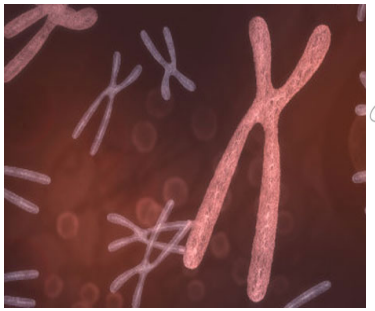
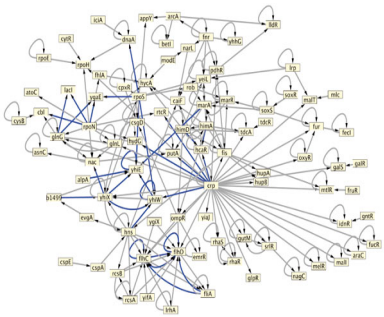
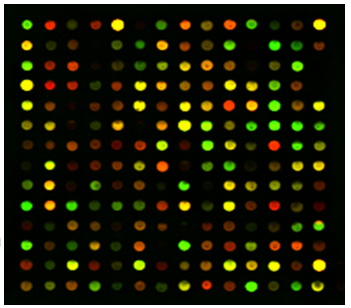
Extended Bayesian scores for reconstructing gene regulatory networks.

Jimmy Vandel

Simon de Givry & Brigitte Mangin







Université Paul Sabatier
TOULOUSE III

ECCS'10
September 16th, 2010

INRA
Institut National de la Recherche Agronomique

OUTLINES :

→ Biological background

→ State of the art

→ Score-based structure learning

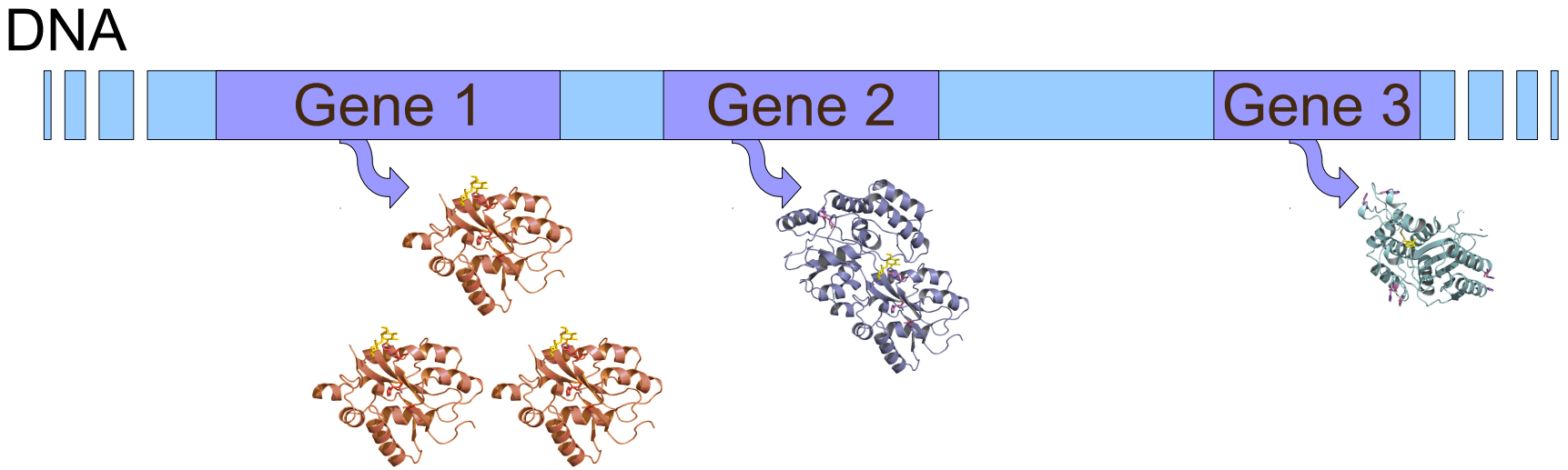
→ Model with extra biological knowledge

→ Results

→ Conclusion and perspectives

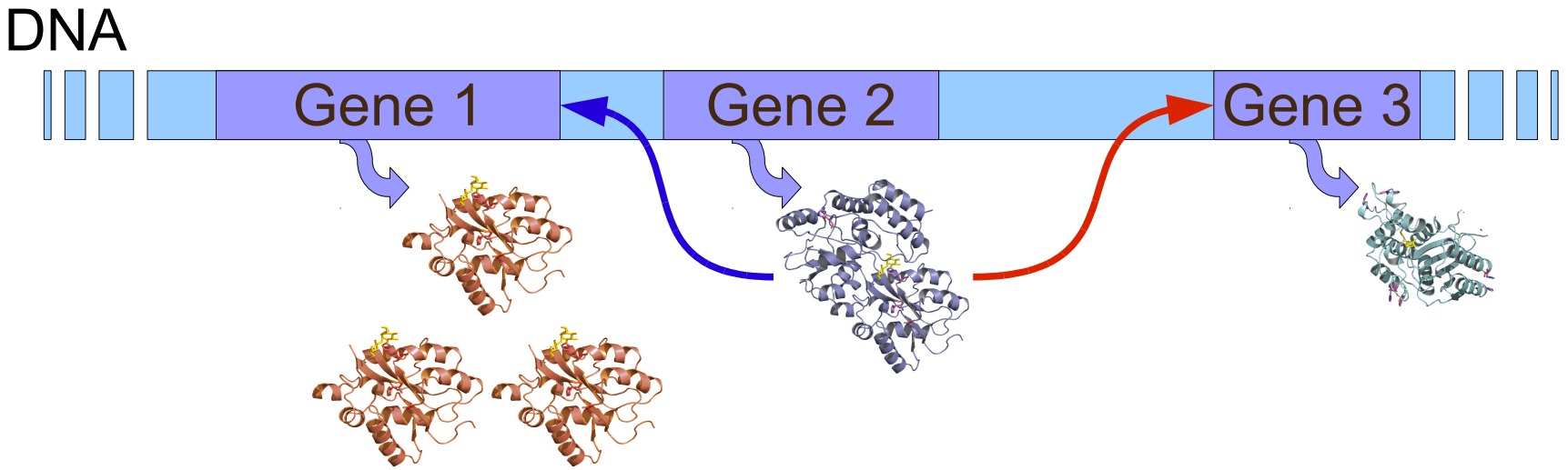
Central dogma of molecular biology

Central dogma of molecular biology



→ gene activities (mRNA concentrations)

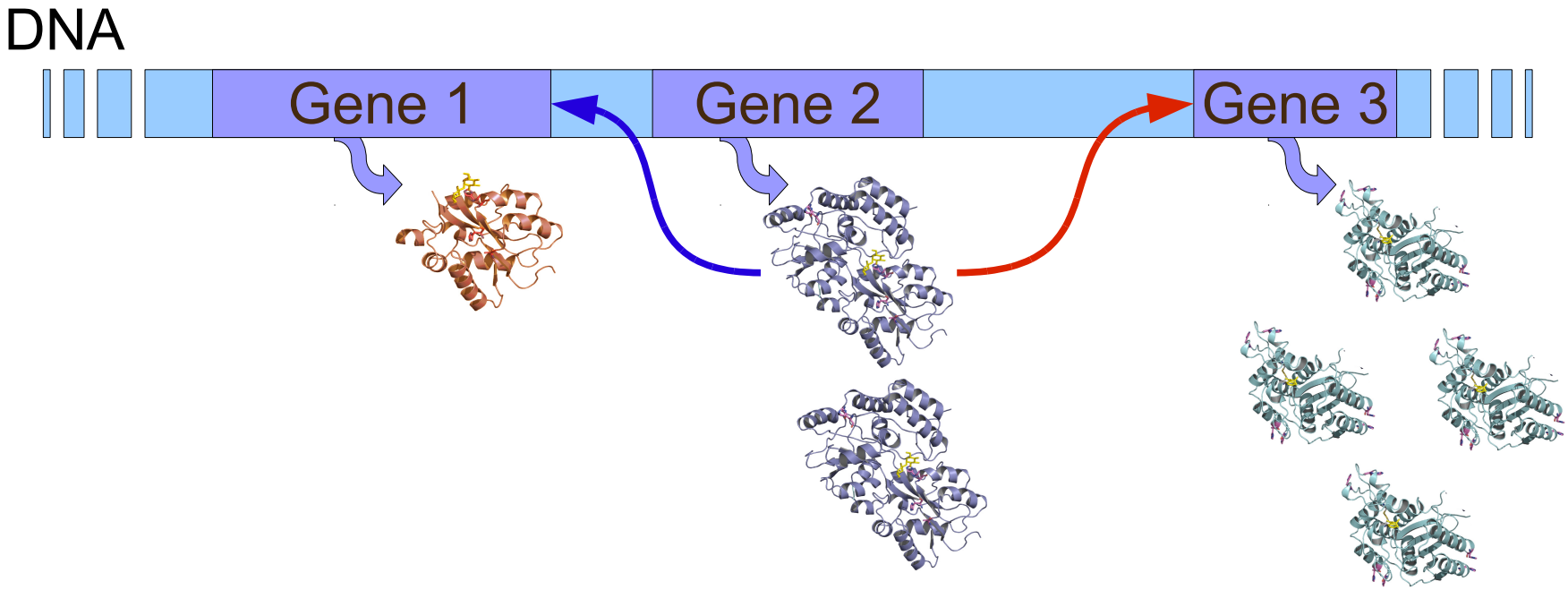
Central dogma of molecular biology



→ gene activities (mRNA concentrations)

