From gene clustering to genetical genomics:

Analyzing or reconstructing biological networks

Matthieu Vignes[1] Jimmy Vandel[1] Nathalie Keussayan[1]
Juliette Blanchet[2] Simon de Givry[1] Brigitte Mangin[1]

[1]BIA Unit - INRA Toulouse
Castanet Tolosan, France

[2]WLF/SLF
Davos, Switzerland

Gensys, ECCS'09 - Warwick, UK
September 2009

Outline

1. **Introduction and biological issues**
   - Causal relationships: from genotype to phenotype
   - Genetical genomics
2. **Gene expression clustering with missing observations in a Markovian setting**
   - Model-based approach with Markovian dependencies
   - Leads to use Markovian modelling in a genetical genomics context
3. **Reconstruction of networks combining genetic and genomics data**
   - Existing methods
   - Artificial data set simulation
   - Learning with Bayesian Networks or with a lasso SEM regression
   - Preliminary results

Outline

Inherited phenotypes have genetic roots

- Phenotype: observed characteristic (anatomical, morphological, molecular, physiological, ethological) or *trait* in a living organism. Many of which are inherited from parents (Mendel's peas...).

- Polymorphisms (*several shapes*) control gene expression or the affinity between a protein and its target. Can be (i) complex and (ii) quantitative ($\neq$ discrete).
- Traits carried out by DNA. Information unit (for constructing and operating an organism) = gene with different forms or alleles whose inheritance is complicated by recombination of chromosomes (diploids).

Inherited phenotypes have genetic roots

- Phenotype: observed characteristic (anatomical, morphological, molecular, physiological, ethological) or *trait* in a living organism. Many of which are inherited from parents (Mendel's peas...).

- Polymorphisms (*several shapes*) control gene expression or the affinity between a protein and its target. Can be (i) complex and (ii) quantitative ($\neq$ discrete).
- Traits carried out by DNA. Information unit (for constructing and operating an organism) = gene with different forms or alleles whose inheritance is complicated by recombination of chromosomes (diploids).

Inherited phenotypes have genetic roots

- Phenotype: observed characteristic (anatomical, morphological, molecular, physiological, ethological) or *trait* in a living organism. Many of which are inherited from parents (Mendel's peas...).

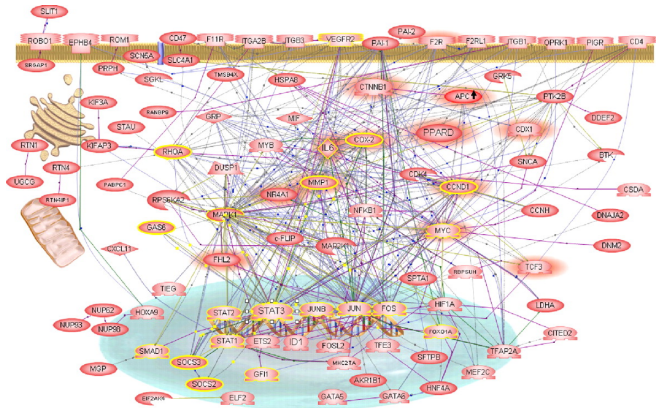
- Polymorphisms (*several shapes*) control gene expression or the affinity between a protein and its target. Can be (i) complex and (ii) quantitative ($\neq$ discrete).
- Traits carried out by DNA. Information unit (for constructing and operating an organism) $=$ gene with different forms or alleles whose inheritance is complicated by recombination of chromosomes (diploids).

Gene Regulatory Networks

- Mutations on DNA seq.: random events that can create a new allele hence new trait(s) when viable → Basis for evolution.
- Links, causal dependencies between genes or genes and their products are represented into a Gene Regulatory Networks (GRN).

Gene Regulatory Networks

- Mutations on DNA seq.: random events that can create a new allele hence new trait(s) when viable → Basis for evolution.
- Links, causal dependencies between genes or genes and their products are represented into a Gene Regulatory Networks (GRN).



Angiogenic signaling network (Adollahi et al. 2007)

Gene Regulatory Networks

- Mutations on DNA seq.: random events that can create a new allele hence new trait(s) when viable $\rightarrow$ Basis for evolution.
- Links, causal dependencies between genes or genes and their products are represented into a Gene Regulatory Networks (GRN).
- Abundance of genomics data (=measurements of cell component activity). Can be directly used to infer GRN (Wehrli et al. 2006, Bansal et al. 2007).

Avowed biological target

Genetical genomics

Combine genetic information (perturbation of the network) and genomics measures (Jansen & Nap 2001) because...

- Biological goal: Understand genetic mechanisms (i) allowing observed diversity and (ii) able to accomplish many diverse functions.
- More pragmatic goal: exploiting genetic context and observed (e-)traits to reconstruct GRN or less ambitiously: identify genes with strong regulatory roles.

With...High levels of measurement replication: each allele at each QTL present in a large number of samples → the effect of the QTL on gene expression will therefore be measured many times.

Avowed biological target

Genetical genomics

Combine genetic information (perturbation of the network) and genomics measures (Jansen & Nap 2001) because...

- Biological goal: Understand genetic mechanisms (i) allowing observed diversity and (ii) able to accomplish many diverse functions.
- More pragmatic goal: exploiting genetic context and observed (e-)traits to reconstruct GRN or less ambitiously: identify genes with strong regulatory roles.

With...High levels of measurement replication: each allele at each QTL present in a large number of samples → the effect of the QTL on gene expression will therefore be measured many times.

Avowed biological target

Genetical genomics

Combine genetic information (perturbation of the network) and genomics measures (Jansen & Nap 2001) because...

- Biological goal: Understand genetic mechanisms (i) allowing observed diversity and (ii) able to accomplish many diverse functions.
- More pragmatic goal: exploiting genetic context and observed (e-)traits to reconstruct GRN or less ambitiously: identify genes with strong regulatory roles.

With...High levels of measurement replication: each allele at each QTL present in a large number of samples → the effect of the QTL on gene expression will therefore be measured many times.

Avowed biological target

Genetical genomics

Combine genetic information (perturbation of the network) and genomics measures (Jansen & Nap 2001) because...

- Biological goal: Understand genetic mechanisms (i) allowing observed diversity and (ii) able to accomplish many diverse functions.
- More pragmatic goal: exploiting genetic context and observed (e-)traits to reconstruct GRN or less ambitiously: identify genes with strong regulatory roles.

With...High levels of measurement replication: each allele at each QTL present in a large number of samples → the effect of the QTL on gene expression will therefore be measured many times.

Avowed biological target

Genetical genomics

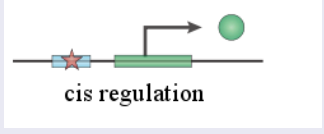
Combine genetic information (perturbation of the network) and genomics measures (Jansen & Nap 2001) because...

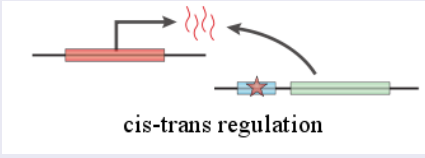
- Biological goal: Understand genetic mechanisms (i) allowing observed diversity and (ii) able to accomplish many diverse functions.
- More pragmatic goal: exploiting genetic context and observed (e-)traits to reconstruct GRN or less ambitiously: identify genes with strong regulatory roles.

