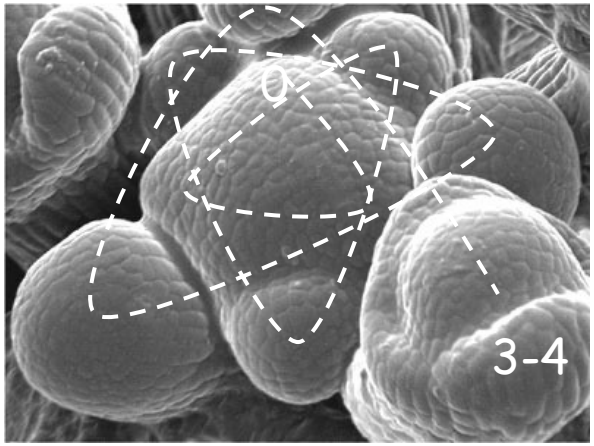
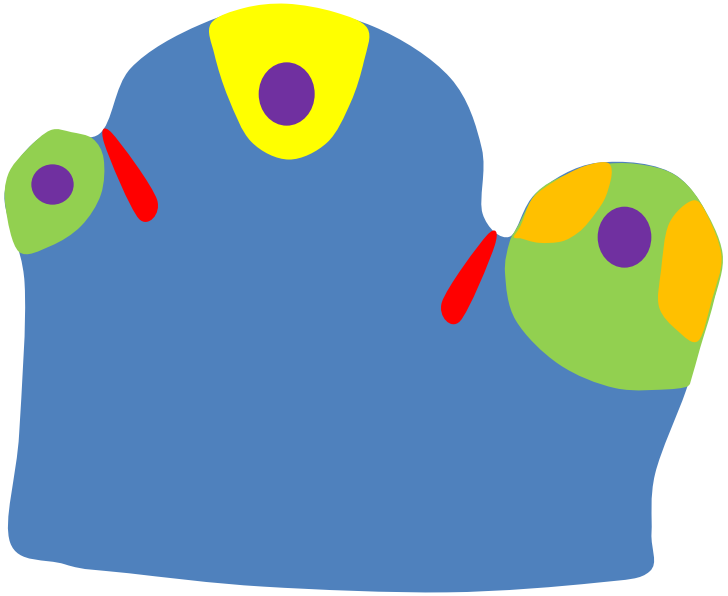
How do molecular networks control flower development?



Cell identities are associated with molecular steady states

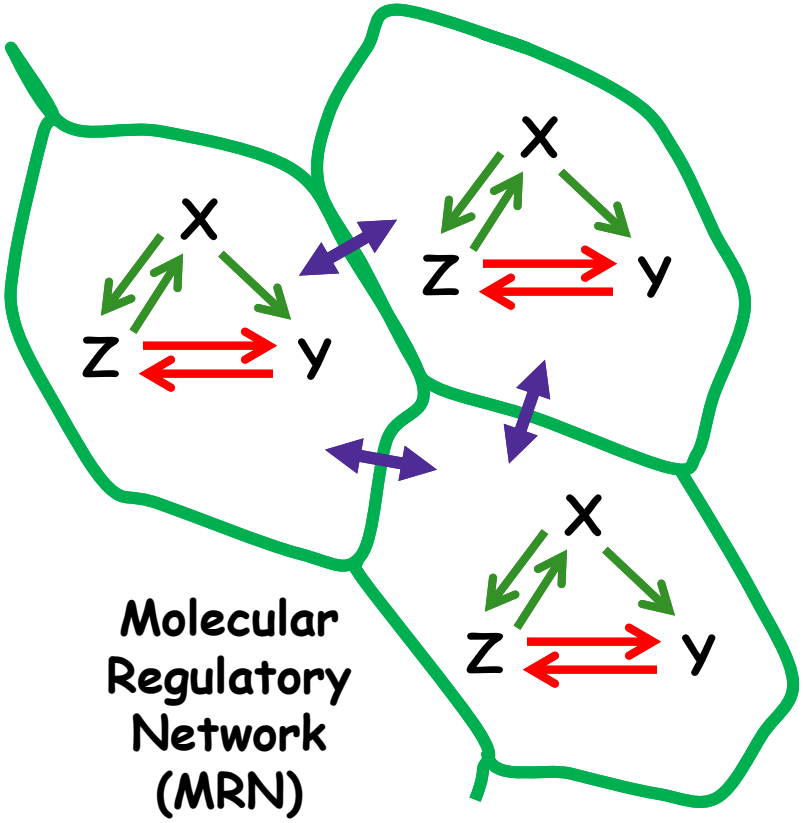


Development can be seen as a succession of molecular steady states evolving through time

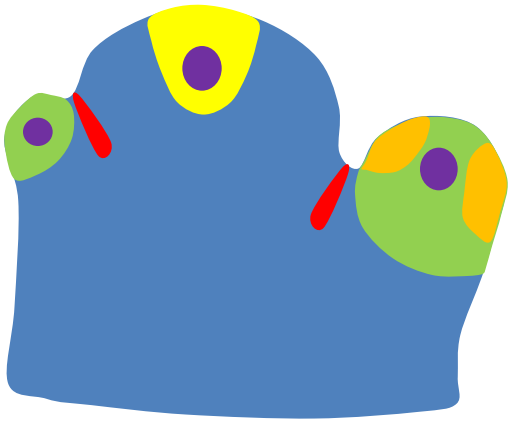
Very important point: this system is self-organised

1-Same interaction network in all the cells

2-Communication between cells affect the state of the network

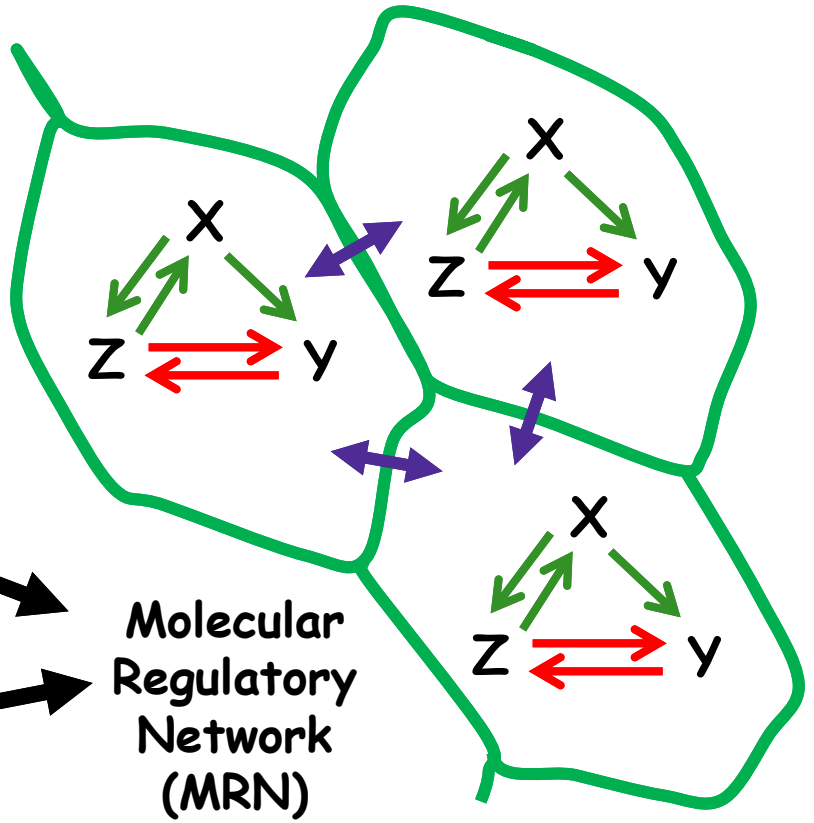


Objective: construct a MRN which is coherent with expression patterns and long range interactions



Expression patterns
(steady states)

Long range
interactions



Molecular
Regulatory
Network
(MRN)

