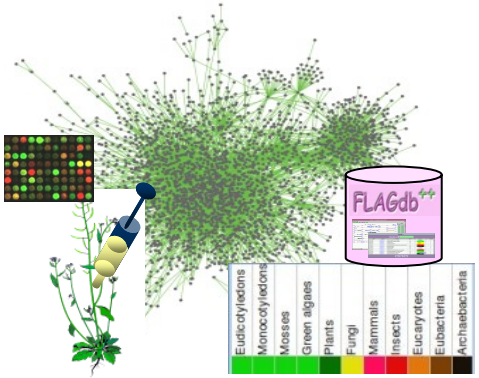


SONATA: STRESS, ORPHAN, NETWORK AND TRANSCRIPTOME IN ARABIDOPSIS

Marie-Laure Martin-Magniette

URGV : Unit of Plant Genomic Research
UMR INRA-UEVE-CNRS
Team : Bioinformatics for predictive genomics

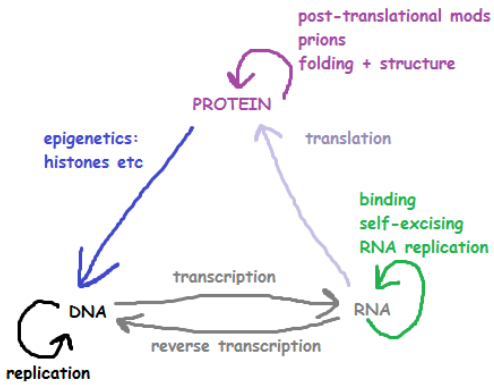
UMR AgroParisTech/INRA Applied Mathematics and Informatics
Team: Statistics and Genome

Current challenge in genomics

- It is now relatively easy to sequence an organism and to localize its genes.
- Nearly 40% of the predicted genes have no assigned function (Hanson et al., 2010)
- New challenge is the functional annotation i.e identifying the function(s) of each gene
- Study of the sequence similarity is not enough since sequence similarity does not necessarily imply a similarity of protein structure or function

EVOLUTION OF THE DOGMA

- One gene-one enzyme hypothesis is now considered as an oversimplication

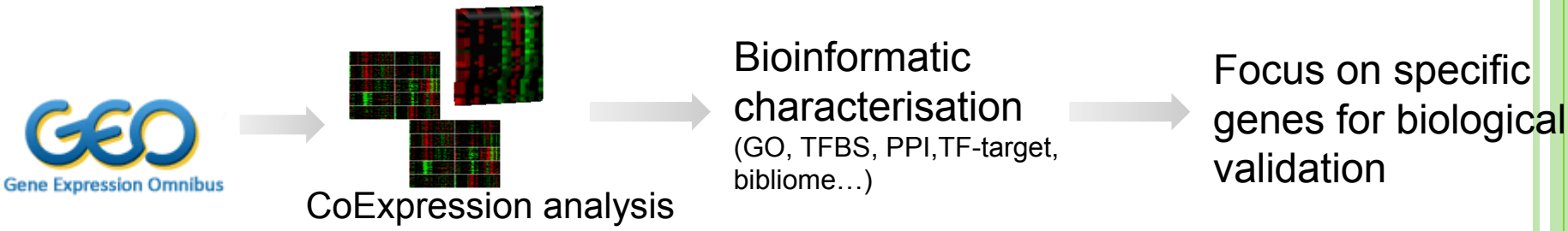


- Nowadays, we prefer to speak about functional complex involving many genes

- High-throughput technologies allow one to have access to the transcriptome (set of the transcribed genes in a given sample)

- Studying transcription in a various set of context allows to identify co-expressed genes, which are good candidates to be involved in a same biological process (Eisen et al, 1998)

