BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Software

Volume 8(11)

FunSys: Software for functional analysis of prokaryotic transcriptome and proteome

Pablo de Sá[1], Anne Pinto[2], Rommel Thiago Jucá Ramos[1], Nilson Coimbra[1], Rafael Baraúna[1], Hivana Dall'Agnol[1], Adriana Carneiro[1], Alex Ranieri[1], Agenor Valadares[1], Vasco Azevedo[2], Maria Paula Schneider[1], Debmalya Barh[3]* & Artur Silva[1]

[1]Instituto de Ciências Biológicas, Universidade Federal do Pará, Belém-PA, Brazil; [2]Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte-MG, Brazil; [3]Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur, WB-721172, India; Debmalya Barh – E-mail: dr.barh@gmail.com, Phone: +91 944 955 0032; *Corresponding author

Received May 14, 2012; Accepted May 20, 2012; Published June 16, 2012

Abstract:
The vast amount of data produced by next-generation sequencing (NGS) has necessitated the development of computational tools to assist in understanding the myriad functions performed by the biological macromolecules involved in heredity. In this work, we developed the FunSys programme, a stand-alone tool with an user friendly interface that enables us to evaluate and correlate differential expression patterns from RNA sequencing and proteomics datasets. The FunSys generates charts and reports based on the results of the analysis of differential expression to aid the interpretation of the results.

Availability: FunSys and a test dataset are freely available at https://sourceforge.net/projects/funsysufpa/. It requires Sun jdk 6 or higher and MySQL server 5.1 or higher.

Keywords: Next-generation sequencing, Gene expression, Bioinformatics, Software

Background:

New sequencing technologies (high-throughput sequencing) such as SOLiD, Ion Torrent, GS FLX and Illumina are characterized by their greater accuracy, higher-coverage sequencing and reduced read size, all of which present challenges in the processing and analysis of these data. These technologies allow scientists to unravel prokaryotic genomes in a single round of sequencing and at reduced costs compared to the Sanger method **[1]**.

The use of these platforms in functional genomic analyses enables transcriptome-wide analyses via the RNAseq technique, which is a promising alternative to microarrays **[2]**. There are many advantages to this new method, including the reduction or even absence of noise, high-coverage sequencing, transcript detection, low cost and reductions in the time and labour required for sample preparation. In addition, this method produces results similar to those of quantitative PCR **[3-5]**.

A commonly used analytical approach to proteomics studies is the separation of polypeptide chains using 2D electrophoresis **[6]** followed by the isolation and tryptic digestion of spots to identify peptides via mass spectrometry **[7]**. The complementarity between 2D electrophoresis and mass spectrometry generates two pieces of information that can be integrated into transcriptome analyses: the relative volume of each spot and the "Locus_tag" of each protein identified.

The data obtained from transcriptomic and proteomic analysis allow for differential expression analyses that are used to determine which genes are hypo or hyper-expressed under different stress conditions compared to a control condition. Due to the need for large-scale processing of data obtained on Next Generation Sequencing (NGS) platforms and the difficulty of integrating transcriptome and proteome data, we developed FunSys, a programme that can be used to analyze differential expression data from RNAseq and can integrate these analyses with data from proteomic studies

Methodology:

Data Collection

We used the bacterium *Corynebacterium pseudotuberculosis* strain *1002* as a model organism. The bacteria were grown under osmotic stress, acid stress, heat stress and control conditions (Unpublished data). Following sample preparation, cDNA was amplified and sequenced. The sequencing results were filtered, and error correction was performed. The *C. pseudotuberculosis* strain *1002* genome was used to map sequences to quantify the expression level of each gene.

Expression level analysis of cistrons was performed with the Bioscope programme (Applied Biosystems) using the WT Analysis pipeline (SOLiD™ WT Analysis Pipeline) specific for the analysis of transcriptome data. These analyses were used to map the sequence reads to the *C. pseudotuberculosis* strain *1002* genome to determine gene expression levels based on the RNASeq technique. This pipeline generated a final output file (counttag.txt) that consists of a list of genes present in the genome and their respective expression levels in Reads Per Kilobase of exon model per Million mapped reads (RPKM) **[8].**

Proteome data were obtained from a sample file from the ImageMaster 2D Platinum programme (http://www.gelifesciences.com). This file was manually changed so that the numbering of each spot was replaced with "Locus_tags" present in *C. pseudotuberculosis* strain *1002*. The relative volume of each spot was maintained, and all other information in the file was deleted. Finally, the file was converted into a tabular format with two columns: the "Locus_tag" and the relative volume of the spot. The generated file was utilized to validate FunSys integration of the results from the transcriptome and proteome studies.

FunSys

FunSys was developed in JAVA programming language, using the paradigm of object orientation and the graph library Swing http://java.sun.com/docs/books/tutorial/uiswing. It receives ".gff" input files that are generated from EMBL files converted using the Artemis programme **[9]** and "counttag.txt" input files that are generated using the Bioscope programme.

