



CBS Genome Atlas Database: a dynamic storage for bioinformatic results and sequence data

Peter F. Hallin and David W. Ussery*

Center for Biological Sequence Analysis (CBS), BioCentrum-DTU, Building 208,
The Technical University of Denmark, DK-2800 Lyngby, Denmark

Received on January 16, 2004; revised on March 22, 2004; accepted on April 3, 2004
Advance Access publication July 15, 2004

ABSTRACT

Summary: Currently, new bacterial genomes are being published on a monthly basis. With the growing amount of genome sequence data, there is a demand for a flexible and easy-to-maintain structure for storing sequence data and results from bioinformatic analysis. More than 150 sequenced bacterial genomes are now available, and comparisons of properties for taxonomically similar organisms are not readily available to many biologists. In addition to the most basic information, such as AT content, chromosome length, tRNA count and rRNA count, a large number of more complex calculations are needed to perform detailed comparative genomics. DNA structural calculations like curvature and stacking energy, DNA compositions like base skews, oligo skews and repeats at the local and global level are just a few of the analysis that are presented on the CBS Genome Atlas Web page. Complex analysis, changing methods and frequent addition of new models are factors that require a dynamic database layout. Using basic tools like the *GNU Make* system, *csh*, *Perl* and *MySQL*, we have created a flexible database environment for storing and maintaining such results for a collection of complete microbial genomes. Currently, these results counts to more than 220 pieces of information. The backbone of this solution consists of a program package written in *Perl*, which enables administrators to synchronize and update the database content. The *MySQL* database has been connected to the CBS web-server via *PHP4*, to present a dynamic web content for users outside the center. This solution is tightly fitted to existing server infrastructure and the solutions proposed here can perhaps serve as a template for other research groups to solve database issues.

Availability: A web based user interface which is dynamically linked to the Genome Atlas Database can be accessed via www.cbs.dtu.dk/services/GenomeAtlas/

Contact: pfh@cbs.dtu.dk

Supplementary information: This paper has a supplemental information page which links to the examples presented: www.cbs.dtu.dk/services/GenomeAtlas/suppl/bioinfdatabase

1 INFRASTRUCTURE

1.1 File structure

The central location for all sources and results are made up by a filesystem indexed according to taxonomic relationships. Each organism is put in a file system as follows:

```
./[kingdom]/[genus]/[species]/[strain]/
[segment]/[segmentid].gbk
```

The [segment] refers to either a chromosome ID or a plasmid. If only one chromosome is present in the genome, [segment] defaults to 'Main'. The [segmentid] is the base name of all files for the current organism and is derived from the genus, species, strain and the segment as this example for *Campylobacter jejuni* strain NCTC 11168 (Parkhill *et al.*, 2000) shows:

```
./Bacteria/Campylobacter/jejuni/
NCTC11168/Main/Cjejuni_NCTC11168_
Main.gbk
```

1.2 File contents

Each directory contains some basic information such as the fasta file with whole genome sequence, fasta files with all open reading frames, all protein sequences derived from the GenBank annotation, etc. In addition, more complex bioinformatic calculations are kept within each folder. Intrinsic curvature, DNA flexibility and DNA stability plays an important role in the promoter analysis of microbial genomes (Pedersen *et al.*, 2000). To visualize these data, we are constructing different types of chromosomal maps (atlases) some of which include the 'Structural Atlas', 'Repeat Atlas' and 'Genome Atlas' (Pedersen *et al.*, 2000). Each atlas is available in either vector graphic format (PS) or compressed bitmap (PNG). The intermediate files used to build these atlases are maintained as well. For each property calculated, there is a corresponding list of numerical values calculated for every base pair in the genome. These lists are available by clicking 'CBS Download' for the Genome segment.

*To whom correspondence should be addressed.

Base composition such as AT content and GC skew, global and local repeats (Jensen *et al.*, 1999), gene length distribution (Skovgaard *et al.*, 2001), genome periodicity (Worning *et al.*, 2000), tRNA predictions by tRNAscan-SE 1.23 (Lowe and Eddy, 1997) and over-represented oligos for origin determination (P.Worning, L.J.Jensen, P.F.Hallin, H.H.Stærfeldt and D.W.Ussery, submitted for publication) are other results of analysis that are kept in this structure.

1.3 The GNU Make system

Whether a given piece of information is the result of complex analysis or just simple conversions or counts, the task remains the same: Define source files needed → Construct a program → Define result files (referred to as targets). This pipeline fits into the *Make* system since it keeps track of which files need to be updated and which programs should be used to build the target files. *Make* was chosen to create and update all the files within the data structure since it enables easy maintenance and development of new calculation procedures. Based on the *Make* rules kept on the servers and the file structure just described, all results and analysis can be performed by typing simple *Make* commands.

