Gene regulatory networks reconstruction from simulated System Genetics data

what we tried, what we learnt

D. Allouche C. Cierco-Ayrolles S. de Givry G. Guillermin B. Mangin T. Schiex J. Vandel M. Vignes

StatSeq meeting - Paris

Friday 29th March 2013

Problem motivation

Modelling a biological system

- Interests are in **industrial** (pharmaceutical, agribusiness, genetic engineering . . .) and **public** (health, environment, research on biological mecanisms and the impact of causal intervention) sectors.

Problem motivation

Modelling a biological system

- Interests are in **industrial** (pharmaceutical, agribusiness, genetic engineering . . .) and **public** (health, environment, research on biological mecanisms and the impact of causal intervention) sectors.
- Computational-aided biological modelling due to the considered complexity of systems: high-dimension, non-linear dependencies, mixture of discrete and continuous measures . . .

Problem motivation

Modelling a biological system

- Interests are in **industrial** (pharmaceutical, agribusiness, genetic engineering . . .) and **public** (health, environment, research on biological mecanisms and the impact of causal intervention) sectors.
- Computational-aided biological modelling due to the considered complexity of systems: high-dimension, non-linear dependencies, mixture of discrete and continuous measures . . .
- Organism (e.g. plant, animal) $\sim$ complex system comprising **many acting entities** (genes, proteins, metabolites), in **interation** with each other: passing messages, integrating information and transforming it. . . The use of a **graph or network** to represent such a system seems adequate.

Problem motivation

Modelling a biological system

- Interests are in **industrial** (pharmaceutical, agribusiness, genetic engineering . . .) and **public** (health, environment, research on biological mecanisms and the impact of causal intervention) sectors.
- Computational-aided biological modelling due to the considered complexity of systems: high-dimension, non-linear dependencies, mixture of discrete and continuous measures . . .
- Organism (e.g. plant, animal) $\sim$ complex system comprising **many acting entities** (genes, proteins, metabolites), in **interation** with each other: passing messages, integrating information and transforming it. . . The use of a **graph or network** to represent such a system seems adequate.
- Issues here: (i) formal adequate modelling framework and (ii) identification of a network that best represents the system.

Problem motivation

Modelling a biological system

- Interests are in **industrial** (pharmaceutical, agribusiness, genetic engineering . . .) and **public** (health, environment, research on biological mecanisms and the impact of causal intervention) sectors.
- Computational-aided biological modelling due to the considered complexity of systems: high-dimension, non-linear dependencies, mixture of discrete and continuous measures . . .
- Organism (e.g. plant, animal) $\sim$ complex system comprising **many acting entities** (genes, proteins, metabolites), in **interation** with each other: passing messages, integrating information and transforming it. . . The use of a **graph or network** to represent such a system seems adequate.
- Issues here: (i) formal adequate modelling framework and (ii) identification of a network that best represents the system.
- Focus in this presentation on (simulated) 'Systems Genetics' or 'Genetical Genomics' data, only looking at the level of genes.

Quick data description

- 72 data sets, 3 repeats (different networks) for each of the 24 configurations.

Quick data description

- 72 data sets, 3 repeats (different networks) for each of the 24 configurations.
- $24 = 8$ (configurations) $\times 3$ (network sizes: $100, 1,000$ and $5,000$ genes).

Quick data description

- 72 data sets, 3 repeats (different networks) for each of the 24 configurations.
- $24 = 8$ (configurations) $\times 3$ (network sizes: $100, 1,000$ and $5,000$ genes).
- $8 = 2^3$ configurations: combinations of (i) 2 sample sizes ($n = 900$ or 300), (ii) 2 gene expression heritability (**H**igh vs **L**ow) and (iii) 2 chromosome densities (**D**ense vs **S**parse)

Quick data description

- 72 data sets, 3 repeats (different networks) for each of the 24 configurations.
- $24 = 8$ (configurations) $\times 3$ (network sizes: $100, 1,000$ and $5,000$ genes).
- $8 = 2^3$ configurations: combinations of (i) 2 sample sizes ($n = 900$ or 300), (ii) 2 gene expression heritability (**H**igh vs **L**ow) and (iii) 2 chromosome densities (**D**ense vs **S**parse)

Dataset generation recipe

Choose simulation parameters, choose a network, generate individual genotypes and then simulate steady-state gene g expression data from:

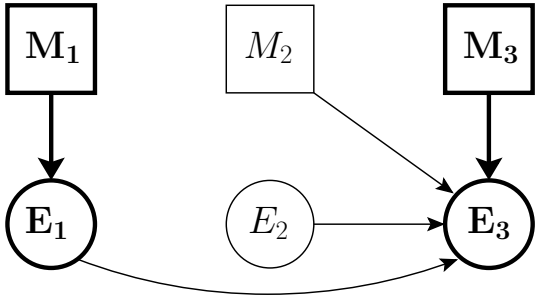
$$\frac{\mathrm{d}\,G_g}{\mathrm{d}\,t} = Z_g^c \cdot V_g \cdot \theta_g^{syn} \cdot \prod_k \left(1 + A_{k,g} \frac{G_k^{h_{k,g}}}{G_k^{h_{k,g}} + (K_{k,g}/Z_k^t)^{h_{k,g}}}\right) - \lambda_g \cdot \theta_g^{deg} \cdot G_g$$

from SysGenSIM, [Pinna et al. 2011]

Listing possible interactions

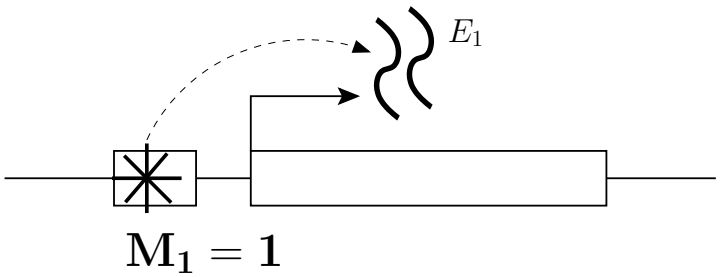
Cis regulation

$$\frac{\mathrm{d}\,G_g}{\mathrm{d}\,t} = \mathbf{Z^c_g} \cdot V_g \cdot \theta_g^{syn} \cdot \prod_k \left(1 + A_{k,g} \frac{G_k^{h_{k,g}}}{G_k^{h_{k,g}} + (K_{k,g}/Z_k^t)^{h_{k,g}}}\right) - \lambda_g \cdot \theta_g^{deg} \cdot G_g$$



