Nucleic Acids Research, 2005, Vol. 33, Database issue ***D317–D320***
doi:10.1093/nar/gki075

BacMap: an interactive picture atlas of annotated bacterial genomes

Paul Stothard, Gary Van Domselaar, Savita Shrivastava, Anchi Guo, Brian O'Neill, Joseph Cruz, Michael Ellison[1,2] and David S. Wishart*

Department of Computing Science and Biological Sciences, [1]Department of Biochemistry and [2]Institute for Biomolecular Design, University of Alberta, Edmonton, AB, Canada T6G 2E8

Received August 14, 2004; Revised and Accepted October 8, 2004

ABSTRACT

BacMap is an interactive visual database containing fully labeled, zoomable and searchable chromosome maps from more than 170 bacterial (archaebacterial and eubacterial) species. It uses a recently developed visualization tool (CGView) to generate high-resolution circular genome maps from sequence feature information. Each map includes an interface that allows the image to be expanded and rotated. In the default view, identified genes are drawn to scale and colored according to coding directions. When a region of interest is expanded, gene labels are displayed. Each label is hyperlinked to a custom 'gene card' which provides several fields of information concerning the corresponding DNA and protein sequences. Each genome map is searchable via a local BLAST search and a gene name/synonym search. BacMap is freely available at http://wishart.biology.ualberta.ca/BacMap/.

INTRODUCTION

Since the first bacterial genome was completed in 1995 (1), more than 170 bacterial genomes have been sequenced (2). Another 500 are currently being sequenced and many will likely be released in the coming year. With current sequencing technology, it is now possible to sequence and assemble an entire bacterial genome in less than a week. This flood of extremely valuable genomic data is threatening to overwhelm our capacity to assimilate and process it. Never before has so much information been available about so many different bacterial species. A growing challenge facing microbiologists and bioinformaticians alike is to find ways to better manage, display and compare these data. Several database resources have appeared over the past few years to help in this regard. GenBank and EMBL now maintain up-to-date microbial genome sequence archives, while MIP's PEDANT (3) and TIGR's CMR (4) provide more detailed annotations and statistics on many microbial genomes. However, the tendency for most consolidated microbial databases is to present sequence-related data in a text-only or in a relatively limited graphic format. This text-only approach ignores the rich information that can be gained by displaying interactive graphical maps of individual genes or open reading frames (ORFs) in their genomic context or by comparing different genomic maps to one another. Given the enormous success of Ensembl's (5) dual textual and graphical approach to presenting metazoan genomes, we decided to investigate the possibility of using a similar concept to present bacterial genomic data.

Here, we describe a highly visual, fully interactive, self-updating, web-enabled database called BacMap. BacMap is a unique resource that allows users to rapidly compare, explore and search through hundreds of bacterial genomes. BacMap is essentially an electronic atlas of bacterial genomes with hundreds of colorful, clickable maps, all linked to thousands of megabytes of detailed annotation. By pre-calculating the image maps and annotations, BacMap allows users to rapidly search, zoom in, rotate and zoom out of its circular genomic maps. Furthermore, no special browser plugins or applets are required, making BacMap fully compatible with a wide range of web browsers, hardware configurations and operating systems.

DATABASE DESCRIPTION

The BacMap homepage contains a list of all publicly released eubacterial and archaebacterial genomes (177 at the time of this writing). This list is presented alphabetically, according to genus, species and strain. Each bacterial name is hyperlinked to a 'species card', which provides detailed information about the organism in tabular format, including its taxonomy, gram staining properties and number of chromosomes. A brief description of the species, discussing its physiology, general characteristics, ecological niche and relevance to human or animal disease is also given. In addition, an image of the

*To whom correspondence should be addressed. Tel: +780 492 0383; Fax: +780 492 1071; Email: david.wishart@ualberta.ca

organism (if available) is provided. Below each genome entry is a list of the genome's constituent chromosomes, sorted by length. Five buttons are provided for each chromosome: 'MAP', 'TEXT SEARCH', 'BLAST', 'STATS' and 'DOWNLOAD'. The 'MAP' button displays a graphical map of the chromosome in a new window. The 'TEXT SEARCH' and 'BLAST' buttons are linked to the text search and BLAST search interfaces for the chromosome. The 'STATS' button displays several graphs concerning the chromosome's genomic and proteomic characteristics. Finally, the 'DOWNLOAD' button opens the data download page for the chromosome.