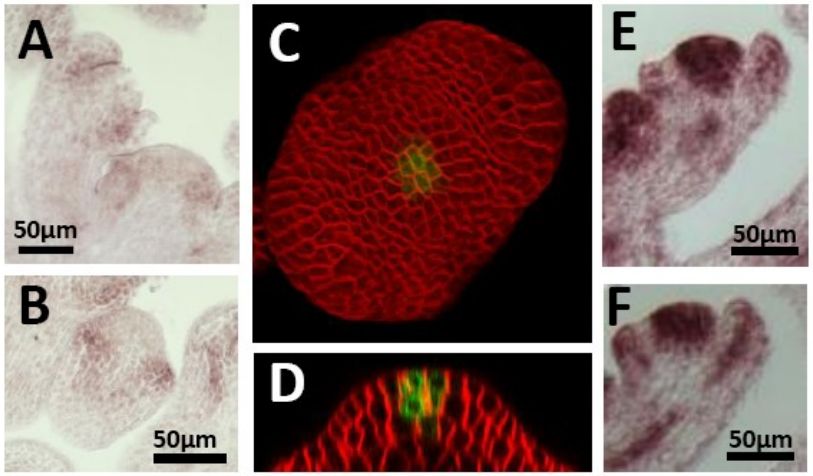
behavior

structure

1- Interaction database (~300 papers)

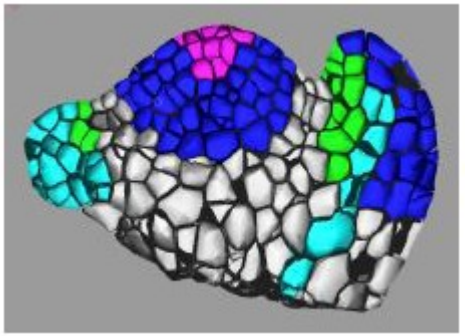
- direct molecular interactions
- induction evidence
- genetic interactions (can be direct or not)

2- Expression database > atlas > molecular states



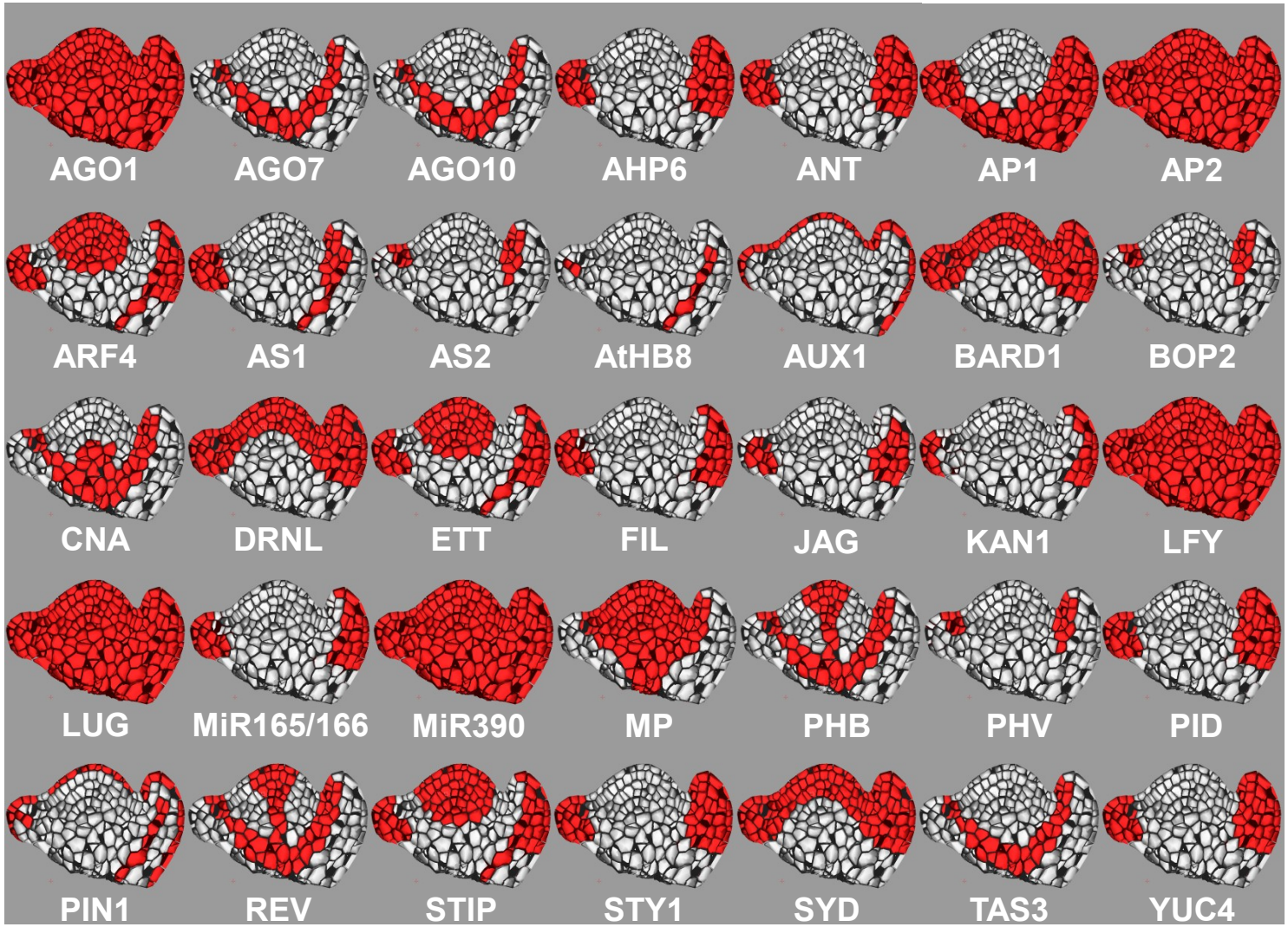
	Zone 1	Zone 2	Zone 3	Zone 4
AS1	1	0	1	0
CLV3	0	0	0	1
ETT	0	1	1	1

Steady states



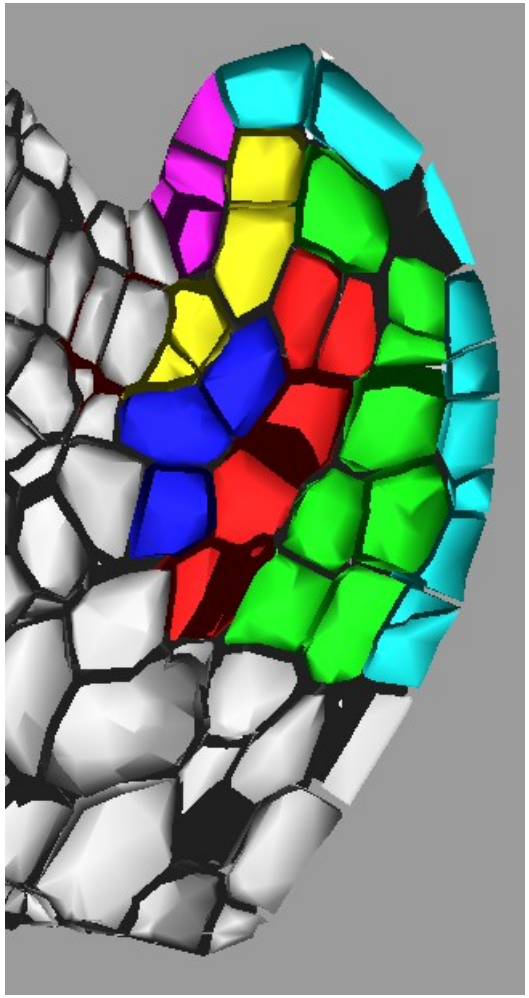
Superimposition

35 sepal genes projected (flower bud stage 3)

