Identification de modules fonctionnels par l'analyse topologique d'un réseau de corégulation

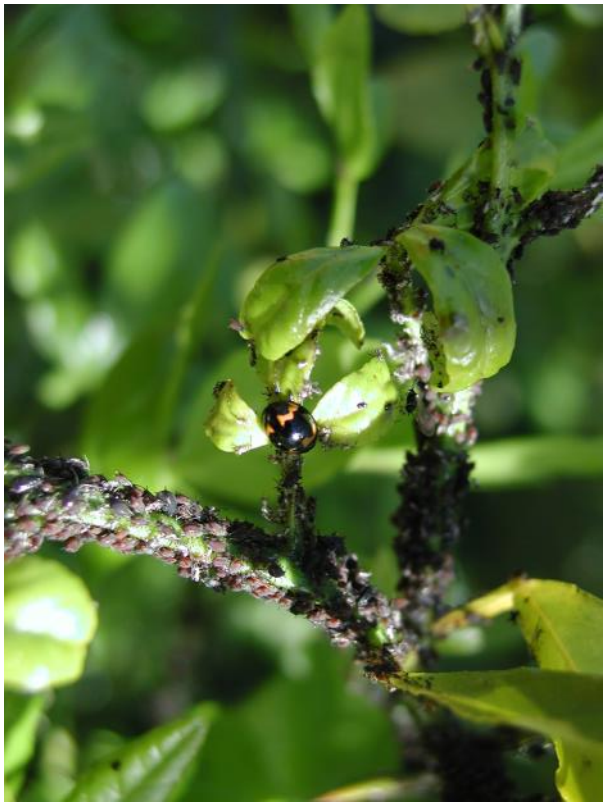
Etienne Delannoy
Pierre Latouche

Netbio, 2019

Biological context

Multiple biotic and abiotic stresses
impacting plant growth

Coordinated response to
stresses in general ?



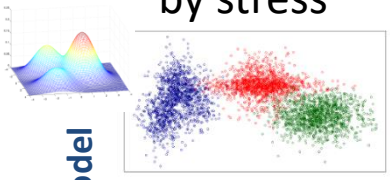
Numerous single stress transcriptomic data sets available
→ Stress gene co-expression network

Coexpression analyses of 18 stress responses

CATdb: 387 comparisons in 18 stress categories: 9 biotic and 9 abiotic

http://tools.ips2.u-psud.fr/CATdb

Matrix
{ genes x log-ratios}
by stress



Gaussian Mixture Model

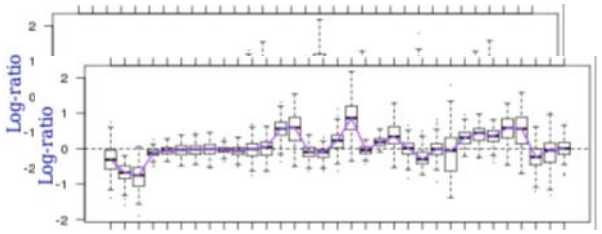
GEM2Net

http://tools.ips2.u-psud.fr//GEM2NET/

Mathematical Criterion to select the cluster number (BIC)

Classification rule based on conditional probabilities

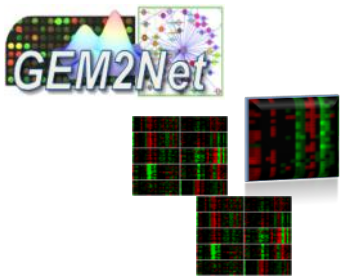
~700 Clusters of coexpressed genes



Stress category	Sample_nb	Gene_nb	Cluster_nb
Drought	17	8167	34
Gamma ray	25	5419	32
Heavy metals	45	10533	57
Nitrogen	46	13807	60
Oxidative stress	16	10027	52
Salt	15	5786	30
Temperature	45	11199	34
UV	7	7903	37
Other abiotic	8	3944	24
Fungi	21	9705	51
Biotrophic bacteria	40	11817	56
Necrotrophic bacteria	26	11030	50
Nematodes	10	7487	29
Oomycetes	14	5591	31
Rhodococcus	7	1965	13
Stifenia	6	1565	17
Virus	33	11685	54
Other biotic	6	3803	20

Zaag R *et al.,* Nucleic Acids Res, 2015

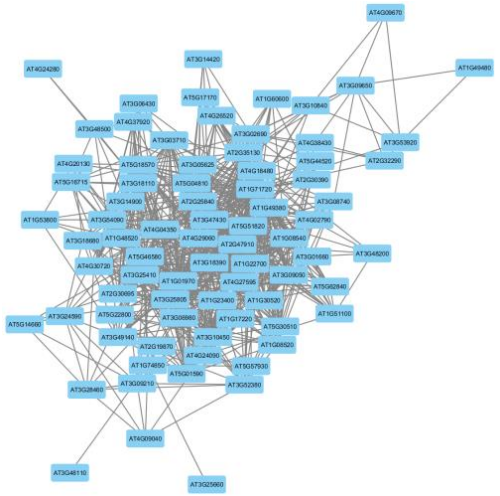
From stress coexpression clusters to stress coregulation gene network



Coexpression clusters for each category of stress

Integration

Occurrence of pairs of coexpressed genes conserved in several stresses among the 18 considered stress categories



Coexpression network

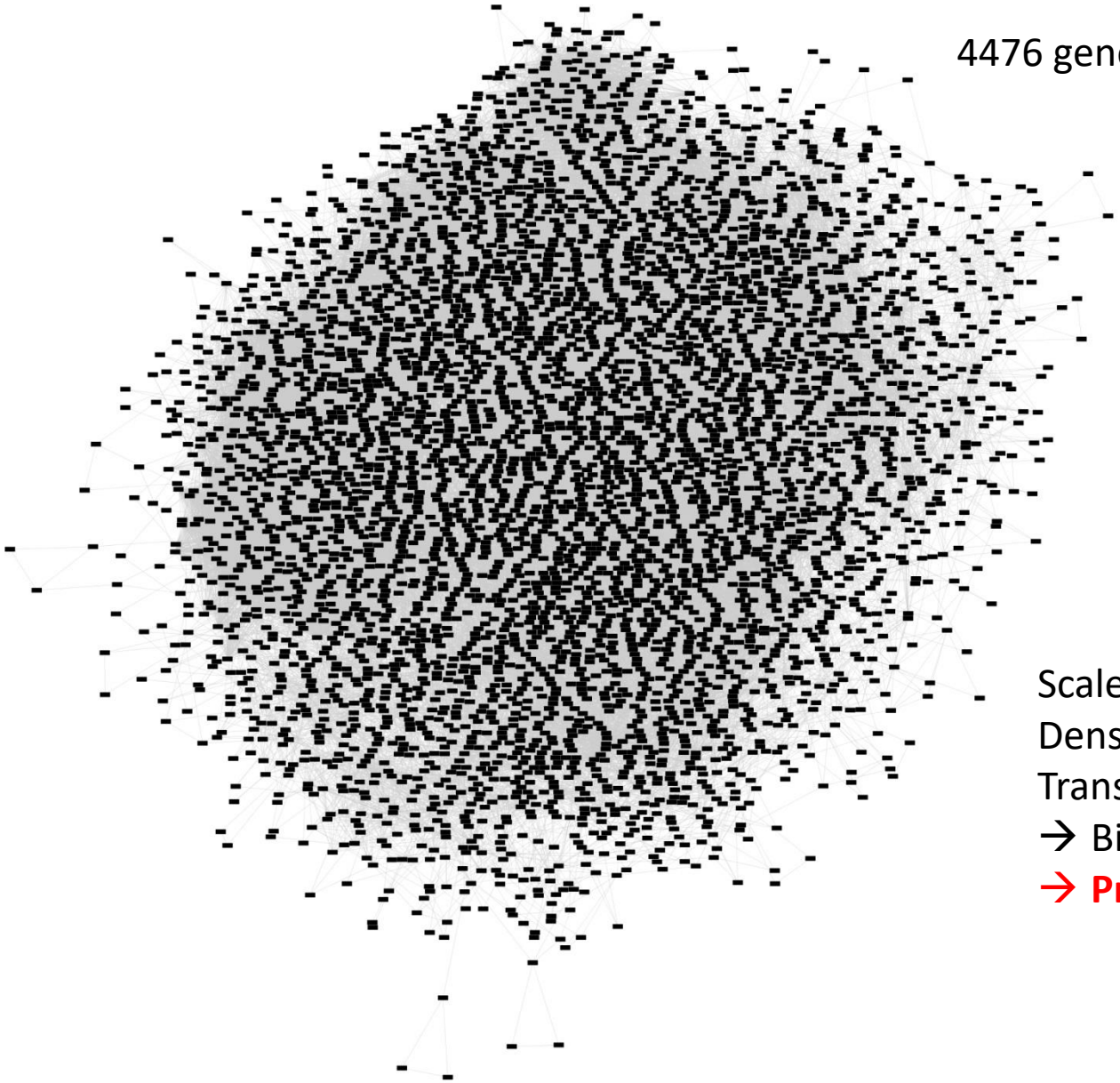
Filters

1) Compared with random networks, only edges providing a FDR<1% were kept

2) Only genes involved in triangles were considered as co-regulated

Coregulation network

Arabidopsis stress co-regulation network



4476 genes and 56487 co-regulation links

86% of the co-regulation links are supported by both biotic and abiotic stresses

Scale-free network
Density = 0.006
Transitivity= 0.54
→ Biological network
→ **Presence of gene clusters**

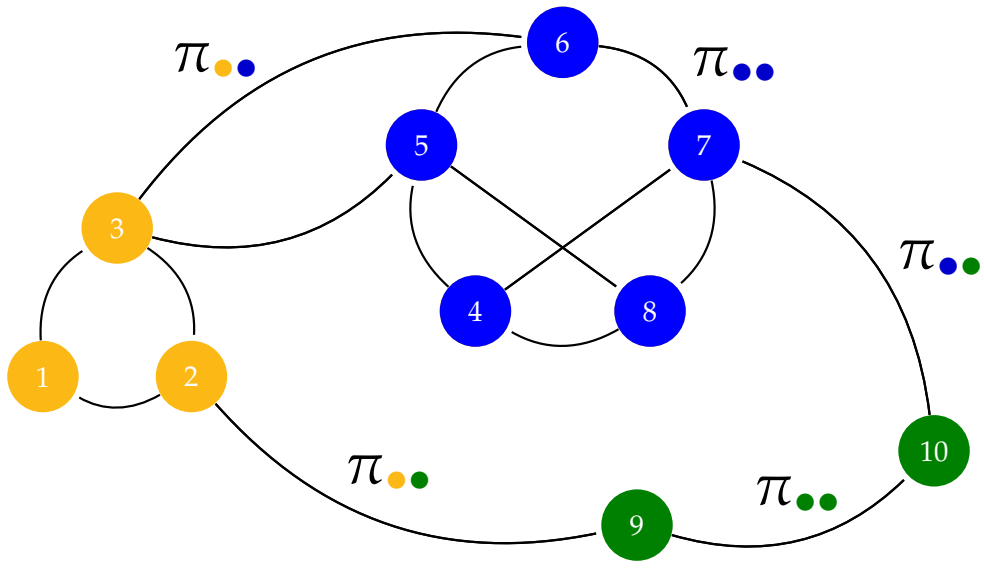
Stochastic Block Model (SBM) [WW87, NS01]

- Z_i independent hidden variables :
  - $Z_i \sim \mathcal{M}\Big(1,\ \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_K)\Big)$
  - $Z_{ik} = 1$: vertex i belongs to class k
- $X|Z$ edges drawn independently :

$$X_{ij}|\{Z_{ik}Z_{jl} = 1\} \sim \mathcal{B}(\pi_{kl})$$

- A mixture model for graphs :

$$X_{ij} \sim \sum_{k=1}^{K}\sum_{l=1}^{K} \alpha_k \alpha_l \mathcal{B}(\pi_{kl})$$


$\pi_{\bullet\bullet}$
$\pi_{\bullet\bullet}$
$\pi_{\bullet\bullet}$
$\pi_{\bullet\bullet}$
$\pi_{\bullet\bullet}$
6
5
7
3
4
8
1
2
10
9

Maximum likelihood estimation

- **Log-likelihoods of the model** :
  - Observed-data : $\log p(X|\alpha, \pi) = \log \{\sum_Z p(X, Z|\alpha, \pi)\}$
    $\hookrightarrow$ K^N terms
- Expectation Maximization (EM) algorithm requires the knowledge of $p(Z|X, \alpha, \pi)$

Problem

$p(Z|X, \alpha, \pi)$ is not tractable (no conditional independence)

Variational EM

Daudin et al. [DPR08]

Maximum likelihood estimation

- **Log-likelihoods of the model** :
  - Observed-data : $\log p(X|\alpha, \pi) = \log \{\sum_Z p(X, Z|\alpha, \pi)\}$
    $\hookrightarrow$ K^N terms
- Expectation Maximization (EM) algorithm requires the knowledge of $p(Z|X, \alpha, \pi)$

Problem

$p(Z|X, \alpha, \pi)$ is not tractable (no conditional independence)

Variational EM

Daudin et al. [DPR08]

Maximum likelihood estimation

- **Log-likelihoods of the model** :
  - Observed-data : $\log p(X|\alpha, \pi) = \log \{\sum_Z p(X, Z|\alpha, \pi)\}$
    $\hookrightarrow$ K^N terms
- Expectation Maximization (EM) algorithm requires the knowledge of $p(Z|X, \alpha, \pi)$

