BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 19 2009, pages 2617–2618
doi:10.1093/bioinformatics/btp464

Gene expression

Genoscape: a Cytoscape plug-in to automate the retrieval and integration of gene expression data and molecular networks

Mathieu Clément-Ziza[1,2,†,*], Christophe Malabat[3,†], Christian Weber[4], Ivan Moszer[5], Tero Aittokallio[6], Catherine Letondal[7] and Sandrine Rousseau[5,*]

[1]Biotechnology Center, TU Dresden, Dresden, Germany, [2]INSERM U-781, Université René-Descartes, Hôpital Necker-Enfants Malades, [3]Institut Pasteur, Unité de Génétique des Interactions Macromoléculaires, CNRS, URA2171, [4]Institut Pasteur, INSERM U786, Biologie Cellulaire du Parasitisme, [5]Institut Pasteur, Plate-Forme Intégration et Analyse Génomiques, [6]Institut Pasteur, Unité de Biologie Systémique and [7]Institut Pasteur, Pôle Informatique, Paris, France

Received on February 9, 2009; revised on July 1, 2009; accepted on July 24, 2009
Advance Access publication August 3, 2009
Associate Editor: Trey Ideker

ABSTRACT

Summary: Genoscape is an open-source Cytoscape plug-in that visually integrates gene expression data sets from GenoScript, a transcriptomic database, and KEGG pathways into Cytoscape networks. The generated visualisation highlights gene expression changes and their statistical significance. The plug-in also allows one to browse GenoScript or import transcriptomic data from other sources through tab-separated text files. Genoscape has been successfully used by researchers to investigate the results of gene expression profiling experiments.

Availability: Genoscape is an open-source software freely available from the Genoscape webpage (http://www.pasteur.fr/recherche/unites/Gim/genoscape/). Installation instructions and tutorial can also be found at this URL.

Contact: Mathieu.clement-ziza@biotec.tu-dresden.de; sandrine.rousseau@pasteur.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Gene expression profiling via microarray experiments is widely used in the field of functional genomics, leading to an increase in the size and complexity of the generated data. Analysing such data sets and deciphering the related biological processes can no longer be achieved by traditional methods. Nevertheless, scientists need to understand how the genes, whose expression was measured, are involved in higher functions of biological systems.

Tools have been developed to tackle the massive amount of data produced and improve management, storage, query and analysis of expression data and associated annotations. GenoScript (http://genoscript.pasteur.fr/, Moreira *et al.*, 2002) is such a web application, it can handle microbial microarray experiment data from small projects or large consortium programs, with a private access to unpublished data. GenoScript follows the MIAME 2.0 international standard (Brazma *et al.,* 2001). Cytoscape (Shannon *et al.*, 2003) is an open-source software platform for the dynamic visualisation of molecular interaction networks as graphs. It enables the assignment of gene expression metrics and any comprehensive data to each network element. The Kyoto Encyclopedia of Genes and Genomes Pathway database (KEGG, Kanehisa *et al.*, 2006) is a collection of manually drawn, organism-specific pathway maps representing known molecular interactions and metabolic reaction networks.

Genoscape was developed for biologists to automate the process of locally (i) retrieving statistically analysed expression data from GenoScript, (ii) retrieving biological pathways from KEGG, (iii) integrating those data into Cytoscape and (iv) modifying the visualisation to highlight the level and significance of expression ratio values.

2 SYSTEM OVERVIEW

Genoscape allows users to browse the GenoScript database over a network connection and select transcriptomic data to be imported into Cytoscape. A tab-separated text file can also be used as the input. Genoscape automatically maps most gene or gene product identifiers to KEGG identifiers, enabling the import of expression data from various sources.

When importing KEGG pathways, elements are filtered in order to keep only those nodes corresponding to genes or enzymes. Moreover, additional gene nodes are created when KEGG pathway elements represent enzyme or protein complexes, in order to unambiguously integrate individual gene expression data.

