Inferring Causal Phenotype Networks Driven by Expression Gene Mapping

Elias Chaibub Neto & Brian S. Yandell

UW-Madison

June 2009

outline

- QTL-driven directed graphs
  - Assume QTLs known, network unknown
  - Infer links (edges) between pairs of phenotypes (nodes)
    - Based on partial correlation
  - Infer causal direction for edges
  - Chaibub et al. (2008 *Genetics*)
  - Software R/qdg available on CRAN
- Causal graphical models in systems genetics
  - QTLs unknown, network unknown
  - Infer both genetic architecture (QTLs) and pathways (networks)
  - Chaibub et al. (2009 *Ann Appl Statist* tent accept)
  - Software R/QTLnet in preparation for CRAN

QTL-driven directed graphs (R/qdg)

- See edited slides by Elias Chaibub Neto
  - BIOCOMP 2008 talk
  - Chaibub Neto, Ferrara, Attie, Yandell (2008) Inferring causal phenotype networks from segregating populations. *Genetics 179*: 1089-1100.
  - Ferrara et al. Attie (2008) Genetic networks of liver metabolism revealed by integration of metabolic and transcriptomic profiling. *PLoS Genet 4*: e1000034.

Introduction

- Our objective is to learn metabolic pathways from data.
- We represent these pathways by directed networks composed by transcripts, metabolites and clinical traits.
- These phenotypes are quantitative in nature, and can be analyzed using quantitative genetics techniques.

Introduction

- In particular, we use Quantitative Trait Loci (QTL) mapping methods to identify genomic regions affecting the phenotypes.

- Since variations in the genotypes (QTLs) cause variations in the phenotypes, but not the other way around, we can unambiguously determine the causal direction

$$\text{QTL} \Rightarrow \text{phenotype}$$

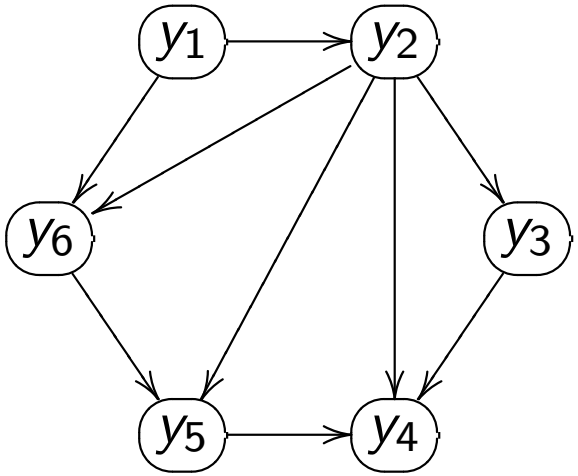
- Knowing that a QTL causally affects a phenotype will allow us to infer causal direction between phenotypes.

PC algorithm

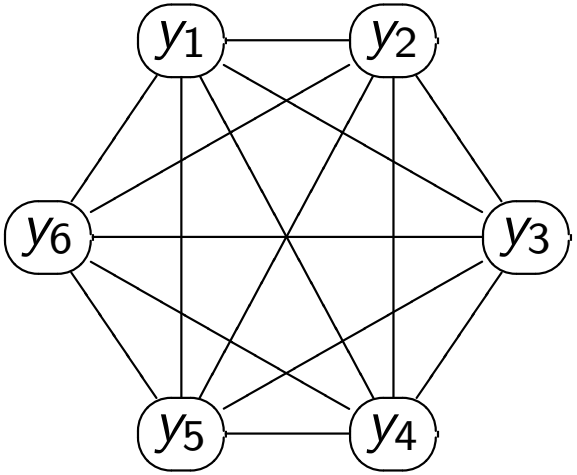
- Causal discovery algorithm developed by Spirtes et al 1993.
- It is composed of two parts:
  1. Infers the skeleton of the causal model.
  2. Partially orient the graph (orient some but not all edges).
- We are only interested in the first part (the "PC skeleton algorithm"). We do **not** use the PC algorithm to edge orientation (we use the QDG algorithm instead).

Step 1 (PC skeleton algorithm)

Suppose the true network describing the causal relationships between six transcripts is



The PC-algorithm starts with the complete undirected graph



and progressively eliminates edges based on conditional independence tests.

Step 1 (PC skeleton algorithm)

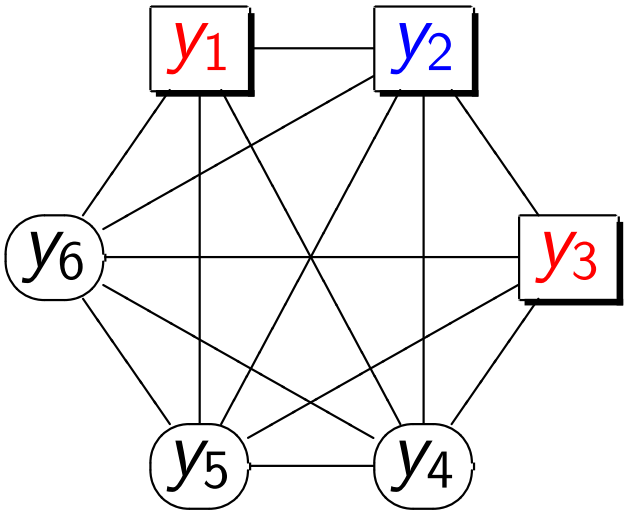
The algorithm performs several rounds of conditional independence tests of increasing order.

It starts with all zero order tests, then performs all first order, second order ...