CLASSICAL FLOWCHART



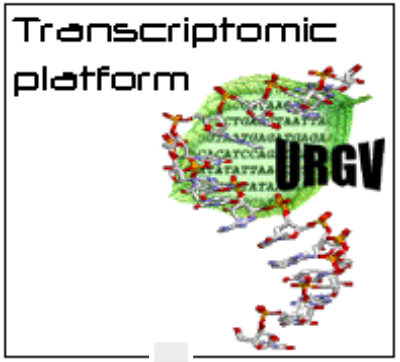
Drawbacks

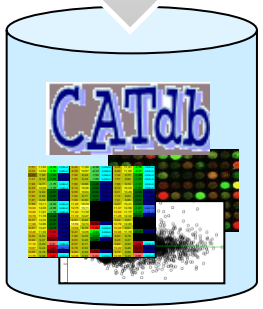
- Data are generally extracted from international repositories
- It leads to heterogeneous data in terms of acquisition and preprocessing
- Co-expression generally done by analyzing gene pairs (Pearson correlation)
- Difficult to interpret since the number of gene pairs is large
- It is a local point of view of a complex question

Arabidopsis thaliana

- First plant sequenced in 2000
- About 25 000 genes
- Only 14% of the genes have a validated function
- 20 % of genes are orphean (no information on their function), generally discarded of the published works …
- Large transcriptomic ressources are available

STRESS, ORPHAN, NETWORK AND TRANSCRIPTOME IN ARABIDOPSIS





Since 2003
> 300 collaborations
> 14 500 hybridizations
110 publications

Gagnot et al., 2008

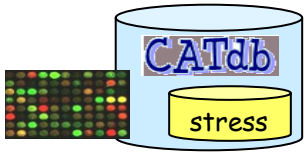
Goal: Explore the orphan gene space to identify **new candidate genes** involved in defense and adaptation process.

Method: Predict co-expression networks using model-based clustering

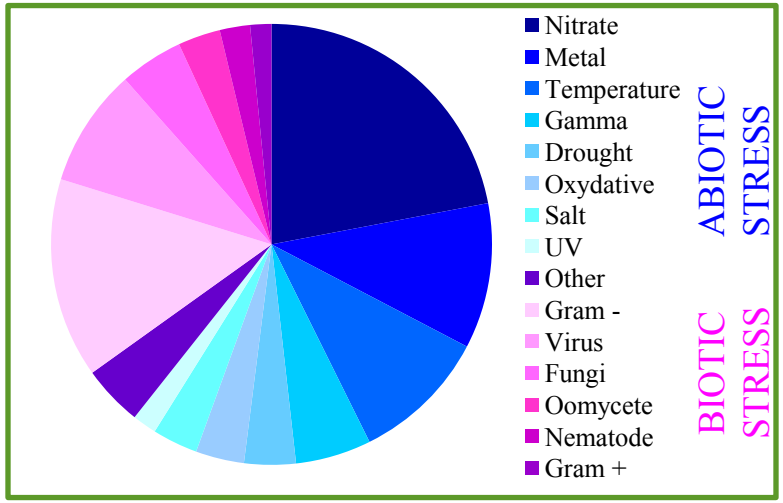
Data: An original transcriptomic resource generated by the platform of URGV, stored in a dedicated database

- Homogeneous transcriptomic data
- ~ 6000 genes not present in Affymetrix chip
- High diversity of biological samples relative to stress.

CREATION OF THE DATASET



424 comparisons performed in dye-swap



Extraction of the raw pvalues calculated for the differential analysis
FWER controlled at 5% across comparisons and genes

60% of the genes (> 18000) have transcription 'impacted' (directly or not) by stress

Large overlap of impacted genes between biotic and abiotic stress

Expected number by the biologists at the beginning of the project : 3 000 genes !!

