BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 9 2007, pages 1161–1163
doi:10.1093/bioinformatics/btl658

Gene expression

FIVA: Functional Information Viewer and Analyzer extracting biological knowledge from transcriptome data of prokaryotes

Evert-Jan Blom[1], Dinne W. J. Bosman[2], Sacha A. F. T. van Hijum[1], Rainer Breitling[3], Lars Tijsma[2], Remko Silvis[1], Jos B. T. M. Roerdink[2] and Oscar P. Kuipers[1,*]

[1]Molecular Genetics, Groningen Biomolecular Sciences, [2]Institute for Mathematics and Computing Science and [3]Groningen Bioinformatics Centre, University of Groningen, PO Box 800, 9700 AV, Groningen, The Netherlands

Received on October 19, 2006; revised on November 24, 2006; accepted on December 19, 2006

Advance Access publication January 19, 2007

Associate Editor: Dmitrij Frishman

ABSTRACT

Summary: FIVA (Function Information Viewer and Analyzer) aids researchers in the prokaryotic community to quickly identify relevant biological processes following transcriptome analysis. Our software assists in functional profiling of large sets of genes and generates a comprehensive overview of affected biological processes.

Availability: http://bioinformatics.biol.rug.nl/standalone/fiva/

Contact: o.p.kuipers@rug.nl

Supplementary information: http://bioinformatics.biol.rug.nl/standalone/fiva/suppMaterials.php

1 INTRODUCTION

Genome-wide expression profiles describing various cellular states are obtained by use of DNA microarrays. Following statistical analysis of the raw gene expression values, data-driven methods such as unsupervised clustering allow grouping of genes based on their (temporal) expression patterns. Genes involved in similar cellular processes are expected to have a high probability of exhibiting similar expression patterns. Analysis and interpretation of these clusters is time-consuming and error-prone. Various applications have been developed to functionally profile differentially expressed genes from DNA-microarray experiments.

Several of these, as reviewed by Khatri *et al.* (2005), overlap with our application in terms of functionality and data sources employed. Many of these focus on higher organisms and therefore lack support for prokaryote gene identifiers. A number of applications support rarely used (Uniprot, GI accession) identifiers (Hosack *et al.*, 2003) or only identifiers for a limited set of organisms (Scheer *et al.*, 2006). Moreover, with few exceptions, these software products use gene ontology as their exclusive data source. In addition, the laborious task of preprocessing the list containing differentially expressed genes must be performed by a researcher. A stand-alone application that focuses on prokaryotes is therefore essential for the fast-growing community of microbiologists making use of a plethora of (confidential) microbial genome sequences.

*To whom correspondence should be addressed.

We have developed FIVA (Functional Information Viewer and Analyzer). It uses several sources of biological annotations to create an extensive functional profile based on gene expression data. Furthermore, FIVA is capable of processing groups of genes assembled by other criteria (e.g. functional grouping of genes which are not available in current annotation modules). The significance of each biological process is calculated to distinguish between significant and spurious occurrences.

2 PROGRAM OVERVIEW

2.1 Input

The input data for FIVA consists of transcriptome data and genome annotation files (e.g. EMBL or Genbank), supplemented with annotation information. FIVA supports a broad variety of prokaryotic gene identifiers from the expression datasets, including locus tags and standard gene names (further details available in Supplementary Materials). Each annotation module uses functional information from one of the following sources to classify the groups of genes and determine any significantly over-represented categories. (i) Gene ontology (ii) Metabolic pathways (iii) COG classes (iv) Regulatory interactions (v) UniProt keywords (vi) InterPro (vii) User-defined functional categories.

2.2 Processing

The analysis in FIVA first involves the partitioning of the gene expression data into up- and down-regulated fractions. Testing different settings to partition the data is not a trivial task. FIVA offers the ability to automatically detect the optimal settings for each individual experiment based on the number of over-represented functional categories. In addition to this partitioning method which is based on thresholds applied to a single experiment, the iGA algorithm (Breitling *et al.*, 2004) is implemented. This algorithm optimizes the parameters for each functional category, which greatly improves the sensitivity of the analysis and increases the number of affected biological processes that can be reliably detected. Furthermore, the analysis of gene expression data can also be applied on user-defined gene lists.

