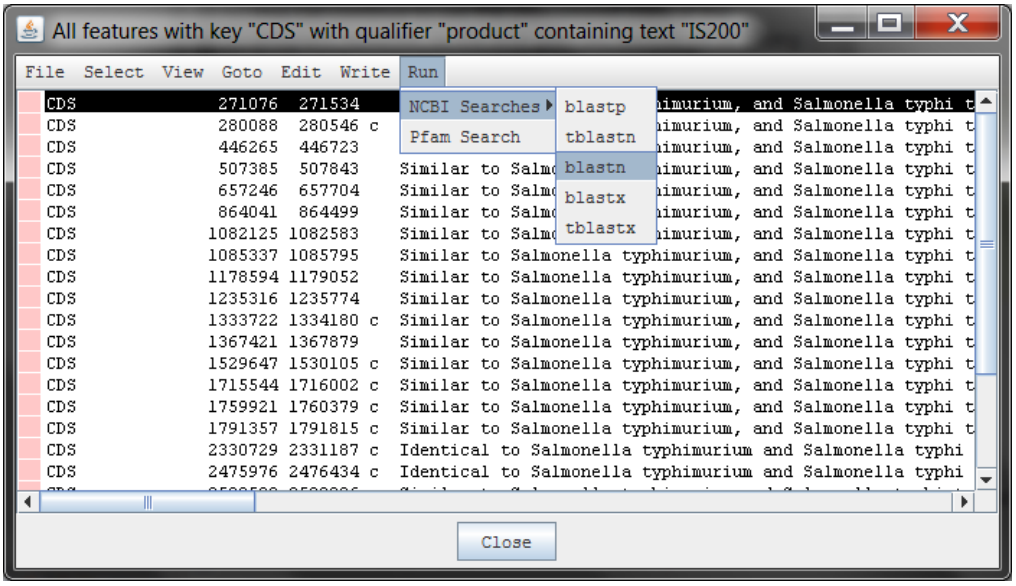


Figure 1.32: *CDS selections for BLAST using Artemis*

A quick definition of what does BLAST do is that this tool compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. What is more, BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

So, according to the above definition, using the output classification (Figure 2.3 in detail), we will now find the regions in genome where different 'taxa' appear, which have different important blast predictions. To do that, in the resulted browser window, we select all the sequences that are relevant with the IS200 transposase element product of the first CDS of CT18, and we save them in two fasta file formatted files, one for the complete sequence and one for the aligned sequences. For instance, after 'feeding' BLAST with the first CDS we got as output 21 different 'taxa' containing IS200 transposase, which contain statistically important blast predictions (Figure 1.3.4). Also, it is worth saying that sequences of all the above CDSs labelled with the same hmm score appear to have the same 'taxa' sequences.

Salmonella typhi mRNA for insertion sequence IS200 transposase (ORF1)	823	823	100%	0.0	99%	AJ301622.1
Salmonella enterica subsp. enterica serovar Heidelberg str. B182, complete genome	820	4484	100%	0.0	99%	CP003418.1
Salmonella enterica subsp. enterica serovar Typhimurium str. 798, complete genome	820	5731	100%	0.0	99%	CP003386.1
Salmonella enterica subsp. enterica serovar Typhimurium str. UK-1, complete genome	820	7380	100%	0.0	99%	CP002614.1
Salmonella enterica subsp. enterica serovar Typhimurium str. ST4/74, complete genome	820	5740	100%	0.0	99%	CP002487.1
Salmonella enterica subsp. enterica serovar Typhimurium str. T000240 DNA, complete genome	820	4920	100%	0.0	99%	AP011957.1
Salmonella enterica subsp. enterica serovar Typhimurium SL1344 complete genome	820	5740	100%	0.0	99%	FQ312003.1
Salmonella enterica subsp. enterica serovar Typhimurium str. 14028S, complete genome	820	8200	100%	0.0	99%	CP001363.1
Salmonella enterica subsp. enterica serovar Typhimurium str. D23580 complete genome	820	6560	100%	0.0	99%	FN424405.1
Salmonella enterica subsp. enterica serovar Typhimurium str. LT2, complete genome	820	4920	100%	0.0	99%	AE006468.1
Salmonella enterica subsp. enterica serovar Heidelberg str. SL476, complete genome	820	3664	100%	0.0	99%	CP001120.1
Salmonella enterica subsp. enterica serovar Typhimurium isolate M1855 putative permease (vbhL) gene, partial cds; mutated intergenic region TM781, complete sequence; and putative integral membrane protein (vbhM) gene, partial cds	820	820	100%	0.0	99%	DQ077525.1
Salmonella enterica serovar Typhi (Salmonella typhi) strain CT18, complete chromosome; segment 19/20	820	820	100%	0.0	99%	AL627283.1
Salmonella enterica serovar Typhi (Salmonella typhi) strain CT18, complete chromosome; segment 10/20	820	2454	100%	0.0	99%	AL627274.1
Salmonella typhimurium strain SARA17 insertion sequence IS200 unknown gene	820	820	100%	0.0	99%	L25848.1
Salmonella typhimurium strain LT2 NADP+-linked malic enzyme (maeB), partial cds; insertion element IS200 transposase, complete cds; eut operon, complete sequence; and unknown genes	820	820	100%	0.0	99%	AF093749.1
S.typhi gene encoding putative IS200 transposase, 1228bp	820	820	100%	0.0	99%	Y09989.1
Salmonella typhimurium insertion sequence IS200 transposase (tnpA) gene, complete cds	820	820	100%	0.0	99%	AF025380.1
Salmonella typhimurium putative IS200 transposase gene, complete cds	820	820	100%	0.0	99%	U44749.1
S.typhimurium fliB,C) genes and insertion sequence IS200	820	820	100%	0.0	99%	Z54217.1
Salmonella enterica subsp. enterica serovar Weltevreden str. 2007-60-3289-1 complete genome, contig 13	814	814	100%	0.0	99%	FR775200.1
Salmonella enterica subsp. enterica serovar Paratyphi A str. AKU_12601 complete genome, strain AKU_12601	814	4703	100%	0.0	99%	FM200053.1
Salmonella enterica subsp. enterica serovar Paratyphi A str. ATCC 9150, complete genome	814	4703	100%	0.0	99%	CP000026.1
S.typhi encoding putative IS200 transposase, 1575bp	814	814	100%	0.0	99%	Y09991.1
Salmonella enterica subsp. enterica serovar Weltevreden str. 2007-60-3289-1 complete genome, contig 53	809	809	100%	0.0	99%	FR775250.1
S.typhi IS200 insertion sequence (between gyrA and rcsC)	809	809	100%	0.0	99%	X91136.1
Salmonella enterica subsp. enterica serovar Weltevreden str. 2007-60-3289-1 complete genome, contig 34	805	805	100%	0.0	99%	FR775221.1
Salmonella enterica subsp. enterica serovar Paratyphi B str. SPB7, complete genome	805	5563	100%	0.0	99%	CP000886.1
S.typhi fliG gene, gene encoding putative IS200 transposase and gene encoding putative RNAseE-like protein	801	801	100%	0.0	99%	Y09990.1
S. typhimurium DNA for insertion sequence IS200	800	800	100%	0.0	99%	X56834.1
Salmonella enterica subsp. enterica serovar Weltevreden str. 2007-60-3289-1, whole genome shotgun sequence, plasmid pSW82 complete sequence, contig 58	792	792	100%	0.0	98%	FR775255.1

Figure 1.33: *BLAST selection of aligned sequences that are relevant IS200 transposase*

1.3.5 IS200 Transposase Phylogenetic Tree using BLAST Analysis Results

In this step we are assigned to create all the phylogenetic trees for transposase IS200, by using all the statistically significant DNA sequences returned from BLAST, for each of the 25 CDSs of S. Typhi CT18.

The steps that we will follow for the creation of the phylogenetic trees will be the same as in the previous subsections, at first creating from the fasta formatted predictions, the MSA, and then from the MSA producing the phylogenetic tree. We are going to perform MSA with ClustalW2 and then create and preview the phylogenetic trees with PHYLIP via UGENE's PLYLIP plugin.

As we can assume by the previous step what we are assigned is to make the phylogenetic trees by using the aligned sequences of the BLAST output and not only the complete. Nonetheless, we tried to make a multiple sequence alignment to the complete sequences just to satisfy our curiosity of what will happen. As expected, nothing happened; we left the browser open for hours so as to find an alignment if existed but this was futile. So, below we present two figures showing the BLAST aligned sequences of the first CT18 CDS's phylogenetic trees (Figures 1.3.5, 1.3.5).

Figure 1.34: *Phylogenetic phylogram tree for 1st CDS tnpA CT18 IS200 aligned BLAST*

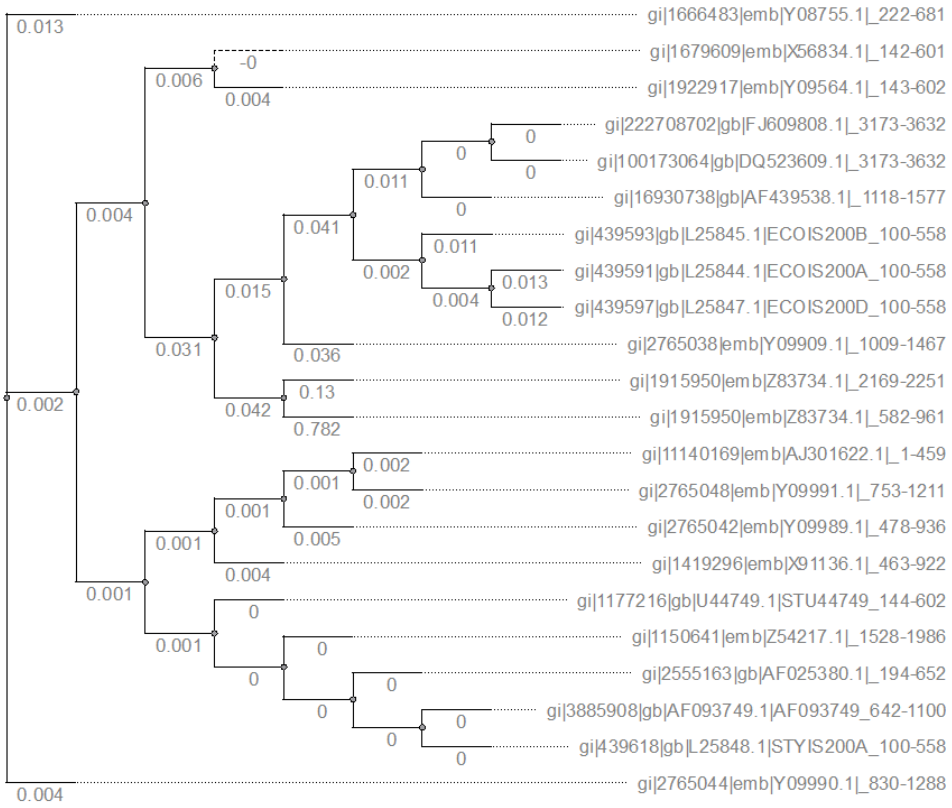


Figure 1.35: *Phylogenetic cladogram tree for 1st CDS tnpA CT18 IS200 aligned BLAST*

As expected, the way that phylogenetic trees are separated is not caused by the different phylogenetic 'taxa' that appeared in blast predictions, but in a different separation due to the alignments' similarities. So, we can conclude that the phylogenetic 'taxa' cannot be separated in a phylogenetic tree. A fact showing that is the minor difference in the branches of the phylogram tree view (shown also in the labels of the branches of both figures).

1.3.6 IS200 Transposase Consensus Tree and Tree Distances Distribution

Having completed the previous step of creating all the phylogenetic trees came from BLAST, we can now create a consensus tree and make comparisons in the different DNA topologies. After that, with the consensus tree data we will create a distribution of the tree distances and an outcome should be reached.

This time we will need to use PHYLIP main distribution program, to use its functional property of producing a consensus tree given all the phylogenetic trees of the previous step.

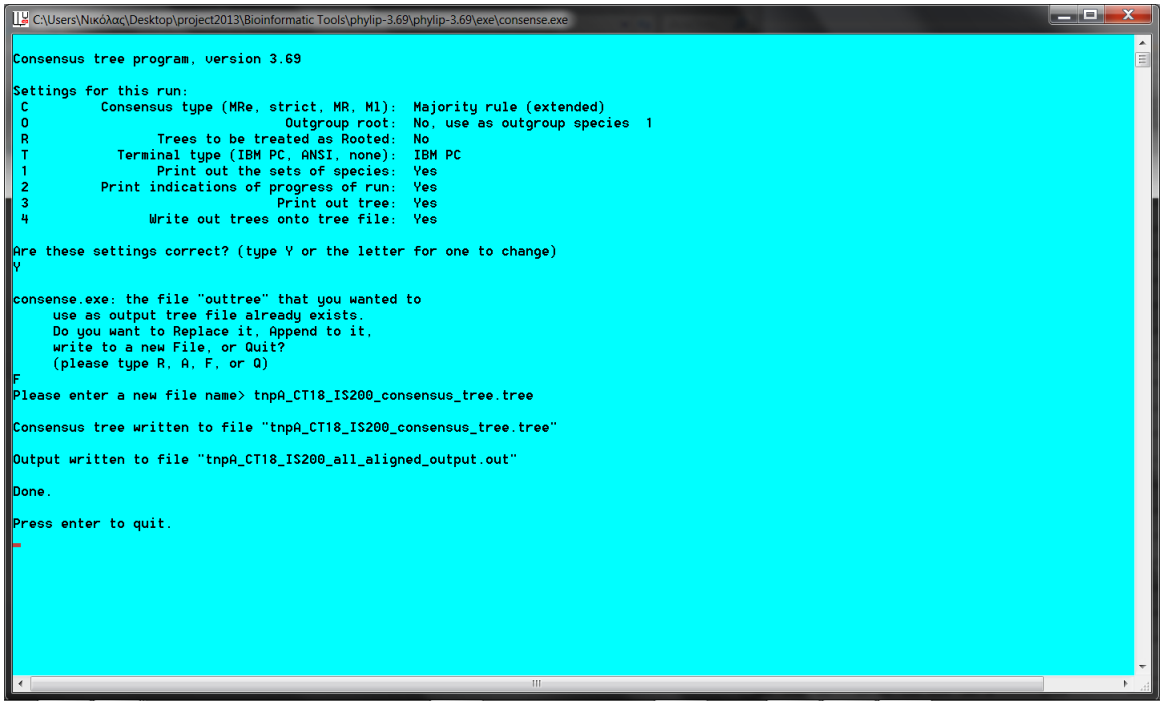


Figure 1.36: *PHYLIP configuration for the consensus tree*

So, we load in Phylip (Figure 1.3.6) the file *tnpA_CT18_IS200_all.nwk* whose format contains all the phylogenetic trees, given from BLAST significant sequences, separated by blank lines. The output result is opened and an 'ascii' created tree can be found in file *tnpA_CT18_IS200_all_output.out*. The consensus tree in a tree format so that we can visualize it, as indicated from the figure above, can be found in file *tnpA_CT18_IS200_consensus_tree.tree*. A visualization of the consensus tree, using UGENE is shown in Figure 1.3.6.

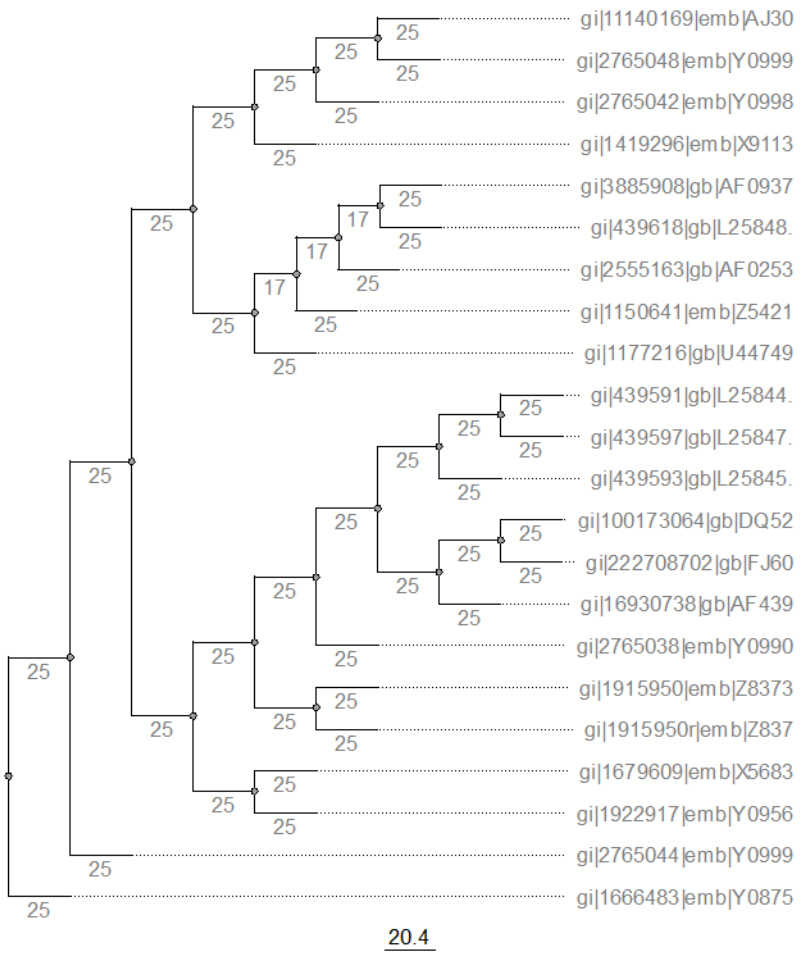


Figure 1.37: *Consensus Tree Visualization using UGENE*

Now that the consensus tree is ready, we will use the 'treedist' PHYLIP's executable tool whose properties compare two phylogenetic trees and find the differences in their topology. We are going to execute 'treedist' with two input files as indicated in the documentation. The initial file is the *tnpA_CT18_IS200_all_aligned.nwk* with all the phylogenetic trees that we used for the construction of the consensus tree, and as the second input file we are going to use the consensus tree itself (*tnpA_CT18_IS200_consensus_tree.tree*). We followed the customization settings proposed from the documentation, as shown in Figure 2.3. The final results is shown in Figure 1.3.6.

tnpA_CT18_IS200_all_aligned_output.out | tnpA_CT18_IS200_consensus_tree.tree | tnpA_CT18_IS200_treedist.out

```
Tree distance program, version 3.69

Symmetric differences between all pairs of trees
   from first and second tree files:

Trees 1 and 26:     0
Trees 2 and 26:     0
Trees 3 and 26:     6
Trees 4 and 26:     0
Trees 5 and 26:     0
Trees 6 and 26:     0
Trees 7 and 26:     0
Trees 8 and 26:     0
Trees 9 and 26:     0
Trees 10 and 26:     0
Trees 11 and 26:     0
Trees 12 and 26:     0
Trees 13 and 26:     0
Trees 14 and 26:     6
Trees 15 and 26:     6
Trees 16 and 26:     0
Trees 17 and 26:     6
Trees 18 and 26:     6
Trees 19 and 26:     6
Trees 20 and 26:     6
Trees 21 and 26:     6
Trees 22 and 26:     0
Trees 23 and 26:     0
Trees 24 and 26:     0
Trees 25 and 26:     0
Trees 26 and 1:     0
Trees 26 and 2:     0
Trees 26 and 3:     6
Trees 26 and 4:     0
Trees 26 and 5:     0
Trees 26 and 6:     0
Trees 26 and 7:     0
Trees 26 and 8:     0
Trees 26 and 9:     0
Trees 26 and 10:     0
Trees 26 and 11:     0
Trees 26 and 12:     0
Trees 26 and 13:     0
Trees 26 and 14:     6
Trees 26 and 15:     6
```

Figure 1.38: *Phylogenetic topology comparisons using 'treedist' PHYLIP's tool. As it is obvious of the comparison results the 25 phylogenetic trees with the consensus tree when all compared in pairs with the consensus tree, a few differences were found, which is our first hind that the consensus tree was created properly.*

Finally, for this step, we are going to create a distribution of the above comparison results to show in a more functional that our conclusion and presentiment about a properly created consensus tree are genuine (Figure 1.3.7).

