

Marius Felder,^{1,2} Alessandro Romualdi,³ Gernot Glöckner,² Matthias Platzer² and Jürgen Sühnel¹

¹ Biocomputing and ² Genome Analysis Groups, Leibniz Institute for Age Research – Fritz Lipmann Institute: FLI,

Jena Centre for Bioinformatics, Beutenbergstr. 11, D-07745 Jena, Germany

³ SIRS-Lab GmbH, Winzerlaer Str. 2, D-07745 Jena, Germany

Introduction

GenColors [1,2] 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. With GenColors dedicated genome browsers containing a group of related genomes can be easily set up and maintained. The tool has been efficiently used for *Borrelia garinii* [3,4] and is currently applied to various ongoing genome projects on *Legionella*, *Pseudomonas* and *E. coli* genomes. Examples for freely accessible GenColors-based dedicated genome browsers are the Spirochetes Genome browser SGB (sgb.fli-leibniz.de), the Photogenome Browser CGB (cgb.fli-leibniz.de) and the Enterobacter Genome Browser ENGINE (engine.fli-leibniz.de). The system has now been adapted to handle also eukaryotic genomes. A first application of this feature is the ongoing annotation and analysis of two fungal species. Another GenColors-based tool is the Jena Prokaryotic Genome Viewer JPGV (jpgv.fli-leibniz.de). Contrary to the dedicated browsers it offers information on almost all finished bacterial genomes. As of July 10, 2008 it includes 1140 genomic elements of 293 species.

The Browser

The available analysis and display options fall into three categories: General information, Search and Genome comparison.

The Spirochetes Genome Browser

Home Genomes Methods Documentation External links Options Contact FLI Jena

QuickSearch Perfect match Go 2008-05-07 14:54:31

General information

- Gene lists**: Browse gene lists of selected genomes.
- COG functional categories**: Select genes of specific COG functional categories across genomes.
- Genome plots**: Generate circular and linear plots.
- (Clickable) linear whole genome views**: Display a clickable linear view of all genes of a genomic element coloured according to the COG functional classification.
- Genome lists**: Generate genome lists of selected genomes.
- Horizontal gene transfer (predicted by SigI-HMM)**: Get information on genes of a genome acquired by horizontal gene transfer.
- Sequence retrieval**: Retrieve genomic sequences or intergenic regions of selected genomic elements.

Search

- Advanced search**: Search for organisms, genes, COGs and external database information (short output with gene name, locus tag, description, location, strand information. Comprehensive output with additional information on best-bidirectional protein hits in all SGB genomes and UniProt and TrEMBL hits).
- Sequence similarity search**: Identify sequence similarities between any sequence stretches and all SGB genomes.

Genome comparison

- Best bidirectional hits (protein-based analysis)**: Generate a list of best bidirectional hits between the proteins of two genomes.
- Gene core sets (protein-based analysis)**: List all related genes of two or more genomes, genomic elements or user defined lists.
- Gene partnerships (DNA-based analysis)**: The alignment of two collinear genomic elements allows the identification of potential gene relationships by their similar localisation. This analysis can identify related gene pairs that cannot be found as protein-based bidirectional hits.
- Protein variations (protein-based analysis)**: List and analyse all best bidirectional protein hits.
- Synteny analysis (protein-based analysis)**: Analyse synteny between different genomes based on the best bidirectional protein hits.
- Gene conservation (protein-based analysis)**: Analyse the gene conservation between one selected genome and all other SGB genomes.
- Codon and amino acid usages (DNA- and protein-based analysis)**: Show or compare the codon and amino acid usages of one or more genomes or genomic elements.
- Alignments of whole reference and target genomic elements (DNA-based analysis)**: Analyse the DNA alignment statistics of reference and target genomes.

You are logged in as alex [Logout](#)

Figure 1: Methods page of the Spirochetes Genome Browser

General information

GenColors offers browsing across whole genomes (Figure 2, left), single genomic elements or according to the COG functional classification (Figure 2, right).

Borrelia garinii, PBI

Generate gene lists for all or specific genomic elements or COG functional categories. Show summary table. Download as GenBank file.

Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130

Generate gene lists for all or specific genomic elements or COG functional categories. Show summary table.

Unclassified [-]

RNA processing and modification [A]

Chromatin structure and dynamics [B]

Energy production and conversion [C]

Cell cycle control, mitosis and meiosis [D]

Amino acid transport and metabolism [E]

Figure 2: Browsing options in SGB

Search

GenColors basically offers two ways of querying the underlying database. There is a QuickSearch option for text strings in gene names, descriptions or locus tags as well as an AdvancedSearch option that allows the combination of 20 different data types. These include gene identifiers/description, gene lengths, genomes or genomic elements, COG categories, PROSITE sequence patterns and the complete external database information provided by UniProt. Sequence based searches are done via BLAST.

Search by locus tag

Search by gene description

Search by gene note

Search by gene length (please use '<', '>' or '=')

Search by protein sequence motif (Prosite format)

Select by external database identifier Aarhus/Ghent-2DPAGE

Figure 3: AdvancedSearch in GenColors (selected categories)

Gene information sheet

The most detailed information on a gene can be found on the Gene information sheets (Figure 4). These sheets start on top with a Gene environment graph. The DNA bases of both strands and their translation are displayed in the Basepair view. In the central part of the sheet general gene information is provided. If the corresponding protein sequence is included in UniProt its description and all the external database links are also shown here.