- Notation: $\perp\!\!\!\perp \equiv$ independence. We read $i \perp\!\!\!\perp j \mid k$ as *i is conditionally independent from j given k*.
- Remark: in the Gaussian case zero partial correlation implies conditional independence, thus

$$i \perp\!\!\!\perp j \mid k \quad \Leftrightarrow \quad cor(i, j \mid k) = 0 \quad \Rightarrow \quad \text{drop } (i, j) \text{ edge}$$

Example (order 1)

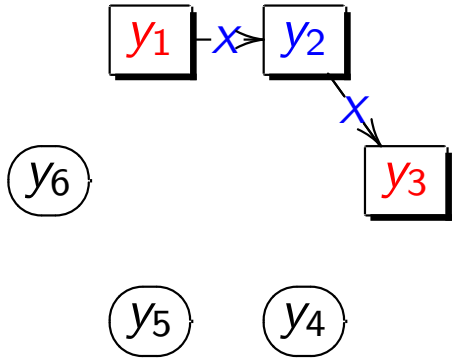


$A(1) \setminus 2 = \{2, 4, 5, 6\}$

$1 \perp\!\!\!\perp 3 \mid 2$

vs

$1 \not\perp\!\!\!\perp 3 \mid 2$



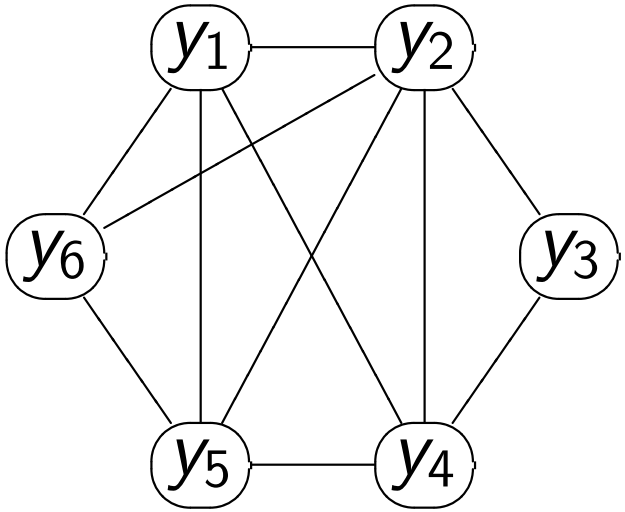
y_2 d-separates y_1 from y_3

$1 \perp\!\!\!\perp 3 \mid 2$

drop edge

move to next edge

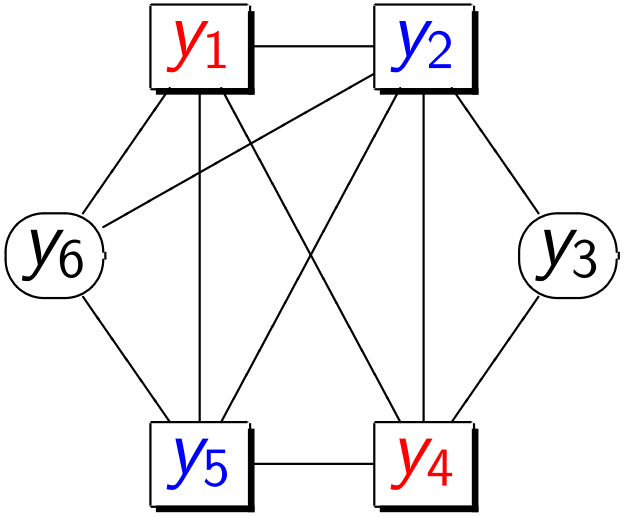
Example (order 1)



After all first order conditional independence tests.

The algorithm then moves to second order conditional independence tests.

Example (order 2)

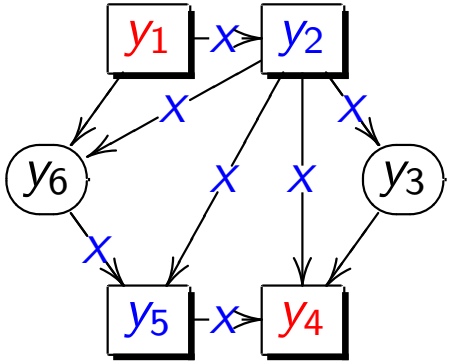


$A(1) \setminus 4 = \{2, 5, 6\}$

$1 \perp\!\!\!\perp 4 \mid 2, 5$

vs

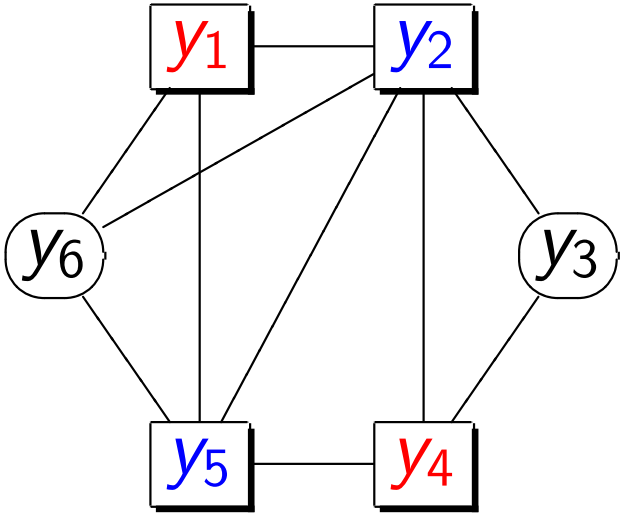
$1 \not\perp\!\!\!\perp 4 \mid 2, 5$



(y_2, y_5) d-separate y_1 from y_4

$1 \perp\!\!\!\perp 4 \mid 2, 5$

Example (order 2)

