



COmplete GENome Tracking (COGENT): a flexible data environment for computational genomics

Paul Janssen[†], Anton J. Enright[‡], Benjamin Audit, Ildelfonso Cases, Leon Goldovsky, Nicola Harte[‡], Victor Kunin and Christos A. Ouzounis*

Computational Genomics Group, The European Bioinformatics Institute, EMBL Cambridge Outstation, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK

Received on May 28, 2002; revised on January 24, 2003; accepted on February 10, 2003

ABSTRACT

Summary: We present a database of fully sequenced and published genomes to facilitate the re-distribution of data and ensure reproducibility of results in the field of computational genomics. For its design we have implemented an extremely simple yet powerful schema to allow linking of genome sequence data to other resources.

Availability: <http://maine.ebi.ac.uk:8000/services/cogent/>

Contact: genomes@ebi.ac.uk

INTRODUCTION

The number of completely sequenced genomes is constantly increasing, with over 80 entire genome sequences published to date. In principle, this unprecedented resource of genomic data opens up new opportunities for computational biology, by expanding the possibilities of genome-wide investigations. Yet, this data avalanche also challenges the current practices with respect to accessibility, reproducibility and useability. The capacity of linking the genomic sequence data with other information, for example functional classes, cellular localization, chromosomal position and expression profiles clearly leads to new knowledge. However, the combination of diverse data resources for genome analysis can be cumbersome, because of the wide variability of syntactic and semantic conventions deployed by different data repositories. Although some operations should be easy to perform, in practice a significant amount of time is spent to achieve them (Stein, 2002).

First, although there exist resources that allow full access to genome sequence information (Bernal *et al.*, 2001), there is a clear lack of a single point of reference that allows flexible and direct access to full-genome information with a few mouse clicks. Most of the web resources are not designed for genome-scale analysis, as

they were generally created for a ‘one gene at a time’ mode of browsing and, as a consequence, the raw data are not readily available. This information is necessary for large-scale computational analyses and should thus be conveniently accessible and re-distributable. It is frequently the case that computational biologists may be working on the same data set using different identifiers and may not be able to easily compare their results.

Furthermore, databases are tremendously heterogeneous so that it is in practice impossible to link their contents in an automated way. The primary reason is that the naming conventions for genes are not adequate and might differ from one resource to another. For example, the COGs database uses gene names as they can be found in EMBL or GenBank entries, e.g. AF1241 and BS_gsaB, found in COG0001 cluster. In this example, gsaB had to be prefixed by ‘BS’ to uniquely point to the *Bacillus subtilis* gsaB gene, since gsaB is not a unique identifier. In order to map this COG cluster to the InterPro domain database, an additional piece of information, such as the SwissProt identifier or accession number, is also necessary. Therefore, simply linking two commonly used databases for protein sequence annotation can be a challenging task.

Finally, even if the data, naming schemes and their mappings are fully available, file parsing and format conversion can still be a bottleneck in bioinformatics research. For example, the mapping of identifiers and other terms to keywords, various classification schemes or species names can be a challenge. In practice, every single user will have to write their own parser to effectively utilize this information. In our view, the difficulties mentioned above raise a major issue about consistency and reproducibility for computational biology. To overcome some of the obstacles that hinder consistency and reproducibility, we have developed a new data environment for complete genome analysis, called COGENT.

IMPLEMENTATION

Some ongoing projects already offer functional frameworks to share software development (e.g. <http://>

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

[‡]Present Address: External Services Group, EMBL-EBI.

//www.bioperl.org) or high-throughput data analysis. For example, the Ensembl project (<http://www.ensembl.org>) provides full access by distributing the software (Perl modules) to query its complex MySQL database either on a local installation or on their server (Hubbard *et al.*, 2002). The aim is to provide a simple and practical tool on which to develop computational genomics projects, to facilitate sharing of data and results and, hopefully, to enable synergy in the field. We are successfully using COGENT in our research group, so that we now advocate its use by the wider community.