$M_i = 0$ or 1

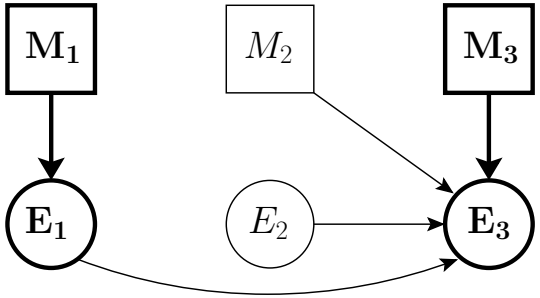
$E_i \in \mathbb{R}$



Listing possible interactions

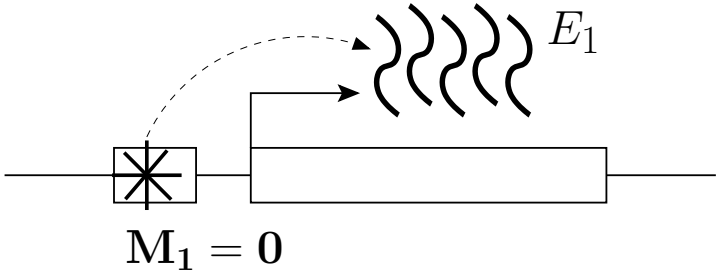
Cis regulation

$$\frac{\mathrm{d}\,G_g}{\mathrm{d}\,t} = \mathbf{Z^c_g} \cdot V_g \cdot \theta_g^{syn} \cdot \prod_k \left(1 + A_{k,g} \frac{G_k^{h_{k,g}}}{G_k^{h_{k,g}} + (K_{k,g}/Z_k^t)^{h_{k,g}}}\right) - \lambda_g \cdot \theta_g^{deg} \cdot G_g$$



$M_i = 0$ or 1

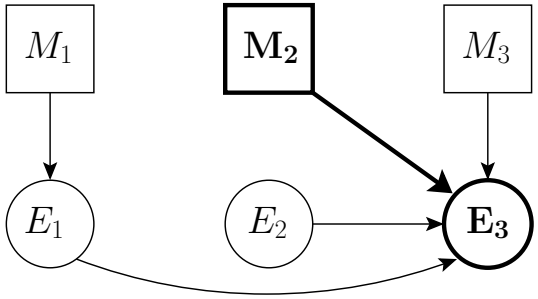
$E_i \in \mathbb{R}$



Listing possible interactions

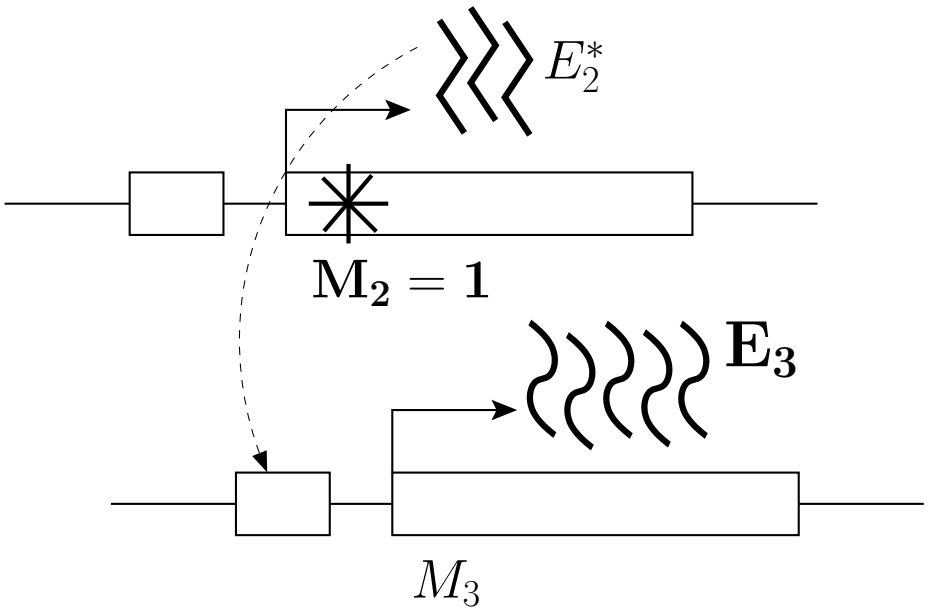
Trans regulation

$$\frac{\mathrm{d}\,G_g}{\mathrm{d}\,t} = Z_g^c \cdot V_g \cdot \theta_g^{syn} \cdot \prod_k \left(1 + A_{k,g} \frac{G_k^{h_{k,g}}}{G_k^{h_{k,g}} + (K_{k,g}/\mathbf{Z_k^t})^{h_{k,g}}} \right) - \lambda_g \cdot \theta_g^{deg} \cdot G_g$$



$M_i = 0$ or 1

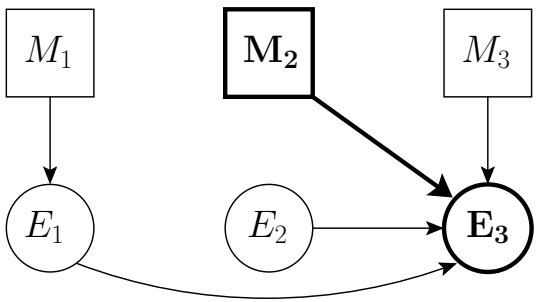
$E_i \in \mathbb{R}$



Listing possible interactions

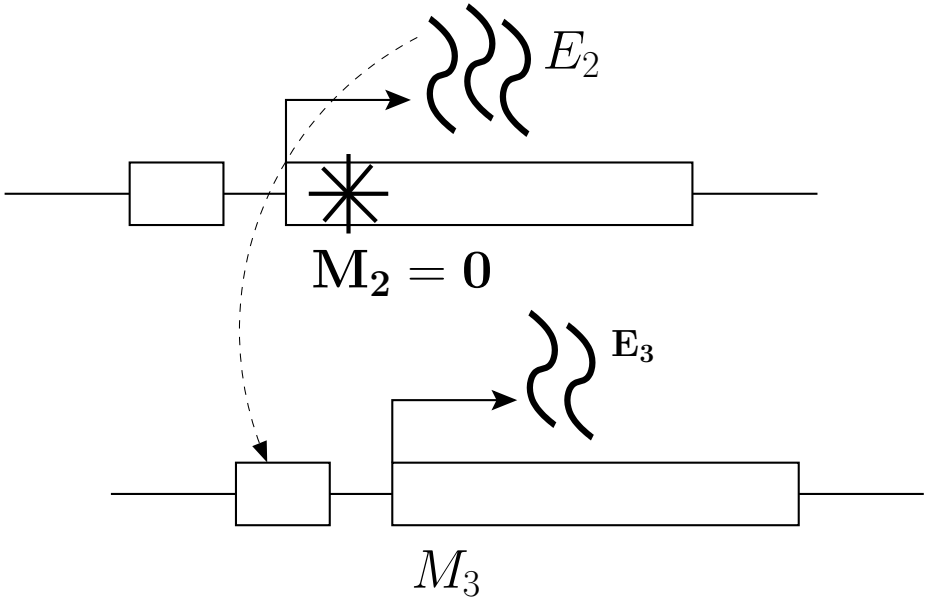
Trans regulation

$$\frac{\mathrm{d}\,G_g}{\mathrm{d}\,t} = Z_g^c \cdot V_g \cdot \theta_g^{syn} \cdot \prod_k \left(1 + A_{k,g} \frac{G_k^{h_{k,g}}}{G_k^{h_{k,g}} + (K_{k,g}/\mathbf{Z_k^t})^{h_{k,g}}} \right) - \lambda_g \cdot \theta_g^{deg} \cdot G_g$$