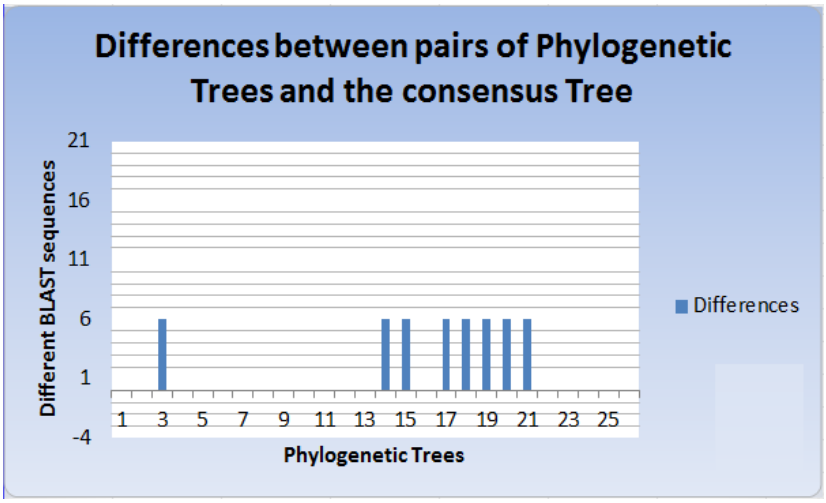


Figure 1.39: *Pair Distribution of Phylogenetic Trees with the consensus tree*

From the above figure the outcome that is reached shows that the differences between the consensus tree and the phylogenetic trees are minor (consider that each phylogenetic tree was built on 21 significant sequences given from BLAST) and as a result, we can now imply that our premonition was definitely true. Furthermore, these so few differences show that correctly with had implied some steps before that from the BLAST alignment we cannot find phylogenetic differences.

1.3.7 IS200 Transposase SuperTree

Finally, having created all the phylogenetic trees, the consensus tree, and their comparison distribution as a last step in this report, we are assigned to create a supertree from the reticulate network to add a different prospect in the phylogenetic history of the 'taxa' that where previously compared.

To do that, we used the SplitsTree[45] v4 Tool as was indicated in Bioinformatics (ITMB, MD, 2012-13) class. To build the Supertree we only thing to do is to 'feed' it with all the phylogenetic trees that we found in a previous step, each emanated from 21 BLAST significant sequences. Figure **??** shows the supertree's visualization. It worths saying that to make sense, the supertree schema, to a common biologist or even a user in general, it is recommended to searched for the annotations of the supertree in a database like NCBI's, where also BLAST API is also found.

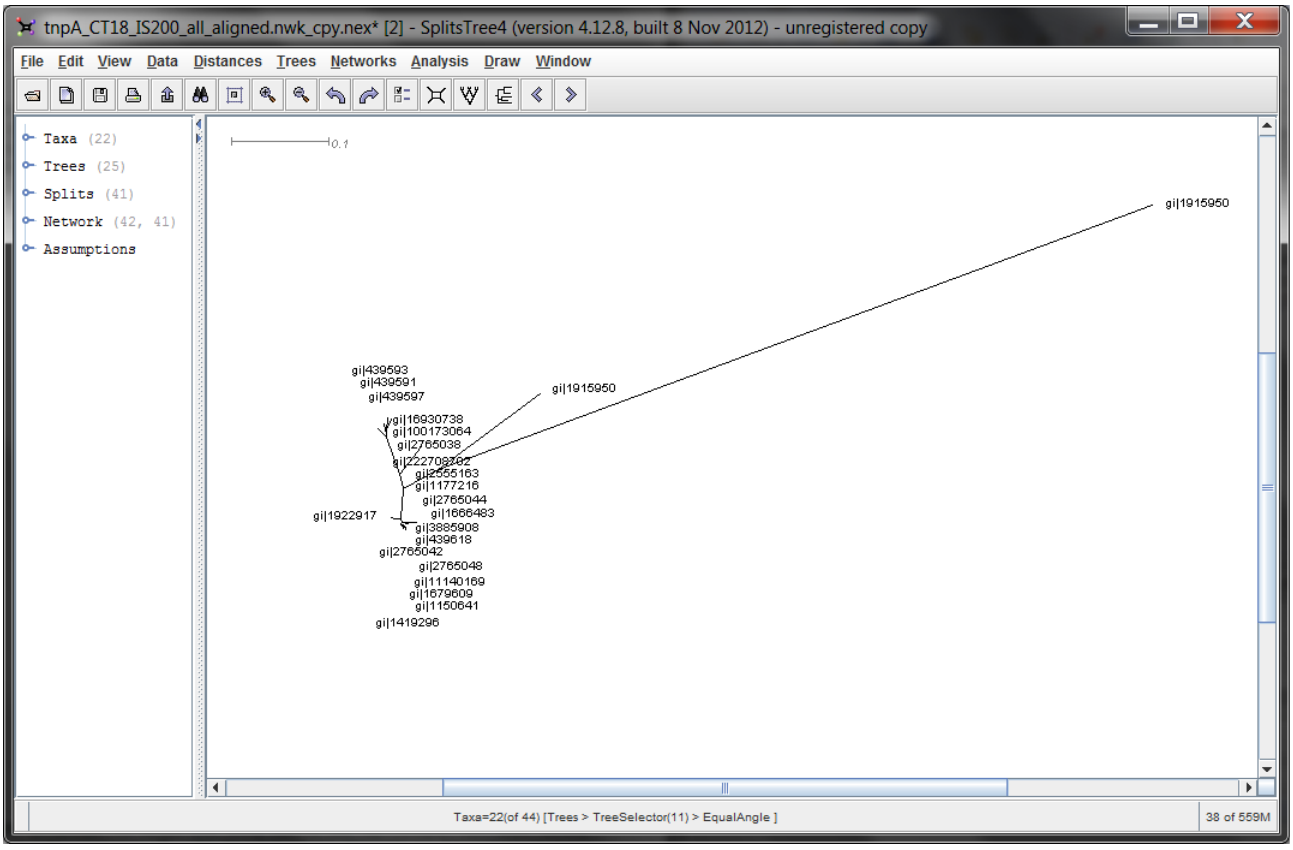


Figure 1.40: *Supertree Visualization*

To sum up, we were assigned with this last step in our report so that to find out that we can inspect the phylogenetic history of 'taxa' via another prospect, which we believe is a very promising prospect view in the evolution of Bioinformatics in Molecular Biology.

[45] http://www.splitstree.org/

CHAPTER 2

Miscellaneous

In this chapter we present some more detailed information that are referred in the first chapter. The ordination of this chapter is analogous to the ordination of the first chapter. It is worth noting that all figures shown in the chapter are screenshots from a lot of relevant to what is requested databases and browsers, which are cited in the caption of each figure.

2.1 Introduction

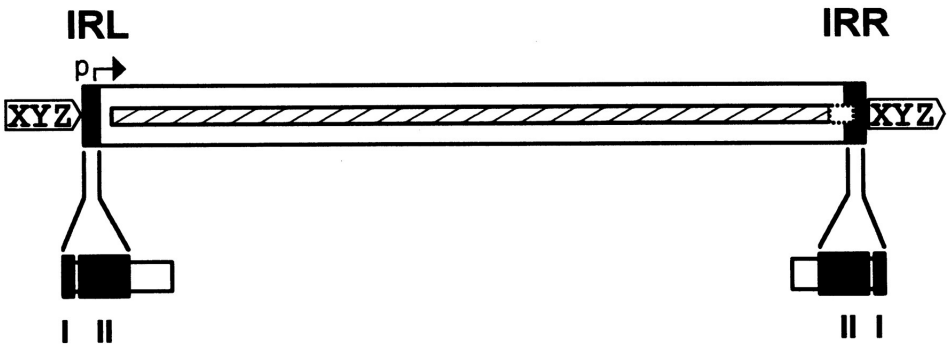


Figure 2.1: *Organization of a typical IS. The IS is represented as an open box in which the terminal IRs are shown as grey boxes labelled IRL (left inverted repeat) and IRR (right inverted repeat). A single open reading frame encoding the transposase is indicated as a hatched box stretching along the entire length of the IS and extending within the IRR sequence. XYZ enclosed in a pointed box flanking the IS represents short DR sequences generated in the target DNA as a consequence of insertion. The Tpase promoter, p, which is partially localized in IRL, is shown by a horizontal arrow. A typical domain structure (grey boxes) of the IRs is indicated beneath. Domain I represents the terminal base pairs at the very tip of the element whose recognition is required for Tpase-mediated cleavage. Domain II represents the base pairs necessary for sequence-specific recognition and binding by the Tpase*[2].

[2] http://mmbr.asm.org/content/62/3/725/F1.expansion.html

Major features of prokaryote IS families

Family	Group(s)	Size range (bp)[a]	DR (bp)[b]	ENDS[c]	IR[d]	No. of ORFs[e]	Comments[f]
IS*1*		770	9 (8–11)	GGT	Y	2	Phage λ integrase?
IS*3*	IS*2*	1,300–1,350	5	TGA	Y	2	DD(35)E
	IS*3*	1,200–1,300	3 (4)			2	
	IS*51*	1,300–1,400	3 (4)			2	
	IS*150*	1,400–1,550	3–5			2	
	IS*407*	1,200–1,250	4			2	
IS*4*		1,300–1,950	9–12	C(A)	Y	1	DDE
IS*5*	IS*5*	1,100–1,350	4	GG	Y	1	DDE
	IS*427*	800–1,000	2–3	Ga/g		2	
	IS*903*	1,000–1,100	9	GGC		1	
	IS*1031*	850–950	3	GAG		1	
	ISH*1*	900–1,150	8			1	
	ISL*2*	800–1,100	2–3			1	
IS*6*		750–900	8	GG	Y	1	DD(34)E
IS*21*		1,950–2,500	4 (5, 8)	TG	Y	2	DDE
IS*30*		1,000–1,250	2–3		Y	1	DD(33)E
IS*66*		2,500–2,700	8	GTA	Y	>3	
IS*91*		1,500–1,850	0		N	1	ssDNA Rep
IS*110*		1,200–1,550	0		N	1	Site-specific recombinase
IS*200*/IS*605*		700–2,000	0		N	1 (2)	Complex organization
IS*256*		1,300–1,500	8–9	Gg/a	Y	1	DDE, eukaryotic relatives
IS*630*		1,100–1,200	2		Y	1	DDE, eukaryotic relatives
IS*982*		1,000	ND[g]	AC	Y	1	DDE
IS*1380*		1,650	4	Cc/g	Y	1	
ISAs*1*		1,200–1,350	8	C	Y	1	
ISL*3*		1,300–1,550	8	GG	Y	1	

Figure 2.2: *Major features of prokaryote IS families.* **(a)** *Size range represents the typical range of each group.* **(b)** *Length of direct target repeats. Less frequently observed lengths are included in parentheses.* **(c)** *Conserved terminal base pairs. Capital letters (and capital letters within parentheses) refer to mostly (and often) conserved bases. Lowercase letters separated by slashes indicate alternative conservation at that position.* **(d)** *Presence (Y) or absence (N) of terminal inverted repeats.* **(e)** *ORF, open reading frame. Number in parentheses indicates the possible involvement of a second ORF in the transposition process.* **(f)** *DDE represents the common acidic triad presumed to be part of the active site of the transposase. ssDNA, single-stranded DNA.* **(g)** *ND, not determined*[4].

[4] http://mmbr.asm.org/content/62/3/725/T2.expansion.html

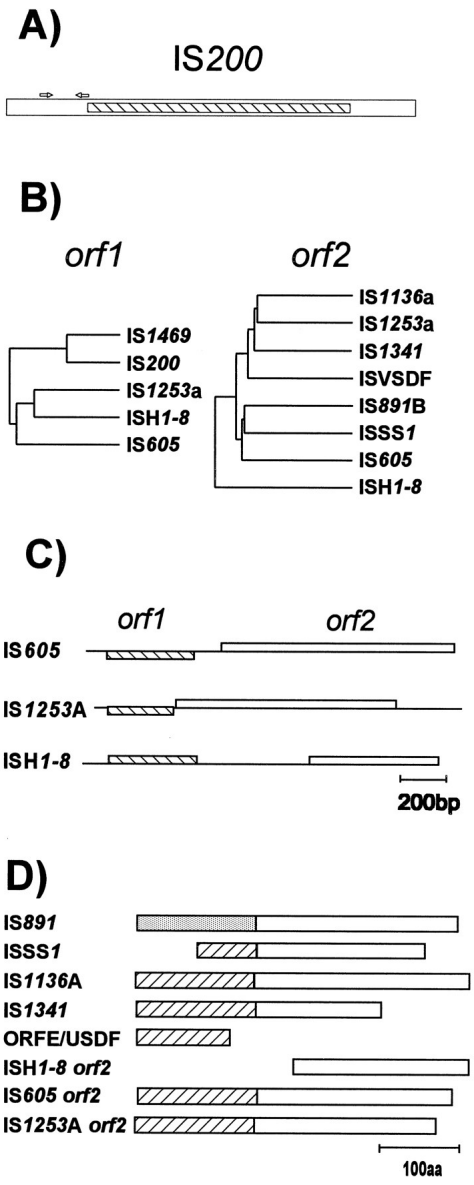


Figure 2.3: *IS200 complex. (A) Organization of IS200. Short IRs (open arrows) are shown at the left end, and the relative position of the potential open reading frame (hatched box) is indicated. (B) Dendrogram of IS200 family Tpases,orf1 (left) and the associated orf2 reading frames (right). (C) Relative localization of orf1 andorf2 in selected examples. The convention for the orientation of each reading frame is that frames shown above the line are transcribed to the right while those below the line are transcribed to the left. (D) Relationship between various examples oforf2 and other IS elements. aa, amino acids*[6].

[6] http://mmbr.asm.org/content/62/3/725/F17.expansion.html

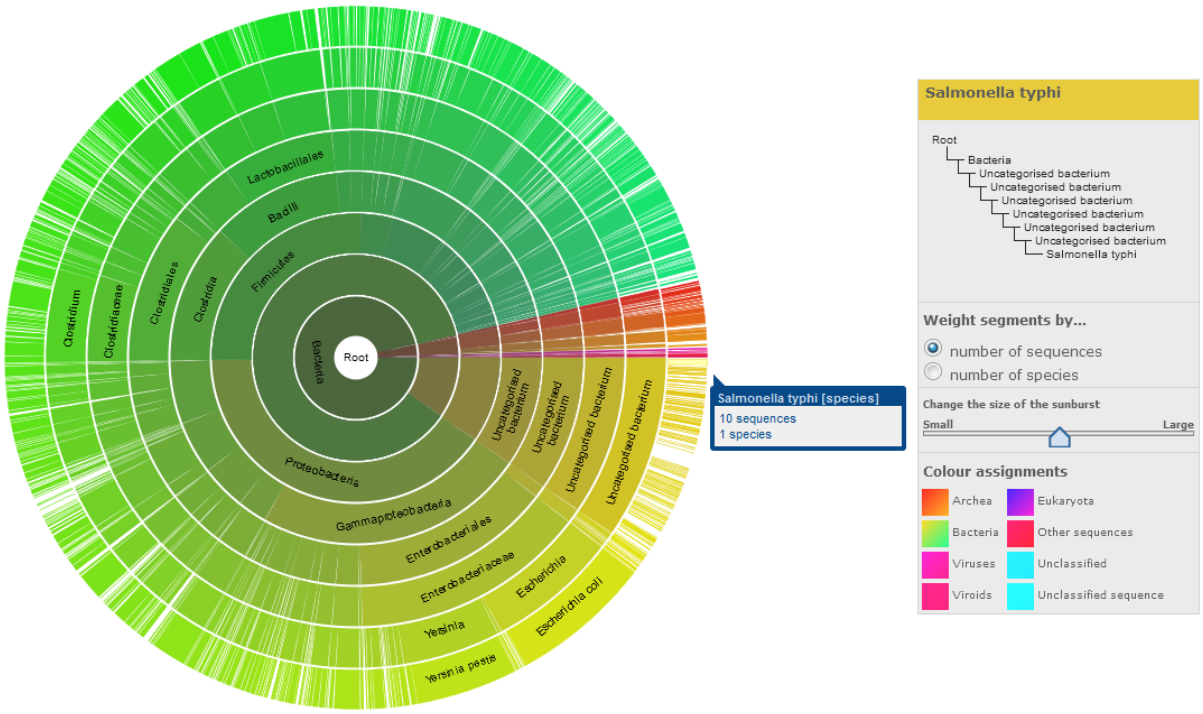


Figure 2.4: *Transposase IS200 - Salmonella Typhi. In this figure there are represented all the 1193 species that encode the IS200 transposase*[8].

