Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics
& Bioengineering (BIBE), Larnaca, Cyprus, 11-13 November 2012

GCVT – A Genome Comparison And Viewing Tool

Shakuntala Baichoo, Parinita Ujoodha, Yovish Bissessur
Department of Computer Science and Engineering
University of Mauritius
shakunb@uom.ac.mu

Abstract—**This paper highlights the features of a simple Genome Comparison and Viewing Tool (GCVT) for comparing bacterial genomes based on their genes, products, GC counts and genome sizes. For selected features from the results of comparison, a facility to perform alignment using the Needleman-Wunsch, Smith – Waterman and ClustalW algorithms is also provided. The software can take as input, genome files in a number of formats namely the Embl, Genbank and Fasta formats. GCVT also provides the facility for visualizing the relationship between chosen features using a guide tree. A number of additional features like the DNA to protein converter amongst many others contribute to ease the tasks of researchers. For the convenience of the users, the software can also take a protein file in the PDB format and display its three-dimensional structure. Compared to existing software like Artemis Comparison Tool (ACT), DNAVis and inGeno, GCVT is much simpler and user-friendly.**

*Keywords – **bioinformatics; bacterial genome; genome comparison; genome viewing.***

I. Introduction

There is a very large number of public domain databases covering a wide range of genomic data available to researchers in bioinformatics [1]. Genomic data are stored in several file formats namely Embl, Genbank or Fasta, amongst many others. Genomes of organisms can thus be compared and viewed using the annotated information stored in these files or the sequences.

A number of tools have been implemented to visualize/compare genomic data based on annotations, namely the Artemis Comparison Tool (ACT) [2], DNAVis [3] and inGeno [4]. ACT is implemented in Java and provides an interface for comparing a number of genome files but users must first provide a comparison file for pairs of genome files and this is not very intuitive to users. DNAVis is another tool implemented in C++ and provides a real-time visualization of DNA sequences and their comparative genome annotations. inGeno is an interactive visualization platform for sequence comparisons between complete genome sequences and all associated annotations and features.

Biologists increasingly need alignment methods that can handle long biological sequences, compare them and provide a rich visual display of the aligned sequences and the corresponding results. The most common types of alignment are the pairwise sequence alignment and the multiple sequence alignment. Some genome comparison tools provide comparison at the sequence level, e.g. Mauve Aligner [5]. Sequence alignment helps to identify regions of similarity between sequences [6]. Thus, the alignment of sequences is a crucial feature for doing Genome Comparison.

A tool doing genome comparison should be able to cater for both annotation-level and sequence-level comparison. Thus researchers in bioinformatics require a good genome comparison and viewing tool which will provide adequate facilities like annotation level comparison, gene identification, and specific feature extraction, circular and linear view of whole genome as well as sequence level comparison in the form of alignment. Hence, keeping all these features in mind, an integrated package, GCVT, has been designed and developed for the efficient viewing and comparison of genomes.

II. Implementation

GCVT is developed on the Java platform together with BioJava as primary library. Embl, Genbank and Fasta files are taken as input in the software for both comparison and viewing purposes. When performing file upload operation, several coding sequence features are extracted from the respective files (Only Embl and Genbank file format since FASTA file format is not annotated) and saved in a MYSQL database. For the time being, the software provides molecular biologists with the facility to compare up to five whole genome files. Basically, GCVT performs two types of comparison namely Annotation Level Comparison and Sequence Level Comparison. In Annotation Level Comparison, features, of uploaded files, can be compared by gene name, gene size, gene location, product description, product size and lastly product location.

On the other hand, in Sequence Level Comparison, the whole genome of the uploaded files can be compared by their genome size and GC Content. Smaller sequences like Protein and DNA sequences of specific common genes and products are compared using alignment algorithms namely Needleman Wunsch [7], Smith Waterman [8] and Clustal W [9]. The Needleman Wunsch algorithm has been used for end-to-end comparison of two sequences while the Clustal W algorithm has been used for multiple sequence alignment. The Smith Waterman algorithm has been used to perform local alignment so as to find the best local match between two sequences. The software gives a rich and colorful display of the alignment of the sequences and lists the percentage identity and other additional information about the alignment as shown in Figure 1.

