

Analysis of microarray data with THEA

<http://thea.unice.fr>

Pasquier C. (claud.pasquier@unice.fr)

Christen R. (richard.christen@unice.fr)

Virtual Biology Lab

Institute of Signaling, Developmental Biology and Cancer Research

Laboratory of Virtual Biology

CNRS UMR 6543 Centre de Biochimie

06108 Nice Cedex 02; France.

Reference:

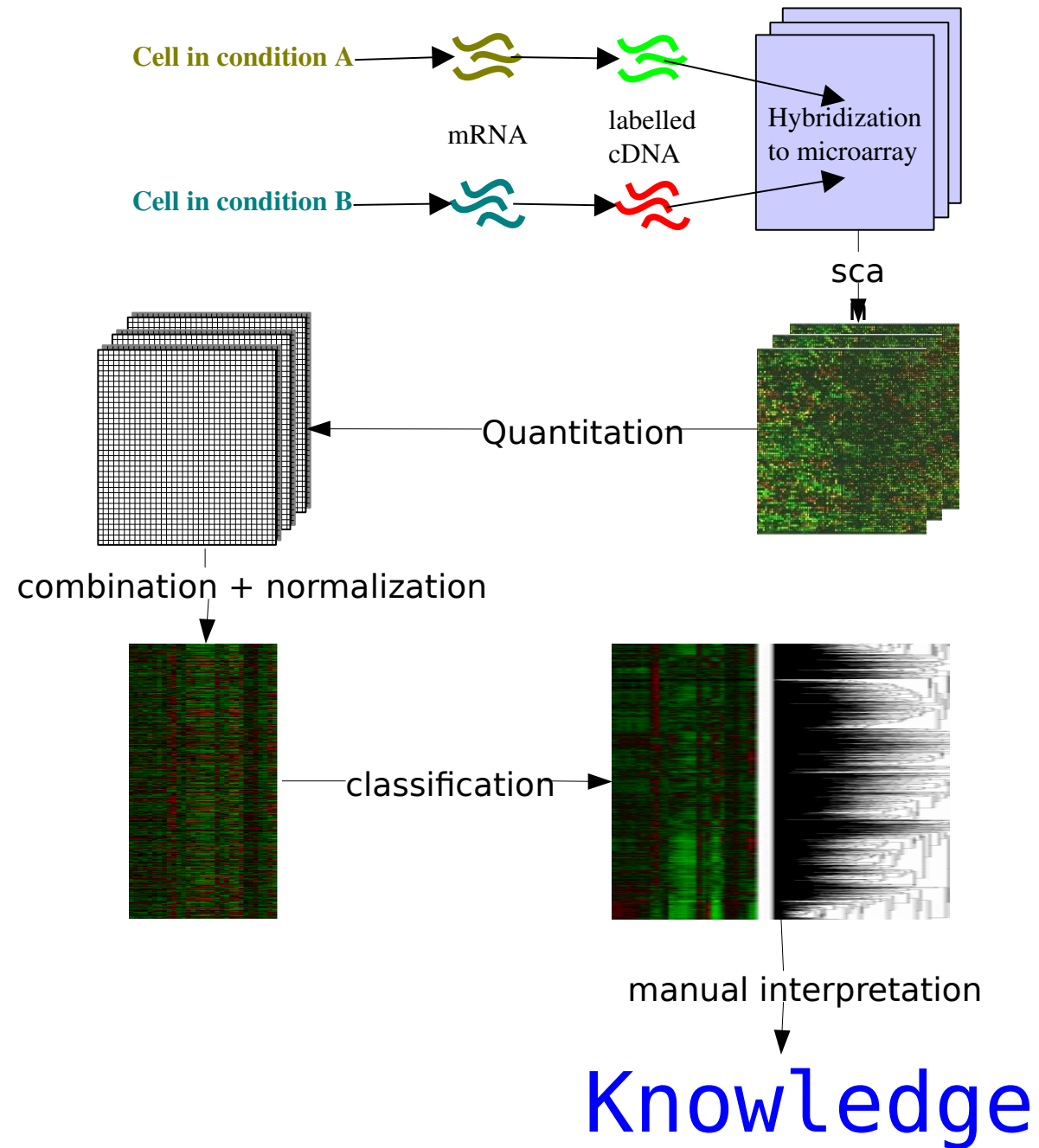
Pasquier C., Girardot F., Jevardat de Fombelle K. and Christen R. THEA: Ontology driven analysis of microarray data. Bioinformatics Advance Access published on May 6, 2004

The problem

Microarray technology makes it possible to measure thousands of variables and to compare their values under hundreds of conditions.

