BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 18 2005, pages 3669–3671
doi:10.1093/bioinformatics/bti606

Genome analysis

GenColors: accelerated comparative analysis and annotation of prokaryotic genomes at various stages of completeness

Alessandro Romualdi[1], Roman Siddiqui[2], Gernot Glöckner[2], Rüdiger Lehmann[2] and Jürgen Sühnel[1,*]

[1]Biocomputing Group and [2]Genome Analysis, Institute of Molecular Biotechnology, Beutenbergstrasse 11, 07745 Jena, Germany

Received on February 22, 2005; revised on revised May 10, 2005; accepted on July 28, 2005
Advance Access publication August 2, 2005

ABSTRACT

Summary: GenColors is a new web-based software/database system aimed at an improved and accelerated annotation of prokaryotic genomes, considering information on related genomes and making extensive use of genome comparison. It offers a seamless integration of data from ongoing sequencing projects and annotated genomic sequences obtained from GenBank. The genome comparison tools determine, for example, best-bidirectional hits, gene conservation, syntenies and gene core sets. Swiss-Prot/TrEMBL hits allow annotations in an effective manner. To further support the annotation base-specific quality data can also be displayed if available. With GenColors dedicated genome browsers containing a group of related genomes can be easily set up and maintained. It has been efficiently used for *Borrelia garinii* and is currently applied to various ongoing genome projects.

Availability: Detailed information on GenColors is available at http://gencolors.imb-jena.de. Online usage of GenColors-based genome browsers is the preferred application mode. The system is also available upon request for local installation.

Contact: jsuehnel@imb-jena.de

1 INTRODUCTION

As of January 18, 2005 the GOLD database lists 197 completed and 508 ongoing bacterial genome projects (Bernal *et al*., 2001). This huge and quickly growing amount of information can substantially contribute to an acceleration and improvement of the annotation process of newly sequenced genomes by genome comparison (Bentley and Parhkill, 2004). However, there are only a few tools available that offer an up-to-date information on prokaryotic genomes with an emphasis on genome comparison, one example being coliBASE (Chaudhuri *et al*., 2004). We have developed and describe here the software/database system GenColors that employs extensive genome comparison for accelerated and accurate annotation of prokaryotic genomes. Initially, GenColors (GENome COmparison by LOw Redundant Sequencing) was designed for the annotation and analysis of new genomes obtained by low redundancy sequencing. However, the actual features make GenColors a valuable tool for the analysis, presentation and annotation of bacterial genomes from the earliest to the final stages of a sequencing project.

2 GENCOLORS FEATURES

The analysis and display options of GenColors fall into three categories: General information, Search and Genome comparison. General information includes a Summary statistics and Gene lists, the browsing across all genomes according to the COG functional classification (Tatusov *et al*., 1997), and a Genome plots option that generates linear and circular plots of complete genomes considering a variety of genome features. Gene lists can be generated for complete genomes, individual genomic elements and as search output. In most cases it is possible to compile and store user-defined gene lists for further work.

The most detailed information on a gene can be found on the Annotation sheets. They start on top with a Gene environment graph (Fig. 1, top). In the Basepair view (Fig. 1, bottom) the DNA bases of both strands, their translation in the six frames and numerical confidence [given as Phred values, Ewing *et al*. (1998)] and coverage data are shown. The menu bar offers information on Best bidirectional hits (BBHs), Gene conservation, Syntenies, Swiss-Prot or TrEMBL hits (Boeckmann *et al*., 2003), DNA or protein sequence BLAST hits (Altschul *et al*., 1997) within the browser database, and Codon and amino acid usage. In the central part of the sheet general gene information is provided that is obtained either from the corresponding GenBank file or from provisional annotations for an ongoing genome project. If the corresponding protein sequence is included in Swiss-Prot its description and all the external database links are also shown here. In the lower part of the annotation sheet BBHs to all other genomic elements included in the browser database and the five best Swiss-Prot/TrEMBL hits as well as related InterPro (Mulder *et al*., 2005) and Gene Ontology (The Gene Ontology Consortium, 2000) classifications are indicated.

There is a simple Quick search option for gene names, descriptions or locus tags as well as an Advanced search option that allows the

[*]To whom correspondence should be addressed.