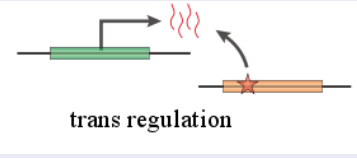
With...High levels of measurement replication: each allele at each QTL present in a large number of samples → the effect of the QTL on gene expression will therefore be measured many times.

Biological ingredients

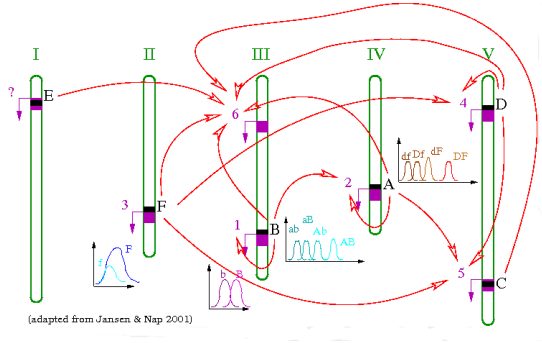
3 mechanisms to link genotype to the observed e-traits



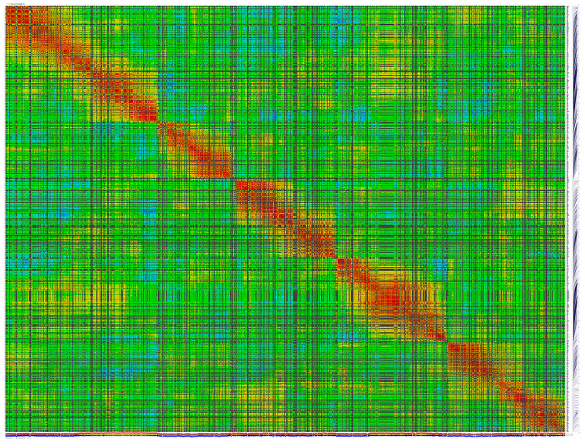




$\oplus$



Physical map



Linkage map

Biological findings (unavowed)

- Unanswered questions so far: (i) number of loci that underlie variation in heritable phenotypes, (ii) distribution of their effect sizes, (iii) their molecular natures, (iv) mechanisms of action and interaction and (v) their dependencies on environmental variables.

- Applications: medical and agricultural genetics, genetic engineering as well as in basic evolutionary biology.

Biological findings (unavowed)

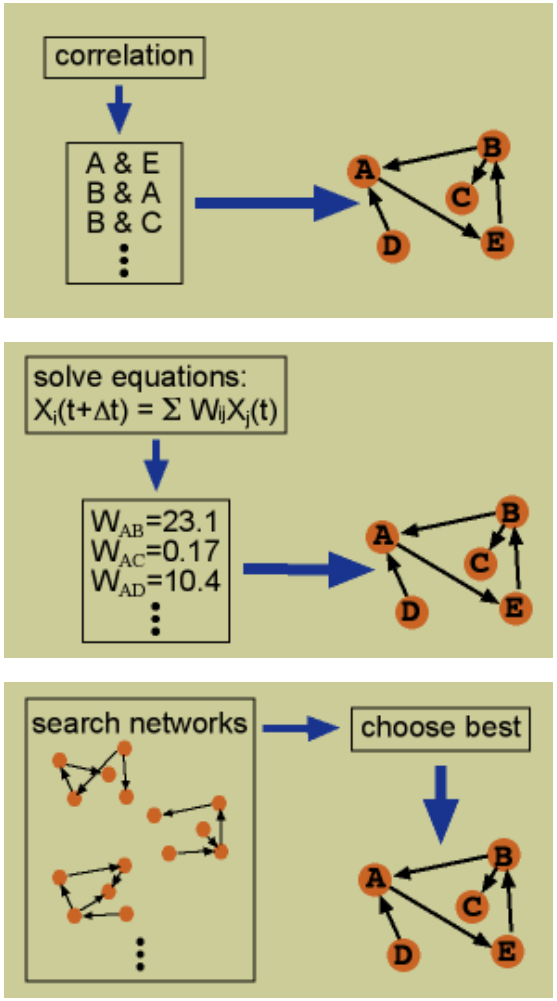
- Unanswered questions so far: (i) number of loci that underlie variation in heritable phenotypes, (ii) distribution of their effect sizes, (iii) their molecular natures, (iv) mechanisms of action and interaction and (v) their dependencies on environmental variables.

- Applications: medical and agricultural genetics, genetic engineering as well as in basic evolutionary biology.

Learning GRN from expression data

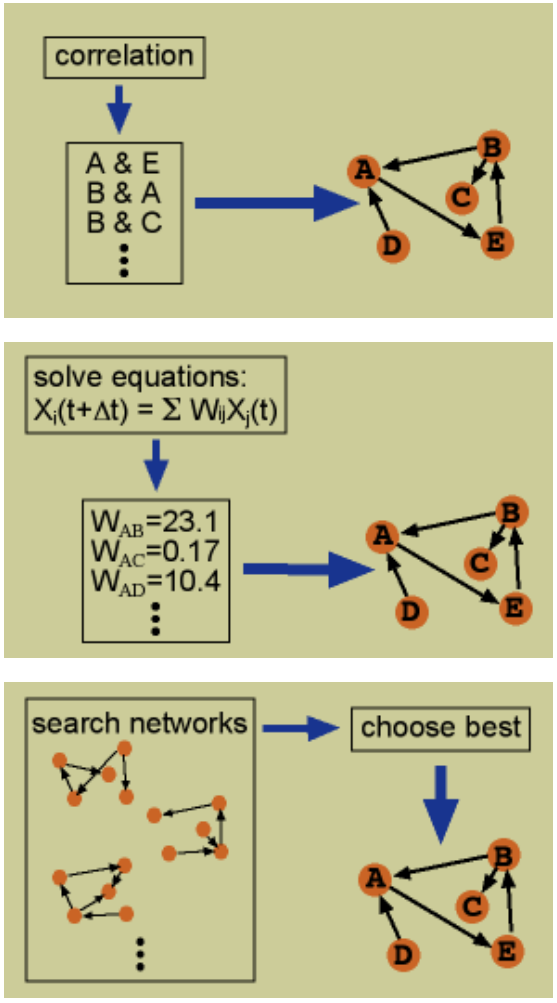
- Pairwise algorithms (correlation, mutual information, hierarchical clustering...).
- Differential equation modelling.
- Network-based algorithms (boolean networks, dynamic/discrete BN...)