GCVT provides different interactive displays of the annotated files such as (i) Genome Maps, (ii) Tree View Structure of features and (iii) Double Stranded Feature View

978-1-4673-4358-9/12/$31.00 ©2012 IEEE

of the genome. The **GView** library has been integrated in the software to facilitate the implementation of the Genome maps while the BioJava Library was used for the other types of views.

The integration of all these features makes the software an intuitive Genome Comparison and Viewing Tool whilst the algorithms used ensures the reliability of comparison and viewing results.

III. Results and Discussion

The main features of GCVT can be divided as follows (i) Genome Comparison (ii) Genome Viewing and lastly (iii) Additional Features.

A. *Genome Comparison*

GCVT is software that allows for both annotation level comparison and Sequence level comparison up to 5 genome files namely in Embl, Fasta and Genbank format.

- ***Annotation level Comparison***

As whole genome files (EMBL and GENBANK only) incorporate several coding sequence (CDS) features from which comparison can be made, biologists can therefore find similarities and differences of the uploaded genomes files without much higher level computations. For instance, a biologist can easily find the location of a common gene in each of the five uploaded files. The results of the comparison are provided in tabular format.

Since all the required information have being already extracted and saved in the MYSQL database in the file upload operation, the respective organism information is queried accordingly and input in a vector data structure for comparison purposes. Similarly, the process is repeated for all the genome files selected and the output vector will hold all the common genes present in all the selected files. For example, if three genome files have been selected for comparison, the first vector will contain all the common genes present in Genome File 1 and Genome File 2 together with some other additional information. The second vector will hold all the information of the Genome File 3. Then, the first vector will be compared with the second one to obtain all the common genes alongside with some other information too. The outcomes of the comparison are input in another final vector and displayed in a Jtable in GCVT as shown in Figure 2.

Moreover, the user is also allowed to open the results displayed in the Jtable in Microsoft excel where s/he can plot charts of selected values and save them accordingly. In addition, all or particular DNA and Protein Sequences can be saved to FASTA file format for future use.

- ***Sequence level Comparison***

GCVT allows the comparison of whole genome sequences by their GC count and their Genome sizes. The whole genome sequence is read from each file uploaded and the GC count and Genome size is computed. Based on the selection of the files, the GC count and Genome size is plotted and displayed to the user as show in Figure 3 and Figure 4 respectively.

GCVT also allows the user to view the detailed nucleotide content (A, T, C, and G) of a selected file in a bar chart. The whole genome sequence is read from the selected file and the percentage of Adenine, Thymine, Cytosine and Guanine is computed. If more bar charts are generated from selected files, the nucleotide contents of the selected files can be compared in more detail. For instance, Figure 5 shows the comparison of Mycobacterium Avium and Mycobacterium Leprae.

The JFree Chart Library has been integrated in the software to make the displays of the graphs more interactive for the users.

Sequence alignment is used for the comparison of smaller sequences. GCVT can perform both global and local alignment. The algorithms used for pairwise sequence alignment include the Needleman Wunsch algorithm for global alignment and Smith Waterman for local alignment. The software also provides the facility to compare the protein/DNA sequences of common genes or common products of the uploaded files by performing Multiple Sequence alignment. The ClustalW algorithm is used to perform multiple sequence alignment of both protein and DNA sequences. GCVT displays the alignment results in a user friendly interface as well as alignment details like the percentage identity, aligned length of the sequences amongst others. The alignment of the sequences of the common gene rpsF can be shown in Figure 6.