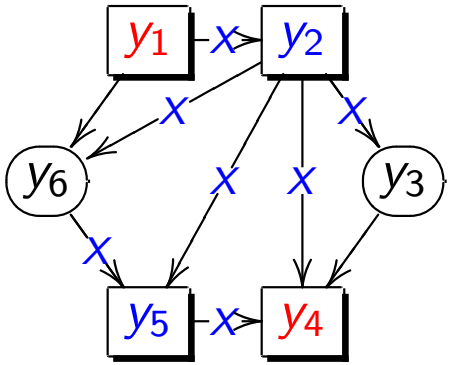


$A(1) \setminus 4 = \{2, 5, 6\}$

$1 \perp\!\!\!\perp 4 \mid 2, 5$

vs

$1 \not\perp\!\!\!\perp 4 \mid 2, 5$



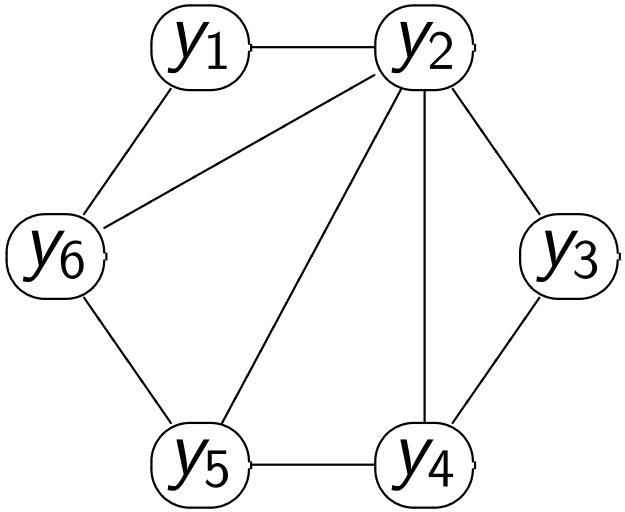
(y_2, y_5) d-separate y_1 from y_4

$1 \perp\!\!\!\perp 4 \mid 2, 5$

drop edge

move to next edge

Example (order 2)



After all second order conditional independence tests.

The algorithm than moves to third order, fourth order ...

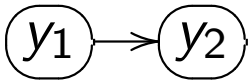
It stops when for each pair (i, j) the cardinality of

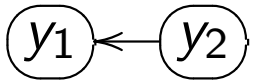
$$A(i) \setminus j$$

is smaller then the order of the algorithm.

Edge orientation

Consider two traits y_1 and y_2. Our problem is to decide between models:

M_1 : 

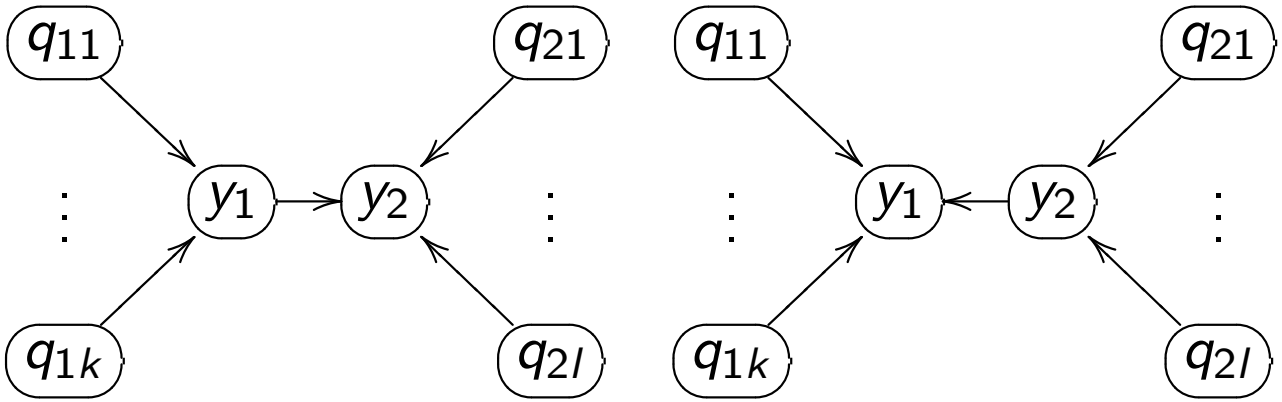
$M2$: 

Problem: the above models are likelihood equivalent,

$$f(y_1)f(y_2 \mid y_1) = f(y_1, y_2) = f(y_2)f(y_1 \mid y_2) \ .$$

Edge orientation

However, models



are *not* likelihood equivalent because

$$f(\mathbf{q}_1)f(y_1 \mid \mathbf{q}_1)f(y_2 \mid y_1, \mathbf{q}_2)f(\mathbf{q}_2)$$
$$\neq$$
$$f(\mathbf{q}_2)f(y_2 \mid \mathbf{q}_2)f(y_1 \mid y_2, \mathbf{q}_1)f(\mathbf{q}_1)$$

Edge orientation

We perform model selection using a direction LOD score

$$LOD = \log_{10} \left\{ \frac{\prod_{i=1}^{n} f(y_{1i} \mid \mathbf{q}_{1i}) f(y_{2i} \mid y_{1i}, \mathbf{q}_{2i})}{\prod_{i=1}^{n} f(y_{2i} \mid \mathbf{q}_{2i}) f(y_{1i} \mid y_{2i}, \mathbf{q}_{1i})} \right\}$$

where $f()$ represents the predictive density, that is, the sampling model with parameters replaced by the corresponding maximum likelihood estimates.

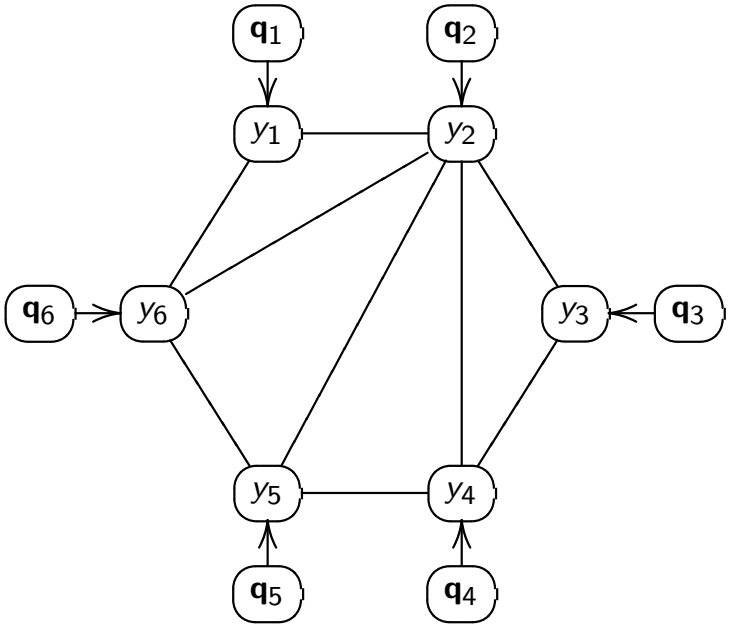
QDG algorithm

QDG stands for QTL-directed dependency graph.
The QDG algorithm is composed of 7 steps:

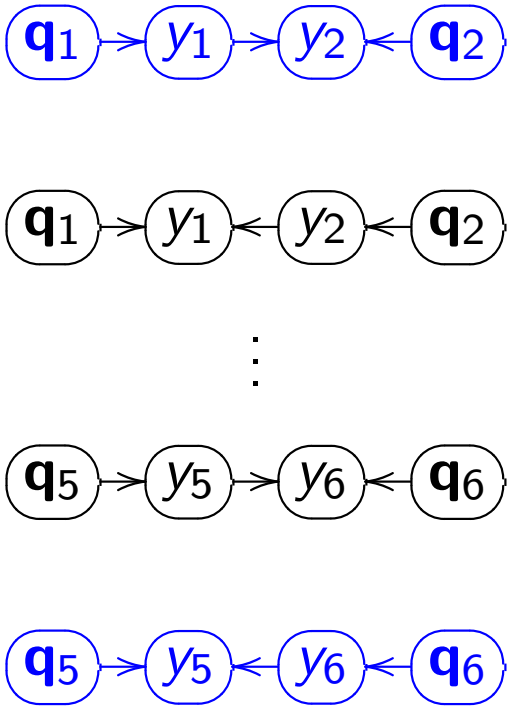
1. Get the causal skeleton (with the PC skeleton algorithm).
2. Use QTLs to orient the edges in the skeleton.
3. Choose a random ordering of edges, and
4. Recompute orientations incorporating causal phenotypes in the models (update the causal model according to changes in directions).
5. Repeat 4 iteratively until no more edges change direction (the resulting graph is one solution).
6. Repeat steps 3, 4, and 5 many times and store all different solutions.
7. Score all solutions and select the graph with best score.

Step 2

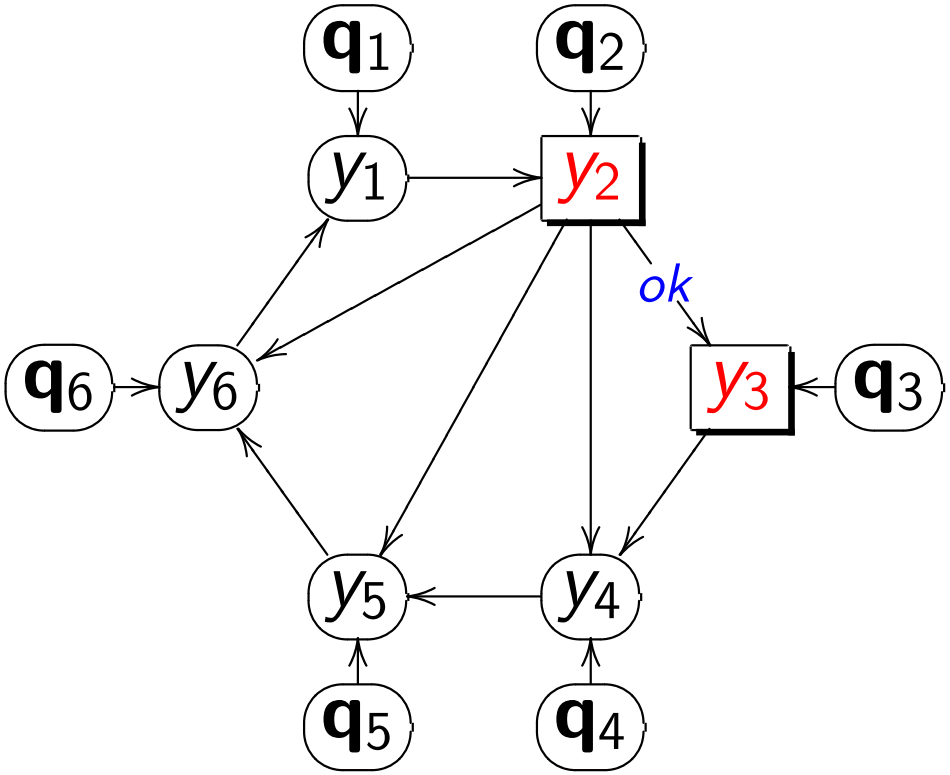
Now suppose that for each transcript we have a set of e-QTLs

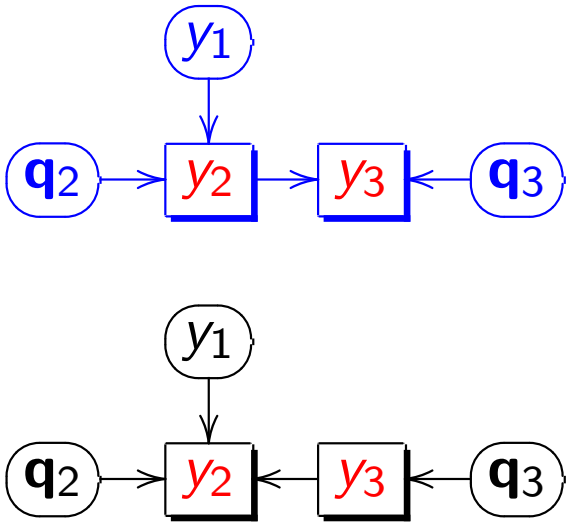


Given the QTLs we can distinguish causal direction:

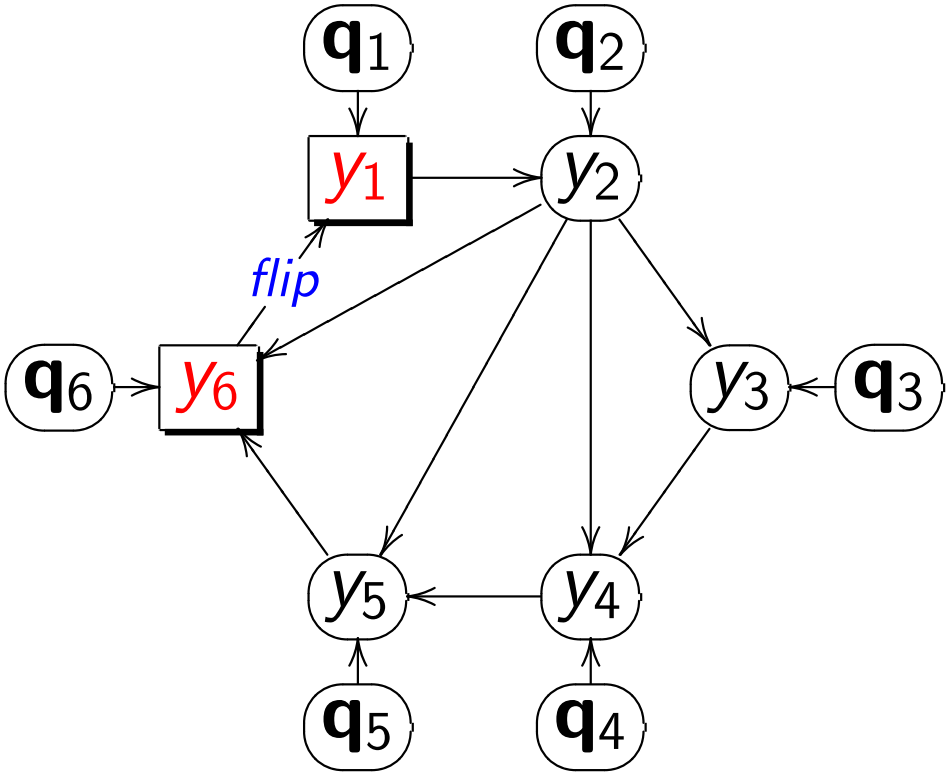


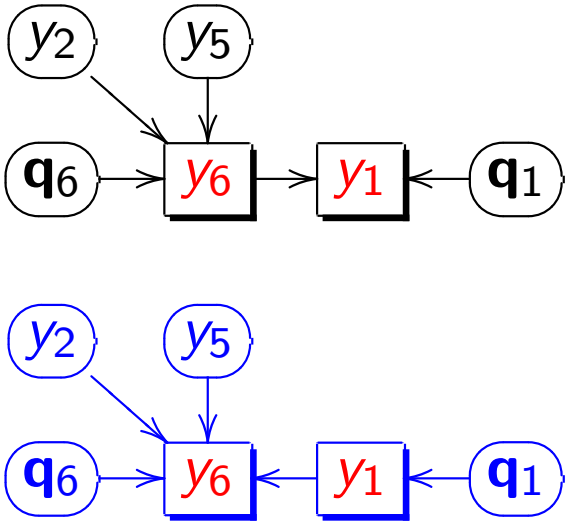
Steps 4 and 5 (first iteration)





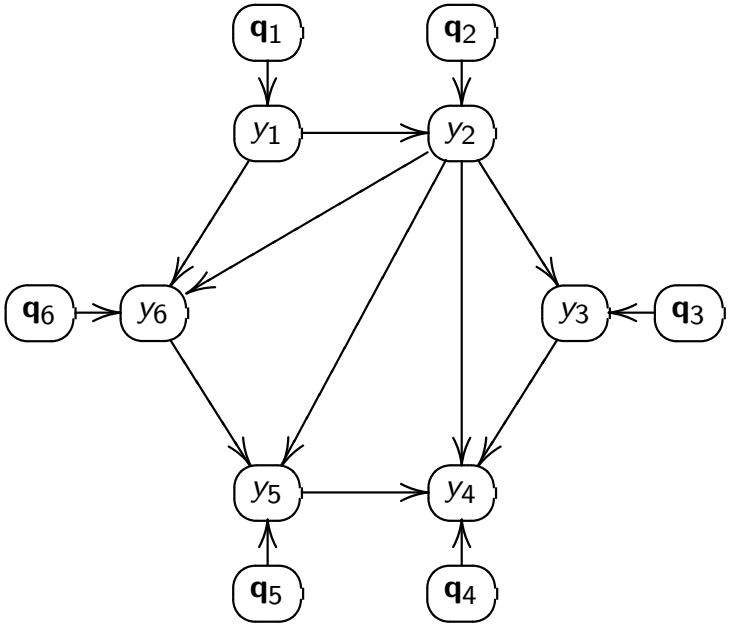
Steps 4 and 5 (first iteration)





Steps 4 and 5 (first iteration)

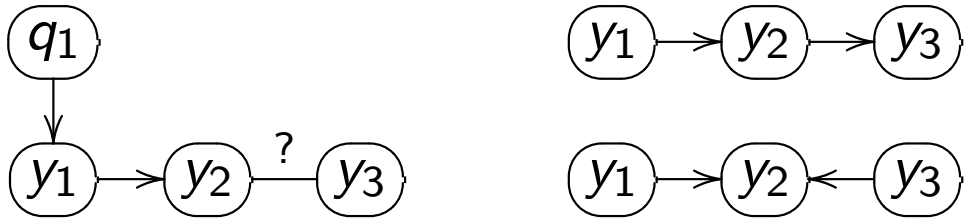
Suppose the updated causal model after the first iteration (DG_1) is



Since some arrows changed direction ($DG_1 \neq DG_0$), the algorithm goes for another round of re-computations.