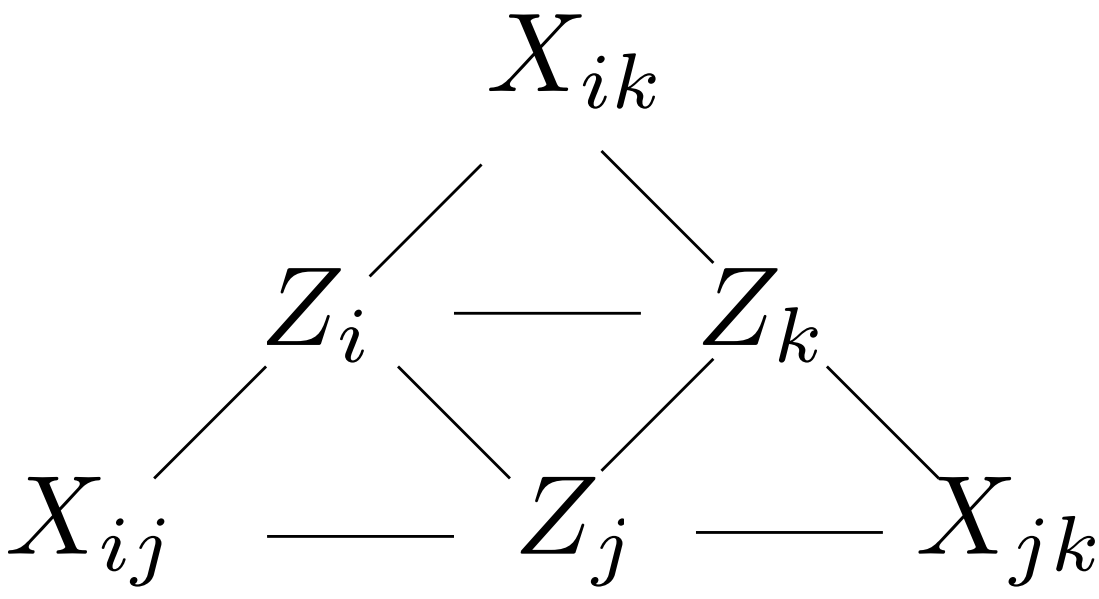
Problem

$p(Z|X, \alpha, \pi)$ is not tractable (no conditional independence)

Variational EM

Daudin et al. [DPR08]

Graphical model and moral graph



Moral graph of SBM

Model selection

Criteria

Since $\log p(X|\alpha, \pi)$ is not tractable, we *cannot* rely on :

- $AIC = \log p(X|\hat{\alpha}, \hat{\pi}) - M$
- $BIC = \log p(X|\hat{\alpha}, \hat{\pi}) - \frac{M}{2} \log \frac{N(N-1)}{2}$

ICL

Biernacki et al. [BCG00] $\hookrightarrow$ Daudin et al. [DPR08]

Variational Bayes EM $\hookrightarrow$ *ILvb*

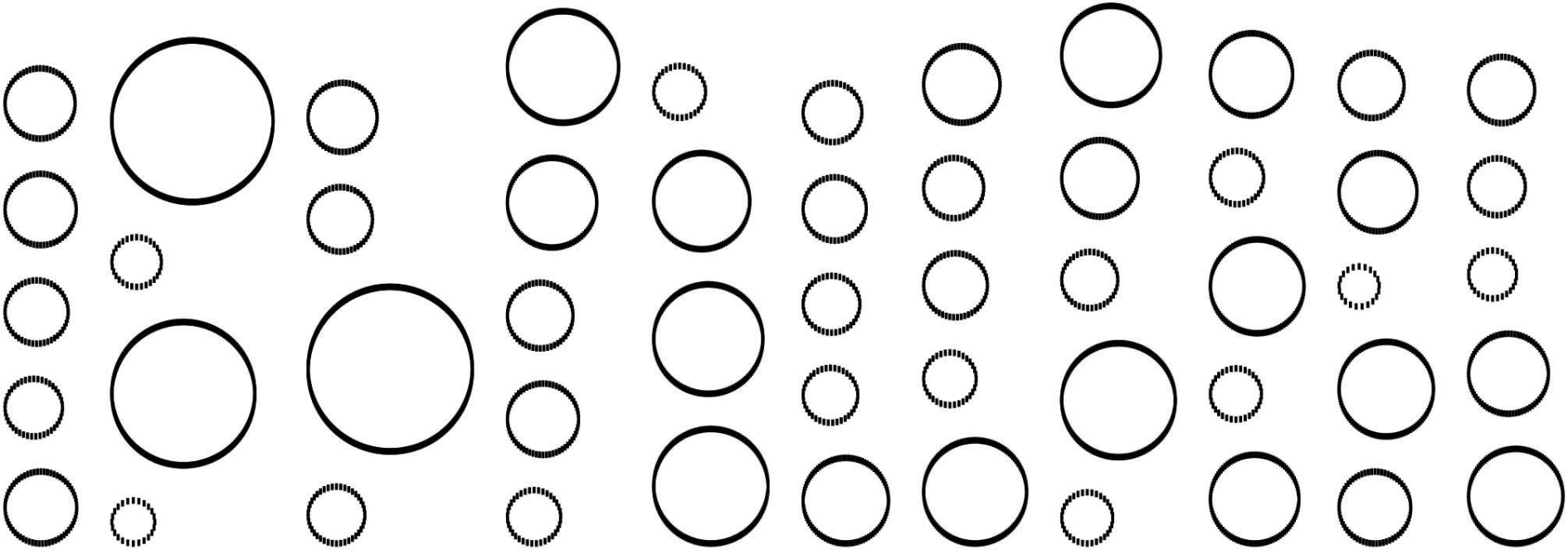
Latouche et al. [LBA12]

Others

McDaid et al. [MDMNH13]

Identification of gene communities within the network

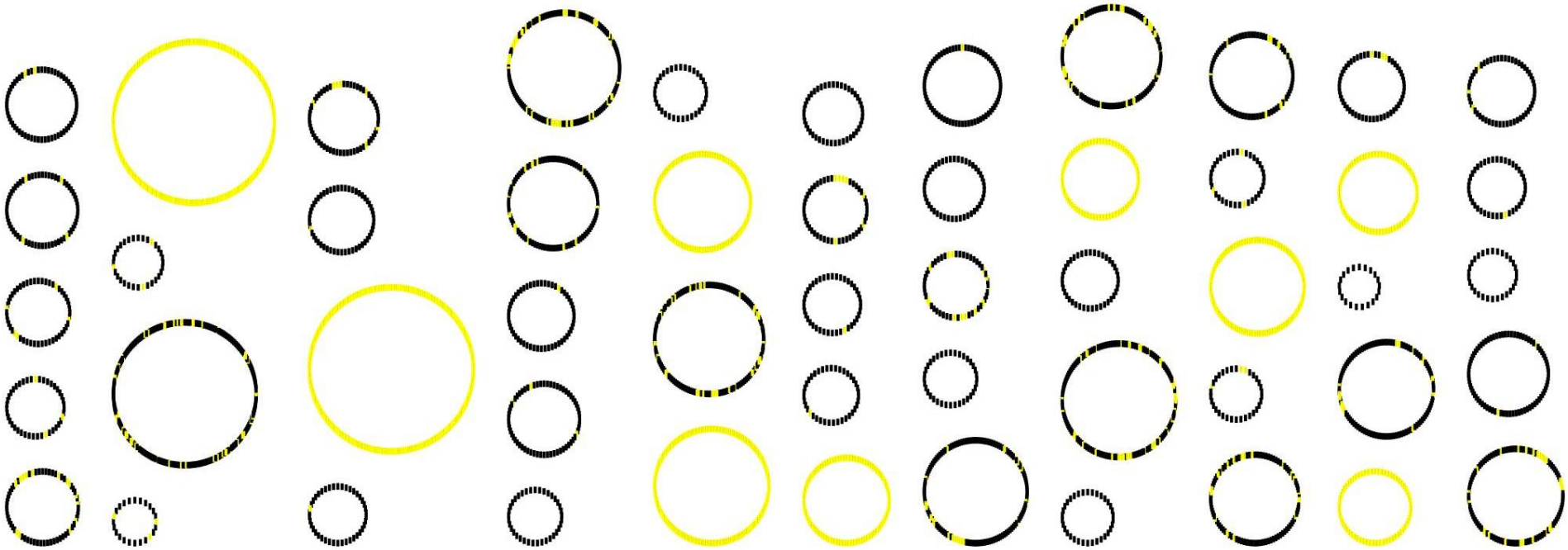
52 communities of 21 to 351 genes



Stability of the communities?

Identification of the common response to stresses

2674 genes in 43 communities describe the common response to stresses

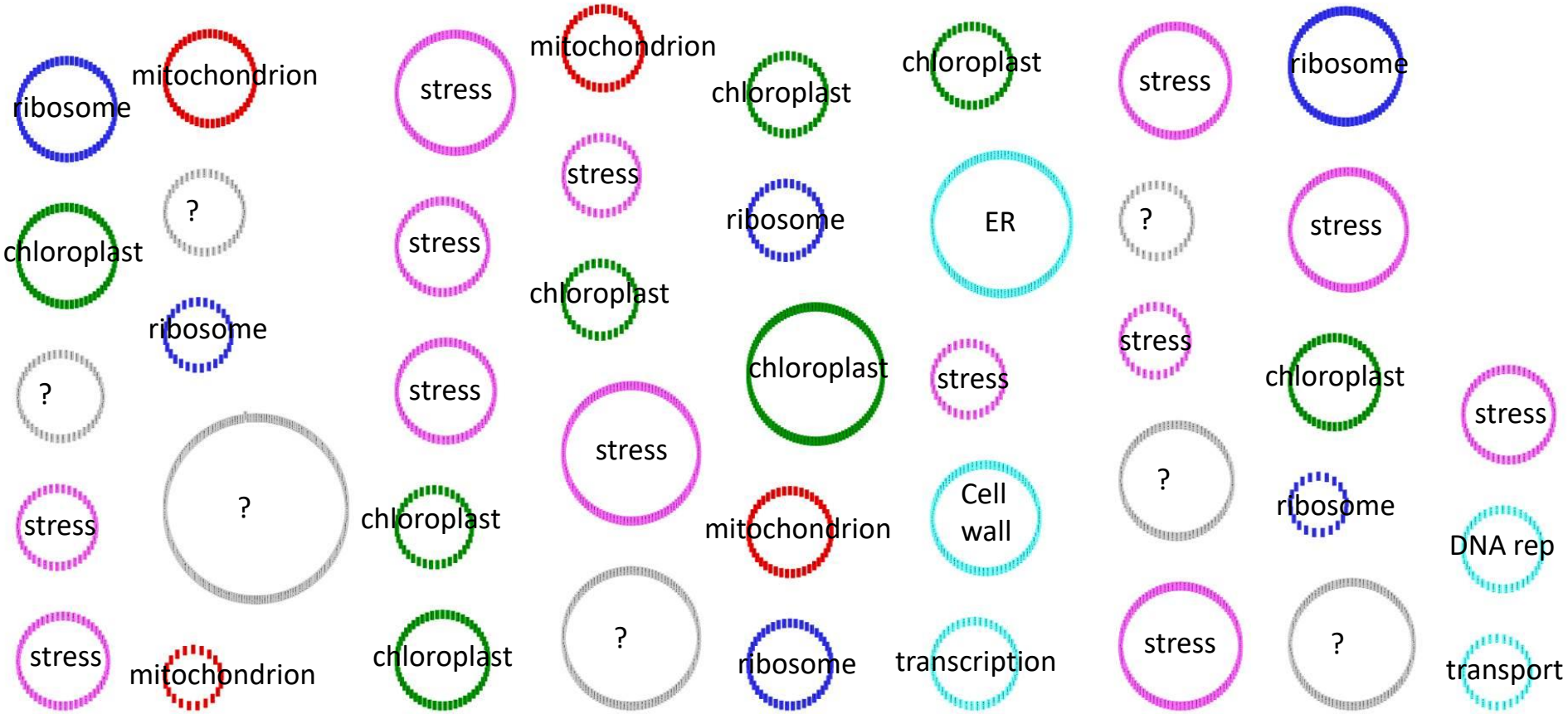


Cross-validation procedure

- For each stress category, create a network from the 17 others
- Find communities using mixture of graphs
- Comparison of these 18 results with the network built from all the categories

