BMC Bioinformatics

BioMed Central

Software

GeneViTo: Visualizing gene-product functional and structural features in genomic datasets

Georgios S Vernikos†, Christos G Gkogkas†, Vasilis J Promponas and Stavros J Hamodrakas*

Address: Department of Cell Biology and Biophysics, Faculty of Biology, University of Athens, Panepistimiopolis, 15701, Athens, Greece

Email: Georgios S Vernikos - gverniko@yahoo.gr; Christos G Gkogkas - mubiol@yahoo.com; Vasilis J Promponas - vprobon@biol.uoa.gr; Stavros J Hamodrakas* - shamodr@cc.uoa.gr

* Corresponding author †Equal contributors

Published: 31 October 2003

BMC Bioinformatics 2003, **4**:53

Received: 22 July 2003

Accepted: 31 October 2003

This article is available from: http://www.biomedcentral.com/1471-2105/4/53

Abstract

Background: The availability of increasing amounts of sequence data from completely sequenced genomes boosts the development of new computational methods for automated genome annotation and comparative genomics. Therefore, there is a need for tools that facilitate the visualization of raw data and results produced by bioinformatics analysis, providing new means for interactive genome exploration. Visual inspection can be used as a basis to assess the quality of various analysis algorithms and to aid in-depth genomic studies.

Results: GeneViTo is a JAVA-based computer application that serves as a workbench for genome-wide analysis through visual interaction. The application deals with various experimental information concerning both DNA and protein sequences (derived from public sequence databases or proprietary data sources) and meta-data obtained by various prediction algorithms, classification schemes or user-defined features. Interaction with a Graphical User Interface (GUI) allows easy extraction of genomic and proteomic data referring to the sequence itself, sequence features, or general structural and functional features. Emphasis is laid on the potential comparison between annotation and prediction data in order to offer a supplement to the provided information, especially in cases of "poor" annotation, or an evaluation of available predictions. Moreover, desired information can be output in high quality JPEG image files for further elaboration and scientific use. A compilation of properly formatted GeneViTo input data for demonstration is available to interested readers for two completely sequenced prokaryotes, *Chlamydia trachomatis* and *Methanococcus jannaschii.*

Conclusions: GeneViTo offers an inspectional view of genomic functional elements, concerning data stemming both from database annotation and analysis tools for an overall analysis of existing genomes. The application is compatible with Linux or Windows ME-2000-XP operating systems, provided that the appropriate Java Runtime Environment is already installed in the system.

Background

The impressive progress in Molecular Biology, enhanced by the development of rapid genome sequencing technologies, led to an exponential growth of the number of available DNA/protein sequences deposited in public databases. Between the early 90's, when the Human

Genome Project began, and 1996 the complete genome sequences of 5 unicellular organisms had been determined. By the time of this writing (September 2003) 160 genomes (including the Human Genome) have been completely sequenced, while 643 genome projects are still in progress [1,2]. On the other hand, the intensive research activity in the field of Bioinformatics generates a large amount of heterogeneous meta-data which, examined on a large-scale, demand further analysis in order to extract valuable biological information.

DNA or protein sequence retrieval from specialized curated databases (GenBank [3], SWISS-PROT [4]) is quite effective, by the means of well-established tools, such as SRS [5] or Entrez [6]. Cross-references between entries from disseminated biological databases are abundant, helping for easy navigation over the World Wide Web, but are unable to offer an overview of the way sequence features are distributed in ordered sequence sets, such as complete genomes.

Moreover, several bioinformatics analysis and prediction tools are available, either as web services or as standalone applications, attempting to give further insight to existing sequence information. These tools produce different output, according to the analysis type, and results representation is mainly oriented towards a per functional element basis. These analyses complement experimental data and guide further research activities.

Once information concerning a genome is obtained (sequence, annotation and meta-data), an integration step is required in order to come to advanced biological conclusions. Such a task is time-consuming and painstaking, as long as data for hundreds/thousands of sequences are "thick-set" in structured text files. Furthermore, the monotonous machine-readable file format does not reveal at once features contained in a set of sequences in an intuitive way. It becomes quite clear, especially in cases of completely sequenced genomes, that organizing data in text files constitutes only a primordial level of presentation. Thus, a more sophisticated approach for easier, efficient, more productive and less chaotic representation is required.

Data visualization, using specialized Computer Graphics Software, act as an intermediate link between raw data and the user for more effective and elaborate manipulation of numerous genomes. Such computational workbenches become even more useful when they incorporate, apart from the already deposited data, additional tools, making large-scale *in silico* experiments easier.

Several powerful genome visualization tools are already available, mainly focused on features related to nucleotide sequences: gff2ps [7], Artemis [8], SeqVista [9], NCBI Map Viewer [10], TIGR Genome browser [11], ENSEMBL project viewer [12], ERGO™ [13]. Each of these methods follows a different philosophy in the type of input data (e.g. sequences, maps, nucleotide sequence features), the accepted formats and the way that features are visualized. Our approach is mainly focused in presenting features related to gene products and their distribution along genomic regions.

We have developed GeneViTo, a JAVA-based computer application to incorporate in a single depiction sequence features existing in annotation records from nucleotide and protein sequence databases (GenBank, SWISS-PROT) and prediction methods output (e.g. PRED-CLASS [14], PRED-TMR2 [15], orienTM [16], SIGNALP [17]).

GeneViTo provides interfaces to additional analysis tools, as well as several search utilities, to easily manipulate and further examine sequenced genomes. Existing annotations may easily be extracted with a mouse click on the color boxes representing structural genes (protein coding regions or functional RNA products).

The GeneViTo working environment attempts to unify the representation of data from genome-proteome resources and bioinformatics meta-data scattered around the Internet. The scope is to achieve a holistic yet detailed image of an entire organism and help with genome annotation, phylogenetic studies and comparative genomics efforts.

The open system architecture combined with object oriented JAVA programming allows future incorporation of numerous bioinformatics tools with various outputs, thus alleviating the need to develop program-specific visualization software for a given biological task.

Implementation

Java-based Object Oriented Design

GeneViTo has been fully implemented in the Java programming language, exploiting its implicit object oriented design and graphical representation capabilities. Two main Java classes (MyFrame and MyPanel) serve as the core modules for data depiction and user interaction respectively, whereas additional classes are used for efficient data handling utilities (such as required transformations, raw data processing and analysis). Object oriented software design enables for ease of code reusability and provides interfaces for linking new software (or software modules) to build systems of extended capabilities.

Furthermore, the Java programming language has become very popular among programmers for its expressive power and elegance and, at the same time, very attractive to users for reasons of portability. Standard Java interpreters do

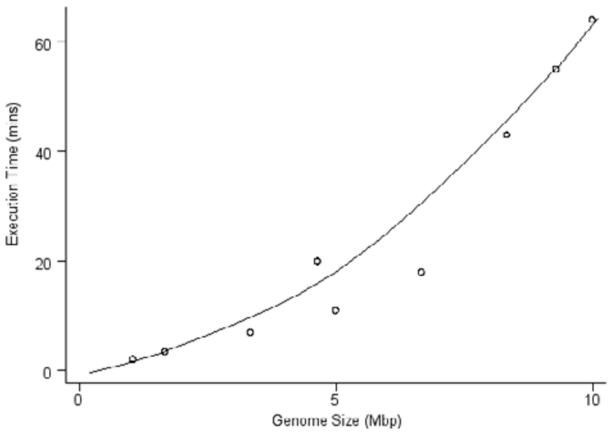


Figure 1
Execution Time of data transformation against Genome Size. This drawing represents the relationship between execution time and genome size. The test has been performed for the genomes of *C. trachomatis* (genome size 1,042,519 bp), *M. jannaschii* (1,664,970 bp), *E. coli* (4,639,221 bp) and six artificially constructed genomes (3,329,940 bp, 4,994,910 bp, 6,659,880 bp, 8,324,850 bp, 9,278,442 bp and 9,989,820 bp) respectively.

not build architecture-specific binaries but bytecode interpreted by the Java Virtual Machine, which makes Java applications, conceptually, cross-platform.

Performance Benchmark Tests

To assess the performance of GeneViTo, we have run several tests to address computational issues, such as memory usage. A summary of these tests is provided in Table 1. Memory usage (in the worst case 80 MBytes of main memory) is not a limiting factor, since it not correlated with the displayed genome's size (data not shown). Figure 1 as well as data available in Table 1 also demonstrate that, in practice, execution time has to be considered only in the stage of initial data transformation, which is only held once for each genome data-set. Nevertheless, a small but not substantial increase in time is expected when the number of genomic elements whose properties have to be incorporated into GeneViTo increases (data not shown). Loading genomic data is quite fast, requiring just a few seconds (Table 1).

Concerning the visualization process, GeneViTo uses true color (a minimum of 256 colors suffices for old graphics cards) and genome browsing is very fast even on typical desktop computers. Graphics displayed on screen can be saved in high-quality JPEG graphics files. The resolution of these snapshots is only limited by the screen resolution.

Table 1: Performance Benchmark Tests

Task	Main Memory Usage (Mbytes)		Execution Time
	Minimum	Maximum	
Preparing a Genome	10.5	31.0	See Figure 1
Loading a Genome	30.0	80.0	1–5 sec
Browsing	30.0	80.0	Instant

Performance tests used to collect the data presented in Table 1 and Figure 1 have been executed on a personal computer with an Intel Celeron 2 GHz processor and 256 MB of main memory.

Input Description

GeneViTo requires data to be specially formatted in order to display all desired information (Figure 2). The application offers an internal automatic procedure to transform available data into GeneViTo native format, which is mainly tab-delimited plain text and XML, and store them in a user-defined directory on the local hard drive. Related genomic data are extracted from GenBank (e.g. .ptt, .fnn files), while proteomic data (sequence, sequence related features) are derived from the EBI Proteome Analysis Server [18] (in SWISS-PROT format). Prediction concerning the number and topology of helical transmembrane segments stem from PRED-TMR2 and orienTM respectively, while predicted protein structural classes originate from the PRED-CLASS algorithm. The SIGNALP server provides prediction on protein signal peptides. Restriction enzyme sites are predicted by Webcutter [19], while COGs database [20] offers information about orthologous gene groups. All the aforementioned web resources (see Table 2) are freely available for public use, but are not limiting GeneViTo's functionality; in practice, any properly formatted user-defined prediction data can be used.

