Towards Arabidopsis thaliana genetic regulatory network using discrete Bayesian network structure learning algorithms

Jimmy Vandel, Brigitte Mangin,
Matthieu Vignes & Simon de Givry.

Université Paul Sabatier
TOULOUSE III

Unité BIA Toulouse
Biométrie et Intelligence Artificielle

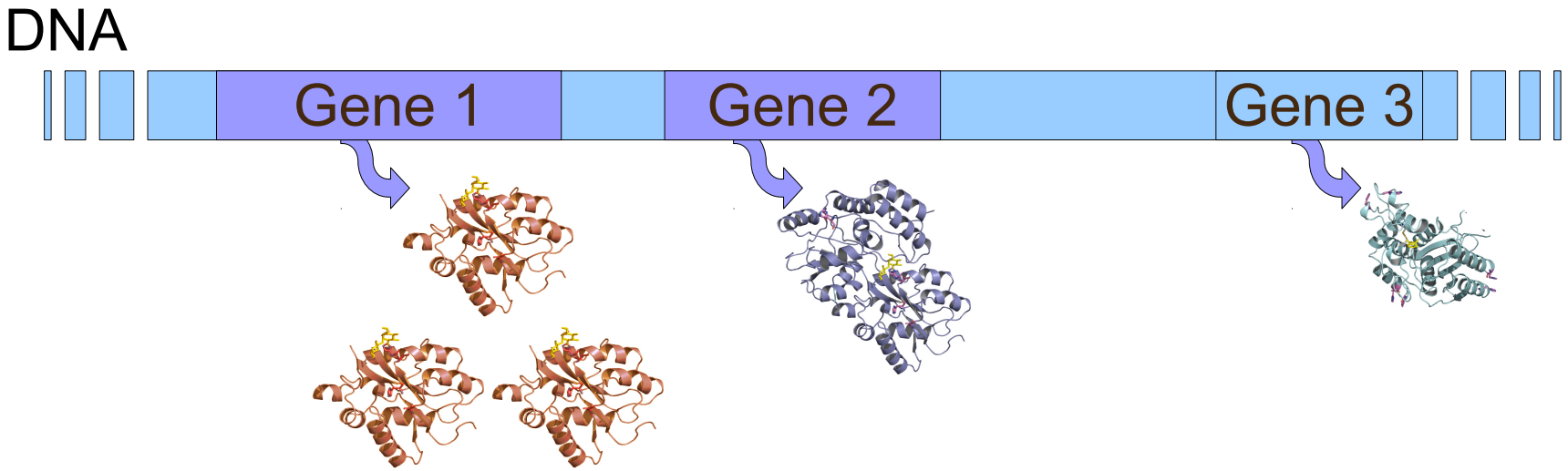
INRA
Institut National de la Recherche Agronomique

Outlines:

- Biological motivation
- Bayesian Networks framework
- Learning Algorithms
- Arabidopsis thaliana data
- Experimentation
- Perspectives

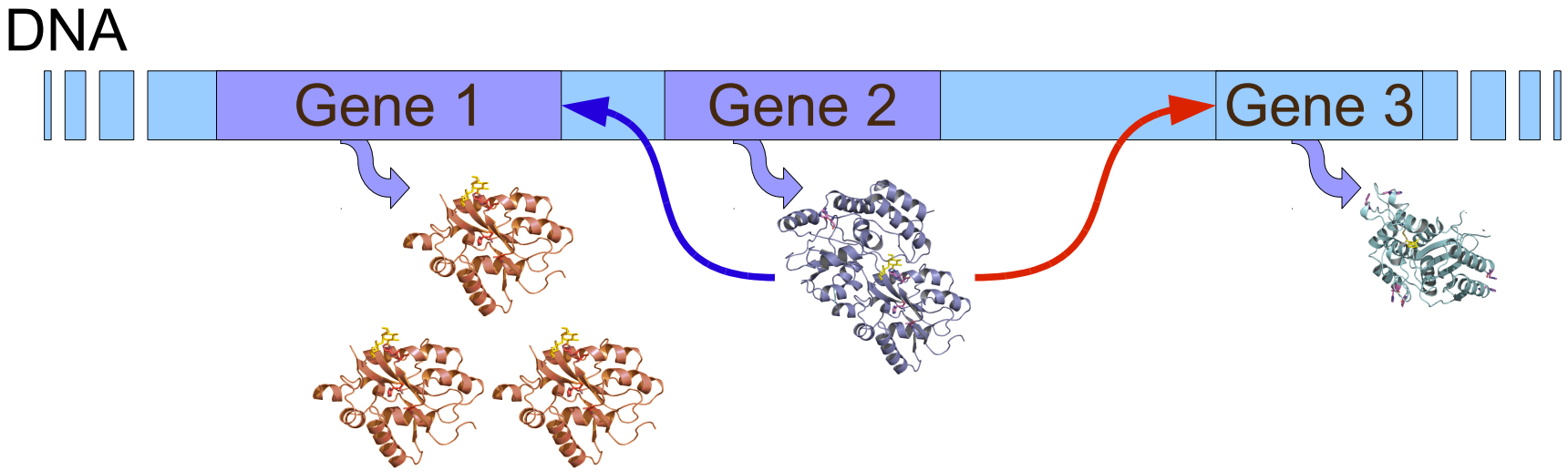
Biological motivation

Biological motivation



→ gene expressions (mRNA concentrations)

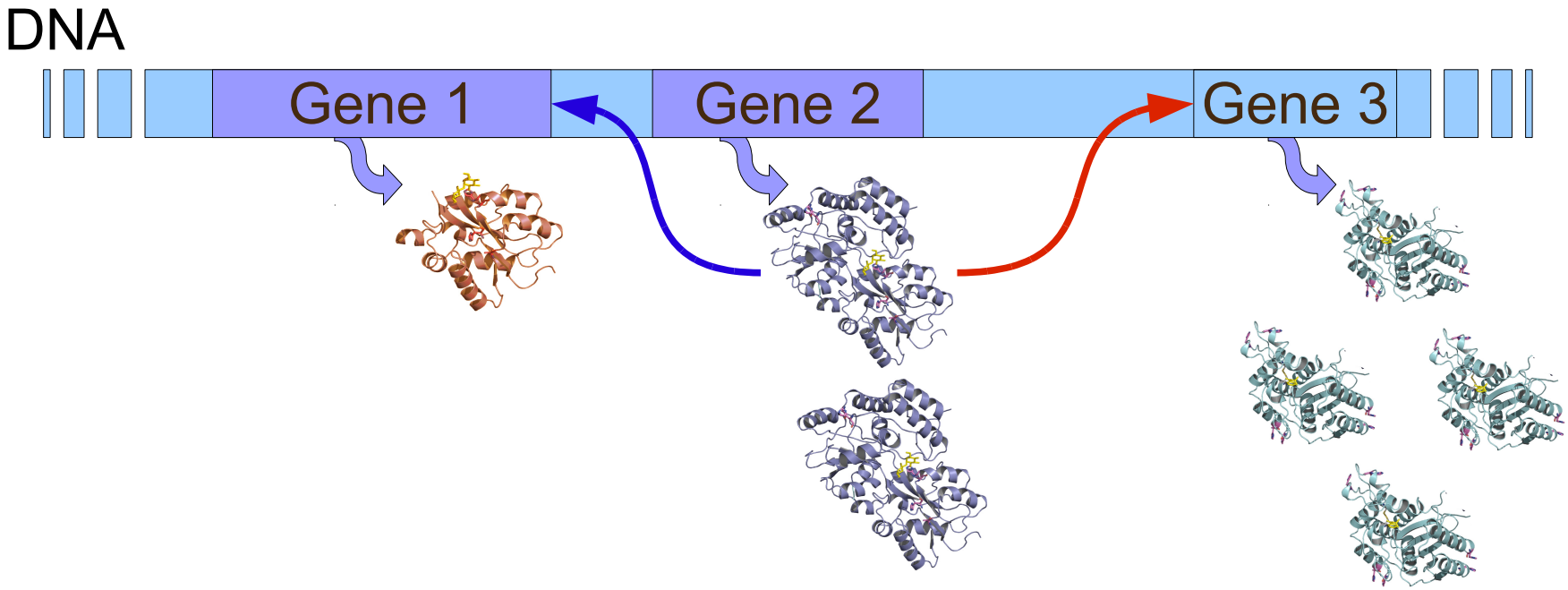
Biological motivation



→ gene expressions (mRNA concentrations)