Clicking on the 'MAP' button generates a full-screen circular image of the entire bacterial chromosome. On the lower edge of the image is a brief synopsis of the chromosome, including the GenBank accession number of the source sequence. The full view map consists of two concentric rings of forward and reverse strand genes (protein and RNA), with tick marks indicating chromosomal position. Some maps may contain additional feature rings (e.g. COG functional classifications) depending on which annotations are currently available for the chromosome. Clicking on any of the tick marks in the sequence ruler expands the map by a predefined step, and centers the view on the base closest to the tick mark that was clicked. The map view can also be manipulated using the control panel located at the bottom of the map. The 'Expand +' button zooms in on the current view, while 'Expand −' returns to a view showing more of the map with reduced detail. The view can also be shifted along the chromosome backbone in the clockwise and counterclockwise directions, using the 'Rotate +' and 'Rotate −' buttons, respectively. Hyperlinked gene labels are visible from the first zoom level onwards. Pointing to a gene label displays the start and stop positions of the gene, as well as its known or predicted function. Clicking on the gene label replaces the map view with the corresponding 'gene card'. The rapid response to these operations is achieved by pre-rendering all the images for each chromosome and the corresponding HTML image maps. The image rendering is done by a recently developed in-house program called CGView, which was designed to generate annotated images of circular chromosomes or plasmids. On average, more than 4000 PNG images and HTML image maps were generated for each chromosome in BacMap. These pre-rendered images typically occupy 100 MB of disk per chromosome.

The chromosome maps can be explored manually, or with the assistance of two search tools integrated with the BacMap database. One tool is a Boolean text search, which can be used to search for specific gene names, protein names, alternate names or partial names. Any matches returned from a text search are shown on a dynamically generated search results map. A textual list of the matching genes is also returned, with hyperlinks provided so that the relevant pre-rendered chromosomal map views and gene cards can be quickly retrieved. The second searching method uses BLAST (6) to identify genes that are similar to a user-supplied sequence. The BLAST method can be used as an alternate route to find genes/proteins if the text search is unsuccessful, or as a means to identify orthologs and paralogs of a sequence of interest. As with the text search, the BLAST results are shown graphically and textually, with hyperlinks provided for accessing the chromosome maps and gene cards. Figure 1 provides a montage of BacMap images to demonstrate the image quality, utility and general operation of the BacMap database.

A particularly useful feature of BacMap is the extensive annotation provided for each gene. These annotations, presented as gene cards, are accessible by clicking on the gene labels displayed on individual chromosome maps, or by using the text and BLAST searches. The cards are built using a variety of public databases, such as UniProt (7) and PDB (8), as well as numerous in-house prediction programs. Indeed, many of the annotation methods employed for BacMap were originally developed and validated in the preparation of the CyberCell database—a comprehensive molecular database on *Escherichia coli* (9). Each time a new bacterial chromosome is added to the BacMap database, an initial set of gene cards is constructed. These preliminary cards, which can be generated rapidly, serve as a source of basic gene and protein information until the more extensively annotated cards have emerged from the analysis pipeline. The final cards typically contain about 50+ fields of annotation, including information on a variety of sequence statistics, potential orthologs and paralogs, predicted function, predicted secondary structure and predicted subcellular location. The annotations are continually updated as more information becomes available and as better prediction programs are developed. A partial listing of BacMap's current annotation fields along with their sources and/or methods is given in Table 1.

In addition to an extensive collection of genome maps and annotations, BacMap also supports flat file downloads of all textual data associated with each genome. This includes complete genomic DNA sequence data, FASTA files of all identified genes, FASTA files of all identified proteins and a flat file of all the BacMap annotations for each gene/protein. These files are accessed by clicking on the 'DOWNLOAD' button associated with each chromosome listed on the BacMap homepage. In addition, BacMap provides a number of charts and graphs for each chromosome, which illustrate nucleotide and amino acid usage, distribution of protein lengths, distribution of functions and distribution of predicted subcellular locations. These charts, which are regularly updated in conjunction with the BacMap gene cards, can be accessed using the 'STATS' buttons. While BacMap does not yet support relational queries for comparative genomics, it does support cross-genome comparisons. By separately querying and displaying chromosome segments from two or more species, users may visually compare bacterial orthologs, operon structure, gene synteny or gene context, chromosomal gene distributions or functional localizations. The BacMap database is automatically updated on a weekly basis using custom updating software. This software obtains new bacterial genome GenBank records from the NCBI and then passes them to the programs that generate the BacMap database content. The only database entries that require some limited manual editing are the 'species cards'.