A guide tree has also been used as a method of comparison to show which sequences are closer. The results of the multiple sequence alignment of the sequences are stored in the guide tree. GCVT writes the respective results in the **newick** format which is interpreted by the TreeViewX software and used to display the tree. The user is also provided with the facility to save the tree for future use. Figure 7 shows the guide tree of the leus common gene found in mycobacterium_avium file, mycobacterium_bovis file, mycobacterium_leprae file and mycobacterium_tuberculosis file.

B. *Genome Viewing*

GCVT provides different means of viewing the features of the genomes namely (i) Viewing annotations and sequences in a tree view, (ii) Viewing the whole genome in a linear or circular view and (iii) Double stranded feature view of the DNA sequence.

The Tree View feature is a user friendly interface which provides biologists the facility to view specific features in the selected file. Details like the Accession ID, Organism name amongst others are displayed to the user as shown in Figure 8. The Tree View has been implemented using the BioJava Library.

Linear and Circular maps have been provided to the user so that they can easily navigate across the genomes. GCVT displays the location and name of the gene when the user places the cursor at a specific location. Coding and Non Coding regions are also demonstrated in the maps. For example the red lines in Figure 9 are coding regions and white spaces are non coding regions. Figure 9 and 10 display the Linear and Circular view of mycobacterium_avium and mycobacterium_leprae respectively. The user can also save the maps in .jpg,.png,.svg and.svgz format for future use. The Gview Library components have been used to construct interactive genome maps for the users.

The double stranded feature view of DNA is useful in viewing genes on both sides of the DNA. In GCVT, the genes in the positive and negative strand are filtered, labeled and plotted linearly to give a rich and colorful display of the genes in the genome as shown in Figure 11.

C. Additional Features

GCVT provides a number of useful features, namely the:

- ***EMBL and FASTA file download***

GCVT allows the download of EMBL and FASTA files from EBI website. The user just has to enter a valid accession id and the file will be downloaded into the respective downloaded files directory. Once downloaded, these files can be further used as input for comparison in GCVT.

- ***Search for a specific feature***

Moreover, users can also search for specific features like gene or product from the uploaded EMBL or GENBANK files to get more detailed information about the respective features. The selected organism table is queried from the database to display the appropriate results in a tabular format. These results can be saved in excel format and the protein or DNA sequences can be exported to the FASTA format for further use.

- ***GC count Variation per file***

GCVT allows the user to view the GC count variation of the genome of the selected file. The genome is split into 1000 intervals and the GC count of each interval is calculated and plotted in a graph. Figure 12 shows the GC count variation of mycobacterium_avium genome. As it can be seen, high peaks demonstrate high GC count. Regions that show huge variations (an instant fall in the peak) can contain potential foreign genes.

- ***3D structure of protein***

GCVT can display the three-dimensional structure of a protein given the PDB file as input. The PDB reader from BioJava Library has been used to read the PDB file and the Jmol library has been used to display the protein structure in three dimensions. This feature can be useful to show the 3D arrangement of the amino acids within a protein.

- ***DNA to protein Converter***

GCVT provides the facility to convert DNA sequences to protein sequences. A DNA sequence and an organism are taken as input to convert the corresponding DNA to a protein sequence. The IUPAC codon table has been used to do the mapping of the transcribed RNA to protein. Using this feature, a user is does not have to manually transcribed the DNA to RNA and then translate it to protein.

- ***Protein download***

Users can also download protein sequences from UNIPROT website given a valid protein accession id. Once the download is over, the latter can save the downloaded protein sequences in FASTA file format for future use.

- ***Web service***

GCVT uses the NCBI BLAST web service to allow the user to BLAST protein or nucleotide sequences. A FASTA file either DNA or protein is taken as input and the corresponding blast program either blast p or blast n is input as parameter. Once the operation is successful, the results are displayed to the user and the latter can save them for future use. This feature is useful if ever, the user wants to get more details about specific DNA or protein sequences.

- ***Visualisation Tool***

GCVT has embedded external software like **DNA plotter** and **Archaeopteryx** to provide additional functionalities in the software. For example, Archaeopteryx can be used to manipulate phylogenic trees.