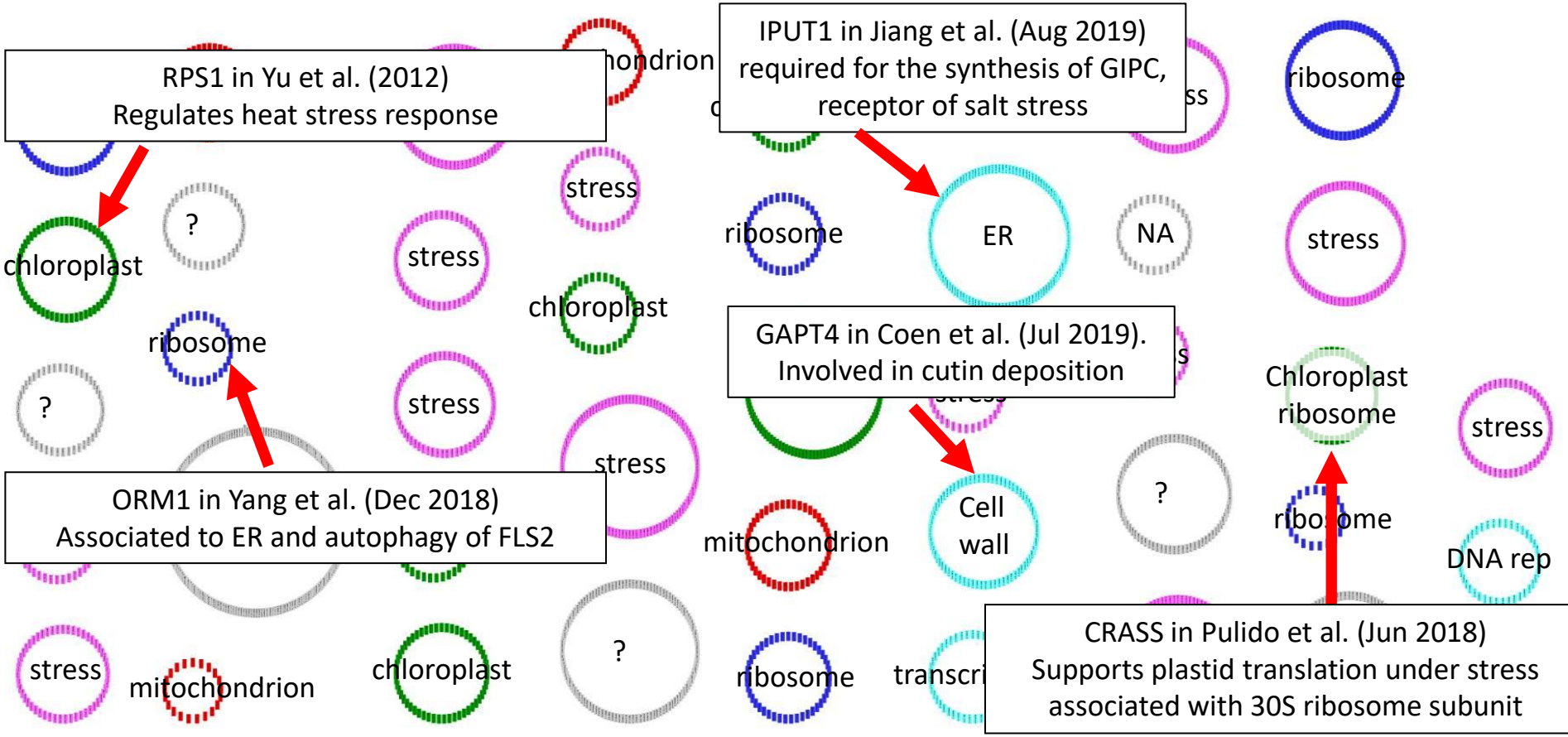
Identification of the common response to stresses

Most communities with GO enrichments



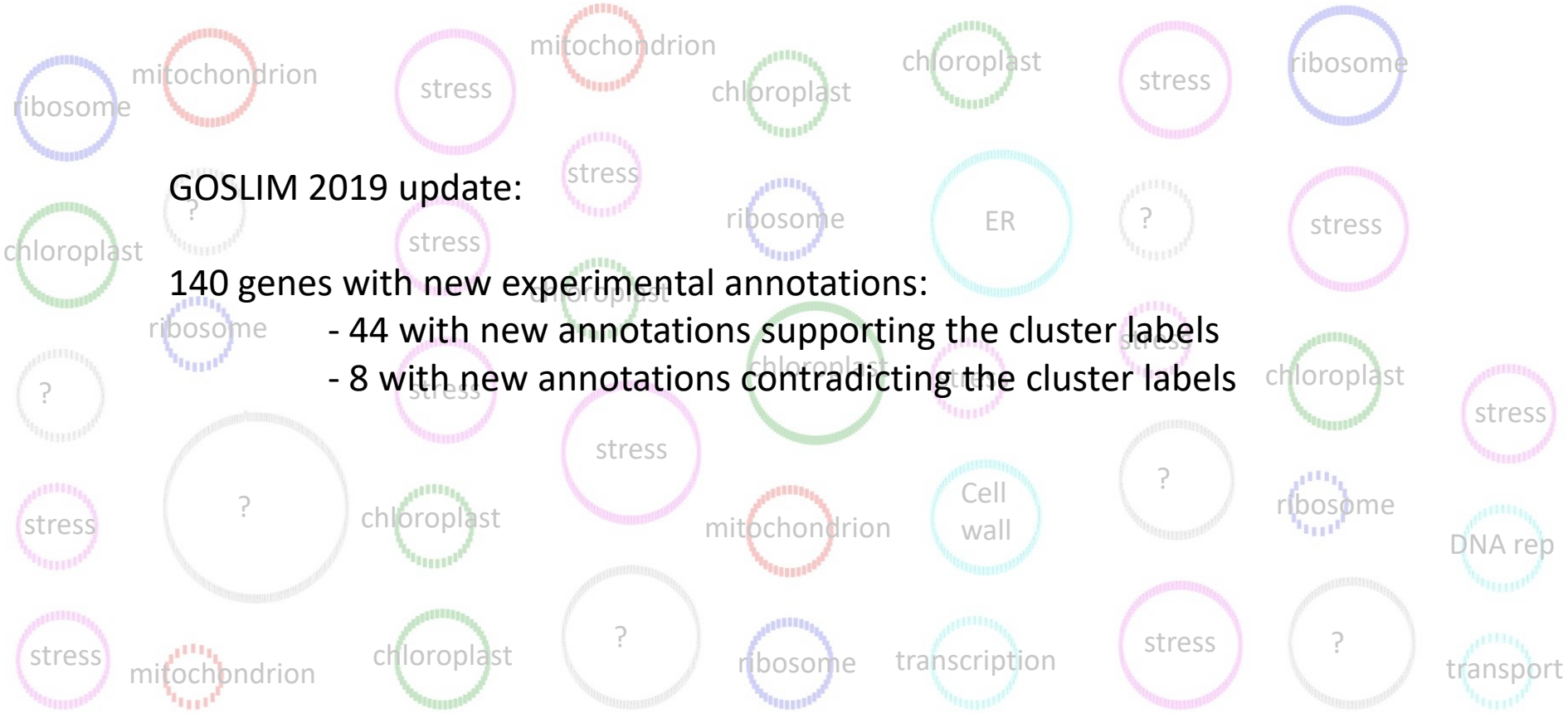
Identification of the common response to stresses

Functional validation of the communities



Identification of the common response to stresses

Functional validation of the communities

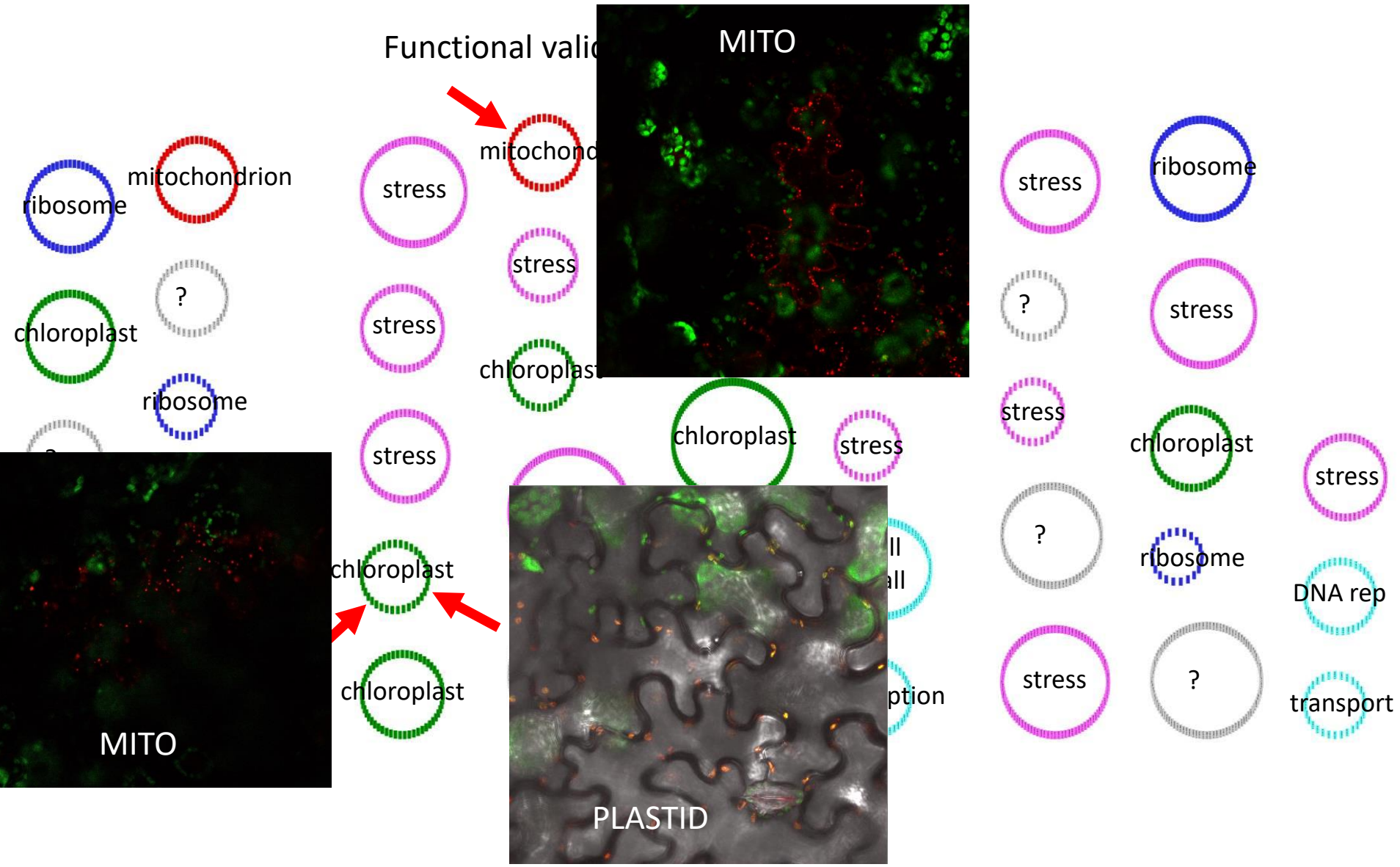


GOSLIM 2019 update:

140 genes with new experimental annotations:

- 44 with new annotations supporting the cluster labels
- 8 with new annotations contradicting the cluster labels

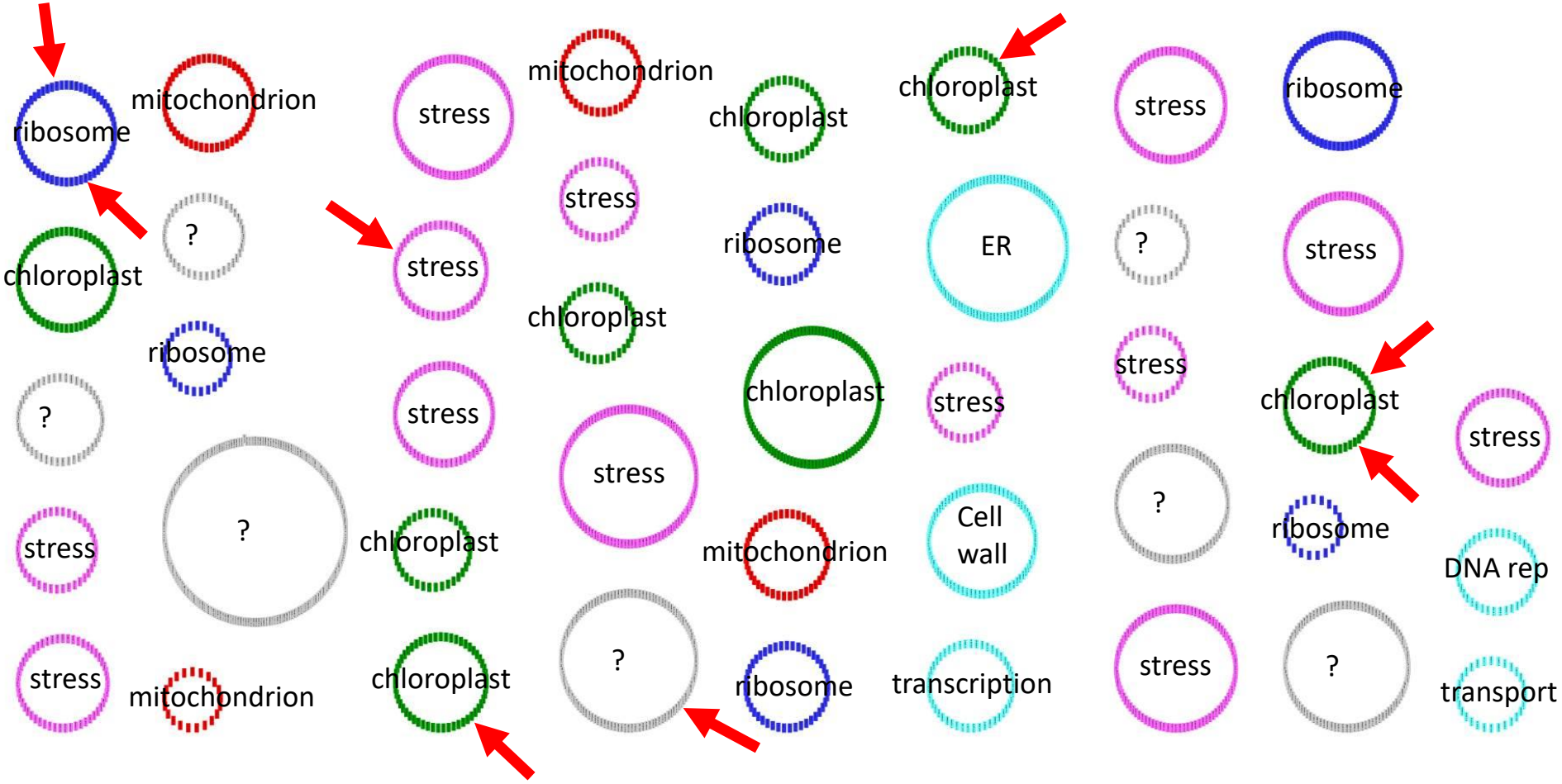
Identification of the common response to stresses