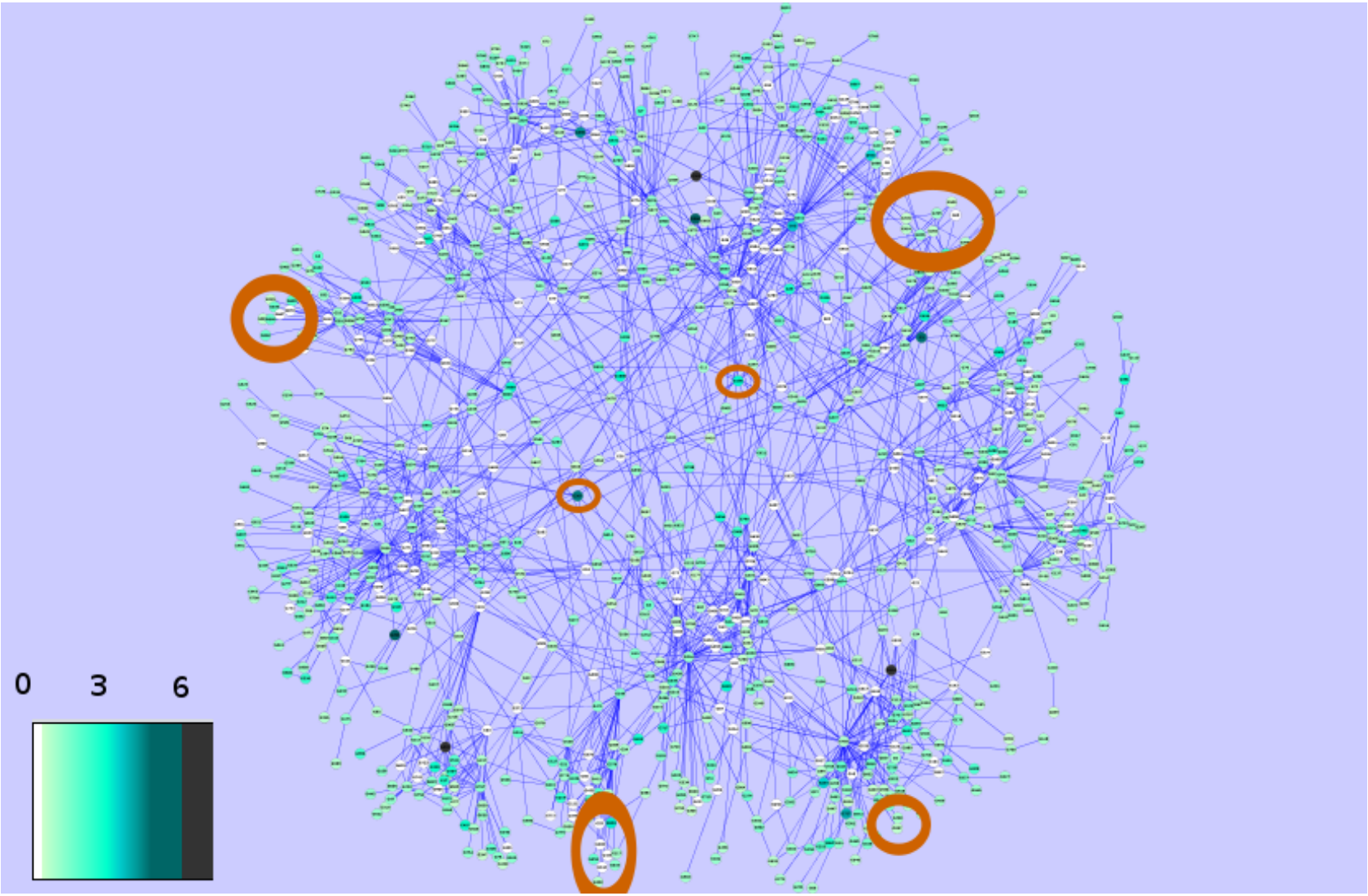


$M_i = 0$ or 1

$E_i \in \mathbb{R}$



Marker multifactorial effect visualisation



Models used

to decipher relationships between variables

- Penalised linear regressions (**lasso**, **Dantzig**) + data bootstrap

Models used

to decipher relationships between variables

- Penalised linear regressions (**lasso**, **Dantzig**) + data bootstrap
- Bayesian networks (**BN**) + data bootstrap

Models used

to decipher relationships between variables

- Penalised linear regressions (**lasso**, **Dantzig**) + data bootstrap
- Bayesian networks (**BN**) + data bootstrap
- random forests (RF; has integrated bootstrap)

Models used

to decipher relationships between variables

- Penalised linear regressions (**lasso**, **Dantzig**) + data bootstrap
- Bayesian networks (**BN**) + data bootstrap
- random forests (RF; has integrated bootstrap)

Data bootstrap [Efron 1981]

- strategy used to get confidence on predictions and overcome noise effect.
- implementation: randomly draw (with replacement) N_{boot} replicate data-sets of identical sample size as the original data, replicate the computation (drawback 1) and store the N_{boot} models to estimate distribution of the desired statistics (e.g. edge weight).
- did not make use of the offered possibility to study the behaviour of any (lack of) fitness function (e.g. likelihood, MSE) from out-of-bootstrap samples (but for RF) since each replicate doesn't use $\sim 37\%$ of the original samples (drawback 2 when n is small).

Penalised linear regressions

Solve individual linear regression for each gene:

$$E_g = \sum_{j=1}^{p} \alpha_{gj} M_j + \sum_{\substack{j=1 \\ j \neq g}}^{p} \beta_{gj} E_j + \varepsilon_g$$

Since $n < p$, assumption that few (α, β)'s are 0 (makes GRN sparse), penalised regression methods such as the lasso or the Dantzig selector were chosen.

Penalised linear regressions

Solve individual linear regression for each gene:

$$E_g = \sum_{j=1}^{p} \alpha_{gj} M_j + \sum_{\substack{j=1 \\ j \neq g}}^{p} \beta_{gj} E_j + \varepsilon_g$$

Since $n < p$, assumption that few (α, β)'s are 0 (makes GRN sparse), penalised regression methods such as the lasso or the Dantzig selector were chosen.

lasso penalisation [Tibshirani 1996]

Both shrinks (bias) and selects variables according to:

$$\begin{aligned}(\hat{\alpha}, \hat{\beta})^{lasso} &= \arg\min_{\alpha,\beta} \mid\mid E - M\alpha - E\beta \mid\mid_{\ell_2}^2 + \lambda \mid\mid (\alpha, \beta) \mid\mid_{\ell_1} \\ &= \arg\min_{\alpha,\beta} \mid\mid E - M\alpha - E\beta \mid\mid_{\ell_2}, \text{ subject to } \mid\mid (\alpha, \beta) \mid\mid_{\ell_1} \leq t\end{aligned}$$

Penalised linear regressions

Solve individual linear regression for each gene:

$$E_g = \sum_{j=1}^{p} \alpha_{gj} M_j + \sum_{\substack{j=1 \\ j \neq g}}^{p} \beta_{gj} E_j + \varepsilon_g$$

Since $n < p$, assumption that few (α, β)'s are 0 (makes GRN sparse), penalised regression methods such as the lasso or the Dantzig selector were chosen.

Dantzig selector [Candès & Tao 2007]

Slightly different constaint (related to gradient of RSS):

$$\begin{aligned}
(\hat{\alpha}, \hat{\beta})^{dantzig} &= \arg\min_{\alpha,\beta} \| (\alpha, \beta) \|_{\ell_1} \text{ s.t. } \| (E, M)^\top (E - M\alpha - E\beta) \|_{\ell_\infty} \leq \delta \\
&= \arg\min_{\alpha,\beta} \| (E, M)^\top (E - M\alpha - E\beta) \|_{\ell_\infty}, \text{ s.t. } \| (\alpha, \beta) \|_{\ell_1} \leq t
\end{aligned}$$

Building weights for edges predictions

Our strategy

- Repeat model fitting for the N_{boot} bootstraps and for a grid of $(q = 10)$ penalties.