(Bansal et al. 2007 and V.A. Smith's website)



Learning GRN from expression data

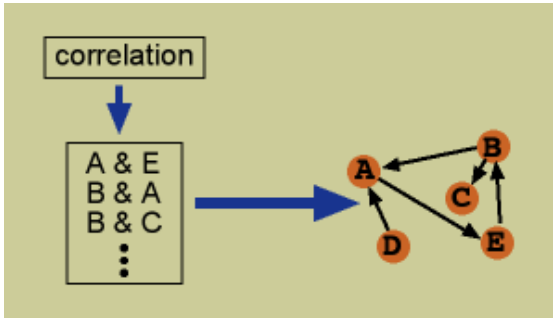
- Pairwise algorithms (correlation, mutual information, hierarchical clustering...).
- Differential equation modelling.
- Network-based algorithms (boolean networks, dynamic/discrete BN...)



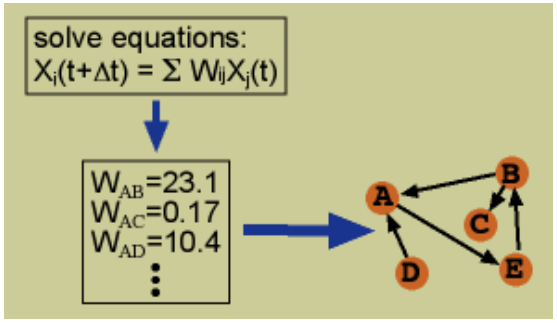
(Bansal et al. 2007 and V.A. Smith's website)

Learning GRN from expression data

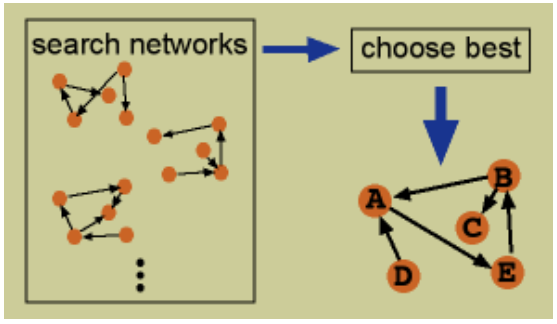
- Pairwise algorithms (correlation, mutual information, hierarchical clustering...).



- Differential equation modelling.



- Network-based algorithms (boolean networks, dynamic/discrete BN...)



(Bansal et al. 2007 and V.A. Smith's website)

Outline

Gene clustering with missing observations in a Markovian setting

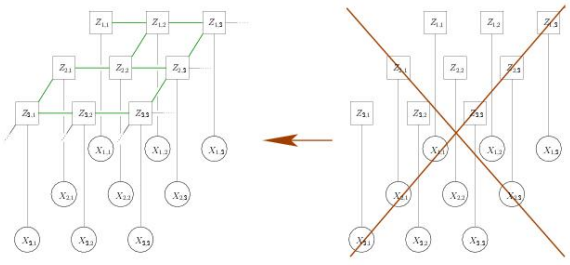
- Data: omics measurements on individual biological entities & interactions between these entities (from experimental evidence or derived: litterature, genomic context, co-expression...).
- Network information in Markov Random Field (MRF).
- Observations modelled conditionally on node status through probabilistic distributions (e.g. Gaussian distribution specifically built for high-dimensional data, Bouveyron et al., Comput. Statist. Data Analysis 2007) so accounting for noise.

Gene clustering with missing observations in a Markovian setting

- Data: omics measurements on individual biological entities & interactions between these entities (from experimental evidence or derived: litterature, genomic context, co-expression...).
- Network information in Markov Random Field (MRF).
- Observations modelled conditionally on node status through probabilistic distributions (e.g. Gaussian distribution specifically built for high-dimensional data, Bouveyron et al., Comput. Statist. Data Analysis 2007) so accounting for noise.

Gene clustering with missing observations in a Markovian setting

- Data: omics measurements on individual biological entities & interactions between these entities (from experimental evidence or derived: litterature, genomic context, co-expression...).
- Network information in Markov Random Field (MRF).
- Observations modelled conditionally on node status through probabilistic distributions (e.g. Gaussian distribution specifically built for high-dimensional data, Bouveyron et al., Comput. Statist. Data Analysis 2007) so accounting for noise.



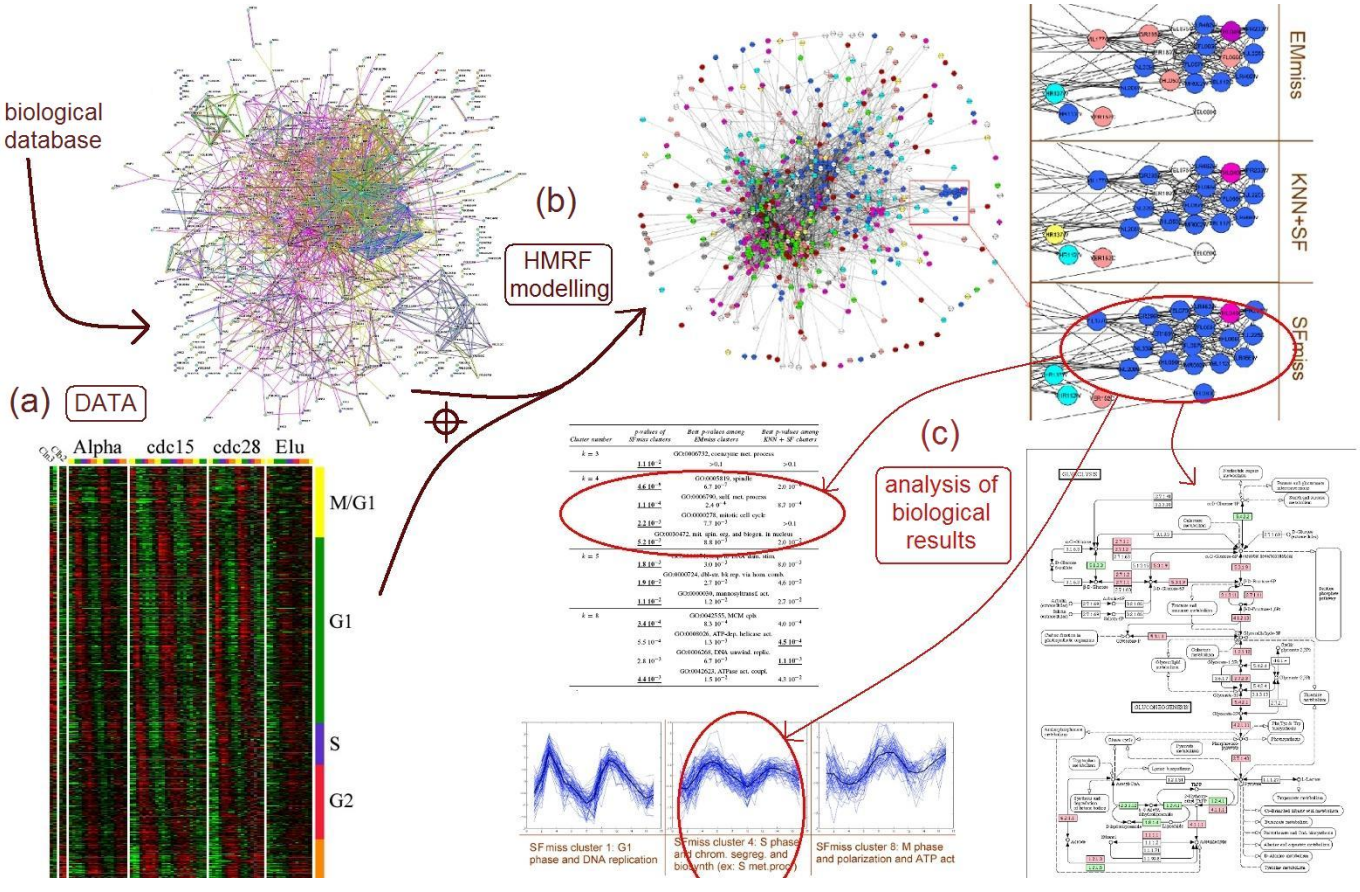
Gene clustering with missing observations in a Markovian setting