Pictures by S. Domenichini

Identification of the common response to stresses

Functional validation of the communities



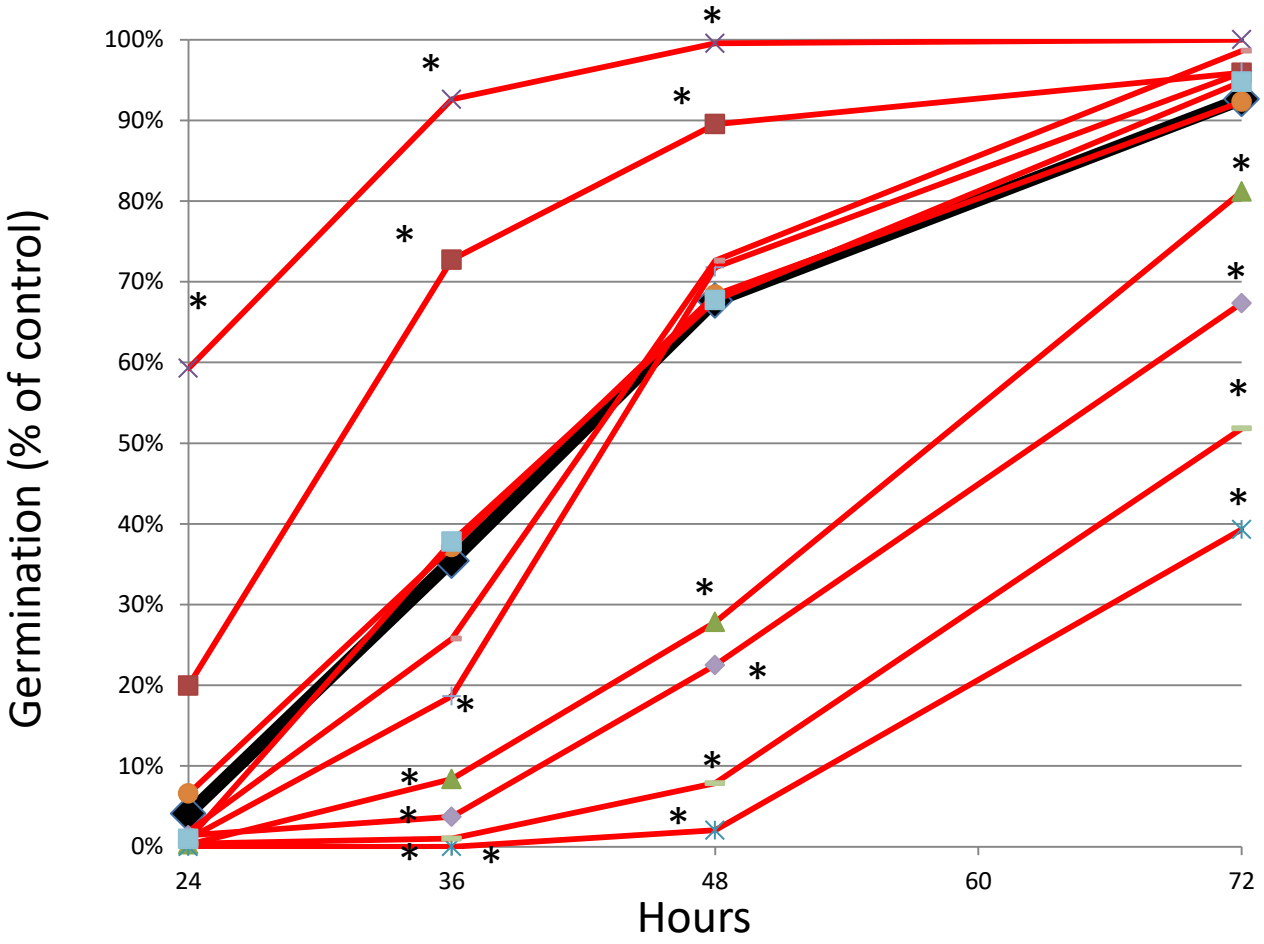
Phenotyping of T-DNA mutants of 8 genes with unknown function

Identification of the common response to stresses

Germination speed on 150mM NaCl

>3 biological replicates with 1 seed lot

Functional validation of the communities



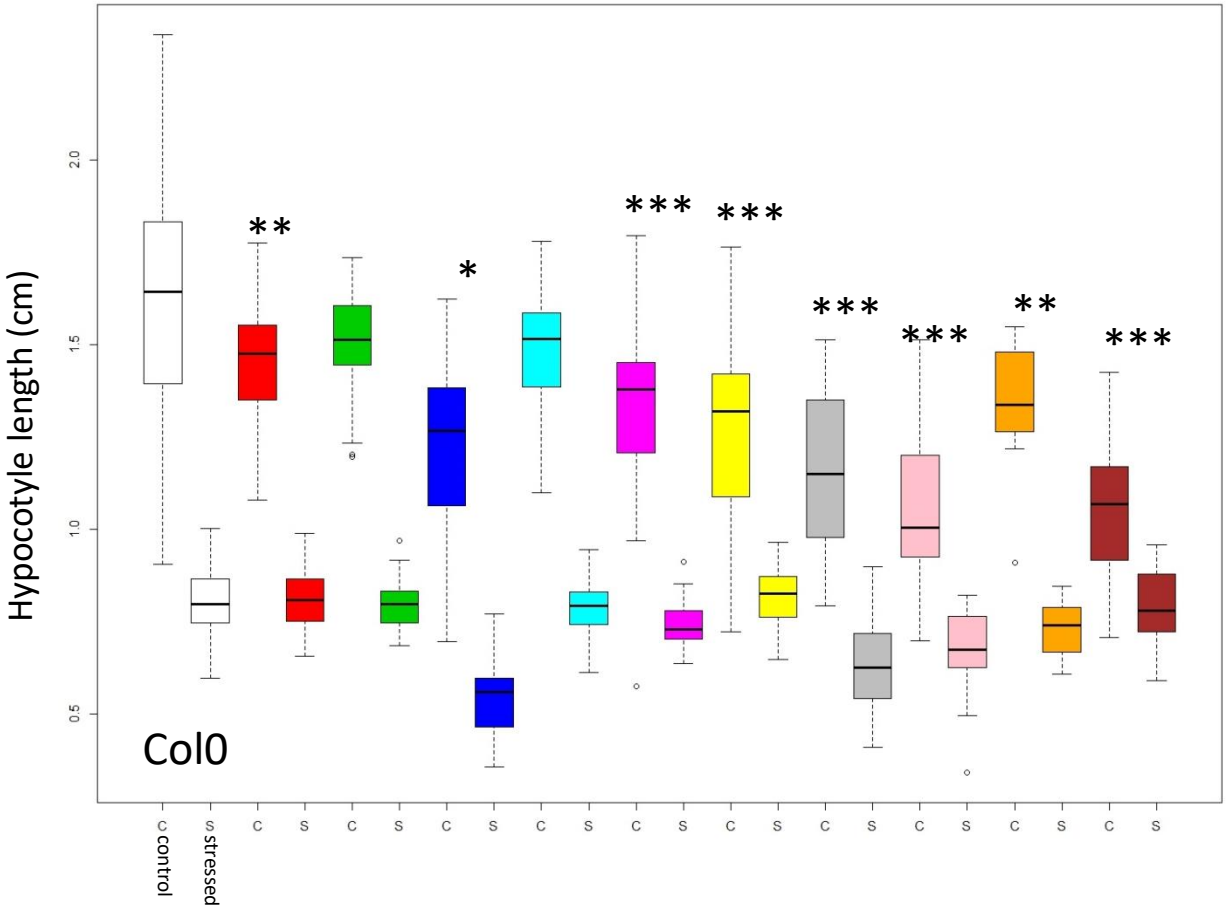
Phenotyping of T-DNA mutants of 8 genes with unknown function

Identification of the common response to stresses

Functional validation of the communities

Heat stress

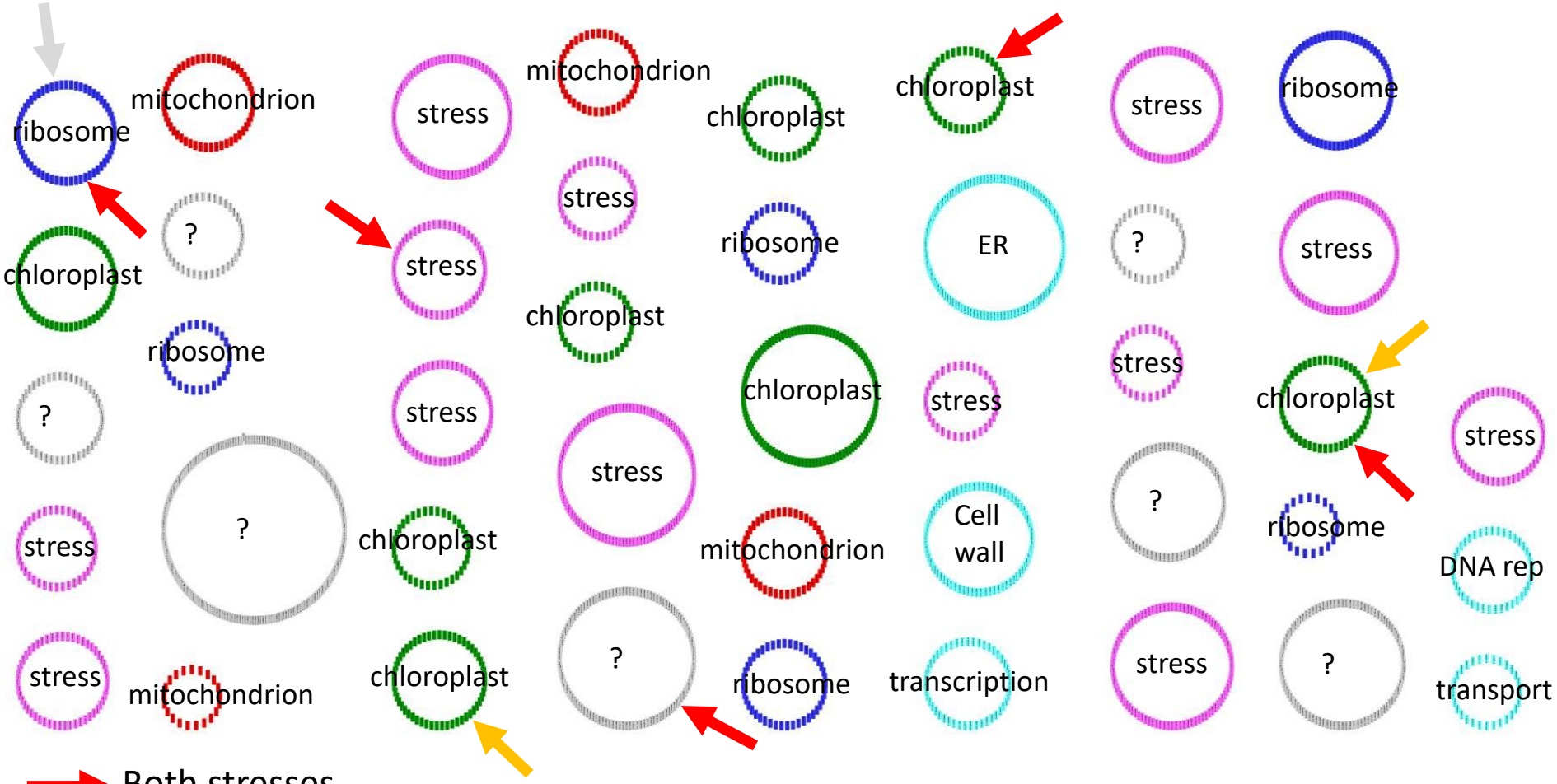
3 days-old etiolated seedlings incubated at 44°C for 1h30 before 3 days of normal conditions.
3 independent exp.