2 THE DATABASE

All sources of results are present within a precise file structure and a database solution should therefore aim to establish a robust link between a database table and file system contents. This has been the primary task for the project. Some data are stored by copying the file content directly into the table records. This is done for simple results such as numerical values and single-line results. When more complex results are generated, e.g. statistical reports from tRNA predictions or intermediate atlas data with millions of lines, it is preferable to have a link to the data rather than copying the entire file to the database. By having this breakdown of storage, large amounts of disk space is saved. This structure is visualized in Figure 1.

The establishment of a link between the file structure and the main database table is done by introducing a smaller database table—the configuration table. This table keeps a detailed description of how data should be collected and stored in the main database table. Each row of the configuration table contains instructions about how a specific result should be obtained. This layout is described in Figure 2.

3 PERL COMPONENTS—SHELL TASKS

Most database maintenance and upgrades are done from the Unix shell prompt. A collection of Perl scripts was written which performs common tasks, and some important scripts are described below:

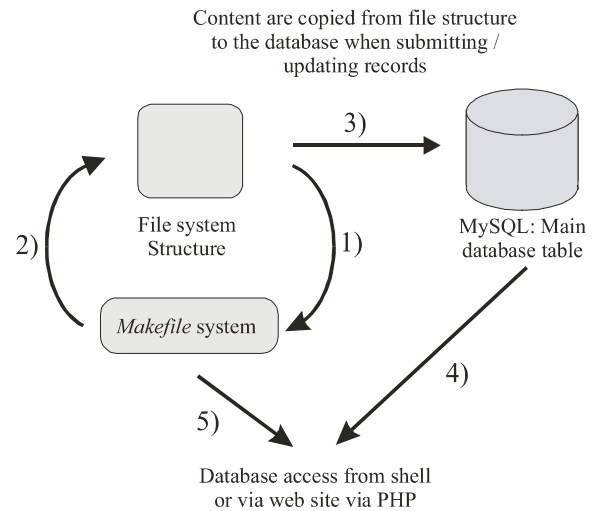


Fig. 1. Principle pipelines of the Genome Atlas Database. For each piece of information that is maintained *Make* will read the source file(s) from the directory (1) and generate the corresponding results (2). After the results are generated, the results are processed and either linked or copied into the database. Local users and Web users will have access to both the directory structure (5, for linked content) and the database content (4, for values directly stored in the database) when using the services.

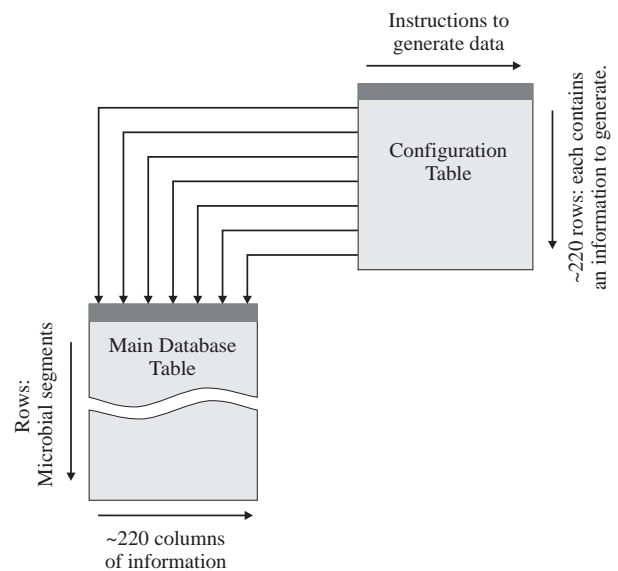


Fig. 2. Each row in the configuration table corresponds to a column in the main database table. Each row of the configuration table has an instruction set that describes how a given information should be obtained. Some of these instructions, include name of target column in main table, MySQL data type of this column, extension of file to read data from, whether to execute *Make* on this file when submitting a genome and whether to store or link the content. Scripts are written which creates/deletes columns in the main database table based on the content of the configurations table.