- Data: omics measurements on individual biological entities & interactions between these entities (from experimental evidence or derived: litterature, genomic context, co-expression...).
- Network information in Markov Random Field (MRF).
- Observations modelled conditionally on node status through probabilistic distributions (e.g. Gaussian distribution specifically built for high-dimensional data, Bouveyron et al., Comput. Statist. Data Analysis 2007) so accounting for noise.

- Novel instantiation of an EM-based algorithm for model estimation: mean-field like approximations and accounting for missing observations (MAR).

Model-based approach with Markovian dependencies

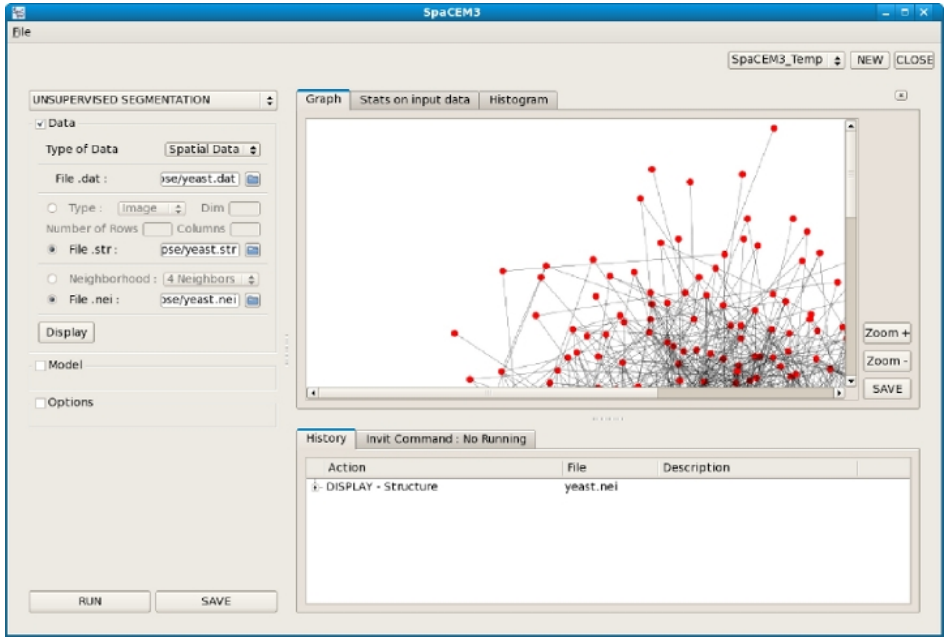
Workflow of a computational biology data analysis with our method



(from Blanchet & Vignes, J. Comput. Biol. 2009)

SpaCEM^3 software

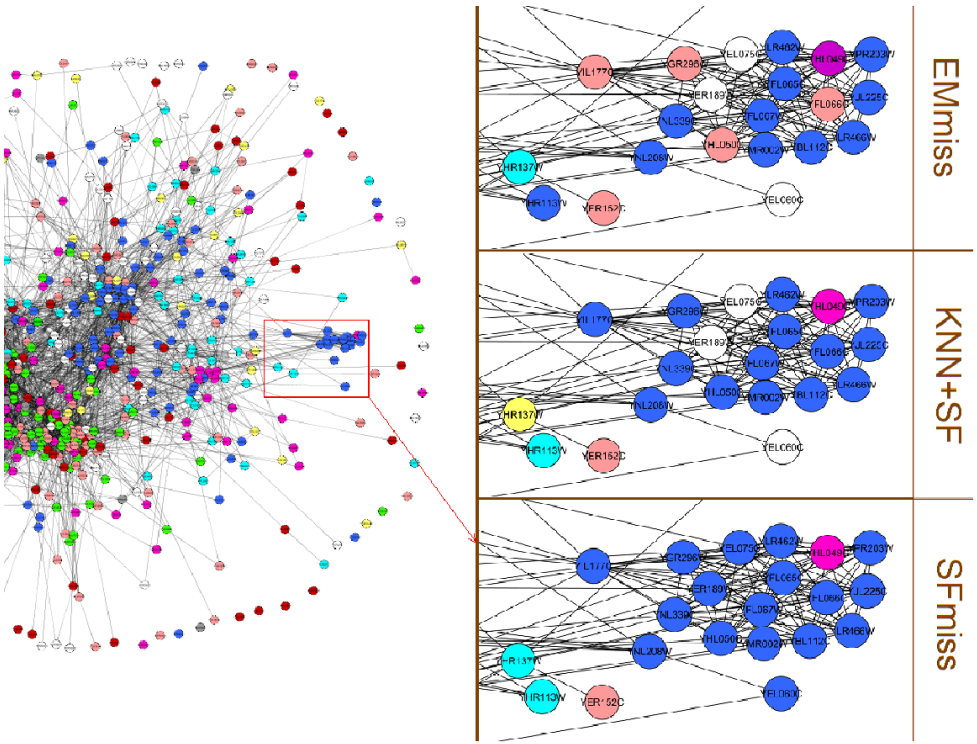
The SpaCEM^3 software allows the user to specify the structure of the model, estimate parameters, select relevant models (BIC, ICL) and visualize the results in the GUI.



(freely available at `http://spacem3.gforge.inria.fr/`)

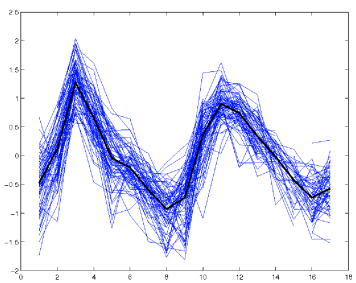
Biological features of clusters

- Modularity
- Interpretability of cluster profiles
- GO term representativity
- Link to metabolic pathways

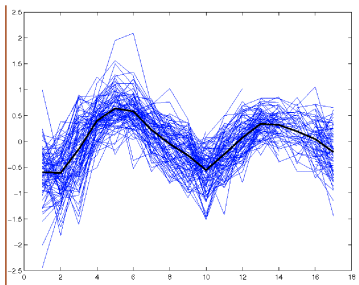


Biological features of clusters

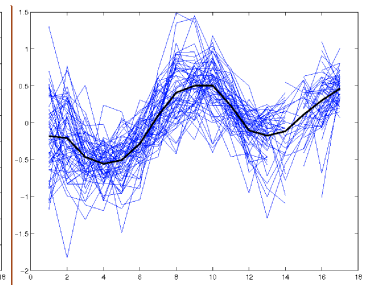
- Modularity
- Interpretability of cluster profiles
- GO term representativity
- Link to metabolic pathways



SFmiss cluster 1: G1 phase and DNA replication



SFmiss cluster 4: S phase and chrom. segreg. and biosynth (ex: S met.proc.)



