Bayesian Networks for Gene Networks Discovery: Parallel and Optimised Learning

Marco Scutari, Genetics Institute, University College London (UCL)

Bayesian Networks as Gene Networks

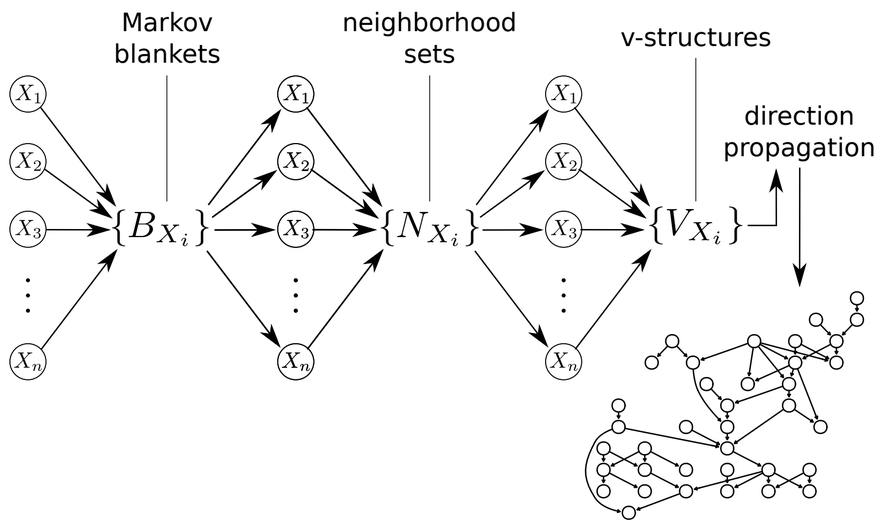
In genetics and systems biology, Bayesian networks (BNs) are used to describe and identify interdependencies among genes and gene products, with the eventual aim to better understand the molecular mechanisms that link them. If we assign each gene to one node in the BN, edges represent the interplay between different genes, and can describe either direct (causal) interactions or indirect influences that are mediated by unobserved genes. BNs can be estimated (learned) with a variety of algorithms, which can all be traced to three approaches:

1. constraint-based, which are based on conditional independence tests;
2. score-based, which are based on goodness-of-fit scores;
3. and hybrid, which combine the previous two approaches.

Score-based algorithms are just the application of general purpose optimisation techniques to BNs, and most are inherently sequential (e.g. each step depends on the previous one). On the other hand, constraint-based algorithms can be parallelised effectively, to the point that it is feasible to learn gene networks from high-dimensional data.

Parallel Constraint-Based Learning

Constraint-based algorithms display a coarse-grained parallelism, because they can be split in parts whose status needs to be updated only two times. Recent algorithms, which learn the Markov blankets of the nodes as an intermediate step, require one additional update.



Therefore, they can all be parallelised as shown above:

1. the Markov blanket of each node can be learned independently from the others;
2. each neighbourhood is a subset of the corresponding Markov blanket and, therefore, can be learned independently from the others. The consistency of the Markov blankets must be checked beforehand. They may not be symmetric for very noisy data, so we need to examine all pairs of nodes and remove them from each other's Markov blanket if they do not appear in both of them;
3. given the neighbourhoods, the v-structures centred on each node (i.e. the one with the converging arcs) can be identified in parallel; using Markov blankets is not required, but reduces the search space considerably. Again, the consistency of the neighbourhoods must be checked beforehand.

Note that the resulting BN is identical to the one obtained from the non-parallel implementation, and the tests they perform are exactly the same.

Optimised Constraint-Based Learning

Although optimisations for constraint-based algorithms have not been explored in detail in literature, some papers (e.g. [6]) suggest using backtracking to reduce the number of conditional independence tests and the size of the conditioning sets. Since Markov blankets and neighbourhoods are symmetric, we can consider those we already learned to initialise the one we are currently learning. For example, for the Markov blanket of a node X_i:

1. we can tentatively include all the nodes whose Markov blankets include X_i, as $X_i \in B_{X_j} \Leftrightarrow X_j \in B_{X_i}$;
2. we can tentatively exclude all the nodes whose Markov blankets do not include X_i.

Nodes that are tentatively included can later be removed by a test (i.e. they may be false positives), and nodes that are excluded can later be included (i.e. they may be false negatives).