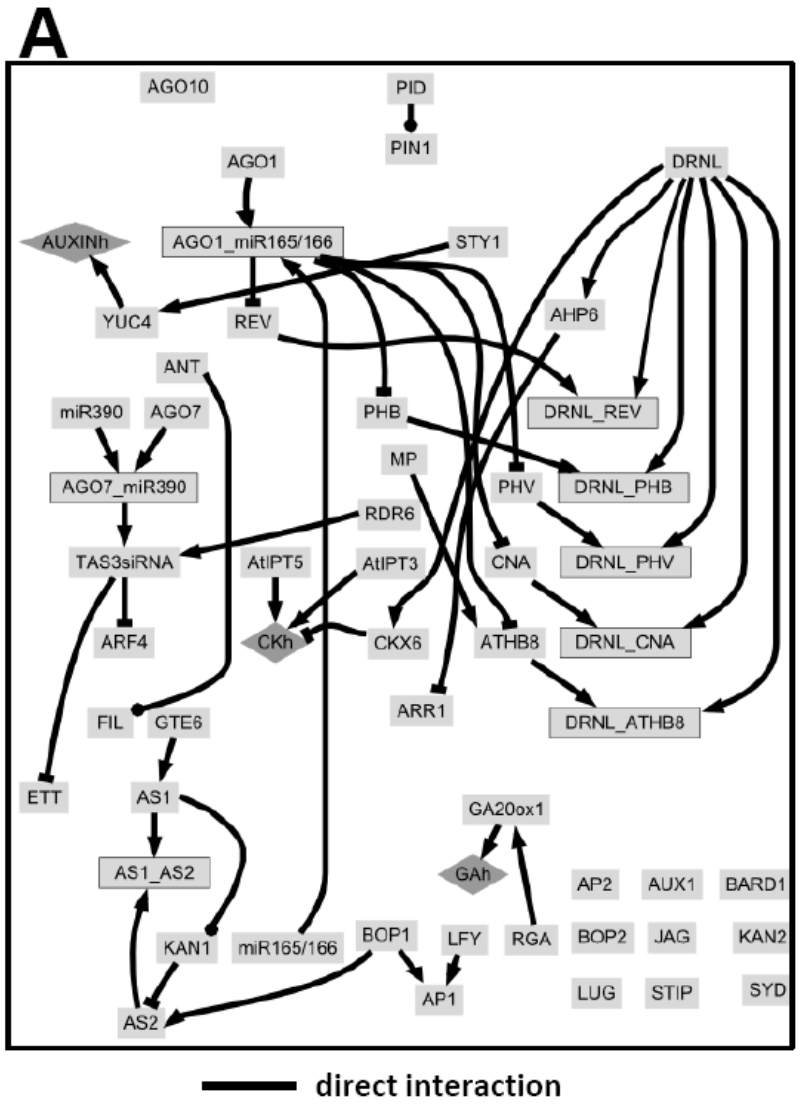


Sepal primordium : 6 molecular states with 35 elements

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5	Zone 6
AGO1	1	1	1	1	1	1
AGO7	1	1	1	0	0	0
AGO10	1	1	1	0	0	0
AHP6	1	1	1	1	1	1
ANT	1	1	1	1	1	1
AP1	1	1	1	1	1	1
AP2	1	1	1	1	1	1
ARF4	0	0	0	1	1	1
AS1	1	1	1	1	0	0
AS2	1	1	0	0	0	0
AtHB8	0	0	0	1	0	0
AUX1	1	0	0	0	1	0
BARD1	1	1	1	1	1	1
BOP2	1	1	0	0	0	0
CNA	1	1	1	0	0	0
DRNL	1	1	1	1	1	1
ETT	0	0	0	1	1	1
FIL	0	0	0	1	1	1
JAG	0	0	1	1	0	1
KAN1	0	0	0	0	1	1
LFY	1	1	1	1	1	1
LUG	1	1	1	1	1	1
MiR165/166	0	0	0	1	1	1
MiR390	1	1	1	1	1	1
MP	1	1	1	0	0	0
PHB	1	1	1	0	0	0
PHV	1	1	0	0	0	0
PID	1	1	1	1	1	1
PIN1	1	0	0	1	1	0
REV	1	1	1	0	0	0
STIP	0	0	0	1	1	1
STY1	1	1	1	1	1	1
SYD	1	1	1	1	1	1
TAS3	1	1	1	0	0	0
YUC4	1	1	1	1	1	1



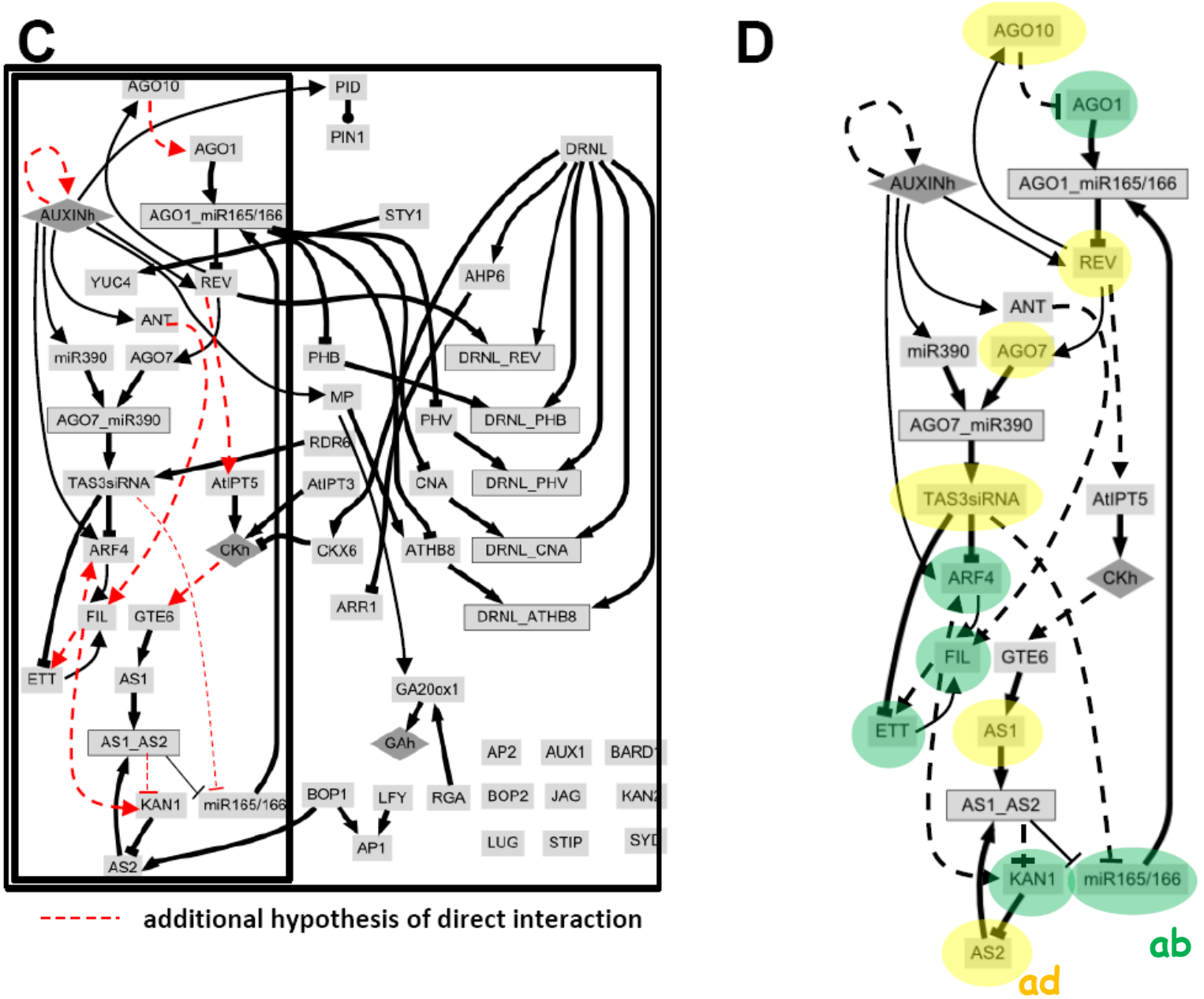
3- Candidate molecular interaction graph



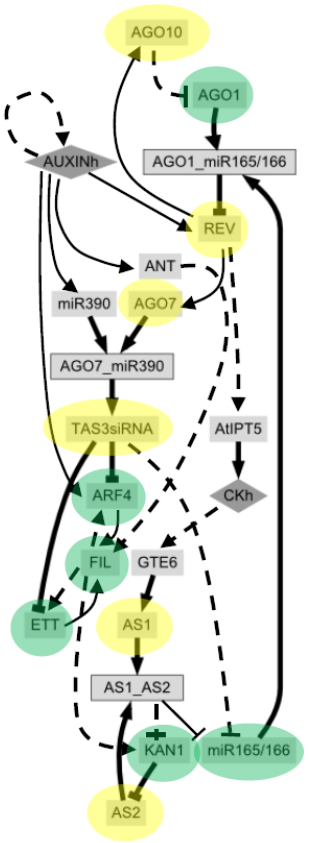
Theoretical requirements

- close circuits
- inputs

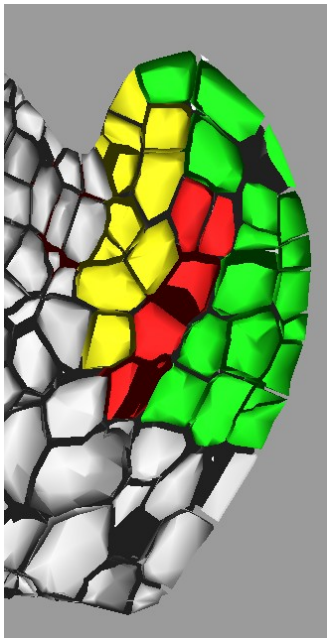
3- Candidate molecular interaction graph