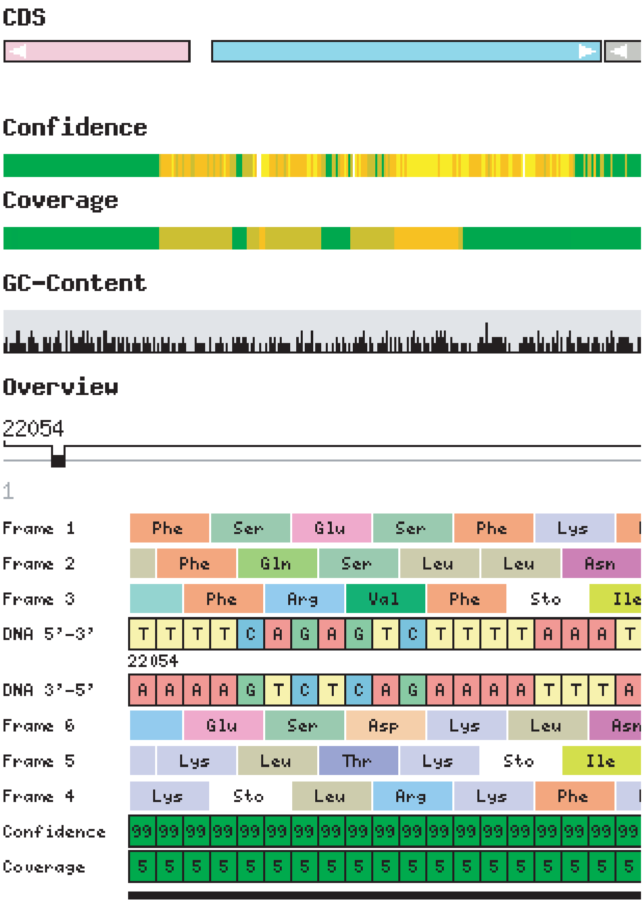
CDS
Confidence
Coverage
GC-Content
Overview
22054
1
Frame 1 Phe Ser Glu Ser Phe Lys
Frame 2 Phe Gln Ser Leu Leu Asn
Frame 3 Phe Arg Val Phe Sto Ile
DNA 5'-3' T T T T C A G A G T C T T T T A A A T
22054
DNA 3'-5' A A A A G T C T C A G A A A A T T T A
Frame 6 Glu Ser Asp Lys Leu Asn
Frame 5 Lys Leu Thr Lys Sto Ile
Frame 4 Lys Sto Leu Arg Lys Phe
Confidence 99 99 99 99 99 99 99 99 99 99 99 99 99 99 99 99 99 99 99
Coverage 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5

Fig. 1. Section of the Gene environment graph (top) and the Basepair view (bottom). Genes are colored according to the COG functional classification. Color-coded bars indicate the technical quality of the sequence.

combination of 16 different data types. They include gene identifiers/description, gene lengths, quality data, genomes or genomic elements, COG categories, PROSITE sequence patterns (Hulo *et al.*, 2004) and the complete external database information provided by Swiss-Prot. Sequence-based searches are done via BLAST, both for DNA sequences and for amino acid sequences.

Genome comparison tools constitute the major part of the GenColors system, many of them being based on BBHs. These are defined as best BLAST hits between all protein sequences of two genomic elements that have at least 30% sequence identity and where the length of the matching region spans at least 30% of the query length. A statistical analysis of sequence differences between two BBH proteins provides information on amino acid differences (insertions/deletions and duplications) as well as on synonymous and non-synonymous base exchanges. Gene core sets are defined as groups of genes with BBHs for all possible pairs of organisms in the data source. Synteny groups consist of two or more neighboring genes in one genomic element that have neighboring BBHs in another genomic element of either the same or a different species. Neighboring genes may be interrupted by up to five genes that have no BBHs in the counterpart genomic element. For pairs of complete genomes/genomic elements the results are shown as graphical representation and in tabular form. For individual genes tables with Synteny groups containing the gene under study are displayed for all possible pairs of genomic elements. The Gene conservation option shows for each gene if there are BBHs to other genes in any of the genomes/genomic elements included in the browser database and if these genes are synteny group members. GenColors performs also a statistical analysis of (start) codon usage and of amino acid composition for one or two genomes. Finally, global alignments obtained from an assembly program such as GAP4 (Bonfield *et al.*, 1995) or from any other programs can be used for the generation of detailed tables and figures reporting the identified sequence differences. The resulting tables contain also subsections for intragenic and extragenic regions as well as for each genome feature present on the genomic element.