Building weights for edges predictions

Our strategy

- Repeat model fitting for the N_{boot} bootstraps and for a grid of $(q = 10)$ penalties.
- Estimate weights $w_{M_j \to E_g}$ by the ratio of $\alpha_{gj}^{boot,pen} \neq 0$ and $w_{E_j \to E_g}$ by $\frac{\#\{\beta_{\mathbf{gj}}^{boot,pen} \neq 0\} + \#\{\beta_{\mathbf{jg}}^{boot,pen} \neq 0\}}{4qN_{boot}}$: post-symetrisation of $E_g \to E_{g\prime}$ edges and higher confidence in $M_g \to E_{g\prime}$ relationships.

Building weights for edges predictions

Our strategy

- Repeat model fitting for the N_{boot} bootstraps and for a grid of $(q = 10)$ penalties.
- Estimate weights $w_{M_j \to E_g}$ by the ratio of $\alpha_{gj}^{boot,pen} \neq 0$ and $w_{E_j \to E_g}$ by $\frac{\#\{\beta_{\mathbf{gj}}^{boot,pen} \neq 0\} + \#\{\beta_{\mathbf{jg}}^{boot,pen} \neq 0\}}{4qN_{boot}}$: post-symetrisation of $E_g \to E_{g\prime}$ edges and higher confidence in $M_g \to E_{g\prime}$ relationships.
- [Bach 2008] established that under "some conditions" (sparsity, size effect and unique λ_n), the bootstrap lasso identifies correct edges with probability 1 and selects false positives with probability < 1 when $n \to \infty$. Our ranking should be related to edge existence !

Bayesian networks

- Defined by a directed acyclic graph and conditional probabilities $P(V \mid \mathrm{Par}_V)$ for all nodes V in the graph ([Pearl 1988] and [Friedman 2000] for use with expression data).

Bayesian networks

- Defined by a directed acyclic graph and conditional probabilities $P(V \mid \mathrm{Par}_V)$ for all nodes V in the graph ([Pearl 1988] and [Friedman 2000] for use with expression data).
- Natural representation of GRN but for cycles. However, cycles can be obtained (restarts, bootstraps).

Bayesian networks

- Defined by a directed acyclic graph and conditional probabilities $P(V \mid \mathrm{Par}_V)$ for all nodes V in the graph ([Pearl 1988] and [Friedman 2000] for use with expression data).
- Natural representation of GRN but for cycles. However, cycles can be obtained (restarts, bootstraps).
- Algorithm for BN inference are of two kinds: based either on independence tests or on **scores**: Bayesian (BD, BDeu. . .) or information theoretic (AIC, BIC . . .).

Bayesian networks

- Defined by a directed acyclic graph and conditional probabilities $P(V \mid \mathrm{Par}_V)$ for all nodes V in the graph ([Pearl 1988] and [Friedman 2000] for use with expression data).
- Natural representation of GRN but for cycles. However, cycles can be obtained (restarts, bootstraps).
- Algorithm for BN inference are of two kinds: based either on independence tests or on **scores**: Bayesian (BD, BDeu. . .) or information theoretic (AIC, BIC . . .).
- NP hard problem (even if indegree ≤ 2 [Chickering 1996]): a simple greedy search is already very computation demanding: number of parents limited to 5.

Bayesian networks

Algorithm

1. Discretise expression into adaptively (2 to 4 states).

Bayesian networks

Algorithm

1. Discretise expression into adaptively (2 to 4 states).
2. Select potential parental set for each node if local BDeu score increased by adding parents separately in comparison to the empty graph.

Bayesian networks

Algorithm

1. Discretise expression into adaptively (2 to 4 states).
2. Select potential parental set for each node if local BDeu score increased by adding parents separately in comparison to the empty graph.
3. Select most influencial marker from sliding window.

Bayesian networks

Algorithm

1. Discretise expression into adaptively (2 to 4 states).
2. Select potential parental set for each node if local BDeu score increased by adding parents separately in comparison to the empty graph.
3. Select most influencial marker from sliding window.
4. Account for biological knowledge: enforce $M_g \rightarrow M_{g+1}$ along the chromosome and forbid $E_g \rightarrow M_{g\prime}$, prior *cis*-reg. effect tested.

Bayesian networks

Algorithm

1. Discretise expression into adaptively (2 to 4 states).
2. Select potential parental set for each node if local BDeu score increased by adding parents separately in comparison to the empty graph.
3. Select most influencial marker from sliding window.
4. Account for biological knowledge: enforce $M_g \rightarrow M_{g+1}$ along the chromosome and forbid $E_g \rightarrow M_{g\prime}$, prior *cis*-reg. effect tested.
5. Selected DAG with highest BDeu score among 3 restarts a Stochastic Greedy Search algorithm with extended local move operators (SGS3, see [Vandel et al. 2012]).

Bayesian networks

Algorithm

1. Discretise expression into adaptively (2 to 4 states).
2. Select potential parental set for each node if local BDeu score increased by adding parents separately in comparison to the empty graph.
3. Select most influencial marker from sliding window.
4. Account for biological knowledge: enforce $M_g \rightarrow M_{g+1}$ along the chromosome and forbid $E_g \rightarrow M_{g\prime}$, prior *cis*-reg. effect tested.
5. Selected DAG with highest BDeu score among 3 restarts a Stochastic Greedy Search algorithm with extended local move operators (SGS3, see [Vandel et al. 2012]).
6. Scores are then simply the ratio of edge detection among the bootstraps.

Network post-processings

Ultimate goal (in our simulated context): list of directed interactions between genes, not list of causal eQTLs and gene expression influences

- First proposition **gen+mark**: $w_{g\ tog\prime} = w_{M_g\ toE_{g\prime}} + w_{E_g\ toE_{g\prime}}$.

Network post-processings

Ultimate goal (in our simulated context): list of directed interactions between genes, not list of causal eQTLs and gene expression influences

- First proposition **gen+mark**: $w_{g\ tog\prime} = w_{M_g\ toE_{g\prime}} + w_{E_g\ toE_{g\prime}}$.
- Second proposition **filt.mark** only relying on markers: $w_{g\ tog\prime} = \max_{h \in \{g-5;g+5\}} w_{M_h\ toE_{g\prime}}$.

Network post-processings

Ultimate goal (in our simulated context): list of directed interactions between genes, not list of causal eQTLs and gene expression influences

- First proposition **gen+mark**: $w_{g\ to g'} = w_{M_g\ toE_{g'}} + w_{E_g\ toE_{g'}}$.
- Second proposition **filt.mark** only relying on markers: $w_{g\ to g'} = \max_{h \in \{g-5;g+5\}} w_{M_h\ toE_{g'}}$.
- Third proposition **gen+mark.filt**: combination of 2nd marker weights and sum just like in 1st proposition.