Expected behavior of our network



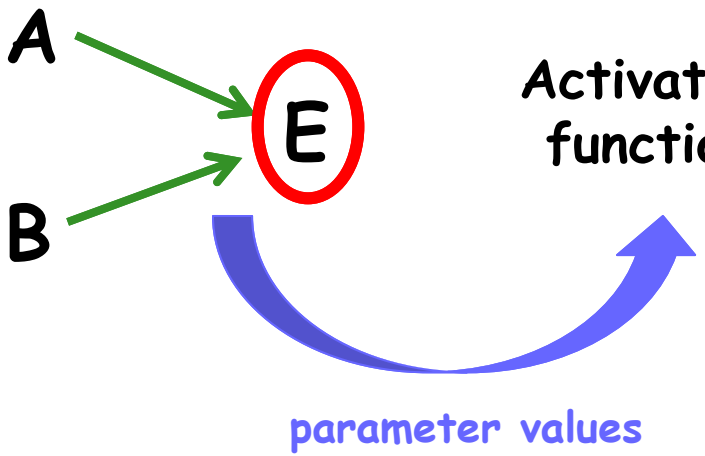
	Zone 1 Adaxial	Zone 2 Vascular	Zone 3 Abaxial
AGO1	1	1	1
AGO10	1	0	0
AGO7	1	0	0
ANT	1	1	1
ARF4	0	1	1
AS1	1	1	0
AS2	1	0	0
ETT	0	1	1
FIL	0	1	1
KAN1	0	0	1
MiR165/166	0	1	1
MiR390	1	1	1
REV	1	0	0
TAS3	1	0	0



A mathematical model is required
Boolean logic has been used

4- Parameter values inferred from expression data > solution(s)

Behavior expressed as an activation function



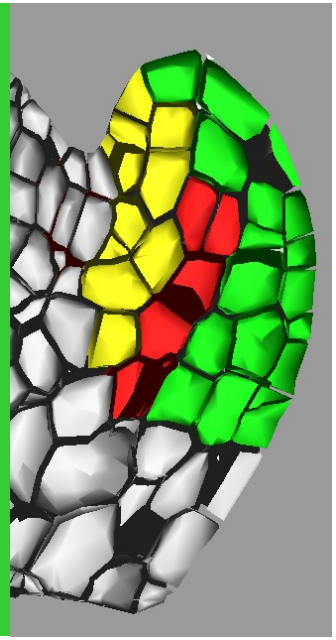
state of the inputs		output
A	B	E
0	0	0 0 0 .....
0	1	0 1 0 .....
1	0	0 1 1 .....
1	1	1 1 1 .....

Activation functions are found for each element

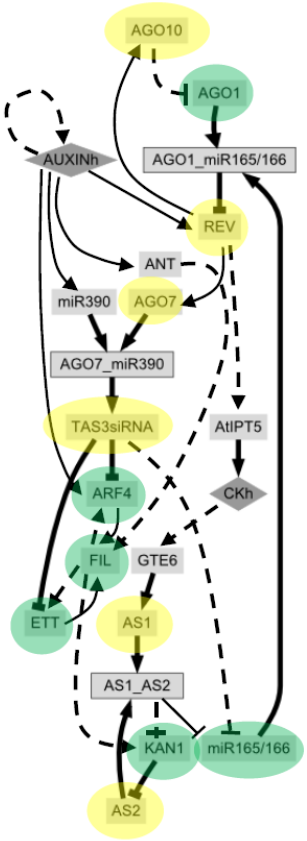
A solution is represented by a set of activation functions

47 solutions were found

	Zone 1 Adaxial	Zone 2 Vascular	Zone 3 Abaxial
AGO1	1	1	1
AGO10	1	0	0
AGO7	1	0	0
ANT	1	1	1
ARF4	0	1	1
AS1	1	1	0
AS2	1	0	0
ETT	0	1	1
FIL	0	1	1
KAN1	0	0	1
MiR165/166	0	1	1
MiR390	1	1	1
REV	1	0	0
TAS3	1	0	0



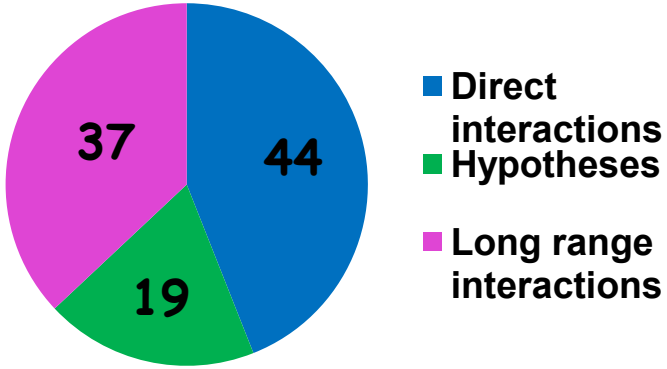
Steady states



Candidate molecular network

Structure

5- Model validation



37 genetic interactions used
to test the dynamics of the MRNs

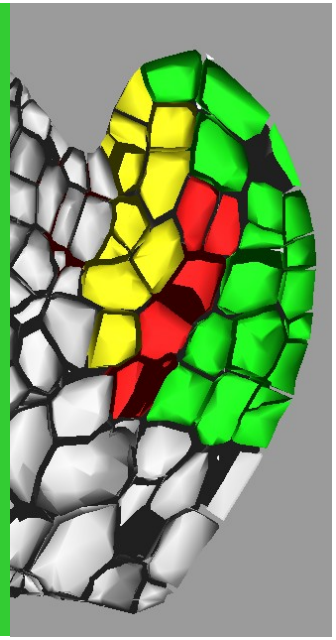
Model predictions ⟷ Experimental observations

35 (out of 37) indirect interactions predicted by the model were supported by experimental observations

The 47 solutions were all equally good

The Plant Cell (2011) La Rota et al.

	Zone 1 Adaxial	Zone 2 Vascular	Zone 3 Abaxial
AGO1	1	1	1
AGO10	1	0	0
AGO7	1	0	0
ANT	1	1	1
ARF4	0	1	1
AS1	1	1	0
AS2	1	0	0
ETT	0	1	1
FIL	0	1	1
KAN1	0	0	1
MiR165/166	0	1	1
MiR390	1	1	1
REV	1	0	0
TAS3	1	0	0

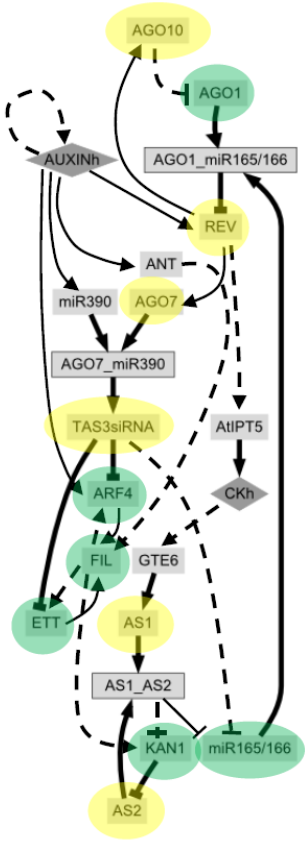


Steady states

- 1- off(AGO10) → up(miR165), down(REV)
- 2- on(AGO10) → up(REV)
- 3- off(AGO7) → unchanged(FIL), up(ARF4,ETT)
- 4- off(AS1) → unchanged(FIL), up(ETT)
-
-
-

Long range interactions

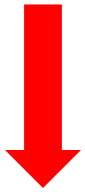
Behavior



Candidate molecular network

Structure

If we considere this network as the reference



Can we infere this network (or a resembling one)
from transcriptomic data?