MODEL-BASED CLUSTERING

- Data are n genes described by Q variables : $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ where $\mathbf{y}_i \in \mathbb{R}^Q$ are iid of unknown density h

- Data are assumed to come from several subpopulations modeled separately and the whole population is the mixture :

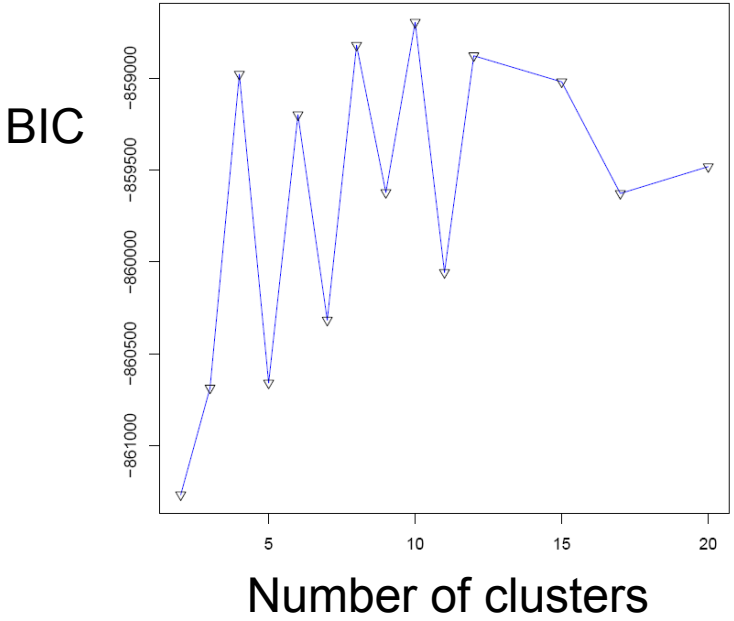
$$f_{\text{clust}}(.|K, m, \alpha) = \sum_{k=1}^{K} p_k \Phi(.|\mu_k, \Sigma_k)$$

with

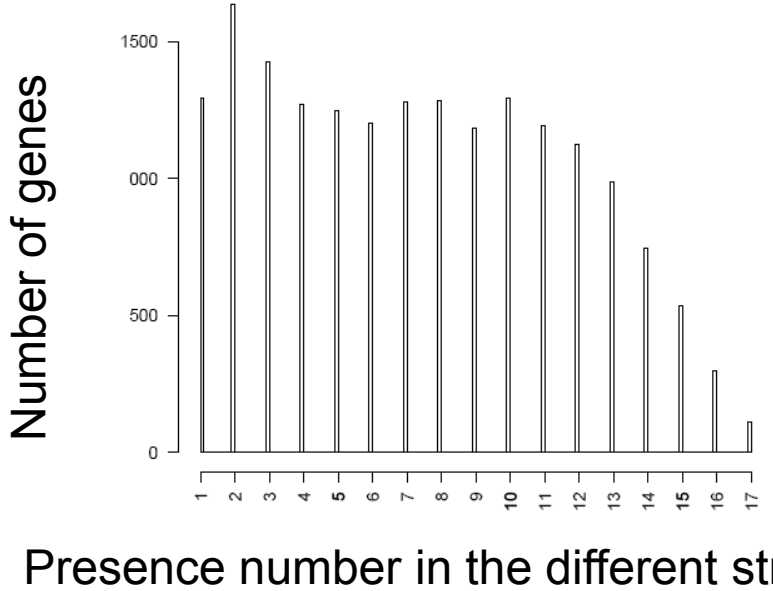
  - K the number of clusters (*i.e. subpopulations)*
  - $\alpha = (\mathbf{p}, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K)$ *where* $\mathbf{p} = (p_1, \ldots, p_K)$, $\sum_{k=1}^{K} p_k = 1$
  - $\Phi(.|\mu_k, \Sigma_k)$ *the density function of* $\mathcal{N}_Q(\mu_k, \Sigma_k)$

- The selected model $(\widehat{K}, \widehat{m})$ maximizes the BIC criterion.

APPLICATION ON THE WHOLE BIOTIC DATASET



BIC varies a lot and the curve is not convex



Large variability in the response according to the stress.