2.2 Salmonella Typhi CT18

St.dna | Start

```
1   FT   CDS             190..255
2   FT                   /class="3.1.18"
3   FT                   /colour=7
4   FT                   /ec_orthologue="LPT_ECOLI"
5   FT                   /gene="STY0001"
6   FT                   /gene="thrL"
7   FT                   /note="Orthologue of E. coli thrL (LPT_ECOLI); Fasta hit
8   FT                   to LPT_ECOLI (21 aa), 86% identity in 21 aa overlap"
9   FT                   /product="thr operon leader peptide"
10  FT   CDS             337..2799
11  FT                   /class="3.1.18"
12  FT                   /colour=7
13  FT                   /ec_orthologue="AK1H_ECOLI"
14  FT                   /gene="STY0002"
15  FT                   /gene="thrA"
16  FT                   /note="Orthologue of E. coli thrA (AK1H_ECOLI); Fasta hit
17  FT                   to AK1H_ECOLI (820 aa), 94% identity in 820 aa overlap"
18  FT                   /product="aspartokinase I/homoserine dehydrogenase I"
19  FT   misc_feature    343..369
20  FT                   /colour=8
21  FT                   /note="PS00324 Aspartokinase signature"

  . . .

63191  FT   CDS             4808327..4809013
63192  FT                   /class="2.2.11"
63193  FT                   /colour=2
63194  FT                   /ec_orthologue="LAST_ECOLI"
63195  FT                   /gene="lasT"
63196  FT                   /gene="STY4949"
63197  FT                   /note="Similar to Escherichia coli hypothetical tRNA/rRNA
63198  FT                   methyltransferase LasT LasT SW:LAST_ECOLI (P37005) (228
63199  FT                   aa) fasta scores: E(): 0, 83.3% id in 228 aa, and to
63200  FT                   Serratia marcescens hypothetical  tRNA/rRNA LasT LasT
63201  FT                   SW:LAST_SERMA (P37006) (196 aa) fasta scores: E(): 0,
63202  FT                   68.5% id in 197 aa, and to Pyrococcus sp spoU protein SpoU
63203  FT                   TR:O74022 (EMBL:D86335) (235 aa) fasta scores: E():
63204  FT                   6.4e-14, 30.4% id in 237 aa"
63205  FT                   /product="putative RNA methyltransferase"
63206  >Salmonella enterica serovar Typhi (Salmonella typhi) strain CT18 complete chromosome (4809037 bp)
63207  agagattacgtctggttgcaagagatcataacaggggaaattgattgaaaataaatatat
63208  cgccagcagcacatgaacaagtttcggaatgtgatcaatttaaaaatttattgacttagg
63209  cgggcagatactttaaccaatataggaatacaagacagacaaataaaaatgacagagtac
63210  acaacatccatgaaccgcatcagcaccaccaccattaccaccatcaccattaccacaggt
```

Figure 2.5: *Start file format*[9].

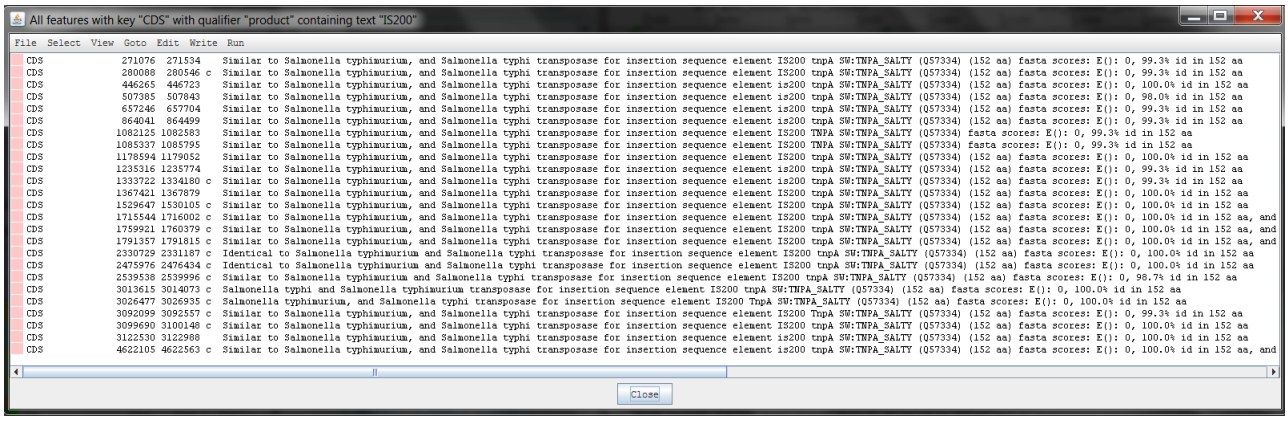
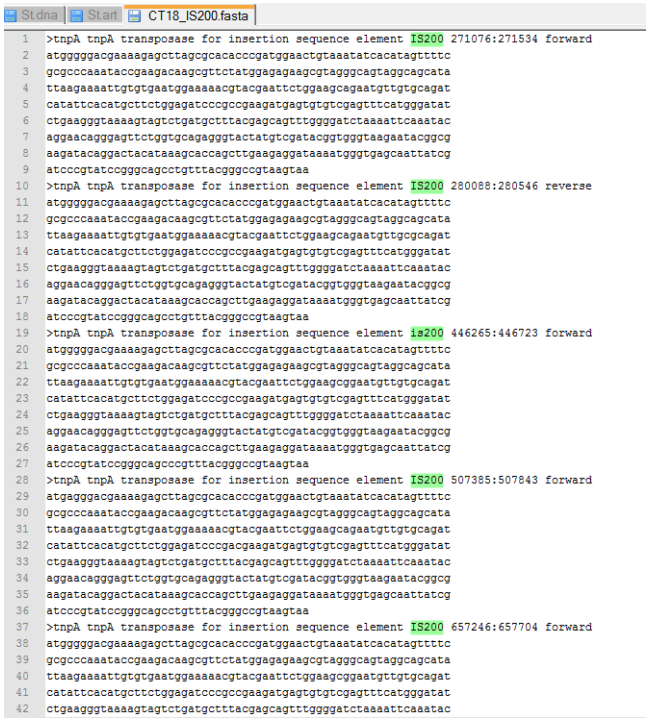


Figure 2.6: *Artemis Feature Selector - View of 25 CDS features producing IS200 transposase*

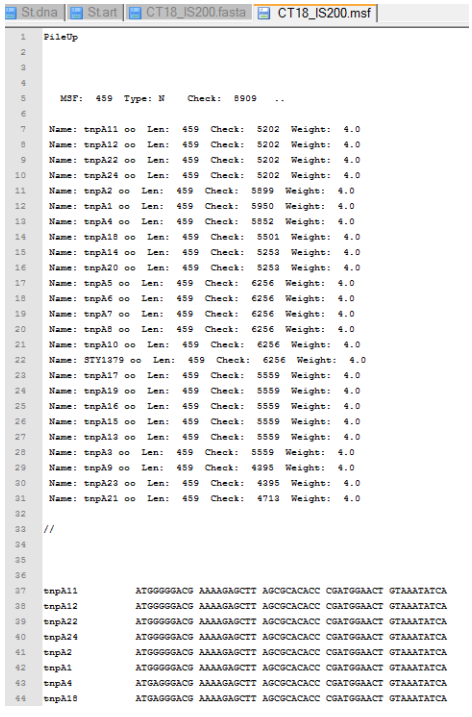
[8] http://pfam.sanger.ac.uk/family/Y1_Tnp#tabview=tab7

[9] ftp://ftp.sanger.ac.uk/pub/pathogens/Salmonella/typhi/St.art



```
St.dna | St.art | CT18_IS200.fasta
>tnpA tnpA transposase for insertion sequence element IS200 271076:271534 forward
atgggggacgaaaagagcttagcgcacacccgatggaactgtaaatatcacatagttttc
gcgcccaaataccgaagacaagcgttctatggagagaagcgtagggcagtaggcagcata
ttaagaaaattgtgtgaatggaaaaacgtacgaattctggaagcagaatgttgtgcagat
catattcacatgcttctggagatcccgccgaagatgagtgtgtcgagtttcatgggatat
ctgaagggtaaaagtagtctgatgctttacgagcagtttggggatctaaaattcaaatac
aggaacagggagttctggtgcagagggtactatgtcgatacggtgggtaagaatacggcg
aagatacaggactacataaagcaccagcttgaagaggataaaatgggtgagcaattatcg
atcccgtatccgggcagcctgtttacgggccgtaagtaa
>tnpA tnpA transposase for insertion sequence element IS200 280088:280546 reverse
atgggggacgaaaagagcttagcgcacacccgatggaactgtaaatatcacatagttttc
gcgcccaaataccgaagacaagcgttctatggagagaagcgtagggcagtaggcagcata
ttaagaaaattgtgtgaatggaaaaacgtacgaattctggaagcagaatgttgcgcagat
catattcacatgcttctggagatcccgccgaagatgagtgtgtcgagtttcatgggatat
ctgaagggtaaaagtagtctgatgctttacgagcagtttggggatctaaaattcaaatac
aggaacagggagttctggtgcagagggtactatgtcgatacggtgggtaagaatacggcg
aagatacaggactacataaagcaccagcttgaagaggataaaatgggtgagcaattatcg
atcccgtatccgggcagcctgtttacgggccgtaagtaa
>tnpA tnpA transposase for insertion sequence element is200 446265:446723 forward
atgggggacgaaaagagcttagcgcacacccgatggaactgtaaatatcacatagttttc
gcgcccaaataccgaagacaagcgttctatggagagaagcgtagggcagtaggcagcata
ttaagaaaattgtgtgaatggaaaaacgtacgaattctggaagcggaatgttgtgcagat
catattcacatgcttctggagatcccgccgaagatgagtgtgtcgagtttcatgggatat
ctgaagggtaaaagtagtctgatgctttacgagcagtttggggatctaaaattcaaatac
aggaacagggagttctggtgcagagggtactatgtcgatacggtgggtaagaatacggcg
aagatacaggactacataaagcaccagcttgaagaggataaaatgggtgagcaattatcg
atcccgtatccgggcagcccgtttacgggccgtaagtaa
>tnpA tnpA transposase for insertion sequence element IS200 507385:507843 forward
atgagggacgaaaagagcttagcgcacacccgatggaactgtaaatatcacatagttttc
gcgcccaaataccgaagacaagcgttctatggagagaagcgtagggcagtaggcagcata
ttaagaaaattgtgtgaatggaaaaacgtacgaattctggaagcagaatgttgtgcagat
catattcacatgcttctggagatcccgacgaagatgagtgtgtcgagtttcatgggatat
ctgaagggtaaaagtagtctgatgctttacgagcagtttggggatctaaaattcaaatac
aggaacagggagttctggtgcagagggtactatgtcgatacggtgggtaagaatacggcg
aagatacaggactacataaagcaccagcttgaagaggataaaatgggtgagcaattatcg
atcccgtatccgggcagcctgtttacgggccgtaagtaa
>tnpA tnpA transposase for insertion sequence element IS200 657246:657704 forward
atgggggacgaaaagagcttagcgcacacccgatggaactgtaaatatcacatagttttc
gcgcccaaataccgaagacaagcgttctatggagagaagcgtagggcagtaggcagcata
ttaagaaaattgtgtgaatggaaaaacgtacgaattctggaagcggaatgttgtgcagat
catattcacatgcttctggagatcccgccgaagatgagtgtgtcgagtttcatgggatat
ctgaagggtaaaagtagtctgatgctttacgagcagtttggggatctaaaattcaaatac
```

Figure 2.7: *Fasta file including bases of 25 CDSs producing IS200 transposase*



```
St.dna | St.art | CT18_IS200.fasta | CT18_IS200.msf
PileUp


  MSF:  459  Type: N    Check:  8909   ..

 Name: tnpA11 oo  Len:  459  Check:  5202  Weight:  4.0
 Name: tnpA12 oo  Len:  459  Check:  5202  Weight:  4.0
 Name: tnpA22 oo  Len:  459  Check:  5202  Weight:  4.0
 Name: tnpA24 oo  Len:  459  Check:  5202  Weight:  4.0
 Name: tnpA2 oo  Len:  459  Check:  5899  Weight:  4.0
 Name: tnpA1 oo  Len:  459  Check:  5950  Weight:  4.0
 Name: tnpA4 oo  Len:  459  Check:  5852  Weight:  4.0
 Name: tnpA18 oo  Len:  459  Check:  5501  Weight:  4.0
 Name: tnpA14 oo  Len:  459  Check:  5253  Weight:  4.0
 Name: tnpA20 oo  Len:  459  Check:  5253  Weight:  4.0
 Name: tnpA5 oo  Len:  459  Check:  6256  Weight:  4.0
 Name: tnpA6 oo  Len:  459  Check:  6256  Weight:  4.0
 Name: tnpA7 oo  Len:  459  Check:  6256  Weight:  4.0
 Name: tnpA8 oo  Len:  459  Check:  6256  Weight:  4.0
 Name: tnpA10 oo  Len:  459  Check:  6256  Weight:  4.0
 Name: STY1379 oo  Len:  459  Check:  6256  Weight:  4.0
 Name: tnpA17 oo  Len:  459  Check:  5559  Weight:  4.0
 Name: tnpA19 oo  Len:  459  Check:  5559  Weight:  4.0
 Name: tnpA16 oo  Len:  459  Check:  5559  Weight:  4.0
 Name: tnpA15 oo  Len:  459  Check:  5559  Weight:  4.0
 Name: tnpA13 oo  Len:  459  Check:  5559  Weight:  4.0
 Name: tnpA3 oo  Len:  459  Check:  5559  Weight:  4.0
 Name: tnpA9 oo  Len:  459  Check:  4395  Weight:  4.0
 Name: tnpA23 oo  Len:  459  Check:  4395  Weight:  4.0
 Name: tnpA21 oo  Len:  459  Check:  4713  Weight:  4.0

//


tnpA11          ATGGGGGACG AAAAGAGCTT AGCGCACACC CGATGGAACT GTAAATATCA
tnpA12          ATGGGGGACG AAAAGAGCTT AGCGCACACC CGATGGAACT GTAAATATCA
tnpA22          ATGGGGGACG AAAAGAGCTT AGCGCACACC CGATGGAACT GTAAATATCA
tnpA24          ATGGGGGACG AAAAGAGCTT AGCGCACACC CGATGGAACT GTAAATATCA
tnpA2           ATGGGGGACG AAAAGAGCTT AGCGCACACC CGATGGAACT GTAAATATCA
tnpA1           ATGGGGGACG AAAAGAGCTT AGCGCACACC CGATGGAACT GTAAATATCA
tnpA4           ATGAGGGACG AAAAGAGCTT AGCGCACACC CGATGGAACT GTAAATATCA
tnpA18          ATGAGGGACG AAAAGAGCTT AGCGCACACC CGATGGAACT GTAAATATCA
```

Figure 2.8: *Output result file of ClustalW2 msa tool - CT18_IS200.msf*

Figure 2.9: *MSA of CT18_IS200 in a full sequence view. In this figure the top window shows the MSAs that resulted from the CT18_IS200.msf that we got as an output of ClastalW2 (in aligned order). And the bottom window shows the MSAs results generated directly from CT18_IS200.fasta (in input order). In order to get a full view of the full sequence alignment we un-zoomed these windows and we can verify in this way that both MSAs consensus are identical.*

Figure 2.10: *UGENE - MSA of CT18_IS200 in directly from fasta file. This figure also shows zoomed and an un-zoomed view of MSAs.*