Phenotyping of T-DNA mutants of 8 genes with unknown function

Identification of the common response to stresses

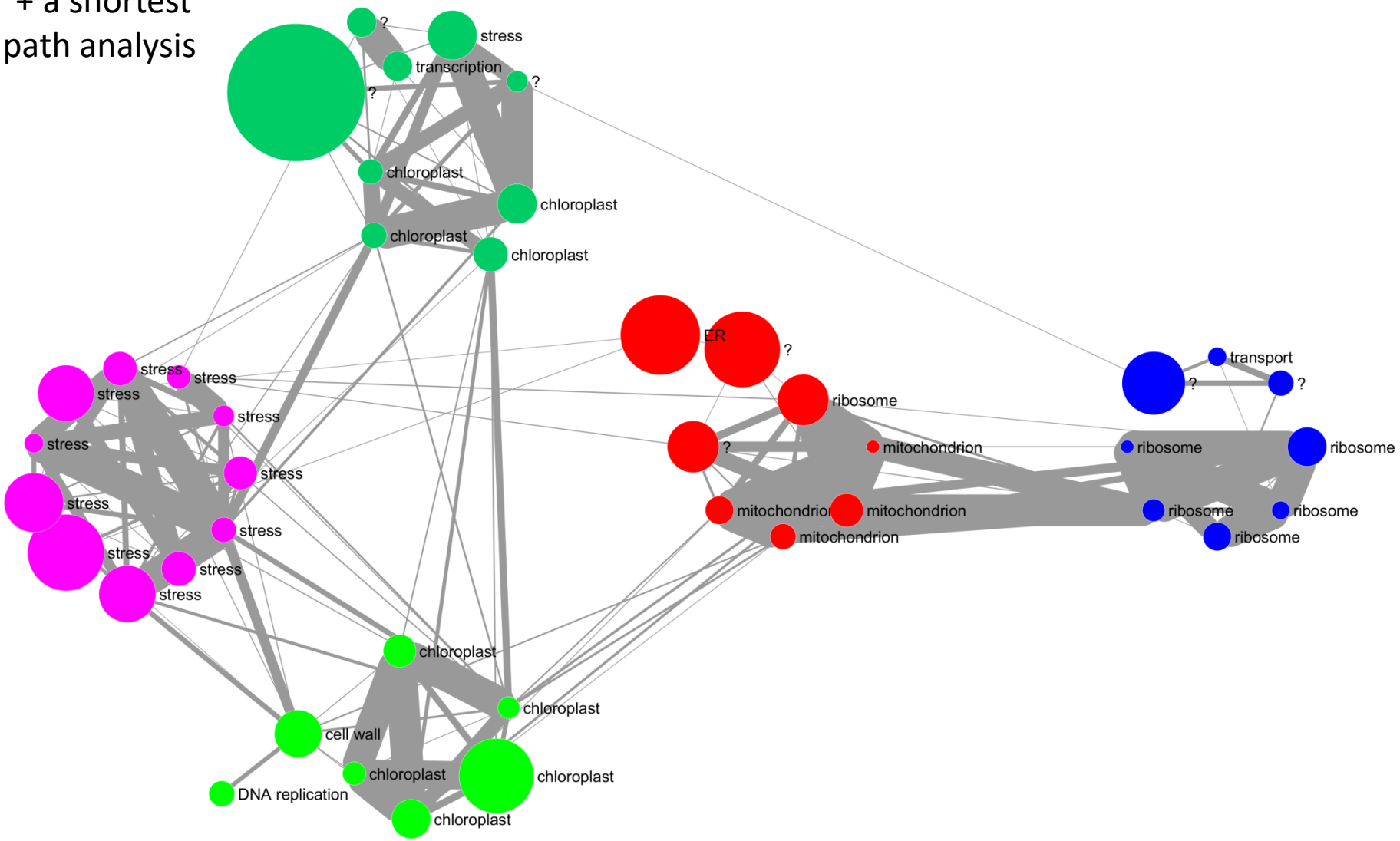
Functional validation of the communities



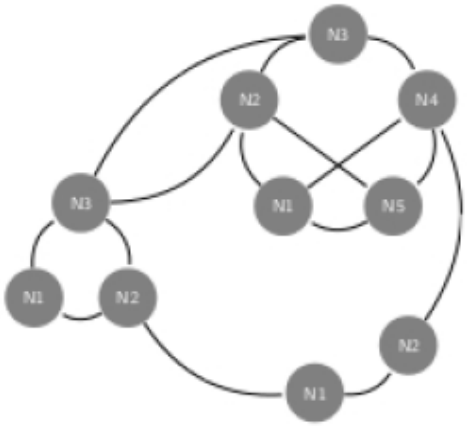
Ongoing biotic stress assays and complementation

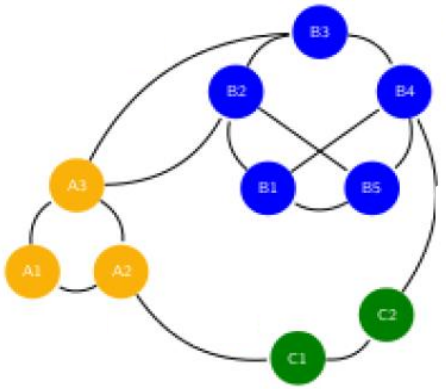
The backbone of plant stress response

+ a shortest path analysis

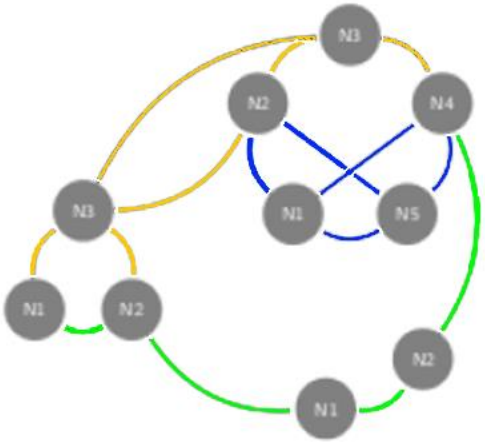


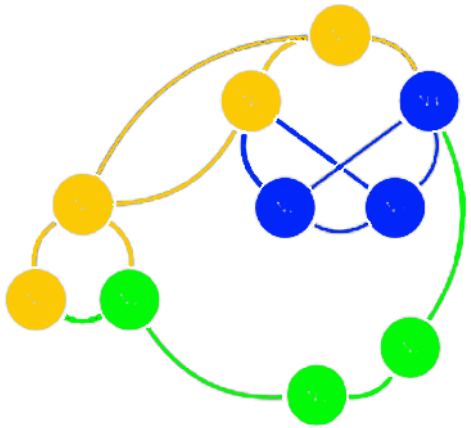
Identification of stress specific communities within the network





Information
on the edges





Introduction

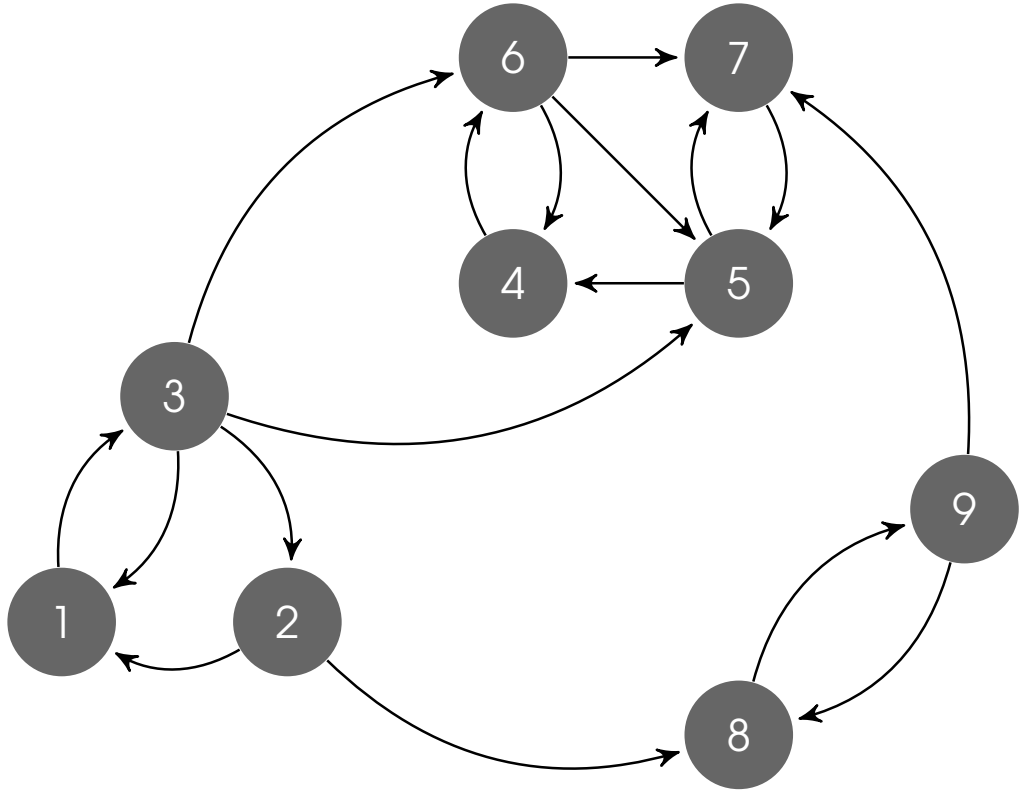


FIGURE – An (hypothetic) email network between a few individuals.

Introduction

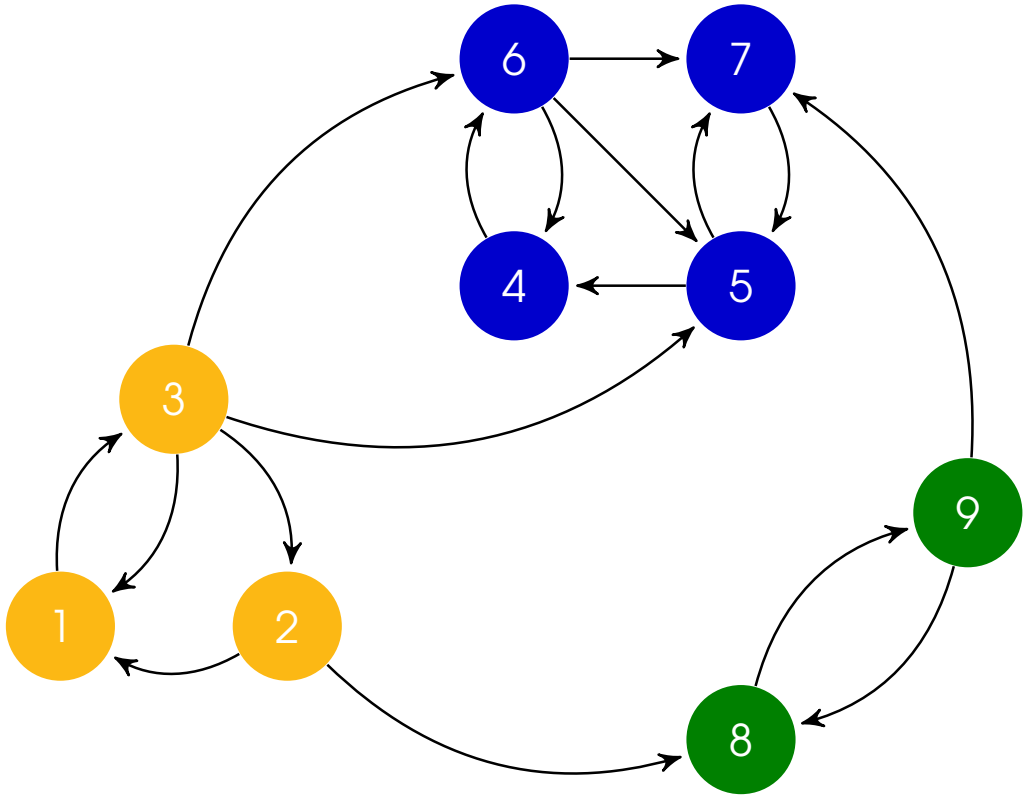


FIGURE – A typical clustering result for the (directed) binary network.

Introduction [BLZar]

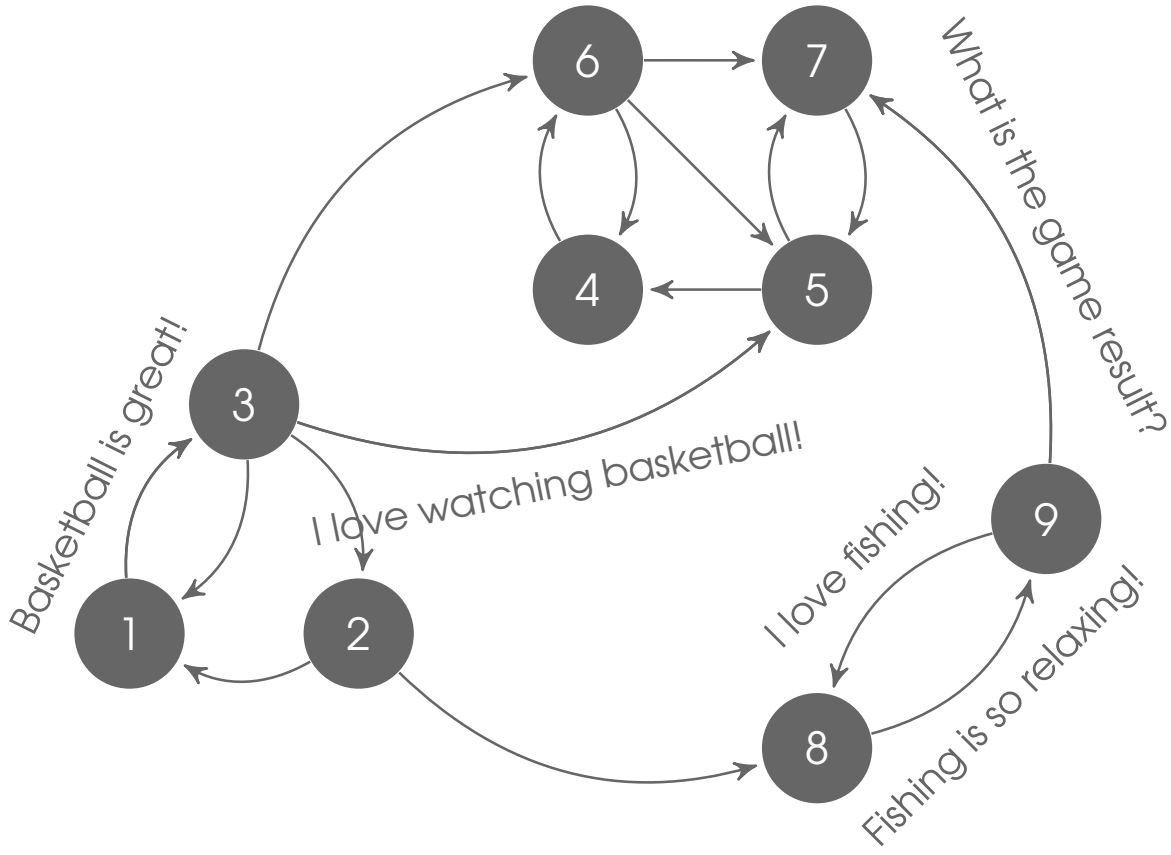


FIGURE – The (directed) network with textual edges.

