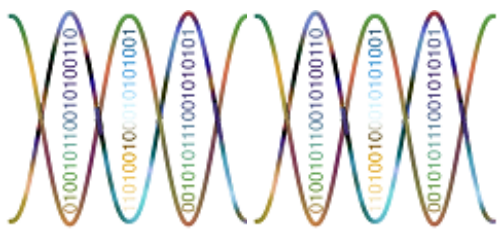
# GAAS

# Gene Array Analyzer Software

*Marco Masseroli, PhD*

*Bioengineering Department, Politecnico di Milano*

*masseroli@biomed.polimi.it*



# Rationale

- Analysis of changes in gene expression profiles can lead to identify transcription signatures and regulatory pathways
- Microarrays experiments enable measuring thousands of gene expression levels at once **but** generate a huge amount of "not very clean" data

- Experiment data need to be efficiently managed, properly analyzed, and results must be adequately and concisely visualized to help uncover new biological knowledge

# Main Issues

- Different high-throughput gene expression methodologies (e.g. nylon filters, spotted cDNA microarrays, oligonucletide chips)

- Variability of each gene expression experiment

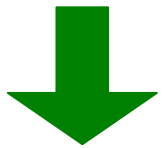
  using several replicate experiments, at least three (Lee et al., 2000)

- Relevant expression intensity variability and noise among experiments

  careful data preprocessing (e.g. background analysis, normalization)

# Requirements

- The large amount of spotted genes in a single array,
- the number of arrays used in replicate experiments,
- the analyses needed for the massive data produced,



- require a software framework able to manage and rapidly analyze thousands of input data simultaneously.
- This framework should be flexible through the use of parameters, and present a friendly graphical user interface.

# Gene Array Analyzer Software

To fit the stated requirements, we designed the Gene Array Analyzer Software (GAAS).

- GAAS is an integrated multi-stage and multi-user software system
- It is structured in management, analysis and visualization sections
- GAAS has been implemented in MS-Visual C++ programming language and interconnected to a relational database system developed using MS-Access 2000

# Management Section

The management section enables each user:

- processing different kind of input data sets
  - input data templates are used to specify the content of any input data column
  - a preprocessing stage uses these templates to reduce any input data format to a built-in database data structure
- setting, storing and retrieving own analysis parameter sets
- defining custom data output configurations for storing analysis results in output databases

# Analysis Section

Data analysis is made flexible by several parameters and subdivided into sequential steps:

1. Automatic topological background and spot quality evaluation, and background correction
2. Within-array data normalization (mean or median intensity of all high quality clones in the array or using a subset of selected clones)
3. Differential gene expression evaluation on expression intensity ratios
   - in a single experiment (automatically determined log-ratio confidence intervals, or interactively defined folding threshold)
   - on multiple replica experiments (conditional probability model, or single experiment gene regulation reproducibility)

# Analysis Section 1

Data analysis is made flexible by several parameters and subdivided into sequential steps:

1. Automatic topological background and spot quality evaluation, and background correction (Nadon and Shoemaker, 2002)
2. Within-array data normalization (mean or median intensity of selected clones)
   - global approach using all high quality clone in the array
   - using a subset of control genes (housekeeping or spikes of heterologous genes) assumed not to vary among the evaluated conditions (Yang et al., 2002)

# Analysis Section 2

3. Differential gene expression evaluation on expression intensity ratios
   - in a single experiment with respect to:
     - automatically determined confidence intervals determined on log-ratio distribution (normality assessed by Kolmogorov-Smirnov test implemented with the Lilliefors correction)
     - an interactively defined folding threshold (Chen et al., 1997; Nadon and Shoemaker, 2002)

# Analysis Section 3

- on multiple replica experiments according to:
  - a conditional probability model (for a given gene, the regulation probability determined in each experiment is used to calculate the regulation conditional probability in the considered replica experiments)
  - single experiment gene regulation reproducibility.

When log-normality of expression intensity ratios is not verified, the user-defined folding threshold and regulation reproducibility cut-off provide acceptable results without any *a priori* assumption on differential expression distribution.

# Visualization Section

Data and analysis results can be visually navigated both in tabular and graphical form.

- Tabular data collect expression levels, quality labels, regulation results, and several gene identifiers.
- Histogram plots allow easy comparison of expression intensity distributions of multiple experiments.
- Scatter plots of expression levels immediately give insights into gene regulation.

Implemented clone search and navigation procedures enable to move interactively from input data to tabular results and plots to investigate single gene behavior.

# Availability

GAAS software package and supplementary information are freely available for academic and non profit use at:

**http://www.medinfopoli.polimi.it/GAAS/**