Among the different methods used
to infer networks, which ones performs the best,
which kind of information is obtained ?

Matériel utilisé

Données transcriptome issues de la plateforme transcriptomique de l'URGV.

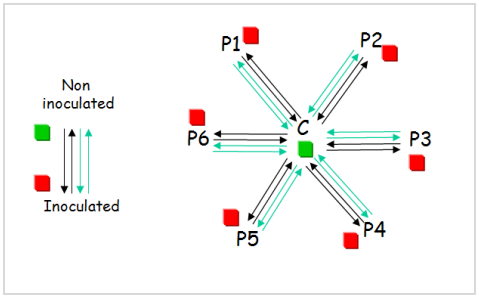
- Même protocole expérimentaux pour toutes les comparaisons
- Même analyses statistiques pour la normalisation et l'analyse différentielle
- Stockage et gestion des données dans CATdb

Permet de

- garantir une certaine homogénéité du jeux de données
- retourner "assez facilement" à la description expérimentale de chaque expérience

Démarche de la plate-forme

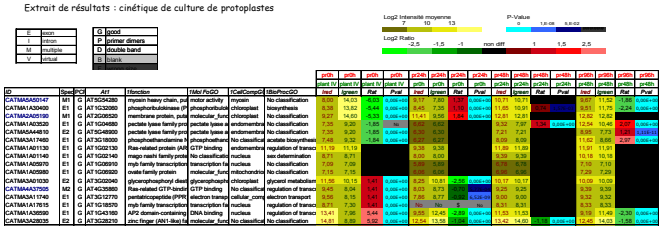
Choix de la puce (méthode)



Dessin des expériences

Hybridations

Analyses statistiques



Stockage dans CATdb

CATdb a Complete Arabidopsis Transcriptome database

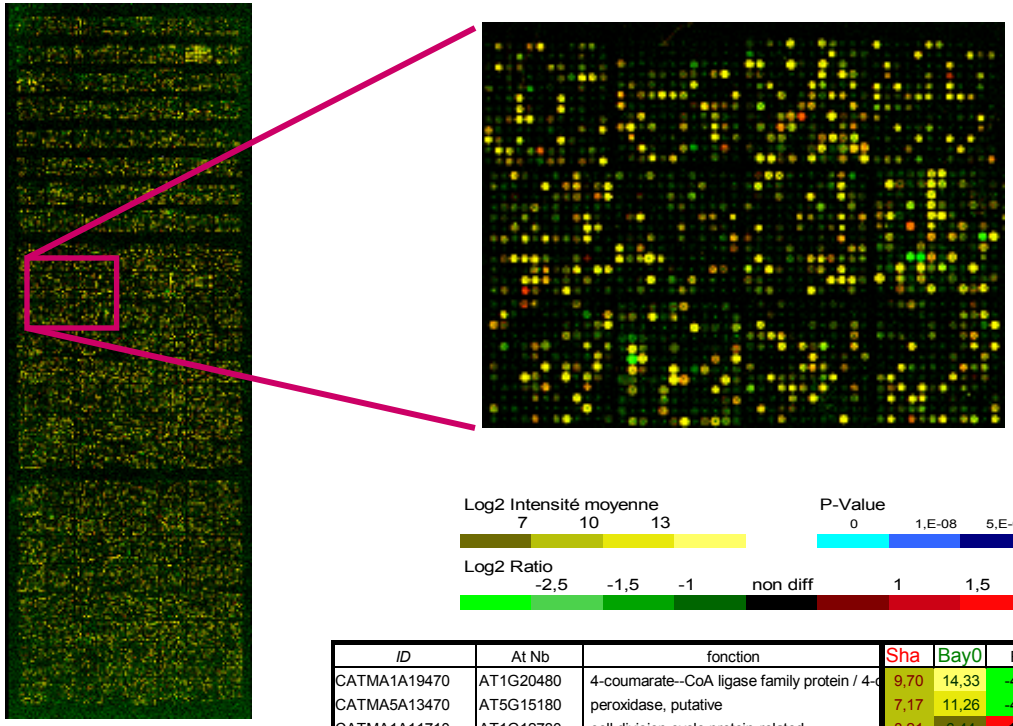
Aide à la lecture des résultats

FLAGdb++

CATdb

La puce CATMA: évolution au cours du temps mais sur toutes les versions

présence de 24576 sondes pour plus de 22K genes



THRESHOLD = p-val < 0.05
after Bonferroni
correction

ID	At Nb	fonction	Sha	Bay0	Rat	Pval	ril140	Sha	Rat	Pval	ril140	ril104	Rat	Pval
CATMA1A19470	AT1G20480	4-coumarate--CoA ligase family protein / 4-(	9,70	14,33	-4,63	0,00E+0	9,62	9,99	-0,37	1,00E+0	9,78	9,73	0,05	1,00E+0
CATMA5A13470	AT5G15180	peroxidase, putative	7,17	11,26	-4,09	0,00E+0	10,52	7,06	3,46	0,00E+0	10,40	6,92	3,48	0,00E+0
CATMA1A11710	AT1G12730	cell division cycle protein-related	8,21	6,44	1,77	0,00E+0	8,51	8,43	0,08	1,00E+0	8,39	8,37	0,02	1,00E+0
CATMA4A16726	AT4G15910	drought-responsive protein / drought-induce	11,12	9,19	1,93	0,00E+0	10,77	10,75	0,02	1,00E+0	10,84	9,25	1,59	0,00E+0
CATMA3A45470	AT3G52540	ovate family protein	9,07	7,04	2,03	0,00E+0	7,02	8,84	-1,82	0,00E+0	7,14	7,41	-0,27	1,00E+0
CATMA4A04850	AT4E04720	eugene prediction	9,03	6,68	2,35	0,00E+0	6,75	8,77	-2,01	0,00E+0	6,80	9,36	-2,55	0,00E+0
CATMA3A49320	AT3G56350	superoxide dismutase (Mn), putative / mang	9,42	7,17	2,26	0,00E+0	7,88	9,42	-1,54	0,00E+0	7,89	7,08	0,81	8,50E-5
CATMA3A43260	AT3G50220	expressed protein	11,36	8,79	2,57	0,00E+0	8,38	11,12	-2,74	0,00E+0	8,37	8,05	0,32	1,00E+0
CATMA3A19130	AT3G19520	expressed protein	10,36	7,85	2,52	0,00E+0	7,94	10,15	-2,21	0,00E+0	7,97	10,68	-2,71	0,00E+0
CATMA5A33260	AT5G37890	seven in absentia (SINA) protein, putative	9,42	6,84	2,58	0,00E+0	9,52	9,57	-0,05	1,00E+0	9,36	9,43	-0,07	1,00E+0
CATMA2A46320	AT2G47900	F-box family protein / tubby family protein	10,04	7,70	2,34	0,00E+0	9,62	9,99	-0,37	1,00E+0	9,57	8,12	1,45	0,00E+0
CATMA3A07710	AT3G08900	reversibly glycosylated polypeptide-3 (RGP3	11,82	11,44			12,01	10,87	1,14	1,68E-11	10,93	11,31		

Les versions correspondent à des ajouts de sondes. En particulier pour étudier les Petits ARNs

Conséquences: données manquantes non aléatoires pour certains gènes car inclus sur la puce dans les plus récentes versions seulement

Description d'une comparaison entre 2 conditions

- Puce 2 couleurs et expérience faite en dye-swap

- Au moment de l'analyse d'image, retrait des sondes ayant un signal non conforme (expertise humaine) → données manquantes au hasard

- Normalisation par lowess, correction d'un effet bloc -> obtention d'un log-ratio

- Partage de la correction pour obtenir une intensité par échantillon : I.ref, I.sample

- Analyse différentielle : variance comune calculée à partir des sondes ayant une variance de la différence d'expression non extrême. Ajustement de la pvalue par Bonferroni.

