Causal Protein Signalling Networks

Hunting for the Lost Pathways

Marco Scutari

m.scutari@ucl.ac.uk
Genetics Institute
University College London

January 25, 2013

Graphical Models

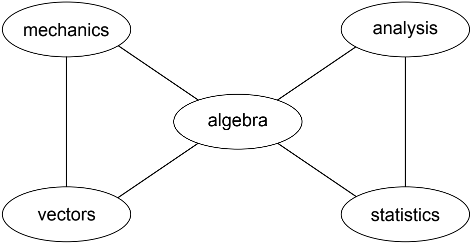
Graphical Models

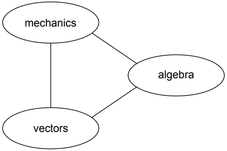
Graphical models are defined by:

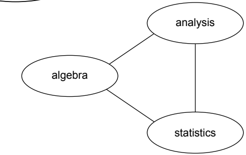
- a network structure, $\mathcal{G} = (\mathbf{V}, E)$, either an undirected graph (Markov networks, gene association networks, correlation networks, etc.) or a directed graph (Bayesian networks). Each node $v_i \in \mathbf{V}$ corresponds to a random variable X_i;
- a global probability distribution, $\mathbf{X}$, which can be factorised into a set of small local probability distributions according to the edges $e_{ij} \in E$ present in the graph.

This combination allows a compact representation of the joint distribution of large numbers of random variables and simplifies inference on the resulting parameter space.

A Gaussian Markov Network (MARKS)



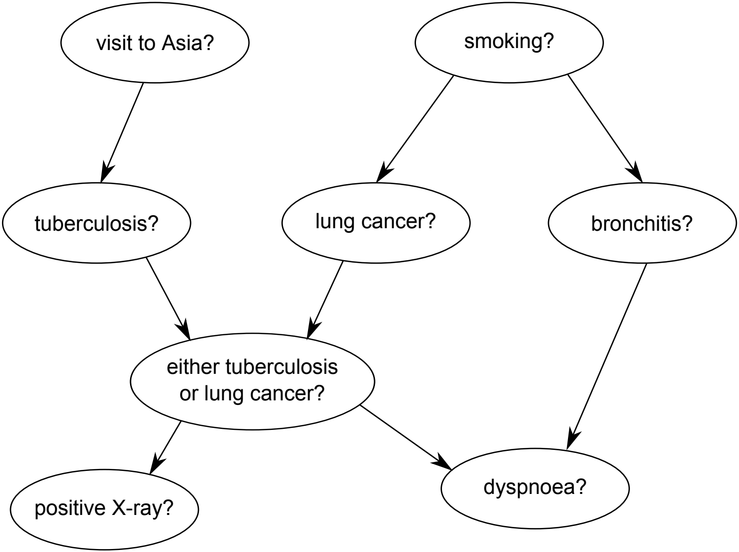




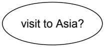
$$\begin{bmatrix} 1.000 & 0.332 & 0.235 \\ 0.332 & 1.000 & 0.327 \\ 0.235 & 0.327 & 1.000 \end{bmatrix}$$

$$\begin{bmatrix} 1.000 & 0.451 & 0.364 \\ 0.451 & 1.000 & 0.256 \\ 0.364 & 0.256 & 1.000 \end{bmatrix}$$

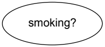
A Discrete Bayesian Network (ASIA)



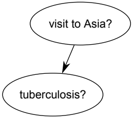
A Discrete Bayesian Network (ASIA)

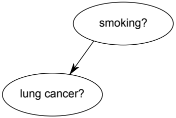


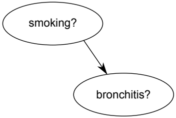
yes	no
0.01	0.99



yes	no
0.50	0.50



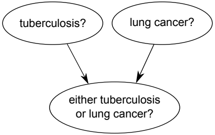


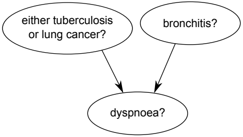


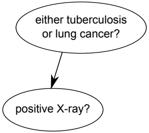
	yes	no
yes	0.05	0.95
no	0.01	0.99

	yes	no
yes	0.10	0.90
no	0.01	0.99

	yes	no
yes	0.60	0.40
no	0.30	0.70







	yes	no
yes:yes	1	0
yes:no	1	0
no:yes	1	0
no:no	0	1

	yes	no
yes:yes	0.90	0.10
yes:no	0.70	0.30
no:yes	0.80	0.20
no:no	0.10	0.90

	yes	no
yes	0.98	0.02
no	0.05	0.95

Causal Protein Signalling Networks

Source

In the following, we try to reproduce (to the best of my ability, and Karen Sachs' recollections about the implementation details that did not end up in the Methods section) the statistical analysis in the following paper:

Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data
Karen Sachs, *et al.*
Science **308**, 523 (2005);
DOI: 10.1126/science.1105809

That's a landmark paper in applying Bayesian Networks because:

- it successfully establishes causality claims;
- it highlights the use of observational vs interventional data;
- results are validated using existing literature.

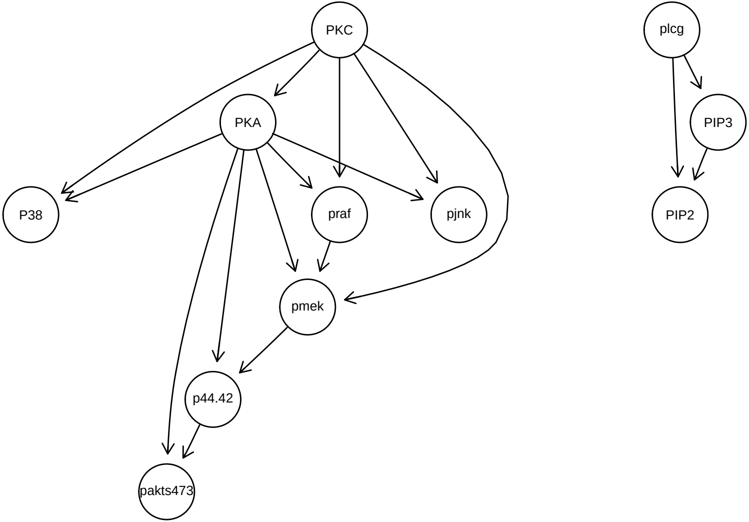
An Overview of the Data

The data consist in the simultaneous measurements of 11 phosphorylated proteins and phospholipids derived from thousands of individual primary immune system cells:

- 1800 (observational) data subject only to general stimulatory cues, so that the protein signalling paths are active;
- 600 (interventional) data with with specific stimulatory & inhibitory cues for each of the following 4 proteins: pmek, PIP2, pakts473, PKA;
- 1200 (interventional) data with specific cues for PKA.

Overall, the data set contains 5400 observations with no missing value. Not all the proteins involved in the modelling pathways are observed.

Network Reconstructed from Literature



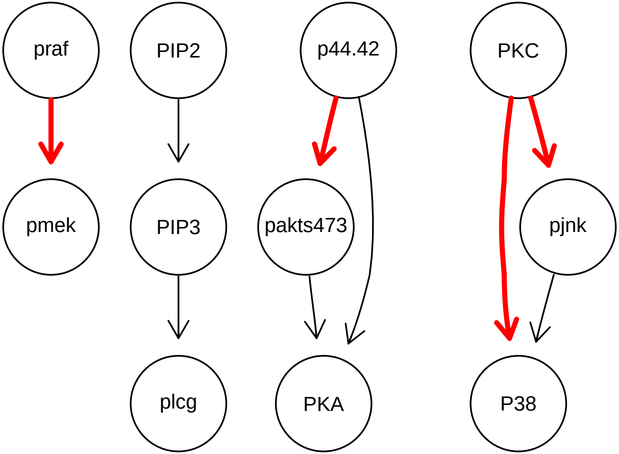
Using Only Observational Data

As a first, exploratory analysis, we can try to learn a network from the data that were subject only to general stimulatory cues. Since these cues only ensure the pathways are active, but do not tamper with them in any way, such data are observational (as opposed to interventional).

```
> library(bnlearn)
> hc(sachs, score = "bge", iss = 5)
```