Some of the GenColors features bear resemblance to the Artemis/ACT system (Berriman and Rutherford, 2003). Note, however, that contrary to GenColors no database is required to use Artemis. So, we consider Artemis a useful supplementary tool to GenColors.

3 ANNOTATION WITH GENCOLORS

The Annotation sheets represent the main starting point for gene annotation. Users can edit and change the annotation for each gene. Most of the tools required for annotation such as BLAST scans of the Swiss-Prot/TrEMBL and of the local browser database as well as analyses on gene conservation and syntenies can be directly accessed. Moreover, the display of the five best Swiss-Prot/TrEMBL hits and of the related InterPro and Gene Ontology information may accelerate the annotation process substantially. For genomes that are already annotated the original GenBank description may have been superseded by a more recent Swiss-Prot/TrEMBL annotation. Therefore, this information is also given. In the Advanced search option it is also possible to generate complete gene lists with information on BBHs and on the five best Swiss-Prot/TrEMBL hits. The annotation is further facilitated by an option where in a BBH list the description of a gene can easily be transferred to the related gene to be annotated by simply clicking on a transfer button.

4 DATA FLOW

GenColors imports files in GenBank format either directly from GenBank or from assembly programs, such as GAP4, in the case of ongoing sequencing projects. Quality data can be entered from the assembly programs using a tab-delimited table format. After various analyses and preliminary annotations sequence data of an ongoing genome project can be returned to the assembly program for further gap closure. We have developed the GenALA toolkit facilitating the data flow between the assembly program GAP4 and GenColors (Lehmann, unpublished data). This iterative process is performed until the final annotated version of the genomic sequence is obtained. Export filters from GenColors in either GenBank or in Sequin tab-delimited table formats are integrated. DNA or protein sequences from gene lists can be exported as multi-FASTA files.

5 IMPLEMENTATION

GenColors includes 85 Perl scripts, 4 Perl modules and 3 relational databases. It requires Apache, MySQL, BioPerl (Stajich *et al.*, 2002) and EMBOSS (Rice *et al.*, 2000) and is running on a SuSE Linux 9.2 server. For speeding up server response some analyses and scans are pre-computed. Automated procedures manage the download of the most recent versions of the Swiss-Prot and TrEMBL databases and the necessary BLAST scans. GenColors provides a user management with different data access privilege levels, but it may also be used for setting up free access browsers.

6 APPLICATIONS

We have recently completed a genome sequencing project on *Borrelia garinii* (Glöckner *et al*., 2004), where the GenColors system has been applied. GenColors allows the easy setup of dedicated genome browsers that include a group of related genomes. The first example, the Spirochetes Genome Browser (SGB) including also *B.garinii* is freely accessible at http://sgb.imb-jena.de

ACKNOWLEDGEMENTS

The authors thank Matthias Platzer and Markus Schilhabel for helpful discussions and critical reading of the manuscript. The help provided by Kerstin Wagner in setting up and maintaning the SGB external link page as well as in icon design is gratefully acknowledged. Friedrich Haubensak was instrumental in the server design. This work was supported by Grant 0312704E of the German Ministry for Education and Research. Funding to pay the Open Access publication charges for this article was provided by institutional funding of the Institute of Molecular Biotechnology.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bentley,S.D. and Parkhill,J. (2004) Comparative genomic structure of prokaryotes. *Annu. Rev. Genet.*, **38**, 771–791.

Bernal,A. *et al*. (2001) Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res.*, **29**, 126–127.

Berriman,M. and Rutherford.K. (2003) Viewing and annotating sequence data with Artemis. *Brief. Bioinformatics*, **4**, 124–132.

Boeckmann,B. *et al*. (2003) The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Bonfield,J.K. *et al*. (1995) A new DNA sequence assembly program. *Nucleic Acids Res.*, **23**, 4992–4999.

Chaudhuri,R.R. *et al*. (2004) coliBASE: an online database for *Escherichia coli*, *Shigella* and *Salmonella* comparative genomics. *Nucleic Acids Res.*, **32**, D296–D299.

Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.

Glöckner,G. *et al*. (2004) Comparative analysis of the *Borrelia garinii* genome. *Nucleic Acids Res.*, **32**, 6038–6046.

Hulo,N. *et al*. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Res.*, **32**, D134–D137.

Mulder,N.J. *et al*. (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201–D205.

Rice,P. *et al*. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.

Stajich,J.E. *et al*. (2002) The Bioperl Toolkit: perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.

Tatusov,R.L. *et al*. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.