To summarize, BacMap is an interactive electronic atlas of bacterial genomes. It builds from and extends upon the very successful visualization concepts originally introduced in Ensembl, providing a 'map-centric' or visually centered approach to exploring bacterial genomic data. The rich annotations in BacMap, coupled with its detailed, color-coded image maps, should permit users to look at bacterial genomes with increasing detail. The BacMap database can be freely accessed at http://wishart.biology.ualberta.ca/BacMap/.

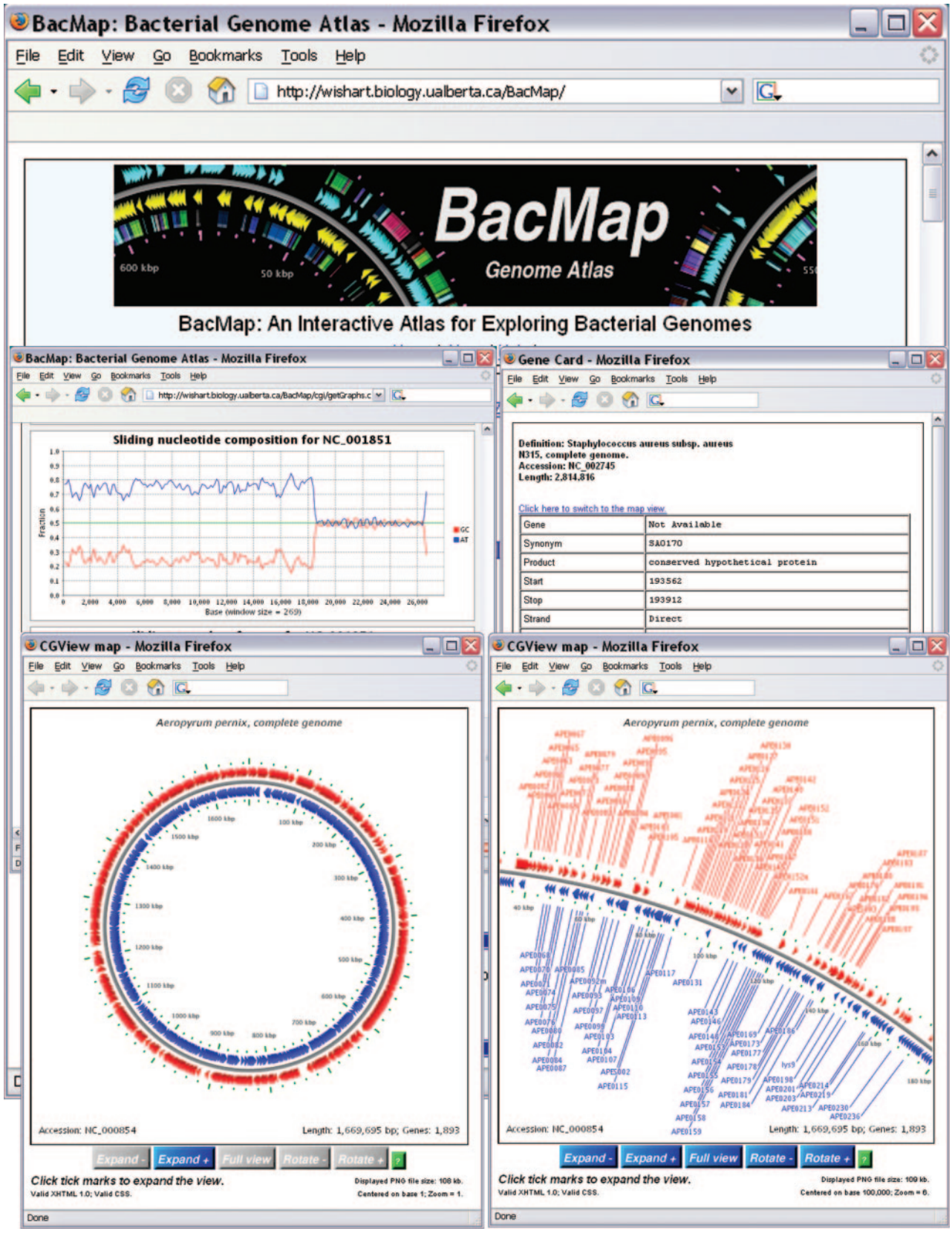

Figure 1. A screenshot montage of the BacMap database. Two different views of the chromosomal map for *Aeropyrum pernix* are shown in the foreground. The full chromosome map is on the left, while an expanded view providing more detail is on the right. Behind the full chromosome map is a nucleotide composition graph for plasmid lp28-1 from *Borrelia burgdorferi*, and next to this graph is a gene card displaying textual information about a predicted gene from *Staphylococcus aureus*. These views of data can be accessed from the BacMap homepage, or from the hyperlinked results of BLAST searches and text searches.