→ gene regulations

Central dogma of molecular biology

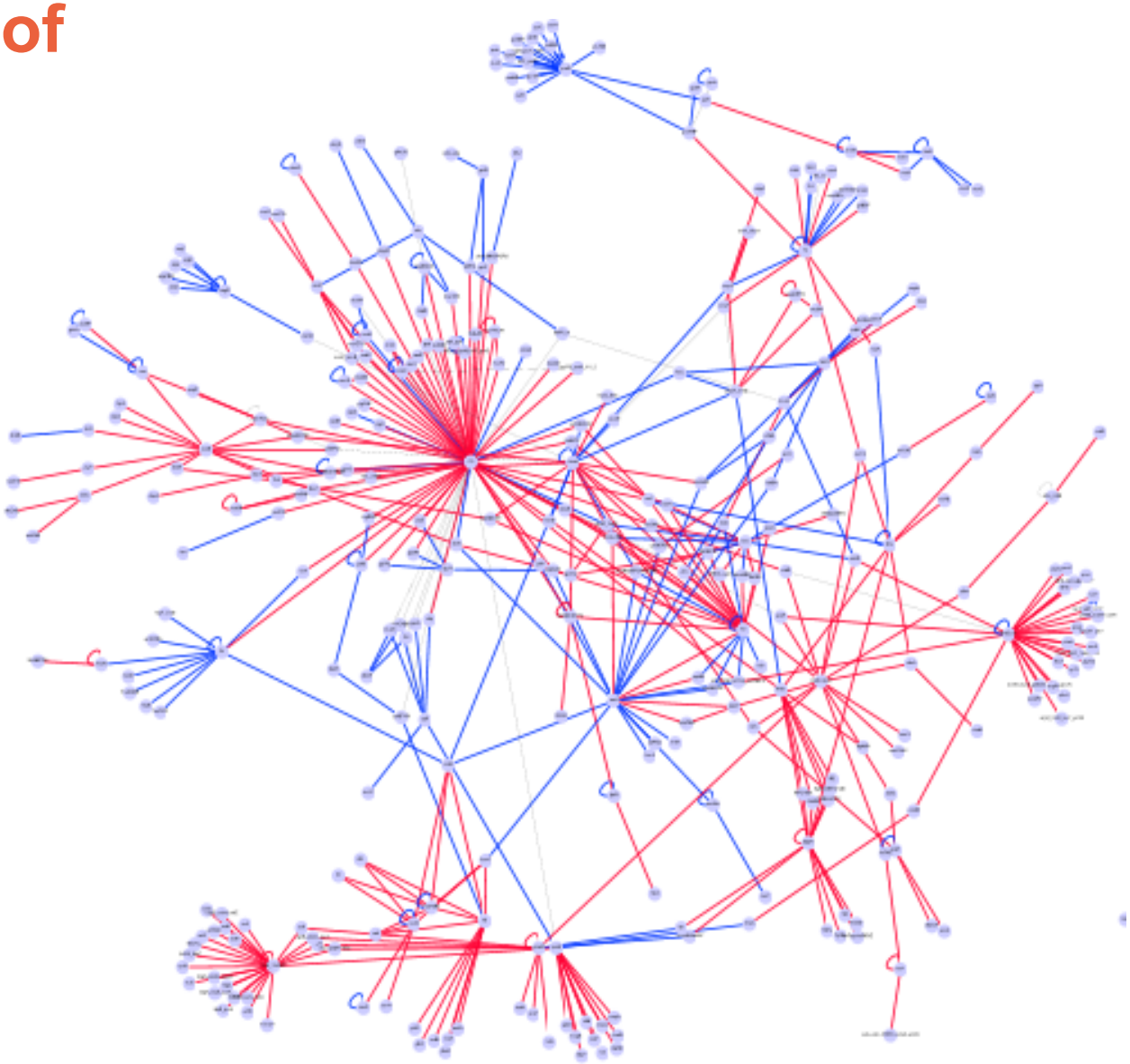


→ gene activities (mRNA concentrations)

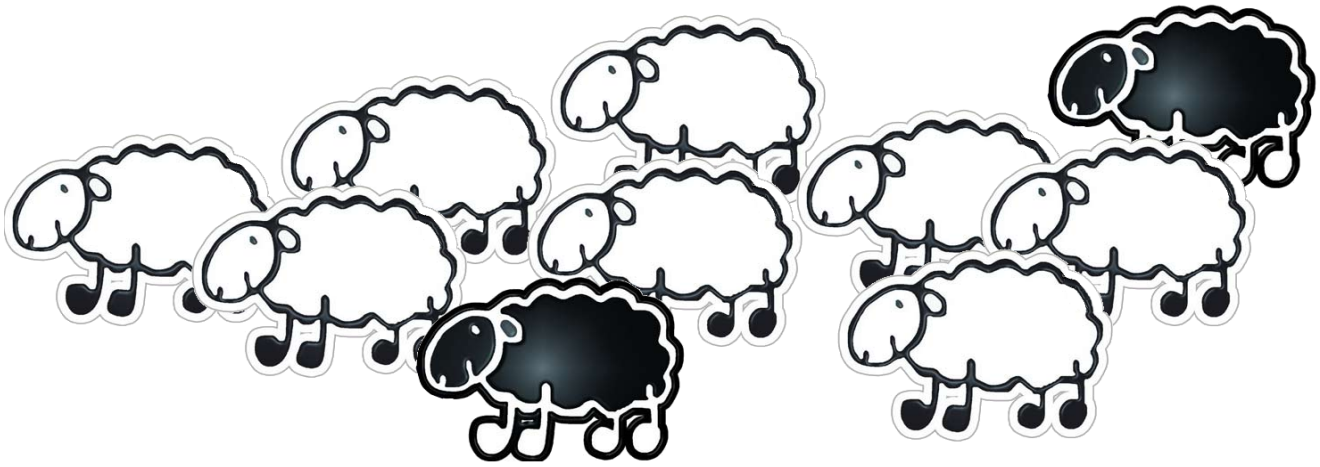
→ gene regulations

Goal : Reconstruction of gene regulatory network.

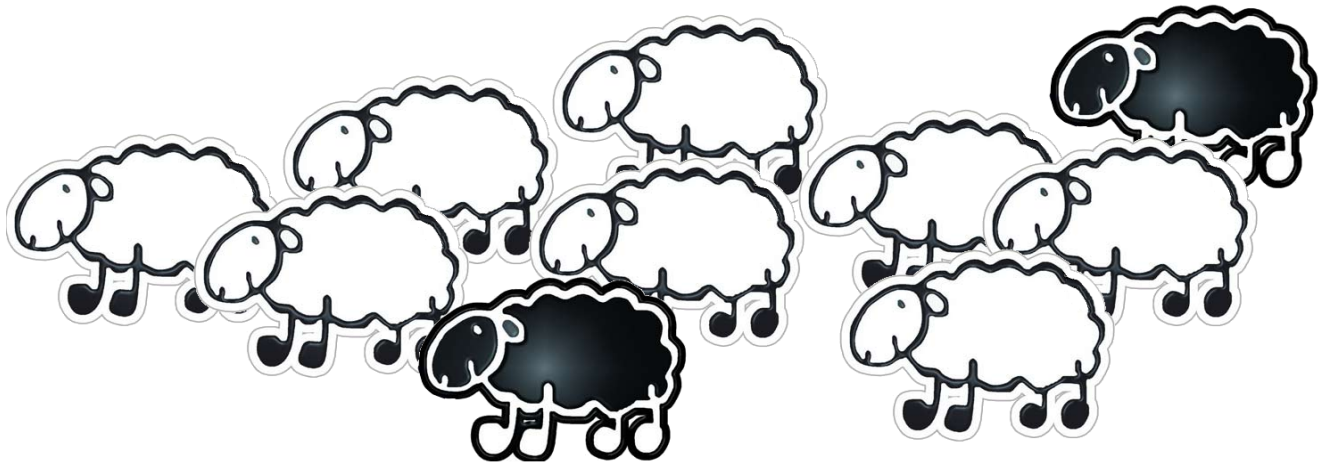
Escherichia coli
(423 genes
578 regulations)