Network post-processings

Ultimate goal (in our simulated context): list of directed interactions between genes, not list of causal eQTLs and gene expression influences

- First proposition **gen+mark**: $w_{g\ to g'} = w_{M_g\ toE_{g'}} + w_{E_g\ toE_{g'}}$.
- Second proposition **filt.mark** only relying on markers: $w_{g\ to g'} = \max_{h \in \{g-5;g+5\}} w_{M_h\ toE_{g'}}$.
- Third proposition **gen+mark.filt**: combination of 2nd marker weights and sum just like in 1st proposition.

Other cleverer post-processings can be built but time was lacking to thoroughly assess them in the different configurations !

Results 1: AUPR for $1,000$ gene networks

	AUPR with edge orientations				AUPR without edge orientations			
	Methods				Methods			
Network/configuration/data-set	Lasso	Dantzig	RF	BN	Lasso	Dantzig	RF	BN
Net4-Conf1-DS25-300SH	11.65	12.02	9.63	**14.20**	15.76	**16.41**	11.06	15.90
Net4-Conf2-DS26-900SH	15.88	15.66	17.95	**18.30**	**21.97**	21.68	20.05	20.08
Net4-Conf3-DS27-300SL	11.20	11.35	3.88	**11.83**	16.64	**17.18**	5.29	15.01
Net4-Conf4-DS28-900SL	21.49	21.78	9.64	**27.28**	32.46	**33.30**	11.31	32.95
Net4-Conf5-DS29-300DH	4.89	5.02	**7.31**	7.13	6.97	7.29	8.41	**8.60**
Net4-Conf6-DS30-900DH	9.68	10.05	13.82	**20.15**	13.81	14.53	15.60	**22.23**
Net4-Conf7-DS31-300DL	8.60	9.57	3.09	**13.18**	13.07	14.95	4.38	**16.59**
Net4-Conf8-DS32-900DL	16.20	17.43	7.39	**23.24**	24.20	26.71	9.12	**28.76**
Net5-Conf1-DS33-300SH	16.05	15.71	16.16	**16.96**	**21.52**	21.27	17.81	18.89
Net5-Conf2-DS34-900SH	22.17	21.71	23.96	**30.46**	31.08	30.64	26.28	**32.25**
Net5-Conf3-DS35-300SL	14.55	**14.61**	5.56	13.28	21.69	**22.10**	7.42	16.89
Net5-Conf4-DS36-900SL	24.57	24.70	13.53	**25.56**	37.38	**37.85**	15.86	31.37
Net5-Conf5-DS37-300DH	6.66	6.74	**9.04**	8.71	9.34	9.63	**10.58**	10.27
Net5-Conf6-DS38-900DH	12.80	12.67	21.76	**23.74**	17.55	17.76	23.73	**25.66**
Net5-Conf7-DS39-300DL	10.71	11.16	3.60	**15.36**	17.10	18.19	5.20	**18.71**
Net5-Conf8-DS40-900DL	17.42	17.92	11.04	**25.57**	26.33	27.75	12.86	**30.71**
Net6-Conf1-DS41-300SH	13.07	12.83	13.34	**15.75**	**17.90**	17.64	15.05	17.72
Net6-Conf2-DS42-900SH	17.54	17.59	23.63	**24.13**	24.81	24.80	25.56	**26.14**
Net6-Conf3-DS43-300SL	12.62	12.72	4.32	**13.40**	19.00	**19.38**	5.64	17.02
Net6-Conf4-DS44-900SL	20.72	**21.07**	10.67	20.14	32.06	**32.72**	12.69	26.12
Net6-Conf5-DS45-300DH	5.43	5.51	**7.41**	5.70	7.79	7.98	**8.83**	6.98
Net6-Conf6-DS46-900DH	8.55	8.43	**15.90**	12.34	11.91	11.95	**17.67**	14.13
Net6-Conf7-DS47-300DL	8.70	9.23	2.57	**10.07**	13.69	**14.84**	3.98	13.42
Net6-Conf8-DS48-900DL	14.68	15.33	7.82	**16.11**	22.86	**24.41**	10.06	21.36

Results 2: Effects of simulation parameters

- Sample size n: the larger, the better !

Results 2: Effects of simulation parameters

- Sample size n: the larger, the better !
- Higher gene expression heritability gives better results.

Results 2: Effects of simulation parameters

- Sample size n: the larger, the better !
- Higher gene expression heritability gives better results.
- Sparse chromosome genetic contents are more easily to unravel.

Results 2: Effects of simulation parameters

- Sample size n: the larger, the better !
- Higher gene expression heritability gives better results.
- Sparse chromosome genetic contents are more easily to unravel.
- BUT this is "in principle":

Results 2: Effects of simulation parameters

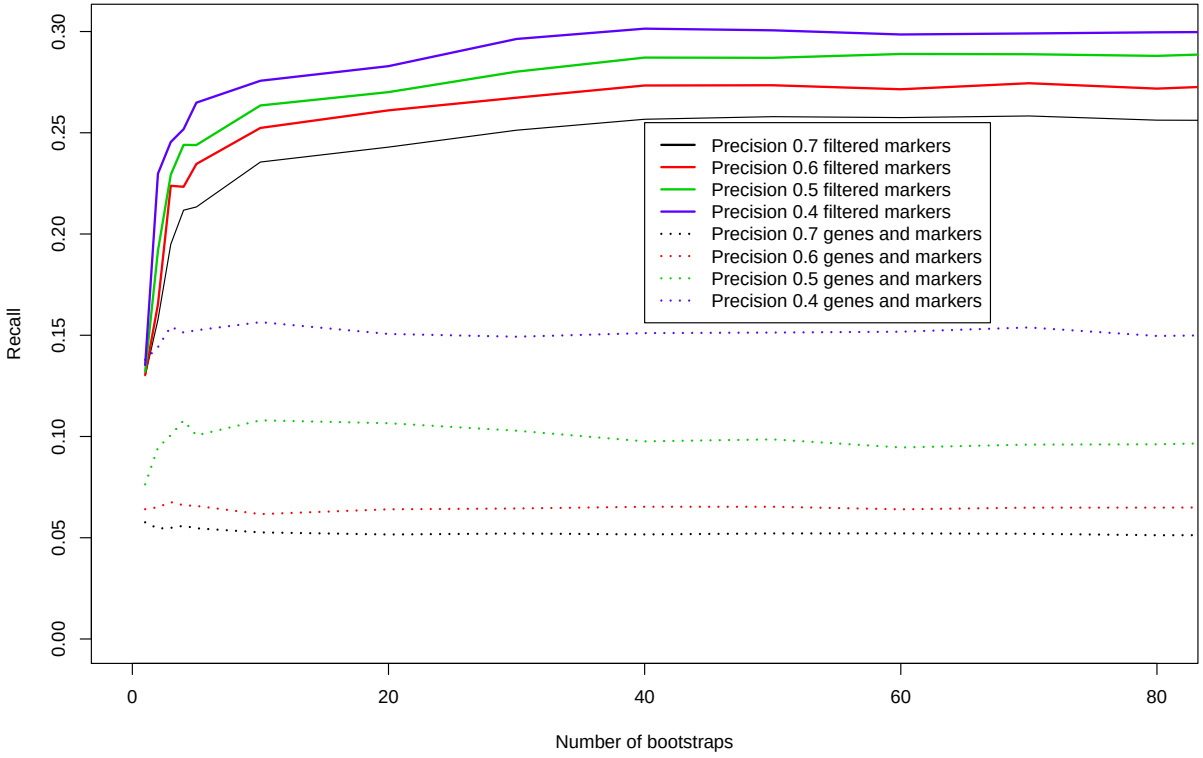
- Sample size n: the larger, the better !
- Higher gene expression heritability gives better results.
- Sparse chromosome genetic contents are more easily to unravel.
- BUT this is "in principle":

Results 2: Effects of simulation parameters

- Sample size n: the larger, the better !
- Higher gene expression heritability gives better results.
- Sparse chromosome genetic contents are more easily to unravel.
- BUT this is "in principle": gene expression heritability and marker density are interlocked:

	Gene expression heritability	
Chromosome density	High	Low
Dense		
Sparse		

Results 3: Effect of bootstraps



Mitigated good news.