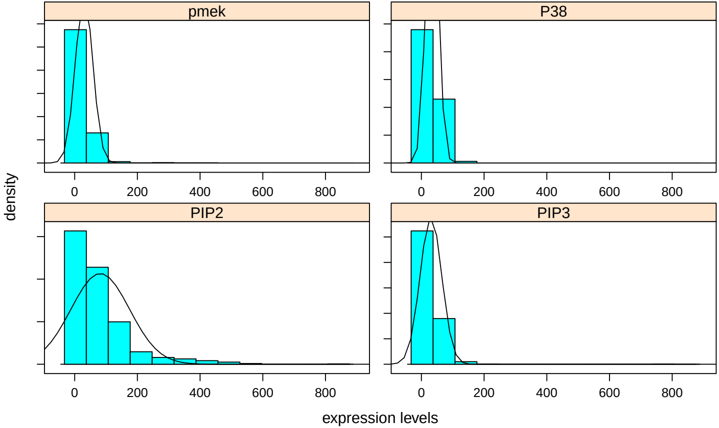
Here we try to learn the network that maximises posterior probability (the "bge" score, when you assume normality) giving very little weight to the uninformative prior (iss = 5) compared to the sample.

Network Reconstructed from the Observational Data



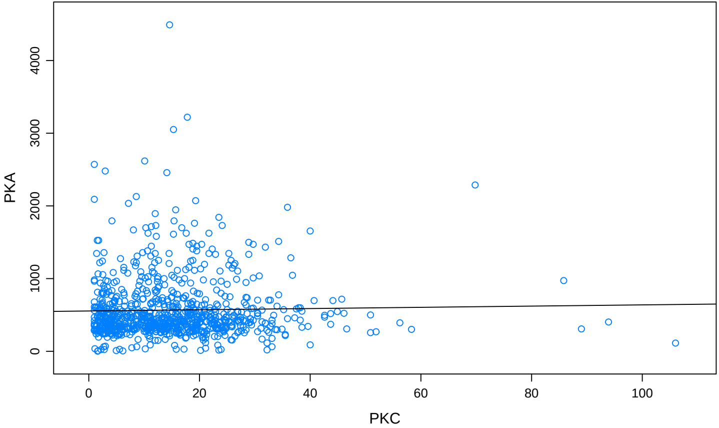
Arcs highlighted in red are also present in the network reconstructed from literature.

Expression Data are not Symmetric



Therefore, assuming a Gaussian distribution is problematic.

Expression Data are not Linked by Linear Relationships



Therefore, tests for correlation are biased and have extremely low power.

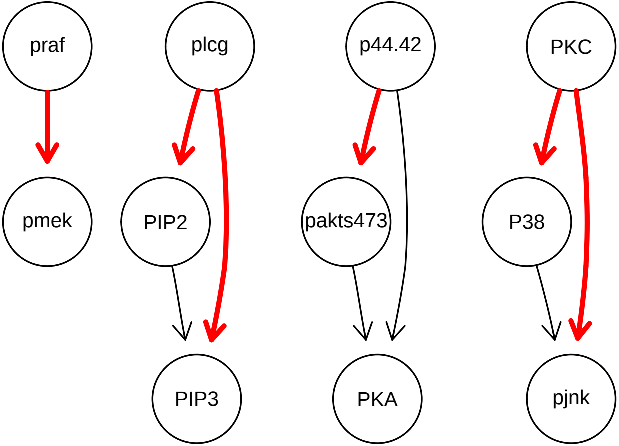
Discretise!

Since we cannot use Gaussian Bayesian networks, we can discretise the data instead. Hartemink's method is designed to preserve pairwise dependencies as much as possible, unlike marginal discretisation methods.

```
> dsachs = discretize(sachs, method = "hartemink",
+             breaks = 3, ibreaks = 60,
+             idisc = "quantile")
```

Variables are first marginally discretised in 60 intervals, which are subsequently collapsed while reducing the mutual information between the variables as little as possible. The process stops when each variable has 3 levels (i.e. low, average and high expression).

Network Reconstructed from the Discretised Data



Two more arcs are correctly identified, but most are still missing.