The original data files are necessary only for the initial formatting step and all essential information is included in the special files created by this procedure. More details about supported file formats, data gathering and organization, are provided through the application's detailed "Help" utility.

Results and Discussion

GUI description

The main window of the application contains 5 panels and a circular map. The "Central Graphics Panel" (Figure 3) contains a graphic display of the selected genome, in which genes are represented by colored boxes placed on a thin black line, marked with the respective length subdivisions (in bp) of the genome. The location and width of each colored box represents the relative position of the

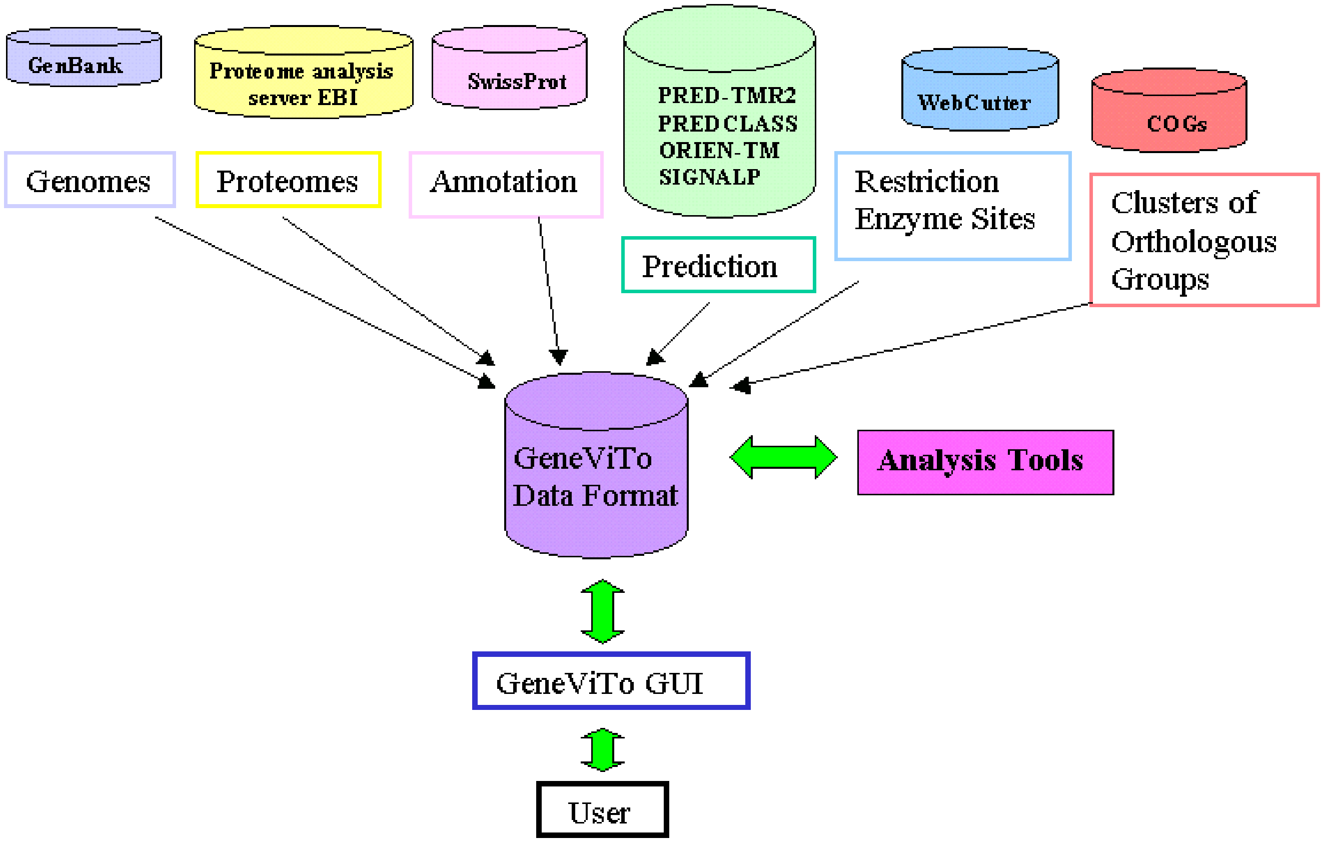


Figure 2
GeneViTo system flowchart.

Table 2: Availability of GeneViTo input data.

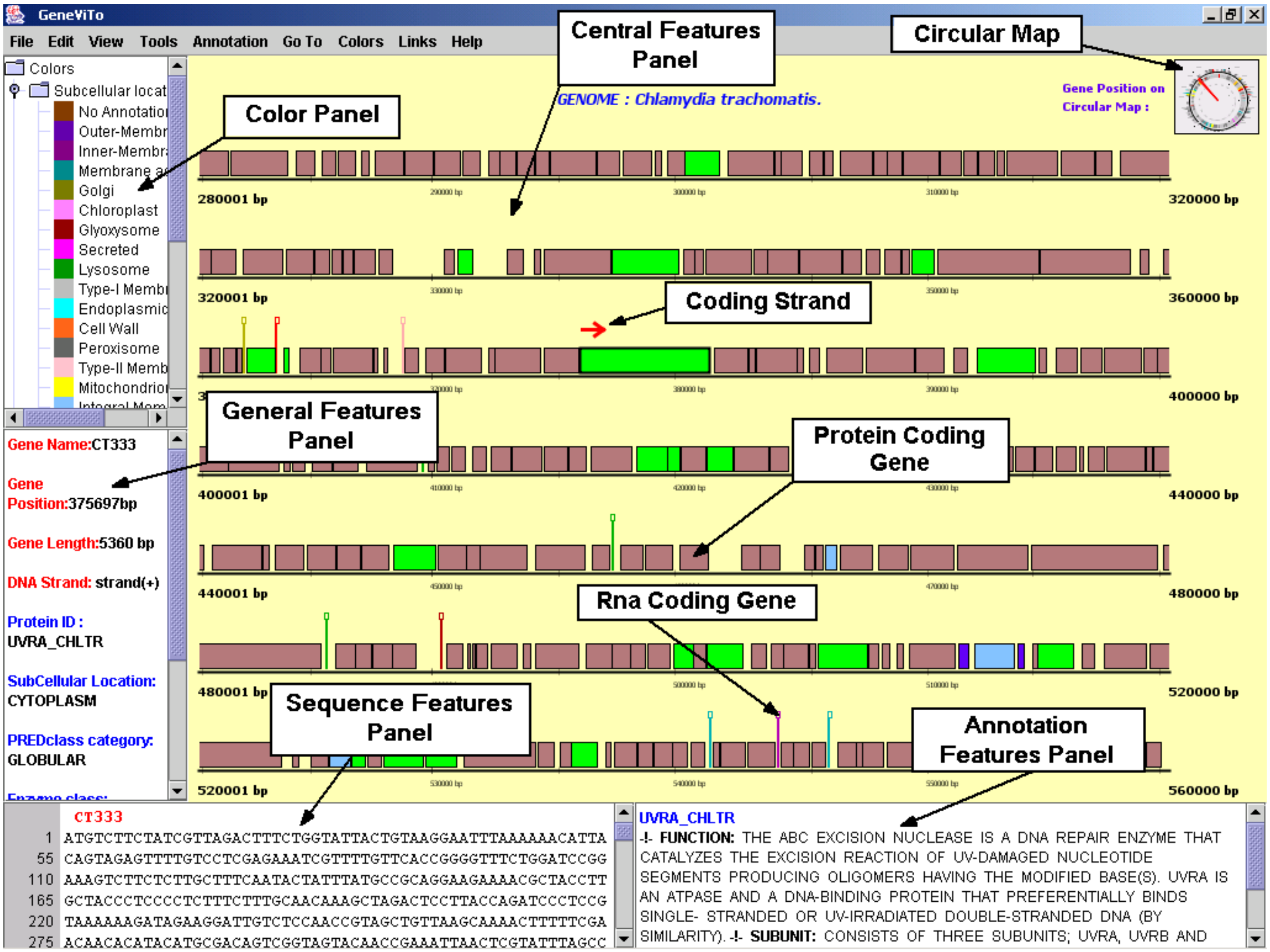
Online Resources	*Type of Data*	*Essential*	*URL*
GenBank	Protein Table	yes	ftp://ftp.ncbi.nih.gov/genbank/genomes/
	FASTA nucleotide coding regions file	yes	ftp://ftp.ncbi.nih.gov/genbank/genomes/
	FASTA Nucleic Acid file	yes	ftp://ftp.ncbi.nih.gov/genbank/genomes/
	Cluster of Orthologous Groups (COGs)	no	http://www.ncbi.nlm.nih.gov/COG/
	Translation table	no	http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=c
Proteome Analysis Server (EBI)	Complete Proteome	yes	http://www.ebi.ac.uk/proteome/index.html
SwissProt			http://us.expasy.org/sprot/
WebCutter	Restriction Enzymes Sites	no	http://www.firstmarket.com/cutter/cut2.html
PRED-CLASS	Protein Structural Classification	no	http://biophysics.biol.uoa.gr/PRED-CLASS/
PRED-TMR2 and orienTM *	Prediction of TM region location and orientation	no	http://biophysics.biol.uoa.gr/PRED-TMR2/, and http://biophysics.biol.uoa.gr/orienTM/
SIGNALP	Signal Peptides	no	http://www.cbs.dtu.dk/services/SignalP/

A list of web-based resources and tools providing input data for visualization with GeneViTo. Essential data types are the minimum required dataset in order to load a Genome into GeneViTo. * PRED-TMR2 and orienTM can be executed in a single step from the orienTM web-server http://biophysics.biol.uoa.gr/orienTM/

corresponding gene and its length. The part of the genome being displayed each time on the "Central Graphics Panel" is a bulk of 280000 bp and all included genomic elements therein. Species information exists on the top of this panel; further on the right resides a clickable "Circular Map" with a red pointer indicating the currently selected genomic element.

