CorrelaGenes: a new tool for the interpretation of the human transcriptome

Paolo Cremaschi[1], Sergio Rovida[2], Lucia Sacchi3, Antonella Lisa[1], Alessandra Montecucco[1], Giuseppe Biamonti[1], Silvia Bione[1], Gianni Sacchi[2]

[1]Institute of Molecular Genetics, National Research Council, Pavia, Italy
[2]Institute of Applied Mathematics and Information Technology "Enrico Magenes", National Research Council, Pavia, Italy
[3]Dipartimento di Ingegneria Industriale e dell'Informazione, University of Pavia, Pavia, Italy

Motivation and Objectives

The comprehension of the molecular mechanisms involved in the physiology of human cells and in the pathogenesis of complex disorders, requires the development of new bioinformatic and biostatistic approaches able to integrate and interpret the huge amount of data derived from different kind of "omics" technologies. Nowadays, the interpretation of the transcriptional state of the cell and its alterations in particular experimental or pathological conditions is of particular interest. To this aim several technologies have been developed to identify and quantify the entire set of cellular transcripts, thus resulting in the availability of expression profiles of many different cell types in many different conditions.

With the aim of contributing to the elucidation of transcriptional dynamics in the cell, we developed CorrelaGenes, a new bioinformatic tool that exploits the expression data available in the Gene Expression Omnibus (GEO http://www.ncbi.nlm.nih.gov/geo/) database. The main goal of this tool is to help identifying sets of genes whose expression appeared simultaneously altered in different experiments, thus suggesting co-regulation or coordinated action in the same biological process.

Methods

CorrelaGenes uses a PostgreSQL (http://www.postgresql.org/) 9.1.3: database initialized using the Curated DataSets in Homo sapiens cell lines publicly available in the GEO archive. The Extract Transform and Load process, described in Figure 1A, was created using the R language 2.14.1 available at The R Project for Statistical Computing (http://www.r-project.org/) .

A total of 978 GEO DataSets were read using the GEOquery R package 2.21.9 (Davis and Meltzer, 2007) and transformed in objects suitable for the subsequent stages of the analysis. The DataSet design was manually analyzed to select 2120 biologically meaningful experimen-

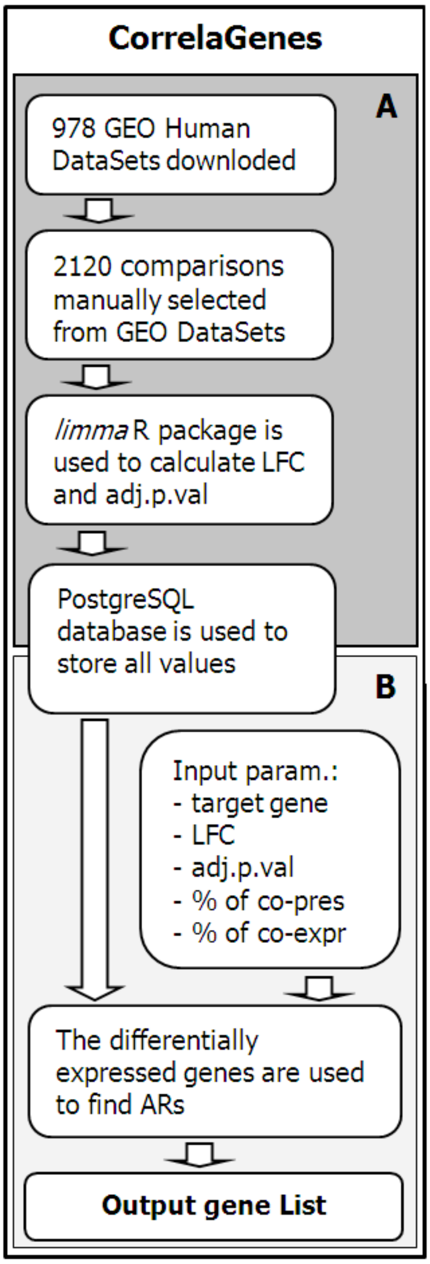
CorrelaGenes
A
978 GEO Human DataSets downloded
2120 comparisons manually selected from GEO DataSets
limma R package is used to calculate LFC and adj.p.val
PostgreSQL database is used to store all values
B
Input param.:
- target gene
- LFC
- adj.p.val
- % of co-pres
- % of co-expr
The differentially expressed genes are used to find ARs
Output gene List

Figure 1: CorrelaGenes workflow. Panel A: PostgreSQL database initialization (R language). Panel B: Data Mining process (Fortran language).

tal comparisons. All the 2120 selected pairwise comparisons were analyzed with the limma R package 3.10.3 (Smyth, 2005) to calculate the log fold change (LFC) and the adjusted p values (adj.p.val) to identify the list of differentially expressed genes. All the values obtained from limma were stored in the PostgreSQL database.

The Association Rule Mining (ARM) is the unsupervised data mining technique that we used to discover genes that are frequently differentially co-expressed (Creighton and Hanash, 2003) in GEO DataSets. We used the standard ARM algorithms to look for Association Rules (ARs) limited to two genes (namely: IF Gene1 is differentially expressed THEN also Gene2 is differentially expressed) one of which is defined as an input parameter fixed for each search (i.e. target gene). The constraints used add a guided approach to the standard ARM technique with the aim of creating a list of genes sharing a coordinated expression with the target.