Whole dataset too heterogeneous
to be analyzed without a priori knowledge

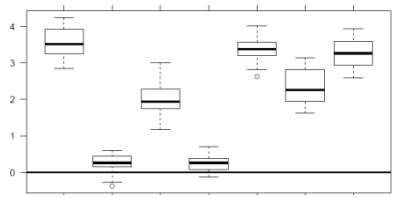
GENE CLUSTERING BY STRESS

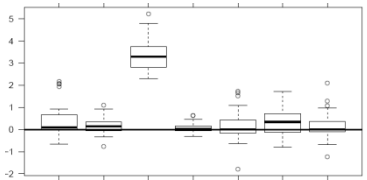
Matrix
{ DE gens x expression differences }

Gaussian Mixture Model

Model selection with BIC
+
Classification rule for controlling the misclassification rate

~700 Clusters of Coexpressed Genes

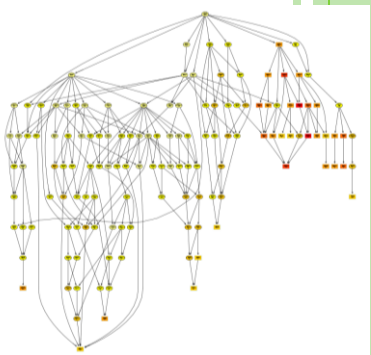




Stress	Genes	Clusters
Nitrogen (root)	14 139	60
Nitrogen(rosette)	13 495	59
Temperature	11 365	34
Heavy metal	10 617	57
Oxydative stress	10 127	52
Drought	8 143	34
UV	7 894	37
Salt	5 729	30
Gamma	5 350	32
Necrotrophic bacteria	11 220	50
Biotrophic bacteria	12 023	56
Virus (rosette)	11 832	54
Fungi	9 773	51
Nematodes	7 413	27
Oomycetes	5 508	31
Rhodococcus	1 900	13
Stifenia	1 525	17

Coexpression clusters linked to functional information

Gene Ontology



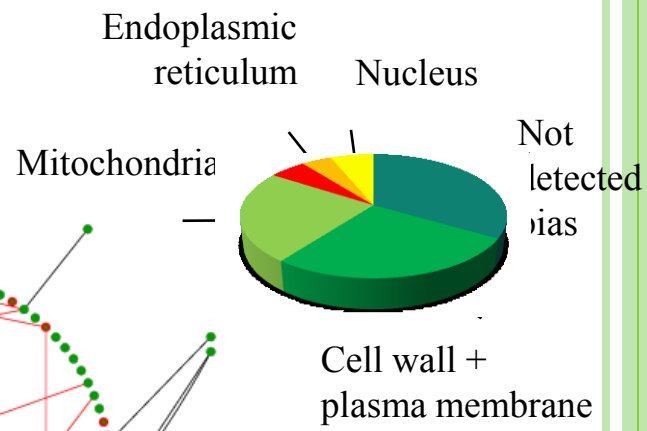
According to GO and literature

61% of the clusters are enriched in genes involved in stress responses

17% of the clusters are enriched in transcription factors

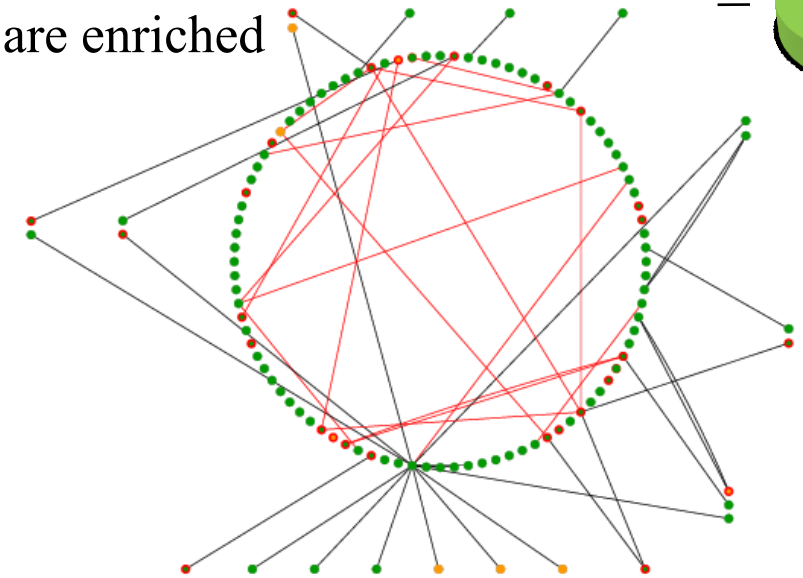
Predicted subcellular localization of proteins