Once microarray data are quantified, normalized and classified, the analysis phase is essentially a manual and subjective task based on visual inspection of classes in the light of the vast amount of information available for each gene.

Currently, data interpretation clearly constitutes the bottleneck of such analyses and there is an obvious need for tools able to fill the gap between data processed with mathematical methods and existing biological knowledge.



The THEA project : overview

The THEA project (Tools for High-throughput Experiments Analysis) is dedicated to the elaboration of tools and methods suited for the analysis of post-genomic data.

The first module developed allows to:

Automatically link gene symbols with selected biological information coming from a knowledge base,

Manually search and browse through these annotations,

Automatically generate meaningful generalizations according to statistical criteria.

THEA makes use of ontologies which constitute a popular way to modelize biological concepts and their relationships. Presently THEA includes:

Gene Ontology (GO), base on a controlled vocabulary which can be used to annotate a gene product in regard to its molecular functions, cellular localizations and biological processes,

Specific vocabularies describing the developmental stages and the anatomy of *Drosophila melanogaster*,

The Medical Subject Headings (MeSH), used for indexing biomedical and health-related documents,

InterPro, a collection of protein families, domains and functional sites.

This example of configuration of THEA's user interface shows six activated plugins.

The **Terms viewer** (top left), allows to graphically browse through the loaded ontologies and select individual terms.

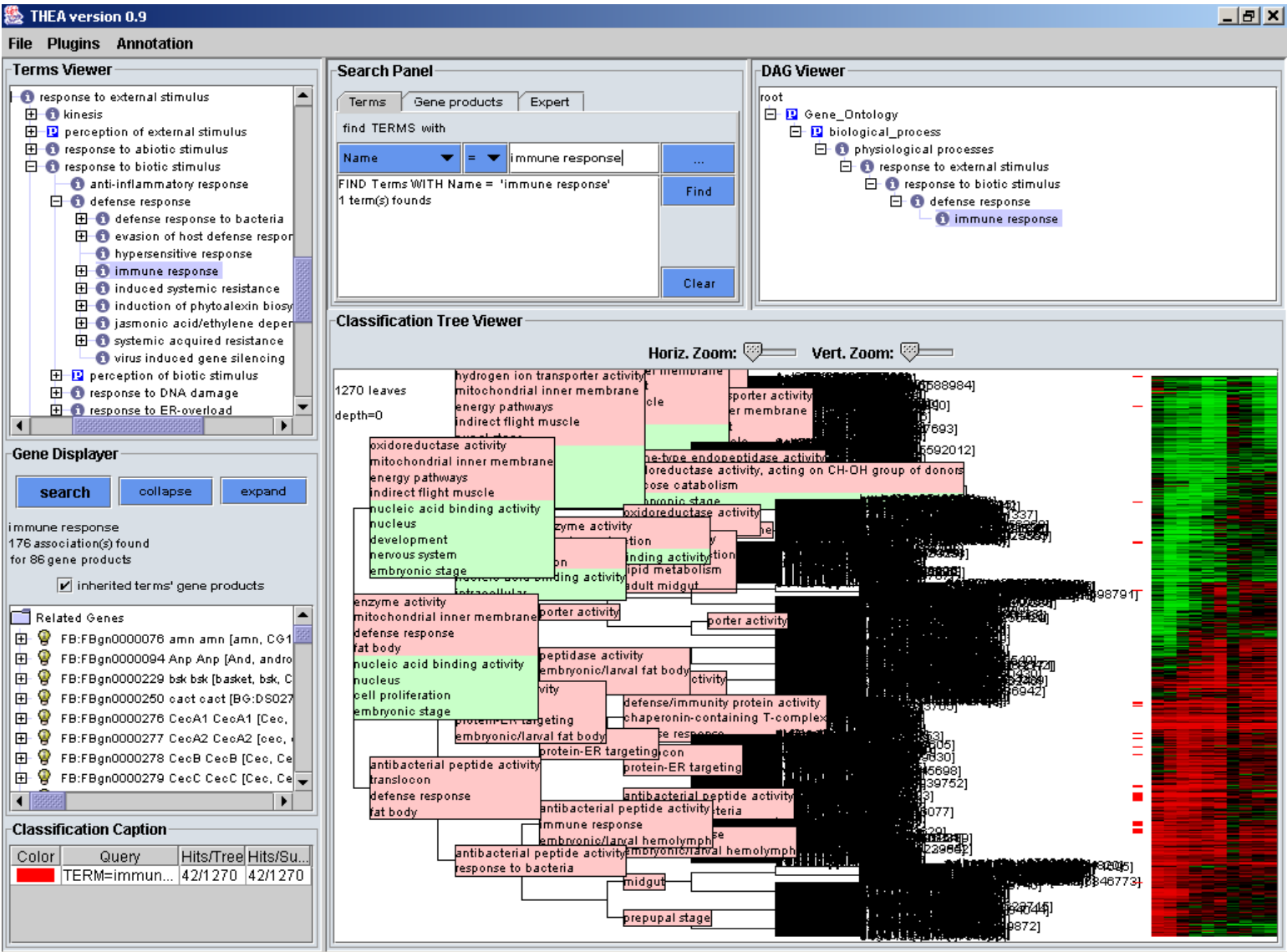
The **Search Panel** (middle top) allows to search for specific terms or character strings among genes or terms. Complex queries are also possible. Here the term “immune response” has been selected.