→ gene regulations

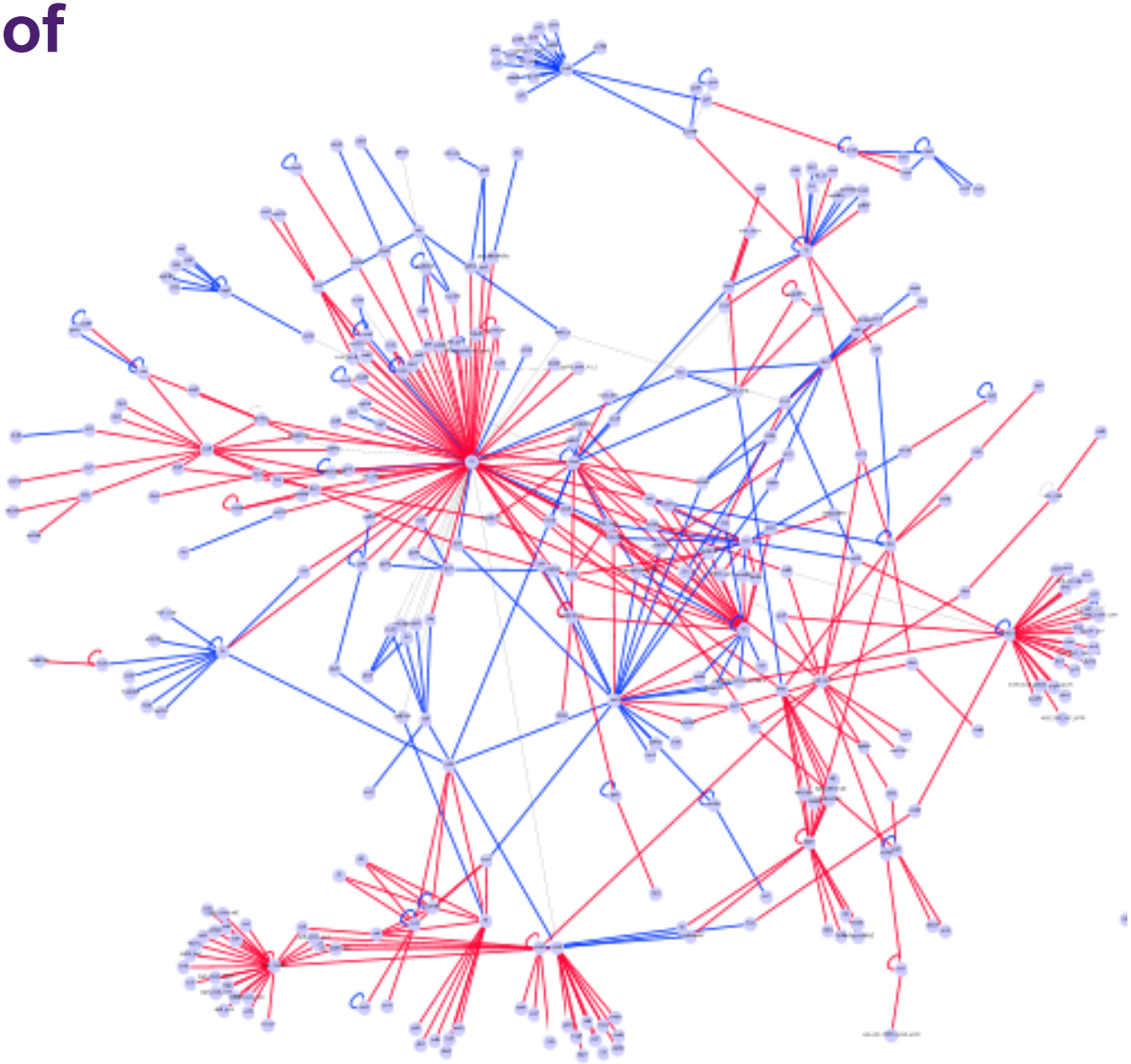
Biological motivation



→ gene expressions (mRNA concentrations)

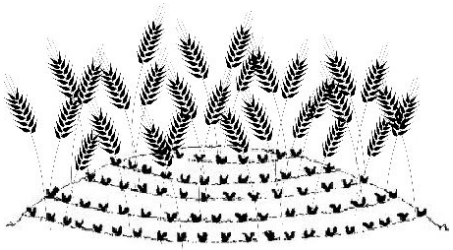
→ gene regulations

Goal : Reconstruction of gene regulatory network.

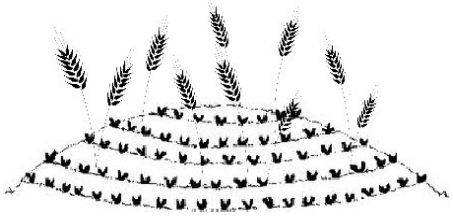


Escherichia coli
(423 genes, 578 regulations)
(SS. Shen-Orr and al., 2002)