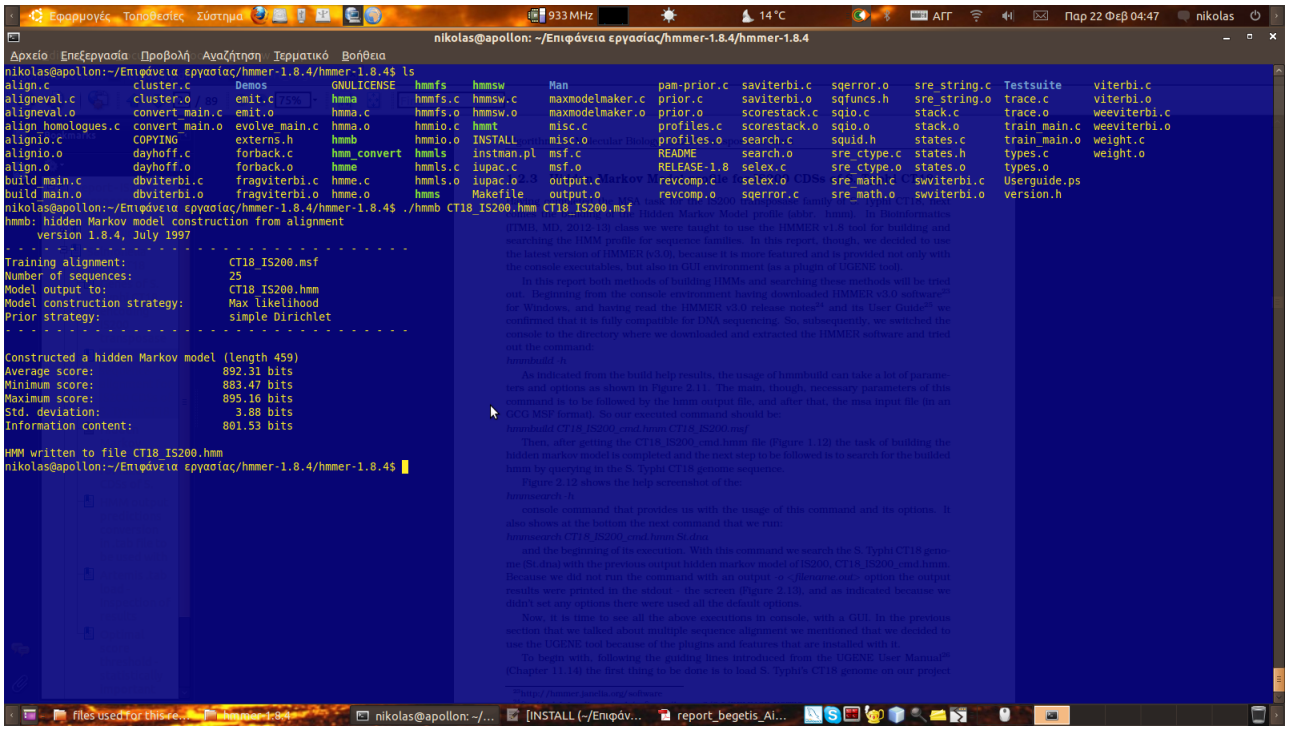


Figure 2.11: *HMMER ubuntu terminal - hmmb*

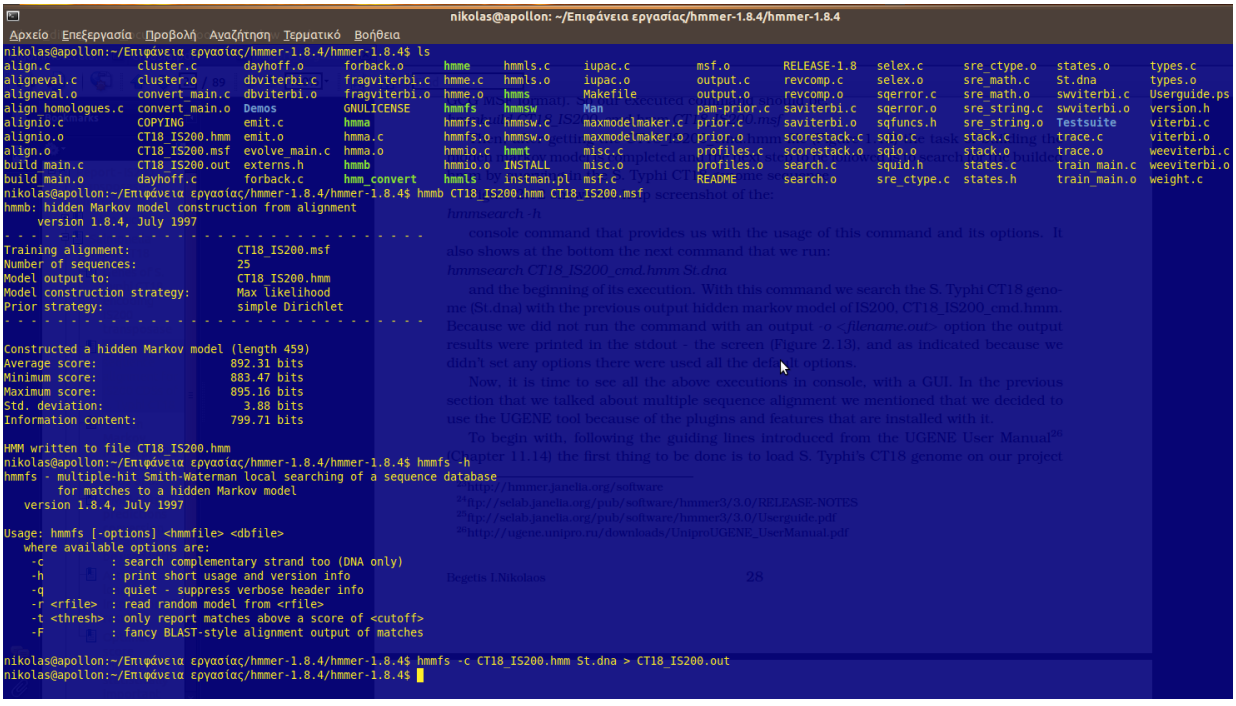


Figure 2.12: *HMMER ubuntu terminal - hmmfs help*

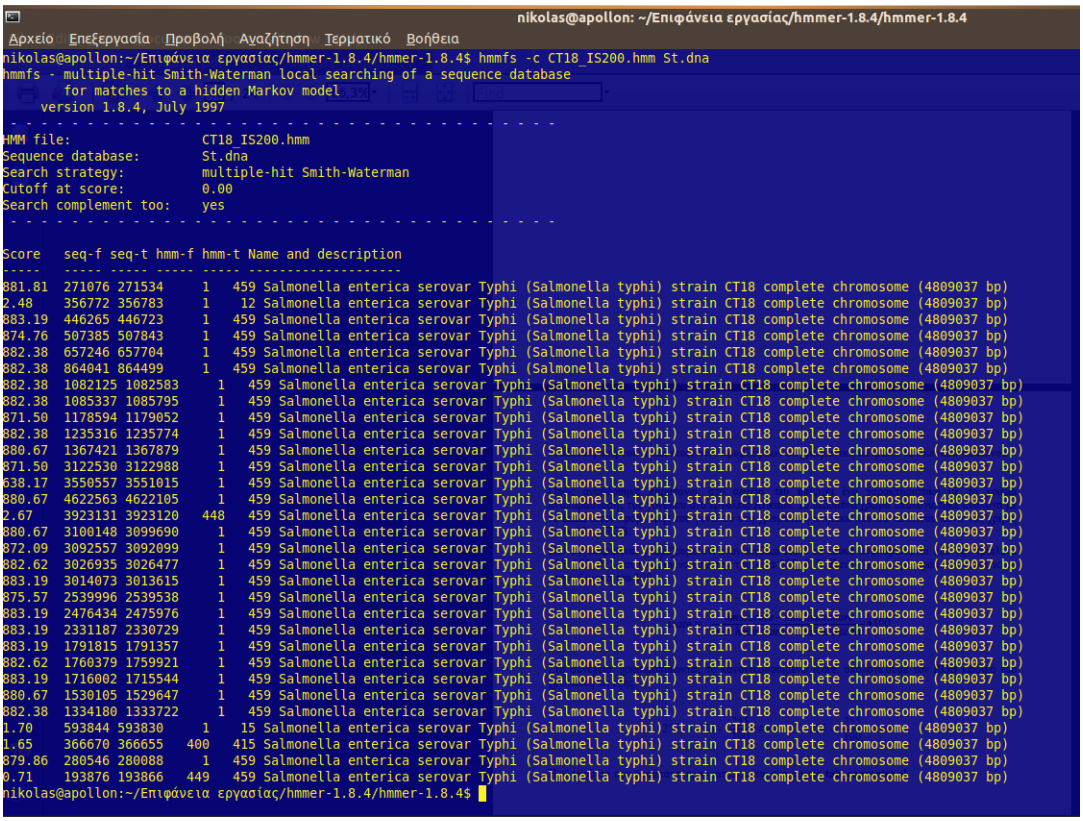


Figure 2.13: *HMMER ubuntu terminal - hmmfs output results*

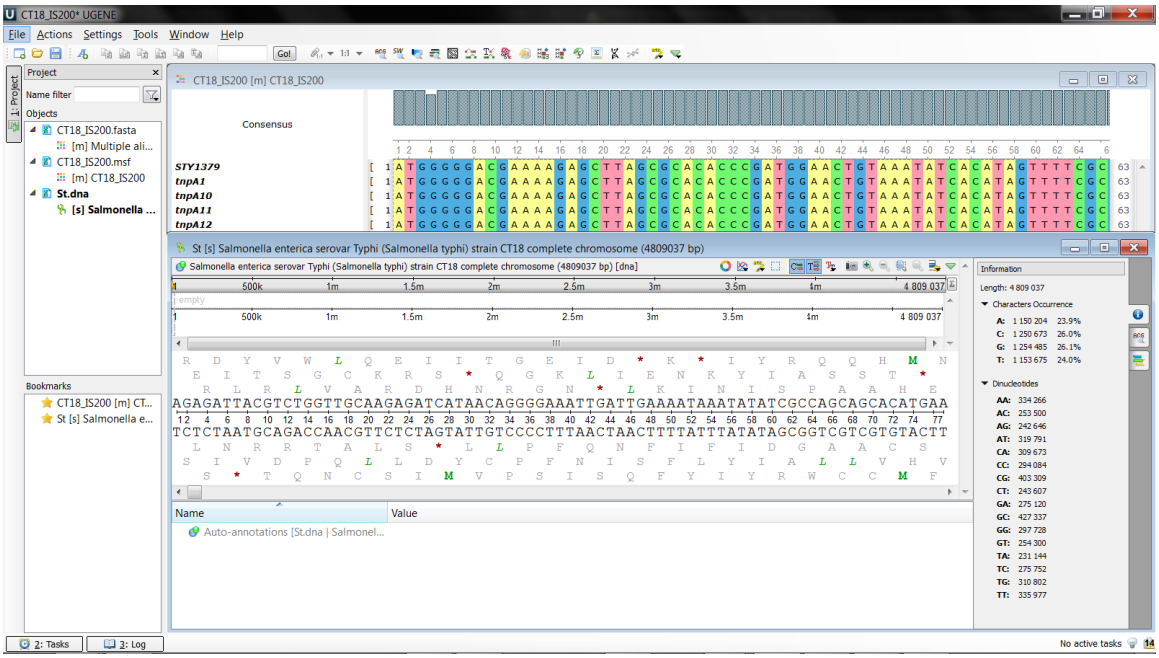


Figure 2.14: *UGENE - S. Typhi CT18 genome sequence, loaded from St.dna file. The DNA sequence is shown in the big right bottom window. A bar on the right indicates some useful information about the S.Typhi CT18 genome, and the bar on the top provides a variety of special options.*

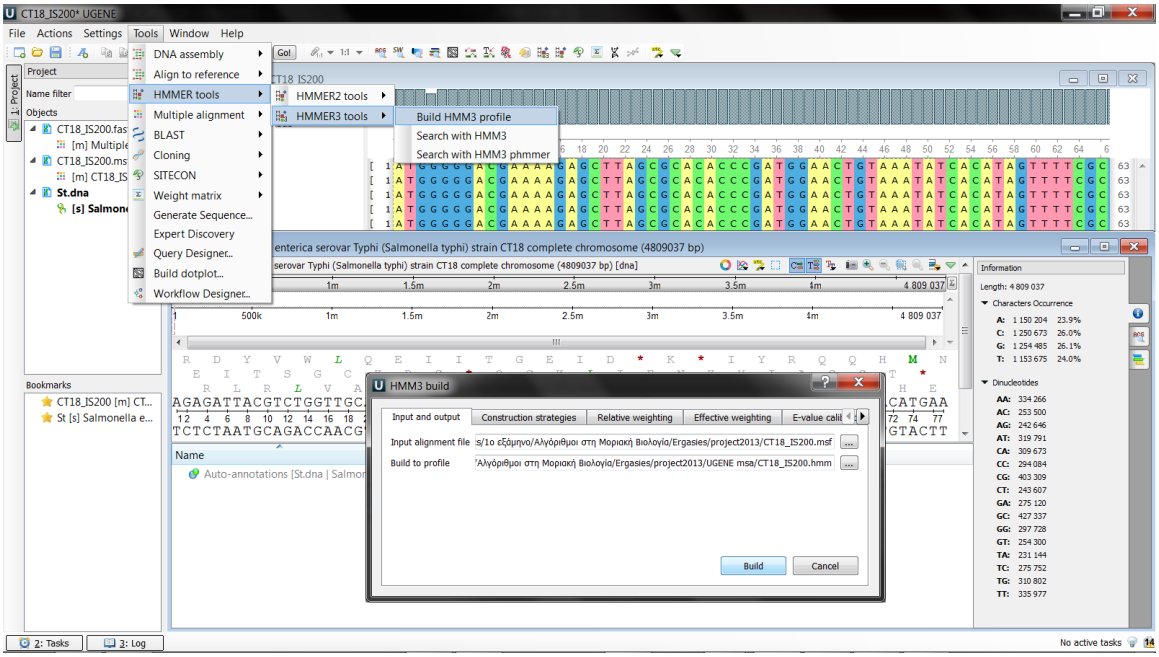


Figure 2.15: *UGENE - HMM3 build for MSAs of CT18_IS200.msf file*

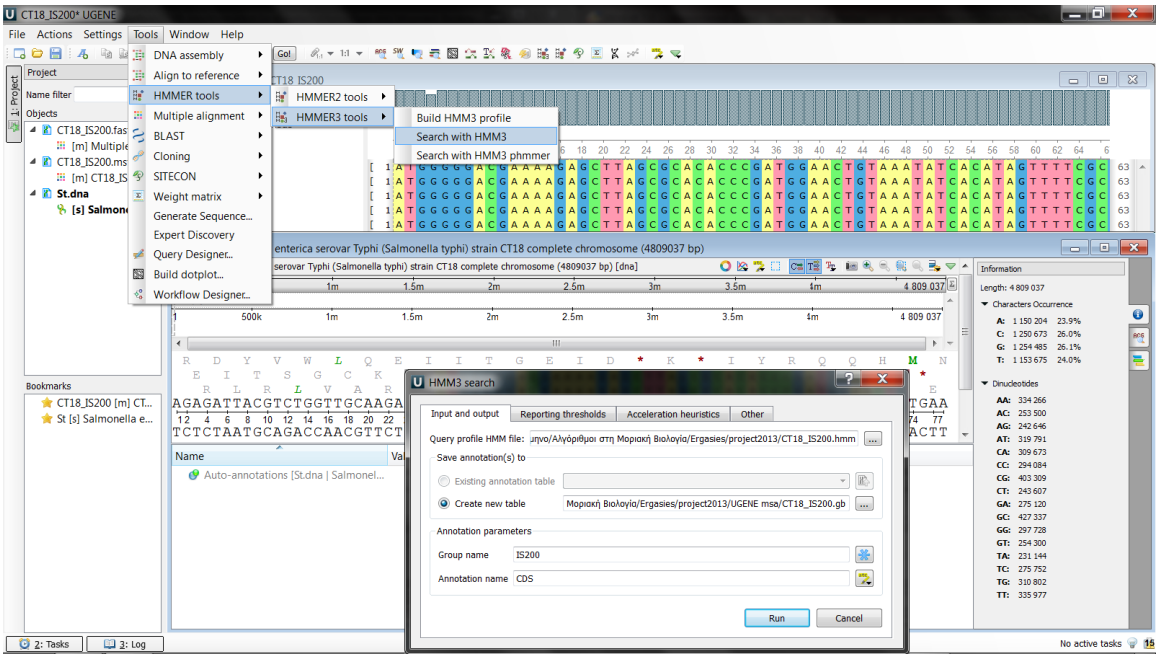


Figure 2.16: *UGENE - HMM3 search S. Typhi CT18 genome using CT18_IS200.hmm*

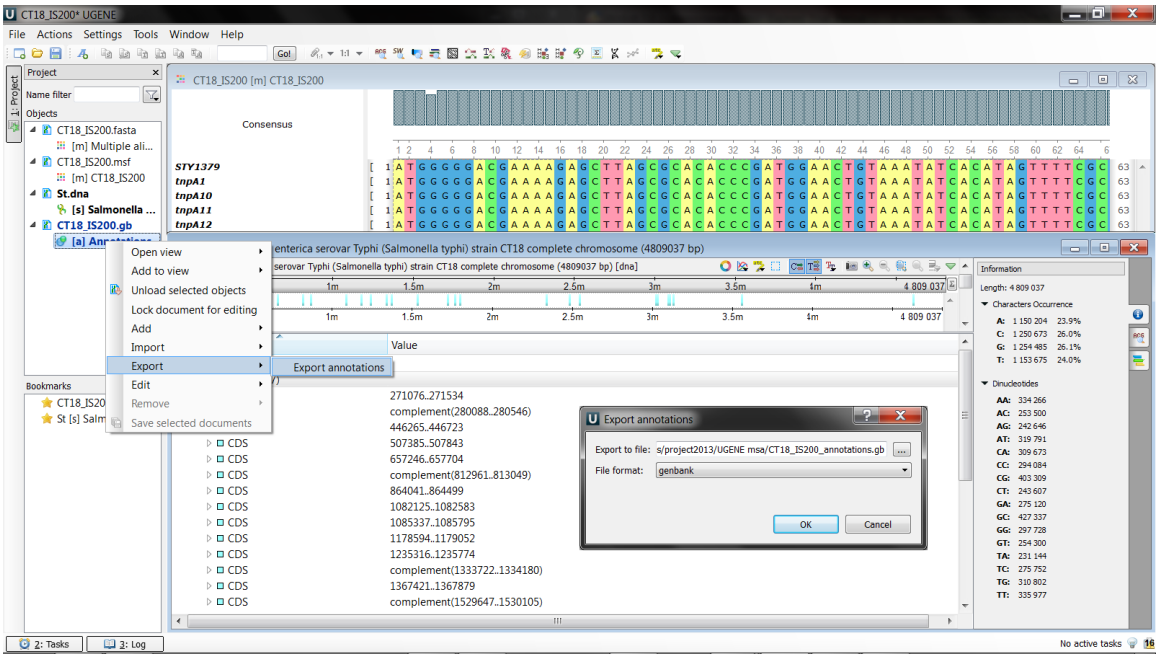


Figure 2.17: *UGENE - HMM3 search result's output predictions - CT18_IS200_predictions.gb. In the main window of S. Typhi genome there have been annotated all the 27 CDSs that resulted from the HMM search on the genome.*

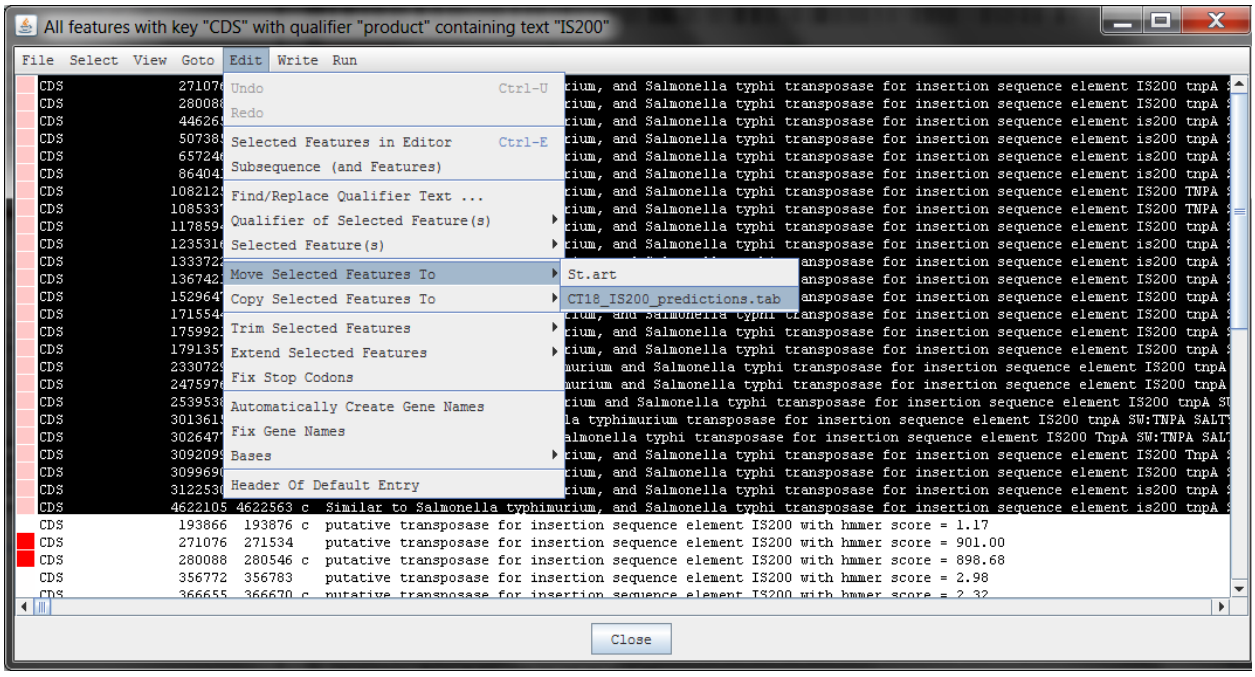


Figure 2.18: *Artemis CT18_IS200 CDSs copied in CT18_IS200_predictions.tab. The selected features are the CDSs that initially belong to S. Typhi's CT18 genome, while the unselected are those the rest that resulted from the query for finding CDSs which produce Is200 transposase, and belong to the .tab predictions file.*