SFmiss cluster 8: M phase and polarization and ATP act.

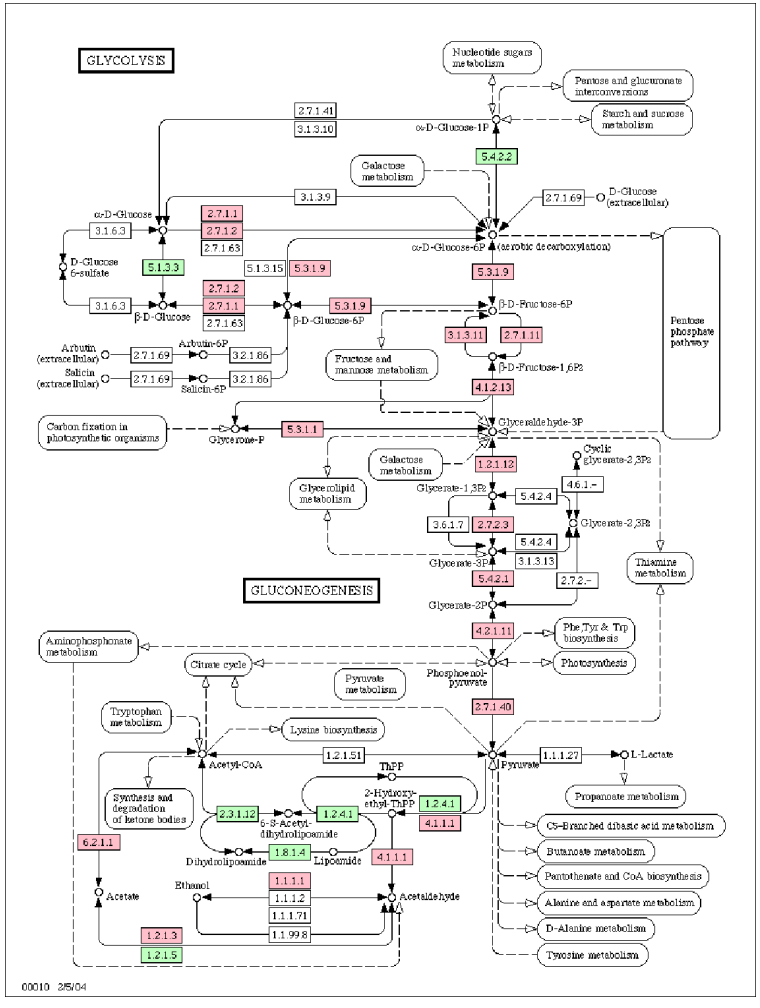
Biological features of clusters

- Modularity
- Interpretability of cluster profiles
- GO term representativity
- Link to metabolic pathways

Cluster number	*p-values of SFmiss clusters*	*Best p-values among EMmiss clusters*	*Best p-values among KNN + SF clusters*
$k = 3$		GO:0006732, coenzyme met. process	
	$\underline{1.1\ 10^{-2}}$	>0.1	>0.1
$k = 4$		GO:0005819, spindle	
	$\underline{4.6\ 10^{-9}}$	$6.7\ 10^{-7}$	$2.0\ 10^{-6}$
		GO:0006790, sulf. met. process	
	$\underline{1.1\ 10^{-4}}$	$2.4\ 0^{-4}$	$8.7\ 10^{-4}$
		GO:0000278, mitotic cell cycle	
	$\underline{2.2\ 10^{-3}}$	$7.7\ 10^{-3}$	>0.1
	GO:0030472, mit. spin. org. and biogen. in nucleus		
	$\underline{5.2\ 10^{-3}}$	$8.8\ 10^{-3}$	$2.0\ 10^{-2}$
$k = 5$		GO:0006974, resp. to DNA dam. stim.	
	$\underline{1.8\ 10^{-3}}$	$3.0\ 10^{-3}$	$8.0\ 10^{-3}$
		GO:0000724, dbl-str. bk rep. via hom. comb.	
	$\underline{1.9\ 10^{-2}}$	$2.7\ 10^{-2}$	$4.6\ 10^{-2}$
		GO:0000030, mannosyltransf. act.	
	$\underline{1.1\ 10^{-2}}$	$1.2\ 10^{-2}$	$2.7\ 10^{-2}$
$k = 8$		GO:0042555, MCM cplx	
	$\underline{3.4\ 10^{-4}}$	$8.3\ 10^{-4}$	$4.0\ 10^{-4}$
		GO:0008026, ATP-dep. helicase act.	
	$5.5\ 10^{-4}$	$1.3\ 10^{-3}$	$\underline{4.5\ 10^{-4}}$
		GO:0006268, DNA unwind. replic.	
	$2.8\ 10^{-3}$	$6.7\ 10^{-3}$	$\underline{1.1\ 10^{-3}}$
		GO:0042623, ATPase act. coupl.	
	$\underline{4.4\ 10^{-3}}$	$1.5\ 10^{-2}$	$4.3\ 10^{-2}$

Biological features of clusters

- Modularity
- Interpretability of cluster profiles
- GO term representativity
- Link to metabolic pathways



Inferring a MRF with genetical genomics data

1. Estimating weights -as a measure of uncertainty- on putative edges and fixing those on edges defined by expert knowledge.
   - ...could lead to the inference of $N(N-1)/2$ parameters.
2. *Triplet Markov fields* (Blanchet & Forbes, IEEE PAMI 2008) allowing objects to be assigned to overlapping subclasses seem an interesting lead to model genetic background of a gene by introducing an additional blanket that could encode genetic dependencies in the population.
   - ...application at present limited to supervised classification. Optimality to include genetics?

Inferring a MRF with genetical genomics data

1. Estimating weights -as a measure of uncertainty- on putative edges and fixing those on edges defined by expert knowledge.
   - ...could lead to the inference of $N(N-1)/2$ parameters.
2. *Triplet Markov fields* (Blanchet & Forbes, IEEE PAMI 2008) allowing objects to be assigned to overlapping subclasses seem an interesting lead to model genetic background of a gene by introducing an additional blanket that could encode genetic dependencies in the population.
   - ...application at present limited to supervised classification. Optimality to include genetics?

Inferring a MRF with genetical genomics data

1. Estimating weights -as a measure of uncertainty- on putative edges and fixing those on edges defined by expert knowledge.
   - ...could lead to the inference of $N(N-1)/2$ parameters.
2. *Triplet Markov fields* (Blanchet & Forbes, IEEE PAMI 2008) allowing objects to be assigned to overlapping subclasses seem an interesting lead to model genetic background of a gene by introducing an additional blanket that could encode genetic dependencies in the population.
   - ...application at present limited to supervised classification. Optimality to include genetics?