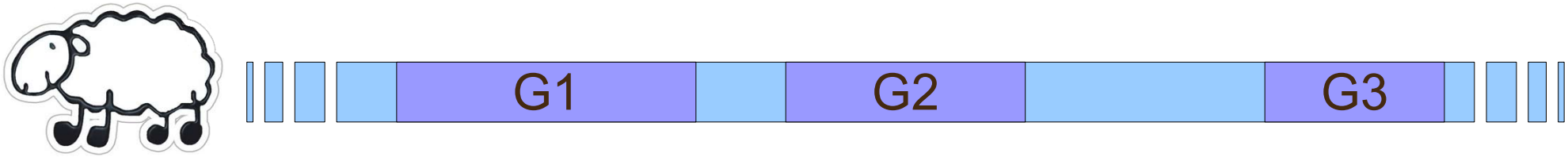


Polymorphism

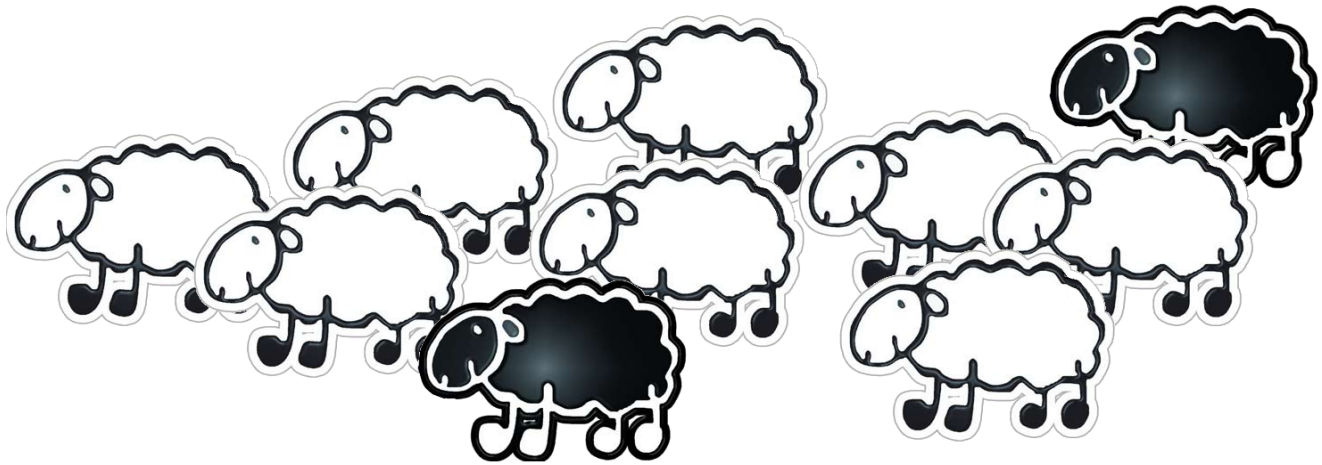


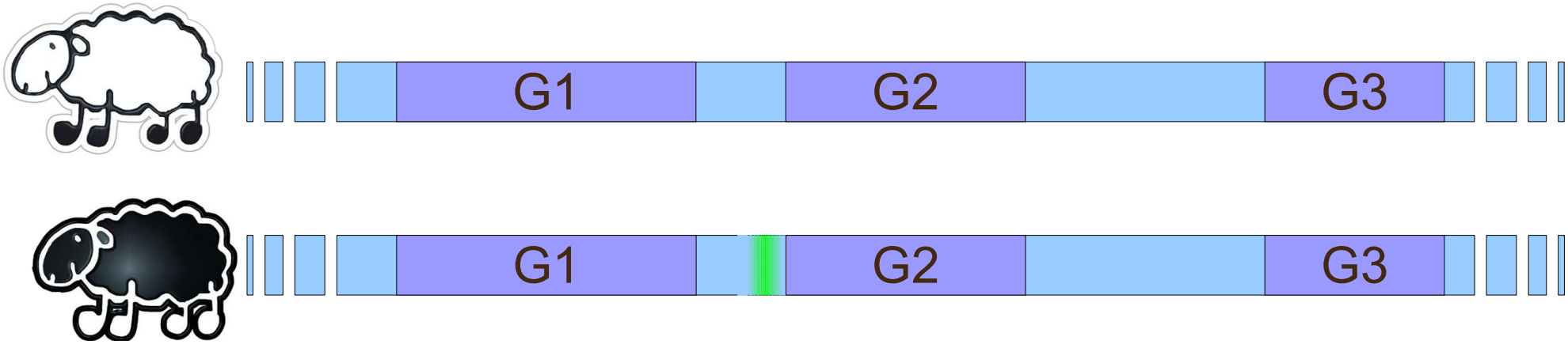
Polymorphism





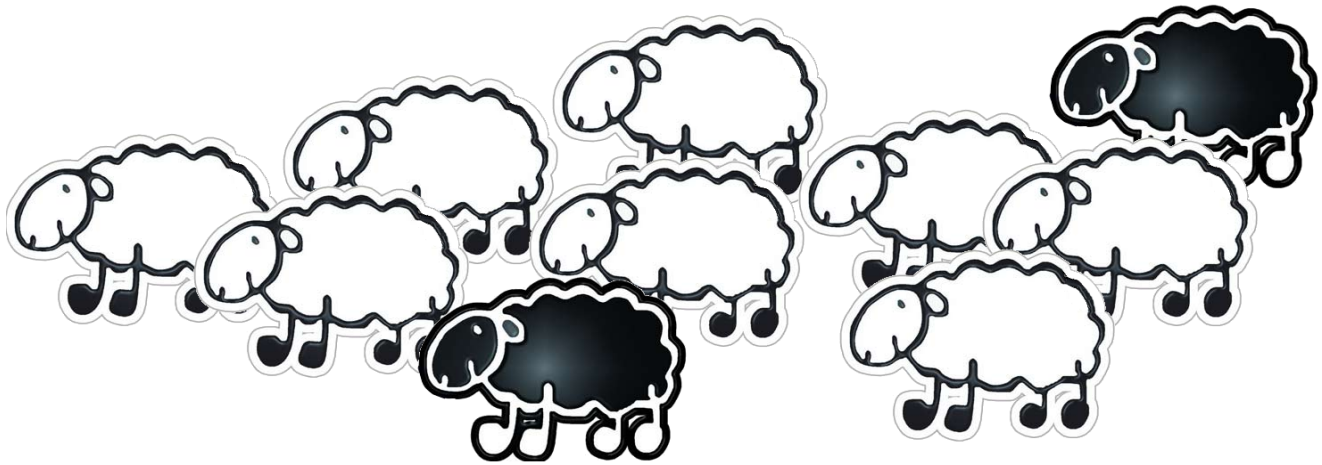
Polymorphism

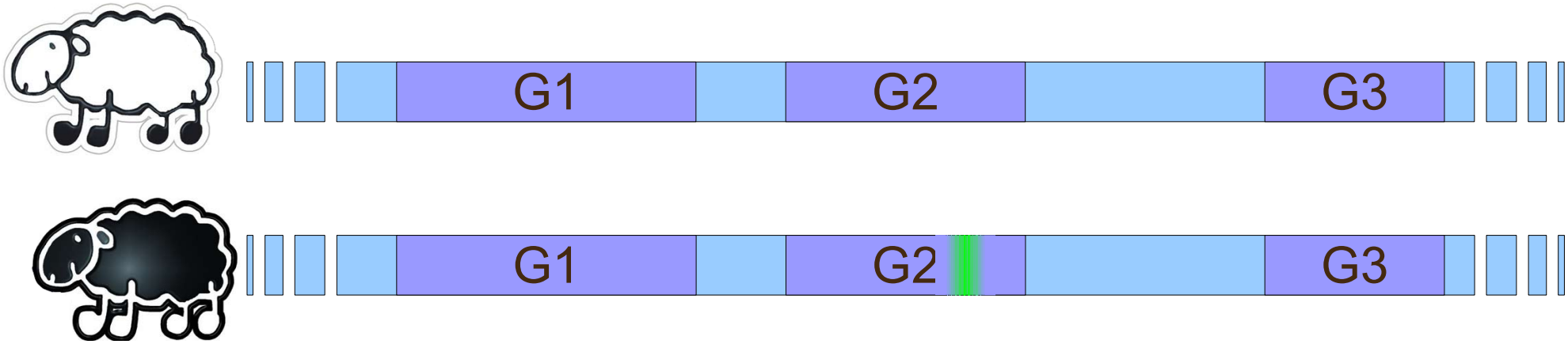




→ DNA mutations in genes - in promoter region (impact on gene activity)

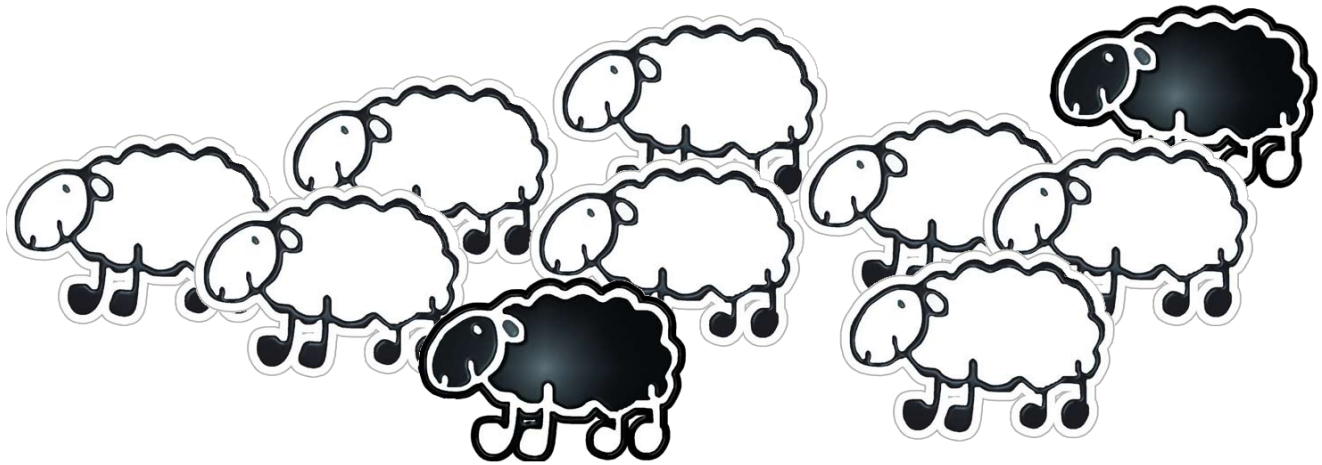
Polymorphism

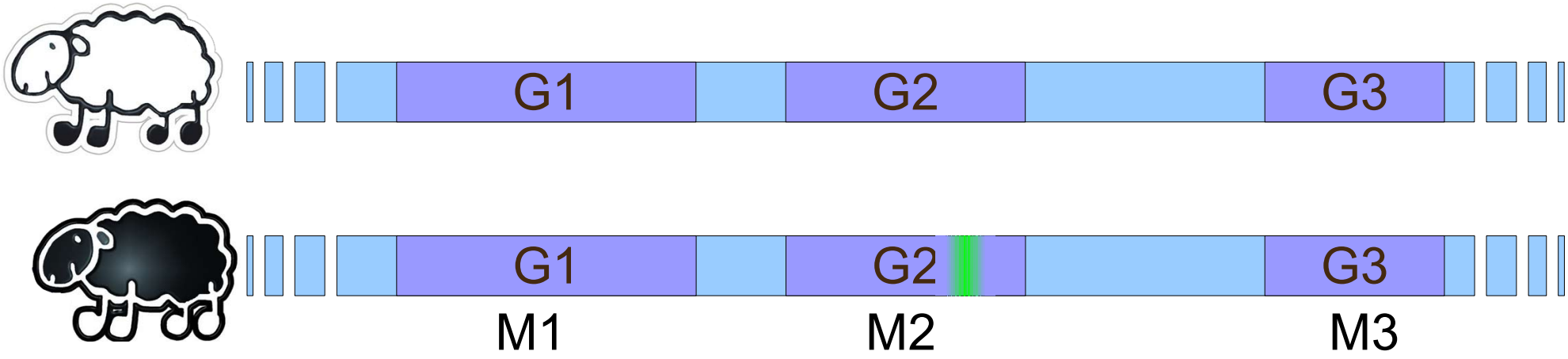




→ DNA mutations in genes - in promoter region (impact on gene activity)
- in coding region (modify protein structure)