The sequence of the selected element is displayed in the "Sequence Features Panel", along with all the relevant information available from nucleotide or protein sequence databases (GenBank, SWISS-PROT) and data available from prediction algorithms (PRED-CLASS, PRED-TMR2, orienTM, SIGNALP). In the "General Features Panel", information about the selected item is displayed in red, such as: gene name, position on the genome (bp), length (bp) and coding strand (5'-3', 3'-5'). Information about the respective protein product (in case of protein coding genes) is available in blue: protein identification, subcellular location, enzyme class, predicted structural class (PRED-CLASS), number of annotated (SWISS-PROT) and predicted transmembrane segments (PRED-TMR2).

The "Color panel" is a tree-like structure indexing the color choices corresponding to the indications displayed in the "Central Graphics Panel", so that users can consult the color patterns each time. Extra information available from SWISS-PROT and user-defined annotation on each protein or gene is displayed in a more detailed way in the "Annotation Features Panel". The "Central Graphics Panel" is dynamically connected to all other panels and the circular map and their information is updated in real time on each selection.



Gene Name:CT333

Gene Position:375697bp

Gene Length:5360 bp

DNA Strand: strand(+)

Protein ID :
UVRA_CHLTR

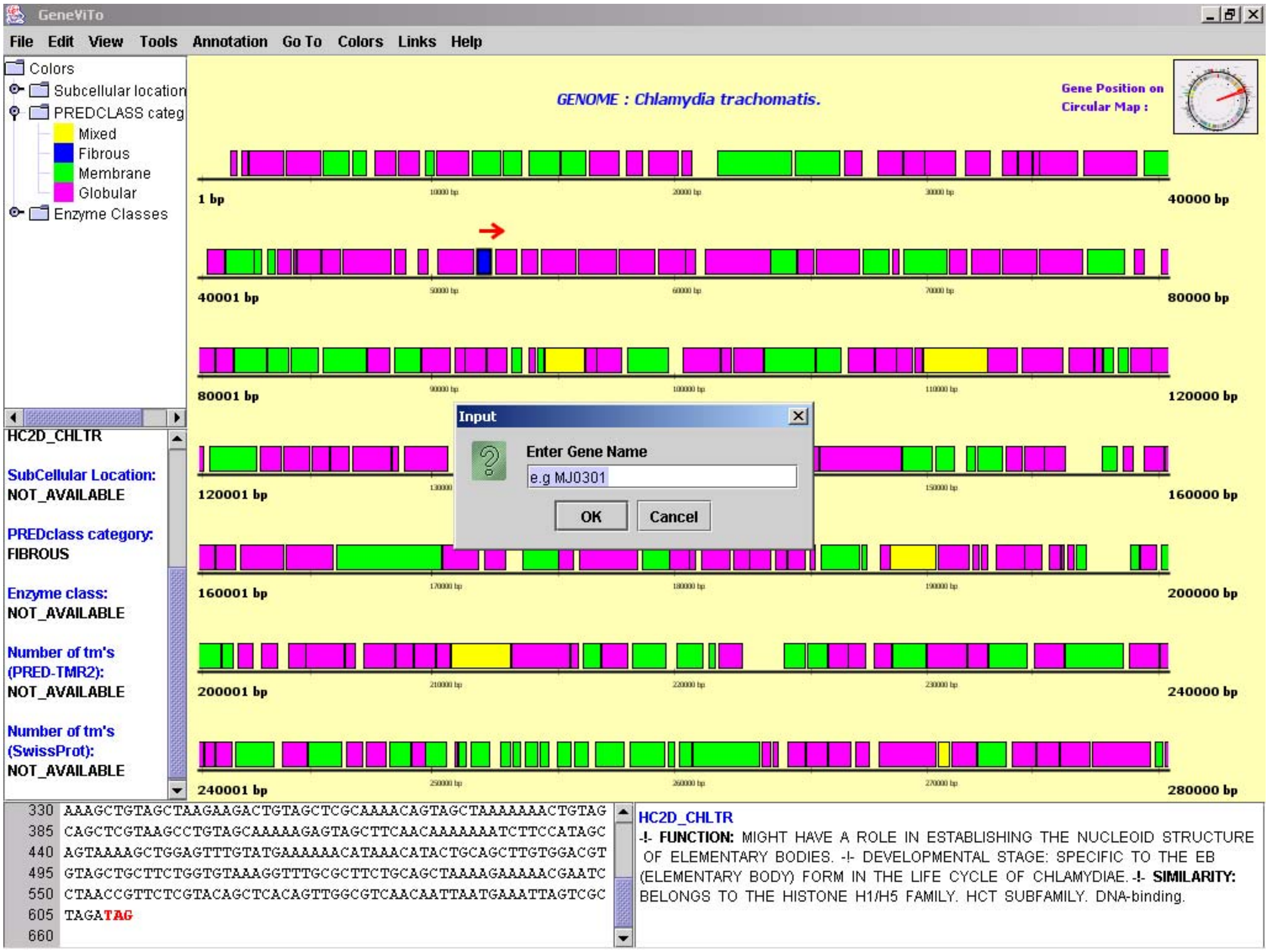
SubCellular Location:
CYTOPLASM

PREDclass category:
GLOBULAR

```
      CT333
  1 ATGTCTTCTATCGTTAGACTTTCTGGTATTACTGTAAGGAATTTAAAAAACATTA
 55 CAGTAGAGTTTTGTCCTCGAGAAATCGTTTTGTTCACCGGGGTTTCTGGATCCGG
110 AAAGTCTTCTCTTGCTTTCAATACTATTTATGCCGCAGGAAGAAAACGCTACCTT
165 GCTACCCTCCCCTCTTTCTTTGCAACAAAGCTAGACTCCTTACCAGATCCCTCCG
220 TAAAAAAGATAGAAGGATTGTCTCCAACCGTAGCTGTTAAGCAAAACTTTTTCGA
275 ACAACACATACATGCGACAGTCGGTAGTACAACCGAAATTAACTCGTATTTAGCC
```

UVRA_CHLTR
-!- **FUNCTION:** THE ABC EXCISION NUCLEASE IS A DNA REPAIR ENZYME THAT CATALYZES THE EXCISION REACTION OF UV-DAMAGED NUCLEOTIDE SEGMENTS PRODUCING OLIGOMERS HAVING THE MODIFIED BASE(S). UVRA IS AN ATPASE AND A DNA-BINDING PROTEIN THAT PREFERENTIALLY BINDS SINGLE- STRANDED OR UV-IRRADIATED DOUBLE-STRANDED DNA (BY SIMILARITY). -!- **SUBUNIT:** CONSISTS OF THREE SUBUNITS; UVRA, UVRB AND

Figure 3
GeneViTo screenshot. A general overview of the graphics interface; feature panels are marked with nametags.



HC2D_CHLTR

SubCellular Location:
NOT_AVAILABLE

PREDclass category:
FIBROUS

Enzyme class:
NOT_AVAILABLE

Number of tm's (PRED-TMR2):
NOT_AVAILABLE

Number of tm's (SwissProt):
NOT_AVAILABLE

```
330 AAAGCTGTAGCTAAGAAGACTGTAGCTCGCAAAACAGTAGCTAAAAAAACTGTAG
385 CAGCTCGTAAGCCTGTAGCAAAAAGAGTAGCTTCAACAAAAAAATCTTCCATAGC
440 AGTAAAAGCTGGAGTTTGTATGAAAAAACATAAACATACTGCAGCTTGTGGACGT
495 GTAGCTGCTTCTGGTGTAAAGGTTTGCGCTTCTGCAGCTAAAAGAAAAACGAATC
550 CTAACCGTTCTCGTACAGCTCACAGTTGGCGTCAACAATTAATGAAATTAGTCGC
605 TAGATAG
660
```

HC2D_CHLTR
-!- **FUNCTION:** MIGHT HAVE A ROLE IN ESTABLISHING THE NUCLEOID STRUCTURE OF ELEMENTARY BODIES. -!- DEVELOPMENTAL STAGE: SPECIFIC TO THE EB (ELEMENTARY BODY) FORM IN THE LIFE CYCLE OF CHLAMYDIAE. -!- **SIMILARITY:** BELONGS TO THE HISTONE H1/H5 FAMILY. HCT SUBFAMILY. DNA-binding.

Figure 4
Searching for a given gene or protein.

Functionality

A genome of interest may be loaded by selecting the option "Open a project" and choosing the folder where the specially formatted files were initially stored. Genomic elements are selected by simple mouse-clicks on the respective boxes. Consequently, all available information is instantly displayed in the relative panels according to the already defined options, while at the same time the red pointer is dragged on the circular map to the respective position of the gene. In addition, an arrow displayed near the selected gene, indicates the DNA strand where the gene is placed on (red left_to_right arrow for the 5'-3' and blue right_to_left arrow for the 3'-5' DNA strand).

"Browsing" a genome

"Browsing" throughout a loaded genome is performed in 280000 bp steps. An alternative search utility offers navigation through the genome, by entering the name, full or partial sequence of a genomic element (gene/protein) in the input dialog provided from the Edit menu (Figure 4); once the above task is initiated, the genome representation "slides" accordingly, and the search result will be highlighted with a pink arrow pointing above it. In the case that the query includes a sequence segment existing in more than one genes or proteins the transition will take place consecutively indicating each gene or protein that matches the search.

Table 3: Subcellular Locations

Outer-Membrane
Inner-Membrane
Membrane associated
Golgi
Chloroplast
Glyoxysome
Secreted
Lysosome
Type-I Membrane Protein
Type-II Membrane Protein
Endoplasmic Reticulum
Cell Wall
Peroxisome
Mitochondrion
Integral Membrane
Cytoplasm
Nucleus
No Annotation

Structural RNA coding genes

Apart from protein coding genes (default option), structural RNA coding genes (t-RNAs, r-RNAs), can also be viewed. Once this option is activated, extra genes are depicted on the genome in the form of "sticks", colored according to the amino acid residue they carry (in case of t-RNAs) or the corresponding ribosomal subunit (in case of r-RNAs). There are several features available on each RNA-coding gene, such as genome position, length, sequence, and the RNA type it belongs to, displayed in the "Sequence Features Panel".