77% of clusters have bias



Protein-protein interactions

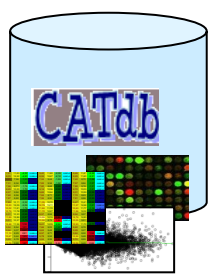
31% of the clusters are enriched

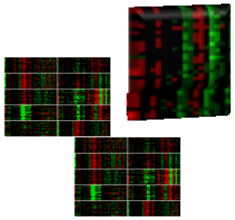


- Transcription factors
- Stress related genes
- PPI intra-cluster
- Other PPI

FIRST CONCLUSIONS

- Model-based clustering helped us to understand that the clustering should not be performed naively on the whole dataset
- An analysis per stress seems obvious but no biologists had told me that it was the correct way to analyze the data ….
- Coexpression depends on the stress conditions meaning that functional modules vary with the environment





CoExpression analysis

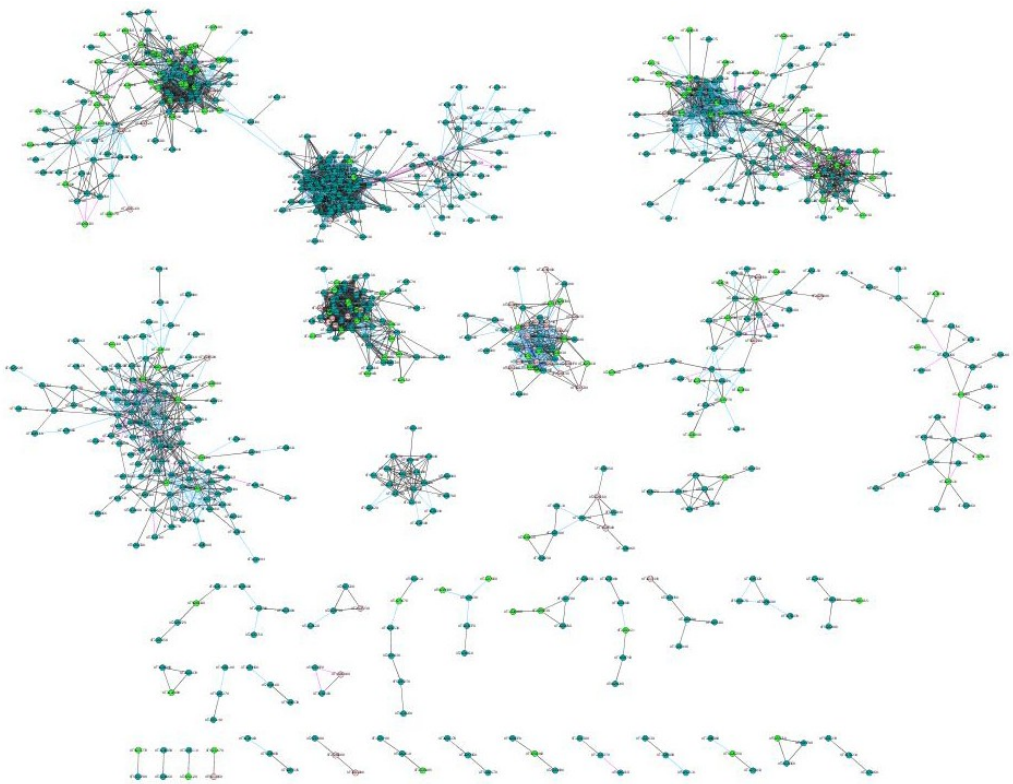
Bioinformatic characterisation
(GO, TFBS, PPI,TF-target, bibliome…)