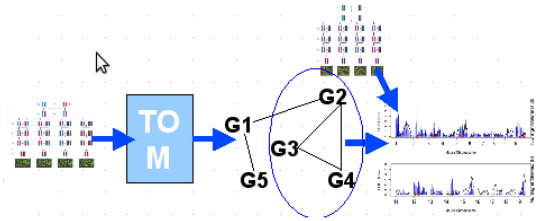
Inferring a MRF with genetical genomics data

1. Estimating weights -as a measure of uncertainty- on putative edges and fixing those on edges defined by expert knowledge.
   - ...could lead to the inference of $N(N-1)/2$ parameters.
2. *Triplet Markov fields* (Blanchet & Forbes, IEEE PAMI 2008) allowing objects to be assigned to overlapping subclasses seem an interesting lead to model genetic background of a gene by introducing an additional blanket that could encode genetic dependencies in the population.
   - ...application at present limited to supervised classification. Optimality to include genetics?

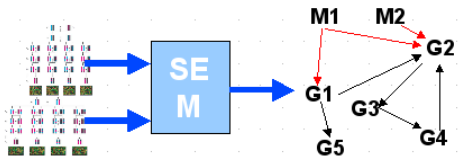
Outline

Learning networks in genetical genomics

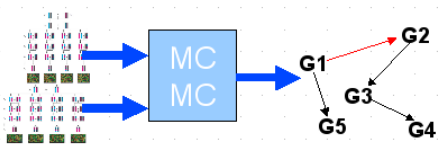
- Pairwise algo. (Ghazalpour et al., PLOS Gen., 2006) co-expression network + module cis-eQTL



- Equation-based algo. (Liu et al., Genetics, 2008): greedy SEM with expr. levels and genotypes as covar., pre-filtered by eQTL info.
  - Nathalie Keussayan's MSc. (with Brigitte Mangin).

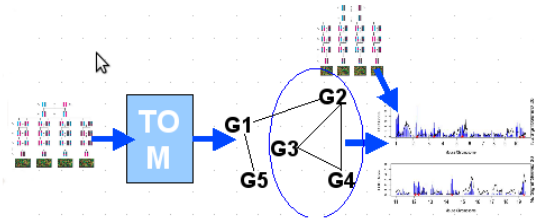


- Network-based algo. (Zhu et al., PLoS Comput. Biol., 2007): MCMC algo. on BN structures with BIC and eQTL info. as a prior.
  - Jimmy Vandel MSc. (with Simon de Givry). Staying with us for a PhD .

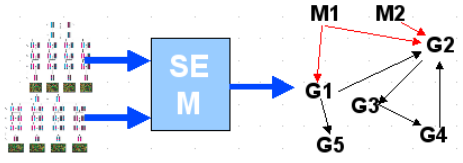


Learning networks in genetical genomics

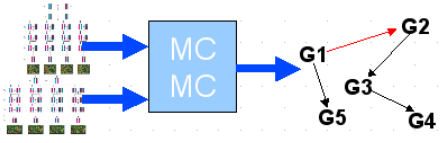
- Pairwise algo. (Ghazalpour et al., PLOS Gen., 2006) co-expression network + module cis-eQTL



- Equation-based algo. (Liu et al., Genetics, 2008): greedy SEM with expr. levels and genotypes as covar., pre-filtered by eQTL info.
  - Nathalie Keussayan's MSc. (with Brigitte Mangin).

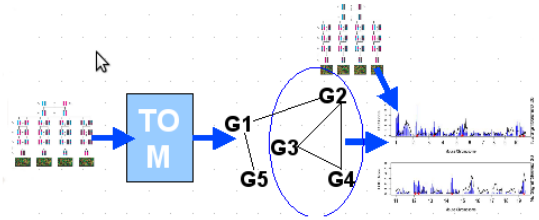


- Network-based algo. (Zhu et al., PLoS Comput. Biol., 2007): MCMC algo. on BN structures with BIC and eQTL info. as a prior.
  - Jimmy Vandel MSc. (with Simon de Givry). Staying with us for a PhD .

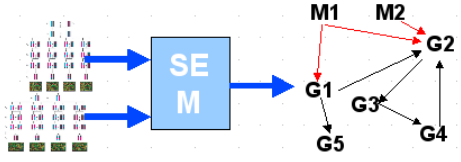


Learning networks in genetical genomics

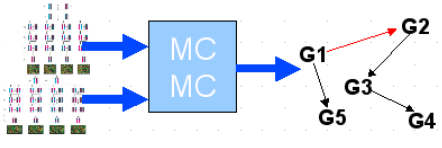
- Pairwise algo. (Ghazalpour et al., PLOS Gen., 2006) co-expression network + module cis-eQTL



- Equation-based algo. (Liu et al., Genetics, 2008): greedy SEM with expr. levels and genotypes as covar., pre-filtered by eQTL info.
  - ▷ Nathalie Keussayan's MSc. (with Brigitte Mangin).

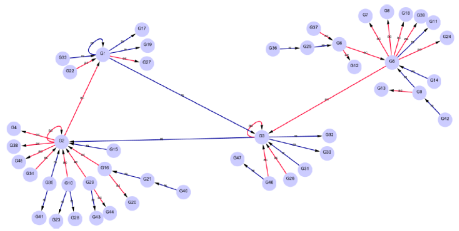


- Network-based algo. (Zhu et al., PLoS Comput. Biol., 2007): MCMC algo. on BN structures with BIC and eQTL info. as a prior.
  - ▷ Jimmy Vandel MSc. (with Simon de Givry). Staying with us for a PhD 🙂.



A recipe for genetical genomics artificial dataset generation

- Choose a network with features as close as possible to know features of realistic biological networks → `http://www.comp-sys-bio.org/AGN/`.

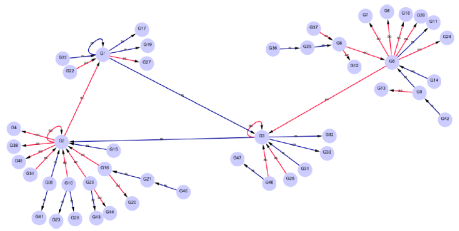


- Simulate genotype from a RIL population: pop size, chromosome size, number and distribution of markers (incl.error and missingness) → CarthaGène.
- Compute gene expression data from gene activity ODE → COmplex PAthway SImulator (COPASI, `http://www.copasy.org/`) for steady-state expression levels.

Note: expr. levels need to be discretized with BN: k-means

A recipe for genetical genomics artificial dataset generation

- Choose a network with features as close as possible to know features of realistic biological networks → `http://www.comp-sys-bio.org/AGN/`.

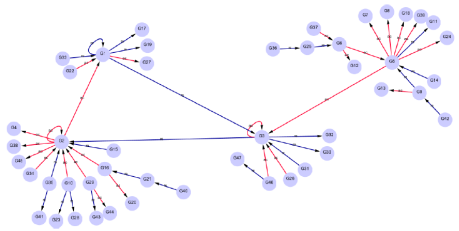


- Simulate genotype from a RIL population: pop size, chromosome size, number and distribution of markers (incl.error and missingness) → CarthaGène.
- Compute gene expression data from gene activity ODE → COmplex PAthway SImulator (COPASI, `http://www.copasy.org/`) for steady-state expression levels.