Directing edges without QTLs

- In general we need to have at least one QTL per pair of phenotypes to infer causal direction.
- In some situations, however, we may be able to infer causal direction for a pair of phenotypes without QTLs. Eg.

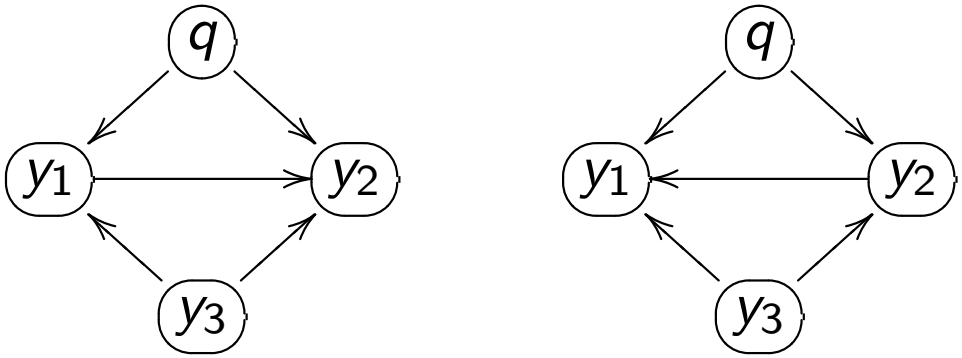


since $f(y_1)\, f(y_2 \mid y_1)\, f(y_3 \mid y_2) \neq f(y_1)\, f(y_2 \mid y_1, y_3)\, f(y_3)$.

- So both QTLs and phenotypes play important roles in the orientation process.

Unresolvable situation

- We cannot infer direction when the phenotypes have exactly same set of QTLs and causal phenotypes

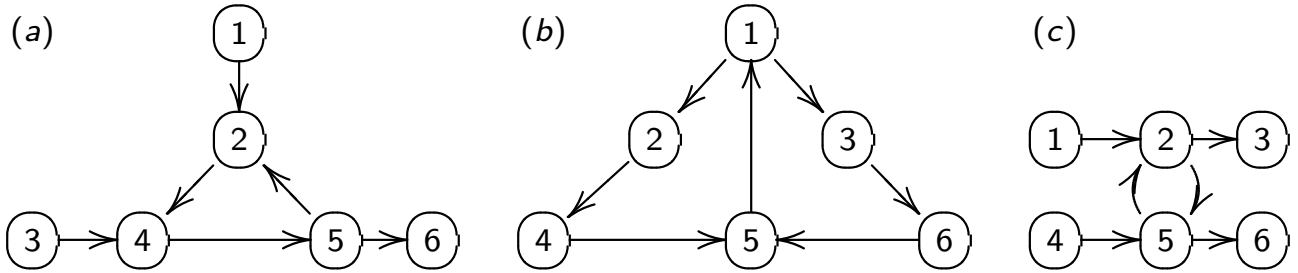


since

$$f(y_1 \mid y_3, q)\, f(y_2 \mid y_1, y_3, q) \;=\; f(y_1 \mid y_2, y_3, q)\, f(y_2 \mid y_3, q)$$

Cyclic networks

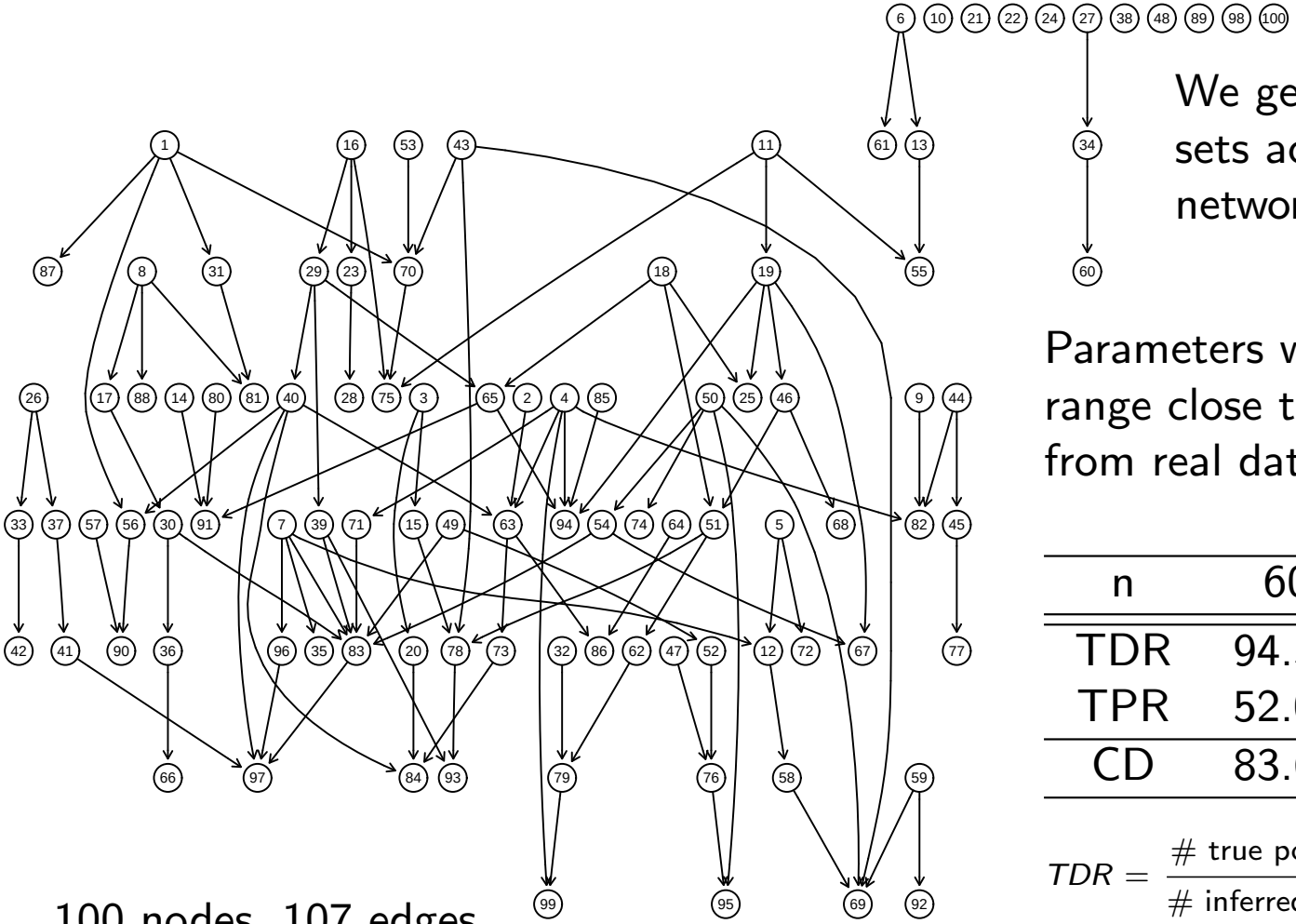
- Our simulations showed good performance with toy cyclic graphs, though.