Results 4: comparison of some inference methods

- lasso $\approx$ Dantzig; good options if not interested in giving edge directions.

Results 4: comparison of some inference methods

- lasso $\approx$ Dantzig; good options if not interested in giving edge directions.
- Bayesian networks is an asset but needs large n (and memory !).

Results 4: comparison of some inference methods

- lasso $\approx$ Dantzig; good options if not interested in giving edge directions.
- Bayesian networks is an asset but needs large n (and memory !).
- slightly disapointed by RF but perhaps not good score (used reduction in precision error, not variance reduction) and integrated both markers and expressions at once.

Results 4: comparison of some inference methods

- lasso $\approx$ Dantzig; good options if not interested in giving edge directions.
- Bayesian networks is an asset but needs large n (and memory !).
- slightly disapointed by RF but perhaps not good score (used reduction in precision error, not variance reduction) and integrated both markers and expressions at once.
- Potential complementarity of the methods; could have tried a meta-analysis [Vignes et al 2011] but (i) did not have a p-value-like score and (ii) this method is now old-fashioned I understand correlations do better ;-) !

Results 4: comparison of some inference methods

- lasso $\approx$ Dantzig; good options if not interested in giving edge directions.
- Bayesian networks is an asset but needs large n (and memory !).
- slightly disapointed by RF but perhaps not good score (used reduction in precision error, not variance reduction) and integrated both markers and expressions at once.
- Potential complementarity of the methods; could have tried a meta-analysis [Vignes et al 2011] but (i) did not have a p-value-like score and (ii) this method is now old-fashioned I understand correlations do better ;-) !

Conclusion here: it depends...

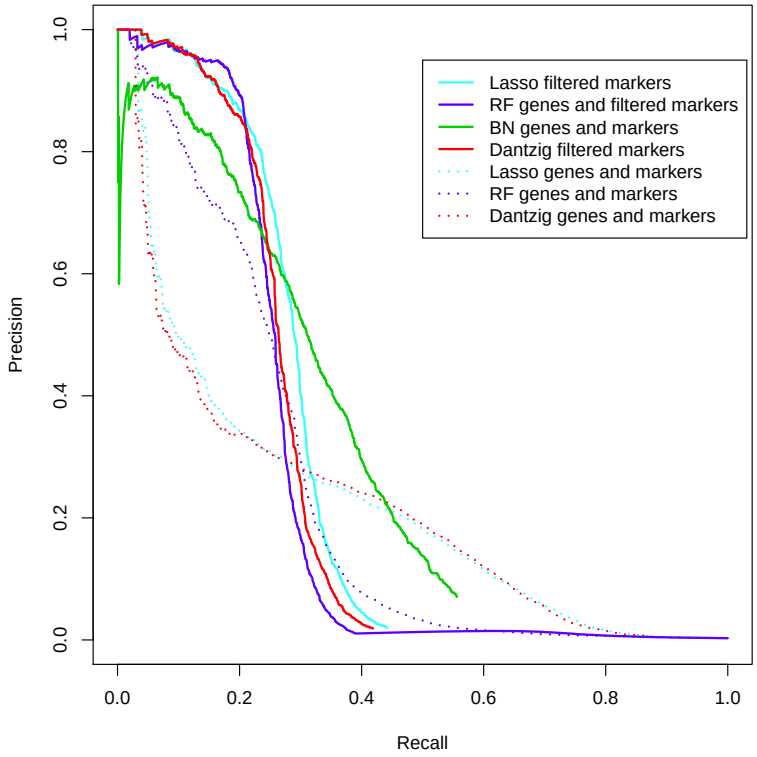
Results 4: comparison of some inference methods

- lasso $\approx$ Dantzig; good options if not interested in giving edge directions.
- Bayesian networks is an asset but needs large n (and memory !).
- slightly disapointed by RF but perhaps not good score (used reduction in precision error, not variance reduction) and integrated both markers and expressions at once.
- Potential complementarity of the methods; could have tried a meta-analysis [Vignes et al 2011] but (i) did not have a p-value-like score and (ii) this method is now old-fashioned I understand correlations do better ;-) !

Conclusion here: it depends...
Don't want to feel to depressed ? [Marbach et al 2012]'s wisdom of crowds: an infinite number of independent (better than random) inference methods is consistent !

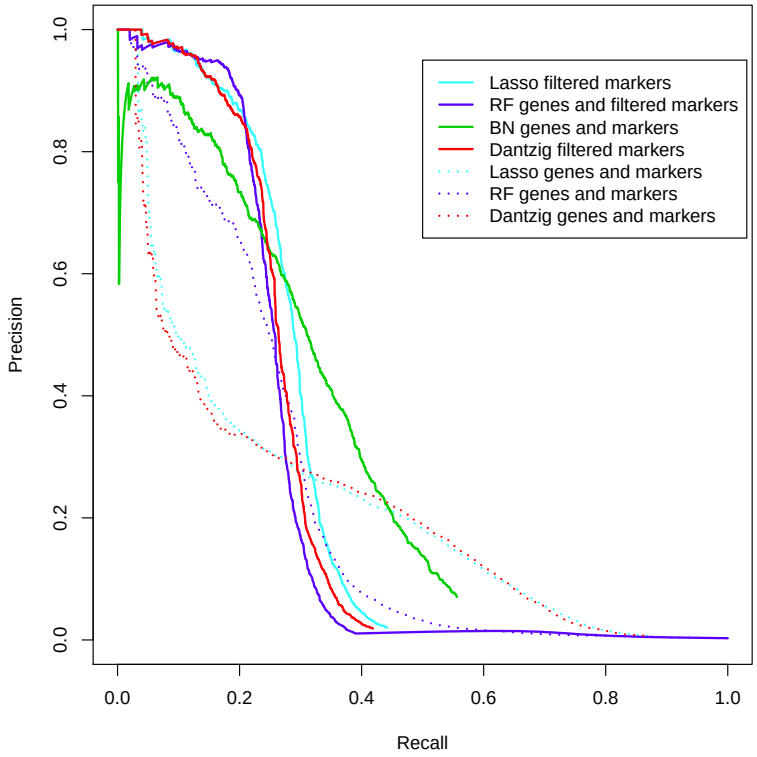
Results 4 encore: comparison of some inference methods

Post-processing also matters !



Results 4 encore: comparison of some inference methods

Post-processing also matters !



