The Microbe browser for comparative genomics

Alexandre Gattiker1,2, Christophe Dessimoz1,3, Adrian Schneider1,3, Ioannis Xenarios1,4, Marco Pagni1,4 and Jacques Rougemont1,2,*

1Swiss Institute of Bioinformatics (SIB), 2School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, 3Institute of Computational Science, ETH Zürich, 8092 Zürich and 4Vital-IT, Bâtiment Génopode, Université de Lausanne, 1015 Lausanne, Switzerland

Received January 27, 2009; Revised April 3, 2009; Accepted April 11, 2009

ABSTRACT
The Microbe browser is a web server providing comparative microbial genomics data. It offers comprehensive, integrated data from GenBank, RefSeq, UniProt, InterPro, Gene Ontology and the Orthologs Matrix Project (OMA) database, displayed along with gene predictions from five software packages. The Microbe browser is daily updated from the source databases and includes all completely sequenced bacterial and archaeal genomes. The data are displayed in an easy-to-use, interactive website based on Ensembl software. The Microbe browser is available at http://microbe.vital-it.ch/. Programmatic access is available through the OMA application programming interface (API) at http://microbe.vital-it.ch/api.

INTRODUCTION
About a thousand complete microbial genomes have been sequenced to date [961 genomes in the Genomes On Line Database (GOLD) on 1 April 2009 (1)], and many different methods have been used to predict genes, yielding large differences in gene annotation even across closely related species. No single computational method yet achieves perfect gene predictions. Furthermore, very few entries have been kept up-to-date in the primary databases such as GenBank (2). We therefore felt that it was important to provide a unified interface to the various gene prediction packages to allow biologists to evaluate them in their genomic and evolutionary contexts.

This leads to another important computational challenge, namely the identification of orthologs. Many studies, such as the prediction of gene function, phylogenetic reconstruction and genomics context analyses, depend on accurate predictions of orthology. Among genes that share a common ancestor, only genes that are separated by a speciation event are actual orthologs (3). To address the need for reliable ortholog sources, several initiatives have been created for better ortholog prediction [see (4) for a comparison]. Among these resources, Orthologs Matrix Project (OMA) stands out by its efficient and robust computational method allowing continuous updating with novel genomes (5) and its ability to exclude non-orthologs, conferring a high reliability in the prediction of true orthologous relationships (4).

Interactive genome browsers have proved invaluable to the community for visualizing genes and experimental data in their genomic context, and as hubs connecting many biomedical databases (6,7). Genome browsers also provide comparative genomics information by displaying homologous regions in a single view. However, most browsers concentrate on eukaryotic genomes, so that biologists working on microbial genomes are restricted to standalone programs such as the Artemis Comparison Tool (8) or web sites such as the Joint Genome Institute’s Integrated Microbial Genomes tools (http://img.jgi.doe.gov/) or GeneDB (http://www.genedb.org) that are more complex to use, can only handle a few genomes at a time and do not integrate as much information via a single interface.

Derived genomic databases that connect and expand reference databases are important in particular for automated analyses such as dataset comparisons. The EBI Genome Reviews database (9) provides complete genome sequence and annotation data, continuously updated and extended with automated and manual annotation in UniProtKB (10). The NCBI RefSeq resource (11) provides a coherent set of sequences, genes and transcripts, some of which have been manually annotated. Frustratingly, the EBI and NCBI resources use distinct sets of identifiers (UniProtKB accession number and protein_id for EBI; Refseq accession number, GeneID and GI number for NCBI) that make it hard to navigate between databases using different references. Furthermore, UniProtKB curators not only extend and uniformize annotation, but they also modify gene sequences, changing translational start site predictions, correcting frameshifts or adding genes missing from the original submission. This information is propagated to Genome Reviews but not to the source DDBJ/EMBL/GenBank entries, which can only be modified by the original

*To whom correspondence should be addressed. Tel: +41 21 693 9573; Fax: +41 21 693 1850; Email: jacques.rougemont@epfl.ch

© 2009 The Author(s)

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
submitter. This introduces an additional divergence
between databases, as it becomes non-trivial to identify
the ‘same’ gene in two different databases where the
gene might have neither the same identifier scheme nor
the same coordinates.