  Note: expr. levels need to be discretized with BN: k-means

A recipe for genetical genomics artificial dataset generation

- Choose a network with features as close as possible to know features of realistic biological networks → `http://www.comp-sys-bio.org/AGN/`.

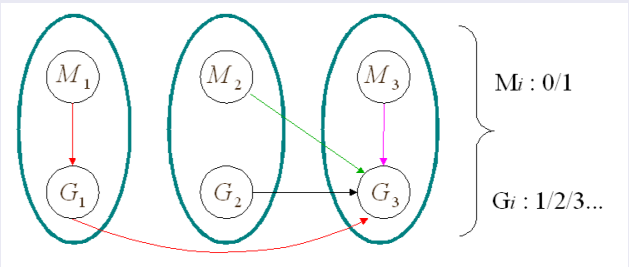


- Simulate genotype from a RIL population: pop size, chromosome size, number and distribution of markers (incl.error and missingness) → CarthaGène.
- Compute gene expression data from gene activity ODE → COmplex PAthway SImulator (COPASI, `http://www.copasy.org/`) for steady-state expression levels.
  Note: expr. levels need to be discretized with BN: k-means,

Bayesian Networks (BN)

Definition of BN

Directed Acyclic Graph (DAG) & $P(V) = \prod_{i=1}^{p} P(V_i \mid V_{pa(V_i)})$, with $V_i := M_i \otimes G_i$. Clever init.: encompassing network with putative eQTL $\rightarrow$ MCQTL `http://carlit.toulouse.inra.fr/MCQTL/`.



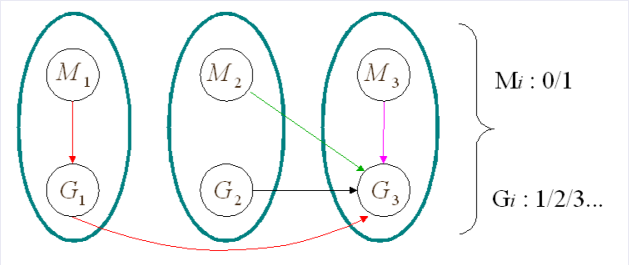
Tested Algorithms (Matlab's BayesNet, K.Murphy and P. Leray)

1. Scoring algorithms: BIC (+ penalty for genetic linkage) with structure exploration strategies: Maximum Weight Spanning Tree (MWST), K2 (node ordering), Greedy Search (GS).
2. Independance algorithms: χ^2 or Likelihood Ratio Test (LRT) with PC or BNPC.

Bayesian Networks (BN)

Definition of BN

Directed Acyclic Graph (DAG) & $P(V) = \prod_{i=1}^{p} P(V_i \mid V_{pa(V_i)})$, with $V_i := M_i \otimes G_i$. Clever init.: encompassing network with putative eQTL $\rightarrow$ MCQTL `http://carlit.toulouse.inra.fr/MCQTL/`.



Tested Algorithms (Matlab's BayesNet, K.Murphy and P. Leray)

1. Scoring algorithms: BIC (+ penalty for genetic linkage) with structure exploration strategies: Maximum Weight Spanning Tree (MWST), K2 (node ordering), Greedy Search (GS).
2. Independance algorithms: χ^2 or Likelihood Ratio Test (LRT) with PC or BNPC.

Structural Equation Modelling (SEM)

$$\triangleright\ Y = Y.B + X.\Theta + \epsilon$$

where:
Y matrix of transcript levels ($n \times p$)
X matrix of genotypes ($n \times q$)
B_{km} direct effect of level of gene k on level of gene m ($B_{ii} = 0$).
Θ_{jm} direct effect of marker j on expression of gene m.

▷ Gene-by-gene regression

$$Y_k = Y_{\setminus k} * \beta_k + X * \Theta_k + \epsilon_k$$

β_k's and Θ_k's need to be estimated as regression coefficients.

▷ Values signif. $\neq 0$ allow us to infer network structure.

Structural Equation Modelling (SEM)

$$\triangleright\ Y = Y.B + X.\Theta + \epsilon$$

where:
Y matrix of transcript levels ($n \times p$)
X matrix of genotypes ($n \times q$)
B_{km} direct effect of level of gene k on level of gene m ($B_{ii} = 0$).
Θ_{jm} direct effect of marker j on expression of gene m.

▷ Gene-by-gene regression

$$Y_k = Y_{\setminus k} * \beta_k + X * \Theta_k + \epsilon_k$$

β_k's and Θ_k's need to be estimated as regression coefficients.

▷ Values signif. $\neq 0$ allow us to infer network structure.

Structural Equation Modelling (SEM)

$$\triangleright\ Y = Y.B + X.\Theta + \epsilon$$

where:
Y matrix of transcript levels ($n \times p$)
X matrix of genotypes ($n \times q$)
B_{km} direct effect of level of gene k on level of gene m ($B_{ii} = 0$).
Θ_{jm} direct effect of marker j on expression of gene m.

▷ Gene-by-gene regression

$$Y_k = Y_{\setminus k} * \beta_k + X * \Theta_k + \epsilon_k$$

β_k's and Θ_k's need to be estimated as regression coefficients.

▷ Values signif. $\neq 0$ allow us to infer network structure.

Lasso estimation of parameters

- Idea 1 *Least Square*: unbiased but variance on estimator becomes a problem since typically $n \ll p$.
- Idea 2 *Biased estimations*: v2.α ridge (not parcimonious), v2.β best subset (fixed number of variables can have coef.$\neq 0$), v2.final Lasso (Tibshirani J. Royal. Statist. Soc B. 1996, selects and reduces variables).

$$\widehat{\beta_k} = \arg\min \left[|Y_k - [Y_{\setminus k} X].\beta_k|_{L_2} + \lambda |\beta_k|_{L_1} \right]$$
$$(|\widehat{\beta_k}|_{L_1} \leq \tau,\ \beta_k =^t [B_k\, \theta_k])$$

We used the Least Angle Regression (LAR) algo. (lars in R) to compute $X.\hat{\beta}$, with cross-validation, BIC and Meinshausen criteria to determine the best λ.

Lasso estimation of parameters

- Idea 1 *Least Square*: unbiased but variance on estimator becomes a problem since typically $n \ll p$.
- Idea 2 *Biased estimations*: v2.α ridge (not parcimonious), v2.β best subset (fixed number of variables can have coef.$\neq 0$), v2.final Lasso (Tibshirani J. Royal. Statist. Soc B. 1996, selects and reduces variables).

$$\widehat{\beta_k} = \arg\min \left[|Y_k - [Y_{\setminus k} X].\beta_k|_{L_2} + \lambda |\beta_k|_{L_1} \right]$$
$$(|\widehat{\beta_k}|_{L_1} \leq \tau,\ \beta_k =^t [B_k\, \theta_k])$$

We used the Least Angle Regression (LAR) algo. (lars in R) to compute $X.\hat{\beta}$, with cross-validation, BIC and Meinshausen criteria to determine the best λ.

