Bioinformatics, 32(7), 2016, 1083–1084
doi: 10.1093/bioinformatics/btv689
Advance Access Publication Date: 24 November 2015
Applications Note

OXFORD

Genome analysis

Synchronized navigation and comparative analyses across Ensembl complete bacterial genomes with INSYGHT

Thomas Lacroix[1,*], Sylvie Thérond[2], Marc Rugeri[2], Pierre Nicolas[1], Annie Gendrault[1], Valentin Loux[1] and Jean-François Gibrat[1]

[1]INRA, UR1404, Unité Mathématiques et Informatique Appliquées du Génome à l'Environnement, 78350 Jouy-en-Josas, France and [2]IDRIS-CNRS, UPS851, 91400 Orsay, France

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on September 11, 2015; revised on November 12, 2015; accepted on November 16, 2015

Abstract

Motivation: High-throughput sequencing technologies provide access to an increasing number of bacterial genomes. Today, many analyses involve the comparison of biological properties among many strains of a given species, or among species of a particular genus. Tools that can help the microbiologist with these tasks become increasingly important.

Results: Insyght is a comparative visualization tool whose core features combine a synchronized navigation across genomic data of multiple organisms with a versatile interoperability between complementary views. In this work, we have greatly increased the scope of the Insyght public dataset by including 2688 complete bacterial genomes available in Ensembl thus vastly improving its phylogenetic coverage. We also report the development of a virtual machine that allows users to easily set up and customize their own local Insyght server.

Availability and implementation: http://genome.jouy.inra.fr/Insyght

Contact: Thomas.Lacroix@jouy.inra.fr

1 Introduction

Faced with an ever-increasing number of sequenced genomes, biologists need efficient and user-friendly tools to assist them in their analyses. In this context, tools that facilitate comparative genomics analyses (i.e. conservation of gene neighborhood, presence/absence of orthologous genes, phylogenetic profiling, etc.) of large amounts of data are needed.

We have recently developed Insyght, a comparative genomic visualization tool (Lacroix *et al.*, 2014) that tightly integrates three complementary views: (i) a table for browsing among homologs, (ii) a comparator of orthologs' functional annotations and (iii) a genomic organization view that improve the legibility of genomic rearrangements and distinctive loci. Insyght benefits from an easy and smooth navigation between these three views and provides users with a powerful search mechanism.

In Lacroix *et al.* (2014), we focused mainly on Insyght visualization capabilities and functionalities but the underlying database contained, at that time, only 400 complete bacterial genomes. Here, we have processed the 2688 complete bacterial genomes available in Ensembl Bacteria when we started this work (June 2014) to provide a more comprehensive coverage of the microbial diversity. We have also developed a virtual machine that allows users to easily set up their own local Insyght server to analyze their genomic data privately.

2 Methods

2.1 Database

Original genome files were obtained from EBI Ensembl Bacteria (Kersey *et al.*, 2014). Data are stored in a PostgreSQL relational

database. This database contains three types of data: primary data such as genomic annotations extracted from the genome files, secondary data that result from the cross comparison of the proteomes using BLASTp, and tertiary data such as the regions of synteny.

2.2 Pipeline

The database is populated by a pipeline of Perl scripts that (i) process the genome files, (ii) run the BLASTp jobs on a supercomputer, (iii) parse the results and (iv) execute the program that identify the syntenies between the >3.5 million pairs of genomes. Two protein coding genes in distinct genomes are considered orthologous if they give rise to a bi-directional best hit (BDBH) when comparing the corresponding proteomes and if the sequence alignment includes more than 50% of the length of the shortest proteins with an *e*-value less than 0.01. Two proteins for which the e-value of the sequence alignment is less than 0.01 and that do not fall into the 'ortholog' category are considered homologous but are not displayed unless they belong to a synteny. Syntenies are delineated with a dynamic programming algorithm that determines the highest scoring paths among all the possible gapped chains of collinear homologues. The scoring scheme that we use (minimum synteny score: 4; ortholog: 4; homolog: 2; mismatch: −4; gap creation: −4; gap extension: −2) allows for the insertion of small gaps.

2.3 Web interface

Insyght makes use of the Asynchronous JavaScript and XML (AJAX) technology to minimize data transfer between server and clients, send simultaneous server requests and transfer most of the processing load on the client side. The graphical rendering uses the HTML5 canvas. We paid special attention to maintain a fast and smooth navigation in spite of a ∼32-fold increase of the database size.