Polymorphism





→ DNA mutations in genes - in promoter region (impact on gene activity)
- in coding region (modify protein structure)

→ observable through one genetic marker for each gene

State of the art of gene regulatory network inference

State of the art of gene regulatory network inference

Pairwise correlation model

- Correlations measurement + threshold

State of the art of gene regulatory network inference

Pairwise correlation model

Correlations measurement + threshold

→ Mutual Information (ARACNE *Margolin* 2006, CLR *Faith* 2007)

State of the art of gene regulatory network inference

Pairwise correlation model

Correlations measurement + threshold

→ Mutual Information (ARACNE *Margolin* 2006, CLR *Faith* 2007)

→ Pearson / Spearman partial correlation (ParCorA *de la Fuente* 2004)

State of the art of gene regulatory network inference

Pairwise correlation model

Correlations measurement + threshold

→ Mutual Information (ARACNE *Margolin* 2006, CLR *Faith* 2007)

→ Pearson / Spearman partial correlation (ParCorA *de la Fuente* 2004)

Linear model

Graphical Gaussian Models (GGM)

State of the art of gene regulatory network inference

Pairwise correlation model

Correlations measurement + threshold

→ Mutual Information (ARACNE *Margolin* 2006, CLR *Faith* 2007)

→ Pearson / Spearman partial correlation (ParCorA *de la Fuente* 2004)

Linear model

Graphical Gaussian Models (GGM)

→ Global search (GeneNet *Schäfer* 2005, SIMoNe *Chiquet* 2008)

State of the art of gene regulatory network inference

Pairwise correlation model

Correlations measurement + threshold

→ Mutual Information (ARACNE *Margolin* 2006, CLR *Faith* 2007)

→ Pearson / Spearman partial correlation (ParCorA *de la Fuente* 2004)

Linear model

Graphical Gaussian Models (GGM)

→ Global search (GeneNet *Schäfer* 2005, SIMoNe *Chiquet* 2008)

→ Local regressions (SEM Lasso *Liu 2008*)

State of the art of gene regulatory network inference

Pairwise correlation model

Correlations measurement + threshold

→ Mutual Information (ARACNE *Margolin* 2006, CLR *Faith* 2007)

→ Pearson / Spearman partial correlation (ParCorA *de la Fuente* 2004)

Linear model

Graphical Gaussian Models (GGM)

→ Global search (GeneNet *Schäfer* 2005, SIMoNe *Chiquet* 2008)

→ Local regressions (SEM Lasso *Liu 2008*)

→ Boosting strategy (GGMSelect *Giraud* 2008)

(Spearman correlation+GGM)

State of the art of gene regulatory network inference

Pairwise correlation model

Correlations measurement + threshold

→ Mutual Information (ARACNE *Margolin* 2006, CLR *Faith* 2007)

→ Pearson / Spearman partial correlation (ParCorA *de la Fuente* 2004)

Linear model

Graphical Gaussian Models (GGM)

→ Global search (GeneNet *Schäfer* 2005, SIMoNe *Chiquet* 2008)

→ Local regressions (SEM Lasso *Liu 2008*)

→ Boosting strategy (GGMSelect *Giraud* 2008)

(Spearman correlation+GGM)

Probabilistic discrete graphical model

→ Bayesian networks on discrete data (*Friedman* 2000, *Zhu* 2007)

Bayesian network

- Directed acyclic graph G composed of n variables X_i with domain size r_i

- Conditional distribution for variable X_i, given its parents Pa_i in G:

$$P_G(X_i/Pa_i^j) = \theta_{ij}$$

- Representation of a joint probability distribution :

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

- We note $Dim(G) = \sum_{i=1}^{n} (r_i - 1) * q_i$ the dimension of the network with $q_i = \prod_{Pa_i^j} r_j$

Score based learning

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

Score based learning

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

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

Score based learning

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

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

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

Score based learning

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 G*i*
  - exact under → hypothesis **B***ayesian* ***D****irichlet score*
  - estimation → ***B****ayesian* ***I****nformation* ***C****riterion* score

Score based learning

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 G*i*
  - exact under → hypothesis **B***ayesian* ***D****irichlet score*
  - estimation → ***B****ayesian* ***I****nformation* ***C****riterion* score
- $P(G_i)$:prior probability of the graph G*i*
  → assumed to be uniform

Scores

Scores

- BD Score *(**D.Heckerman** Machine learning 1995)*
  - A priori on conditional probability θ_{ij} following a Dirichlet distribution with parameter α_{ijk}

$$BD(G) = \prod_i^n \prod_j^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_k^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}$$

Scores

- BD Score *(**D.Heckerman** Machine learning 1995)*
  - A priori on conditional probability θ_{ij} following a Dirichlet distribution with parameter α_{ijk}

$$BD(G)=\prod_i^n \prod_j^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij}+N_{ij})} \prod_k^{r_i} \frac{\Gamma(\alpha_{ijk}+N_{ijk})}{\Gamma(\alpha_{ijk})}$$

- BIC score (**G.Schwartz** *Annals of statistics* 1978)
  - Laplace approximation

$$BIC(G)=\log(P(D/G,\hat{\theta}))-\frac{1}{2}Dim(G)\log(nb_{sample})$$

Scores

- BD Score *(**D.Heckerman** Machine learning 1995)*
  - A priori on conditional probability θ_{ij} following a Dirichlet distribution with parameter α_{ijk}

$$BD(G)=\prod_i^n \prod_j^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij}+N_{ij})} \prod_k^{r_i} \frac{\Gamma(\alpha_{ijk}+N_{ijk})}{\Gamma(\alpha_{ijk})}$$

- BIC score (**G.Schwartz** *Annals of statistics* 1978)
  - Laplace approximation

$$BIC(G)=\log(P(D/G,\hat{\theta}))-\frac{1}{2}Dim(G)\log(nb_{sample})$$

- fNML score *(**T.Silander** International journal of approximate reasoning 2010)*
  (**f**actorized **N**ormalized **M**aximum **L**ikelihood)
  - How much G explain D compare to other possible data sets D'

$$fNML(G)=\log(P(D/G,\hat{\theta}))-\sum_{i=1}^{n}\sum_{j=1}^{q_i} logC_{N_{ij}}^{r_i} \quad with \quad C_{N_{ij}}^{r_i} the\, normalizing\, sum$$

Extended Score

- What about $P(G_i)$?
  - Uniform for all G_i

Extended Score

- What about $P(G_i)$?
  - Uniform for all G_i

    How many restricted graph to node j with $\left|Pa_j^{G_i^j}\right| = l$ exist ?

Extended Score

- What about $P(G_i)$?
  - Uniform for all G_i

How many restricted graph to node j with $\left|Pa_j^{G_i^j}\right| = l$ exist ?

$\tau_l(G_i^j) = 1$ for $l = 0$

Extended Score

- What about $P(G_i)$?
  - Uniform for all G_i

How many restricted graph to node j with $\left|Pa_j^{G_i^j}\right| = l$ exist ?

$$\tau_l(G_i^j) = \begin{cases} 1 & \text{for } l = 0 \\ = (n-1) & l = 1 \end{cases}$$

Extended Score

- What about $P(G_i)$?
  - Uniform for all G_i

How many restricted graph to node j with $\left|Pa_j^{G_i^j}\right| = l$ exist ?

$$\tau_l(G_i^j) = 1 \quad \text{for } l=0$$
$$= (n-1) \quad l=1$$
$$= (n-1) * \left(\frac{n-2}{2}\right) \quad l=2$$

Extended Score

- What about $P(G_i)$?
  - Uniform for all G_i

How many restricted graph to node j with $\left|Pa_j^{G_i^j}\right| = l$ exist ?

$$\tau_l(G_i^j) = 1 \quad \text{for } l = 0$$
$$= (n-1) \quad l = 1$$
$$= (n-1) * \left(\frac{n-2}{2}\right) \quad l = 2$$
$$\vdots$$
$$= \left\{ {n-1 \atop p} \right\} \quad l = p$$

Extended Score

- What about $P(G_i)$?
  - Uniform for all G_i

How many restricted graph to node j with $\left|Pa_j^{G_i^j}\right| = l$ exist ?

$$\tau_l(G_i^j) = \begin{cases} 1 & \text{for } l=0 \\ (n-1) & l=1 \\ (n-1)*\left(\frac{n-2}{2}\right) & l=2 \\ \vdots & \\ \begin{Bmatrix} n-1 \\ p \end{Bmatrix} & l=p \end{cases}$$

For $l < \frac{n-1}{2}$

$\rightarrow \tau_{l-1}(G_i^j) < \tau_l(G_i^j)$

Extended Score

- What about $P(G_i)$?
  - Uniform for all G_i

How many restricted graph to node j with $\left|Pa_j^{G_i^j}\right| = l$ exist ?

$$\tau_l(G_i^j) = \begin{cases} 1 & \text{for } l=0 \\ (n-1) & l=1 \\ (n-1)*\left(\frac{n-2}{2}\right) & l=2 \\ \vdots & \\ \left\{ {n-1 \atop p} \right\} & l=p \end{cases}$$

For $l < \frac{n-1}{2}$

$$\rightarrow \tau_{l-1}(G_i^j) < \tau_l(G_i^j)$$

$$P_{l-1}(G_i^j) < P_l(G_i^j)$$

Against principle of parsimony ?

