A User-friendly Web Bioinformatics Tool to Store, Analyse and Visualize Affymetrix Genechip Human Exon 1.0 ST Array Data of Disease Experiments

Consiglio A (1), Delle Foglie G (1), Giovannelli C (1), Grillo G (1), Ianigro M (2), Licciulli F (1), Liuni S (1)

(1) Institute for Biomedical Technologies (ITB) - Bari National Research Council - CNR, via Amendola 122/D, 70126 Bari, Italy
(2) Institute of Intelligent Systems for Automation (ISSIA) National Research Council - CNR, via Amendola 122/D, 70126 Bari, Italy

Motivation
Recent studies have shown that more than 90% of human multi-exon genes are subjected to Alternative Splicing (AS), a key molecular mechanism in which multiple transcripts may be generated from a single gene. The breakdown in AS mechanisms is widely recognized to play an important role in cellular differentiation and pathologies [1, 2]. PCR, microarray and sequencing technologies have been applied to the study of transcript diversity arising from alternative expression [3]. Last generation Affymetrix GeneChip Human Exon 1.0 ST Arrays offer a more detailed view of the gene expression profile providing information on the AS pattern. Exon arrays, with more than six million data points, can detect approximately one million exons, and allow to perform analyses at both gene and exon level. In particular, the access to genome-wide exon data can provide answers to challenging issues concerning the identification of tissue-specific exons, biomarkers and isoforms involved in different pathological mechanisms. In the last years many statistical algorithms have been developed for extracting accurate information from exon arrays data, both open source and commercial, but they are complex to use and do not allow an immediate interpretation of results. In the present study we developed an integrated user-friendly bioinformatics tool to store, analyse and visualize exon arrays data, based on a data-warehouse approach combined with a number of useful statistical tools to assess differential expression between normal and tumoral tissue. The bioinformatics platform is applied to a new data set of exon arrays. The data set is made up of fourteen couples of normal and tumoral specimens extracted from fourteen patients affected by colon cancer, the third most common cancer diagnosed in both men and women.

Methods
Our bioinformatics tool, through the use of a data warehouse built with the Infobright database engine [4], can manage a huge amount of data and integrate the experimental
data produced by exon arrays with the related annotations from several databases such as ASPicDB [5], RefSeq, HGNC, GO and KEGG. The system imports via web interface the raw Affymetrix CEL files, estimates exon level and gene level signal through some automated preprocessing steps, and stores them in the data warehouse. When medical record data are available, they can also be stored to add further information useful to tag experiment data in a more accurate way. In order to analyze the huge number of data produced by this step, data are enriched with all the standard statistical tools useful for extracting AS events. The tool is completed with a web interface that provides advanced querying tools to allow a user-friendly access to data and statistical results. Statistical plots and alternative transcripts images coming from RefSeq transcripts and ASPic predicted transcripts, give the support for the discovery and the analysis of novel AS events.

Results
At the present, the data warehouse contains data from experiments on colon cancer, which can be queried by gene name, accession, Affymetrix transcript cluster or probe-set id, GO, chromosomal position and KEGG pathway. In particular, we stored gene and exon level expression intensities and medical records. For each gene, the tool shows: alternative names, accession IDs, GO terms and biochemical pathways. Then it plots exon level and gene level expression, highlighting all possible AS events, and provides statistical support to these results by standard methods, such as DABG, Student’s t-test, ANOVA, and MiDAS. The user can customize the plots choosing the statistics parameters and exploiting the medical record data of the samples. The exon level plots are supplied with the graphical representation of RefSeq and ASPic transcript variants, which is useful for an immediate interpretation of AS events. Gene level expression data can be grouped by GO terms or biochemical pathway.

References

Contact email
arianna.consiglio@ba.itb.cnr.it