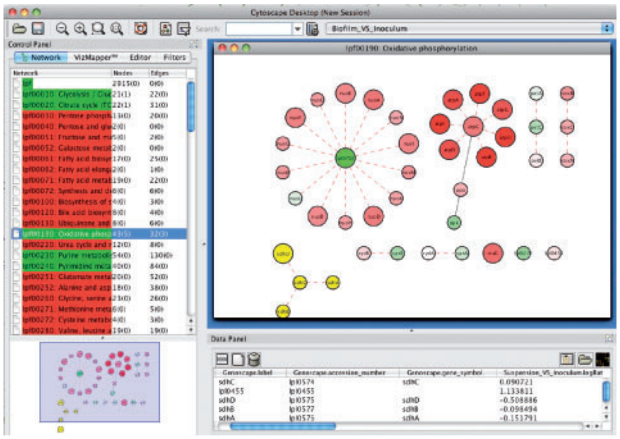
Using Genoscape, KEGG pathways are displayed as Cytoscape networks. Each pathway element is represented as a node. Genoscape generates a visualisation style that highlights gene expression changes and their statistical significance (Fig. 1). Cytoscape graphs produced by Genoscape can be visualised, laid out, modified, and saved in various ways using the built-in Cytoscape features, such as filtering options, or one of the automatic layout algorithms.

3 METHODS AND IMPLEMENTATION

The overall architecture of the plug-in is presented in Supplementary Figure 2. Genoscape was implemented in Java as a Cytoscape plug-in.

[*]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.



Network	Nodes	Edges
lpf	2815(0)	0(0)
lpf00010: Glycolysis / Gluc	21(1)	22(0)
lpf00020: Citrate cycle (TC	22(1)	31(0)
lpf00030: Pentose phosph	13(0)	20(0)
lpf00040: Pentose and glu	2(0)	0(0)
lpf00051: Fructose and ma	5(0)	2(0)
lpf00052: Galactose metal	2(0)	0(0)
lpf00061: Fatty acid biosyn	17(0)	25(0)
lpf00062: Fatty acid elonga	2(0)	1(0)
lpf00071: Fatty acid metab	19(0)	22(0)
lpf00072: Synthesis and de	6(0)	6(0)
lpf00100: Biosynthesis of s	4(0)	3(0)
lpf00120: Bile acid biosynt	8(0)	4(0)
lpf00130: Ubiquinone and	6(0)	6(0)
lpf00190: Oxidative phosp	43(5)	32(3)
lpf00220: Urea cycle and m	12(0)	8(0)
lpf00230: Purine metaboli	54(0)	130(0)
lpf00240: Pyrimidine meta	40(0)	84(0)
lpf00251: Glutamate meta	20(0)	52(0)
lpf00252: Alanine and asp	18(0)	38(0)
lpf00260: Glycine, serine a	23(0)	26(0)
lpf00271: Methionine meta	6(0)	5(0)
lpf00272: Cysteine metabo	4(0)	3(0)
lpf00280: Valine, leucine a	19(0)	19(0)

Genoscape.label	Genoscape.accession_number	Genoscape.gene_symbol	Suspension_VS_Inoculum.logRat
sdhC	lpl0574	sdhC	0.090721
lpl0455	lpl0455		1.133811
sdhD	lpl0575	sdhD	-0.508886
sdhB	lpl0577	sdhB	-0.098494
sdhA	lpl0575	sdhA	-0.151791

Fig. 1. Genoscape user interface. In this example, expression data from a GenoScript experiment and the KEGG pathway Glycolysis/Gluconeogenesis have been integrated into Cytoscape using Genoscape. The nodes represent genes and are coloured with a classical red/green gradient according to the expression ratio level. The size of the nodes is enlarged if the corresponding expression ratio is labeled as statistically significant.

GenoScript relies on a SQL relational database and implements a login procedure to ensure data privacy. All Genoscape data requests are performed through a Common Gateway Interface (CGI). This layer was developed to manage and secure access to the GenoScript database as the user cannot be directly connected to the relational database.