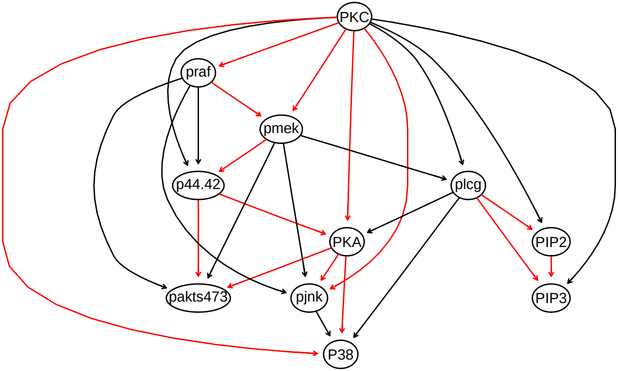
Considering Interventional Data

It is apparent from the previous networks that most signalling paths are not statistically identifiable unless we inhibit or stimulate the expression of at least some of the proteins in the network. Therefore, we include the interventional data in the analysis.

```
> INT = sapply(1:11, function(x)
+         { which(isachs$INT == x) })
> names(INT) = names(isachs)[1:11]
> hc(isachs[, 1:11], score = "mbde",
+     exp = INT, iss = 5)
```

Since the standard posterior probability ("bde") does not take interventions into account, we use a modified BDe score ("mbde") that disregards any causal influence for the proteins that have been inhibited or stimulated.

Network Reconstructed from the Interventional Data



More arcs are included, but there are many false positives.

Removing Noisy Arcs with Model Averaging

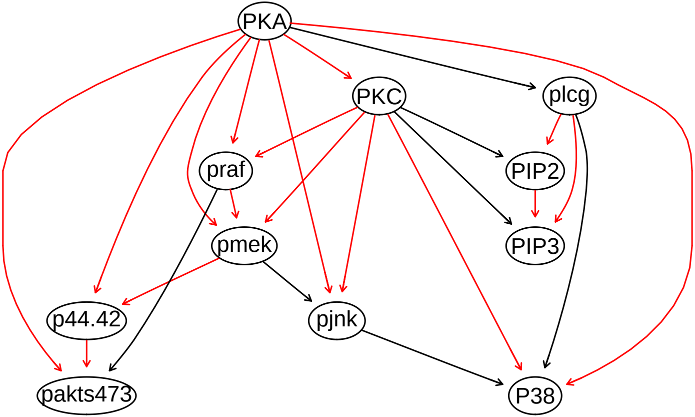
Two simple steps can be taken to remove noisy arcs:

- average multiple networks obtained using different starting points when learning the structure of the graph;
- use TABU search (tabu) instead of Hill-Climbing (hc).

```
> start = random.graph(nodes = nodes,
+   method = "melancon", num = 500, burn.in = 10^5,
+   every = 100)
> netlist = lapply(start, function(net) {
+   tabu(isachs[, 1:11], score = "mbde", exp = INT,
+     iss = 10, start = net, tabu = 50) })
> arcs = custom.strength(netlist, nodes = nodes)
```

A similar approach was chosen as the best performing in Sachs *et al.* [5], with minor differences in results.

Interventional Data with Model Averaging



All the arcs supported by literature are present in the network.

Thanks!

References

References I

D. Koller and N. Friedman.
Probabilistic Graphical Models: Principles and Techniques.
MIT Press, 2009.

K. Korb and A. Nicholson.
Bayesian Artificial Intelligence.
Chapman and Hall, 2nd edition, 2009.

G. Melançon, I. Dutour, and M. Bousquet-Mélou.
Random Generation of DAGs for Graph Drawing.
Technical Report INS-R0005, Centre for Mathematics and Computer Sciences, Amsterdam, 2000.

J. Pearl.
Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.
Morgan Kaufmann, 1988.

K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan.
Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data.
Science, 308(5721):523–529, 2005.

References II

- M. Scutari.
  Learning Bayesian Networks with the bnlearn R Package.
  Journal of Statistical Software, 35(3):1–22, 2010.

- M. Scutari and K. Strimmer.
  Introduction to Graphical Modelling.
  In D. J. Balding, M. Stumpf, and M. Girolami, editors, *Handbook of Statistical Systems Biology*. Wiley, 2011.
  In print.

- P. Spirtes, C. Glymour, and R. Scheines.
  Causation, Prediction, and Search.
  MIT Press, 2000.