Figure 2.19: *Artemis CDS shown in comparison to the next figure for overlapping*

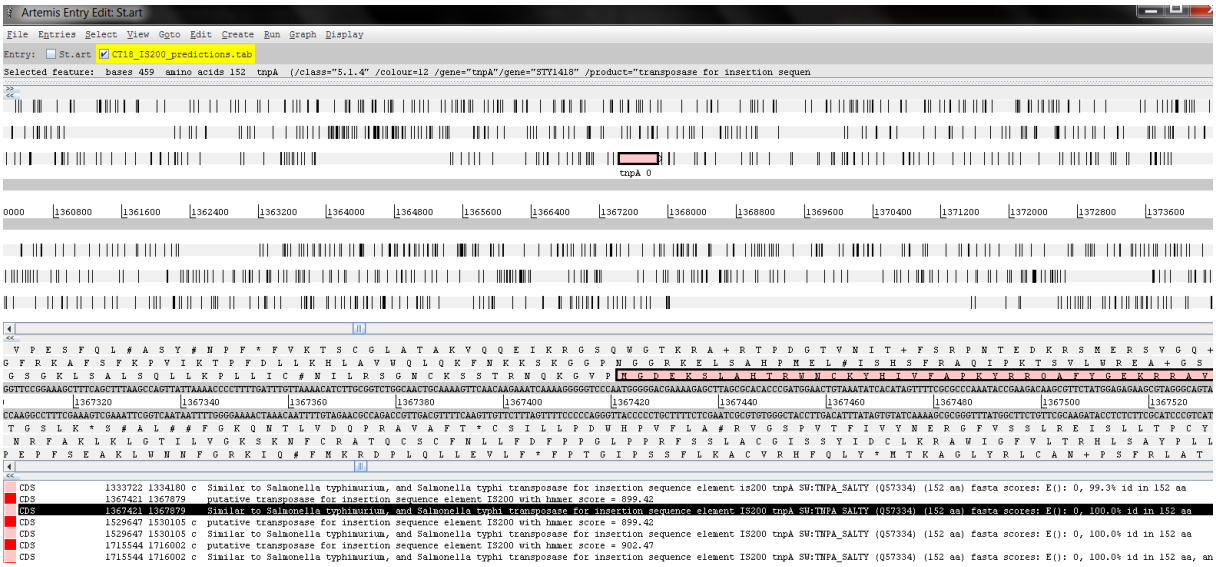


Figure 2.20: *Artemis CDS shown in comparison to the previous figure for overlapping*

2.3 Salmonella Typhi TY2

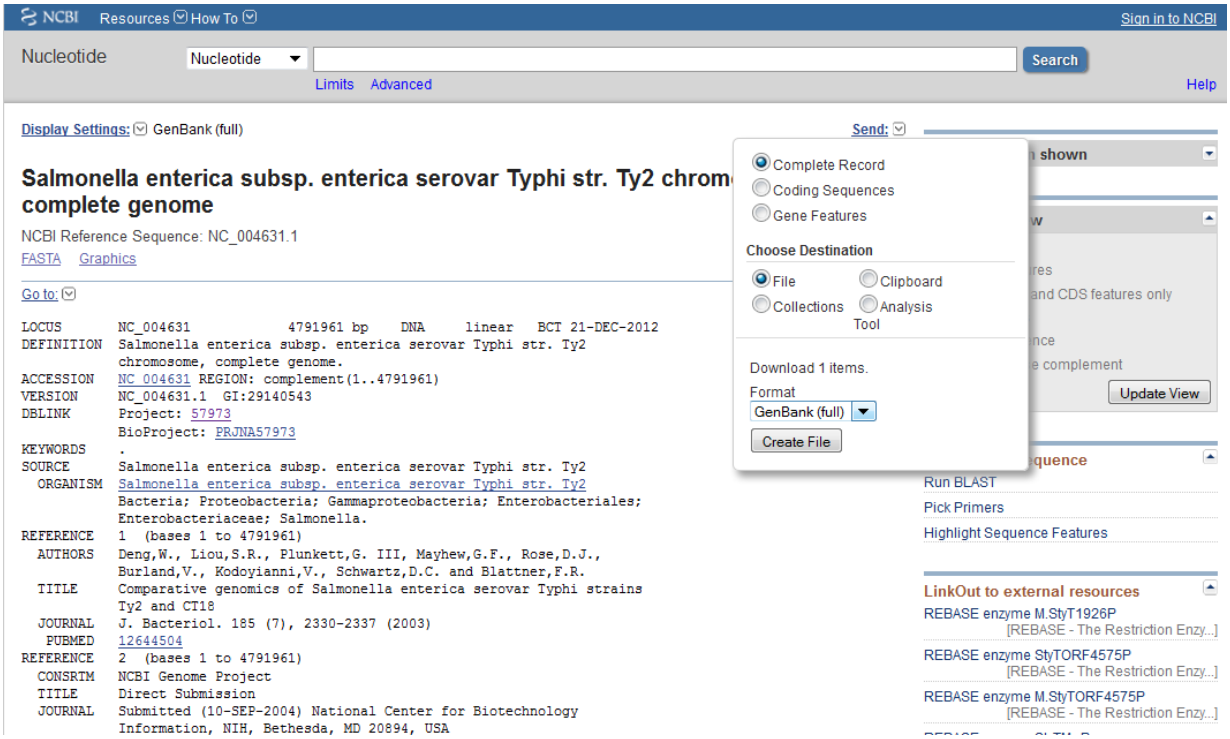


Figure 2.21: *S. Typhi's TY2 genome resource*[10]

[10] http://www.ncbi.nlm.nih.gov/nuccore/NC_004631.1?report=gbwithparts&log$=seqview

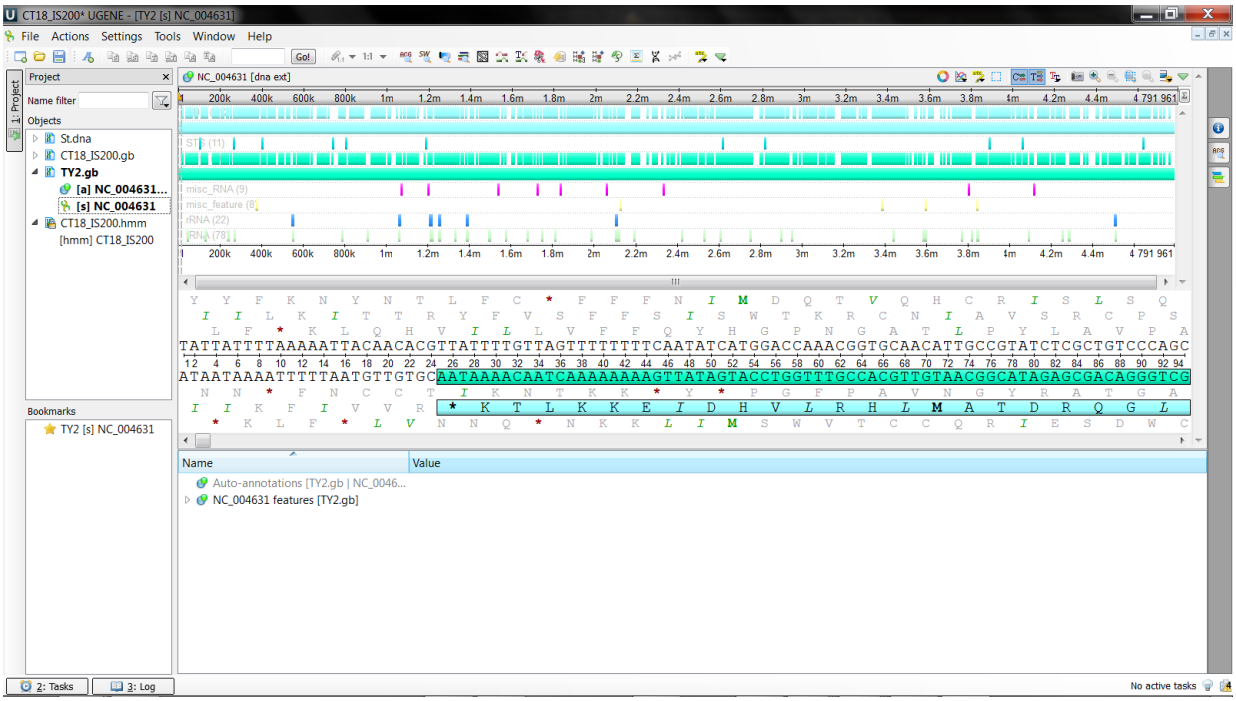


Figure 2.22: *S. Typhi's TY2 genome shown in main window of UGENE tool*

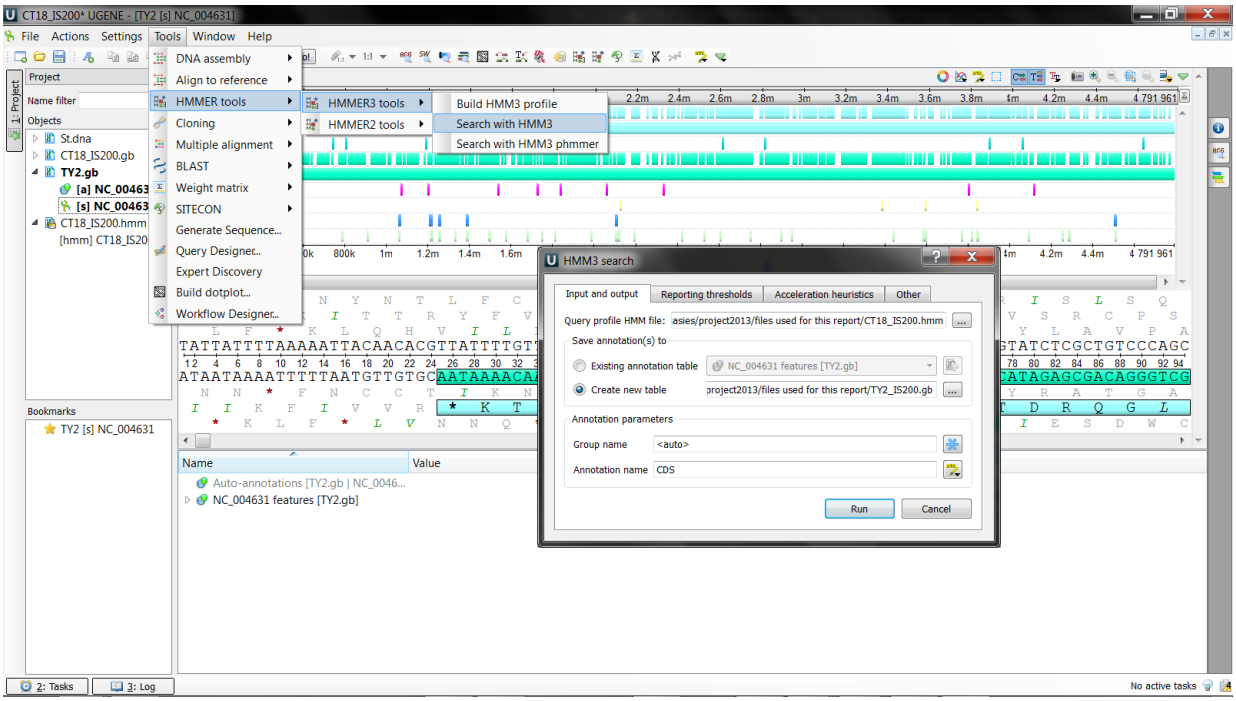


Figure 2.23: *S. Typhi's TY2 genome searched with CT18 hmm*

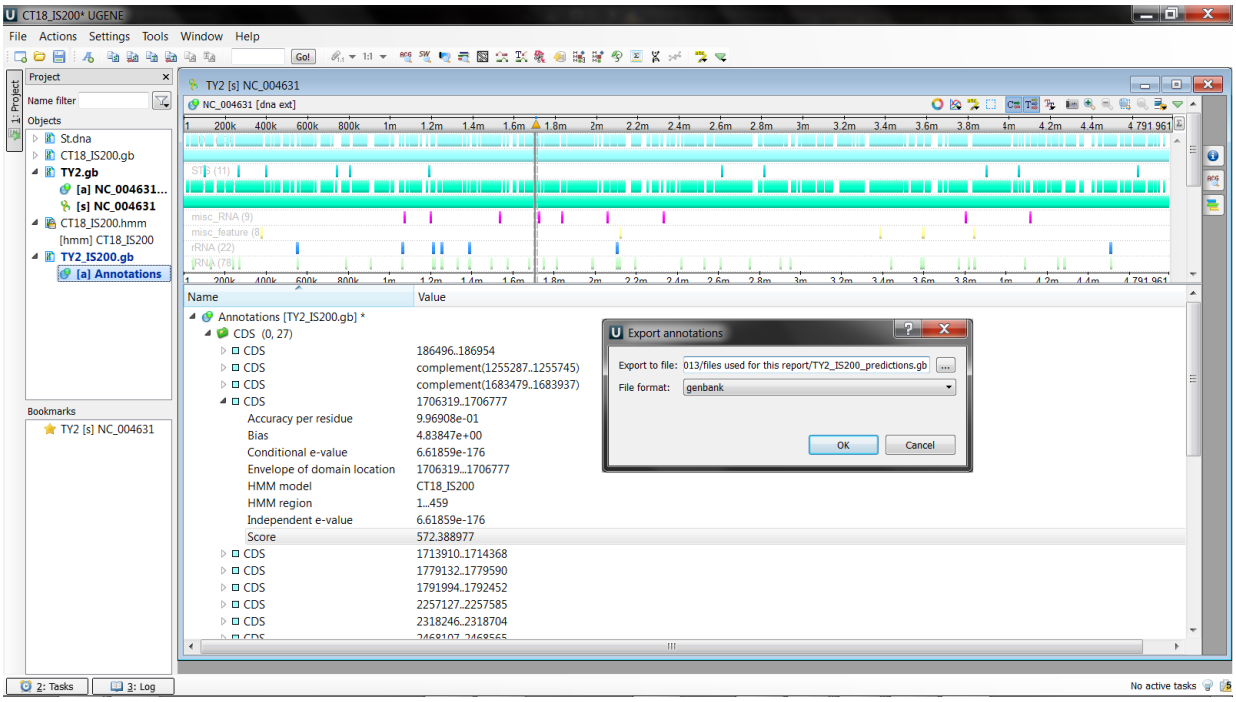


Figure 2.24: *S. Typhi's TY2 genome result predictions searched with CT18 hmm*

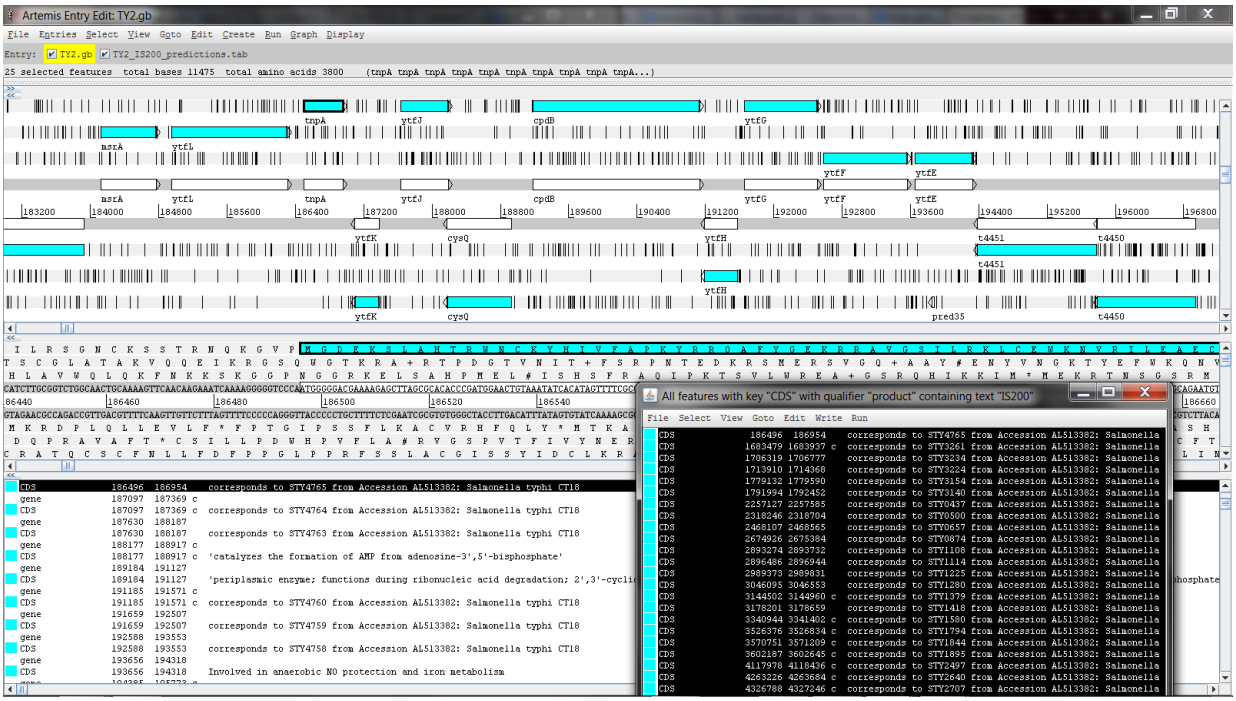


Figure 2.25: *S. Typhi's TY2 genome and TY2 predictions in Artemis. Figure shows both entries and highlights the CDSs that produce IS200 transposase.*