Introduction [BLZar]

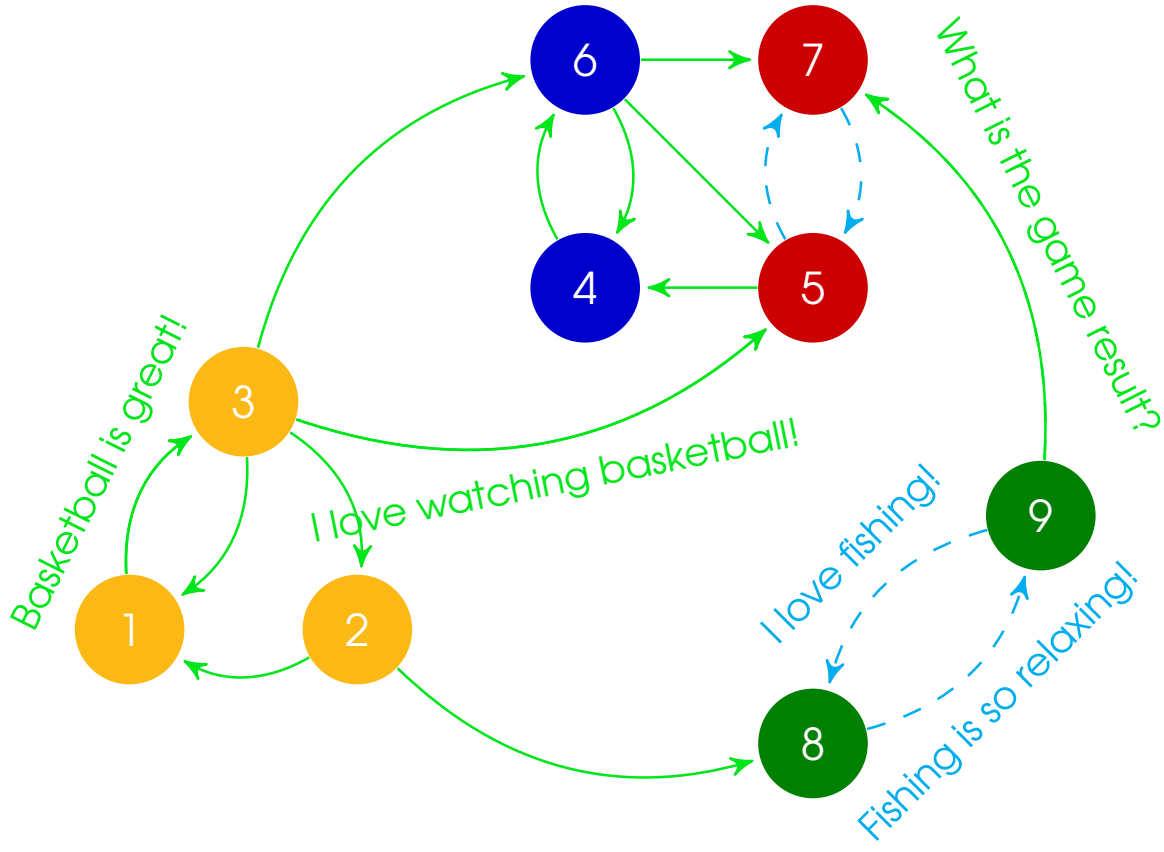


FIGURE – Expected clustering result for the (directed) network with textual edges.

Context and notations

We are interesting in clustering the nodes of a (directed) network of M vertices into Q groups :

- the network is represented by its $M \times M$ adjacency matrix A :

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between i and j} \\ 0 & \text{otherwise} \end{cases}$$

- if $A_{ij} = 1$, the textual edge is characterized by a set of D_{ij} documents :

$$W_{ij} = (W_{ij}^1, ..., W_{ij}^d, ..., W_{ij}^{D_{ij}}),$$

- each document W_{ij}^d is made of N_{ij}^d words :

$$W_{ij}^d = (W_{ij}^{d1}, ..., W_{ij}^{dn}, ..., W_{ij}^{dN_{ij}^d}).$$

Modeling of the edges

Let us assume that edges are generated according to a SBM model :

- each node i is associated with an (unobserved) group among Q according to :

$$Y_i \sim \mathcal{M}(\rho),$$

where $\rho \in [0,1]^Q$ is the vector of group proportions,

- the presence of an edge A_{ij} between i and j is drawn according to :

$$A_{ij} | Y_{iq} Y_{jr} = 1 \sim \mathcal{B}(\pi_{qr}),$$

where $\pi_{qr} \in [0,1]$ is the connection probability between clusters q and r.

Modeling of the documents

The generative model for the documents is as follows :

- each pair of clusters (q, r) is first associated to a vector of topic proportions $\theta_{qr} = (\theta_{qrk})_k$ sampled from a Dirichlet distribution :

$$\theta_{qr} \sim \text{Dir}(\alpha),$$

such that $\sum_{k=1}^{K} \theta_{qrk} = 1, \forall (q, r)$.

- the nth word W_{ij}^{dn} of documents d in W_{ij} is then associated to a latent topic vector Z_{ij}^{dn} according to :

$$Z_{ij}^{dn} | \left\{ A_{ij} Y_{iq} Y_{jr} = 1, \theta \right\} \sim \mathcal{M}\left(1, \theta_{qr}\right).$$

- then, given Z_{ij}^{dn}, the word W_{ij}^{dn} is assumed to be drawn from a multinomial distribution :

$$W_{ij}^{dn} | Z_{ij}^{dnk} = 1 \sim \mathcal{M}(1, \beta_k = (\beta_{k1}, \ldots, \beta_{kV})),$$

where V is the vocabulary size.

STBM at a glance...

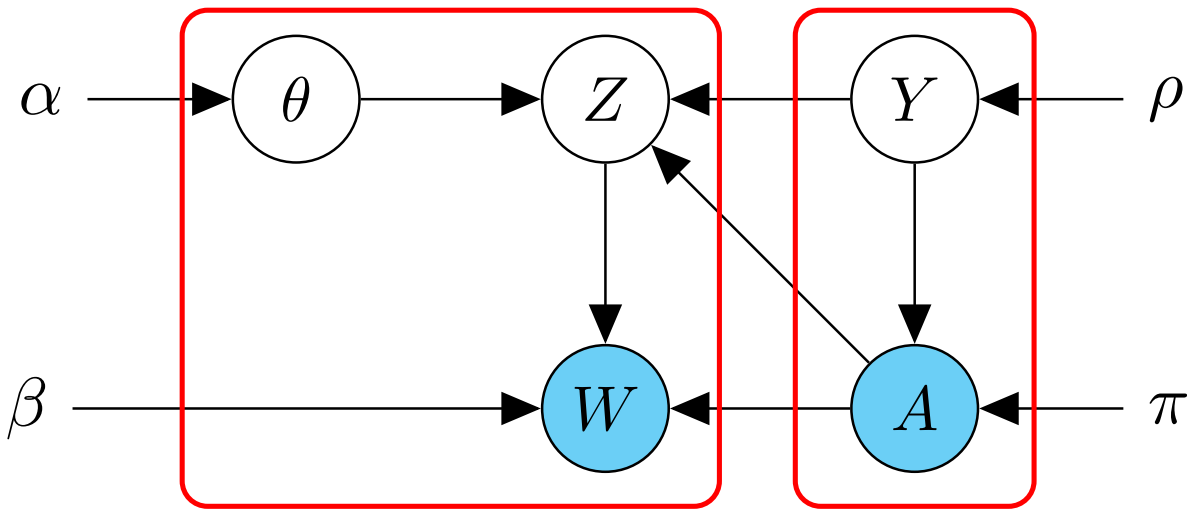


FIGURE – The stochastic topic block model.

STBM at a glance...

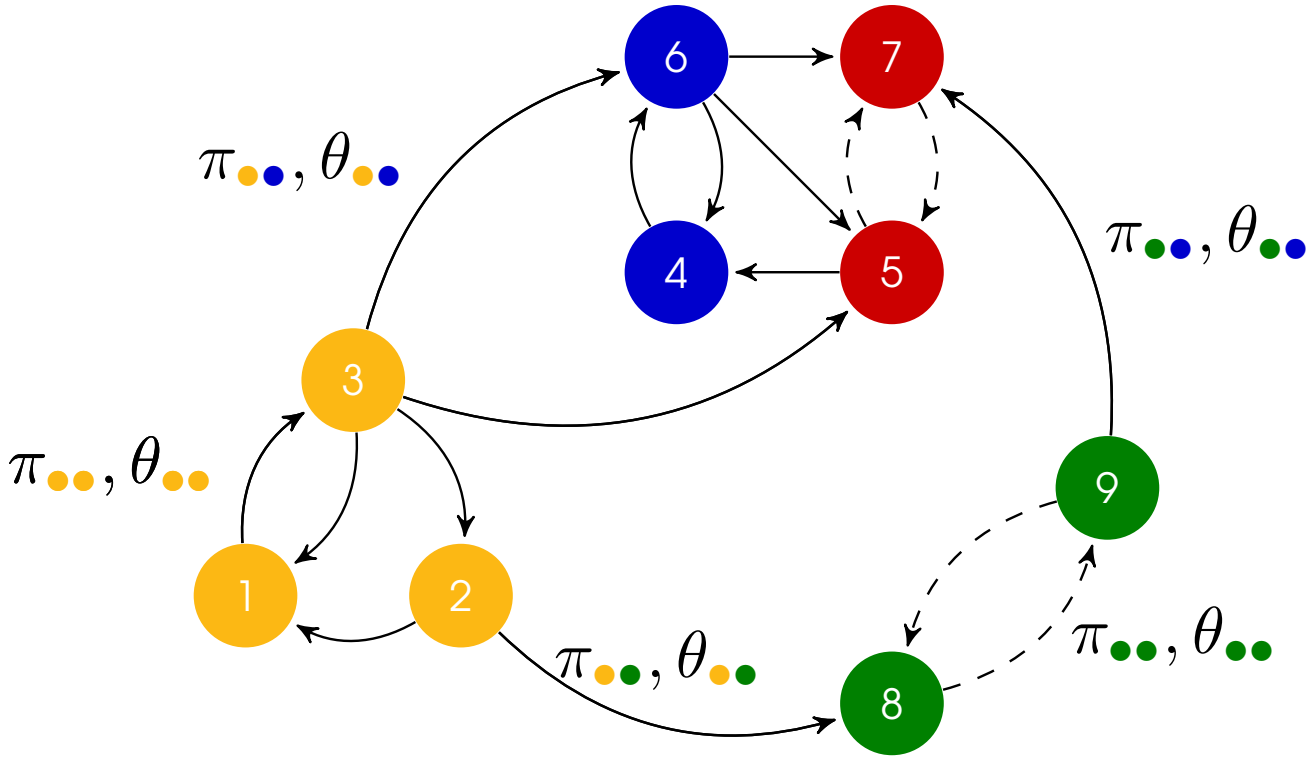


FIGURE – The stochastic topic block model.

Inference

The full joint distribution of the STBM model is given by :

$$p(A, W, Y, Z, \theta|\rho, \pi, \beta) = p(W, Z, \theta|A, Y, \beta)p(A, Y|\rho, \pi).$$

A key property of the STBM model :

- let us assume that Y is observed (groups are known),
- it is then possible to reorganize the documents $D = \sum_{i,j} D_{ij}$ documents W such that :

$$W = (\tilde{W}_{qr})_{qr} \text{ where } \tilde{W}_{qr} = \left\{ W_{ij}^d, \forall (d, i, j), Y_{iq} Y_{jr} A_{ij} = 1 \right\},$$

- since all words in $\tilde{W}_{qr}$ are associated with the same pair (q, r) of clusters, they share the same mixture distribution,
- and, simply seeing $\tilde{W}_{qr}$ as a document d, the sampling scheme then corresponds to the one of a LDA model with $D = Q^2$ documents.