Script	Function
dbget	<p>Extracts information from database</p> <p><i>Syntax:</i> dbget[column_to_search] [content_to_match]([output])</p> <p><i>Example:</i> dbget species jeju% ``#SEGMENTID#\t#ATCONTENT#\n``</p> <p>This extracts AT content of species matching 'jeju%' and shows it as a tab formatted list:</p> <pre>Cjejuni_Strain_pVir 0.74111 Cjejuni_Strain_pCJ01 0.66481 Cjejuni_NCTC11168_Main 0.69451</pre>
dbsync	<p>Walks through the entire file system structure synchronizing content with database.</p>
dbsubmit	<p>Submits a data directory to database record</p> <p><i>Syntax:</i> dbsubmit segmentid-file [options]</p> <p><i>Example:</i> dbsubmit ./segmentid</p>

Currently, there is no support for automatic synchronization with existing databases, such as GenBank, EMBL or others. When a new fully annotated sequence is available, a filesystem directory is manually created as described in Section 1.1. Once this is created, dbsubmit is used to insert a record in the database that links to this directory. This script automatically launches all Make commands needed to produce all targets. This approach ensures independence upon other databases, since it requires only an EMBL or GenBank file containing the complete genome sequence. With the current rate of a few genomes published per month, the tasks are small but future steps might be needed to synchronize the content automatically.

4 THE WEB INTERFACE

All of the information described in this paper is presented in the Genome Atlas Web page. Within each kingdom all segments in the database can be viewed in a single table. The Web page is supplied with search options which enables a search within each kingdom. Figure 4 shows an example of a search within the kingdom of Bacteria. This location points to different tables presenting the database content of all the organisms that are currently stored in the database. The pages gives the possibility of searching by genus, species, strain, segment and/or taxonomy group, as shown in Figure 3.

5 DATABASE APPLICATIONS

In the following sections, we will give a few examples of how the Genome Atlas Database is used for the analysis of the bacterial chromosomes currently stored in the database. All sources and targets involved in these calculations are subjected to the same pipeline described in the earlier sections.

5.1 Origin of replication

A new method for determining the origin of replication in bacterial chromosomes has been proposed (P.Worning,

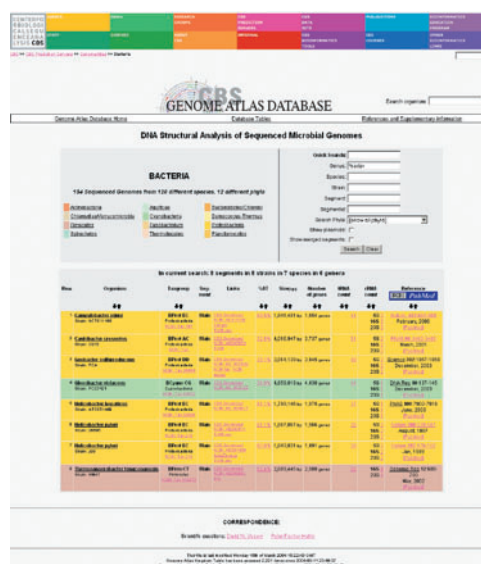


Fig. 3. The Web page enable users to search on different columns in the main database table. Columns like genus, species, strain, segment and phyla are searchable. Here, we have searched for all organisms having genus matching '*bacter'. Where relevant, columns are supplied with a descending/ascending sort option.

L.J.Jensen, P.F.Hallin, H.H.Stærfeldt and D.W.Ussery, submitted for publication), utilizing skews of oligos in the length 1–8 bp. Programs were developed and implemented in the Make system as described above, which generated the predictions for the entire collection of bacterial chromosomes. The database enabled fast visual inspection of all origin plots, and Figure 4a and b show an example of such an origin plot for *Clostridium perfringens* strain 13 (Shimizu *et al.*, 2002)—a Firmicute—and *Buchnera aphidicola* strain BBp (van Ham *et al.*, 2003)—a Proteobacteria.

Differences in the composition of the DNA polymerase holoenzyme are suggested to account for the change in direction of the AT skew that are observed between Proteobacteria and Firmicutes (P.Worning, L.J.Jensen,