Benchmark Data Sets and Algorithms

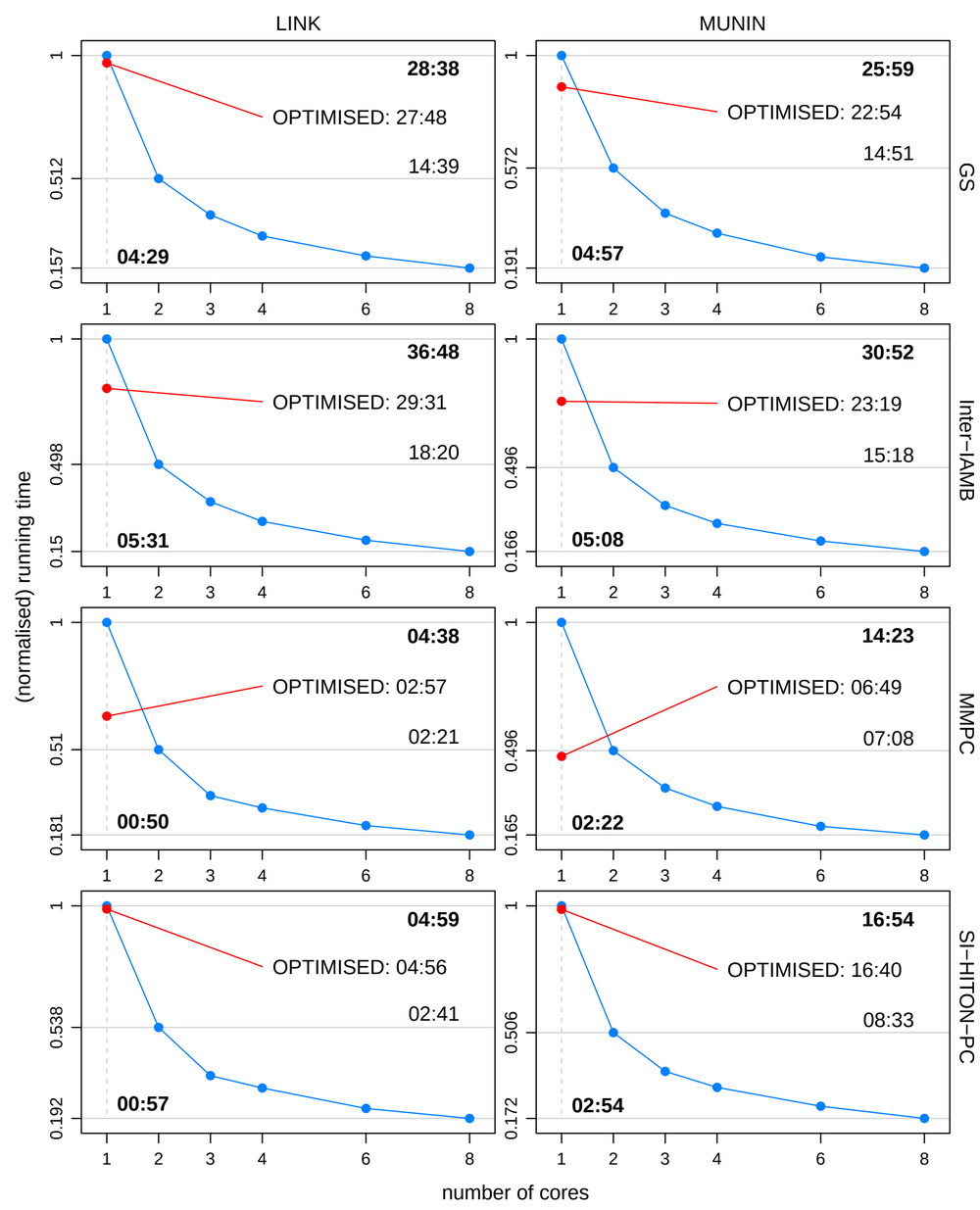
Using **bnlearn** [5], we assessed the parallel and optimised implementations of:

- the Grow-Shrink (GS) and the Interleaved IAMB (Inter-IAMB) learning algorithms [4], which learn complete BNs starting from their Markov blankets;
- the Max-Min Parents-Children (MMPC) [4] and the Semi-Interleaved HITON-PC (SI-HITON-PC) [6] algorithms, which learn the undirected graph underlying the BN;