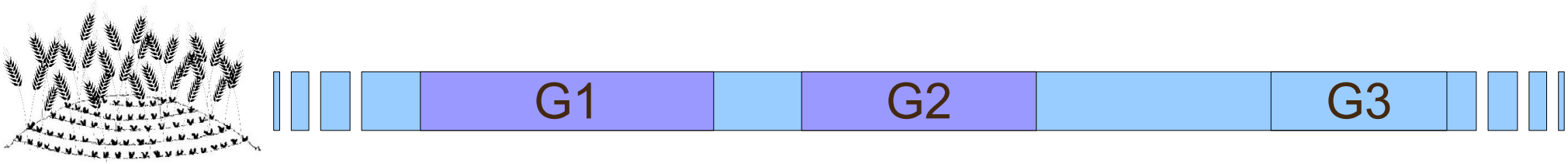
Polymorphism



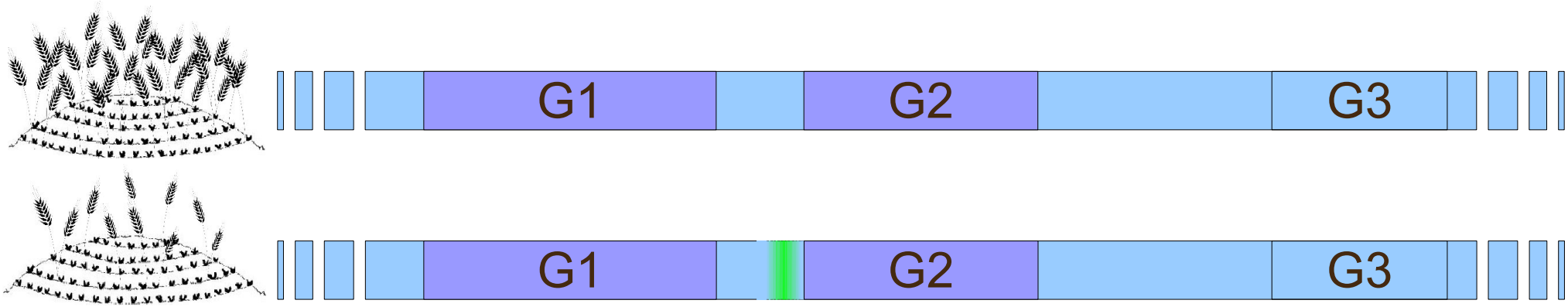




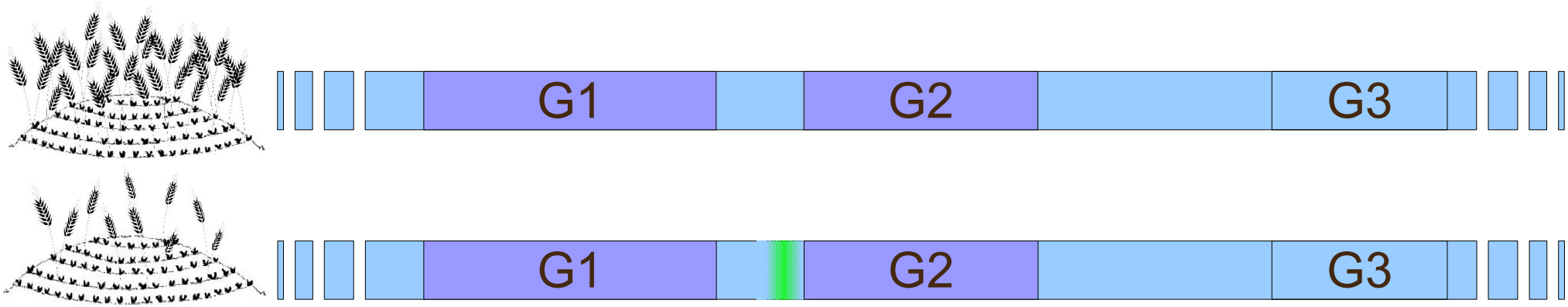
Polymorphism



Polymorphism

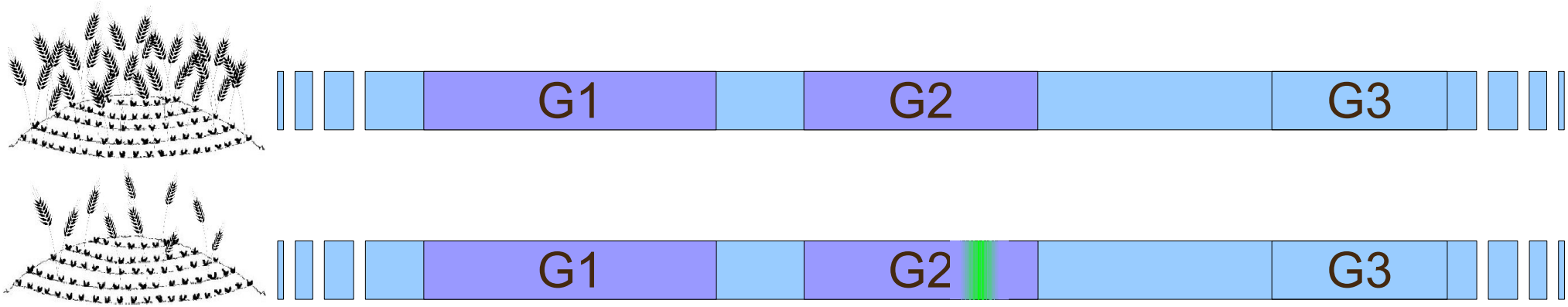


Polymorphism



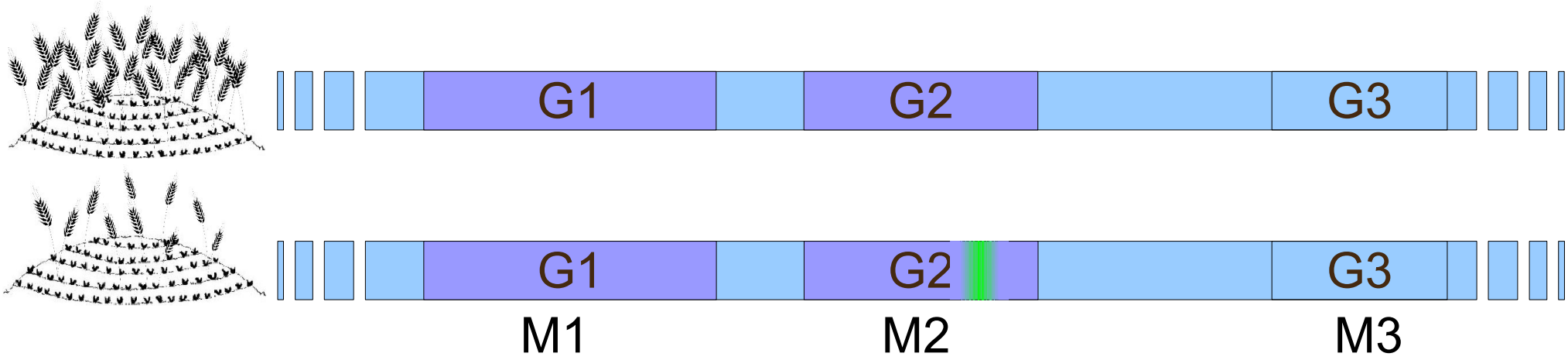
DNA mutations in genes : in promoter region → impact on its gene activity

Polymorphism



DNA mutations in genes: in promoter region → impact on its gene activity
in coding region → impact on others gene activities

Polymorphism

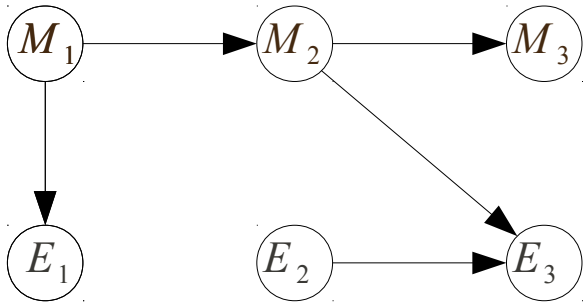


DNA mutations in genes: in promoter region → impact on its gene activity
in coding region → impact on others gene activities

Genetic data: one genetic marker (SNP) per gene

Discrete Bayesian Network

Directed acyclic graph G composed of n variables $X_i = \{E_i, M_i\}$

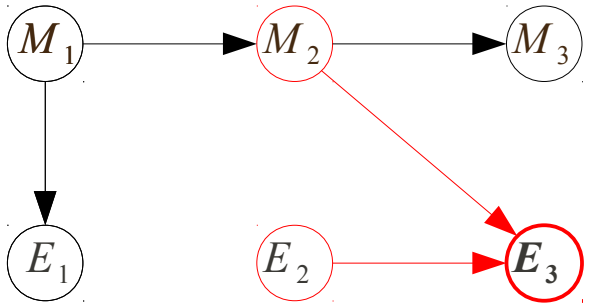


Genetic data

Gene expressions

Discrete Bayesian network

Directed acyclic graph G composed of n variables $X_i = \{G_i, M_i\}$



Genetic data

Gene expressions

Conditional distribution $P_G(E_3 / E_{2,} M_2)$

		E_3	$!E_3$
E_2	M_2	0.72	0.28
E_2	$!M_2$	0.59	0.41
$!E_2$	M_2	0.63	0.37
$!E_2$	$!M_2$	0.10	0.90

Graphical representation of a joint probability distribution:

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

Learning strategy

We look for the graph $G_{score} = argmax_{G_i} P(G_i / D)$ with dataset D.

$$P(G_i / D) = \frac{P(D / G_i) P(G_i)}{P(D)}$$

Learning strategy

We look for the graph $G_{score} = argmax_{G_i} P(G_i / D)$ with dataset D.

$$P(G_i/D) = \frac{P(D/G_i)\,P(G_i)}{P(D)}$$

$$\propto P(D/G_i)\,P(G_i)$$

- $P(D/G_i)$:marginal likelihood of *Gi*
- $P(G_i)$:prior probability of the graph G*i*
  → assumed to be uniform

Learning strategy

We look for the graph $G_{score} = argmax_{G_i} P(G_i / D)$ with dataset D.

$$P(G_i/D) = \frac{P(D/G_i)\,P(G_i)}{P(D)}$$

$$\propto P(D/G_i)\,P(G_i)$$

- $P(D/G_i)$:marginal likelihood of *Gi*
- $P(G_i)$:prior probability of the graph G*i*
  → assumed to be uniform

Objective function easy to evaluate and avoids over-fitting

- decomposable and penalized scores
  - BDeu score (***D.Heckerman*** *Machine learning 1995)*
  - *BIC score (**G.Schwartz** Annals of statistics 1978)*

Local search components

1. Search space
   - Directed Acyclic Graph
   - Partial DAG (PDAG)

   - variable orders

Local search components

1. Search space
   - Directed Acyclic Graph
   - Partial DAG (PDAG)
   - variable orders

2. Initial structure
   - empty structure
   - random structure
   - informed structure (MWST, expert...)

Local search components

1. Search space
   - Directed Acyclic Graph
   - Partial DAG (PDAG)
   - variable orders

2. Initial structure
   - empty structure
   - random structure
   - informed structure (MWST, expert...)

3. Neighborhood operators
   - addition of an edge
   - deletion of an edge
   - reversal of an edge
   - k look-ahead
   - optimal reinsertion

Local search components

1. Search space
   - Directed Acyclic Graph
   - Partial DAG (PDAG)
   - variable orders

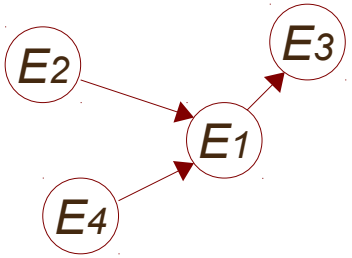
2. Initial structure
   - empty structure
   - random structure
   - informed structure (MWST, expert...)

3. Neighborhood operators
   - addition of an edge
   - deletion of an edge
   - reversal of an edge
   - k look-ahead
   - optimal reinsertion

4. Meta-heuristics
   - hill climbing
   - tabu search
   - simulated annealing
   - MCMC
   - genetic algorithms
   - ...

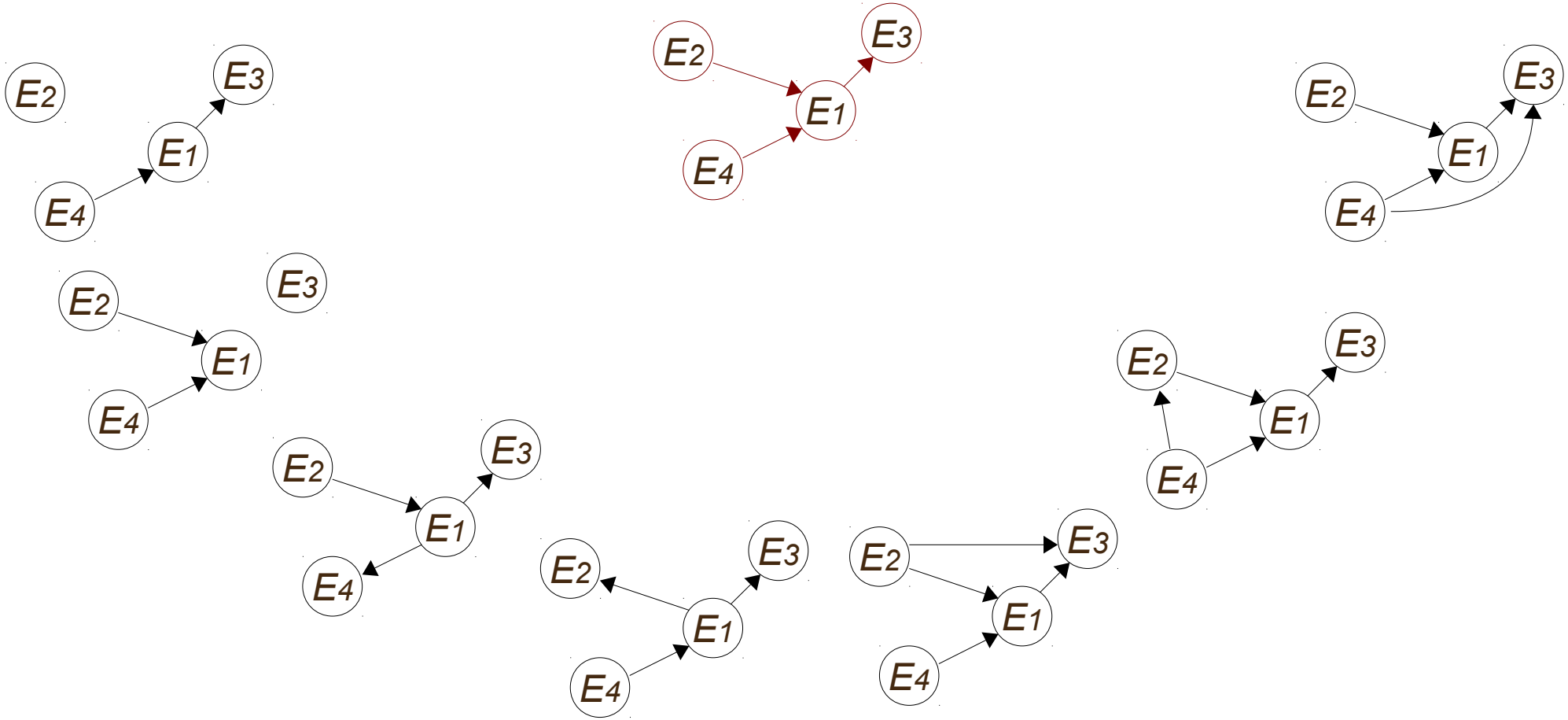
Hill Climbing Algorithm

- **Greedy search**
  - *Start with an initial network (empty graph, a priori graph)*
  - *Score all possible local modifications (addition / deletion / reversal of one edge) and select the best of them (if it exist)*



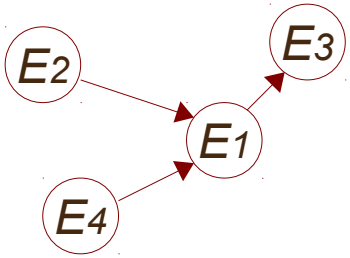
Hill Climbing Algorithm

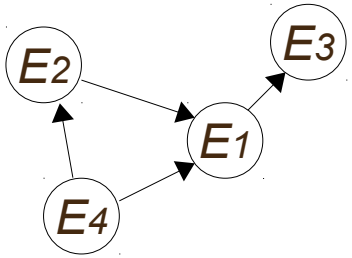
- **Greedy search**
  - *Start with an initial network (empty graph, a priori graph)*
  - *Score all possible local modifications (addition / deletion / reversal of one edge) and select the best of them (if it exist)*