Protein sequence related features

Organizing and grouping information cited in SWISS-PROT features in GeneViTo format, facilitates an "along the genome" depiction of such element distribution. Annotation referring to the subcellular location of a protein has been organized into 18 distinct categories (Table 3). Similarly, using the Enzyme Code (E.C.) listed in the SWISS-PROT annotation, the six known general enzyme classes (Ligases, Transferases, Hydrolases, Lyases, Oxydoreductases and Isomerases) are also embodied for representation. Each genomic element may be colored according to the above criteria, thus giving the opportunity to acquire an overall image about the location and function of proteins distributed along the genome. For example, Figure 5 illustrates a cluster of *M. jannaschii* protein coding genes whose products are annotated as "Hydrolases", indicating possible operon structures along the genome [21]. Functional annotation about the protein product of each gene is available in detail, offering a more integrated view of the specific protein, including: possible function, metabolic pathways, subunits, coenzymes, similarity with other proteins, catalytic activity, cofactors etc. Moreover annotated transmembrane segments, enzyme active sites and signal peptides of a protein are highlighted on the sequence.

Predictions for protein sequences

Prediction algorithms provide information on protein sequences, concerning: possible transmembrane domains (PRED-TMR2) along with their topology (orienTM), structural classification into four distinct classes (PRED-CLASS; Membrane, Globular, Fibrous, Mixed), signal peptide and cleavage site prediction (SIGNALP). These features are a valuable mean of annotating genomic Open Reading Frames with no structural or functional assignment or as a mean of evaluation.

View Clusters of Orthologous groups (COGs)

The NCBI COGs database offers a phylogenetic classification of proteins in an entire genome. Each COG consists of proteins that likely share a common function or domain, which in turn has a role in a given cellular process (or processes). Different COGs functional categories can be highlighted in a different color in order to view the genes that encode such proteins. Once this option is activated, interesting conclusions can be inferred about organization and topological characteristics of the genome (Figure 6).

View restriction sites on the genome

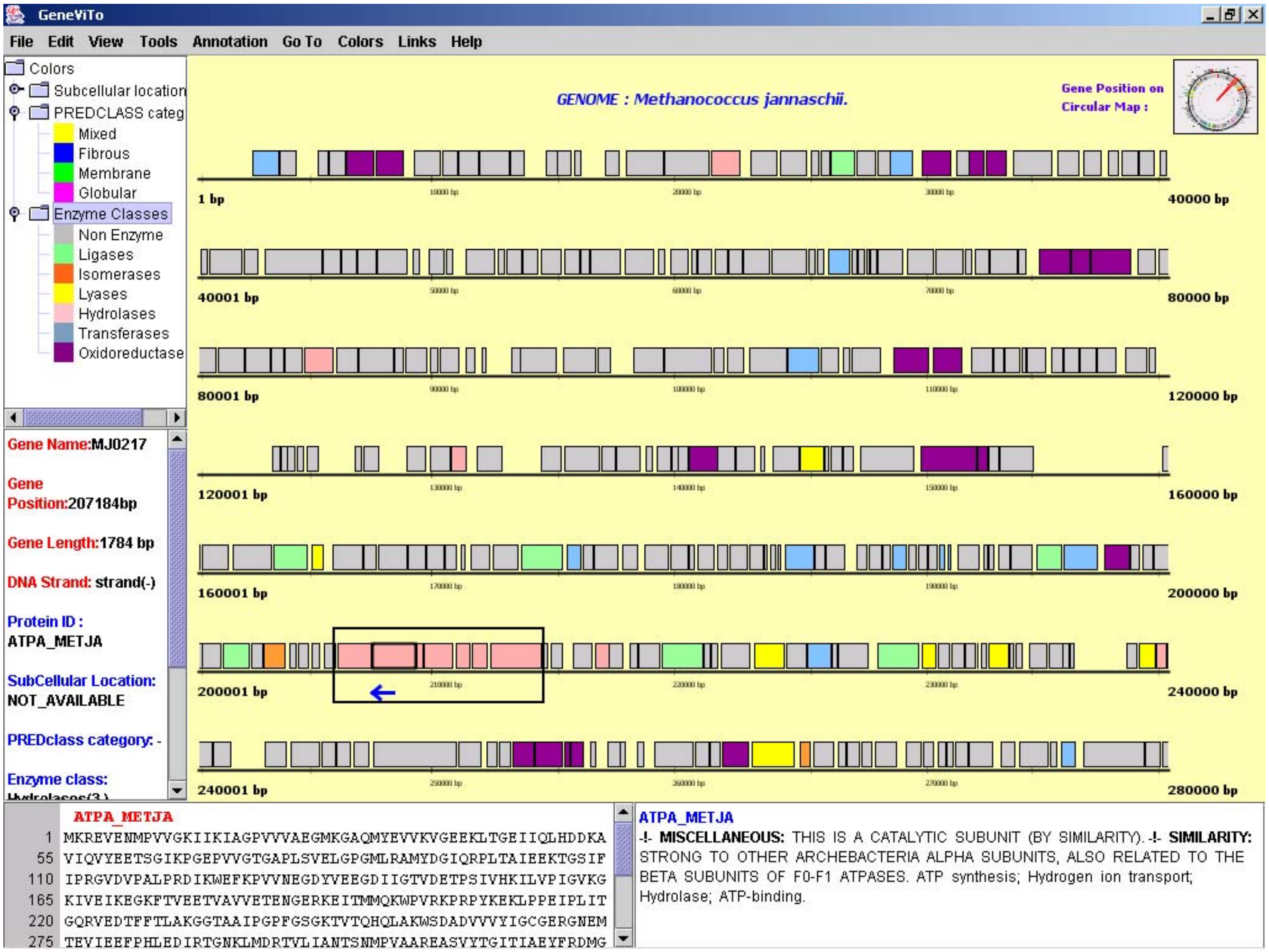
Possible recognition sites of restriction enzymes are highlighted on the whole genome, using Webcutter (Figure 7). Several restriction enzyme sites are available by this application. A pink triangle indicates recognition positions for restriction endonucleases, while clicking on it provides the respective position (in bp) along with the sequence pattern recognized by the selected restriction enzyme.

User defined annotation

Adding personal annotation for a specific gene or protein is supported. By doing so, each time the genome is loaded this user-defined information will also be available, with the option of modifying (erasing or updating) it. This feature allows storage of personal proteomic-genomic research data that can easily be retrieved.

View the whole genome in a circular map

By mouse clicking on the miniature of the circular map on the top-right site of the Graphics Panel, a circular map of the genome is depicted in a new window. In this case, all genes are colored according to the 26 COGs categories (Figure 8). Selection of a particular region on this map will highlight the selected region on the main window in pink color.



Gene Name:MJ0217

Gene Position:207184bp

Gene Length:1784 bp

DNA Strand: strand(-)

Protein ID : ATPA_METJA

SubCellular Location: NOT_AVAILABLE

PREDclass category: -

Enzyme class: Hydrolases(3.)

```
    ATPA_METJA
  1 MKREVENMPVVGKIIKIAGPVVVAEGMKGAQMYEVVKVGEEKLTGEIIQLHDDKA
 55 VIQVYEETSGIKPGEPVVGTGAPLSVELGPGMLRAMYDGIQRPLTAIEEKTGSIF
110 IPRGVDVPALPRDIKWEFKPVVNEGDYVEEGDIIGTVDETPSIVHKILVPIGVKG
165 KIVEIKEGKFTVEETVAVVETENGERKEITMMQKWPVRKPRPYKEKLPPEIPLIT
220 GQRVEDTFFTLAKGGTAAIPGPFGSGKTVTQHQLAKWSDADVVVYIGCGERGNEM
275 TEVIEEFPHLEDIRTGNKLMDRTVLIANTSNMPVAAREASVYTGITIAEYFRDMG
```

ATPA_METJA

-!- **MISCELLANEOUS:** THIS IS A CATALYTIC SUBUNIT (BY SIMILARITY). -!- **SIMILARITY:** STRONG TO OTHER ARCHEBACTERIA ALPHA SUBUNITS, ALSO RELATED TO THE BETA SUBUNITS OF F0-F1 ATPASES. ATP synthesis; Hydrogen ion transport; Hydrolase; ATP-binding.

Figure 5

Illustrating the distribution of gene-products belonging to the main enzyme classes. A screenshot where genes are colored according to the annotated E.C. number of their protein product, in *M. jannaschii*. A "cluster" of contiguous genes coding hydrolases is highlighted in black color.