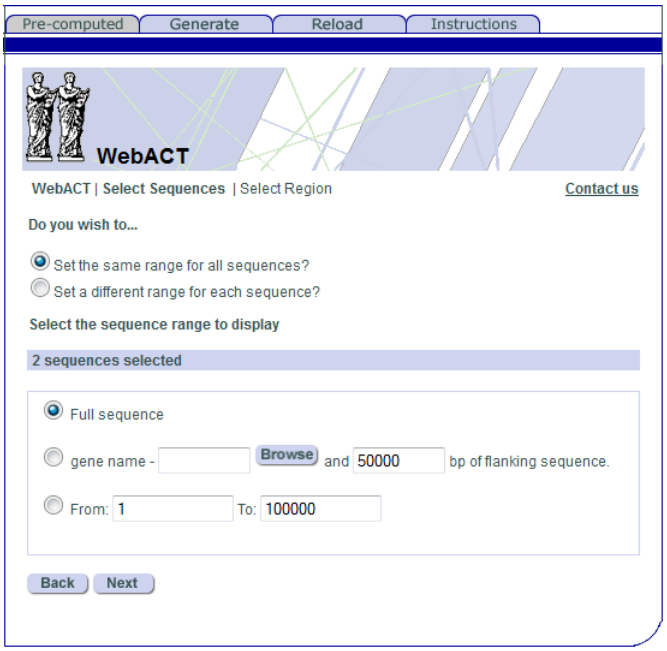


Figure 2.26: *WebACT S. Typhi's CT18 and TY2 part2 of selection*

Figure 2.27: *WebACT S. Typhi's CT18 and TY2 part3 of selection*

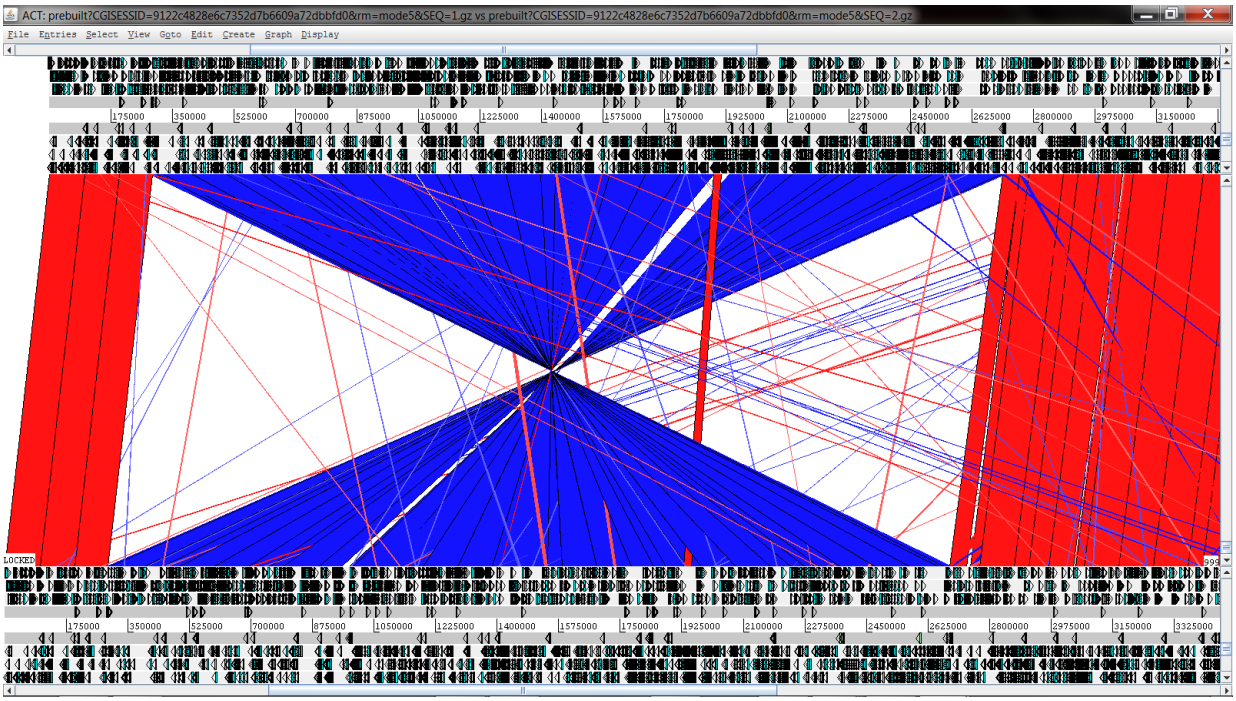


Figure 2.28: *WebACT S. Typhi's CT18 and TY2 comparison bigger threshold*

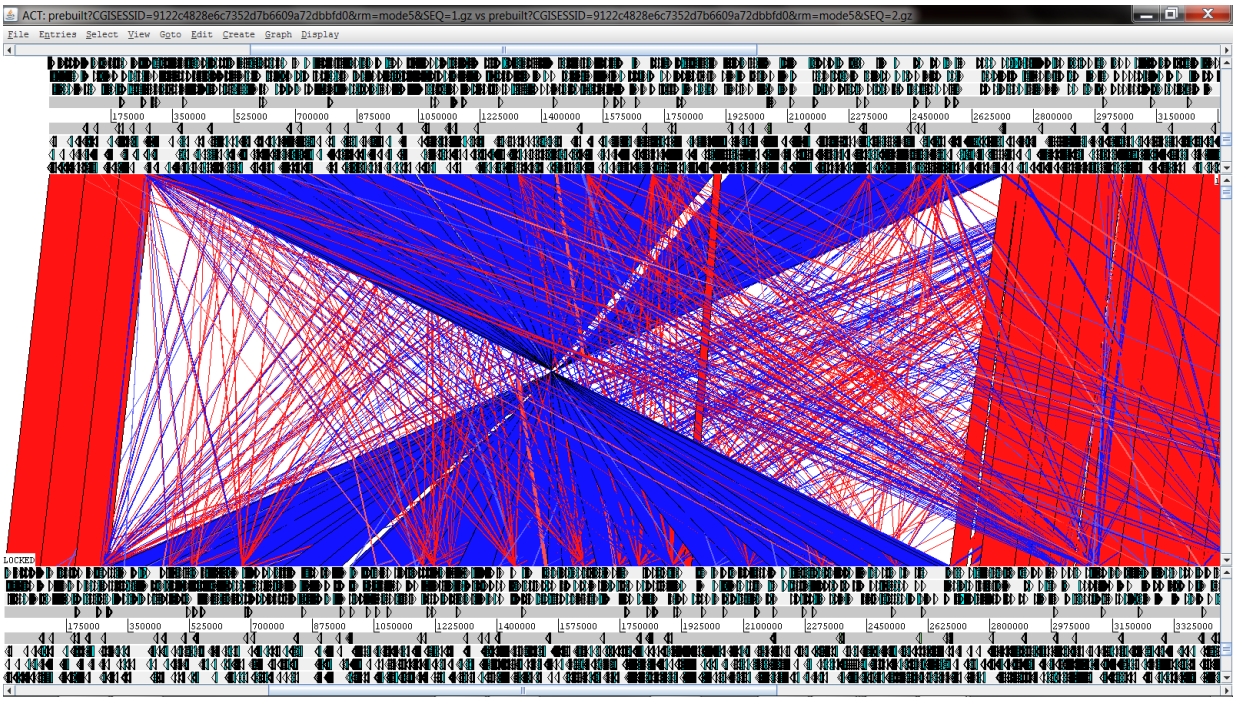


Figure 2.29: *WebACT S. Typhi's CT18 and TY2 comparison smaller threshold*

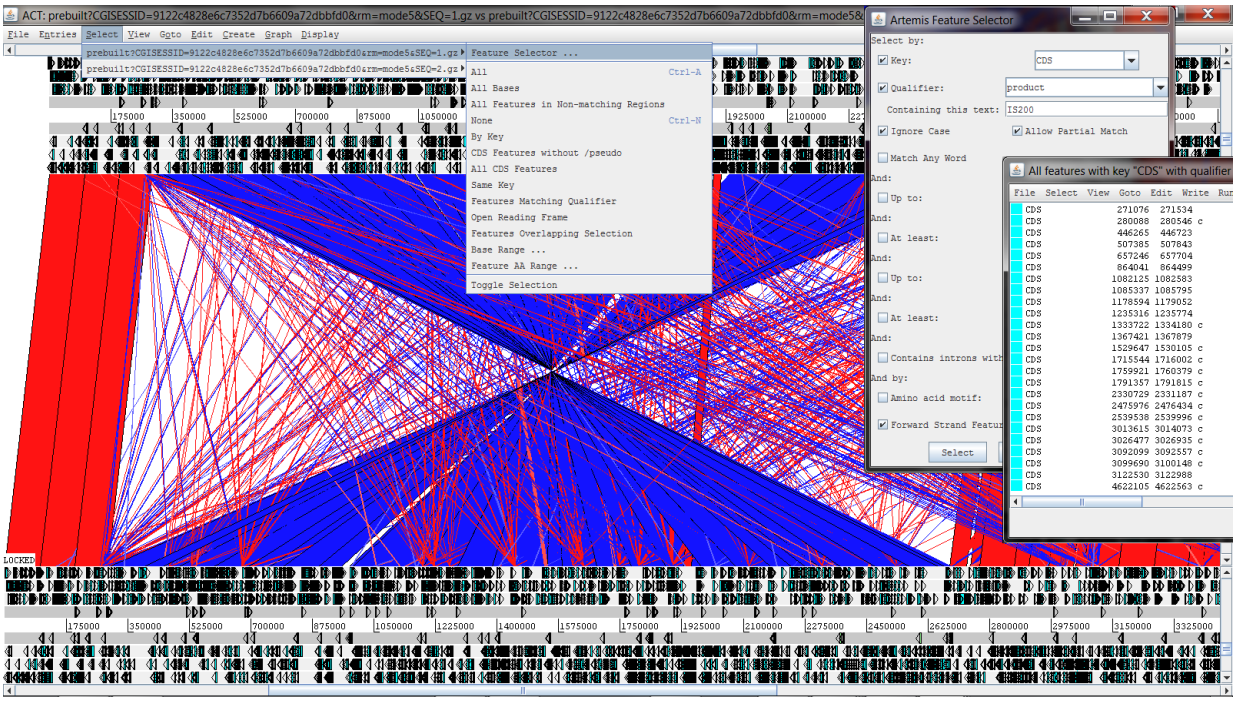


Figure 2.30: *WebACT S. Typhi's CT18 and TY2 IS200 selection*

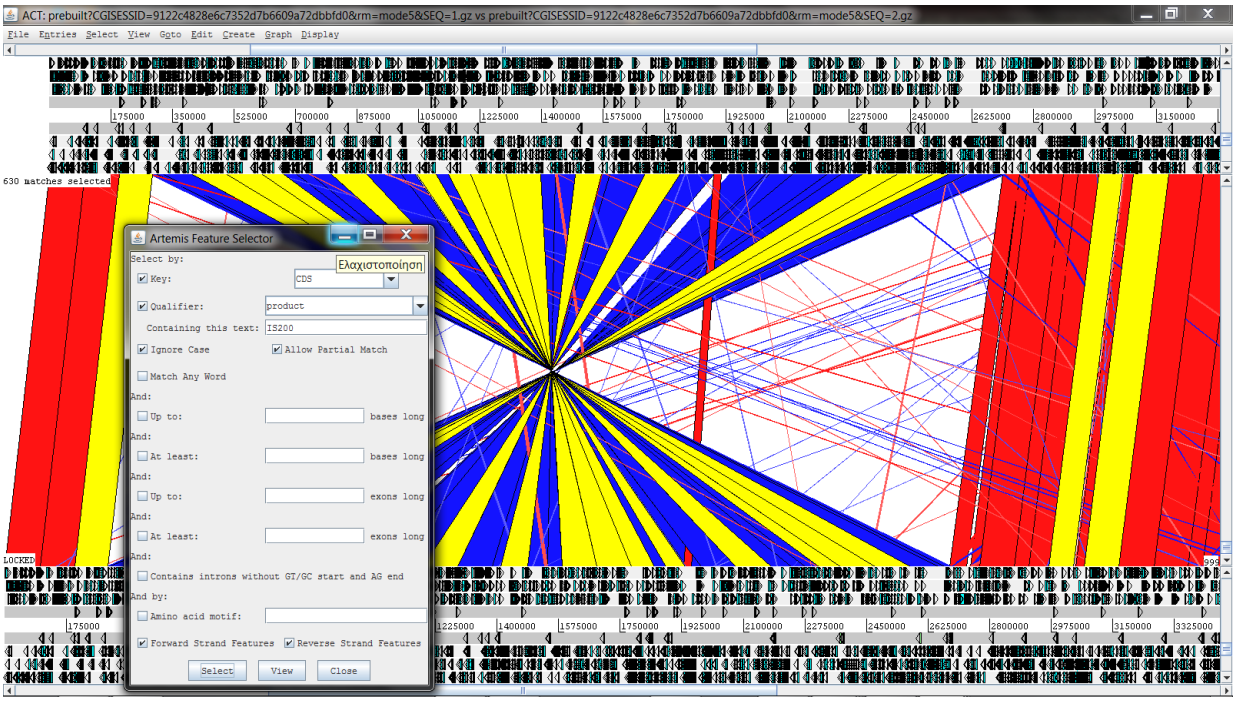


Figure 2.31: *WebACT S. Typhi's CT18 and TY2 IS200 selection result with bigger threshold*

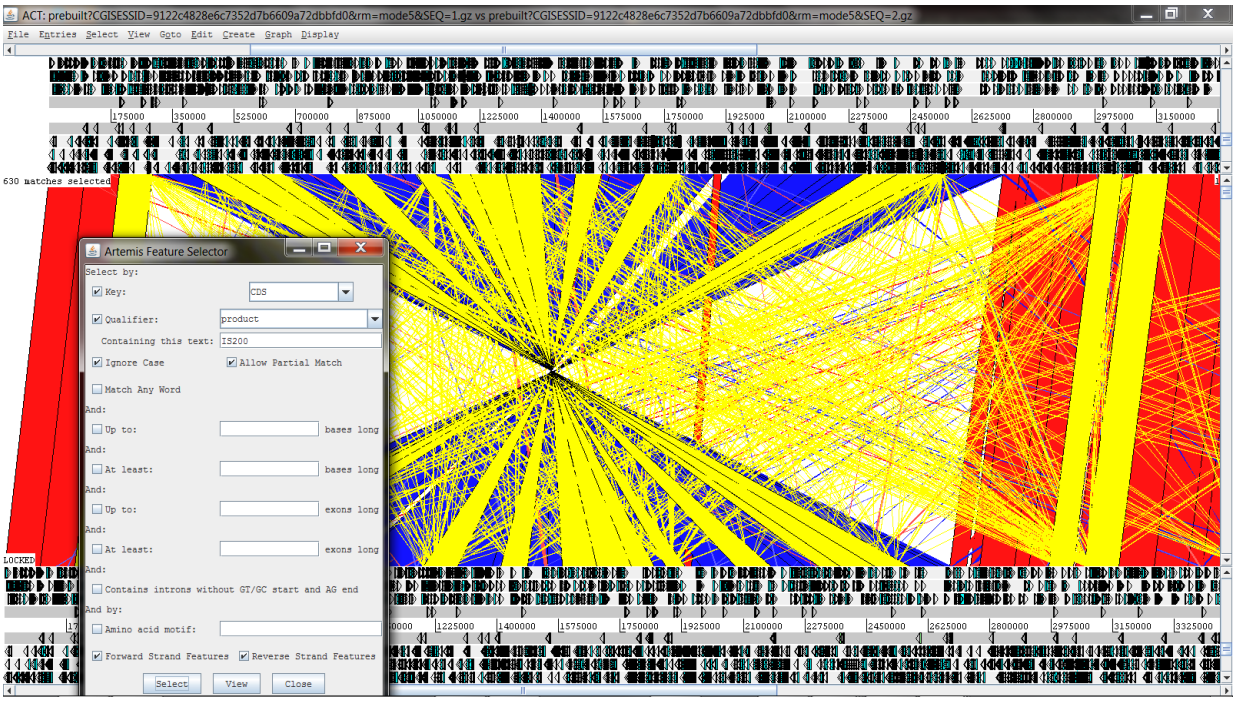


Figure 2.32: *WebACT S. Typhi's CT18 and TY2 IS200 selection result with smaller threshold*

