Data and text mining

VIS-O-BAC: exploratory visualization of functional genome studies from bacteria

Guido Dieterich, Uwe Kärst, Jürgen Wehland and Lothar Jänsch*
Department of Cell Biology, Research Centre for Biotechnology (GBF), Mascheroder Weg 1, 38124 Braunschweig, Germany

Received on September 22, 2005; revised on November 15, 2005; accepted on December 1, 2005
Advance Access publication December 8, 2005
Associate Editor: Martin Bishop

ABSTRACT
Summary: The visualization-aided exploration of complex datasets will allow the research community to formulate novel functional hypotheses leading to a better understanding of biological processes at all levels. Therefore, we have developed a web resource termed VIS-O-BAC designed for the functional investigation of expression data for model systems, such as bacterial pathogens based on a graphical display. Genome-scale datasets derived from typical ‘omic’ approaches can directly be explored with respect to three biologically relevant aspects, the genome structure (operon organization), the organization of genes in pathways (KEGG) and the gene function with Gene Ontology (GO) terms. The integrated viewers can be used in parallel and combine expression data and functional annotations from different external data repositories. The graphical visualizations evidently accelerate both the validation of regulatory information and the detection of affected biological processes.

Availability: http://leger2.gbf.de/cgi-bin/vis-o-bac.pl
Contact: lja@gbf.de
Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION
The plethora of data are now available from bacterial genome sequencing projects [for an overview see Bernal et al. (2001), http://www.genomesonline.org/] has opened a wealth of new research opportunities. Together with data from functional studies of genes and proteins collected over the past half century and comparative genome analyses, a function could be assigned in bacterial genomes for up to 80% of the detected genes (Keseler et al., 2005).

Analyses of genome-scale datasets require resources to transfer experimental results to a concise biological interpretation. Resources such as UniProt (Bairoch et al., 2005) provide a detailed representation of functional data for a given gene, but they are not suited to explore a group of genes or permit the use of expression ratios that are generated routinely in proteome and transcriptome studies.

In recognition of the value of data visualization, several tools or modules have been developed (e.g. GMOD, http://www.gmod.org and BioPerl, http://www.bioperl.org). They find their application in generally web-based programs [e.g. GBrowse (Stein et al., 2002) and WormBase (O’Connell, 2005), www.wormbase.org].

*To whom correspondence should be addressed.

We present the web resource VIS-O-BAC designed for the functional investigation and for the exploratory visualization of high-throughput datasets from bacterial experiments, drawing on three widely accepted functional annotations of genes, the genetic locus (source: GenBank), the metabolic and regulatory pathway (KEGG) and the classification in the context of Gene Ontology (GO). VIS-O-BAC is a powerful resource in which disparate, but related, forms of data are organized, presented and linked by the use of different well-established graphical tools.

2 METHODS
All programs were implemented in Perl making use of the BioPerl libraries (Stajich et al., 2002) and the GD graphics module (http://search.cpan.org/ dist/GD/). Popup windows were realized by overLIBJavaScript library (http://www.bosrup.com/web/overlib/).

Predicted operons and rho-independent transcription terminators in bacterial genomes were downloaded from http://www.tigr.org/tigr-scripts/ operons/operons.cgi and http://www.tigr.org/software/transterm.html, respectively.

We also used the GO-TermFinder (Boyle et al., 2004) distribution. Several of these modules were modified to meet our requirements (colour by expression value, popup window, etc.). GO data were extracted from the UniProt GO annotation (GOA) (ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/ UNIPROT/gene_association.goa_uniprot.gz) and stored in separate files.

3 INTERFACE
The VIS-O-BAC website is composed of a general information start page and query/browse interfaces for the three modules. These are a Genome, a KEGG and a GO Viewer, which are web applications for the visualization and interpretation of large-scale gene expression data. On each viewer page the user can upload or enter a list of regulated genes or proteins for a selected species. A simple table format allows the use of data from different sources, such as proteomics and microarray data. Once uploaded into one of the modules, the table can then also be used in the other viewers, thus visualizing the same gene/protein expression list in different contexts. The gene names are coloured according to a common convention for the experimental parameter (e.g. red or green for relative gene expression ratios below and above 1).