IV. Merits and limitations of GCVT

The merits of GCVT are the ease of searching and viewing the annotations of the uploaded files, the user-friendly interface for viewing of genomes, the comparison of up to 5 whole genome files. Moreover, the additional features contribute to the enhancement of the software and make it easy for non-computer literate researchers to use. Since GCVT has been implemented in Java, it is platform-independent but it was only tested on Windows and Unix. Most software segregates comparison and viewing while GCVT integrates both aspects. The limits of GCVT are that it requires some software to be preinstalled to be able to run correctly, for example MYSQL, TreeViewX amongst others. However a user manual has been provided to explain the users how to perform specific operations.

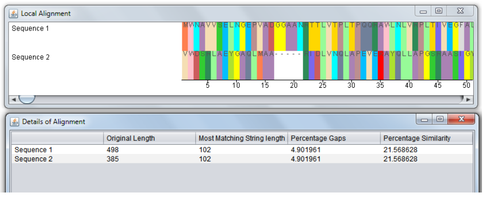
V. Conclusion

GCVT is a user-friendly software that complements the work of biologists in the sub areas of Comparative Genomics and Genome Viewing. As long as new organisms will be discovered and their genomes sequenced, the software will continue to be very useful in terms of genome comparison and viewing. Thus, it will help biologists to have a better insight in biological research.

ACKNOWLEDGMENT

The authors wish to thank Prof. Y.Jaufeerally-Fakim from Faculty of Agriculture, University of Mauritius for her constructive advice on this work. A special thanks goes to the BioJava team for creating such a powerful and yet easy to use library.

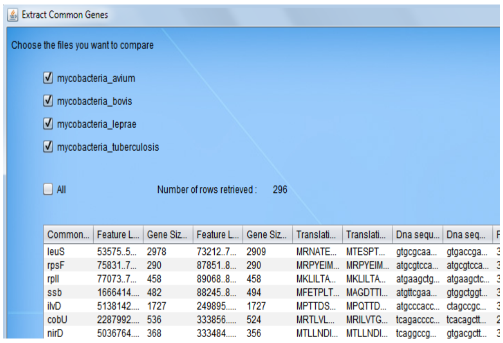
FIGURES



```
Sequence 1   MWNAVVSELNGEPVADGGAANRTTLVTPLTPQQRAWLNLVRPLTI VEGFAL
Sequence 2   VWDSRLAEYGAQLMAA-----RI DLVNQLAPEVERAYQLLAPGSRAASI GY
                 5    10   15   20   25   30   35   40   45   50
```

	Original Length	Most Matching String length	Percentage Gaps	Percentage Similarity
Sequence 1	498	102	4.901961	21.568628
Sequence 2	385	102	4.901961	21.568628

Figure 1. Pairwise sequence alignment results



Choose the files you want to compare

☑ mycobacteria_avium

☑ mycobacteria_bovis

☑ mycobacteria_leprae

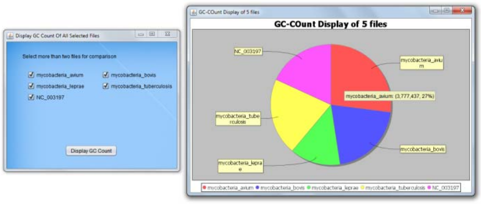
☑ mycobacteria_tuberculosis

□ All

Number of rows retrieved : 296

Common...	Feature L...	Gene Siz...	Feature L...	Gene Siz...	Translati...	Translati...	Dna sequ...	Dna seq...	F
leuS	53575..5...	2978	73212..7...	2909	MRNATE...	MTESPT...	gtgcgcaa...	gtgaccga...	3
rpsF	75831..7...	290	87851..8...	290	MRPYEIM...	MRPYEIM...	atgcgtcca...	atgcgtcca...	3
rplI	77073..7...	458	89068..8...	458	MKLILTA...	MKLILTA...	atgaagctg...	atgaagctc...	3
ssb	1666414....	482	88245..8...	494	MFETPLT...	MAGDTTI...	atgttcgaa...	gtggctggt...	3
ilvD	5138142....	1727	249895.....	1727	MPTTDS...	MPQTTD...	atgcccacc...	ctagccgc...	3
cobU	2287992....	536	333856.....	524	MRTLVL...	MRILVTG...	tcagacccc...	tcacagctt...	2
nirD	5036764....	368	333484.....	356	MTLLNDI...	MTLLNDI...	tcaggccg...	gtgacgctt...	3