COGENT can be considered a realistic solution, because it is simple, highly flexible with a small usage overhead. Working with COGENT entails the use of: (i) the core COGENT database of complete genome sequences, which defines a common naming convention for genomes and proteins (identifiers); and (ii) SQL tables as an exchange format, as well as a potential working environment. These few and simple guidelines allow the convenient exchange of results pertaining to complete genome sequences, necessary for high-throughput computation. Table joining enables the linking of results from different origins and the indexing mechanisms available in relational databases such as MySQL ensure efficient query and retrieval.

DESIGN

The complete design details are available on the COGENT web pages, along with Perl DBI-based scripts and modules to perform basic operations, such as retrieving a sequence or a complete genome peptide file in FASTA format. Nevertheless, it is noteworthy that the core database is composed of just two tables. The *genomes* table contains genome-related information (Table 1). A design decision was made to keep track of the relative timing of genome sequences in order of their appearance in the literature (*rel_order*). Along with the *genome_id*, a mnemonic *species_code* is generated and used as a prefix to construct unique protein identifiers. This mnemonic notation is based on the genus, species and strain name as well as version number for this genome, in order to enable updates e.g. HINF-KW2-01 corresponds to *Haemophilus influenzae* strain KW2, version 01. The *proteins* table holds the amino acid sequence data (Table 2). The unique *protein_ids* are constructed by the *species_code* followed by a dash and a number. The use of mnemonic identifiers enable to clearly distinguish genomes and proteins from a given genome in a simple manner. As new genomes are made available, they are added to the *genomes* and *proteins* tables by the database curators, upon release. Whenever possible, genome sequences are downloaded directly from the sequencing centre web sites—otherwise GenBank is used as a source.

We endeavour to maintain consistency between the content of the database and published complete genomes.

Table 1.

Field	Example
<i>rel_order</i>	1
<i>genome_id</i>	1
<i>fullname</i>	<i>Haemophilus influenzae</i> , KW20
<i>source</i>	TIGR
<i>date_sequenced</i>	28/07/95
<i>species_code</i>	HINF-KW2-01
<i>total_genes</i>	1707
<i>tax_class</i>	Bacteria; Proteobacteria; ...
<i>source_url</i>	ftp://ftp.tigr.org/pub/data/h_influenzae/
<i>size_mb</i>	1.83
<i>curator</i>	janssen
<i>date_added</i>	05/07/2001

Table 2.

Field	Example
<i>protein_id</i>	HINF-KW2-01-000001
<i>genome_id</i>	1
<i>old_name</i>	HI0001
<i>length</i>	339
<i>sequence</i>	MAIKIGINGFGRIG...
<i>annotation</i>	glyceraldehyde-3-phosphate ...

Currently (April 2003), the COGENT core database contains all 114 published complete genomes (89 Eubacteria, 15 Archaea and 10 Eukarya), with a total of more than 450 000 protein sequences.

USAGE AND FUTURE PLANS

The database has already been used effectively in a number of ongoing projects, including genome sequence clustering, genome annotation and phylogenetic analysis.

GeneQuiz annotation in COGENT

As an example of the capability of COGENT, we have mapped the data obtained by the automatic genome annotation system GeneQuiz for 60 complete genomes to the COGENT working environment. By making these corresponding tables available, we will enable COGENT users to benefit from an expanded query capability for this vast amount of data on a genome-wide context.

Future uses of COGENT

We anticipate that the computational biology community will benefit from the advantages offered by this open-source environment. All data and SQL tables can be easily downloaded from the web site. In the future, we hope that new developments might be openly exchanged by other groups using the COGENT data environment.

REFERENCES

- Bernal,A., Ear,U. and Kyrpides,N. (2001) Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res.*, **29**, 126–127.
- Hubbard,T., Barker,D., Birney,E., Cameron,G. Chen,Y. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
- Stein,L. (2002) Creating a bioinformatics nation. *Nature*, **417**, 119–120.