Hill Climbing Algorithm

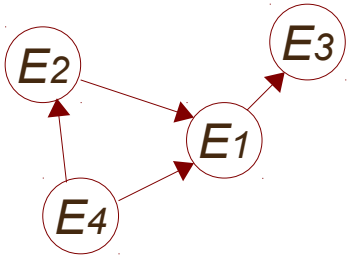
- **Greedy search**
  - *Start with an initial network (empty graph, a priori graph)*
  - *Score all possible local modifications (addition / deletion / reversal of one edge) and select the best of them (if it exist)*





Hill Climbing Algorithm

- **Greedy search**
  - *Start with an initial network (empty graph, a priori graph)*
  - *Score all possible local modifications (addition / deletion / reversal of one edge) and select the best of them (if it exist)*



Arabidopsis thaliana

- ID
  - Genome size : 135 Mb (estimated)
  - Protein coding genes 25000+
  - Chromosomes 5

Arabidopsis thaliana

- ID
  - Genome size : 135 Mb (estimated)
  - Protein coding genes 25000+
  - Chromosomes 5

- Experimental data (*Simon et al., 2008*)
  - RIL Population 158 individuals (Cvi/Col)
  - CATMA chip 34660 probes (22089 genes)
  - SNP markers 89

  → eQTL analysis detects 5035 probes with genetic response (LOD>2)

Data preprocessing

- Missing expression values

  · Missing value of gene m for individual i

  $$D^i_{m/N} = \mu_{E_m} + Cov_{E_m, E_N} Cov^{-1}_{E_N} (D^i_{E_N} - \mu_{E_N})$$

  with N set from 10 best predictor genes (without missing values)

Data preprocessing

- Missing expression values

  · Missing value of gene m for individual i

  $$D^i_{m/N} = \mu_{E_m} + Cov_{E_m, E_N} Cov^{-1}_{E_N} (D^i_{E_N} - \mu_{E_N})$$

  with N set from 10 best predictor genes (without missing values)

- Marker inference (with « *qtl* » R package)

  · Filling in missing genotypes

  · Pseudo markers inference (1cM equally-spaced)

  → 590 markers (including 89 reals)

Data preprocessing

- Missing expression values

  · Missing value of gene m for individual i

  $$D^i_{m/N} = \mu_{E_m} + Cov_{E_m, E_N} Cov^{-1}_{E_N} (D^i_{E_N} - \mu_{E_N})$$

  with N set from 10 best predictor genes (without missing values)

- Marker inference (with « *qtl* » R package)

  · Filling in missing genotypes

  · Pseudo markers inference (1cM equally-spaced)

  → 590 markers (including 89 reals)

- Expressions and pseudo markers discretization (max 4 classes)

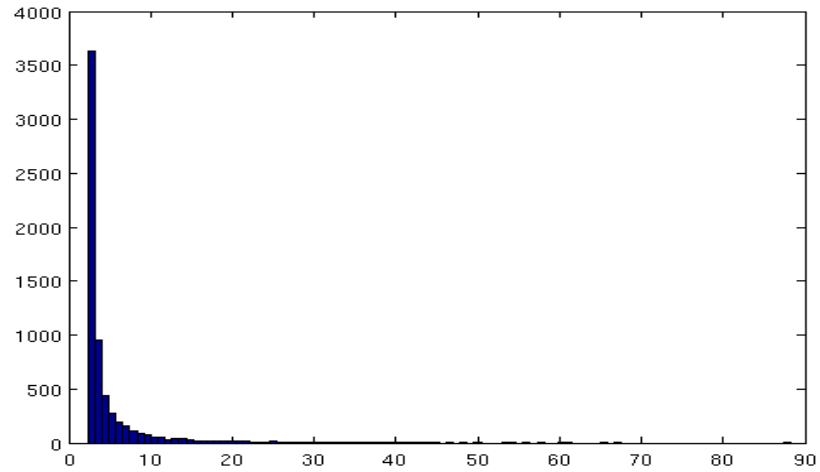
Data selection

- Expression selection

→ 34660 CATMA probes

→ eQTL detected for 5035 probes

→ LOD >= 2.5 for 4176 probes



LOD score of detected eQTLs

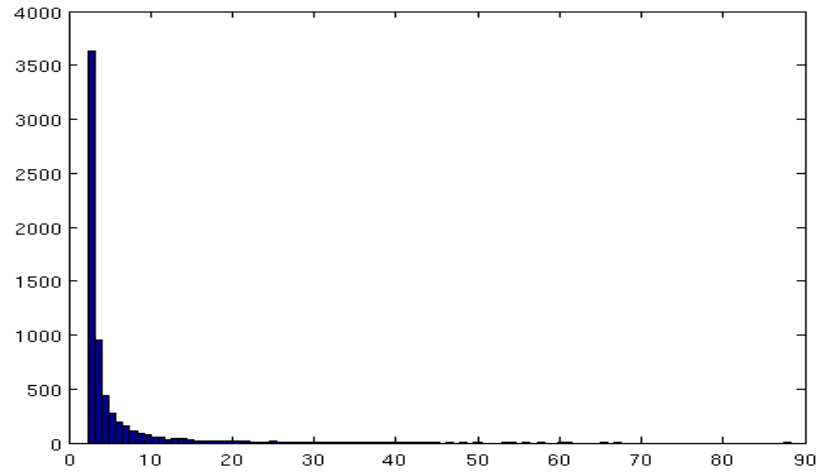
Data selection

- Expression selection

  → 34660 CATMA probes

  → eQTL detected for 5035 probes

  → LOD >= 2.5 for 4176 probes



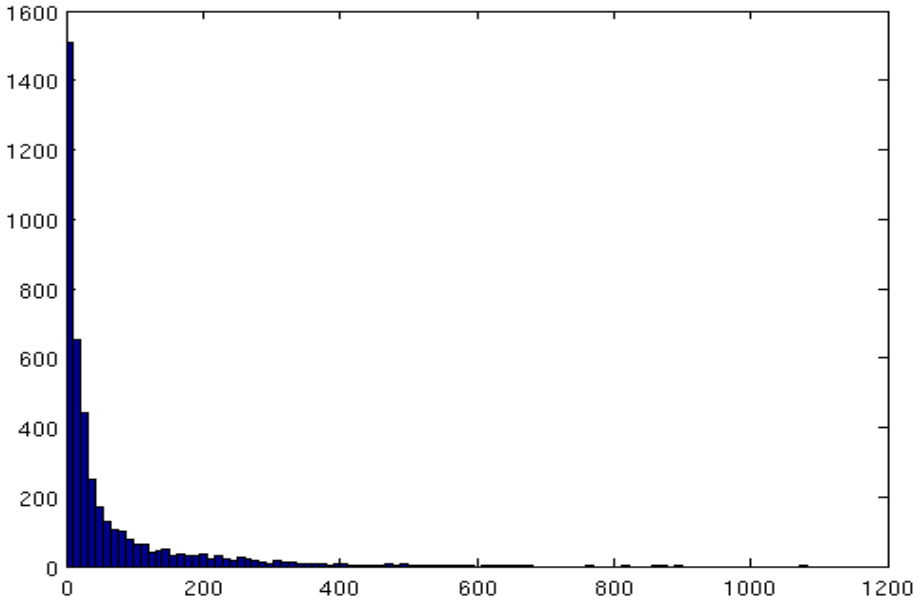
LOD score of detected eQTLs

- No marker restriction

  → 590 markers

Hill climbing restriction

- Limit on the number of parents during the search: 4
- Pre-filtering candidate parents under 2 conditions
  - $score_{BDeu}(E_i / Parent) > score_{BDeu}(E_i)$
  - only one best marker in a sliding window of 10 cM