Figure 2. Common genes together with additional information present in Mycobacteria Avium , Mycobacteria Bovis , Mycobacteria leprae and Mycobacteria tuberculosis


Display GC Count Of All Selected Files
GC-COunt Display of 5 files
GC-COunt Display of 5 files
NC_003197
mycobacteria_avium
mycobacteria_avium: (3,777,437, 27%)
mycobacteria_tuberculosis
mycobacteria_bovis
mycobacteria_leprae
● mycobacteria_avium ● mycobacteria_bovis ● mycobacteria_leprae ● mycobacteria_tuberculosis ● NC_003197
Display GC Count

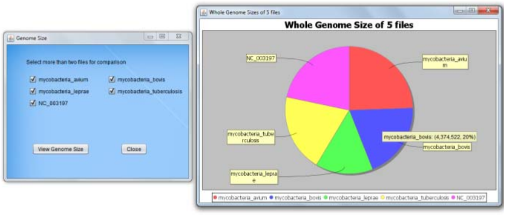
Select more than two files for comparison

☑ mycobacteria_avium ☑ mycobacteria_bovis

☑ mycobacteria_leprae ☑ mycobacteria_tuberculosis

☑ NC_003197

Figure 3. GC Count of Mycobacteria Avium , Mycobacteria Bovis , Mycobacteria leprae, Mycobacteria tuberculosis and nc_003197


Genome Size
Whole Genome Sizes of 5 files
Whole Genome Size of 5 files
NC_003197
mycobacteria_avium
mycobacteria_tuberculosis
mycobacteria_bovis: (4,374,522, 20%)
mycobacteria_bovis
mycobacteria_leprae
● mycobacteria_avium ● mycobacteria_bovis ● mycobacteria_leprae ● mycobacteria_tuberculosis ● NC_003197
View Genome Size
Close

Select more than two files for comparison

☑ mycobacteria_avium ☑ mycobacteria_bovis

☑ mycobacteria_leprae ☑ mycobacteria_tuberculosis

☑ NC_003197

Figure 4. Genome size of Mycobacteria Avium , Mycobacteria Bovis , Mycobacteria leprae, Mycobacteria tuberculosis and nc_003197

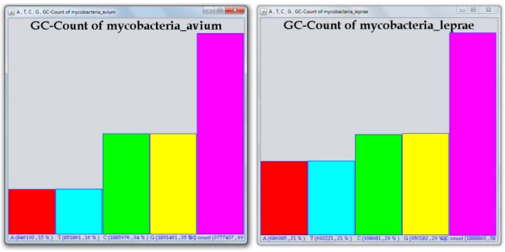
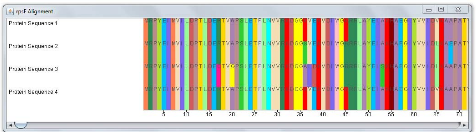


Figure 5. Nucleotide contents of Mycobacteria Avium and Mycobacteria leprae



```
Protein Sequence 1   MRPYEIMVILDPTLDERTVAPSLETFLNVVRKDGGKVEKVDIWGRRRLAYEIAKHAEGIYVVIDVKAAPATY
Protein Sequence 2   MRPYEIMVILDPTLDERTVAPSLETFLNVVRKDGGSVEKVDIWGRRRLAYEIAKHAEGIYVVIDLKAEPATY
Protein Sequence 3   MRPYEIMVILDPTLDECTVGPSLETFLNVVRKDGGATDKVDIWGRRRLAYEISKHAEGIYVVIDLKAAPATY
Protein Sequence 4   MRPYEIMVILDPTLDERTVAPSLETFLNVVRKDGGKVEKVDIWGRRRLAYEIAKHAEGIYVVIDVKAAPATY
                         5    10   15   20   25   30   35   40   45   50   55   60   65   70
```