20 coexpression analyses

Interpretation is not straightforward

Coregulation networks

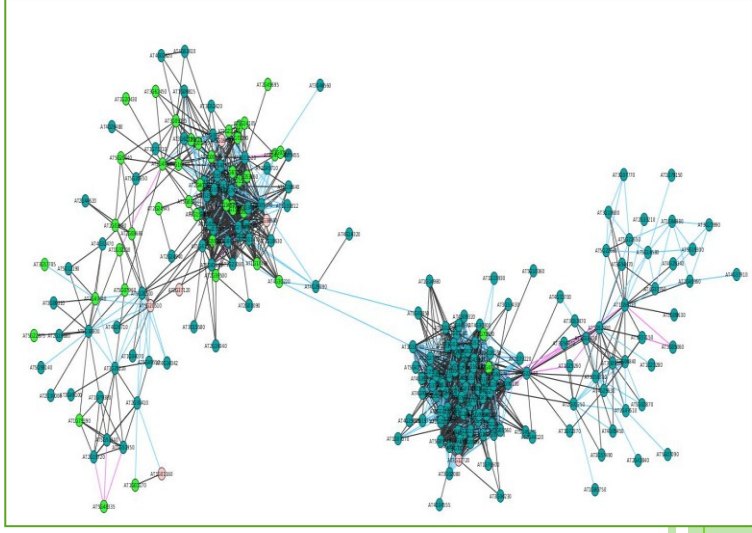
Calculation of occurrence number in a same cluster for each gene pair based on the 20 stress coexpression analyses



Resampling shows that a pair observed more than 4 times in a same cluster is biologically significant

First interpretation with gene pairs conserved in at least 7 stresses

FIRST RESULTS

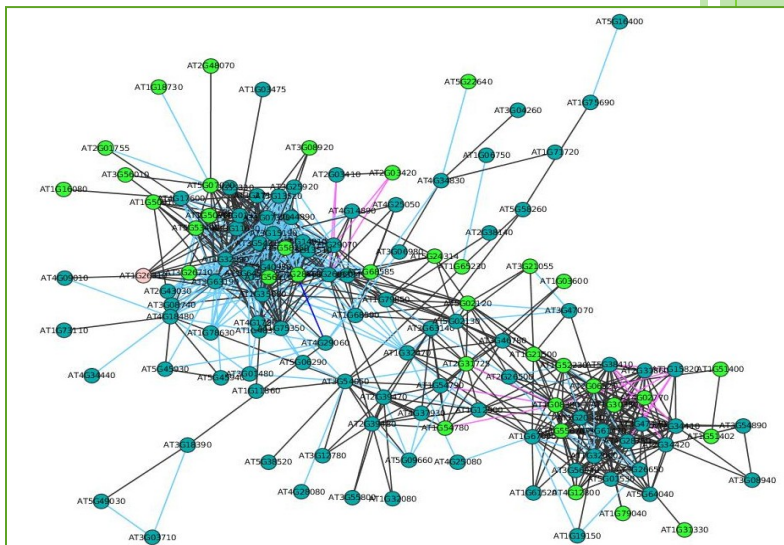


- 1094 genes with 5222 interactions

- 221 Orphans identified in the gene pairs

- 34 connected components more homogeneous than coexpression cluster in term of biological information
  - Some genes inside a component are known to be functionally related
  - More validations are requested

- Most pairs are not specific to a biotic or abiotic stress. So there exists ubiquitous response to stress



Final conclusions

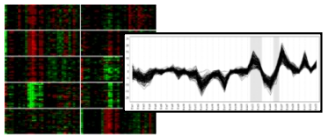
- Model-based clustering allows to better understand the data than pair-based methods
- It provides a global point of view and a way to determine functional modules
- Working with homogeneous data is really the ideal framework
- Results per stress are stored in a database which will be public in July
- Such work provides a general view of the genome activity and should help the biologists to precise the biological questions.
- Investigation around the coregulation network is in progress

Acknowledgements

URGV **Bioinformatics for predictive genomics**

S. Aubourg
V. Brunaud
G. Rigaill
R. Zaag
J.-P. Tamby
C.Guichard
Z.Tariq

Statistics



G. Celeux (Orsay)
C. Maugis (Toulouse)
T. Mary-Huard (INRA)

Biology

E. Delannoy

H. Hirt , N. Frei

And thank you for your attention !

Funding :