Table 1. Annotations added to gene entries in the BacMap database

Annotation	Source and/or method
DNA sequence	GenBank record parsing
Downstream 100 bases	GenBank record parsing
Following gene	GenBank record parsing
GC content	Calculated from sequence
Gene Ontology	InterPro (10)
Gene position	GenBank record parsing
NCBI GI number	GenBank record parsing
Orthologues	BLAST (6) with heuristics
Paralogues	BLAST (6) with heuristics
Pfam family	Pfam (11)
Preceding gene	GenBank record parsing
PROSITE families and domains	PROSITE (12)
Protein length	Calculated from sequence
Protein molecular weight	Calculated from sequence
Protein sequence	GenBank record parsing
Secondary structure	PSIPRED (13)
Subcellular location	PSORTb (14)
Theoretical pI	Calculated from sequence
Transmembrane domains	PredictTM
Upstream 100 bases	GenBank record parsing

The annotation fields are added in several phases, as results are obtained from the various database searching and sequence analysis programs. The fields and their contents may change as new programs are added to the BacMap annotation pipeline.

ACKNOWLEDGEMENTS

Funding for this project was provided by the Alberta Science Research Authority, Western Economic Diversification, Canada, and Genome Prairie (a division of Genome Canada).

REFERENCES

1. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M. *et al*. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.
2. 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.
3. Frishman,D., Mokrejs,M., Kosykh,D., Kastenmuller,G., Kolesov,G., Zubrzycki,I., Gruber,C., Geier,B., Kaps,A., Albermann,K. *et al*. (2003) The PEDANT genome database. *Nucleic Acids Res*., **31**, 207–211.
4. Peterson,J.D., Umayam,L.A., Dickinson,T., Hickey,E.K. and White,O. (2001) The Comprehensive Microbial Resource. *Nucleic Acids Res*., **29**, 123–125.
5. Birney,E., Andrews,D., Bevan,P., Caccamo,M., Cameron,G., Chen,Y., Clarke,L., Coates,G., Cox,T., Cuff,J. *et al*. (2004) Ensembl 2004. *Nucleic Acids Res*., **32**, D468–D470.
6. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipmann,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–420.
7. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al*. (2004) UniProt: the universal protein knowledgebase. *Nucleic Acids Res*., **32**, D115–D119.
8. Bourne,P.E., Addess,K.J., Bluhm,W.F., Chen,L., Deshpande,N., Feng,Z., Fleri,W., Green,R., Merino-Ott,J.C., Townsend-Merino,W. *et al*. (2004) The distribution and query systems of the RCSB Protein Data Bank. *Nucleic Acids Res*., **32**, D223–D225.
9. Sundararaj,S., Guo,A., Habibi-Nazhad,B., Rouani,M., Stothard,P., Ellison,M. and Wishart,D.S. (2004) The CyberCell Database (CCDB): a comprehensive, self-updating, relational database to coordinate and facilitate *in silico* modeling of *Escherichia coli*. *Nucleic Acids Res*., **32**, D293–D295.
10. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al*. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res*., **31**, 315–318.
11. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L., Studholme,D.J., Yeats,C. and Eddy,S.R. (2004) The Pfam protein families database. *Nucleic Acids Res*., **32**, D138–D141.
12. Hulo,N., Sigrist,C.J., Le Saux,V., Langendijk-Genevaux,P.S., Bordoli,L., Gattiker,A., De Castro,E., Bucher,P. and Bairoch,A. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Res*., **32**, D134–D137.
13. McGuffin,L.J., Bryson,K. and Jones,D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–405.
14. Gardy,J.L., Spencer,C., Wang,K., Ester,M., Tusnady,G.E., Simon,I., Hua,S., deFays,K., Lambert,C., Nakai,K. and Brinkman,F.S. (2003) PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res*., **31**, 3613–3617.