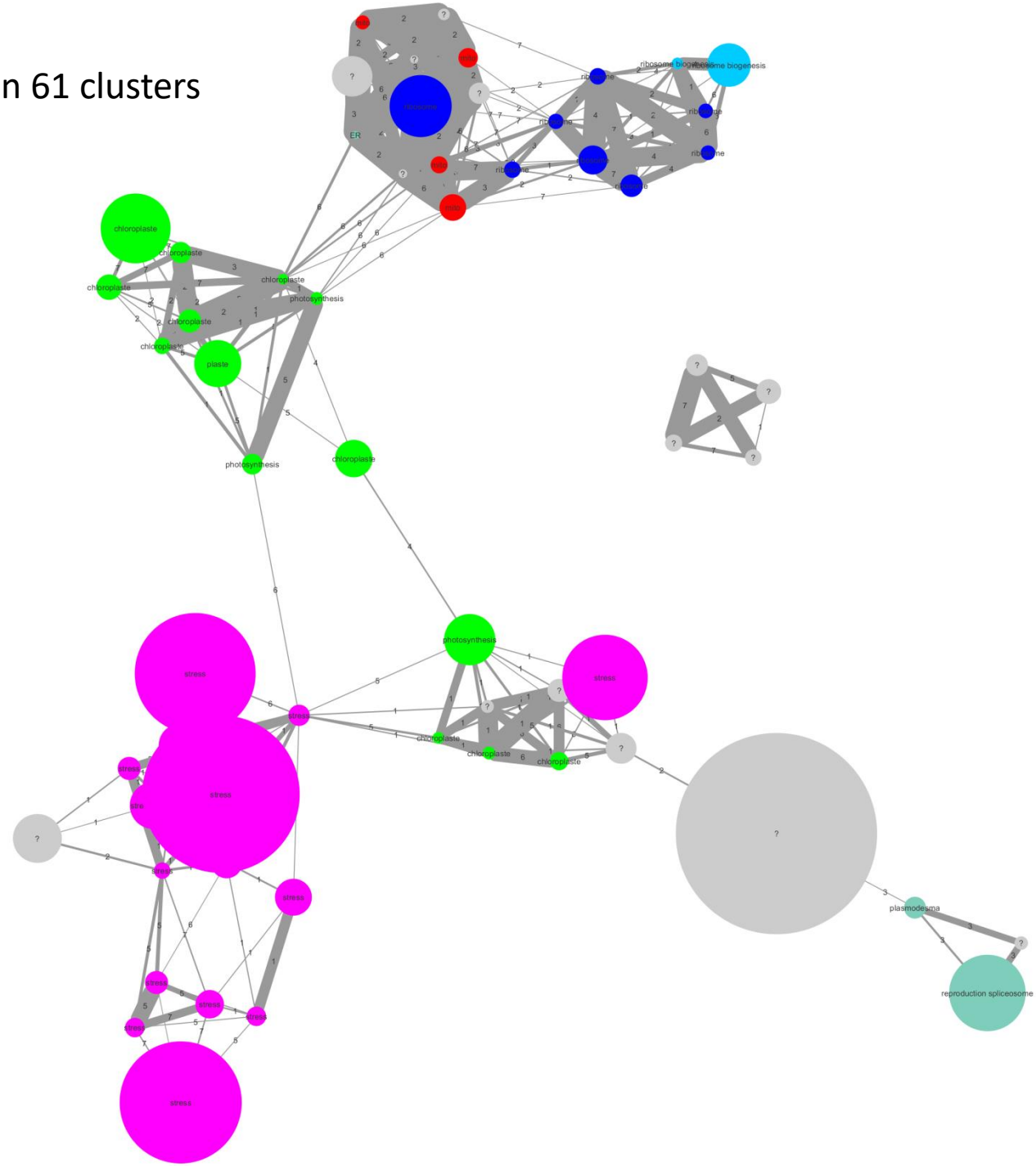
Inference

Given the above property of the model, we propose for inference to maximize the complete data log-likelihood :

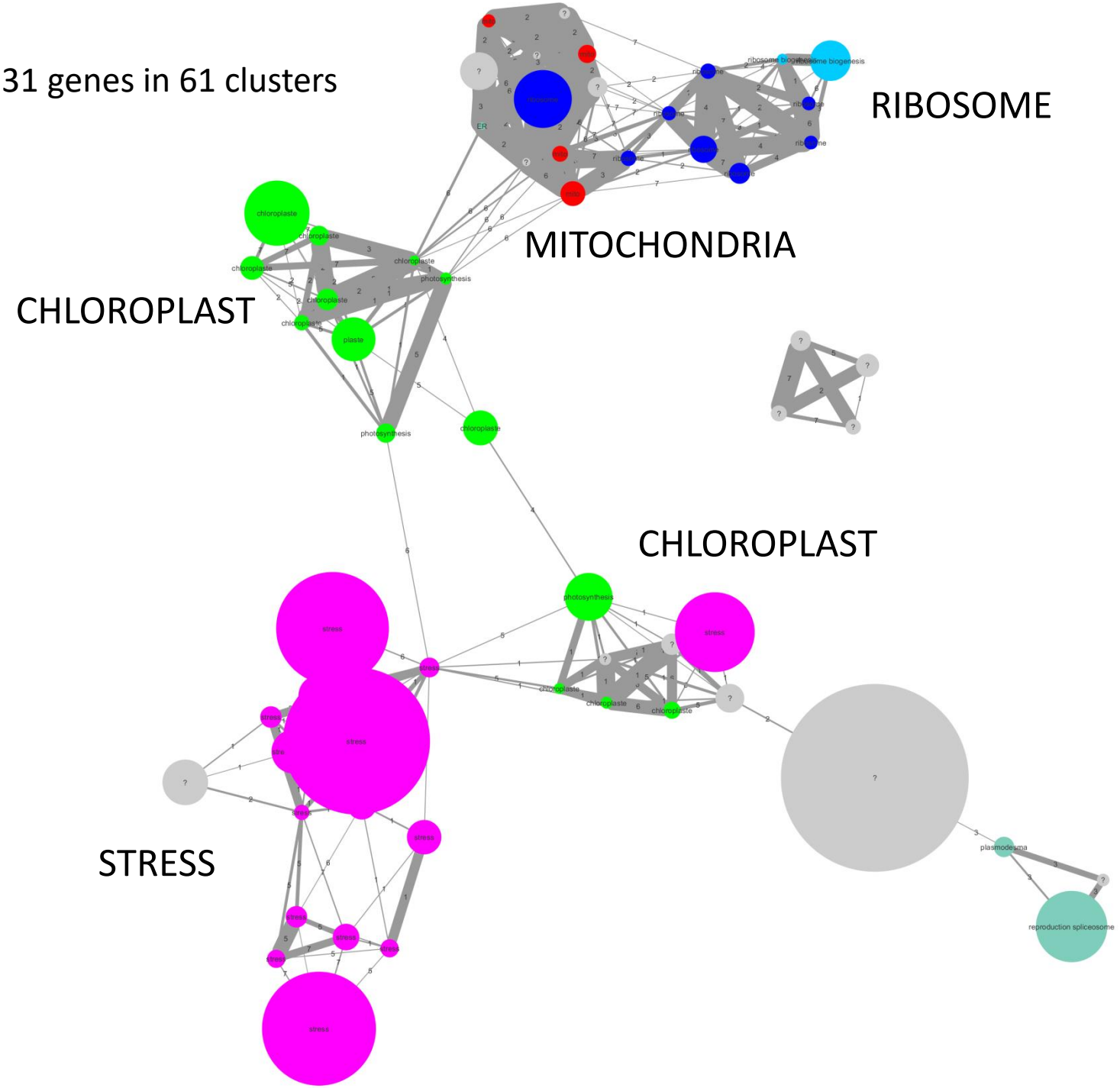
$$\log p(A, W, Y|\rho, \pi, \beta) = \log \sum_{Z} \int_{\theta} p(A, W, Y, Z, \theta|\rho, \pi, \beta) d\theta,$$

with respect to (ρ, π, β) and $Y = (Y_1, \ldots, Y_M)$.