A Fisher exact test is used to calculate *P*-values for each cluster. This *P*-value describes the probability of observing a specific enrichment of genes from a functional category in a cluster by chance. The number of false positives, due to the large number of statistical tests performed, are controlled by four multiple testing corrections (Benjamini/HochBerg, Bonferroni Step-down, Bonferroni and Benjamini Yekutieli). These are implemented to adjust the raw *P*-values (see Supplementary Website).

2.3 Output

For each of the classification modules, a graphical representation of the over-represented categories is generated. A preview is created from these results, from which a selection of the results can be made (Fig. 1). In order to conveniently compare biological phenomena occurring in different experiments, multiple experiments can be loaded simultaneously and are displayed as columns. Clickable links are present for each category in the individual graphical map, providing detailed information for each cluster that contains an enrichment of genes from this category. Furthermore, FIVA uses the KEGG API (http://www.genome.jp/kegg/soap/) to communicate with the KEGG database to color pathways based on the gene distribution in the clusters.

2.4 Implementation and availability

FIVA was programmed as a stand-alone application in Java using the Eclipse (http://www.eclipse.org/) framework and runs on all Java-supporting operating systems (Mac OS, MS Windows, UNIX and Linux). The graphical output can also be viewed by all web browsers that are able to process scalable vector graphics (SVG) or, to ensure portability