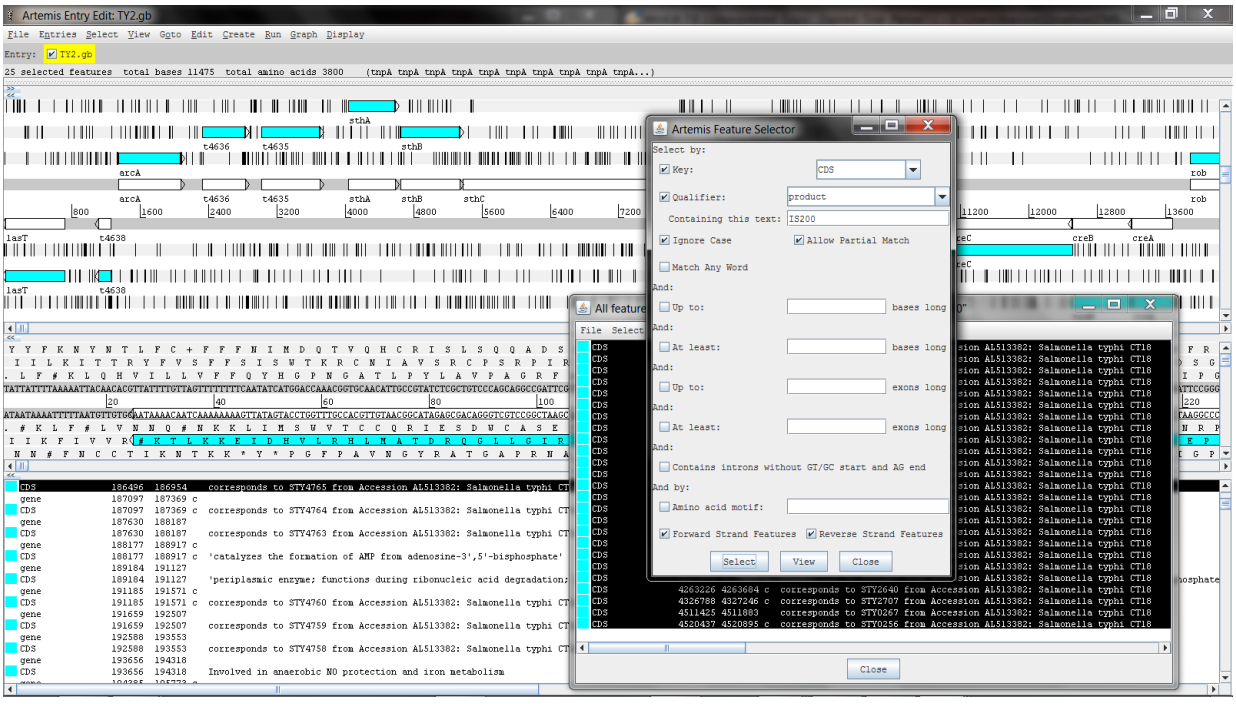


Figure 2.33: *S. Typhi's TY2 CDSs producing IS200 transposase selection*

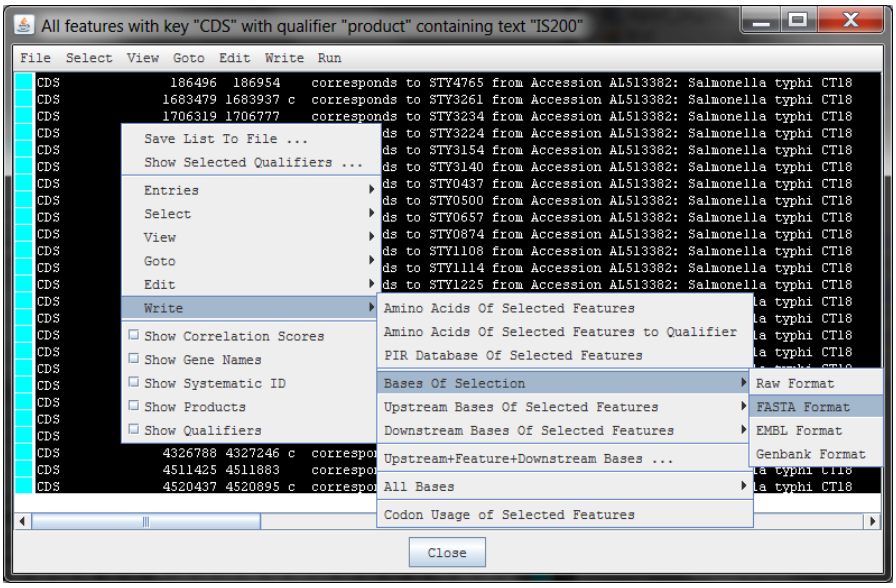


Figure 2.34: *S. Typhi's TY2 CDSs producing IS200 transposase FASTA export*

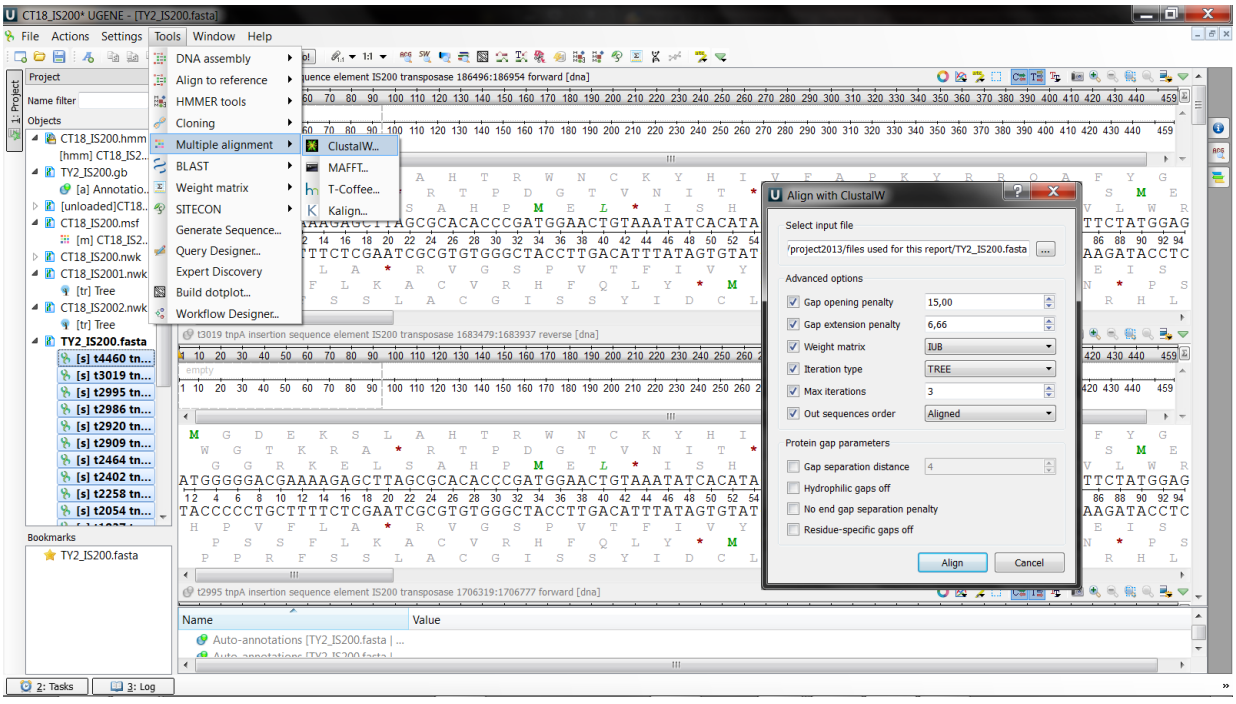


Figure 2.35: *S. Typhi's TY2 MSA using ClustalW2 UGENE's plugin*

Figure 2.36: *S. Typhi's CT18 and TY2 MSAs shown in UGENE*

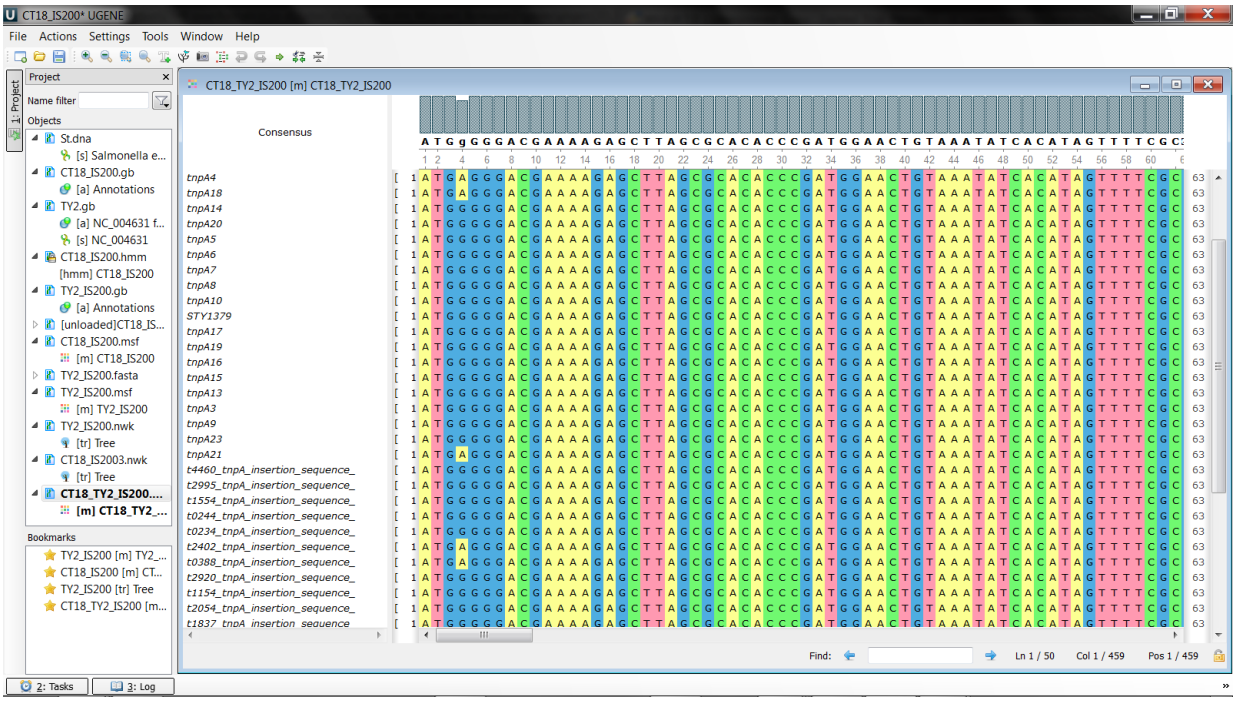


Figure 2.37: *S. Typhi's CT18 and TY2 MSAs merged shown in UGENE*

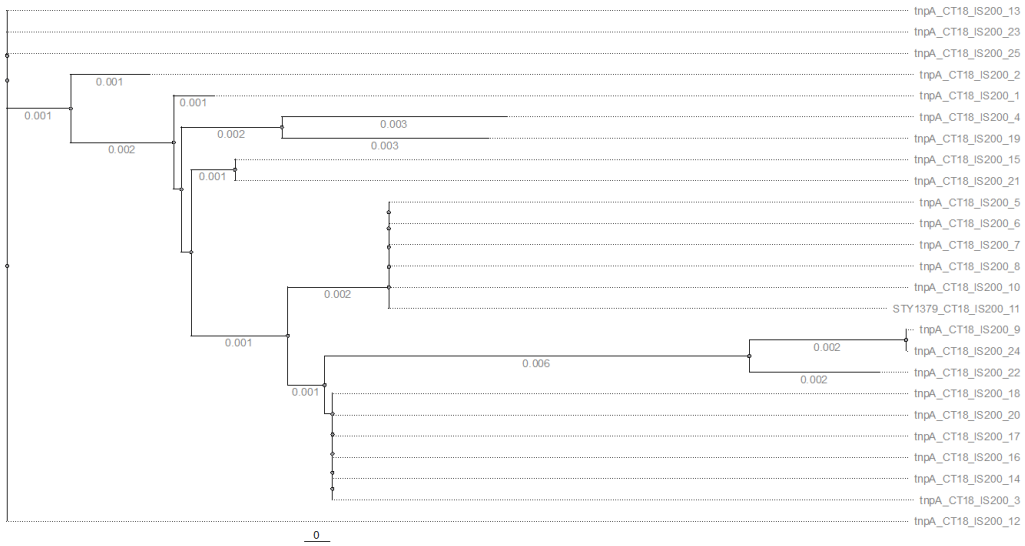


Figure 2.38: *CT18 Phylogenetic Tree. Figure shows the phylogenetic Tree in a Rectangular and Phylogram view, with tree settings set to high width and minimum height.*

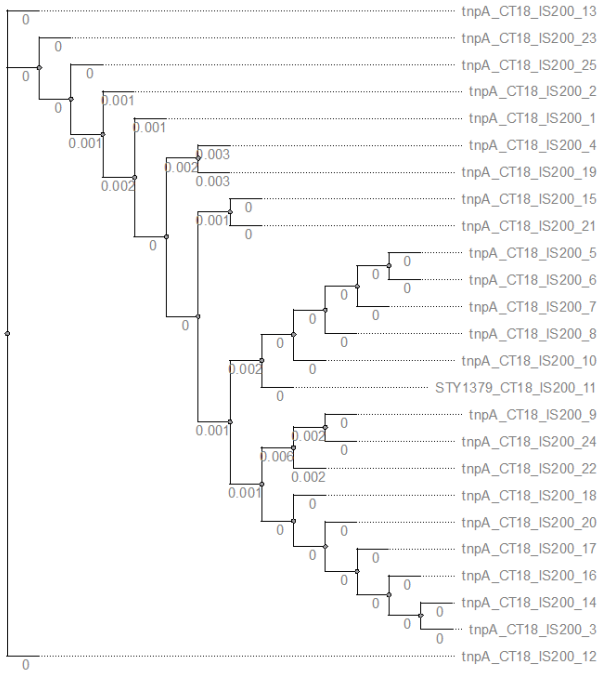


Figure 2.39: *CT18 Phylogenetic Tree. Figure shows the phylogenetic Tree in a Rectangular and Cladogram view, with tree settings set to high width and minimum height.*

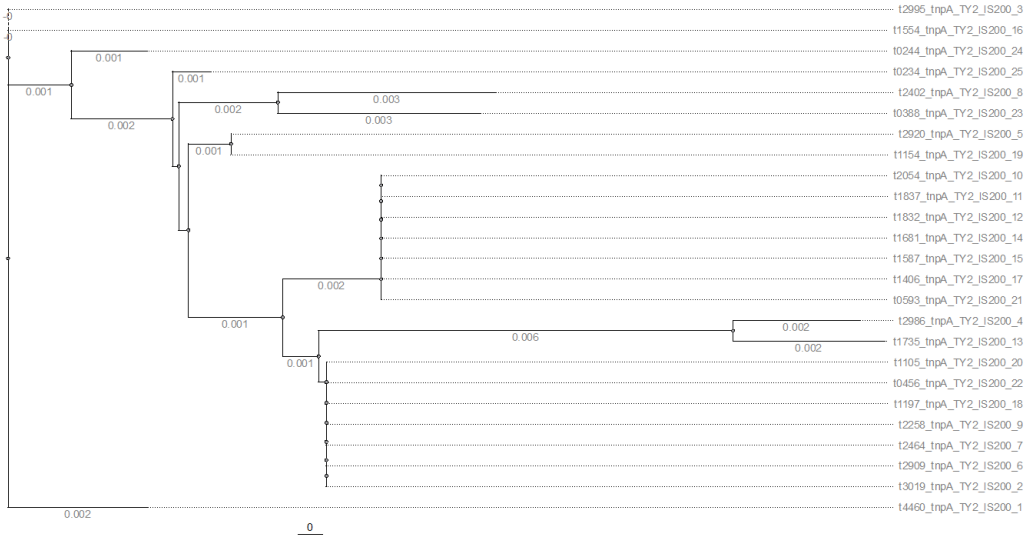


Figure 2.40: *TY2 Phylogenetic Tree. Figure shows the phylogenetic Tree in a Rectangular and Phylogram view, with tree settings set to minimum width and height.*

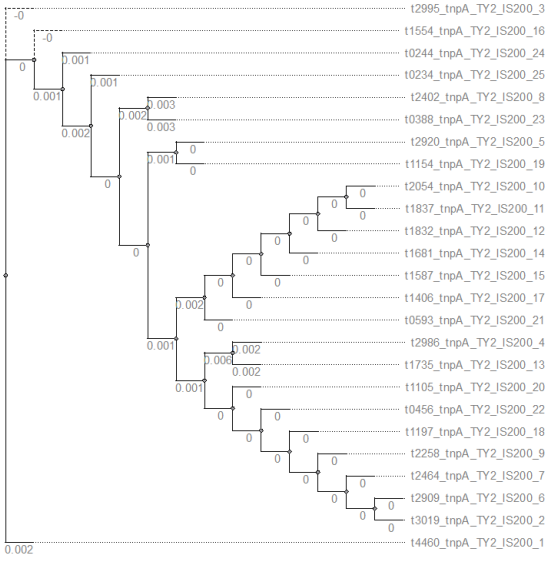


Figure 2.41: *TY2 Phylogenetic Tree. Figure shows the phylogenetic Tree in a Rectangular and Cladogram view, with tree settings set to minimum width and height.*

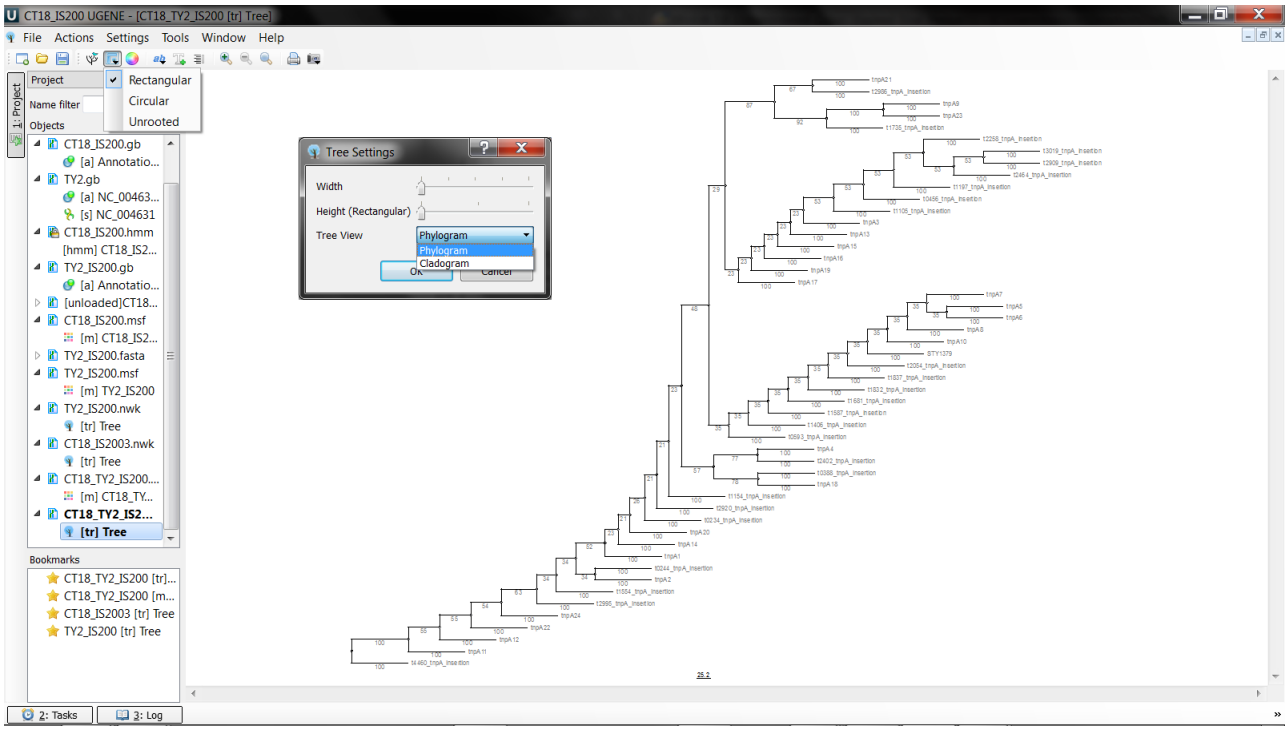


Figure 2.42: *CT18 and TY2 Phylogenetic Tree Configuration. There exist 3 tree layouts: Rectangular, Circular and Unrooted. The tree can also be set to be either in a Phylogram view or in a Cladogram view. The width defines the width that tree's leaves should have, while height defines the distance between the leaves.*