Extended Score

- What about $P(G_i)$?
  - Uniform for all G_i

How many restricted graph to node j with $\left|Pa_j^{G_i^j}\right| = l$ exist ?

$$\tau_l(G_i^j) = \begin{cases} 1 & \text{for } l=0 \\ (n-1) & l=1 \\ (n-1)*\left(\frac{n-2}{2}\right) & l=2 \\ \vdots & \\ \left\{ {n-1 \atop p} \right\} & l=p \end{cases}$$

For $l < \frac{n-1}{2}$

$$\rightarrow \tau_{l-1}(G_i^j) < \tau_l(G_i^j)$$

$$P_{l-1}(G_i^j) < P_l(G_i^j)$$

Against principle of parsimony ?

  - Extend BIC idea (**Chen** *Biometrika* 2008)

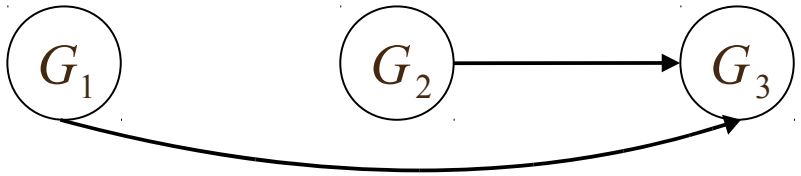
$$P(G_i^j) \propto \tau_l^{-\gamma}(G_i^j) \quad with \quad \gamma \in [0,1]$$

$$\log(P(G_i)) \simeq -\gamma \log\left(\prod_{j=1}^{n} \tau_l(G_i^j)\right)$$

→ applicable for all scores seen previously

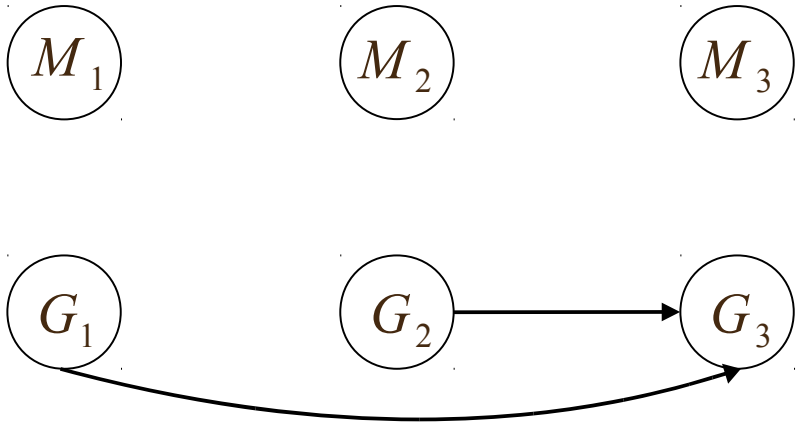
Model description and extra biological knowledge

Full observed data



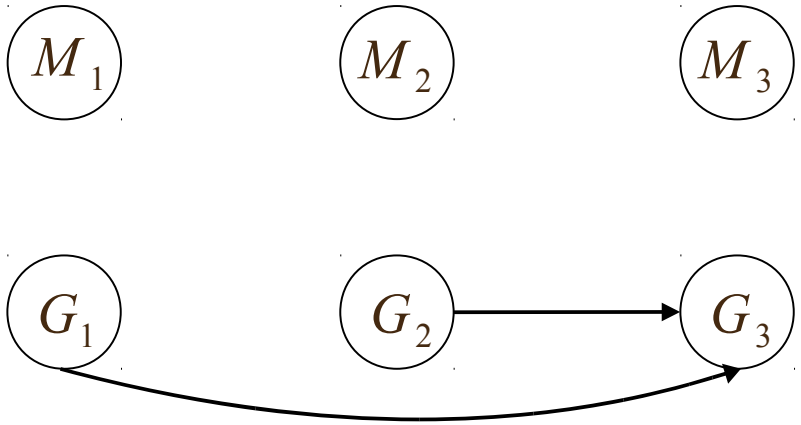
Model description and extra biological knowledge

Full observed data
One marker per gene



Model description and extra biological knowledge

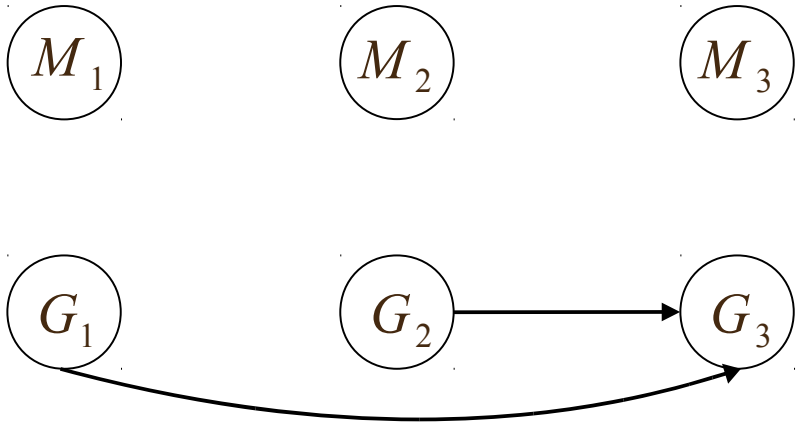
Full observed data
One marker per gene
Discrete expression data



G*i* : [1,4]

Model description and extra biological knowledge

Full observed data
One marker per gene
Discrete expression data
SNP markers

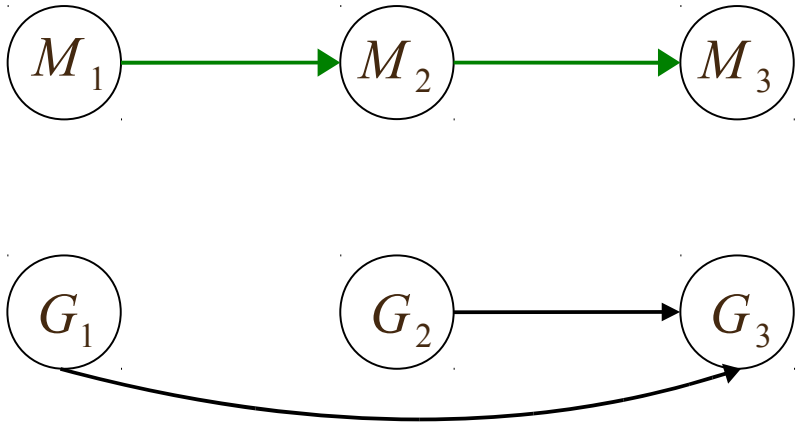


M*i* : 0/1

G*i* : [1,4]

Model description and extra biological knowledge

Full observed data
One marker per gene
Discrete expression data
SNP markers



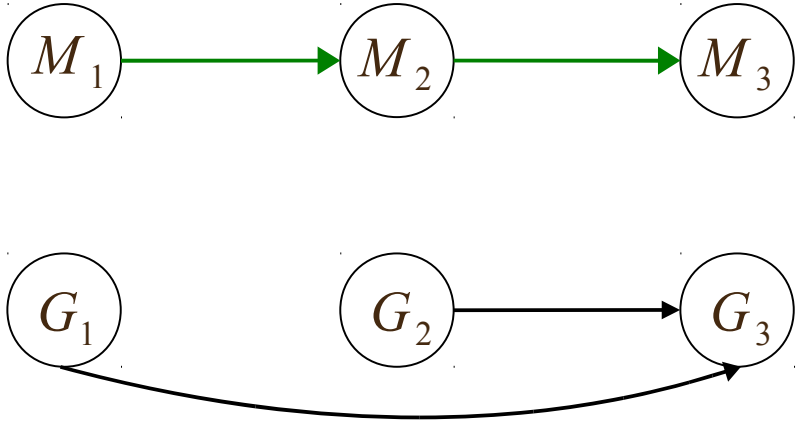
M*i* : 0/1

G*i* : [1,4]

→ Genetic linkage between markers

Model description and extra biological knowledge

Full observed data
One marker per gene
Discrete expression data
SNP markers



M_i : 0/1

G_i : [1,4]

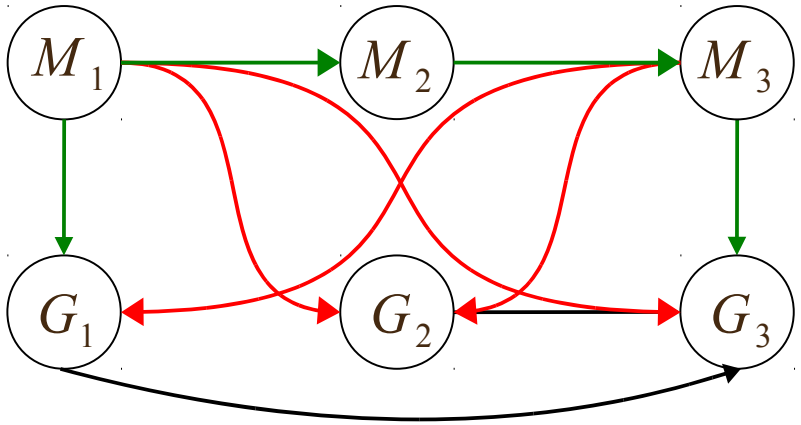
→ Genetic linkage between markers

→ Mutation in promoter region of G_i *(ex G_1 and G_3)*

- Force link $M_i \rightarrow G_i$
- Forbid links $M_i \rightarrow G_j$ $\quad \forall\, j \neq i$