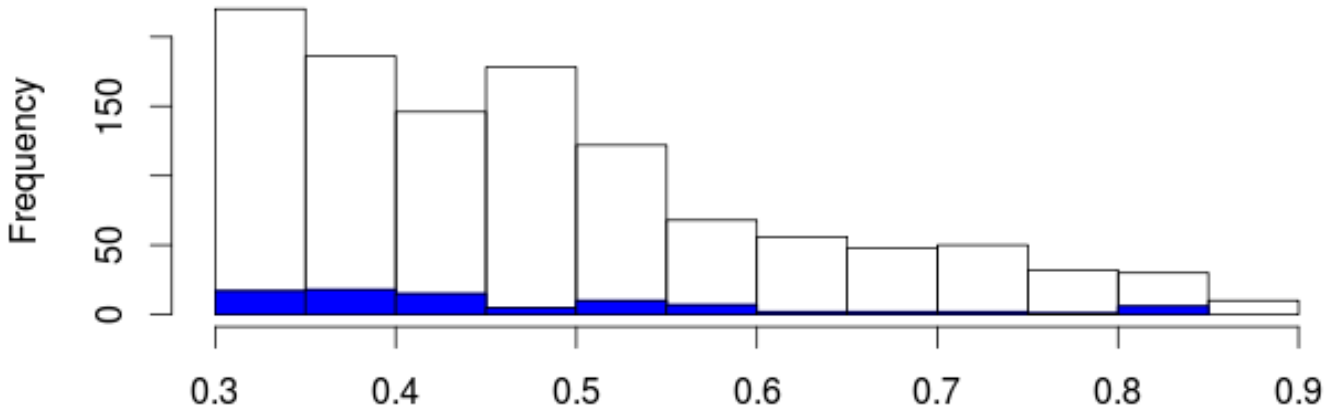
But this may be the other way round on another configuration !

Which edges are we NOT able to grab ?

It's not really an issue of edge direction.

Which edges are we NOT able to grab ?

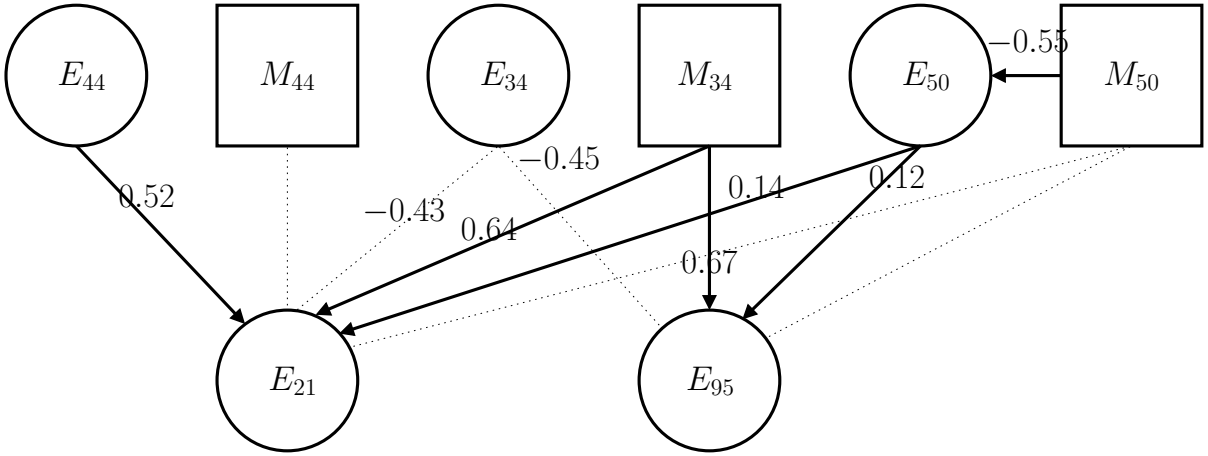
It's not really an issue of edge direction.
But what are we trying to infer: (absolute correlations between gene expressions)



Same situations for correlations between markers and gene expressions

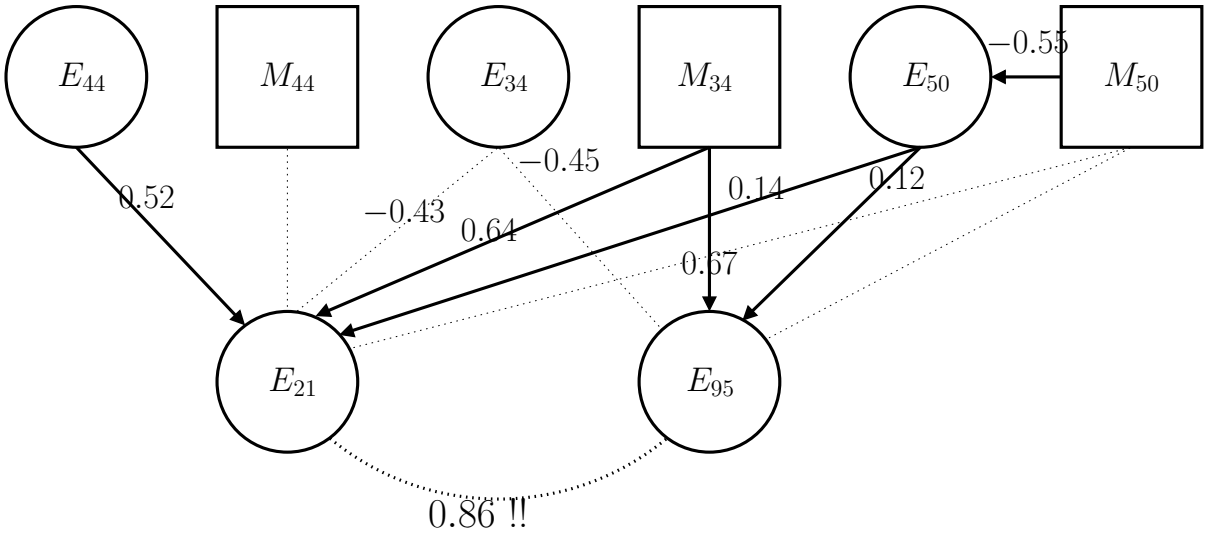
Which edges are we NOT able to grab ?

Into details:



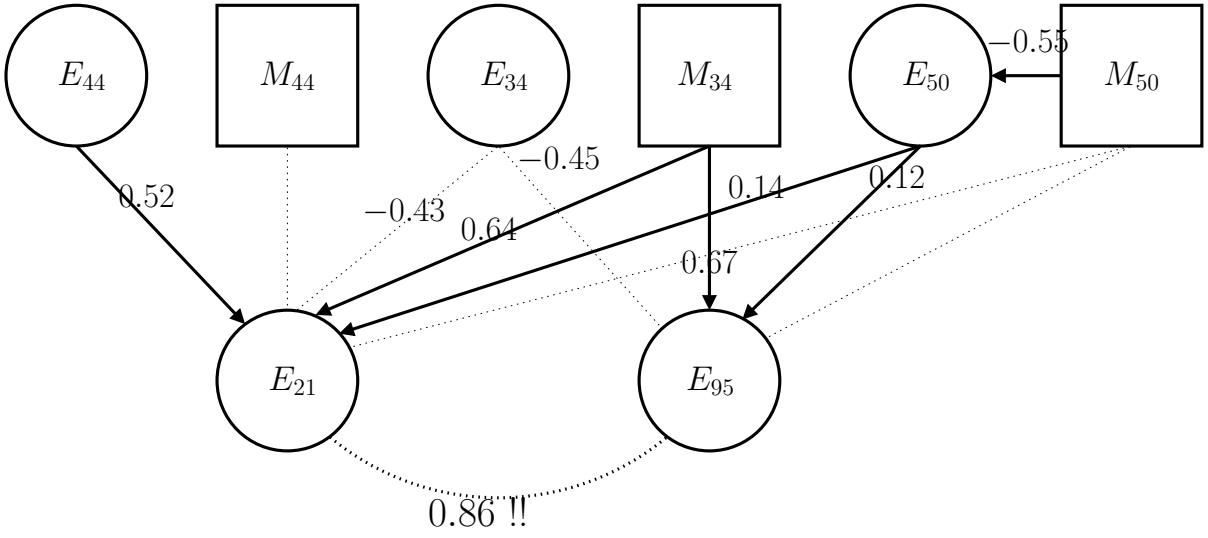
Which edges are we NOT able to grab ?