- The spurious edges in graph (c) were detected at low rates.

- QDG approach cannot detect reciprocal interactions. In graph (c) it orients the edge (2)—(5) in the direction with higher strength.

Simulations



100 nodes, 107 edges

2 or 3 QTLs per phenotype (not shown)

We generated 100 data sets according to this network.

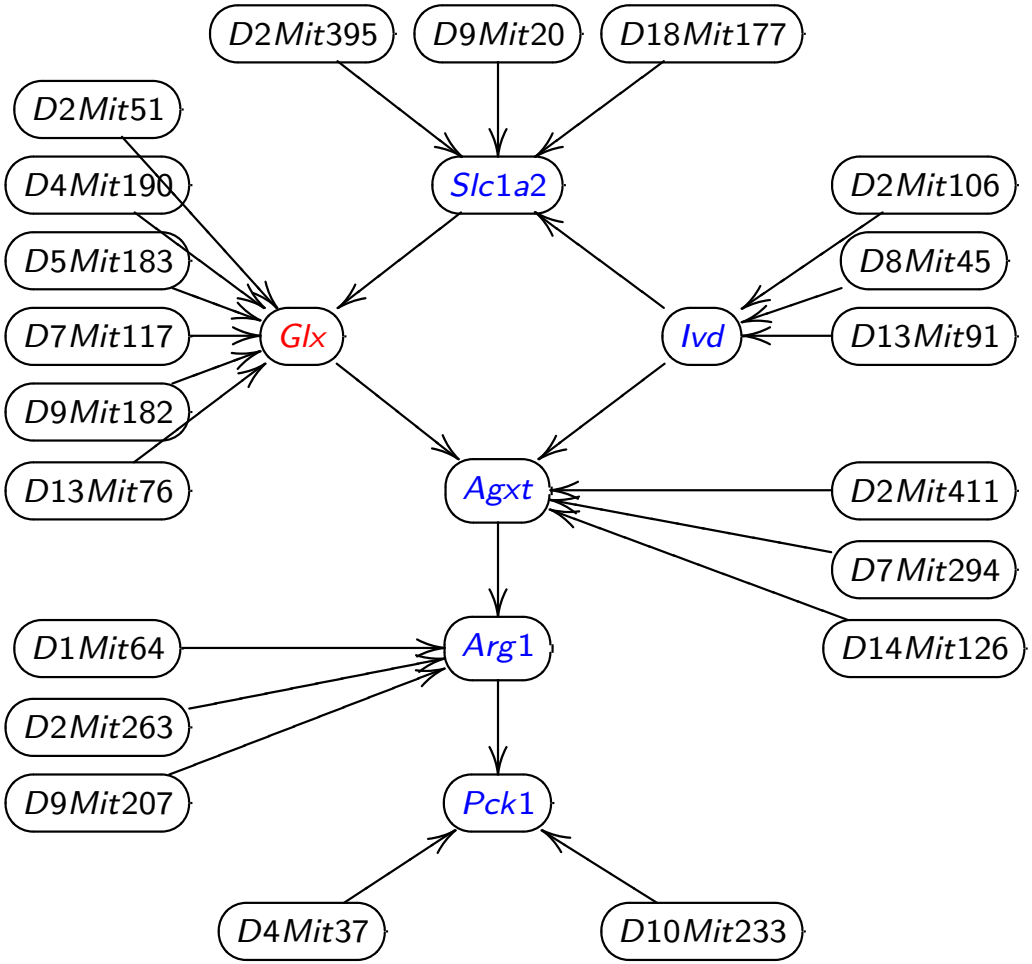
Parameters were chosen in a range close to values estimated from real data.

n	60	300	500
TDR	94.53	95.18	91.22
TPR	52.07	87.33	93.64
CD	83.65	98.58	99.63

$$TDR = \frac{\#\text{ true positives}}{\#\text{ inferred edges}}, \quad TPR = \frac{\#\text{ true positives}}{\#\text{ true edges}}$$

CD: correct direction

Real data example



- We constructed a network from metabolites and transcripts involved in liver metabolism.
- We validated this network with *in vitro experiments* (Ferrara et al 2008). Four out of six predictions were confirmed.

Software and references

The *qdg* R package is available at CRAN.

References:

- Chaibub Neto et al 2008. Inferring causal phenotype networks from segregating populations. Genetics 179: 1089-1100.
- Ferrara et al 2008. Genetic networks of liver metabolism revealed by integration of metabolic and transcriptomic profiling. PLoS Genetics 4: e1000034.
- Spirtes et al 1993. Causation, prediction and search. MIT press.