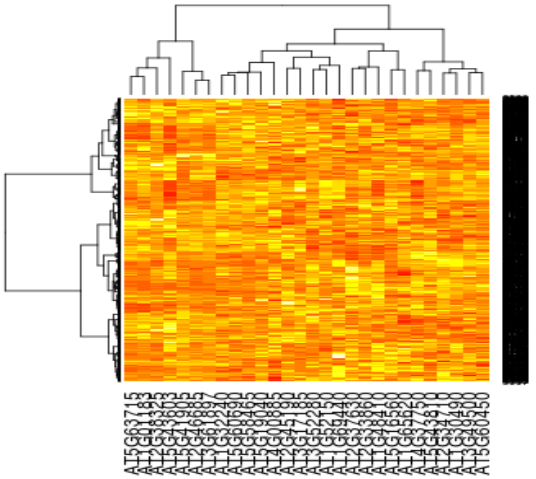
Extraction de CATdb

- Toutes les données publiques le 21 juin 2013 : 2171 comparaisons donc 4342 échantillons disponibles

- Information concernant les comparaisons
  - Organe : 32 modalités avec 4 principales (aerial, rosette, roots, whole plant)
  - Stade de développement : 82 stades
  - L'écotype : 40 dont 1077 columbia
  - Le génotype : 486 modalités
  - Présence de mutation : 4 modalités (yes/no, no, yes, no/yes) : 1083 no

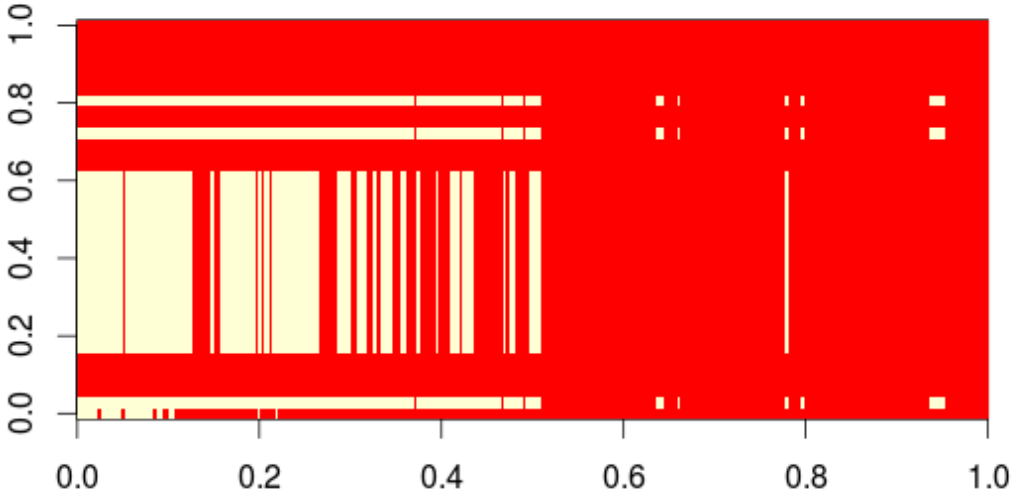
Data 1

- 38 selected probes annotated with an AtG ID corresponding to 28 genes in ref. network.
- 1,558 selected experiments among 2,171 according to sample origin (e.g. 'Leaf', 'whole plant' YES whilst 'Root', 'cell culture' NO)
- → compatibility of plant physiology with organ (sepal) polarity ?
- Data was standardised.



Data 2

- Problem with NA's → ~~data imputation or~~ discarding experiments with missing values because they are not missing at random



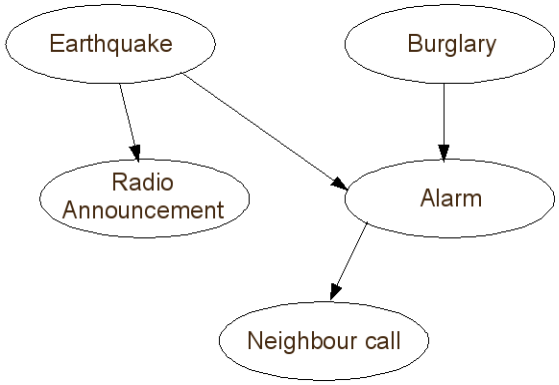
- 21(/37) probes have missing values, 3 with ~half are missing, 17 with ~1/3 missing and 1 with 1/10th missing
- Roughly (less than) half of the experiments have no missing data but 15% have missing values for 3 genes only and 35% have over 20 missing measures (over 37)

Discrete Bayesian network

Static Bayesian networks *(Friedman et al., Plos comput. bio., 2000)*

- Directed Acyclic Graph (DAG)
- Conditional probability distribution of X_i, given its parents Pa_i in G: $P_G(X_i/Pa_i^j)=\theta_i^j$

  (independence of these local probabilities)



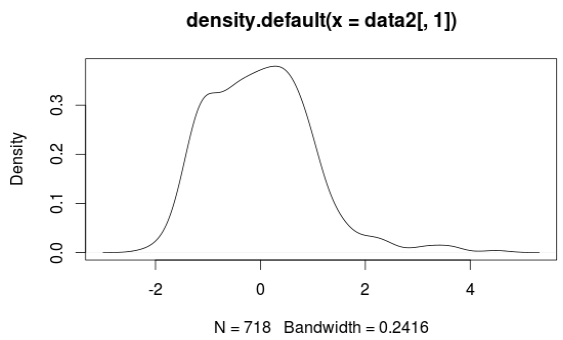
Graphic representation of a joint probability distribution:

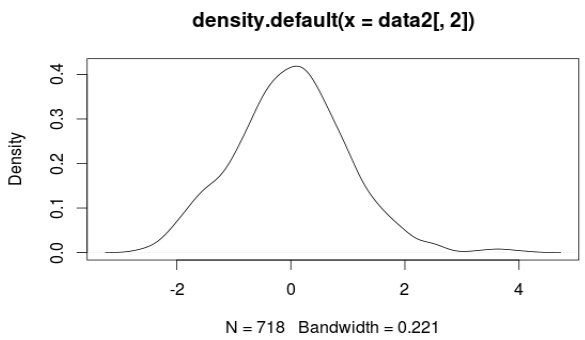
$$P_G(X)=\prod_{i=1}^{n} P_G(X_i/Pa_i)$$

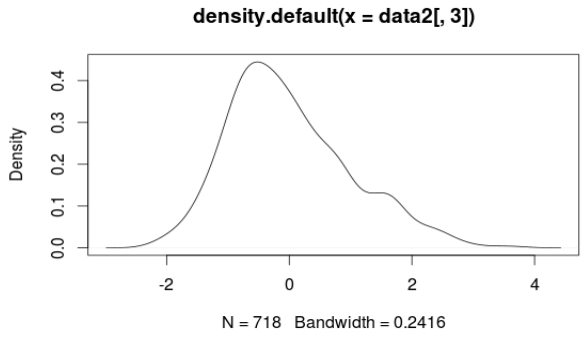
Probability Distribution for the Alarm Node given the events of "Earthquakes" and "Burglaries"			
E	B	P(A | E,B)	P(!A | E,B)
E	B	0.90	0.10
E	! B	0.20	0.80
! E	B	0.90	0.10
! E	! B	0.01	0.99

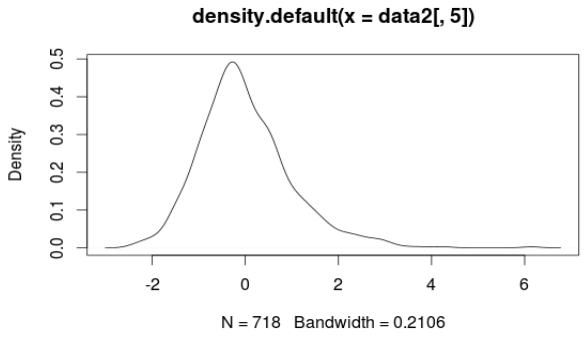
Gaussian graphical model

- $X \sim N(0,\Sigma)$, Σ or more specifically Σ^{-1} encodes direct relationships between probes.
- Data (samples) assumed iid.









Bootstrap

- Counts of |partial corr.| > |random partial corr.|

in an augmented Σ^{-1} over n_{boot}=10,000 bootstrap repeats (stability with n_{boot}=1,000, slightly more stringent).

- Keep edges having counts = n_{boot}.