Model description and extra biological knowledge

Full observed data
One marker per gene
Discrete expression data
SNP markers



M_i : 0/1

G_i : [1,4]

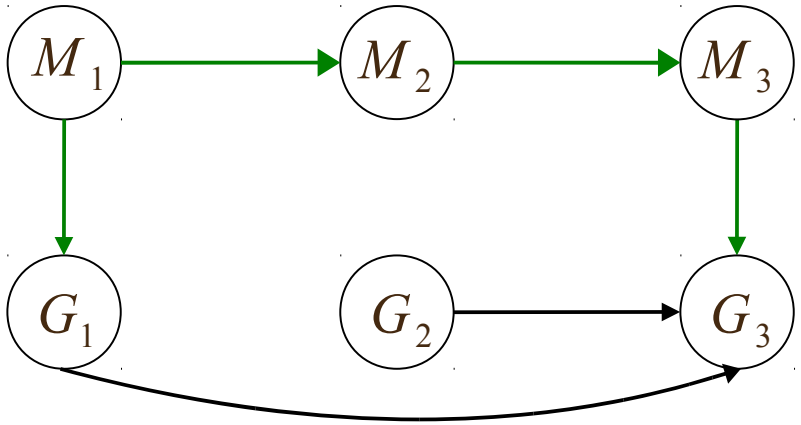
→ Genetic linkage between markers

→ Mutation in promoter region of G_i *(ex G_1 and G_3)*

- Force link $M_i \rightarrow G_i$
- Forbid links $M_i \rightarrow G_j$ $\quad \forall\, j \neq i$

Model description and extra biological knowledge

Full observed data
One marker per gene
Discrete expression data
SNP markers



M_i : 0/1

G_i : [1,4]

→ Genetic linkage between markers

→ Mutation in promoter region of G_i *(ex G_1 and G_3)*

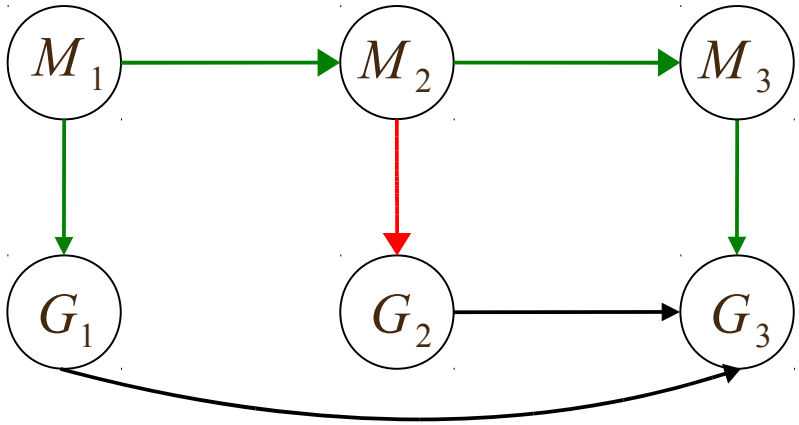
- Force link $M_i \rightarrow G_i$
- Forbid links $M_i \rightarrow G_j$ $\quad \forall\, j \neq i$

→ Mutation in coding region for G_i *(ex G_2)*

- Forbid link $M_i \rightarrow G_i$

Model description and extra biological knowledge

Full observed data
One marker per gene
Discrete expression data
SNP markers



M_i : 0/1

G_i : [1,4]

→ Genetic linkage between markers

→ Mutation in promoter region of G_i *(ex G_1 and G_3)*

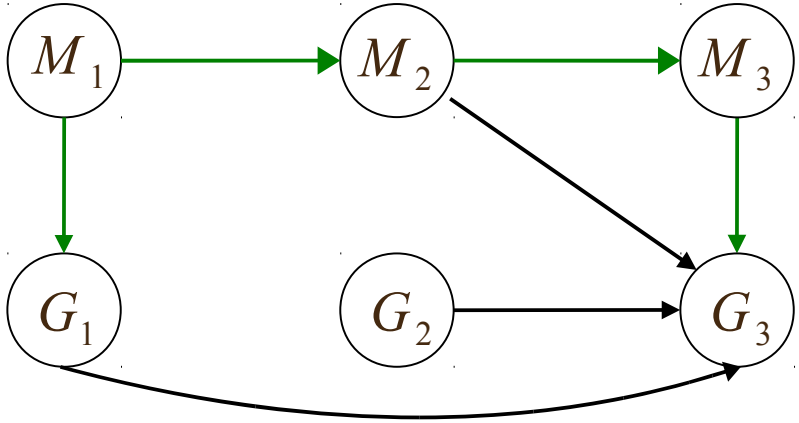
- Force link $M_i \rightarrow G_i$
- Forbid links $M_i \rightarrow G_j$ $\quad \forall\, j \neq i$

→ Mutation in coding region for G_i *(ex G_2)*

- Forbid link $M_i \rightarrow G_i$

Model description and extra biological knowledge

Full observed data
One marker per gene
Discrete expression data
SNP markers



M_i : 0/1

G_i : [1,4]

→ Genetic linkage between markers

→ Mutation in promoter region of G_i *(ex G_1 and G_3)*

- Force link $M_i \rightarrow G_i$
- Forbid links $M_i \rightarrow G_j$ $\quad \forall\, j \neq i$

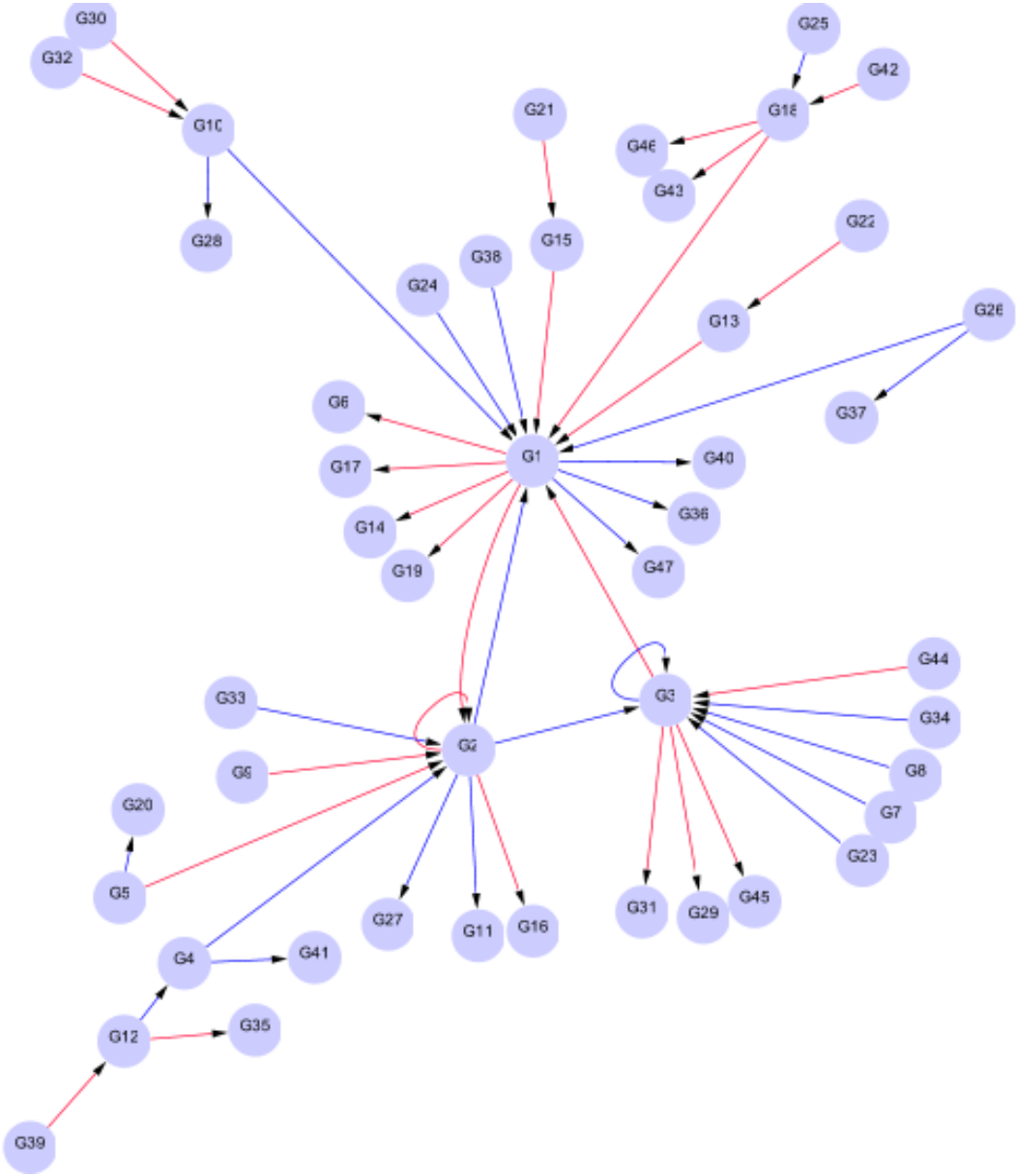
→ Mutation in coding region for G_i *(ex G_2)*

- Forbid link $M_i \rightarrow G_i$

Network studied

Artificial gene networks
(Mendes et al., Bioinformatics, 2003)

- *similar structure of known networks*
- *few hubs*
- *50 nodes / 50 edges*



Network evaluation

- Mean over 50 artificial networks
- Sample size between 50 and 500 individuals
- In our model we project all $M_i \rightarrow G_j$ as a $G_i \rightarrow G_j \quad \forall\, j \neq i$
- We use Greedy search (implemented in **Banjo** *Hartemink* 2005)
- Evaluation doesn't takes into account edges orientation
- 2 common metrics Predictive value / Sensitivity