Acknowledgements

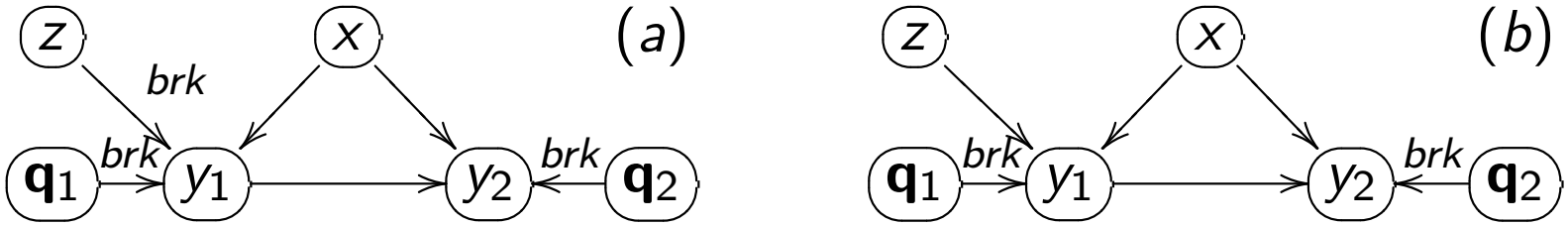
Co-authors:

- Alan D. Attie
- Brian S. Yandell
- Christine T. Ferrara

Funding:

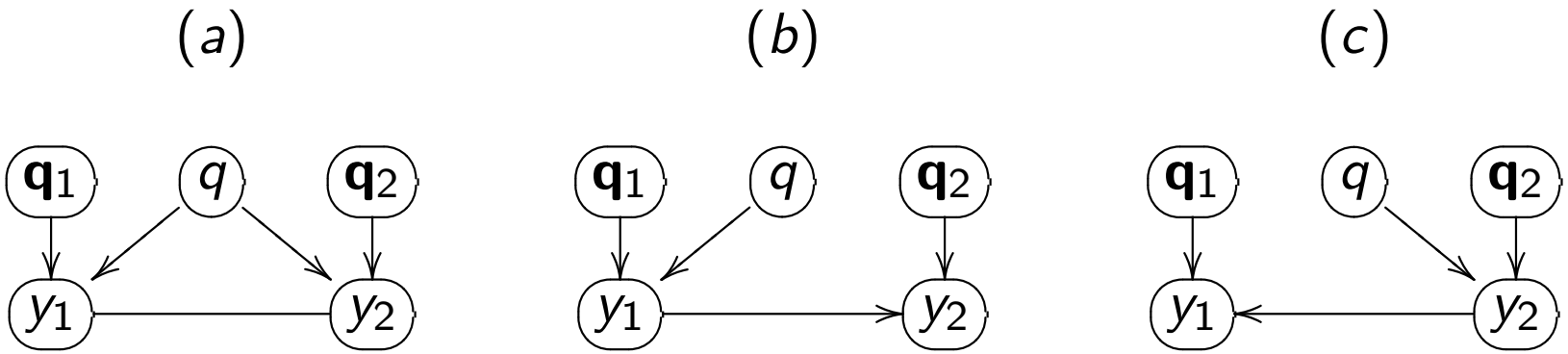
- CNPq Brazil
- NIH grants DK66369, DK58037 and DK06639

Permutation p-values



- To break the connections (brk) that affect direction of an edge, we permute the corresponding pair of nodes (and their common covariates) as a block.
- In panel (a) we permute (y_1, y_2, x) as a block breaking the connections with z, $\mathbf{q}_1$ and $\mathbf{q}_2$;
- In panel (b) we incorrectly keep z in the permutation block.

Direct versus indirect effects of a common QTL



- A strong QTL directly affecting an upstream trait may also be (incorrectly) detected as a QTL for a downstream phenotype.
- To resolve this situation we apply a generalization of Schadt et al. 2005 allowing for multiple QTLs.
- Model (a) supports both traits being directly affected by the common QTL q. Model (b) implies that q directly affects y_1 but should not be included as a QTL of phenotype y_2. Model (c) supports the reverse situation.

causal graphical models in systems genetics

- Chaibub Neto, Keller, Attie , Yandell (2009) Causal Graphical Models in Systems Genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. *Ann Appl Statist* (tent. accept)
- Related references
  - Schadt et al. Lusis (2005 *Nat Genet*); Li et al. Churchill (2006 *Genetics*); Chen Emmert-Streib Storey(2007 *Genome Bio*); Liu de la Fuente Hoeschele (2008 *Genetics*); Winrow et al. Turek (2009 *PLoS ONE*)
- Jointly infer unknowns of interest
  - genetic architecture
  - causal network

Basic idea of QTLnet

- Genetic architecture given causal network
  - Trait y depends on parents pa(y) in network
  - QTL for y found conditional on pa(y)
    - Parents pa(y) are interacting covariates for QTL scan
- Causal network given genetic architecture
  - Build (adjust) causal network given QTL

MCMC for QTLnet

- Propose new causal network with simple changes to current network
  - Change edge direction
  - Add or drop edge
- Find any new genetic architectures (QTLs)
  - Update phenotypes whose parents pa(y) change in new network
- Compute likelihood for new network and QTL
- Accept or reject new network and QTL
  - Usual Metropolis-Hastings idea

Future work

- Incorporate latent variables
  - Aten et al. Horvath (2008 *BMC Sys Biol*)
- Allow for prior information about network
  - Werhli and Husmeier (2007 *SAGMB*); Dittrich et al. Müller (2008 *Bioinfo*); Zhu et al. Schadt (2008 *Nat Genet*); Lee et al. Koller (2009 *PLoS Genet*); Thomas et al. Portier (2009 *Genome Bio*); Wu et al. Lin (2009 *Bioinfo*)
- Improve algorithm efficiency
  - Ramp up to 1000s of phenotypes
- Extend to outbred crosses, humans