**GENOMES DATA IN ENTREZ: REPRESENTATION AND ANALYSIS.**

TATUSOVA T.A.+, OSTELL J.

National Center for Biotechnology Information, 8600 Rockville Pike, BETHESDA, MD 20894, USA;
e-mail: tatiana@ncbi.nlm.nih.gov

+Corresponding author

Keywords: genome data, sequence analysis, sequence data representation

The release of the first microbial genome (*H.influenzae*in 1995 [1]) marked a new era of megabase sequence data. Large-scale sequencing efforts have now produced a number of completely sequenced genomes or chromosomes from a variety of organisms. Genomes data is a new form of information that requires new approaches to be useful and accessible. There are a number of sequence analysis utilities and a lot of preanalysis information that is available for these genomes. But to make good use of them a system is needed which is capable of smoothly and rapidly examining the data at varying levels of detail. The integrated system combining all software tools developed at NCBI to represent, manipulate and analyze the data is designed to accommodate exactly these needs.

**1. Introduction**

In September 1995 National Center for Biotechnology Information (NCBI) created a new Genomes division of GenBank for handling the data obtained from large-scale sequencing of genomes and chromosomes [2]. Currently the Genome division contains 679 entries including information on completely sequenced genomes and chromosomes as well as contiged sequence maps, and integrated genetic and physical maps from higher eukaryotes. Entrez provides access to these records and allows the user to visualize the sequence information at varying levels of detail either graphically or as text. The chromosome views are tightly linked to DNA and protein sequence records, MEDLINE (PubMed) citations, and the three dimensional protein structure division (MMDB). Though the access and distribution of the large amount of data in a convenient and timely manner is of a great importance to scientific community the real challenge is to make effective use of this information. This system offers a number of ways of looking at the data and a variety of tools for sequence analysis. Here we discuss the benefits of using the combination of viewing and analysis tools for exploring complete microbial genome sequences data. Entrez Genomes can be accessed on the WWW at<http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html>.

**2. Data organization**

***2.1. Multiple Integrated Maps***

Relatively small parts of chromosomes of higher eukaryotes have been sequenced. In these cases, the various genetic and physical maps for a particular organism have been compiled and mapped onto a common coordinate system. The beginning of a sequence map for the chromosome is made using contigs of sequence from the same region and organism, then placing the composite sequence onto the coordinate system provided by the integration of the maps.

***2.2. Small Genomes (single GenBank records)***

The small genomes branch typically consists of genomes which are in a single GenBank record. These include viruses and organelles. Due to a large degree of redundancy (multiple versions, population variants, partial sequence entries) in the GenBank records, NCBI selected a reference sequence and then aligned other versions of the sequence with the reference sequence.

***2.3. Large-scale Complete Genomes (virtual chromosomes)***

The complete sequence of chromosomes from some lower eukaryotes, such as yeast and bacteria, are available but exist as smaller, overlapping records in GenBank. Rather than creating single large entries, genome-size submissions are divided into several entries, each no more than 350 KB long (a limit set by the International Nucleotide Database Collaboration). "Virtual records" in Genome division define the method for assembling the long sequence. The individual segments can be assembled by retrieval software so that users can view on demand the complete genome, chromosome, or other unit of interest.

**3. Data representation and analysis**

WebEntrez provides integrated access to the data of various types (single sequence, integrated maps, alignments and tables of markers and protein products) in graphical and text view modes [3]. The user can visualize the sequence information at varying levels of detail: the large-scale organization of very long stretches of contiguous DNA, a region of chromosome of particular interest, or an individual gene. Basically three groups of actions on every level of representation: navigation, reports and analysis; each specific for that level.

***3.1. Reports***

WebEntrez provides integrated access to the data of various types (single sequence, integrated maps, alignments and tables of markers and protein products) in graphical and text view modes [3]. The user can visualize the sequence information at varying levels of detail: the large-scale organization of very long stretches of contiguous DNA, a region of chromosome of particular interest, or an individual gene. Basically three groups of actions on every level of representation: navigation, reports and analysis; each specific for that level.

***3.2. Table of annotated proteins (ProtTable)***

ProtTable is a list of proteins derived from the annotated genomic sequence (or selected region). The information presented in this table includes the position of the coding region, the length of the protein, gene locus name, gene synonym, protein standard name and protein description or function and protein sequence identification number PID. Using the browser find mechanism one can search for a protein of a given function. A number of hot links provide the user with additional information. Clicking on a gene name shows the gene and surrounding region on the main graphical view. The PID is linked to the text view of the sequence recorded in Entrez protein database. One can also get a GenBank format view of the nucleotide sequence that contains the selected coding region or a FASTA format of the region. The FASTA view page provides the opportunity to submit the nucleotide or protein sequence of the selected gene directly to BLAST [5]. The whole set of proteins can be saved to a file in nucleotide or protein FASTA format and used as a database for further research. ProtTable is generated in a separate window in addition to protein graphical view for the same region shown in the main window. Clicking on a gene name will highlight the selected gene on the graphical view.

***3.3. Taxonomic distribution of protein homologs (TaxTable)***

TaxTable is a table generated (precomputed) for complete microbial genomes. All proteins of a selected genome are searched against the nonredundant (nr) database. The detected homologs are classified into three sets representing the primary domains of life (Archaea, Eubacteria, Eukaryota). Each column includes the GI number for the most similar protein for the respective domain and the corresponding random expectation value (in parentheses). The annotation is extracted from the GenBank records. The number of hits with e-value below 1e-04 is indicated for each domain. The table reveals biologically relevant sequence similarities and therefore may be used for functional prediction and evolutionary studies.

***3.4. ORFFinder***

The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool that finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database. This tool identifies all open reading frames using the appropriate genetic codes. The deduced amino acid sequence can be searched against the sequence database using the WWW BLAST server. For sequence records already in the database annotated proteins will be marked with a different color. One may restrict the search for open reading frames to intergenic regions only. One may choose to use a precomputed table for best BLAST hits in intergenic regions for the whole genome and starting from that table select the region for ORF Finder.

***3.1. Pattern search***

Pattern search is implemented in a number if ways. The user-defined pattern can be found in a selected region of sequence or from the whole genome. The search against functional site databases like PROSITE database or BLOCKS database is relevant for the region or individual gene. The results of pattern search are shown in a graphical view. Patterns can describe small motifs or larger regions containing several motifs and can also contain gaps.

**4. Summary**

We describe an integrated and interactive approach to whole genome exploration and analysis. Starting from a graphical overview of a whole genome one can locate a region of interest by a variety of means and get all the information available in the database and related Web resources in a large number of useful forms. In the same pages, direct access is provided to a variety of genome analysis tools to carry the process from genome data retrieval to new scientific discovery by the end user.

**Acknowledgements**

The World Wide Web Entrez Genomes project represents the efforts of many NCBI staff members including Jinghui Zhang, John Kuzio, Greg Schuler, along with the collective contributions of many dedicated scientists world wide.

**References**

1. Fleischmann RD et al Science 269(5223), 496-512 (1995)
2. Kuzio, J. Trends Genet. 12, 321-322 (1996)
3. Benson D.A., Boguski M., Lipman D.J. and Ostell J. Nucleic Acids Res. 25, 1-6 (1997)
4. Pearson W.R. and Lipman D.J. Proc. Natl. Acad. Sci. USA 85, 2444-2448 (1988)
5. Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman Nucleic Acids Res. 25:3389-3402 (1997)