The Integr8 database (9) aggregates curated informa-
tion on completely sequenced genomes, including taxon-
omy down to the precise strain level, and cross-references
to all chromosomes and plasmids comprising the complete
genome.

We introduce the Microbe browser, a web server that
uses the Integr8 database to organize and correlate geno-
mic sequences and annotation from the GenBank,
Genome Reviews and RefSeq databases. We use the
powerful Ensembl web code (7) to present the resulting
data in a fully interactive, user-friendly and platform-
independent manner.

METHODS

Source data are retrieved daily from primary public
servers. Integr8 and Genome Reviews are the source of
genome data, including curated gene sets and annotation
and cross-references to UniProtKB, InterPro, Gene
Ontology and the Protein Data Bank. GenBank and
RefSeq are the source of NCBI cross-references (RefSeq
accession, GeneID and GI number). The OMA database
provides orthology predictions for pairs of genes. Pre-
computed gene predictions from the Glimmer (12),
GeneMark, GeneMarkHMM (13) and Prodigal (http://
compbio.ornl.gov/prodigal) packages are provided by
the NCBI, and predictions by the EasyGene method (14)
are downloaded from the EasyGene web site (http://
servers.binf.ku.dk/cgi-bin/easygene/search).

The Genome Reviews data are used as a reference,
because it incorporates substantial automatic and
manual annotation from the gold standard UniProtKB
knowledgebase (10). Cross-references from GenBank
and RefSeq genes are merged into Genome Reviews
records based on the position of the 3'-end of the genes.
This allows to correctly map not only genes for which no
cross-references exist between the databases, but also
those for which the 5'-end (start site) has been possibly
changed by UniProtKB curators.

USAGE

The Microbe browser home page is used for organism
selection and search term input, which can be a gene
name or a cross-reference to any of the source databases.
Several view pages are available, the three most informa-
tive are detailed below. The user can easily navigate across
those pages and detailed online help is available.

The gene report page integrates data on gene sequence
and annotation, orthologs and cross-references to the
major biological databases.

The chromosome view pages (Figure 1) display the
original genome annotation submitted in the DDBJ/
EMBL/GenBank source databases, the modified annota-
tion from UniProtKB (via Genome Reviews) and the gene
predictions of several popular packages.

The chromosome comparison pages (Figure 2) display
regions surrounding orthologous genes in two or more
organisms, highlighting orthology relationships between
them, and reveal cases of synteny (co-localized orthologs).
This display scales up to comparing a few species with
detailed positional information, while specialized software
has been proposed to visualize synteny across dozens of species in a summarized display (15).

For software developers, programmatic access to the orthology relationships is available via web services through the OMA APIs at http://microbe.vital-it.ch/api.

CONCLUSION

Designed primarily for biomedical researchers, the Microbe browser runs an easy-to-use, interactive view allowing to visualize gene predictions, orthology and synteny relationships and to navigate across databases. Data originates from established bioinformatics databases: DDBJ/EMBL/GenBank source genomic data, annotation and cross-references to the major biological databases retrieved from Genome Reviews and RefSeq, pairwise gene orthology predictions from OMA, and alternative gene predictions from several prediction packages. Future developments will include fungal genomes and metagenomic data.

ACKNOWLEDGEMENTS

We thank R. Fabbretti and V. Flegel for IT support, and also A. Auchincloss, T. Lima and A. Kapopoulou for critical reading.

FUNDING

Swiss Institute of Bioinformatics. Funding for open access charge: Swiss Institute of Bioinformatics and Ecole Polytechnique Fédérale de Lausanne.

Conflict of interest statement. None declared.

REFERENCES