Number of candidate parents per expression variable

Results

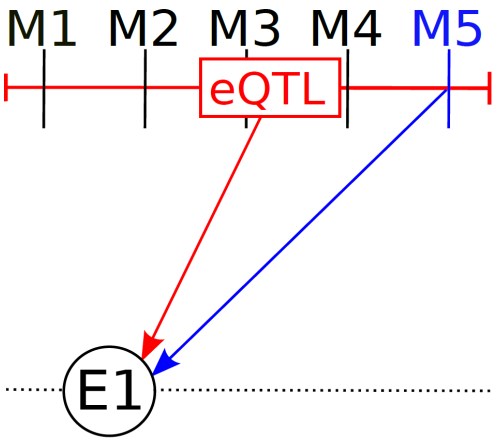
- Learnt Bayesian network
  - 4766 variables
  - 6137 edges (284 $M_i \rightarrow E_j$ / 5853 $E_i \rightarrow E_j$)
- In-Out degrees

Number	0	1	2	3	4	5	6	7	8	9	10+
$\rightarrow$ E	263	2008	1627	237	41	-	-	-	-	-	-
E $\rightarrow$	2164	844	465	260	118	98	58	46	26	29	68
M $\rightarrow$	457	67	29	15	9	7	2	1	3	-	-

(max 65)

Comparison with eQTL analysis

- Situation description
  - cis-eQTL detected for E1

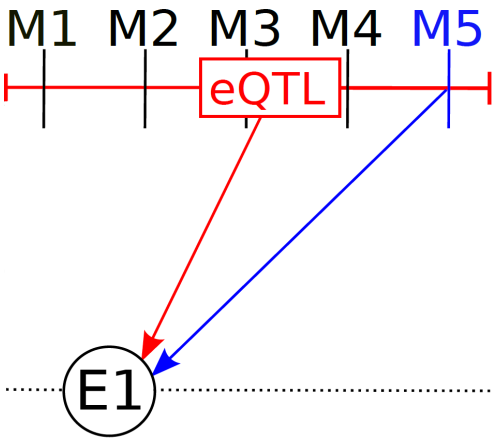


Case 1

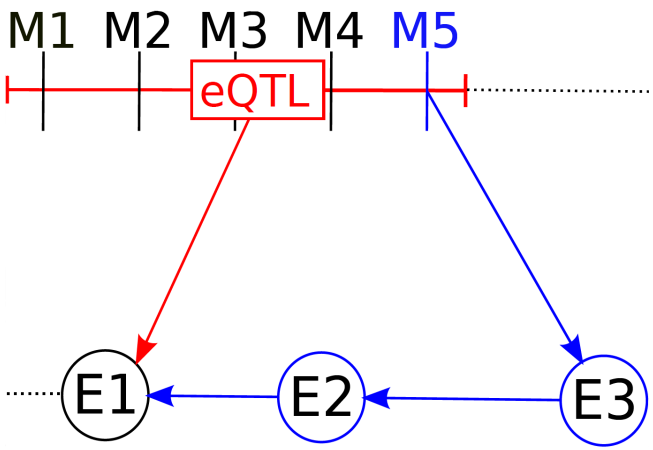
→ Margin of error: 2 cM

Comparison with eQTL analysis

- Situation description
  - cis-eQTL detected for E1



Case 1

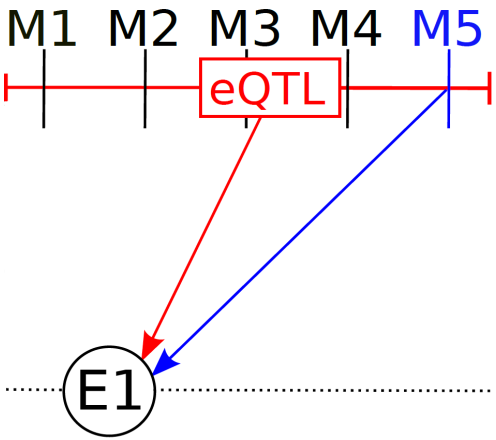


Case 2

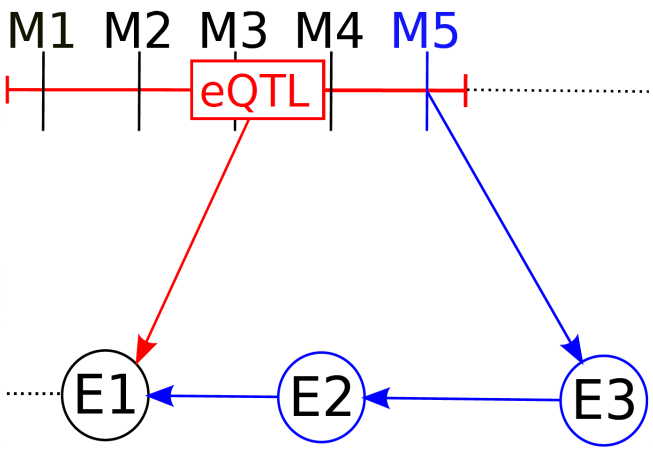
→ Margin of error: 2 cM

Comparison with eQTL analysis

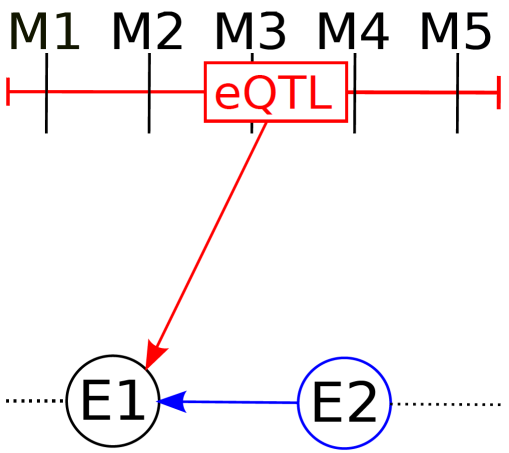
- Situation description
  - cis-eQTL detected for E1