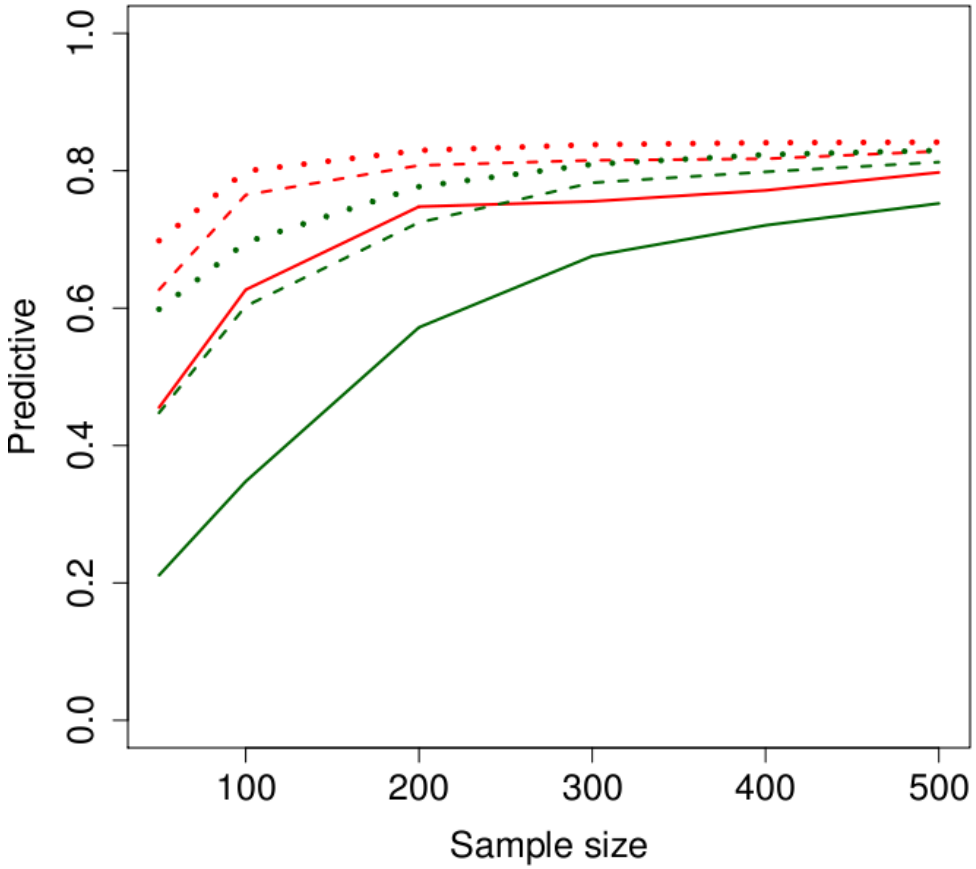
$$Predictive\,value\,(=1-FDR): \frac{TP}{TP+FP}$$

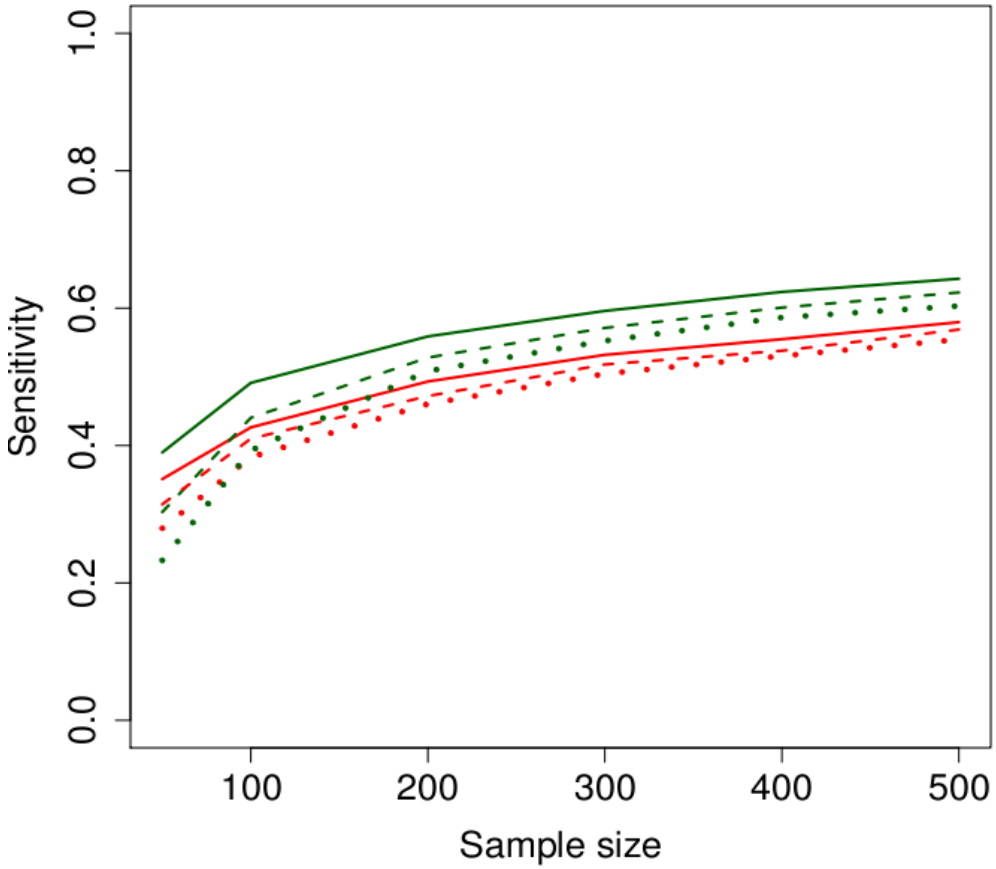
$$Sensitivity: \frac{TP}{TP+FN}$$

TP : number of correct learned edges
FP : number of wrong learned edges
FN : number of missed edges