Into details:



Which edges are we NOT able to grab ?

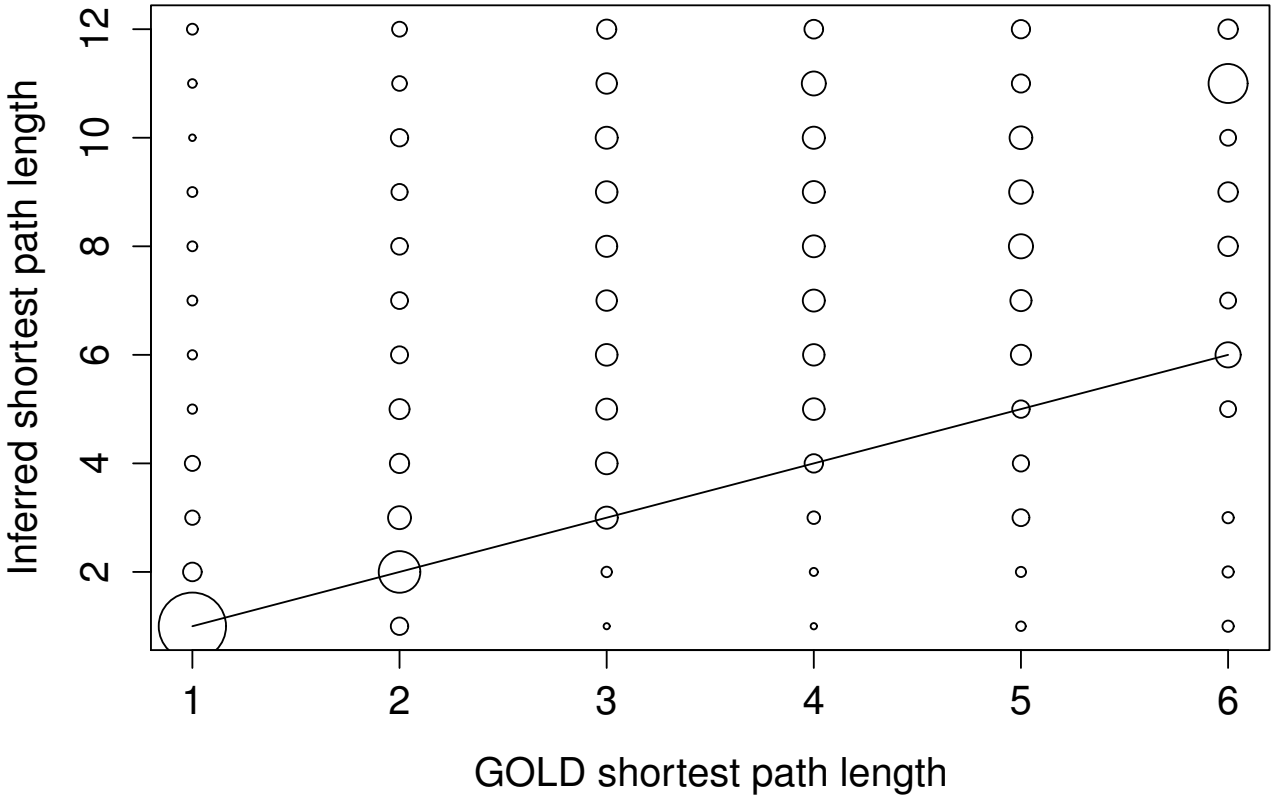
Into details:



Logically, we get $G_{21} \leftrightarrow G_{95}$ as prediction No. 12, $G_{34} \rightarrow G_{95}$ as No. 19, $G_{44} \rightarrow G_{21}$ as No. 192 (reverse is No. 214), $G_{34} \rightarrow G_{21}$ as No. 252...$G_{50} \rightarrow G_{21}$ is prediction No. 2449, $G_{50} \rightarrow G_{95}$ is No. 6559 . . . and many more FP inbetween !!

Building longer/shorter paths ?

Shorter path length comparison



Conclusions 1

Critical look on our work

- Let's be honest, I thought methodology was almost ready for a nice package that would propose state-of-the-art efficient GRN recovery from Systems Genetics data.

Conclusions 1

Critical look on our work

- Let's be honest, I thought methodology was almost ready for a nice package that would propose state-of-the-art efficient GRN recovery from Systems Genetics data.
- Lessons from this data set analysis: not there yet ! Was it too difficult ? Too exotic ? At least it kept us occupied full-time during summer 2012 !

Conclusions 1

Critical look on our work

- Let's be honest, I thought methodology was almost ready for a nice package that would propose state-of-the-art efficient GRN recovery from Systems Genetics data.
- Lessons from this data set analysis: not there yet ! Was it too difficult ? Too exotic ? At least it kept us occupied full-time during summer 2012 !
- Penalised linear regressions, BN, RF . . . are nice models, which actually capture some important interactions with an acceptable degree of precision but: room for improvement ??

Conclusions 2

Future work

- Still some **work to be done**: try data transform (no magic remedy) ? Other naive/sophisticated Machine Learning tools (neural networks, SI algorithms . . .) ?
  Assess the impact of missing information (function, gene) in the system ? Evaluate the potential to find direct **causal** relationships on other data sets ?

Conclusions 2

Future work

- Still some **work to be done**: try data transform (no magic remedy) ? Other naive/sophisticated Machine Learning tools (neural networks, SI algorithms . . .) ?
  Assess the impact of missing information (function, gene) in the system ? Evaluate the potential to find direct **causal** relationships on other data sets ?
- Biologists might have the answer: we want to do computational biology, not (in fact we do that from time to time) pure mathematics we stick to biological problems. The more complex the biological phenomenon to account for, the more granularity in the model ? At least forward and backward (and vice versa) movement between modelling and experimental validation ! Compare model and biological reality it should represent !!

Conclusions 2

Future work

- Still some **work to be done**: try data transform (no magic remedy) ? Other naive/sophisticated Machine Learning tools (neural networks, SI algorithms . . .) ?
  Assess the impact of missing information (function, gene) in the system ? Evaluate the potential to find direct **causal** relationships on other data sets ?
- Biologists might have the answer: we want to do computational biology, not (in fact we do that from time to time) pure mathematics we stick to biological problems. The more complex the biological phenomenon to account for, the more granularity in the model ? At least forward and backward (and vice versa) movement between modelling and experimental validation ! Compare model and biological reality it should represent !!
- Many thanks for your attention !