Case 1



Case 2

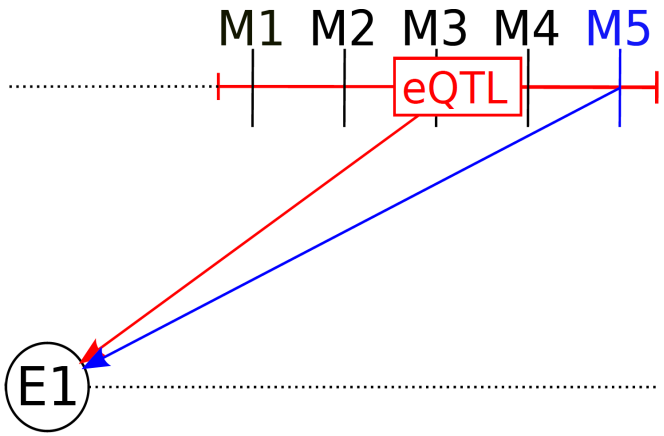


Case 3

→ Margin of error: 2 cM

Comparison with eQTL analysis

- Situation description
  - trans-eQTL detected for E1

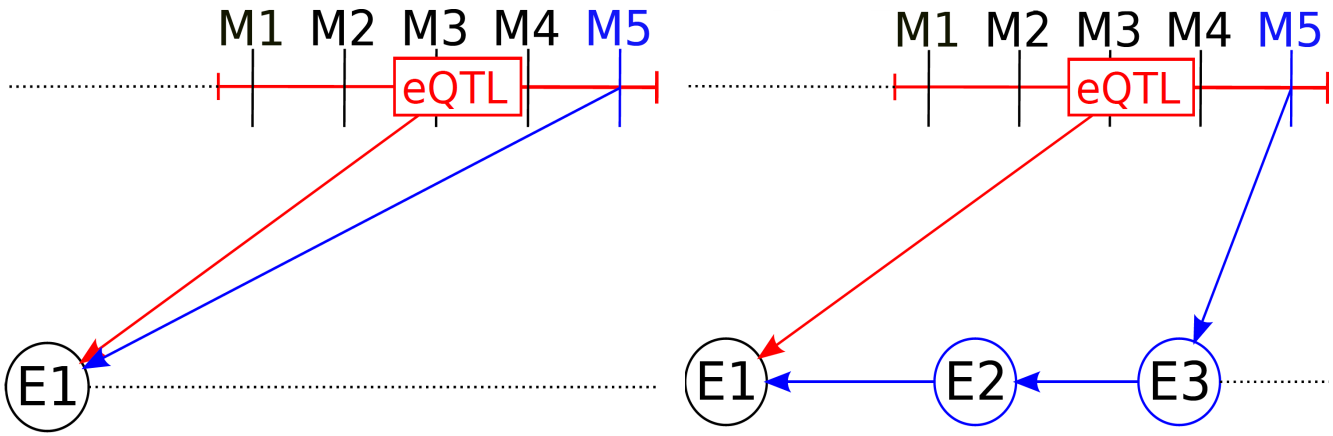


Case 4

→ Margin of error : 2 cM

Comparison with eQTL analysis

- Situation description
  - trans-eQTL detected for E1

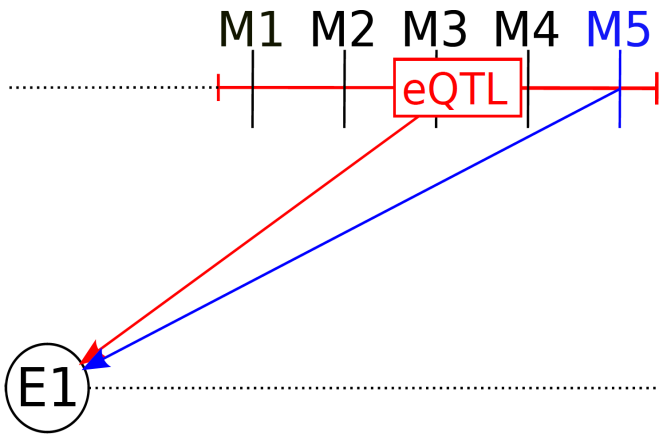


Case 4 Case 5

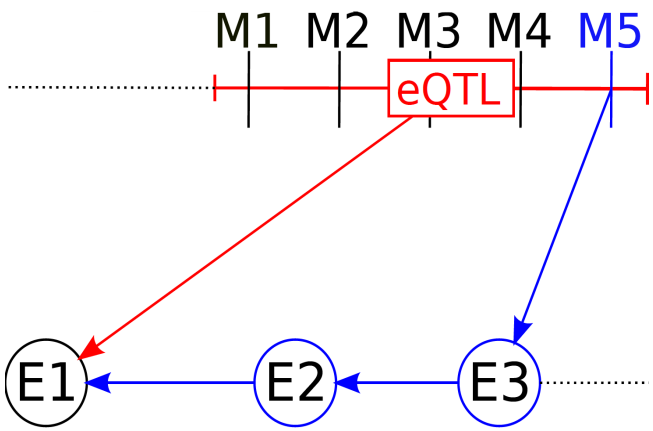
→ Margin of error : 2 cM

Comparison with eQTL analysis

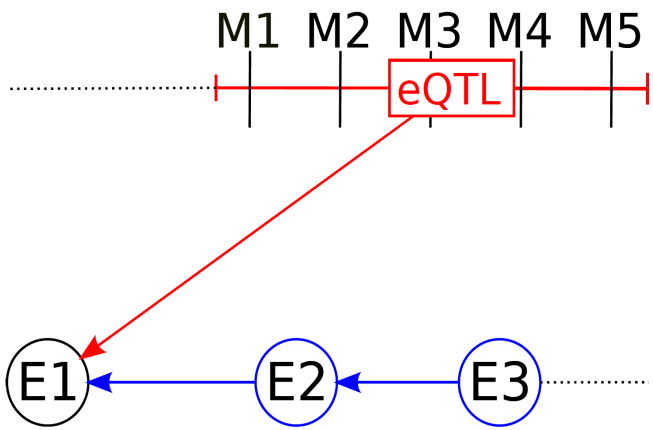
- Situation description
  - trans-eQTL detected for E1



Case 4



Case 5



Case 6

→ Margin of error : 2 cM

Comparison with eQTL analysis

- Analysis on the most significant eQTLs (FDR 1%)

  → 1269 eQTLs

	Case 1	Case 2	Case 3	**Tot-Cis**	Case 4	Case 5	Case 6	**Tot-Trans**
BN	202	435	243	**880**	68	161	69	**298**
eQTL				**938**				**331**

→ Margin of error : 2 cM

Datamining validation

- Select first 200 edges with highest BDeu score improvement in Hill Climbing
- Keep only 126 edges corresponding to interactions between two different chromosomes
- Look at GO biological processes enrichments using Genomatix Pathway System

Datamining validation

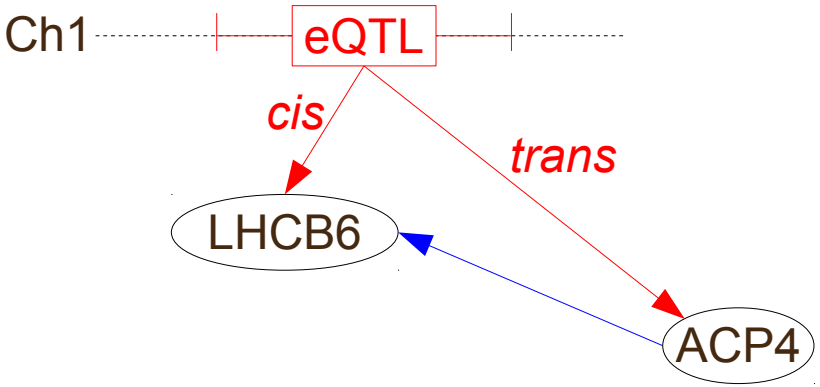
- Select first 200 edges with highest BDeu score improvement in Hill Climbing
- Keep only 126 edges corresponding to interactions between two different chromosomes
- Look at GO biological processes enrichments using Genomatix Pathway System
- The 7th edge : AT4G25050 → AT1G15820
  - AT4G25050 (ACP4) : response to light stimulus
  - AT1G15820 (LHCB6): photosynthesis