In the differential expression analysis, FunSys calculates fold-change values for each gene, taking into consideration the RPKM ratio between the stress and control conditions (when the ratio is less than one, the negative inverse ratio is listed). These values are used to determine whether the expression of a particular gene is induced, repressed or unchanged, as defined by the user input criteria. By default, FunSys adopts the following parameters: a fold-change > 3.0 indicates induced gene expression; a fold-change < -3.0 indicates repressed gene expression; and genes with fold-changes between -3.0 and 3.0 are considered to have unchanged expression levels **[10]**.

After loading the file containing the proteome information, FunSys can integrate the transcriptome expression data (RPKM) and the proteome data using the "Locus_tag".

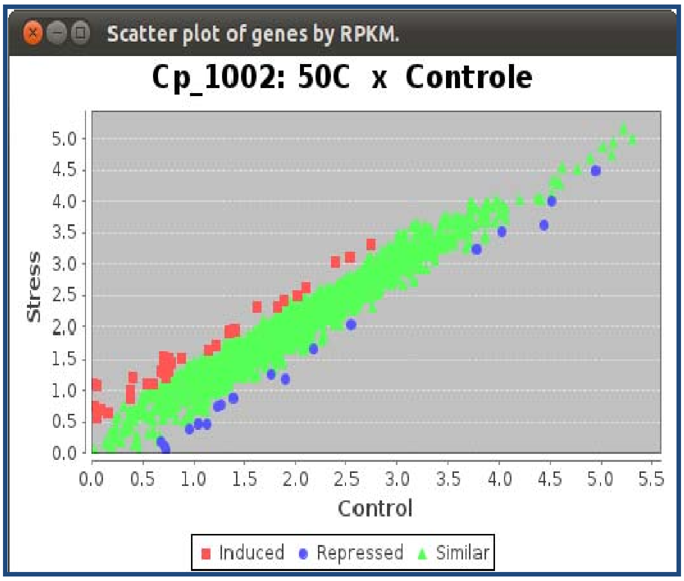

Figure 1: scatter plot of the RPKM values of *C. Pseudotuberculosis* strain *1002* genes for the heat stress and control conditions. In the graph, the RPKM values were converted into a $\log_{10}$ scale (X- and Y-axes). Following conversion, log values less than 0 were excluded because they represent very low RPKM values. The values for the control condition are plotted on the X-axis, and the values for the stress condition (50c) are plotted on the Y-axis.

Discussion:

The sequencing analysis yielded 18,783,810 reads for the osmotic stress condition; 21,622,844 reads for the heat stress condition; 17,393,077 reads for the acid stress condition and 25,235,478 reads for the control condition.

Using FunSys, we analyzed the differential expression of each gene in *C. pseudotuberculosis* strain *1002* between the stress and control conditions and calculated the fold-changes for each gene using the FunSys default parameters. These data were used to determine whether a particular gene was induced, repressed or unchanged. From this analysis, we identified 160 differentially expressed genes (induced and repressed) under the osmotic stress condition, 69 genes under the thermal stress condition and 169 genes under the acid stress condition.

Following the differential expression analysis, FunSys generates a pie chart quantifying the induced, repressed and unchanged genes. A scatter plot of the genes **(Figure 1),** based on RPKM values between the stress and control conditions for each gene, is also generated to provide a graphical representation of the distribution of the RPKM values according to their level of expression.

Following analysis with FunSys, the information generated can be stored in both a tabular format and a graphical format (.png). In addition, a report is also generated of the overall differential expression levels of all genes for all conditions (stress and control conditions), based on a filter available in FunSys.

Conclusion:

FunSys simplifies the differential expression analysis of transcriptome data and integrates these data with proteomics data, thus facilitating the analysis and identification of transcribed and translated genes. In addition, the reports and graphs generated by FunSys will assist users in analyzing their results and generating images for publication.

Acknowledgement:

This work was part of the Rede Paraense de Genômica e Proteômica supported by Fundação de Amparo a Pesquisa do Estado do Pará & CNPq Pronex Núcleo Amazônico de Excelência em Genômica de Microorganismos. Funding: FAPESPA, CNPq e CAPES.

References:

[1] Schuster SC, *Nature*.2008 **5:** 16 [PMID: 18165802]

[2] Teng X & Xiao H, *Sci China C Life Sci*. 2009 **52**: 7 [PMID: 19152079]

[3] Wang Z *et al*. *Nat Rev Genet*. 2009 **10**: 57 [PMID: 19015660]

[4] Morozova O & Marra M, *Genomics*. 2008 **92**: 255[PMID: 18703132]

[5] Pinto AC *et al*. *Genet Mol Res*. 2011 **10:** 1707 [PMID: 21863565]

[6] Rabilloud T *et al*. *J Proteomics*. 2010 **73**: 2064 [PMID: 20685252]

[7] Albaum SP *et al*. *Proteome Sci*. 2011 **9:** 30. [PMID: 21663690]

[8] Mortazavi A *et al*. *Nat Methods* 2008 **5:** 1 [PMID: 18516045]

[9] Rutherford K *et al*. *Bioinformatics*. 2000 **16:** 944 [PMID: 11120685]

[10] Isabella *et al*. *BMC genomics*. 2011 **12:** 51 [PMID: 21251255]

Edited by P Kangueane

Citation: **de Sá *et al.*** 8(11): 529-531 (2012)