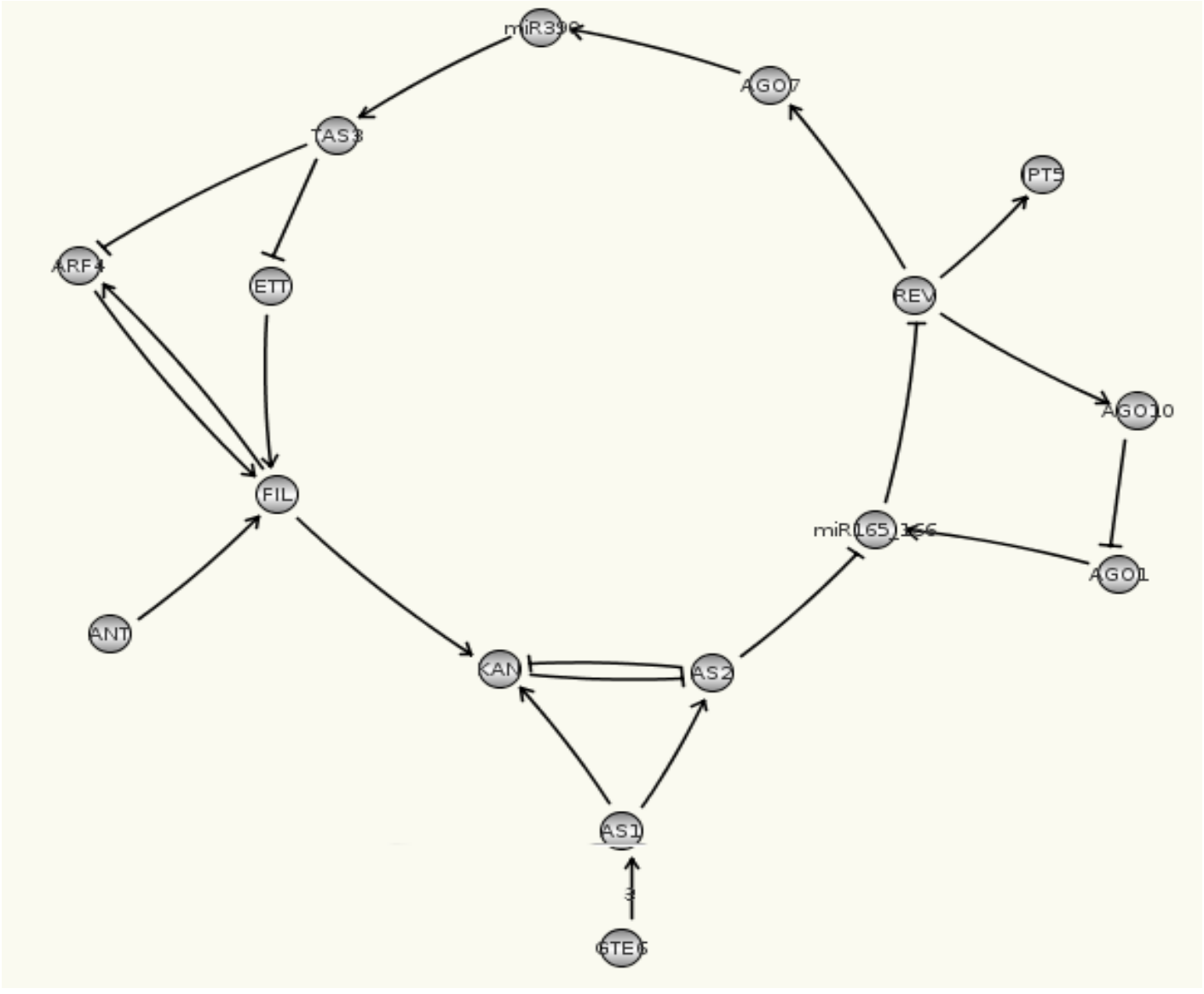
→ list of ~50 edges (between probes).

Group of genes

- Merge genes known to have similar targets
  - [REV, PHB, PHV, CNA] = REV
  - [KAN1, KAN2] = KAN
  - [TAS3, RDR6] = TAS3
- Complex of proteins
  - [AGO1_miR165/166] = miR165/166
  - [AGO7_miR390] = miR390
  - [DRNL_REV] = REV
  - [AS1_AS2] = AS2

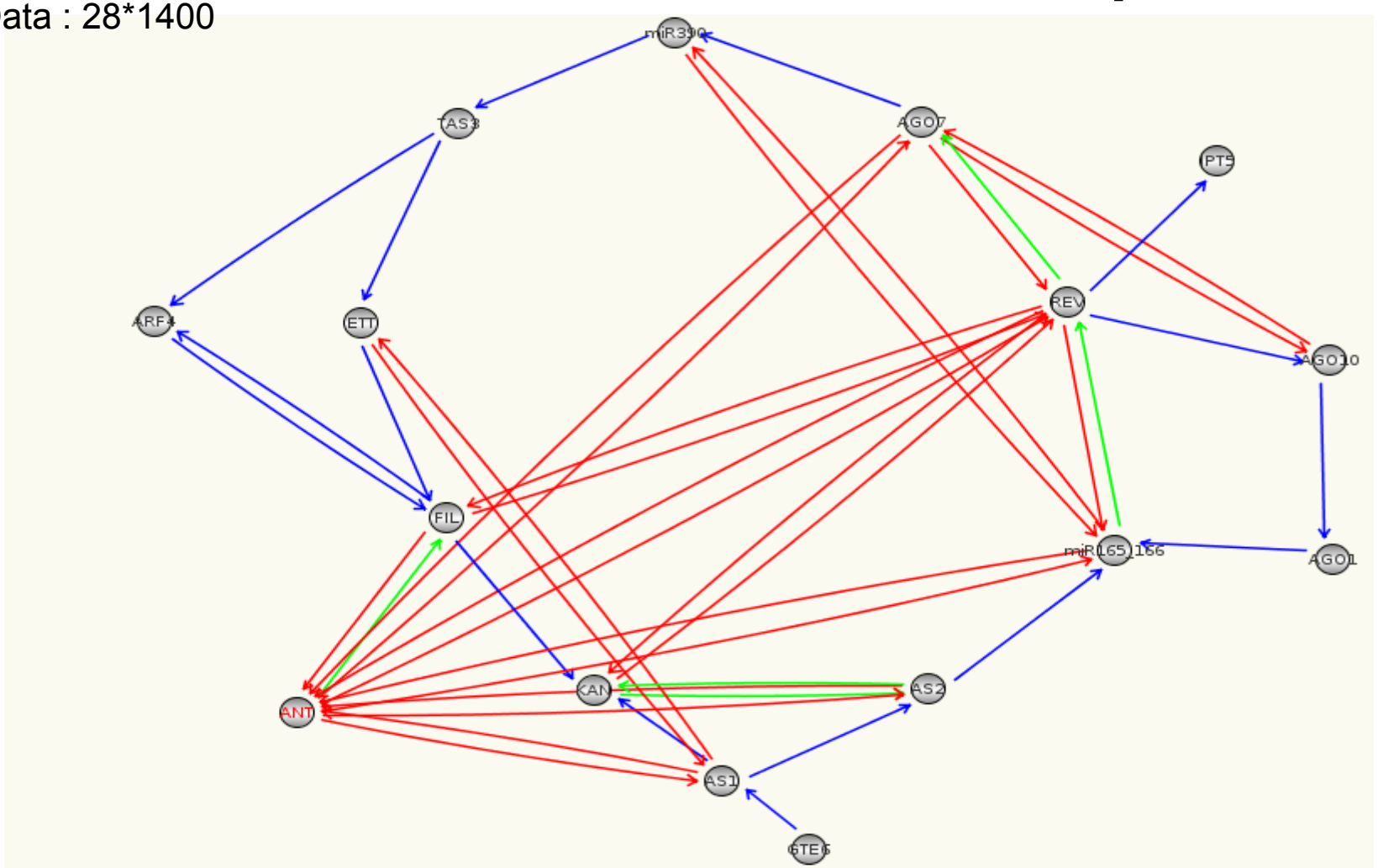
Resulting in 16 classes for 28 transcripts/probes