Saving a Jpeg snapshot

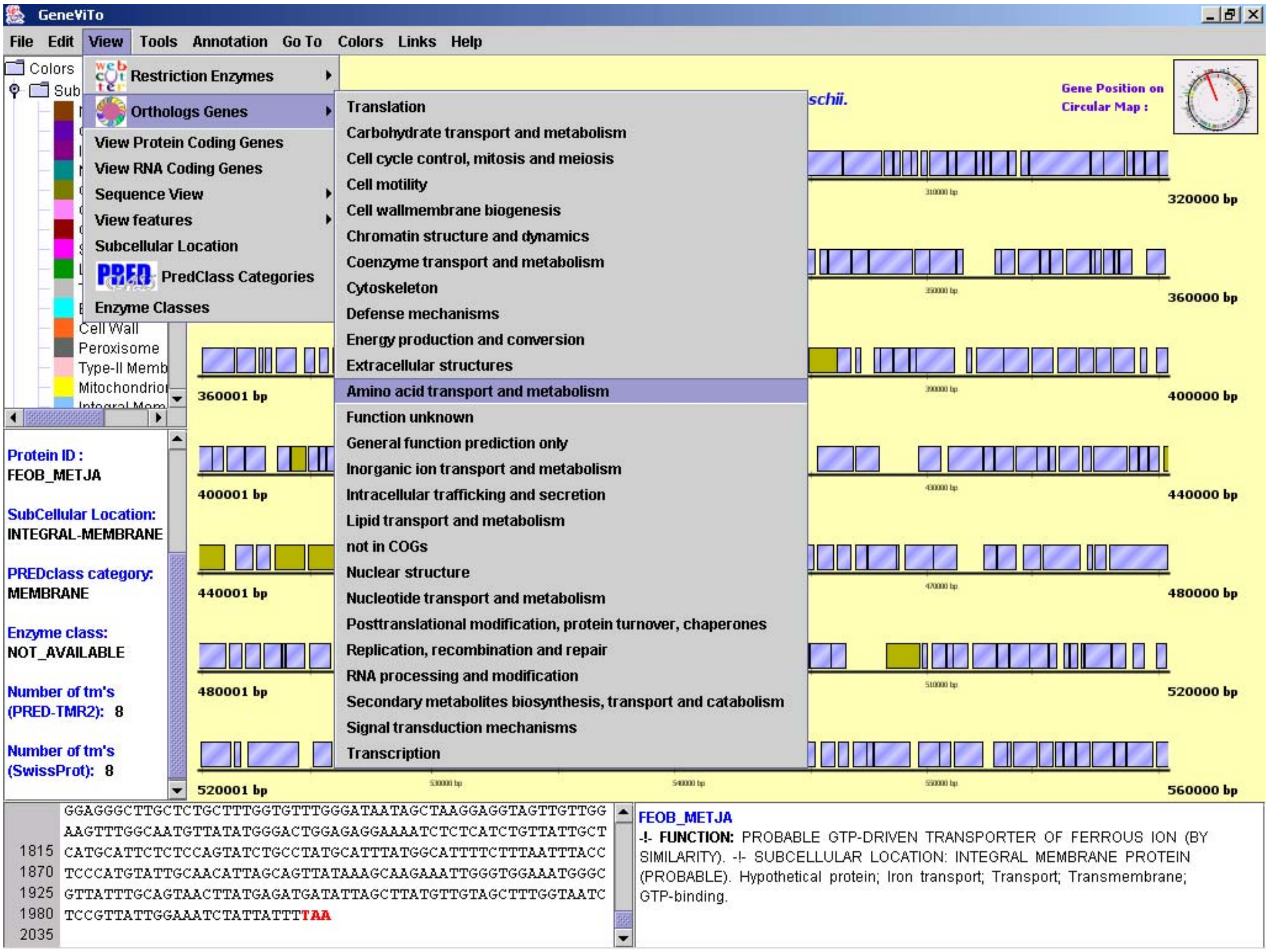
An additional feature provided by GeneViTo is the ability to save the current display in a high-resolution color image file in JPEG format, suitable for direct incorporation into scientific publications or further inspection. Additionally, a color palette enables the modification of the graphics panel background color, according to the user's personal taste.

Searching for similarities

Possible local sequence similarities can be acquired for any selected element running the BLAST [22] algorithm (BlastN, BlastP, BlastX) on a particular sequence (DNA or protein) against some default databases. Any database of choice can also be defined, given that it follows the Fasta format. In this case, GeneViTo offers a utility to format the given file according to the BLAST standards. BLAST results are automatically displayed in a new window after the procedure has been completed.

Searching for a protein sequence motif

An additional tool that is incorporated in the GeneViTo workbench, gives the opportunity of searching for particular sequence motifs in aminoacid sequences throughout the entire coding part of the genome. This procedure is invoked after entering the motif of interest into the relative Input dialog. The output window indicates all protein sequences found to comply with the given motif (Figure 9).


GeneViTo
File
Edit
View
Tools
Annotation
Go To
Colors
Links
Help
Colors
Sub
Restriction Enzymes
Orthologs Genes
View Protein Coding Genes
View RNA Coding Genes
Sequence View
View features
Subcellular Location
PredClass Categories
Enzyme Classes
Translation
Carbohydrate transport and metabolism
Cell cycle control, mitosis and meiosis
Cell motility
Cell wallmembrane biogenesis
Chromatin structure and dynamics
Coenzyme transport and metabolism
Cytoskeleton
Defense mechanisms
Energy production and conversion
Extracellular structures
Amino acid transport and metabolism
Function unknown
General function prediction only
Inorganic ion transport and metabolism
Intracellular trafficking and secretion
Lipid transport and metabolism
not in COGs
Nuclear structure
Nucleotide transport and metabolism
Posttranslational modification, protein turnover, chaperones
Replication, recombination and repair
RNA processing and modification
Secondary metabolites biosynthesis, transport and catabolism
Signal transduction mechanisms
Transcription
Cell Wall
Peroxisome
Type-II Memb
Mitochondrion
Integral Mem
schii.
Gene Position on Circular Map :
320000 bp
360000 bp
360001 bp
400000 bp
400001 bp
440000 bp
440001 bp
480000 bp
480001 bp
520000 bp
520001 bp
560000 bp

Protein ID :
FEOB_METJA

SubCellular Location:
INTEGRAL-MEMBRANE

PREDclass category:
MEMBRANE

Enzyme class:
NOT_AVAILABLE

Number of tm's (PRED-TMR2): 8

Number of tm's (SwissProt): 8

```
      GGAGGGCTTGCTCTGCTTTGGTGTTTGGGATAATAGCTAAGGAGGTAGTTGTTGG
      AAGTTTGGCAATGTTATATGGGACTGGAGAGGAAAATCTCTCATCTGTTATTGCT
1815  CATGCATTCTCTCCAGTATCTGCCTATGCATTTATGGCATTTTCTTTAATTTACC
1870  TCCCATGTATTGCAACATTAGCAGTTATAAAGCAAGAAATTGGGTGGAAATGGGC
1925  GTTATTTGCAGTAACTTATGAGATGATATTAGCTTATGTTGTAGCTTTGGTAATC
1980  TCCGTTATTGGAAATCTATTATTTTAA
2035
```

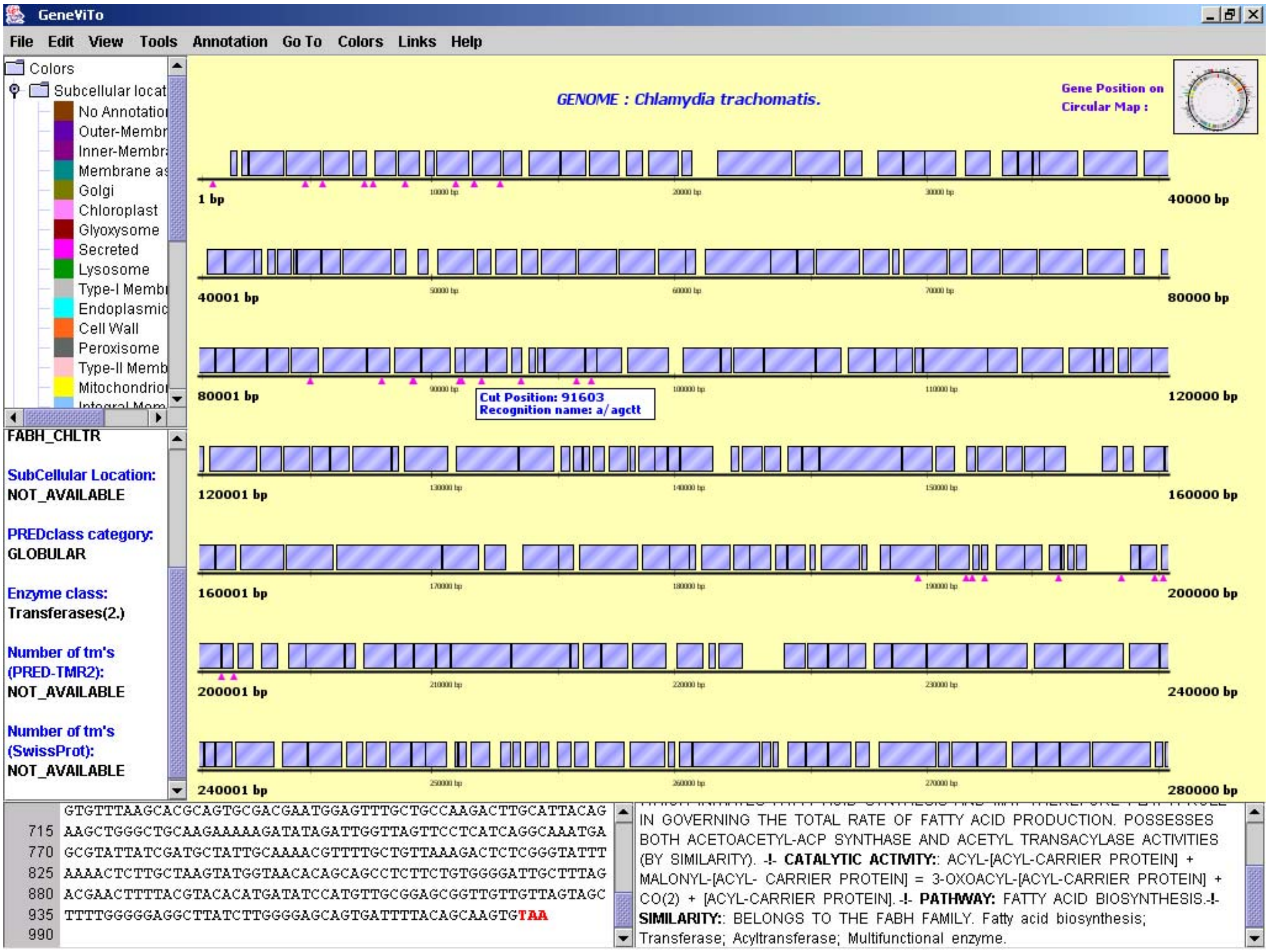
FEOB_METJA
-!- **FUNCTION:** PROBABLE GTP-DRIVEN TRANSPORTER OF FERROUS ION (BY SIMILARITY). -!- SUBCELLULAR LOCATION: INTEGRAL MEMBRANE PROTEIN (PROBABLE). Hypothetical protein; Iron transport; Transport; Transmembrane; GTP-binding.

Figure 6
Illustrating Clusters of Orthologous Groups (COGs). A screenshot depicting the 26 COGs functional categories. The "Amino acid transport and metabolism" group has been selected and the results are highlighted in light-green color, in the *M. jannaschii* genome.