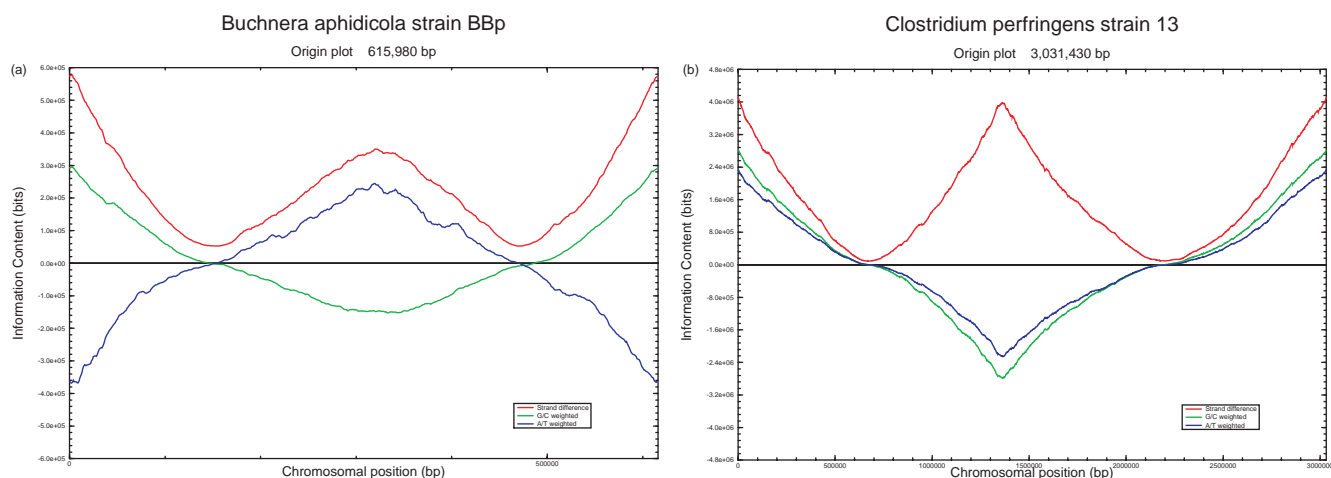


Fig. 4. (a) The origin plot of *Buchnera aphidicola* strain Bbp—a proteobacteria. AT and GC skew are pointing in apposite directions. (b) The origin plot of *Clostridium perfringens* strain 13—a firmicute. Note that AT and GC skew are pointing in same direction—an evidence of the different replication mechanisms in Firmicutes and Proteobacteria.

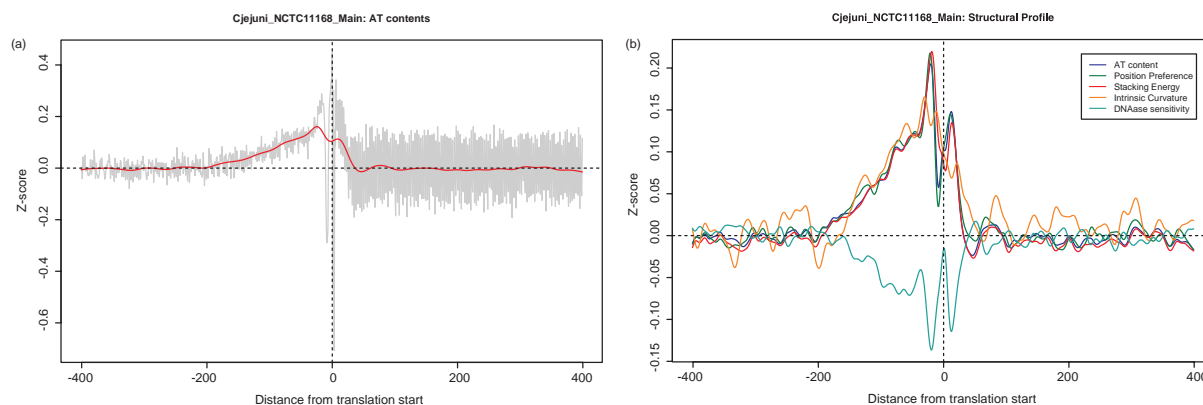


Fig. 5. (a) AT content profile of ± 400 bp window relative to translation start in *C. jejuni* NCTC 11168. Data are smoothed over a ± 8 bp wide Gaussian distribution. Fluctuating raw data downstream of translation start is a consequence of the codon usage preference. (b) Structural profile of ± 400 bp window relative to translation start in *C. jejuni* NCTC 11168. There seems to be a clear correlation between the structural parameters and the AT content. Data are smoothed over a ± 5 bp wide Gaussian distribution. The evident change in structural parameters are signs of a promoter region.

P.F.Hallin, H.H.Stærfeldt and D.W.Ussery, submitted for publication). These and other examples of origin plots can be obtained via the Supplemental Web page.

5.2 DNA structural profiles

Structural profiles are generated for each organism showing average DNA properties. For every chromosome all CDS regions are oriented in the same direction, aligned to translation +1 and a ± 400 bp window extracted. For each position in this alignment, we have calculated the average for different structural parameters such as stacking energy, curvature, position and preference. Figure 4a shows the AT content profile of all CDSs in *C. jejuni* NCTC 11168. The high peak at +2 immediately followed by a low peak at +3 corresponds to the common start codon, ATG. The highly variable AT

content within the coding region is a result of the codon usage preference by the organism. In Figure 5b, the average structural parameters for the same organism shows a clear change indicating the presence of a promoter. It also appears that this change is to some extent correlated with the AT content. All parameters in Figure 5b are smoothed over a ± 5 bp window using a Gaussian distribution. The AT content is smoothed over a ± 8 bp window. For the plots in Figure 5b, a Z-score is used which uses the average and SD of the whole chromosome.