Direct interaction graph in sepal



GGM with bootstrap

Data : 28*1400

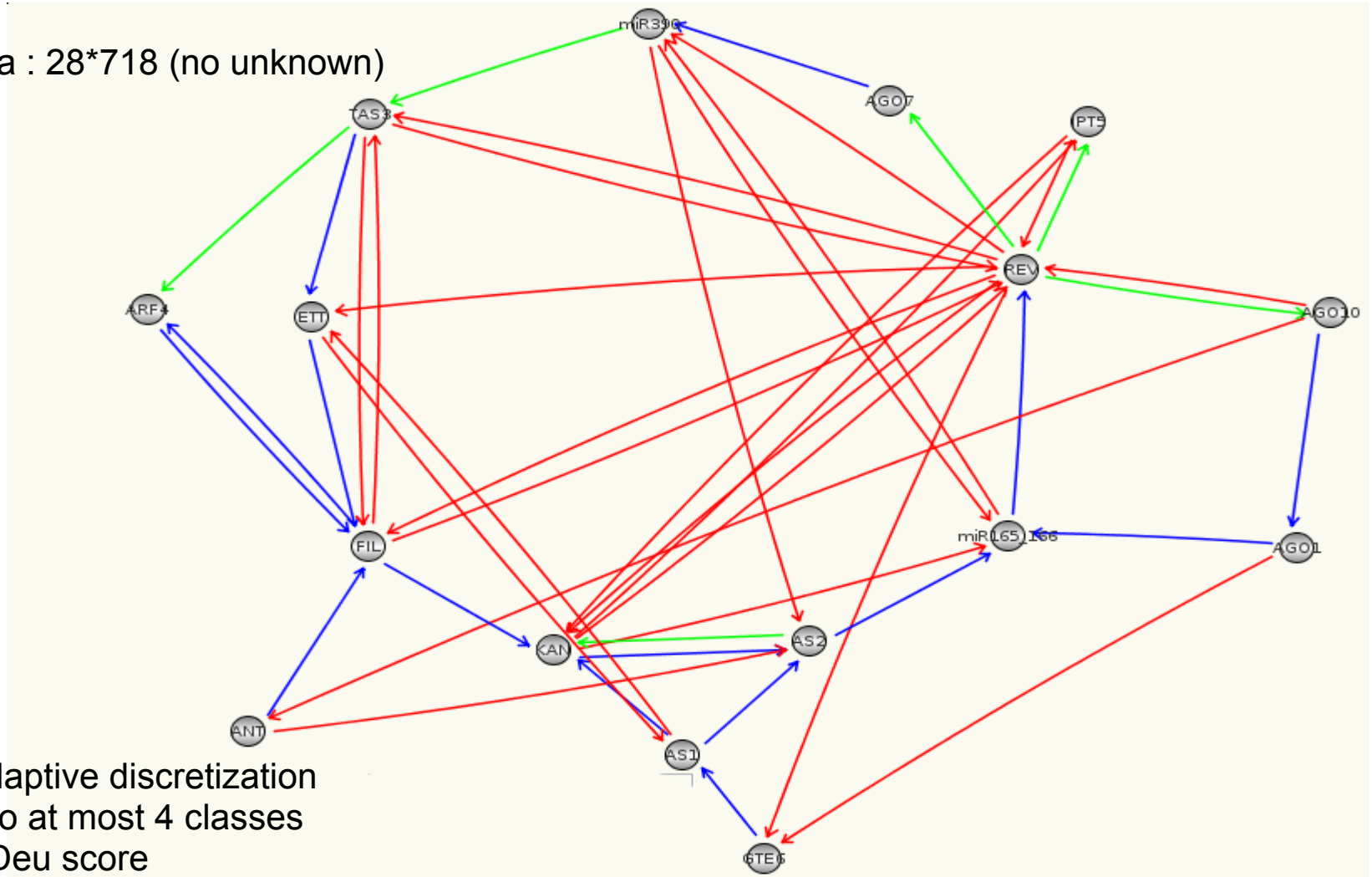


Nonoriented edges present in 1000 bootstraps compared to the sepal graph

http://www.cs.york.ac.uk/aig/sw/gobnilp/

Bayesian Network with bootstrap

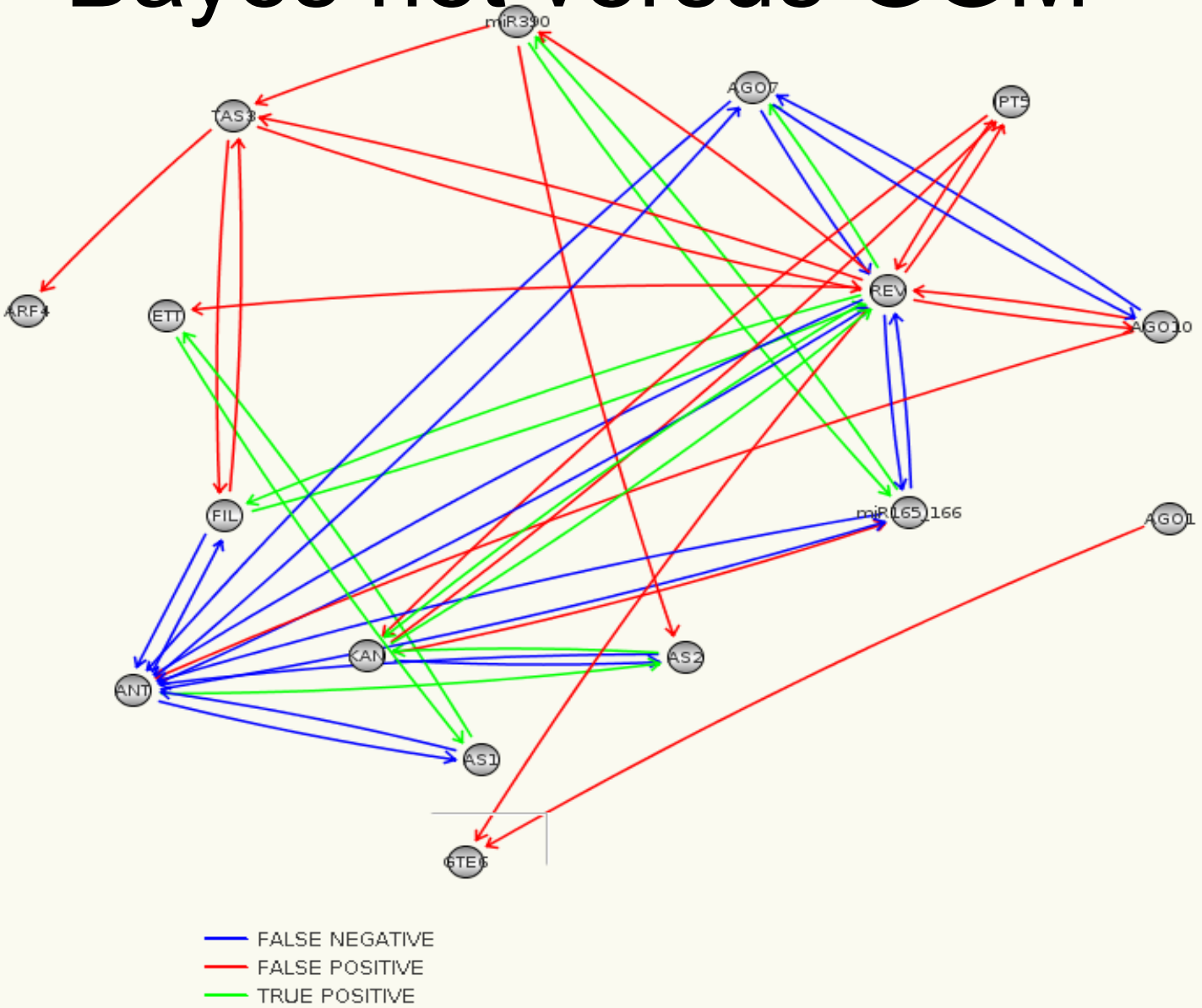
Data : 28*718 (no unknown)



- Adaptive discretization into at most 4 classes
- BDeu score with alpha=1
- Exact solving with GOBNILP (maxp=3)

Nonoriented* edges present in +70/100 bootstraps compared to the sepal graph
(* oriented if >70% in the same direction)

Bayes net versus GGM



Open questions

- Experiment selection: how is it linked to the reference network ?
- Elements of the network not included in the data ?
- Other inference methods ? But why would the tested 2 produce so different results: complementarity or (very) noisy predictions ?
- Room for causal inference from obs. Data ?