Data import from GenoScript to Cytoscape is performed using specific XML and tab-separated text formats. The tab-separated text format is used to transfer expression data. The XML format is used to browse GenoScript from Cytoscape, as it encloses the relational organisation of experiments and analyses in the GenoScript database.

KEGG pathways and the mapping of KEGG identifiers to external references are retrieved *via* the KEGG FTP server. Several tab-separated text files are generated: one for each pathway and one for the pathway list. To avoid unnecessary requests, pathway data are saved locally, improving the plug-in efficiency in subsequent runs. The correspondence between KEGG pathway elements and a GenoScript or user-generated gene list is automatically achieved by scanning KEGG identifier mapping tables, which include identifiers such as KEGG, Entrez-gene, Ensembl or other widely used species-specific identifiers. To update the visualisation according to expression level changes, Genoscape builds a dedicated customizable VisualStyle that maps the expression data to visual properties.

4 DISCUSSION

An existing KEGG tool (http://www.genome.jp/kegg/tool/color_pathway.html) and Pathway Visualization tool (Arakawa *et al.*, 2005) can be used to colour KEGG pathways but (i) the integration of expression data requires data transformations and (ii) the visualization of networks is static and non editable. Other software such as VisANT (Hu *et al.*, 2005) or KGML-ED (Klukas and Schreiber, 2007) offer a dynamic visualization of KEGG pathways but the integration of expression data is hardly possible, and they do not provide most of the useful Cytoscape features, such as interactive filtering and layout. A Cytoscape plugin, BioNetBuilder (Avila-Campillo *et al.*, 2007), can also be used to import KEGG data into Cytoscape but (i) it merges all KEGG interactions into a single network, making its exploration difficult, (ii) it does not provide a simple way to integrate expression data and (iii) it does not produce a visualisation highlighting gene expression changes.

Genoscape addresses the increasing need for researchers working in the field of gene expression profiling to get a dynamic and unified visualisation of comprehensive biological networks and expression data. This tool makes it possible to efficiently integrate the workflow going from the analysis to the mining of microarray expression datasets in an automated and seamless way.

Genoscape has been successfully used to explore expression data of the eukaryotic organism *Entamoeba histolytica*. *Entamoeba histolytica* is the causative agent of amoebiasis, a parasitic infection of the human intestine and liver. Transcription profiles of the parasite under stress conditions were determined and genes involved in stress response were identified using DNA microarrays (Weber *et al.*, 2006). Genoscape allowed the identification of pathways containing modulated genes, facilitating the analysis of networks and opening up new avenues for studies in basic biology, diagnosis of infectious diseases, and drug development.

Conflict of Interest: none declared.

REFERENCES

Arakawa,K. *et al.* (2005) KEGG-based pathway visualization tool for complex omics data. *In Silico Biol.*, **5**, 419–423.

Avila-Campillo,I. *et al.* (2007) BioNetBuilder: automatic integration of biological networks. *Bioinformatics*, **23**, 392–393.

Brazma,A. *et al.* (2001) Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat Genet.*, **29**, 365–371.

Hu,Z. *et al.* (2005) VisANT: data-integrating visual framework for biological networks and modules. *Nucleic Acids Res.*, **33**, W352–W357.

Kanehisa,M. *et al.* (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.

Klukas,C. and Schreiber,F. (2007) Dynamic exploration and editing of KEGG pathway diagrams. *Bioinformatics*, **23**, 344–350.

Moreira,S. *et al.* (2002) SubScript, une base de données dédiée aux expériences de transcriptome chez *Bacillus subtilis*. *Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM'2002)*, J. Nicolas & C. Thermes, INRIA, pp. 309–316.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Weber,C. *et al.* (2006) Stress by heat shock induces massive down regulation of genes and allows differential allelic expression of the Gal/GalNAc lectin in *Entamoeba histolytica*. *Eukaryot. Cell*, **5**, 871–875.