5.3 Comparing number of tRNAs

As a last example of utilization for our database, we will mention the prediction of tRNA loci. In 14 cases of sequenced bacterial chromosomes with valid GenBank entries, tRNA

annotations were missing. Using tRNAscan-SE, we implemented *Make* rules that generate a prediction list, a statistical report and a final count of hits. Additional scripts were made that compare these predictions with existing annotations. In the Supplemental Web page, examples are given for *C.jejuni* NCTC 11168 where one additional arginine tRNA has been detected on bottom strand from 878 305 to 878 380.

6 CONCLUSION

The implementation of a filesystem supported database has proven extremely helpful for creating and maintaining sequence data and bioinformatic analysis for complete microbial genomes. An important goal has been achieved enabling future research results to be easily integrated with the database and presented in the Genome Atlas Database pages. For each chromosome or plasmid, we also provide links to other Web pages that contain additional information about sequenced genomes, e.g. NCBI's Entrez (Schuler *et al.*, 1996) and TIGR's Comprehensive Microbial Resources (Peterson *et al.*, 2001). In contrast to these databases, the Genome Atlas pages were developed to provide links to the visualization of DNA structural properties of sequenced microbial chromosomes. For example, seven different structural atlases are presented for each segment, which allows an overview of the entire genome in a single image. We also provide promoter analysis, base composition properties, DNA repeats and a variety of other chromosomal information. The majority of the information currently stored are presented in tables that enable users to perform any kind of sorting and/or searching within each kingdom; a given property is thereby seen in its context which could be the phyla or the kingdom. The Web pages are constructed in such a way that any intermediate result can be made available for any organism of interest. This is done through the link 'CBS Download' in the 'Link' column.

ACKNOWLEDGEMENTS

We would like to thank Hans-Henrik Stærfeldt for discussion on database and Web design, Peter Wad Sackett for the implementation of PHP and discussion, Karin Lagesen for help on

tRNA predictions and useful suggestions, Kristoffer Rapacki for support on server infrastructure. This work was supported by a grant from the Danish Center for Scientific Computing (DCSC).

REFERENCES

- van Ham,R.C., Kamerbeek,J., Palacios,C., Rausell,C., Abascal,F., Bastolla,U., Fernandez,J.M., Jimenez,L., Postigo,M., Silva,F.J. *et al.* (2003) Reductive genome evolution in *Buchnera aphidicola*. *Proc. Natl Acad. Sci., USA*, **100**, 581–586.
- Jensen,L.J., Friis,C. and Ussery,D.W. (1999) Three views of microbial genomes. *Res. Microbiol.*, **150**, 773–777.
- Lowe,T. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
- Parkhill,J., Wren,B.W., Mungall,K., Ketley,J.M., Churcher,C., Basham,D., Chillingworth,T., Davies,R.M., Feltwell,T., Holroyd,S. *et al.* (2000) The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature*, **403**, 665–668.
- Pedersen,A.G., Jensen,L.J., Brunak,S., Stærfeldt,H.H. and Ussery,D.W. (2000) A DNA structural atlas for *Escherichia coli*. *J. Mol. Biol.*, **299**, 907–930.
- Peterson,J.D., Umayam,L.A., Dickinson,T.M., Hickey,E.K. and White,O. (2001) The comprehensive microbial resource. *Nucleic Acids Res.*, **29**, 123–125.
- Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) Entrez: molecular biology database and retrieval system. *Methods Enzymol.*, **266**, 141–162.
- Shimizu,T., Ohtani,K., Hirakawa,H., Ohshima,K., Yamashita,A., Shiba,T., Ogasawara,N., Hattori,M., Kuhara,S. and Hayashi,H. (2002) Complete genome sequence of *Clostridium perfringens*, an anaerobic flesh-eater. *Proc. Natl Acad. Sci., USA*, **99**, 996–1001.
- Skovgaard,M., Jensen,L.J., Brunak,S., Ussery,D.W. and Krogh,A. (2001) On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet.*, **17**, 425–428.
- Worning,P., Jensen,L.J., Nelson,K.E., Brunak,B. and Ussery,D.W. (2000) Structural analysis of DNA sequence: evidence for lateral gene transfer in *Thermotoga maritime*. *Nucleic Acids Res.*, **28**, 706–709.