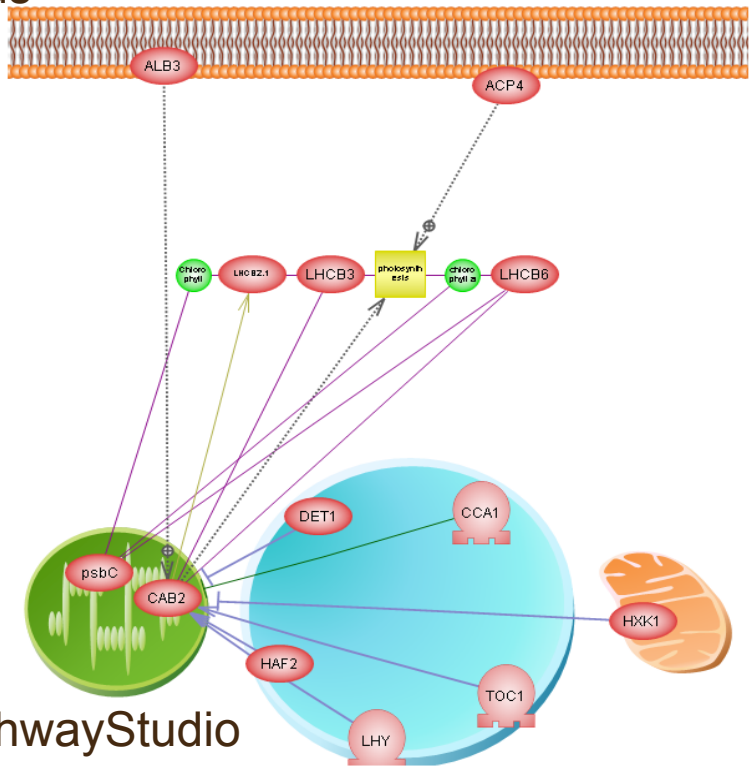
  (PNAS May 5, 2009 vol. 106 no. 18)

Datamining validation

- Select first 200 edges with highest BDeu score improvement in Hill Climbing
- Keep only 126 edges corresponding to interactions between two different chromosomes
- Look at GO biological processes enrichments using Genomatix Pathway System
- The 7th edge : AT4G25050 → AT1G15820
  - AT4G25050 (ACP4) : response to light stimulus
  - AT1G15820 (LHCB6): photosynthesis

(PNAS May 5, 2009 vol. 106 no. 18)





PathwayStudio

Datamining validation

- The far far away interaction (572th) : AT3G56400 → AT2G14560
  found in common with a list of known interactions (Pathway Studio Demo)

  · AT3G56400 (WRKY70) : defence response to fungus

  · AT2G14560 (LURP1): defence response to fungus

  (Plant J. 2008 Jul;55(1):53-64)

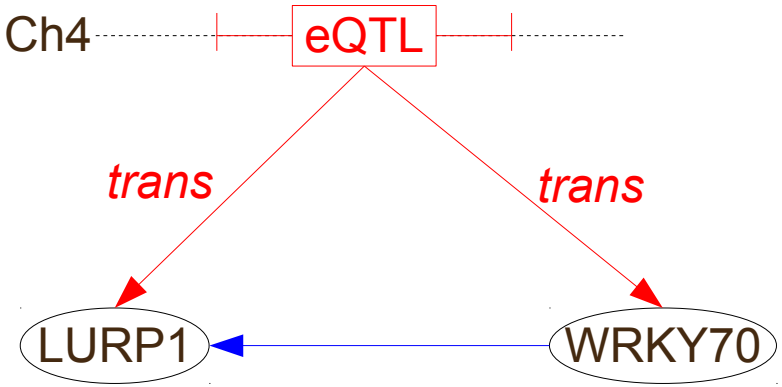
Datamining validation

- The far far away interaction (572th) : AT3G56400 → AT2G14560 found in common with a list of known interactions (Pathway Studio Demo)

  · AT3G56400 (WRKY70) : defence response to fungus

  · AT2G14560 (LURP1): defence response to fungus

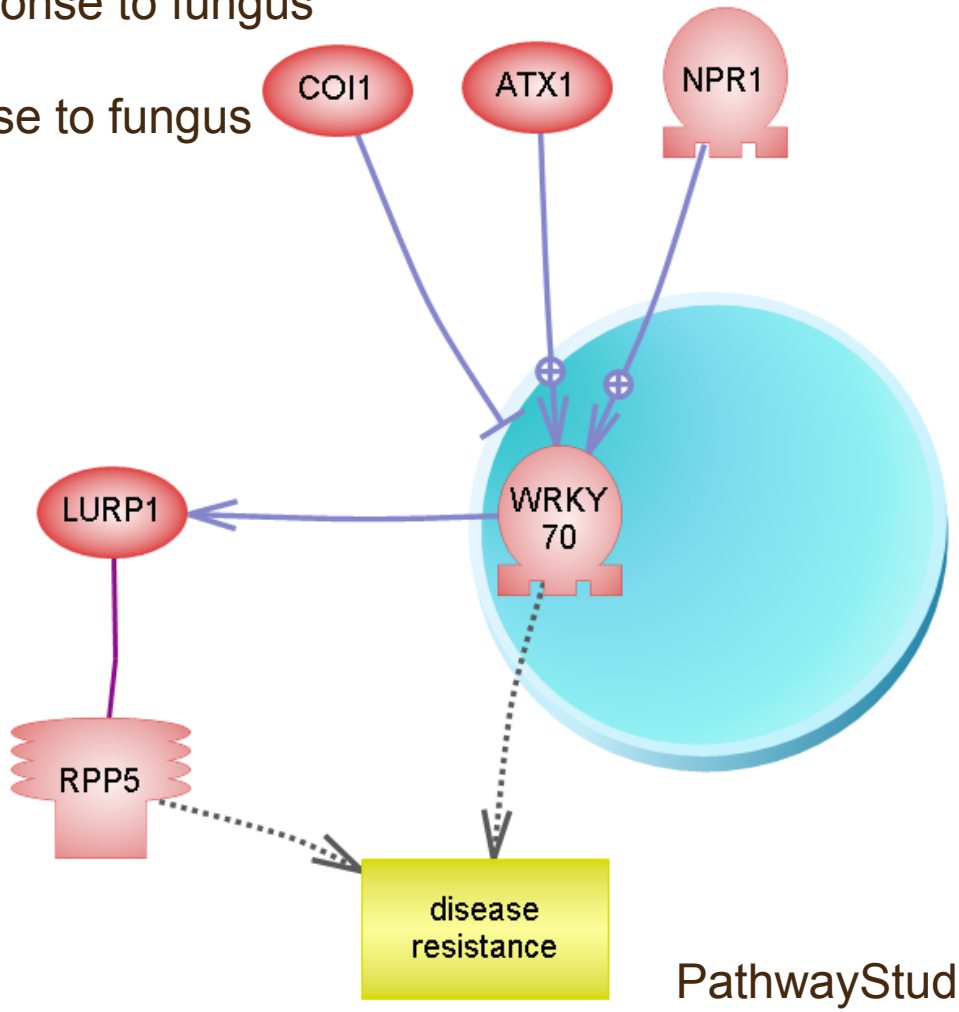
  (Plant J. 2008 Jul;55(1):53-64)



RPP5 ?

→ Chromosome 4

→ in confidence interval



PathwayStudio

Perspectives

- improve robustness with bootstrap strategy
- try new local operator (swap*)
- use multiple discretizations policy
- combine several methods in meta analysis (Lasso, Dantzig)
- validate interactions with text-mining...?