Extra Graphs

A number of extra graphs, especially suitable for an overall image of the genome-proteome, are available to the user, either in the form of a Bar Chart or in the form of a Pie Chart. Being more specific, length distribution of the non-transmembrane segments (loops) is displayed in a Bar Chart (Figure 10), offering information about the length of the transmembrane proteins or the possible domains built by those loops. A Pie Chart displays the percentage of the transmembrane proteins whose C-terminal and N-terminal ends appear with a particular topology (N-in, N-out, C-in, C-out). Transmembrane N, C termini topology can lead to valuable conclusions, for instance, the fact that in most organisms there is a strong bias for both the N-terminal and C-terminal ends to be located inside the cell [23].

Additionally, Pie Charts display the fraction of: a) proteins that have been predicted as membrane, globular, fibrous or mixed by the PRED-CLASS algorithm, b) proteins that belong in one of the 6 Enzyme classes (Figure 11). Moreover, the number of the transmembrane proteins with a given number of transmembrane segments is displayed in a Bar Chart, while the percentage of the proteins located in a certain subcellular location is displayed in a Pie Chart. Composition analysis of nucleotide or amino acid sequences, or the entire genome (e.g. G+C content) is provided. In addition, the percentage of the



Cut Position: 91603
Recognition name: a/agctt

FABH_CHLTR

SubCellular Location:
NOT_AVAILABLE

PREDclass category:
GLOBULAR

Enzyme class:
Transferases(2.)

Number of tm's (PRED-TMR2):
NOT_AVAILABLE

Number of tm's (SwissProt):
NOT_AVAILABLE

```
     GTGTTTAAGCACGCAGTGCGACGAATGGAGTTTGCTGCCAAGACTTGCATTACAG
715  AAGCTGGGCTGCAAGAAAAAGATATAGATTGGTTAGTTCCTCATCAGGCAAATGA
770  GCGTATTATCGATGCTATTGCAAAACGTTTTGCTGTTAAAGACTCTCGGGTATTT
825  AAAACTCTTGCTAAGTATGGTAACACAGCAGCCTCTTCTGTGGGGATTGCTTTAG
880  ACGAACTTTTACGTACACATGATATCCATGTTGCGGAGCGGTTGTTGTTAGTAGC
935  TTTTGGGGGAGGCTTATCTTGGGGAGCAGTGATTTTACAGCAAGTGTAA
990
```

IN GOVERNING THE TOTAL RATE OF FATTY ACID PRODUCTION. POSSESSES BOTH ACETOACETYL-ACP SYNTHASE AND ACETYL TRANSACYLASE ACTIVITIES (BY SIMILARITY). -!- **CATALYTIC ACTIVITY:**: ACYL-[ACYL-CARRIER PROTEIN] + MALONYL-[ACYL- CARRIER PROTEIN] = 3-OXOACYL-[ACYL-CARRIER PROTEIN] + CO(2) + [ACYL-CARRIER PROTEIN]. -!- **PATHWAY:** FATTY ACID BIOSYNTHESIS.-!- **SIMILARITY:**: BELONGS TO THE FABH FAMILY. Fatty acid biosynthesis; Transferase; Acyltransferase; Multifunctional enzyme.

Figure 7
Illustrating Restriction Enzyme Sites. A screenshot in which the recognition sites of HindIII restriction enzyme are highlighted with a pink triangle on the genome of *C. trachomatis*. The recognition site at the position 91603 bp has been selected.

protein-coding, RNA-coding and non-coding part of the genome, is a feature available in a Pie Chart.

Detailed help for all options and tools supported by GeneViTo is organized in a tree-like structure, through the «Help» menu.

Comparison to other genome browsers

Availability of huge amounts of genomic data during the last decade urged the development of computer applications for the analysis and visualization of this enormous information. Consequently, several approaches have been made to the problem of genome visualization, resulting in diverse (as far as both the capabilities and the overall philosophy) genome browsers. Each system seems to have been developed primarily to serve investigators' needs in a 'project' environment, paying more attention to incorporate features needed to accomplish specific tasks. This fact makes direct comparisons of such software difficult and, to some extent, arbitrary. In this section, we compare GeneViTo to three well known and widely used browsers: the TIGR genome browser [11], the ENSEMBL project viewer[12] and ERGO™ [13] (actually its freely available version ERGO Light).

The TIGR genome browser is freely available through the TIGR web server, providing access to all TIGR in-house data. This genome browser comes with some very useful features, such as the graphical display of alternative sources of annotations (e.g. multiple gene-finder predic-

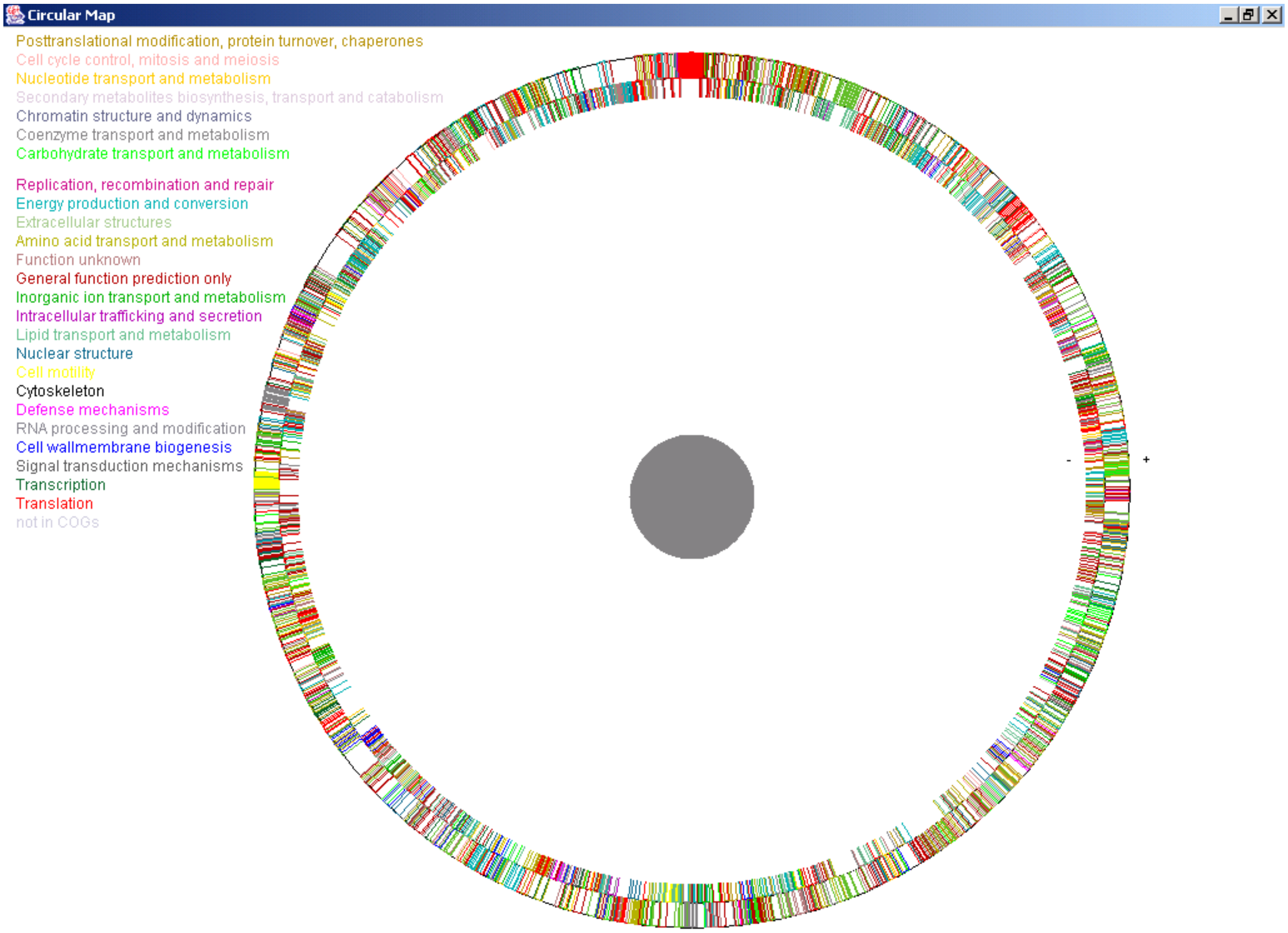

Circular Map
Posttranslational modification, protein turnover, chaperones
Cell cycle control, mitosis and meiosis
Nucleotide transport and metabolism
Secondary metabolites biosynthesis, transport and catabolism
Chromatin structure and dynamics
Coenzyme transport and metabolism
Carbohydrate transport and metabolism
Replication, recombination and repair
Energy production and conversion
Extracellular structures
Amino acid transport and metabolism
Function unknown
General function prediction only
Inorganic ion transport and metabolism
Intracellular trafficking and secretion
Lipid transport and metabolism
Nuclear structure
Cell motility
Cytoskeleton
Defense mechanisms
RNA processing and modification
Cell wallmembrane biogenesis
Signal transduction mechanisms
Transcription
Translation
not in COGs
-
+