For each gene additional data from different sources and links to further resources are displayed within the page (e.g. promoter sequences and FASTA formatted sequences in Genome Viewer) or in popup windows (functional annotations). The application is designed to be open and extensible.
The Genome Viewer displays the distribution and regulation of expressed genes/proteins—both depicted as short, coloured vertical lines under its relative position in the genome—within a bacterial genome and in parallel—at a higher level of detail—in an individual region. In the region view, transcriptional termination signals (rho-independent) and putative operons are shown as blue lines above the genes. A KEGG track indicates whether individual enzyme functions are assigned to one or more KEGG pathways.

The integration of existing KEGG pathway maps into the concept of VIS-O-BAC opens the possibility to improve an accepted graphical presentation for the exploration of metabolic and signaling processes. The KEGG viewer displays the distribution and the regulation of expressed genes/proteins along biochemical pathways originally published by KEGG. The KEGG Viewer shows experimental results in a table format where the left column indicates the pathways and the right column summarizes all genes from the dataset that can be assigned to each corresponding pathway. Additionally the user can select KEGG maps. The assigned gene products are highlighted in blue, whereas all functionally annotated gene products of the selected bacterial genome basically are displayed in green (as default). Downregulation of a gene is indicated by an inverted red vertical arrow and an upregulation of a gene is indicated by a normal green vertical arrow.

The GO viewer module can group and hierarchically display differentially expressed genes according to their functional GO. The user can choose between three general types of classifications (biological process, molecular function and cellular component). The GO Viewer provides quantitative and statistical output and the user can select between three kinds of representation for the concise visualization of the results: (1) a tree view with different levels of GO terms specificity; (2) a table view of GO terms that are significantly overrepresented in the dataset, ranked by \( P \)-value; and (3) a textual, hierarchical tree view of GO terms [as in (2)] with expandable levels, ranked by \( P \)-value. In the Supplementary material we provide a case study presenting typical views generated by VIS-O-BAC.

4 DISCUSSION

Projects analysing gene expression at the mRNA or protein level tend to produce large amounts of data, especially when carried out systematically at genome scales. We have developed an integrative data exploration and visualization package to support the functional genome studies of bacteria in particular. Thereby VIS-O-BAC simplifies the evaluation of large-scale expression studies and aids the researchers in assigning their individual results to biological pathways and extract meaningful and novel relations and implications. Since many genes have been annotated repeatedly, VIS-O-BAC may also reveal inconsistencies by compiling annotations from different sources for comparison. Currently, 25 bacteria species from 14 families—highly relevant in the field of infection medicine—are available but the integration of further bacterial genomes is possible and intended as requested by the users.

Three modules were established for the visualization of regulatory data. All of them can access and present the same supportive annotations collected from GenBank, KEGG and UniProt. This integrative approach already distinguishes VIS-O-BAC from general-purpose genome browsers. The KEGG Viewer will probably be the initial choice for the exploration of large-scale expression studies. The direct visualization of regulatory results on the maps allows fast and intuitive recognition of the functional connections. The use of both KEGG and Genome Viewer is possible and beneficial even without submitting experimental data. Further modules will be considered in later versions of VIS-O-BAC. The generation of functional hypothesis will be enhanced by the planned integration of MineBlast (Dieterich et al., 2005), a tool that supports the process of systematic reannotation of gene functions by a comprehensive literature search.

ACKNOWLEDGEMENTS

We are very grateful to Victor Wray for critical proofreading of the manuscript. This work was funded by the BMBF, ‘Verbundvorhaben: Intergenomics—Bioinformatische Modellierung der Wechselwirkung von Genomen’ (031U110A/031U210A).

Conflict of Interest: none declared.

REFERENCES