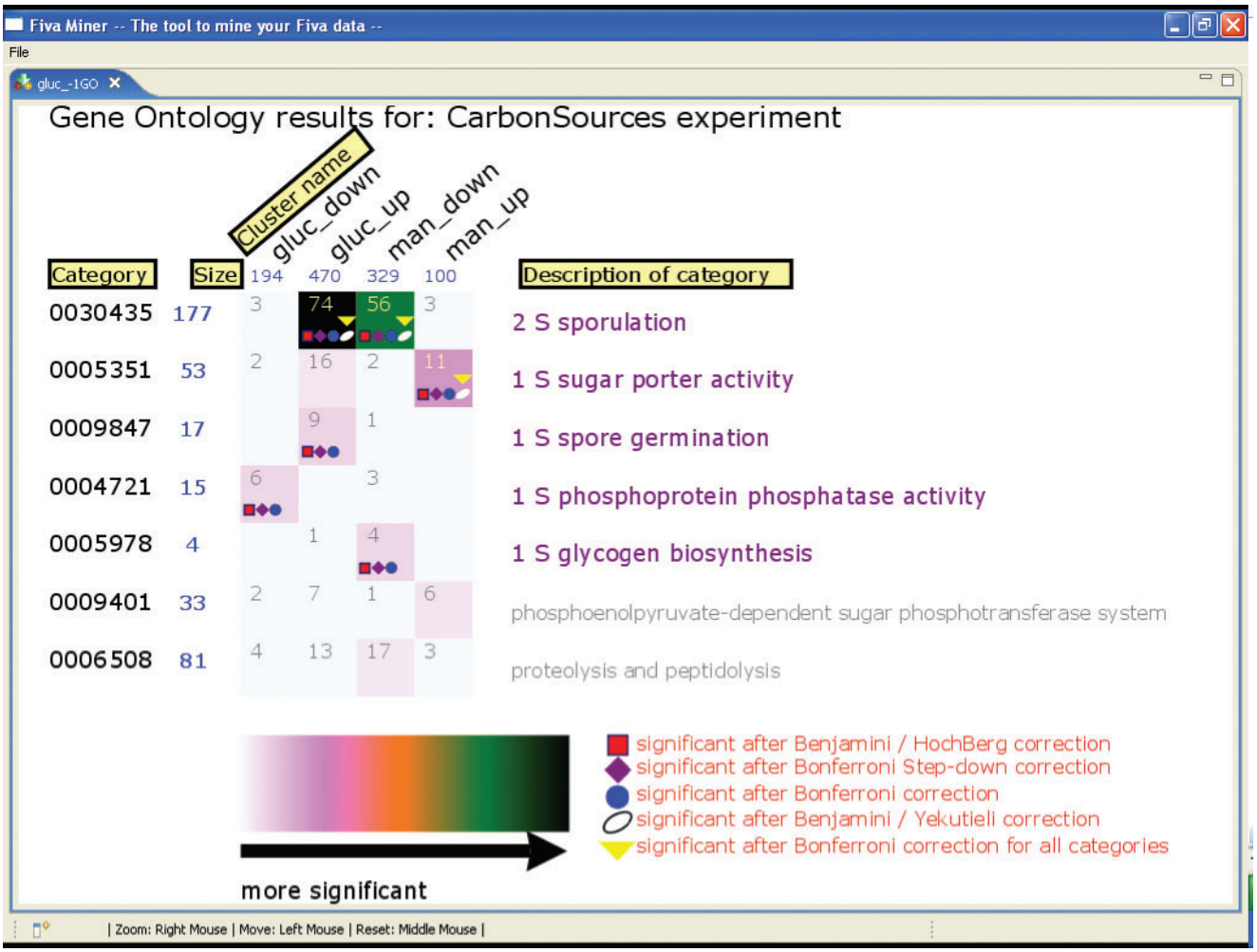

Fig. 1. Graphical output of a single annotation module. Genes from two DNA-microarray datasets (gluc: growth on glucitol compared to growth on glucose, man: growth on mannitol compared to glucose) were partitioned into up- and down-regulated clusters. The size of each cluster is displayed in blue underneath the cluster name. Numbers in each rectangle represent absolute values of occurrences. The significance of occurrences is visualized in a colour gradient which is displayed at the bottom of the plot. The description of each category is placed at the right. S: annotations that are significant after multiple testing correction. Multiple testing correction results are visualized using five different symbols to distinguish between the individual corrections. The number of symbols placed in each rectangle corresponds to the number of multiple testing corrections after which the annotation is found significant.

of the results, portable network graphics. More information on the functionality of FIVA, as well as the results of several test cases, can be found under the Supplementary Materials.

3 CONCLUSION

A full information analysis was performed to assess the overlap between the annotation modules (see Supplementary Website). The gene ontology module is the most informative annotation type for our test organism *Bacillus subtilis* and covers a large portion of the information present in the other types. However, the utilization of multiple modules yields relevant areas which are not shared by any of the other modules. For our test cases, several relevant categories were identified by the metabolic pathways modules but were missed by the GO module (see Supplementary Website for more information on this analysis). We conclude that combining multiple annotation sources into one tool is advantageous compared to using only one or a few sources. The combination of various complementary annotation sources, together with the dynamic visualization and elaborate statistical analysis, allows a richer and more objective exploration of prokaryote expression data than any other available tool provides.

ACKNOWLEDGEMENTS

This study was fully supported by a grant from The Netherlands Organization for Scientific Research and industrial partners in the NWO-BMI project number 050.50.206 on Computational Genomics of Prokaryotes and by Center IOP Genomics. Work performed by SvH was supported by grant QLK3-CT-2001-01473 under the EU programme 'Quality of life and management of living resources: The cell factory'. We thank J.W.Veening for useful suggestions on experimental procedures and G. te Meerman for expert advice on the statistical analysis. Funding to pay the Open Access charges was provided by the Molecular Genetics department of the University of Groningen.

Conflict of Interest: none declared.

REFERENCES

Breitling,R. *et al*. (2004) Iterative group analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC Bioinformatics*, **5**, 34.

Hosack,D.A. *et al*. (2003) Identifying biological themes within lists of genes with ease. *Genome Biol*., **4**, R70.

Khatri,P. *et al*. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.

Scheer,M. *et al*. (2006) JProGO: a novel tool for the functional interpretation of prokaryotic microarray data using Gene Ontology information. *Nucleic Acids Res*., **34**, W510–515.