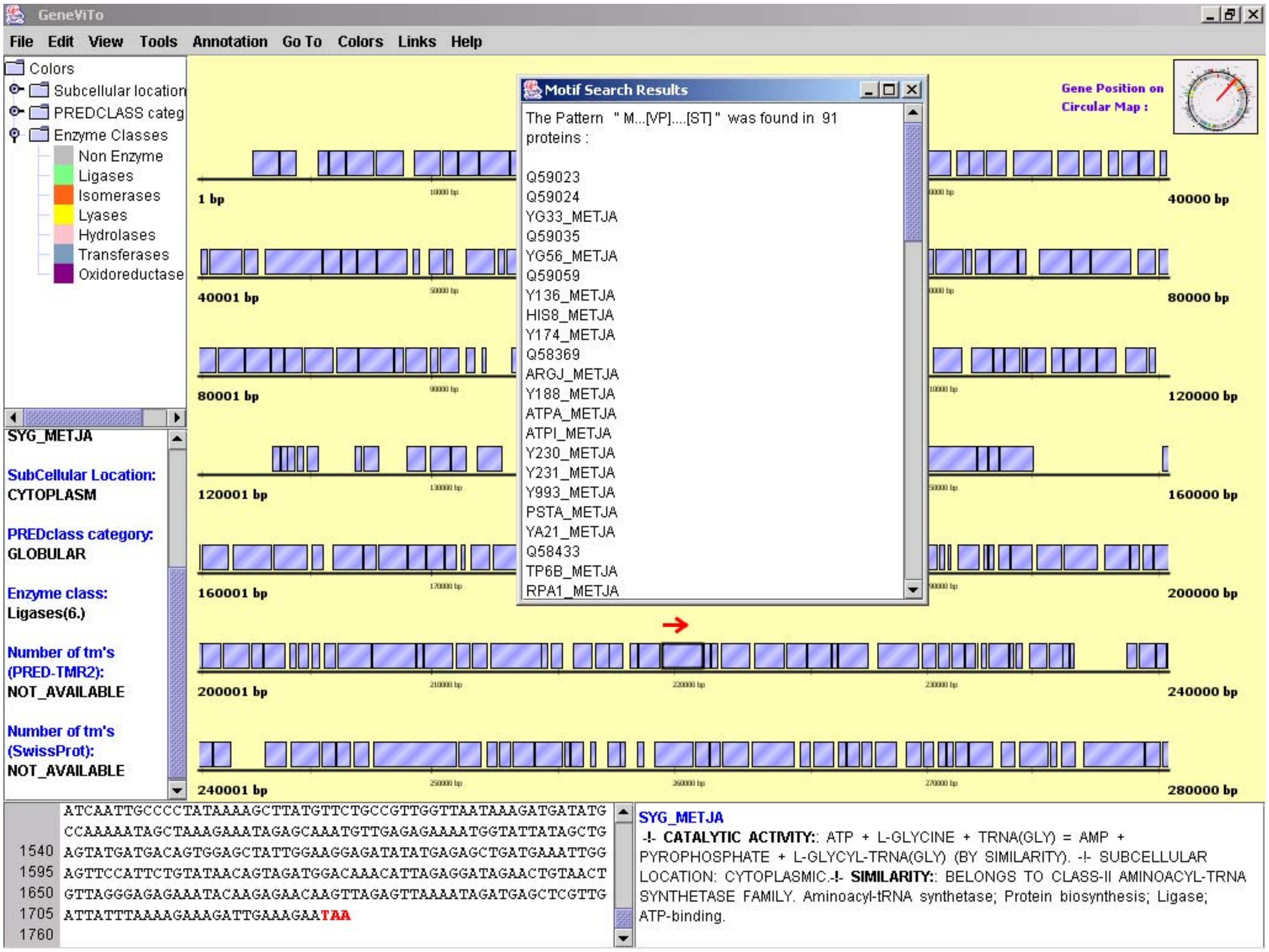
Figure 8
Circular representation of a bacterial genome. A screenshot of the circular map representing the whole genome of *M. jannaschii*; genes are colored according to the COGs categories.

tions), matches of gene products to characteristic sequence motifs and so on. Again, the software system is not available to install locally and users are not able to upload their genomic data along with annotations or predictions, so it has to be considered as an interface to TIGR genomic data.

The ENSEMBL genome browser is a valuable set of genome analysis and visualization tools, offering a functional working environment for web-based data mining and information viewing. Individual gene products are linked to automatically created annotation, while users can save and alter annotations when new experimental evidence become available. Moreover the entire package can be downloaded to be used for individual research purposes.

ERGO is a commercial software suite with excellent capabilities, including metabolic pathway information, but it is mostly data-centric. More important to its great computational power is the underlying data annotation handled by a team of experts. The freely available ERGO Light version, offers free access to a smaller amount of data, yet the same computational tools, but users cannot upload their own data to the server.

GeneViTo, as well as all the above software resources, come with intuitive user friendly GUIs, allowing for easy navigation through the vast amount of genomic data.



The Pattern " M...[VP]....[ST] " was found in 91 proteins :

Q59023
Q59024
YG33_METJA
Q59035
YG56_METJA
Q59059
Y136_METJA
HIS8_METJA
Y174_METJA
Q58369
ARGJ_METJA
Y188_METJA
ATPA_METJA
ATPI_METJA
Y230_METJA
Y231_METJA
Y993_METJA
PSTA_METJA
YA21_METJA
Q58433
TP6B_METJA
RPA1_METJA

SYG_METJA

SubCellular Location:
CYTOPLASM

PREDclass category:
GLOBULAR

Enzyme class:
Ligases(6.)

Number of tm's (PRED-TMR2):
NOT_AVAILABLE

Number of tm's (SwissProt):
NOT_AVAILABLE

```
     ATCAATTGCCCCTATAAAAGCTTATGTTCTGCCGTTGGTTAATAAAGATGATATG
     CCAAAAATAGCTAAAGAAATAGAGCAAATGTTGAGAGAAAATGGTATTATAGCTG
1540 AGTATGATGACAGTGGAGCTATTGGAAGGAGATATATGAGAGCTGATGAAATTGG
1595 AGTTCCATTCTGTATAACAGTAGATGGACAAACATTAGAGGATAGAACTGTAACT
1650 GTTAGGGAGAGAAATACAAGAGAACAAGTTAGAGTTAAAATAGATGAGCTCGTTG
1705 ATTATTTAAAAGAAAGATTGAAAGAATAA
1760
```

SYG_METJA
-!- **CATALYTIC ACTIVITY:**: ATP + L-GLYCINE + TRNA(GLY) = AMP + PYROPHOSPHATE + L-GLYCYL-TRNA(GLY) (BY SIMILARITY). -!- SUBCELLULAR LOCATION: CYTOPLASMIC.-!- **SIMILARITY:**: BELONGS TO CLASS-II AMINOACYL-TRNA SYNTHETASE FAMILY. Aminoacyl-tRNA synthetase; Protein biosynthesis; Ligase; ATP-binding.

Figure 9
A screenshot of a motif search results. Proteins complying with the given motif are listed in a separate window.

GeneViTo, as a stand alone application, enables the incorporation of user defined-data: genomes, annotations and/or computational predictions. Its main advantage, is the clear presentation of sequence feature distribution along genomic regions. In the input layer GeneViTo uses data from public databases and free bioinformatics tools, so in most of the cases users will be able to easily visualize available data through a simple mouse-click. Simultaneous display of computationally predicted features along with available annotations (often derived by computational means as well) provides a useful environment, which may complement the already existing tools for genome annotation and visualization.

Conclusions

Genome wide analysis is a difficult task due to the large amount of primordial data and the "non-productive" way in which they are stored and displayed. Several genome viewers already exist, each one of them serving different needs and research interests. GeneViTo offers an easy to use computer environment that incorporates experimental data combined with prediction algorithms results. Primordial data and meta-data are all embodied in a clear display that offers instantly an intuitive aspect of a genome and a large amount of biological information at hand. The information offered can lead to valuable conclusions and cover a wide variety of biology issues concerning entire organisms.

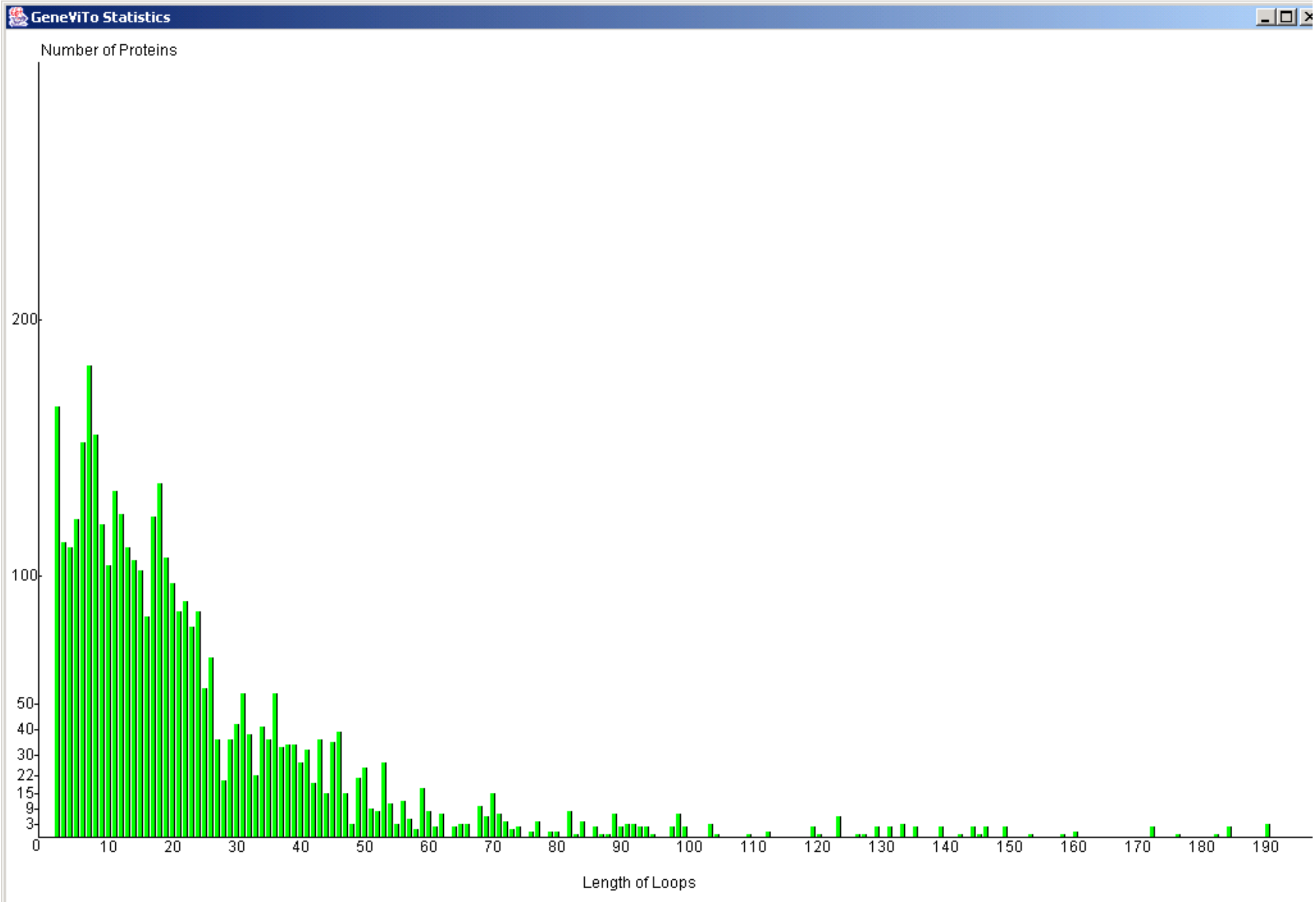


Figure 10
Length distribution of non-transmembrane segments (loops), in *M. jannaschii.*

GeneViTo has already been applied to visualize the genomes of two microbial organisms: the bacterium *C. trachomatis* and the archaeon *M. jannaschii.* Future plans to extend the software platform include the ability to handle multichromosomal genomes as unique sets or the simultaneous display and analysis of complete genomes. In order to achieve that goal, some modifications will be necessary, as we have to efficiently handle the differences in eukaryotic genome organization (e.g. exon-intron structure). Such issues should be taken into account both in the data storage and handling processes and in the visualization philosophy. Computational issues, such as memory requirements, are not raised, as clearly illustrated in Table 1.