CDS

Confidence

Coverage

GC-Content

Overview

887770 89476 90424

1

Frame 1 Gly Ala Lys Leu Lys Leu Leu Thr Phe Thr Phe Phe Phe Leu Gln Phe Leu Leu Arg Leu Phe Leu Pro Arg

Frame 2 Val Pro Arg Ser Phe Phe Phe Ile Phe Ile Phe Ser Cys Arg Arg Phe Ile Thr Ile Thr Phe Tyr Gln Ile

Frame 3 Cys Gln Ile Glu Thr Ser Leu Phe Leu Phe Phe Leu Ala Ile Ser Phe Glu Ser Ile Ser Thr Lys

DNA 5'-3' C C A G C C A A A T T G A A C T T G T T A T T T T A T T T T T G T G C A A T T T C T T T T A G C A A A T

DNA 3'-5' G G C C G G C C A A A T T G A A C T T G T T A T T T T A T T T T T G T G C A A T T T C T T T T A G C A A A T

Frame 6 Ala Leu Arg Phe Phe Arg Ala Lys Lys Lys Cys Arg Arg Phe Arg Lys Phe Arg Arg Glu Phe

Frame 5 His Thr Ile Ser Val Glu Lys Lys Lys Arg Ala Ile Lys Arg Lys Ser Arg Ile Glu Val Leu Arg

Frame 4 Thr Gly Phe Gln Phe Lys Lys Lys Ile Lys Lys Glu Gln Leu Lys Lys Gln Ile Phe Lys Thr Thr Ile

Description: exoinuclease ABC, subunit B

Locus tag: B60861

Strand: +

Begin: 890270

End: 892264

Figure 4: Gene environment graph and basepair view

Genome comparison

Genome comparison tools constitute the major part of the GenColors system. Many of them are 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. BBH lists can further facilitate annotation as the description of a related gene can easily be transferred to the gene to be annotated by simply clicking on a transfer button.

Protein-coding genes in: **Borrelia garinii**

Name	Locus tag	GenBank description
NA	BGB01	conserved hypothetical protein
NA	BGB02	conserved hypothetical protein

BBH in: **Borrelia burgdorferi**

View	Transfer alignment name	Transfer description	Name	Locus tag	GenBank description
<input type="checkbox"/>			BBB01		conserved hypothetical protein
<input type="checkbox"/>			BBB02		B. burgdorferi predicted coding region BBB02

Figure 5: List of best bidirectional hits

Align-Reference	Name / Locus tag / GenBank description	Query	Name / Locus tag / GenBank description	Ins	Del	Dup	Exc	Sd	Nd	S	N	ps*	pn*	ds*	dn*	ds/dn*	(Sd-Nd)/(Sd+Nd)
<input type="checkbox"/>	BB0002: beta-N-acetylhexosaminidase, putative (1029 bp, 342 aa)	<input type="checkbox"/>	NAV/BG0002: beta-N-acetylhexosaminidase, putative (1020 bp, 339 aa)	0	1	0	23	51	28	204	813	0.25	0.034	0.304	0.035	8.625	0.291
<input type="checkbox"/>	BB0003: B. burgdorferi predicted coding region BB0003 (1365 bp, 454 aa)	<input type="checkbox"/>	NAV/BG0003: hypothetical protein (1479 bp, 492 aa)	4	0	0	68	70	80	263	1098	0.267	0.073	0.33	0.077	4.262	-0.067

Figure 6: List of protein variations

Synteny groups consist of two or more neighbouring genes in one genomic element that have neighbouring BBHs in another genomic element of either the same or a different species. Neighbouring genes may be interrupted by up to five genes that have no BBH in the counterpart genomic element.

28: Treponema denticola ATCC 35405

200: Borrelia garinii

Figure 7: Overview (left) of all synteny groups identified between the chromosomes of *Treponema denticola* ATCC 35405 and *Borrelia garinii*. Detailed views of two synteny groups (right) with conserved and inverted gene order.

Genome plots

GenColors generates circular and linear genome plots (PNG, PDF, Postscript) using annotation data and calculating quantities such as GC-content, GC-skew and purine and keto excess. Different features of one or more genomes can be displayed in one representation facilitating comparative analysis.

4446355 GC content 193319

4252035

4059715

3862395

3673076

3478766

3286436

3093196

2899796

2706477

2513157

2318837

2126517

1933197

1739878

1548258

1353238

1168918

966598

779959

579959

375279

168639

Escherichia coli K12

Jena Prokaryotic Genome Viewer

Fig 8: Circular plot of features of the *Escherichia coli* K12 chromosome: different gene types and precomputed data (GC-content, purine excess) are displayed on different orbits. Colouring of the CDS orbits corresponds to COG functional categories.

References

[1] Romualdi, Siddiqui, Glöckner, Lehmann, Sühnel. GenColors: accelerated comparative analysis and annotation of prokaryotic genomes at various stages of completeness. *Bioinformatics* **2005**, 15, 3669-71.

[2] Romualdi, Felder, Rose, Gausmann, Schilhabel, Glöckner, Platzer, Sühnel. GenColors: Annotation and Comparative Genomics of Prokaryotes Made Easy. *Methods Mol. Biol.* **2007**, 395, 75-96.

[3] Glöckner, Lehmann, Romualdi, Pradella, Schulte-Spechtel, Schilhabel, Wilske, Sühnel, Platzer. Comparative analysis of the *Borrelia garinii* genome. *Nucleic Acids Res.* **2004**, 32, 6038-46.

[4] Glöckner, Schulte-Spechtel, Schilhabel, Felder, Sühnel, Wilske, Platzer. Comparative genome analysis: selection pressure on the *Borrelia vls* cassettes is essential for infectivity. *BMC Genomics* **2006**, 7, 211.