Lasso estimation of parameters

- Idea 1 *Least Square*: unbiased but variance on estimator becomes a problem since typically $n \ll p$.
- Idea 2 *Biased estimations*: v2.α ridge (not parcimonious), v2.β best subset (fixed number of variables can have coef.$\neq 0$), v2.final Lasso (Tibshirani J. Royal. Statist. Soc B. 1996, selects and reduces variables).

$$\widehat{\beta_k} = \arg\min \left[|Y_k - [Y_{\setminus k} X].\beta_k|_{L_2} + \lambda |\beta_k|_{L_1} \right]$$
$$(|\widehat{\beta_k}|_{L_1} \leq \tau,\ \beta_k =^t [B_k\, \theta_k])$$

We used the Least Angle Regression (LAR) algo. (lars in R) to compute $X.\hat{\beta}$, with cross-validation, BIC and Meinshausen criteria to determine the best λ.

Lasso estimation of parameters

- Idea 1 *Least Square*: unbiased but variance on estimator becomes a problem since typically $n \ll p$.
- Idea 2 *Biased estimations*: v2.α ridge (not parcimonious), v2.β best subset (fixed number of variables can have coef.$\neq 0$), v2.final Lasso (Tibshirani J. Royal. Statist. Soc B. 1996, selects and reduces variables).

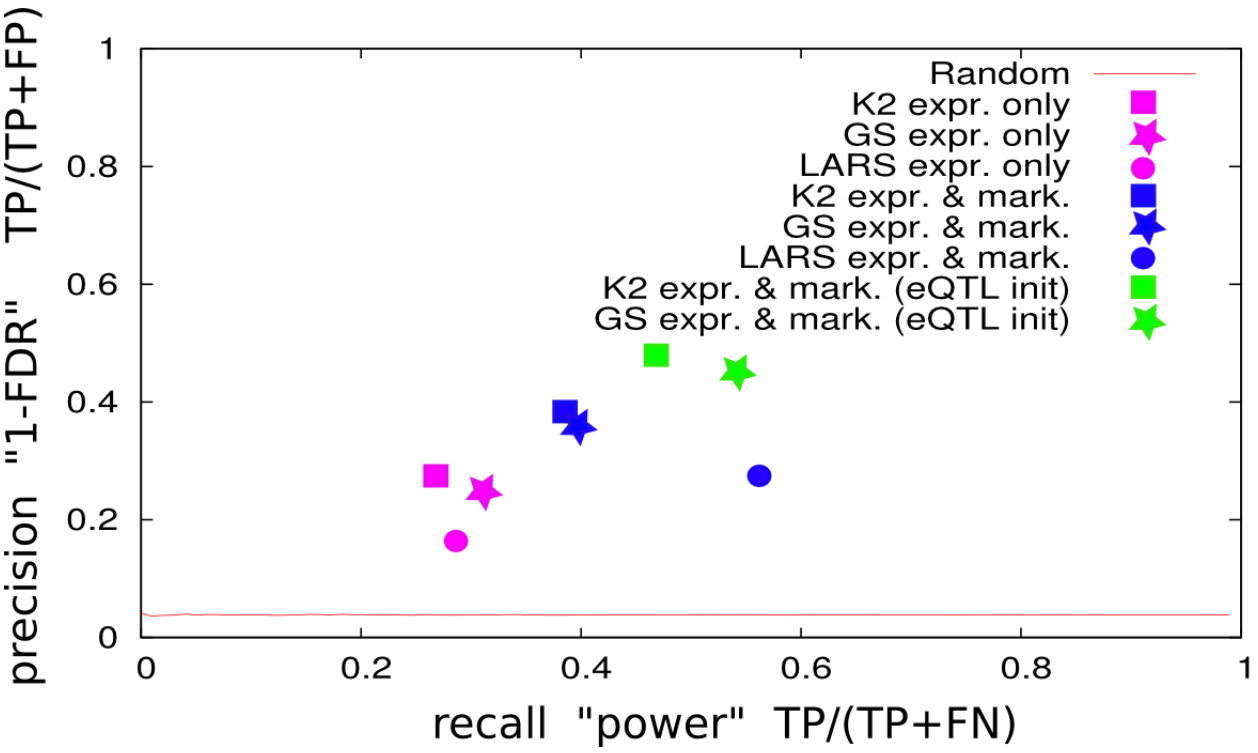
$$\widehat{\beta_k} = \arg\min \left[|Y_k - [Y_{\setminus k} X].\beta_k|_{L_2} + \lambda |\beta_k|_{L_1} \right]$$
$$(|\widehat{\beta_k}|_{L_1} \leq \tau,\ \beta_k =^t [B_k\, \theta_k])$$

We used the Least Angle Regression (LAR) algo. (lars in R) to compute $X.\hat{\beta}$, with cross-validation, BIC and Meinshausen criteria to determine the best λ.

BN vs. SEM: advantages and drawbacks

	BN	SEM
Computational time	☹	☺
Continuous data	☹	☺
Modelling cycles	☹	☺
Param./likelihood estim.	☺	☹
Non-linear dependencies	☺	☹

Results: (i) BN vs. SEM and (ii) with or without genotypes



Network recovery performances on 9 artificial datasets

Summary

Summary

- Panel of different methods to deal with genetical genomics data.
- Plausible synthetic data generation (room for improvement!).
- Obvious gain in using genetic information

Open Problems

- Validate/assess algorithms (any others? Elastic Net?) for network structure recovery in genetical genomics.
- Try these methods on a real gold standard dataset (mice, yeast, thaliana ok...What if sunflower or strawberries).

Summary

Summary

- Panel of different methods to deal with genetical genomics data.
- Plausible synthetic data generation (room for improvement!).
- Obvious gain in using genetic information

Open Problems

- Validate/assess algorithms (any others? Elastic Net?) for network structure recovery in genetical genomics.
- Try these methods on a real gold standard dataset (mice, yeast, thaliana ok...What if sunflower or strawberries).


R. Tibshirani, Regression shrinkage and selection via the lasso, *J. Royal. Statist. Soc B.*, 58:267-88 (1996).


R. Jansen and J. Nap, Genetical genomics: the added value from segregation, *Trends Gen.*, 17:388-91 (2001).


J. Zhu et al., Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations, *PLoS Comput. Biol.*, 3:e64 (2007).


J. Blanchet and F. Forbes. Triplet Markov fields for the supervised classification of complex structured data, *IEEE PAMI*, 30:1055-67 (2008).


B. Liu et al., Gene network inference via structural equation modeling in genetical genomics experiments, *Genetics*, 178:1763-76 (2008).


M.V. Rockman, Reverse engineering the genotype-phenotype map with natural genetic variation, *Nature*, 456:738-44 (2008).


J. Blanchet and M. Vignes, A model-based approach to gene clustering with missing observations reconstruction in a Markov Random Field framework, *J. Comput. Biol.*, 16:475-86 (2009).

Thanks a lot for your attention!

Questions?