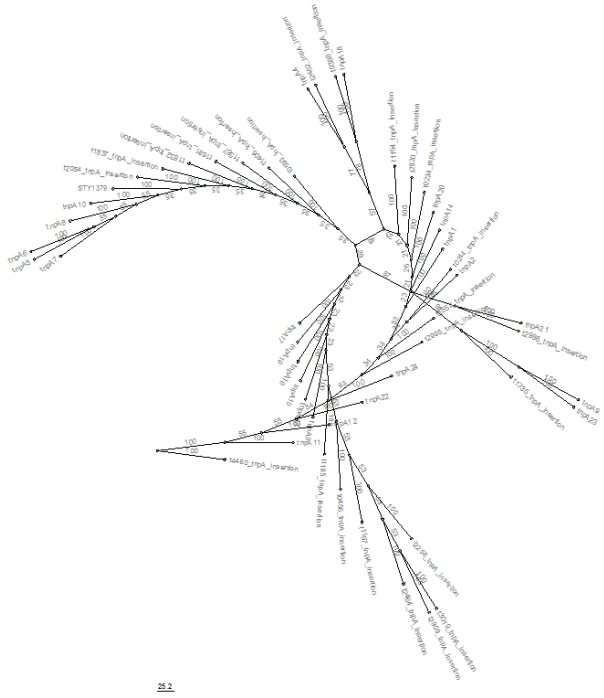


Figure 2.43: *CT18 and TY2 Phylogenetic Tree in an Unrooted Layout.*

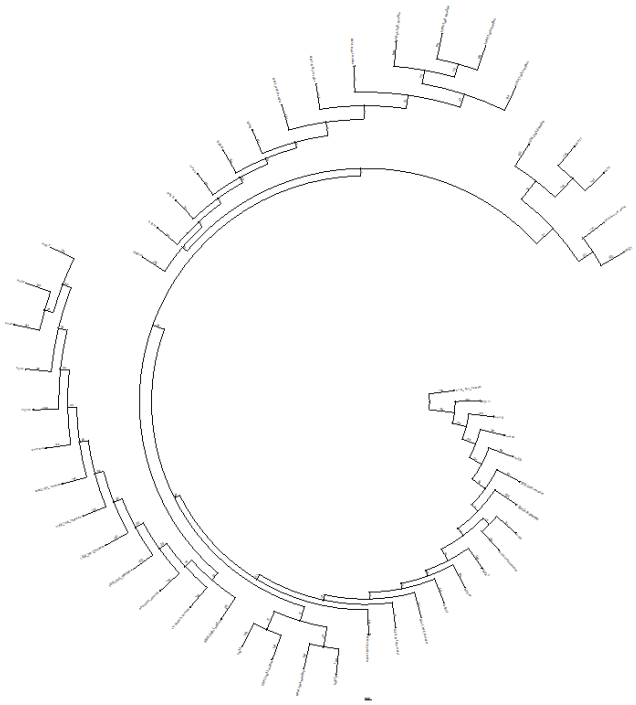


Figure 2.44: *CT18 and TY2 Phylogenetic Tree in a Circular Layout.*

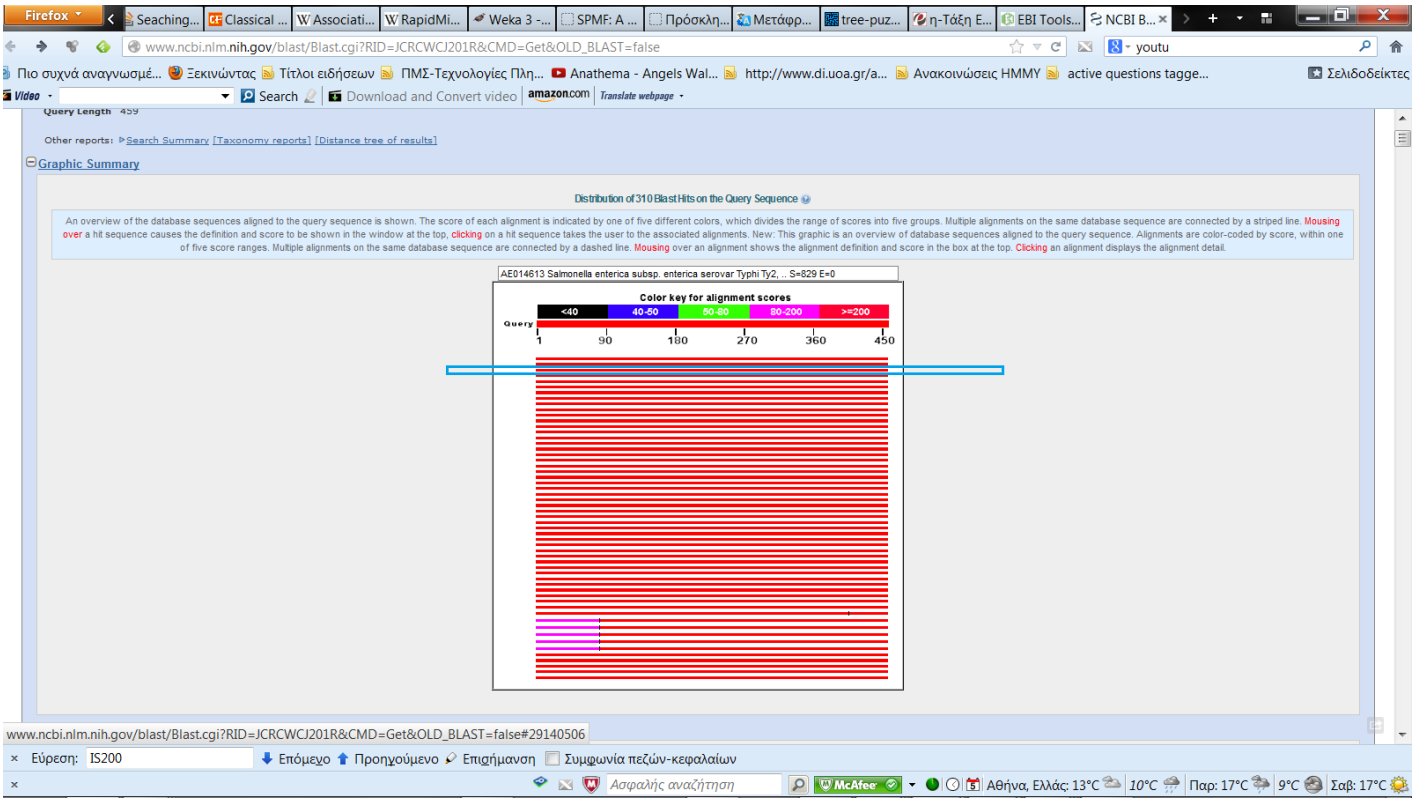


Figure 2.45: *BLAST result for querying the first CDS of S. Typhi CT18. This figure shows in highlighted in color the scores of several CDSs of other gene families. For example we have circled the sequence of S.Typhi TY2 which gives the score 829. More details can be shown in the upper caption of the image indicated.*

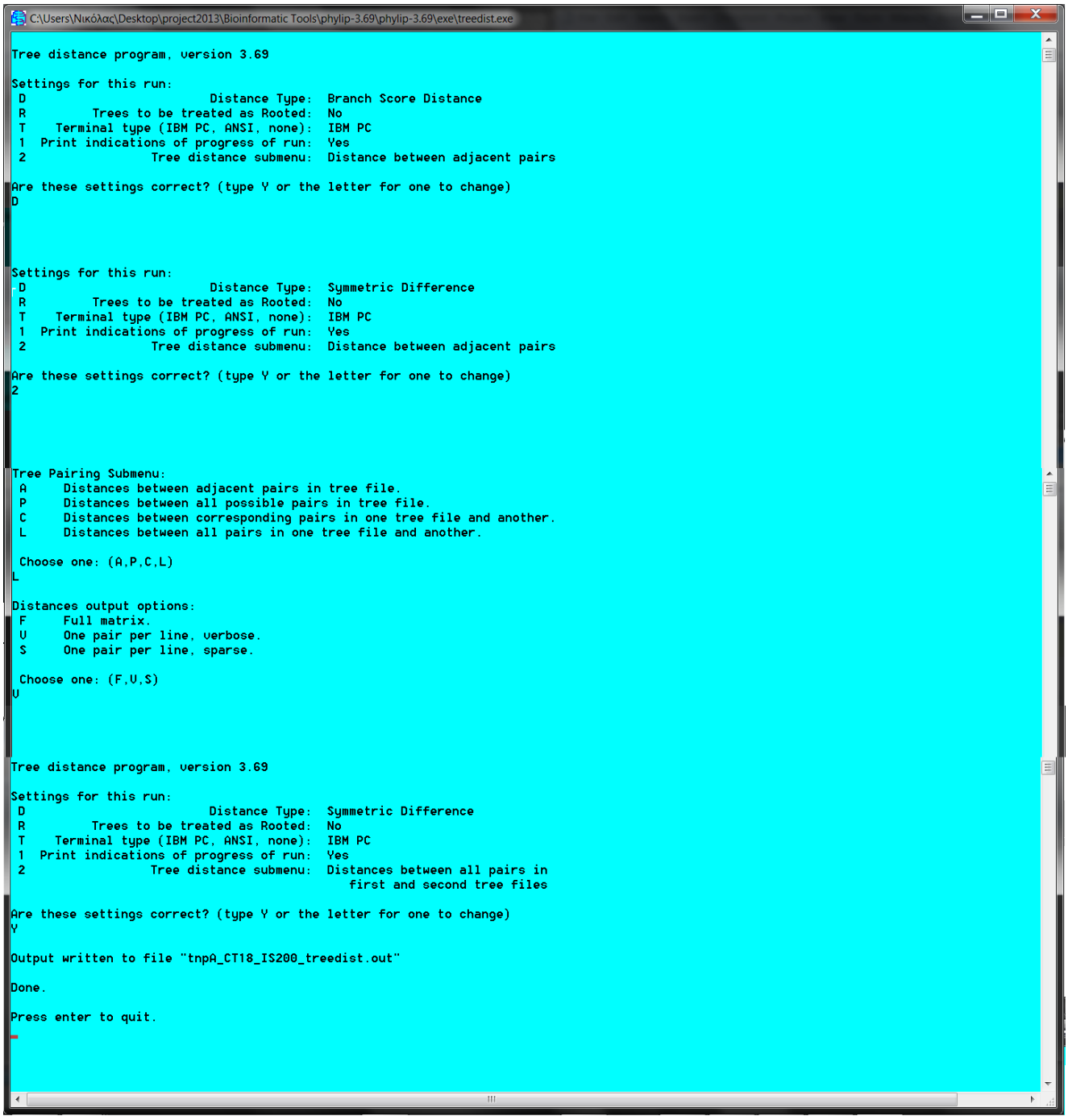


Figure 2.46: *PHYLIP's 'treedist' Tool customization*

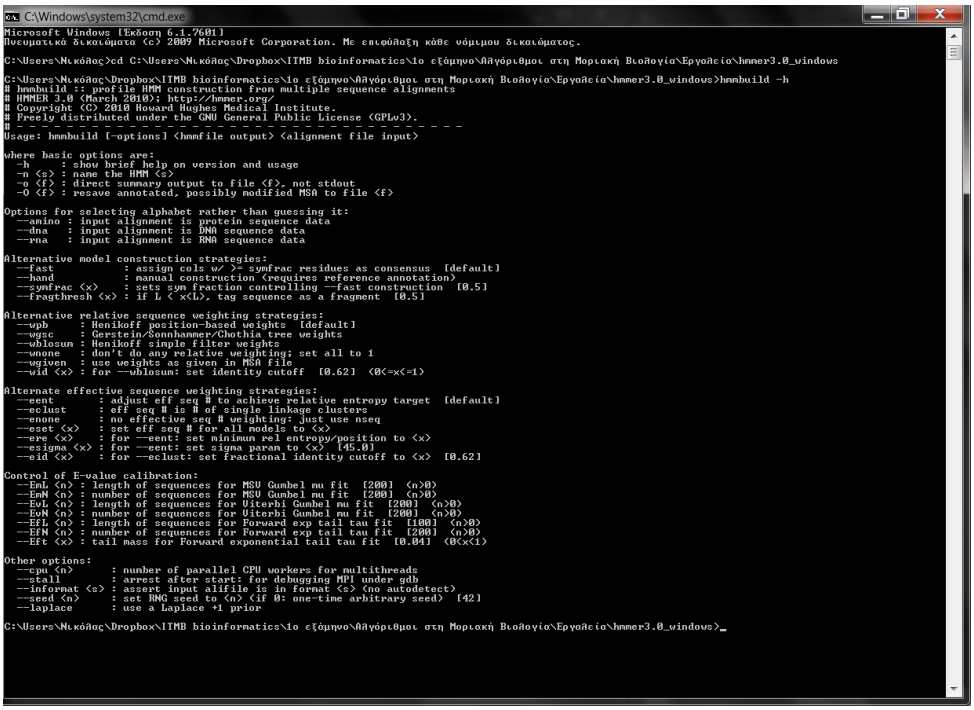


Figure 2.47: *HMMER v3.0 windows console - hmmbuild help*

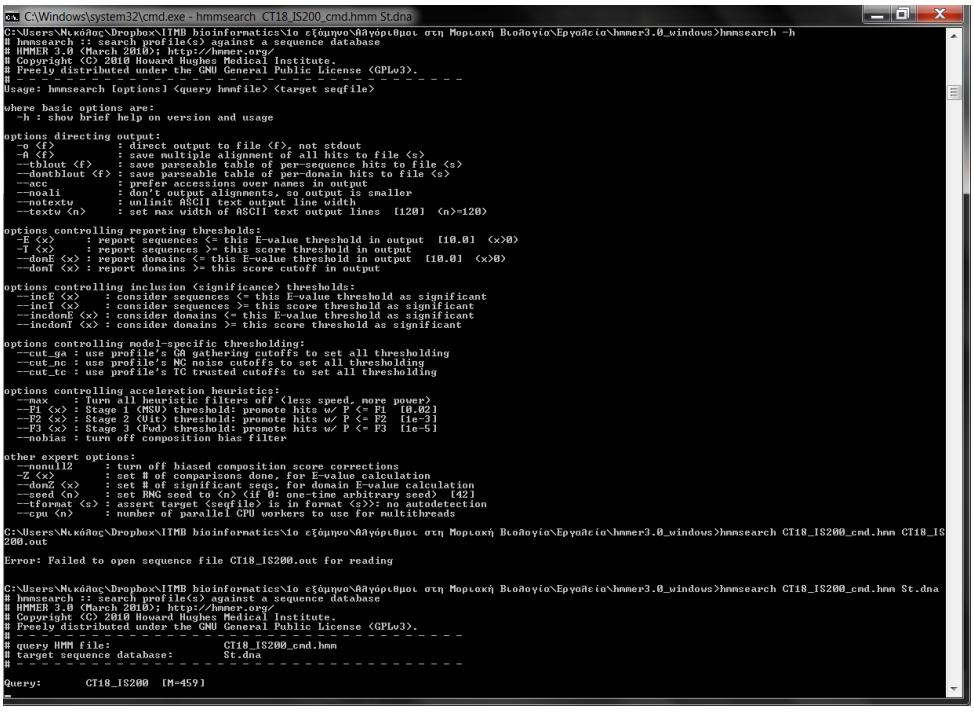


Figure 2.48: *HMMER v3.0 windows console - hmmsearh help*

Figure 2.49: *HMMER v3.0 windows console - hmmsearch output results*

CHAPTER 3

Conclusion

This year's project assignment covered everything we learned in class in an extensive way. Our class was assigned to make a full analysis of Salmonella Typhi str. CT18 genome for finding coding DNA sequences that when encoded the produce the IS200 transposase.

As we were induced, we followed a big path of consecutive steps, and in this way we learned in practice how to use and connect abstractly, yet also, specifically all the few data we possess in a in the optimum way.

To become more specifically, this report's topic was to use hidden markov models to gather all the information we are capable using only these. So, in brief, from the annotated genomes of S. Typhi CT18 and TY2 we found all the CDSs that produce IS200 transposase. Then, with these CDSs we created a Multiple Sequence Alignment, and with the MSAs we built the Hidden Markov Models. Subsequently, we 'fed' these HMMs in Artemis and found the statistically important predictions of HMMs. We noticed that all the statistically important annotations fully overlapped with the CDSs of the genuine genome and we added some features as proposed from the assignment.

Given the above CT18 score threshold we found the HMMs for TY2 and we then compared CT18's and TY's genomes with ACT. On the follow, we created a plylogenetic for IS200 and these two genomes and we concluded that we could not extract any significant conclusion, as these two genomes had almost all of their CDSs identical but in a different genome location.

In the next step of the assignment, we used the IS200 statistically significant sequence predictions that we found with HMMs and we 'fed' them all, one by one, in BLAST. BLAST found all the significant local aligned sequences with other organisms (bacteria mostly), and with theses sequences of BLAST's output we constructed 25 phylogenetic trees (numbered after the 25 HMM significant predictions) of 21 significant local alignment sequences each one of them. We then realised that again we could not separate the 'taxa' in a phylogenetic order, because all the BLAST sequence are alike HMM sequences that we noticed at the beginning of the report that we could not extract this kind of information.

Finally, we created a consensus tree, based on the 25 phylogenetic trees with PHYLIP and we tested it, and it was proved to have been constructed properly by the algorithm that implements this construction. In the end, as the last step of this report we were asked to

create a supertree showing the different organism where IS200 is produced. This final step came up to deliver new questions about 'taxa' and how useful may they become is the are used properly.

To sum up, I found pretty interesting this assignment, because except for learning how to use all these Bioinformatic Tools, I manager to categorize and classify all the semester's information, and even more I think I mined into the way of thinks in the 'bioinformatics' way.

In addition to the above, the miscellaneous chapter was created to place some figures that may be considered as more specialized than requested from the assignment.

Abbreviations - Acronyms

Abbreviations/Acronyms	Full Evolvent
TE	Transposase Element
IS	Insertion Sequence
S. Typhi	Salmonella Typhi
HMM	Hidden Markov Models
CDS	Coding DNA Sequence
EBI	European Bioiformatics institute
MSA	Multiple Sequence Alignment
FT	Table Feature
ACT	Artemis Comparison Tool
BLAST	Basic Local Alignment Sequence Tool
bp	base pairs
MD	Master Degree
ITMB	Information Technologies in Medicine and Biology

Bibliography

Bibliography

[1] Lowary, P. T., and J. Widom. "New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning." Journal of molecular biology 276.1 (1998): 19-42.

[2] Mahillon, Jacques, and Michael Chandler. "Insertion sequences." Microbiology and Molecular Biology Reviews 62.3 (1998): 725-774.

[3] Bisercic, M., and Howard Ochman. "Natural populations of Escherichia coli and Salmonella Typhimurium harbor the same classes of insertion sequences." Genetics 133.3 (1993): 449-454.

[4] McClelland, Michael, et al. "Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of Salmonella enterica that cause typhoid." Nature genetics 36.12 (2004): 1268-1274.

[5] Deng, Wen, et al. "Comparative genomics of Salmonella enterica serovar Typhi strains Ty2 and CT18." Journal of bacteriology 185.7 (2003): 2330-2337.

[6] Baker, Stephen, and Gordon Dougan. "The genome of Salmonella enterica serovar Typhi." Clinical Infectious Diseases 45.Supplement 1 (2007): S29-S33.

[7] Kathryn Holt. "Genomic variation and evolution of Salmonella enterica serovars Typhi and Paratyphi A" Sanger Institute (2009).