3 Results

3.1 Public dataset

We inserted in the database the 2688 complete genomes of Ensembl Bacteria available early 2014, carried out the cross comparisons of their proteomes with BLASTp, and computed the synteny regions for all >3.5 million pairs of genomes. We report ∼2.4 billion orthologs; on average, their alignments have a coverage of ∼90%, an identity of ∼43%, a score of ∼241 and an *e*-value of $\sim 2e-05$. We report ∼1.85 billion singleton orthologs and ∼140 million syntenies that comprise ∼550 million orthologs and ∼190 million non-BDBH homologs. On average, alignments of non-BDBH homologs have a coverage of ∼76%, an identity of ∼35%, a score of ∼145 and an *e*-value of $\sim 8e-05$. BLASTp jobs generated 1.2 TB of raw, compressed data (tabular format). The database size is 3.5 TB, the largest table having ∼5.9 billion rows and occupying ∼2 TB with indexes. The dataset is available at http://genome.jouy.inra.fr/Insyght.

3.2 Web interface

The INSYGHT web interface is composed of three interconnected views. The 'orthologs' table view provides a familiar layout with genes as columns and organisms as rows. It allows to easily check for the presence, absence or multiple copies of homologs. Genes are colored according to the presence of a synteny region if any. It is possible to browse whole genomes, set of genes, or sort the table. The 'annotation comparator' view allows users to compare annotations (GO-terms) for a set of reference genes and their homologs. Annotations are classified into three categories: 'shared', 'unique' and 'missing'. It is possible to restrict the set of organisms considered or filter homologs using various criteria. The 'genomic context' view allows users to browse among homologies, conserved syntenies and loci insertions. The combination of a symbolic and a proportional (i.e. trapezoidal) representations facilitates the display of syntenies and insertion/deletion events at different scales along the genome. It is possible to synchronize the navigation among genomes.

We verified that the performances did not deteriorate when using the new, much larger, database. Start-up time and most loading times still take a few seconds at whole genome scale and for multiple comparisons.

3.3 Virtual machine

We provide a virtual machine (VM) to allow for users' private analysis in a local environment. Users can customize which genomes to compare, the blast parameters, etc. The VM is based on the Ubuntu OS and is pre-configured with all the software (Perl scripts, relational database, web application) and dependencies (PostgreSQL, tomcat, etc.) that are required to set up a private Insyght server. The '.ova' image file (∼3.3 GB) can be imported into a virtualization management environment (e.g. VirtualBox). Documentation on how to run the pipeline and manage the data is provided. Since our publication in 2014, the pipeline provided in the VM has been revamped: it is now ∼2.5× faster, suitable for draft genomes (many contigs), and allows analysis of different genomes with identical taxonomic identifiers.

4 Conclusion

In this note, we report the successful integration and availability of a non-trivial amount of complete genomes in Insyght. Nevertheless, it will be increasingly difficult for us to process the huge number of newly sequenced genomes, in particular with the advent of 3rd generation sequencing technologies that are capable of providing completely assembled genomes. To tackle this issue, we provide a customizable version of Insyght via the virtual machine described above that will allow the user to import part of our public database and to add a set of private genomes, either complete or in draft. In practice, we plan to restrict the public dataset to a single 'best' representative for each group of closely related genomes and to allow users to extract only species in a particular taxonomic node (of course limiting this possibility to 'manageable' amounts of transferred data).

Acknowledgements

The authors would like to thank the IDRIS-CNRS user support team for the help with the computations on the E-biothon BlueGene/P machine and the MIGALE platform for providing computer resources and support for the database and server.

Funding

The E-biothon platform is supported by CNRS, IBM, INRIA, SysFera and the Institut Français de Bioinformatique. IBM provided the BlueGene/P machine. PN received support from the ANR-12-ADAP-0018 PathoBactEvol.

Conflict of Interest: none declared.

References

Kersey,P.J. *et al.* (2014) Ensembl Genomes 2013: scaling up access to genome-wide data. *Nucleic Acids Res.*, **42**, D546–D552.

Lacroix,T. *et al.* (2014) Insyght: navigating amongst abundant homologues, syntenies and gene functional annotations in bacteria, it's that symbol!. *Nucleic Acids Res.*, **42**, 21.