We defined two different criteria to select the most relevant ARs:

- percentage of co-presence (% of co-pres): as not all the comparisons include the same set of probes or some probes could be discarded for a not significant adjusted p value, we created an index to evaluate the percentage of comparisons where a gene is measured in relation of the whole number of comparisons where the target gene is measured;
- percentage of co-expression (% of co-expr): to evaluate the significance of the relationship between a gene and the target, we calculated the percentage of comparisons in which both genes are differentially expressed in relation of the number of comparisons where they are both measured.
- The procedure to perform the co-expression analysis, described in Figure 1B and implemented by a serial Fortran90 prototype code, can be summarized as follows:
- choice of the target gene and setup of the user defined indices for the analysis;
- initialization of the data structures (LFC and adj.p.val);
- identification of differentially expressed probes (a matrix of integer flags is defined, in order to select up-, down- and not-regulated or not-significant probes);
- selection of probes and comparisons associated to the target gene;
- evaluation of the percentage of significant values of both co-pres and co-expr for each single gene;
- creation of the list of all genes matching the selected criteria.

Results and Discussion

A total of 15 target genes (ACTG1, AFF3, APOE, APP, CDC5L, DIAPH2, EMD, FOXO1, HIF1A, IL8, MAPT, PRFP19, PSEN1, PSEN2, PTPN22) were used for the preliminary validation of the procedure with the following criteria: (i) adj.p.val <= 0.05, (ii) absolute value of LFC >= 0.65 (iii), % of co-pres >= 40% and (iv) % of co-expr>= 30%.

The simulations were carried out using a single blade of the CentOS IBM Cluster at IGM-CNR in Pavia. The cluster consists in six computational nodes, interconnected by Gigabit Ethernet and 10G Fiber Channel. Each node is a two processors Intel Xeon E5640 2.66 GHz, sharing 48 GB of RAM. The performance of the algorithm was evaluated using the execution time.

Averaging on the considered 15 genes, the whole procedure requires a mean execution time of 1221 sec for the co-expression analysis of a single gene. We evaluated the average cost of each phase as percentage of the total execution time. The profiling of the code showed that 64.3% of the total time is spent initializing the data, 35.5% is spent creating the different gene lists and only the 0.2% is actually spent gene-rating the ARs. The analysis algorithm exhibits an intrinsic data-parallelism at the level of the processing of the gene, a feature that will be further investigated in order to improve the performance of the whole procedure. A naïve approach to the parallelization consists in the multithreaded implementation for the creation of the gene lists by means OpenMP directives. Anyway, as the limiting step is the data initialization, a brand new approach to overcome this problem could be considered.

The gene lists created starting from the selected 15 target genes, were analyzed for their biological content in order to assess the relevance of the results obtained.

A first observation regards the highly variable number of associated genes extracted for each target gene (i.e. ranging from 99 to 2951) that could be due both to the different number of comparisons in which the target gene was modulated or to the different transcriptional behavior of the genes in the cell. Moreover, we found a quite large number of genes shared by all the 15 lists. This could either reflect the presence of constitutively modulated genes eventually involved in basic cell processes or be the consequence of a too tolerant choice of the parameters used in the simulation.

Some more detailed biological characterization was performed for the 2014 genes of the list extracted with PRPF19 as target the Database for Annotation, Visualization and Integrated Discovery (DAVID, http://david.abcc.ncifcrf.gov/). We used the Database for Annotation, Visualization and Integrated Discovery (http://david.abcc.ncifcrf.gov/) to query the Gene Ontology (GO, http://www.geneontology.org/) for the Biological Process subset of terms. Consistently with the literature data, the GO terms found significantly enriched (Benjamini corrected p value < 0,05) were related to the main known functions of PRPF19 in the cell (i.e. cell cycle, apoptosis, pre-mRNA splicing, DNA damage repair). We also investigated the gene list extracted for CDC5L (n=2794), a gene known to interact with PRPF19 in the pre-mRNA splicing complex (Grote et al., 2010). Despite the fact that the two genes were not selected as associated, a large overlap was found between the two lists (1531 genes in common). This list contains mainly genes related to cell cycle and splicing process. Moreover, an analysis with data obtained with the GeneMANIA (http://www.genemania.org/) and the STRING (http://string-db.org/) web tools for the two genes gave an independent confirmation for a number of genes extracted by our CorrelaGenes tool. Finally, a set of five genes involved in the pathogenesis of Alzheimer disease (Carter, 2007), a common human neurodegenerative disorder, were included in our simulation (APOE, APP, PSEN1, PSEN2 and MAPT). A group of 952 genes were found in common among the five extracted lists thus suggesting the presence of shared pathways that could be exploited for further investigation of pathogenetic mechanisms.

The preliminary results of the simulation showed how CorrelaGenes could contribute to the characterization of transcriptional profiles in the cell and in the definition of molecular pathways and biological process. Moreover, it integrates expression results obtained from other available tools. The good performances shown during the simulation phase encourage us to plan wider validation steps to enhance the accuracy and the reliability of our instrument.

Acknowledgements

This work was supported by Cariplo Foundation grant n. 2010/0253 and by Progetto di interesse CNR-MIUR "Invecchiamento".

References

1. Carter CJ (2007) Convergence of genes implicated in Alzheimer's disease on the cerebral cholesterol shuttle: APP, cholesterol, lipoproteins, and atherosclerosis. Neurochem Int. 50(1) 12-38
2. Creighton C, and Hanash S (2003) Mining gene expression databases for association rules. Bioinformatics 19(1) 79-86. doi:10.1093/bioinformatics/19.1.79
3. Davis S, and Meltzer PS (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics 23(14):1846-1847. doi: 10.1093/bioinformatics/btm254
4. Grote M, Wolf E, Lemm I, Agafonov DE, Schomburg A, et al. (2010) Molecular Architecture of the Human Prp19/CDC5L Complex. Mol Cell Biol. 30(9) 2105-2119. doi: 10.1128/MCB.01505-09
5. Smyth GK (2005) Limma: linear models for microarray data. In: Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W (Eds.) Bioinformatics and Computational Biology Solutions using R and Bioconductor. Springer, pp. 397-420