Figure 6. multiple sequence alignment of the common rpsF Gene

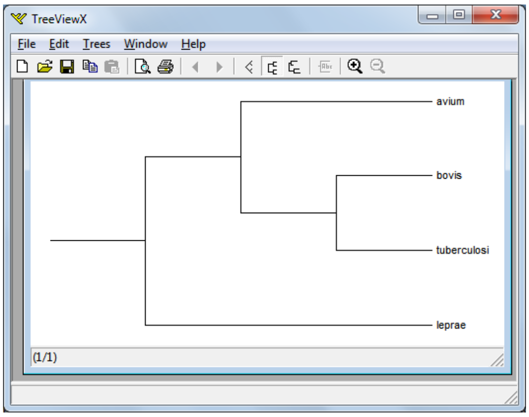
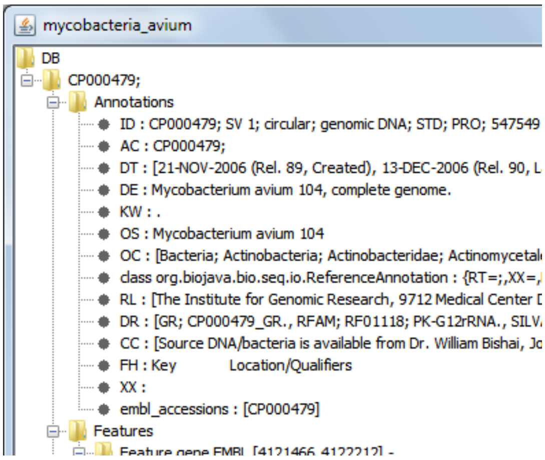
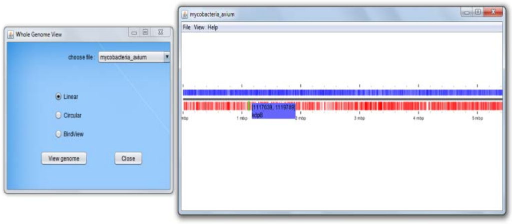


Figure 7. Guide tree of leuS common gene



```
DB
CP000479;
  Annotations
    ID : CP000479; SV 1; circular; genomic DNA; STD; PRO; 547549
    AC : CP000479;
    DT : [21-NOV-2006 (Rel. 89, Created), 13-DEC-2006 (Rel. 90, L
    DE : Mycobacterium avium 104, complete genome.
    KW : .
    OS : Mycobacterium avium 104
    OC : [Bacteria; Actinobacteria; Actinobacteridae; Actinomycetal
    class org.biojava.bio.seq.io.ReferenceAnnotation : {RT=;,XX=,
    RL : [The Institute for Genomic Research, 9712 Medical Center D
    DR : [GR; CP000479_GR., RFAM; RF01118; PK-G12rRNA., SILV
    CC : [Source DNA/bacteria is available from Dr. William Bishai, Jo
    FH : Key            Location/Qualifiers
    XX :
    embl_accessions : [CP000479]
  Features
    Feature gene EMBL [4121466 4122212] -
```