GeneViTo will be available to download upon request to the authors. Download instructions, along with the "ready to run" microbial genome files accompanied with a detailed online manual for preparing and viewing genomic data are freely available at the URL http://bioinformatics.biol.uoa.gr/GENEVITO/index.html.

Availability and Requirements

Project Name: GeneViTo

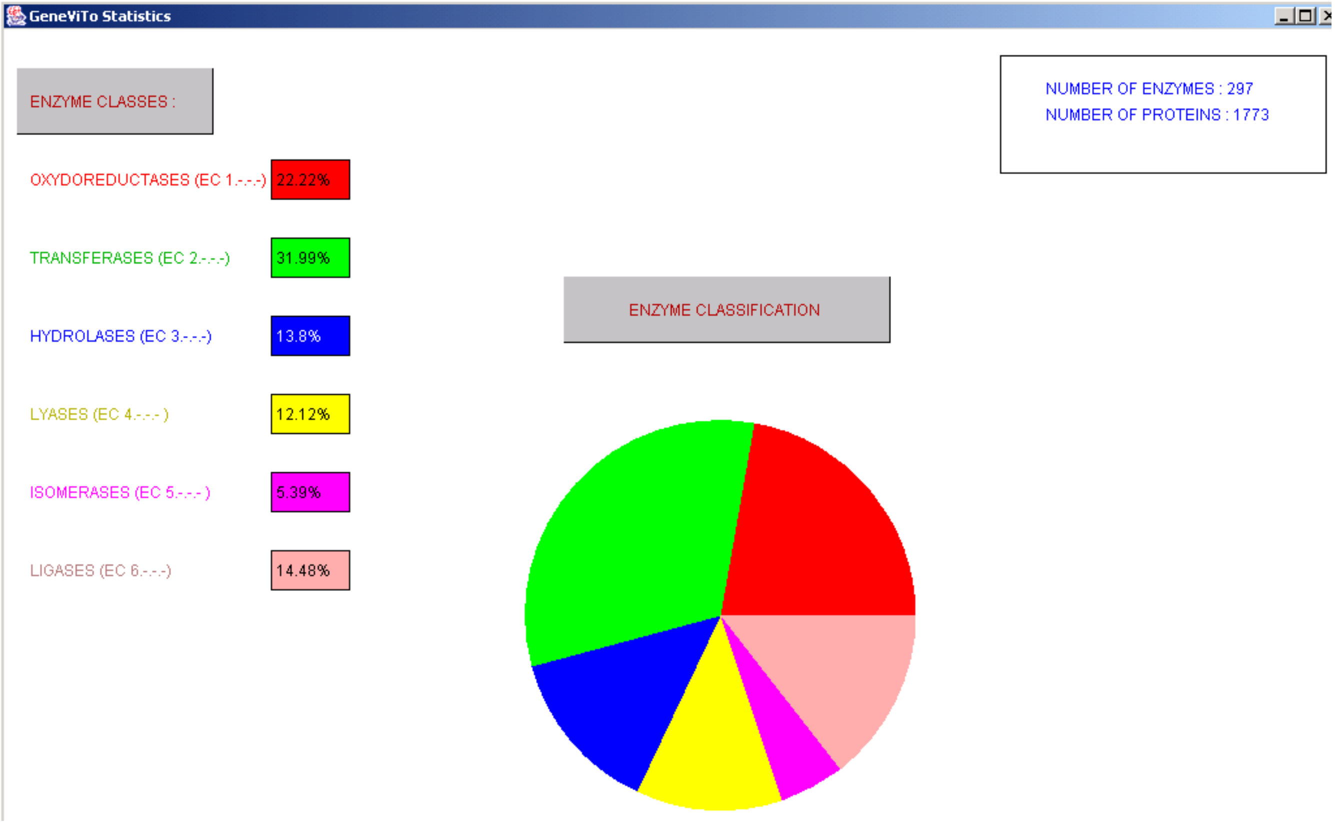
Project Home Page: http://bioinformatics.biol.uoa.gr/GENEVITO/index.html

Operating System(s): Extensively tested on Windows, Linux (Intel). Theoretically, GeneViTo should work on any other platform with Java Runtime Environment (JRE) 1.4.1 installed.

Programming Language: JAVA

Other requirements: Installation of JRE 1.4.1.

License: Free for Academic use.



NUMBER OF ENZYMES : 297
NUMBER OF PROTEINS : 1773

OXYDOREDUCTASES (EC 1.-.-.-) 22.22%
TRANSFERASES (EC 2.-.-.-) 31.99%
HYDROLASES (EC 3.-.-.-) 13.8%
LYASES (EC 4.-.-.-) 12.12%
ISOMERASES (EC 5.-.-.-) 5.39%
LIGASES (EC 6.-.-.-) 14.48%

Figure 11
Percentage of proteins annotated to belong to the different main enzyme classes in *M. jannaschii*.

Any restrictions: None.

Abbreviations

GUI: Graphical User Interface

COGs: Clusters of Orthologous Groups

Authors' contributions

GV created the Graphical User Interface and part of the genome processing programming, CG carried out most of the genome processing programming, VP has tested the software proposing useful improvements and SH coordinated the whole project, suggesting the general directions and innovating features of the application. All authors have read and accepted the final manuscript.

Acknowledgements

The authors wish to thank the two anonymous referees for their valuable comments and suggestions that substantially helped to improve the final manuscript. Pantelis Bagos MSc kindly provided help on performing and presenting the Benchmark tests.

References

1. Bernal A, Ear U and Kyrpides N: **Genomes OnLine Database (GOLD): a monitor of genome projects world-wide.** *Nucleic Acids Res* 2001, **29:**126-7.
2. **GOLD: Genomes OnLine Database** [http://wit.integratedgenomics.com/GOLD/]
3. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J and Wheeler DL: **GenBank.** *Nucleic Acids Res* 2003, **31:**23-7.
4. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S and Schneider M: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2003.** *Nucleic Acids Res* 2003, **31:**365-70.
5. Zdobnov EM, Lopez R, Apweiler R and Etzold T: **The EBI SRS server – recent developments.** *Bioinformatics* 2002, **18:**368-73.
6. Schuler GD, Epstein JA, Ohkawa H and Kans JA: **Entrez: molecular biology database and retrieval system.** In: *Methods in Enzymology Volume 266.* Edited by: *Doolittle RF. San Diego: Academic Press*; 1996:141-62.
7. Abril JF and Guigó R: **gff2ps: visualizing genomic annotations.** *Bioinformatics* 2000, **16:**743-4.
8. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream M-A and Barrell B: **Artemis: sequence visualization and annotation.** *Bioinformatics* 2000, **16:**944-5.
9. Hu Z, Frith M, Niu T and Weng Z: **SeqVISTA: a graphical tool for sequence feature visualization and comparison.** *BMC Bioinformatics* 2003, **4:**1.
10. **NCBI Map Viewer** [http://www.ncbi.nih.gov/mapview/]
11. **TIGR Genome Browse** [http://www.tigr.org/tigr-scripts/CMR2/choose_genome.spl]

12. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H, Lijnzaad P, Melsopp C, Mongin E, Pettett R, Pocock M, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I and Clamp M: **The Ensembl genome database project.** *Nucleic Acids Res* 2002, **30:**38-42.
13. Overbeek R, Larsen N, Walunas T, D'Souza M, Pusch G, Selkov E Jr, Liolios K, Joukov V, Kaznadzey D, Anderson I, Bhattacharyya A, Burd H, Gardner W, Hanke P, Kapatral V, Mikhailova N, Vasieva O, Osterman A, Vonstein V, Fonstein M, Ivanova N and Kyrpides N: **The ERGO™ genome analysis and discovery system.** *Nucleic Acids Res* 2003, **31:**164-79.
14. Pasquier C, Promponas VJ and Hamodrakas SJ: **PRED-CLASS: cascading neural networks for generalized protein classification and genome-wide applications.** *Proteins: Structure, Function, and Genetics* 2001, **44:**361-9.
15. Pasquier C. and Hamodrakas SJ: **An hierarchical artificial neural network system for the classification of transmembrane proteins.** *Protein Eng* 1999, **12:**631-4.
16. Liakopoulos TD, Pasquier C and Hamodrakas SJ: **A novel tool for the prediction of transmembrane protein topology based on a statistical analysis of the SwissProt database: the orienTM algorithm.** *Protein Eng* 2001, **14:**387-90.
17. Nielsen H, Engelbrecht J, Brunak S and von Heijne G: **Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites.** *Protein Eng* 1997, **10:**1-6.
18. **EBI Proteome Analysis Server** [http://www.ebi.ac.uk/proteome/index.html]
19. **Webcutter** [http://www.firstmarket.com/cutter/cut2.html]
20. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND and Koonin EV: **The COG database: new developments in phylogenetic classification of proteins from complete genomes.** *Nucleic Acids Res* 2001, **29:**22-28.
21. **TIGR, Gene pairs for Methanococcus jannaschii** [http://www.tigr.org/tigr-scripts/operons/pairs.cgi?taxon_id=57]
22. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-402.
23. Krogh A, Larsson B, von Heijne G and Sonnhammer ELL: **Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes.** *J Mol Biol* 2001, **305:**567-80.