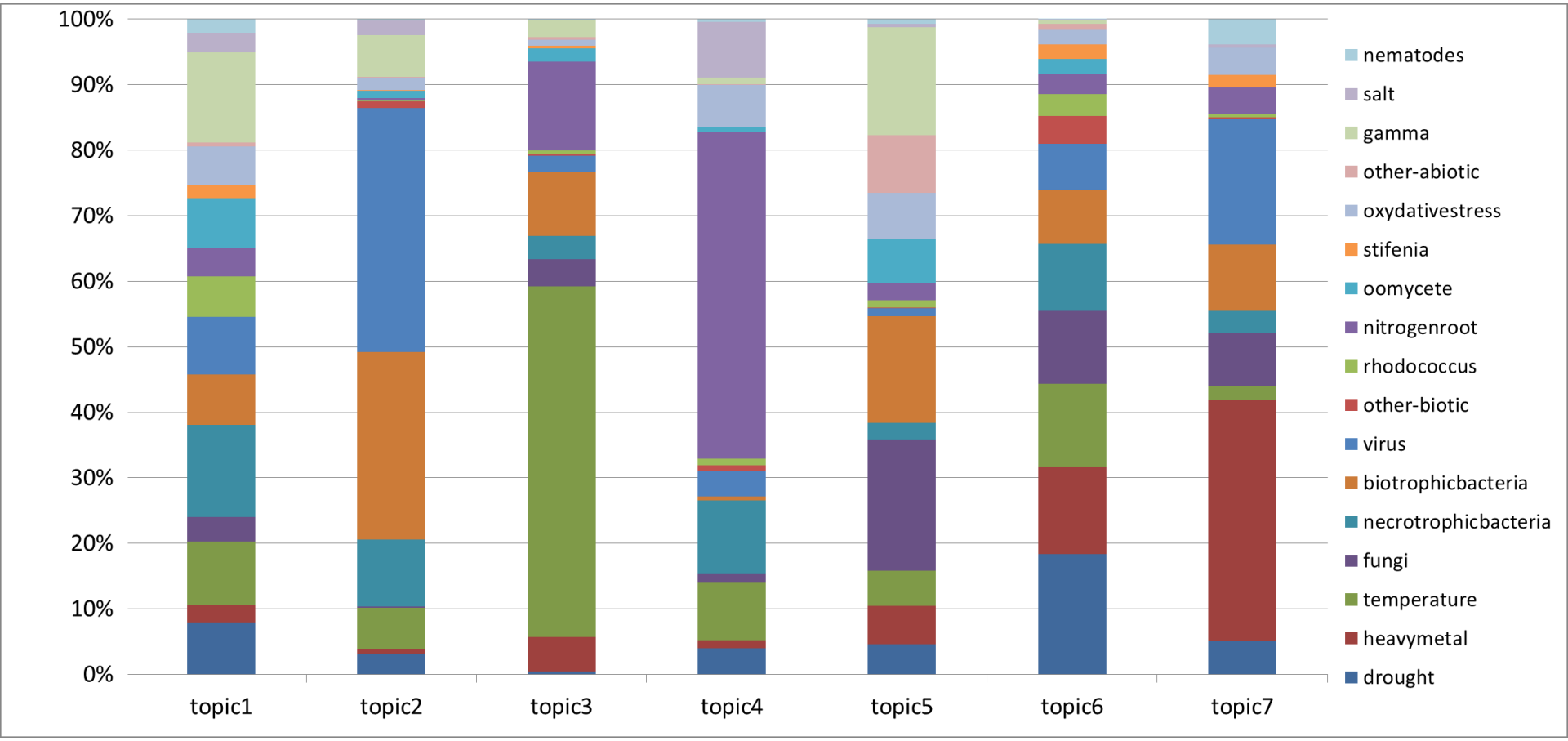
2931 genes in 61 clusters



2931 genes in 61 clusters



7 topics



All stresses

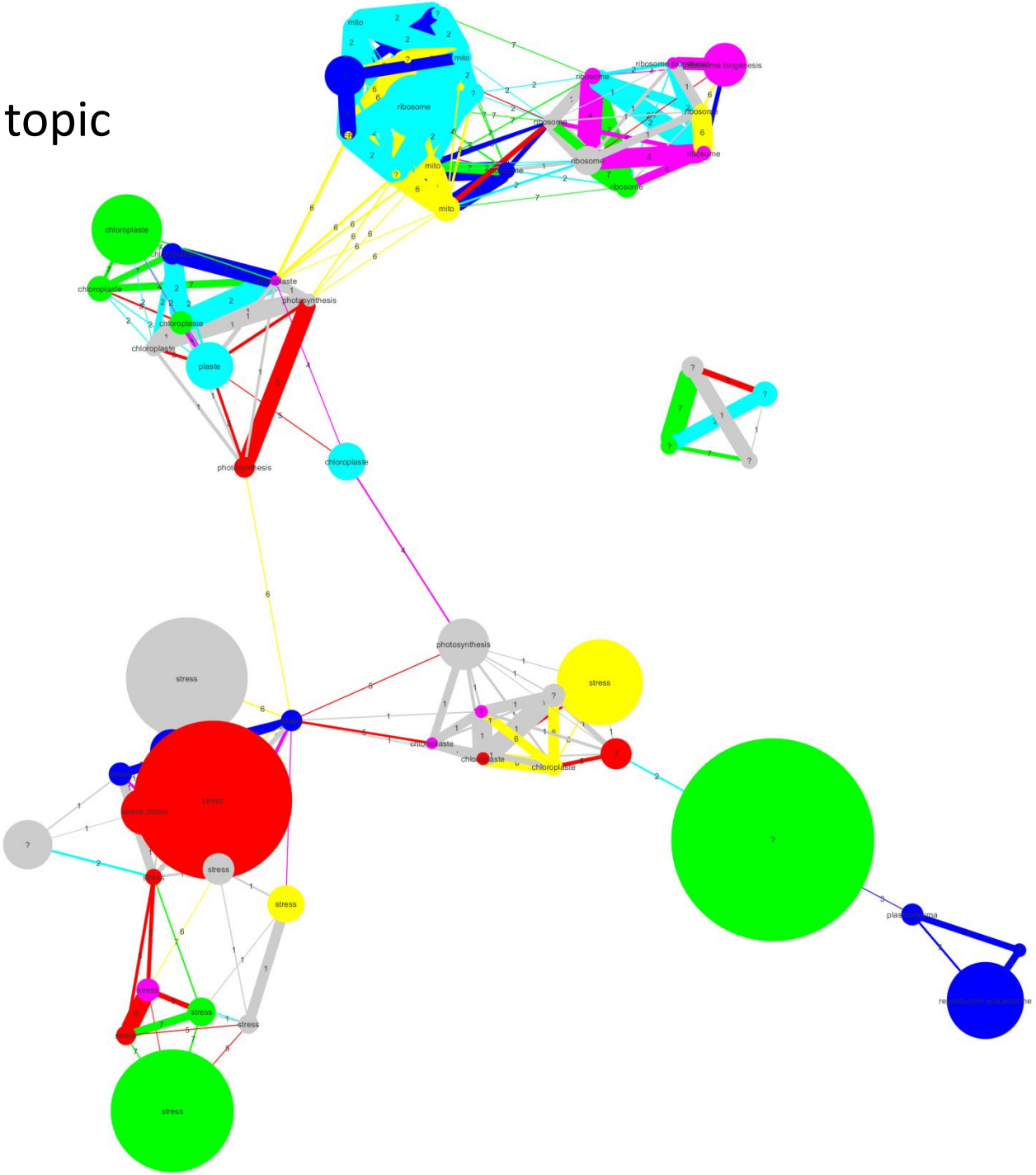
Viruses & bacteria

Temperature

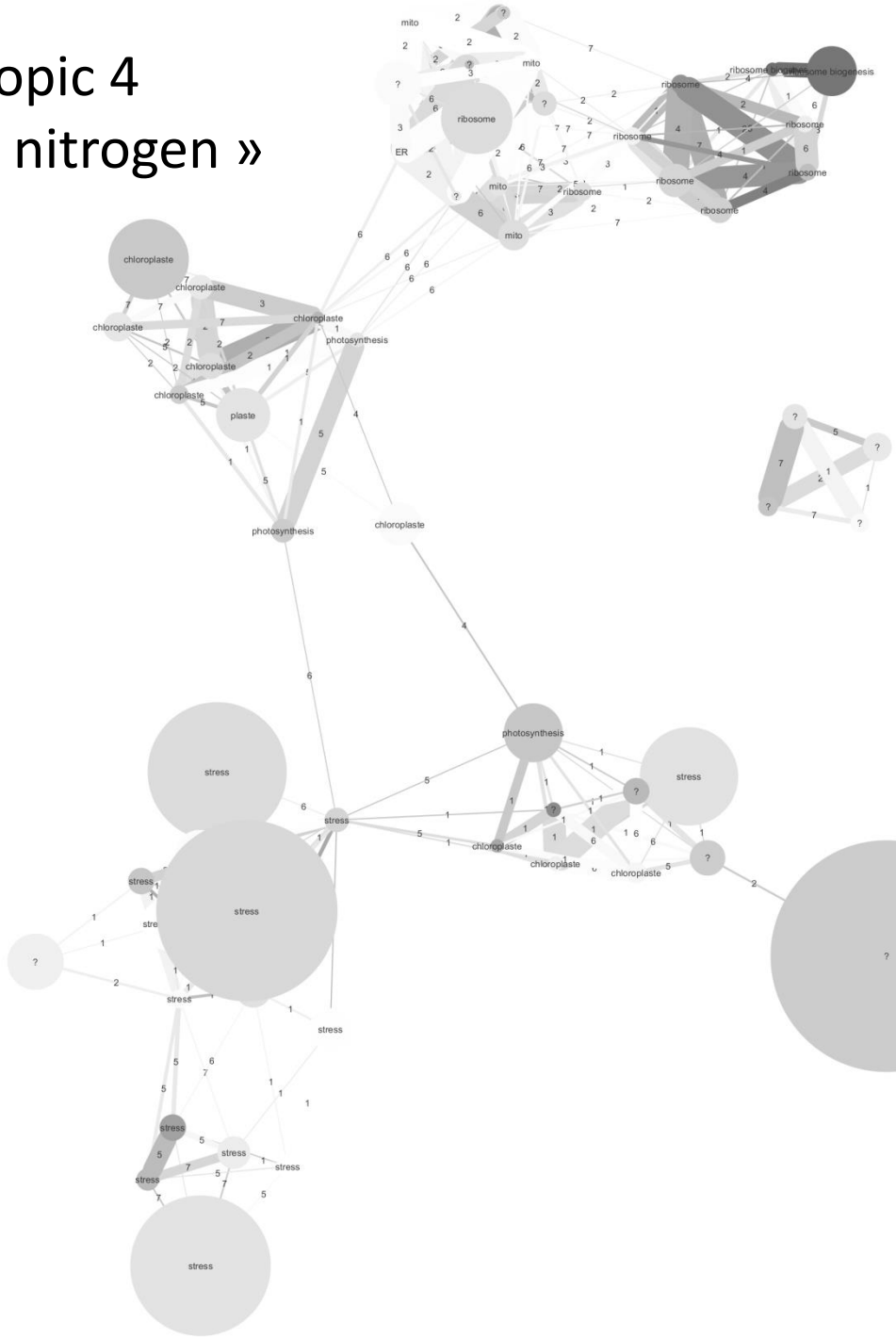
Nitrogen

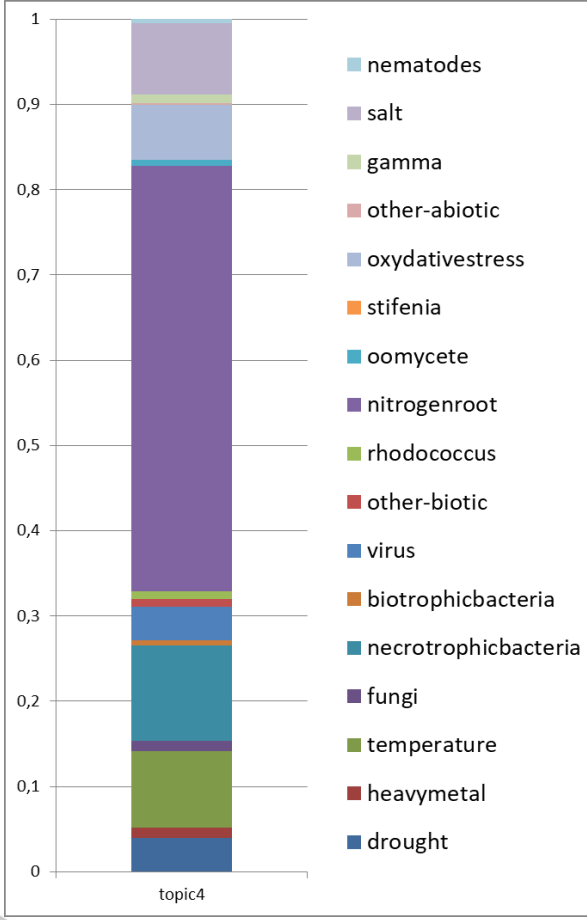
Heavy metals & viruses

Main topic

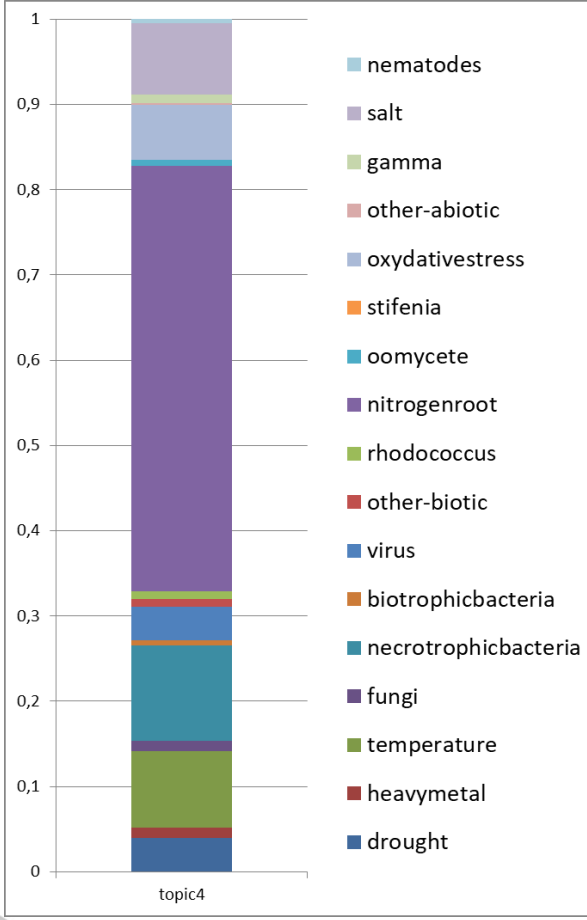

mito
ribosome
ribosome biogenesis
chloroplaste
photosynthesis
plaste
chloroplaste
stress
photosynthesis
stress
chloroplaste
stress
stress
stress
stress
stress
stress
?
?
plasmodesma

Topic 4
« nitrogen »





Topic 4
« nitrogen »



Which gene is doing what in which condition

Predictive genomics

Conclusions of a biologist

CATdb

Transcriptomic data

Modeling of co-expression

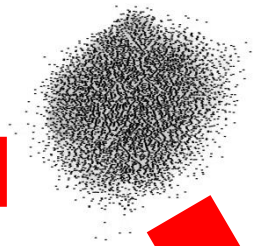
Gaussian Mixture Model

GEM2Net

Co-expression

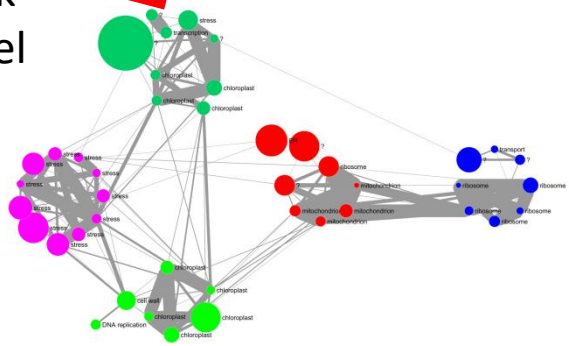
Modeling of co-regulation

integration



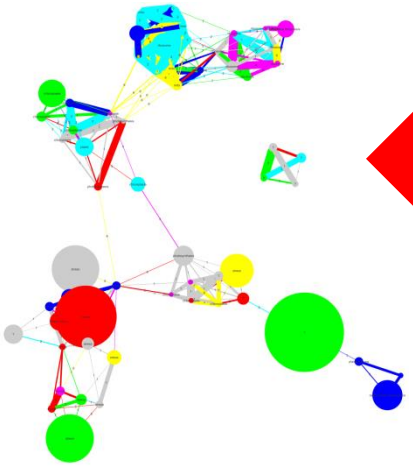
Modeling of network topology

Stochastic Block Model



Modeling of context-dependent topology

Stochastic Topic Block model



Conclusions of a biologist

CATdb

Transcriptomic data

Modeling of co-expression

Gaussian Mixture Model

GEM2Net

Co-expression

Modeling of co-regulation

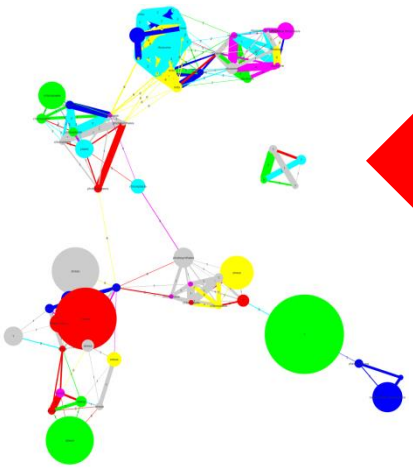
integration

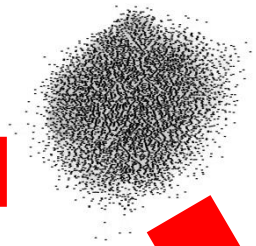
Modeling of context-dependent topology

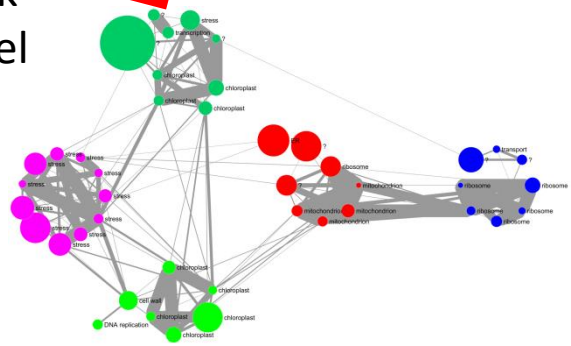
Stochastic Topic Block model

Modeling of network topology

Stochastic Block Model







Statistical modeling is a relevant approach to formulate biological hypotheses

Acknowledgements

IPS2

Marie-Laure Martin-Magniette
Guillem Rigaill
Rim Zaag
Nathalie Rézé
The IPS2 transcriptomic platform
The GNet team