Extended scores impact

Legend: BIC score
fNML score

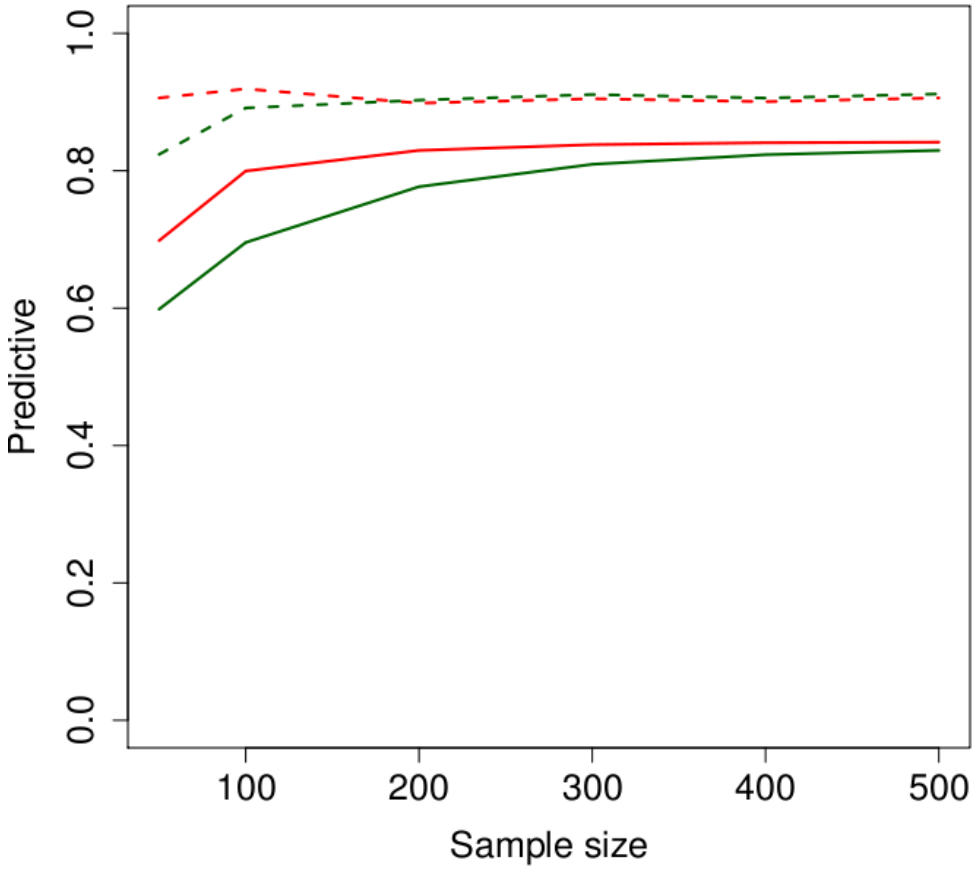


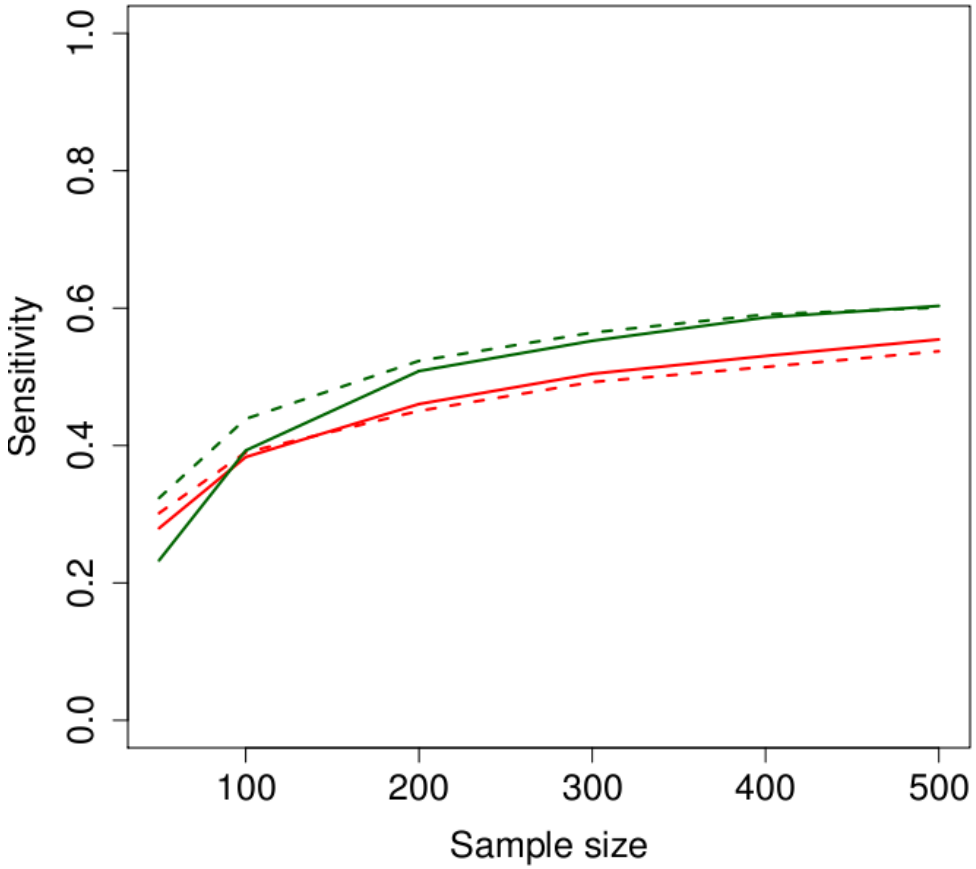


- We compare BIC and fNML scores with several γ values
  - $\gamma = 0 \rightarrow$ classic scores (solid line)
  - $\gamma = 0.5$ (dashed line)
  - $\gamma = 1 \rightarrow$ same probability over connectivity classes (dotted line)

Biological knowledge impact

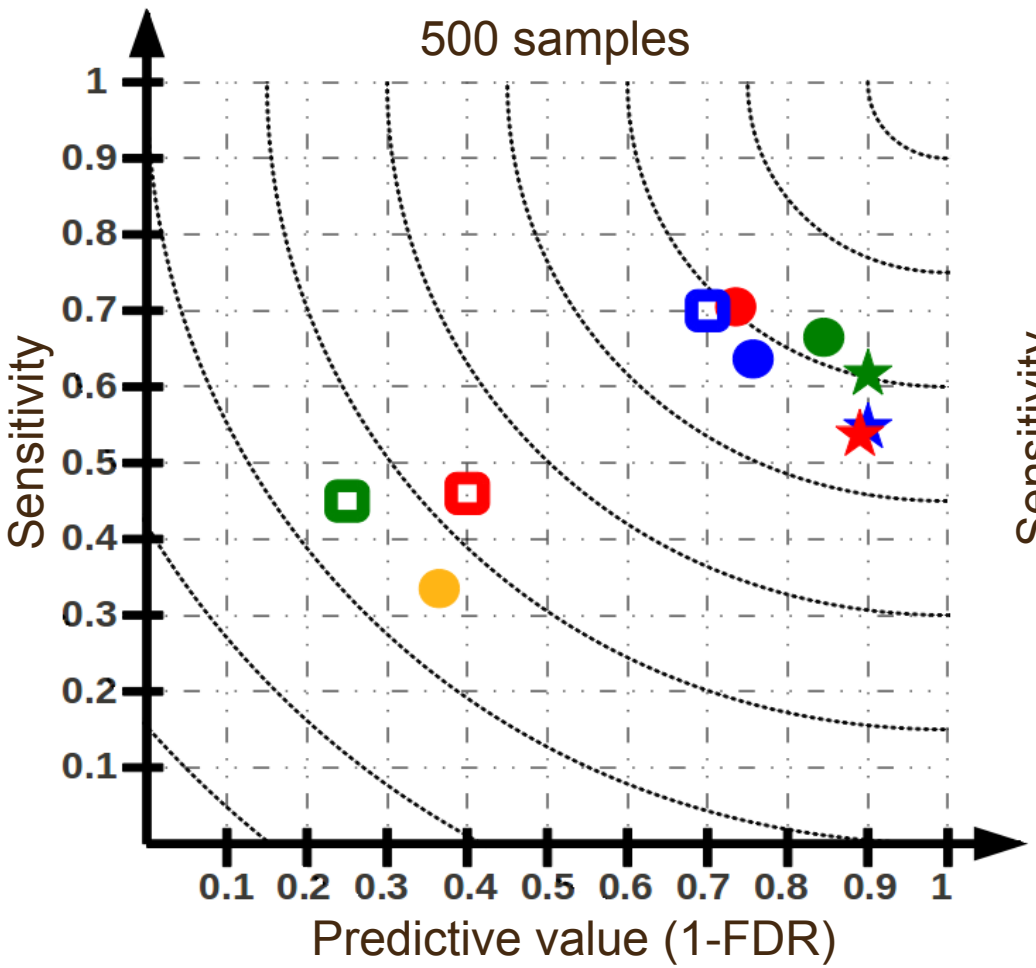
Legend: BIC score
fNML score

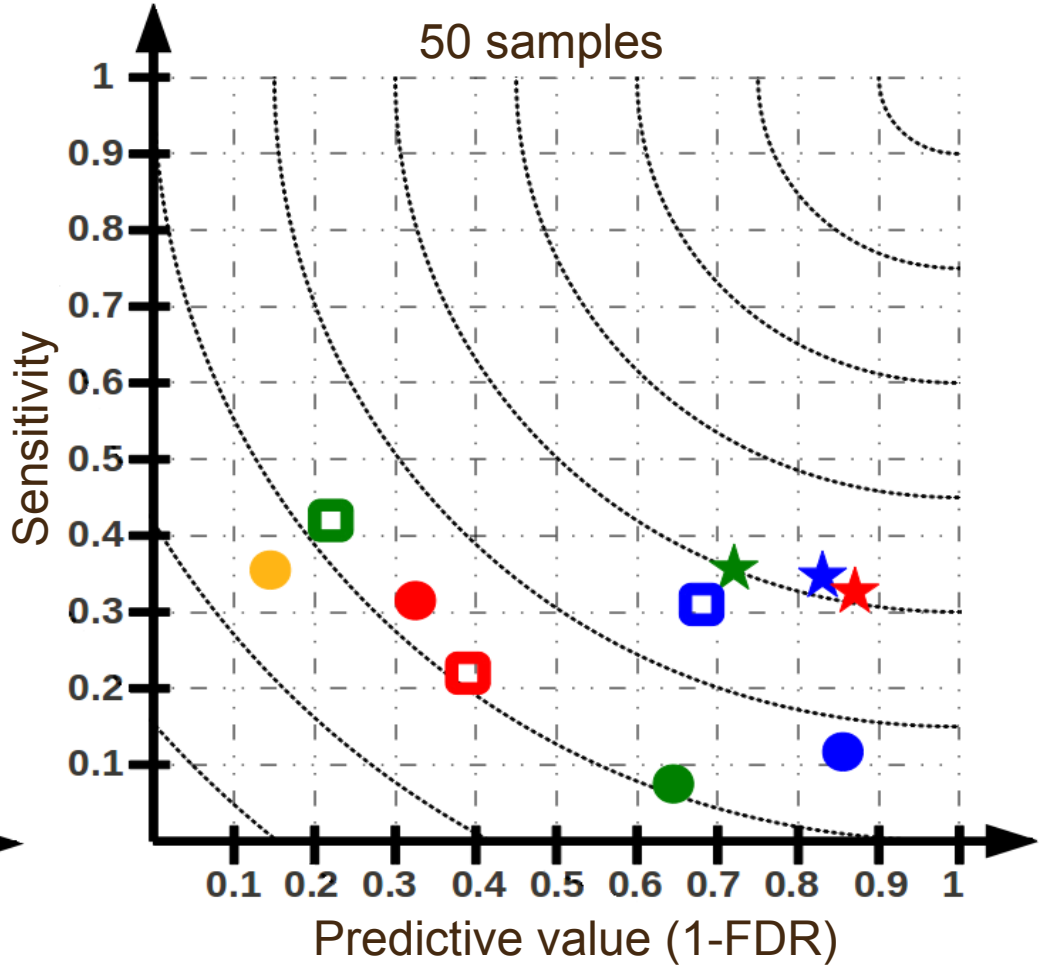




· We fix / ban relations in regulation network inference with γ=1
  · without extra biological knowledge (solid line)
  · with extra biological knowledge (dashed line)

Comparative results





Bayesian network inference : Banjo Greedy Search ($BDeu_1$ ★ / BIC_1 ★ / $FNML_1$ ★)

Linear model : SEM Lasso ●, GGMselect ●, GeneNet ●, SIMoNe ●)

Pairwise correlation : CLR □, ARACNE □, ParCorA □

Conclusion

- Scoring criteria study with uniform node in-degree prior
- Model description taking into account specific Biological knowledge
- Comparison with several regulatory network inference methods
- Robustness of Bayesian networks for sparse graphs

Perspectives

- Improve learning algorithm performance
- Study causality
- Try on real data

Bibliography

- **D.M. Chickering** *'Efficient Approximations for the Marginal Likelihood of Incomplete Data Given a Bayesian Network' 1997*
- ***J. Chen*** *'Extended Bayesian information criteria for model selection with large model spaces' 2008*
- ***T.Silander*** *'Learning locally minimax optimal Bayesian Networks' 2010*
- ***N.Meinshausen*** *'High dimensional graphs and variable selection with the lasso' 2006*
- ***M.Bansal*** *'How to infer gene networks from expression profiles' 2006*
- ***B.Liu*** *'Gene network inference via structural equation modeling in genetical genomics experiments' 2008*
- ***J.Zhu*** *'Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations' 2007*