The **DAG Viewer** (top right), displays a trimmed ontology leading to the selected term.

The **Gene Displayer** (middle left) displays the genes associated to a given term and, eventually, their children (here the 86 genes associated to the term ‘immune response’ and their children are displayed) and provides access to the relevant publications and databases *via* Internet.

The **Classification Caption** (bottom left) displays the legends of the color-coded labels as well as some statistical information.

The **Classification Tree Viewer** (bottom right) allows the display of hierarchical clustering results of gene expression profiles. Genes associated to the selected term and its children are labeled with colored boxes. The names automatically assigned to each cluster are displayed in this plugin as well.



Analysis of a dataset consisting of 1270 regulated genes (De Gregorio, E., Spellman, P.T., Rubin, G.M. and Lemaitre, B. (2001) Genome-wide analysis of the Drosophila immune response by using oligonucleotide microarrays. Proc Natl Acad Sci U S A, 98, 12590-12595).

The THEA project : features and facts

Thea v1.2.3, released September 2004, includes 11 plugins.

The application can be used to analyze a single list of genes or complete classification trees in 3 different formats (Newick, Eisen, Sota)

Groups of co-expressed genes can be automatically labelled with terms found to be over- and/or under -represented in the set using several (parametrable) methods: terms count, density of terms, statistical calculation (Binomial or Hypergeometric law).

Calculations can be relative to gene products present in the classification, known genes in a genome or a user-specified list of genes.

The associated database contains:

250 871 genes from 13 organisms (*S. cerevisiae*, *C. briggsae*, *C. elegans*, *A. gambiae*, *D. melanogaster*, *A. mellifera*, *D. rerio*, *G. gallus*, *P. troglodytes*, *H. sapiens*, *M. musculus*, *R. norvegicus* & *T. nigroviridis*)

61 690 genes annotated with at least one ontology term

1 668 845 known identifiers in 36 databases, including Affymetrix, EMBL/GenBank, LocusLink, RefSeq, UniGene, UniProt, Ensembl, Hugo

Collaborations

Biological labs

CNRS, ESCPI (UMR7637), Marie-Claude Potier
CNRS, UNSA (UMR 6078), Frank Delaunay
CNRS, IPMC (UMR 6097), Pascal Barbry

Software environment

INRIA, SmartTools project, Didier Parigot

Complexe manipulations of tree structures

IRD (UMR2724), François Chevenet

Data Mining

CNRS, I3S (UMR6070), Martine Collard

Text Mining

CNRS, UNSA (UMR6543), Florence Horn