Figure 8. Tree View of Mycobacteria Avium



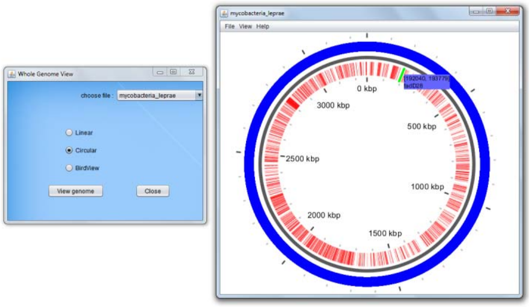
choose file : mycobacteria_avium

● Linear

○ Circular

○ BirdView

Figure 9. Linear View of Mycobacteria Avium file



Whole Genome View

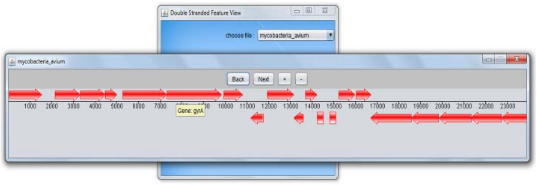
choose file : mycobacteria_leprae

○ Linear

● Circular

○ BirdView

Figure 10. Circular View of Mycobacteria laprae file



Double Stranded Feature View

choose file : mycobacteria_avium

Figure 11. Double Stranded Feature View of the selected Mycobacteria Avium file

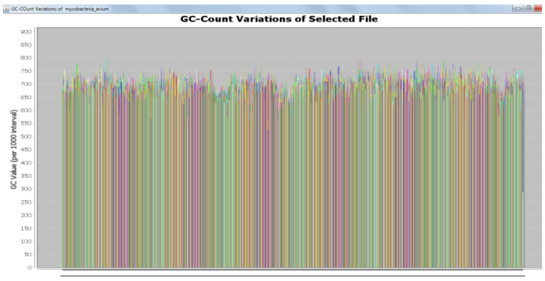


Figure 12. GC count variation in Mycobacteria Avium

REFERENCES

[1] Michael Brudno, Sanket Malde, Alexander Poliakov, Chuong B. Do1, Olivier Couronne, Inna Dubchak and Serafim , “Glocal alignment: finding rearrangements during alignment”, Oxford Journals, Bioinformatics (2003) 19 (suppl 1): i54-i62.

[2] Needleman, S. B. & Wunsch, C. D. (1970), “A General Method applicable to the search for similarities in the amino acid sequence of two proteins". Journal of Molecular Biology 48 (3): 443–453”.

[3] Smith, Temple F.; and Waterman, Michael S. (1981). "Identification of Common Molecular Subsequences", Journal of Molecular Biology, 147: 195–197.

[4] Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007). "ClustalW and ClustalX version 2". Bioinformatics 23 (21): 2947–2948

[5] Pocock M, Down T, Hubbard T., 2000, “BioJava: Open Source Components for Bioinformatics”, ACM SIGBIO Newsletter 20(2), 10-12.

[6] Zvelebil M., Baum J.O., “Understanding Bioinformatics”, Garland Science, ISBN 0-8153-4024-9, 2008

[7] Tim Carver, Matthew Berriman, Adrian Tivey, Chinmay Patel, Ulrike Böhme, Barclay G. Barrell, Julian Parkhill and Marie-Ad`ele Rajandream, “Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database”, BIOINFORMATICS APPLICATIONS NOTE, Bioinformatics (2008) Vol. 24 no. 23, pages 2672–2676

[8] Mark W. E. J. Fiers, Huub van de Wetering, Tim H. J. M. Peeters, Jarke J. van Wijk and Jan-Peter Nap, “DNAVis: interactive visualization of comparative genome annotations”, BIOINFORMATICS APPLICATIONS NOTE, Vol. 22 no. 3 2006, pages 354–355

[9] Thompson JD, Higgins DG, Gibson TJ (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice". Nucleic Acids Res 22 (22)

Sample Genome Files used for testing from EBI:

1. Mycobacterium avium 104, complete genome: CP000479 (Embl format)
2. Mycobacterium bovis BCG Pasteur 1173P2, complete genome: AM408590 (Embl format)
3. Mycobacterium leprae Br4923, complete genome sequence: FM211192 (Embl format)
4. Mycobacterium tuberculosis H37Ra, complete genome: CP000611 (Embl format)
5. Salmonella enterica subsp. enterica serovar Typhimurium str. LT2, complete genome: NC_003197 (Genbank format)