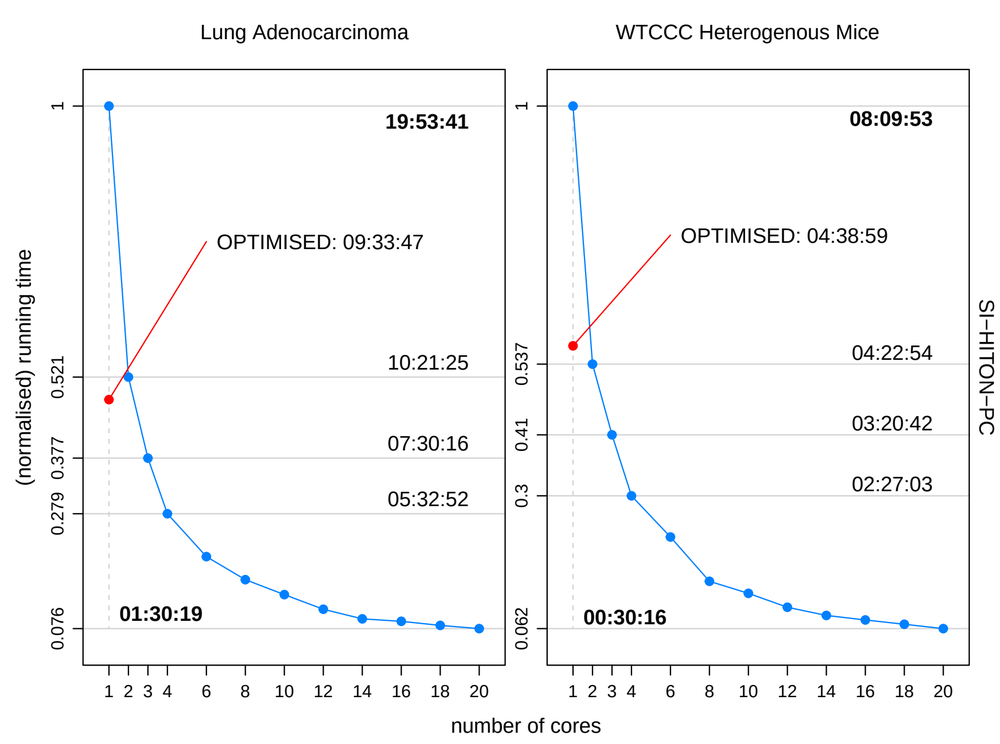
on samples of 20K observations generated from the MUNIN ([1], 1041 nodes, 81K parameters) and the LINK ([3], 724 nodes, 14K parameters) reference BNs. The only algorithm that scales well to genetic data is SI-HITON-PC, which we applied to:

- the lung adenocarcinoma gene expression data (86 obs., 7131 nodes) from [2];
- the WTCCC heterogeneous mice SNP data (1940 obs., 12545 nodes) from [7].

Benchmarks on Reference Bayesian Networks



Benchmarks on Real World Genetic Data



Conclusions

- Parallel implementations of constraint-based BN learning algorithms scale linearly in the number of cores/processors, with little overhead.
- Considering that any modern computer (even desktops) has at least two cores, optimised implementations of constraint-based algorithms are not competitive with the corresponding parallel implementations, even on a single machine.

References

[1] S. Andreassen, F. V. Jensen, S. K. Andersen, B. Falck, U. Kjærulff, M. Woldbye, A. R. Sørensen, A. Rosenfalck, and F. Jensen. MUNIN - an Expert EMG Assistant. In *Computer-Aided Electromyography and Expert Systems*. Elsevier, 1989.

[2] D. G. Beer, S. L. R. Kardia, C.-C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, L. Lin, G. Chen, T. G. Gharib, D. G. Thomas, M. L. Lizyness, R. Kuick, S. Hayasaka, J. M. G. Taylor, M. D. Iannettoni, M. B. Orringer, and S. Hanash. Gene-expression Profiles Predict Survival of Patients with Lung Adenocarcinoma. *Nature Medicine*, 8:816–824, 2002.

[3] C. S. Jensen and A. Kong. Blocking Gibbs Sampling for Linkage Analysis in Large Pedigrees with Many Loops. *The American Journal of Human Genetics*, 65(3):885–901, 1999.

[4] R. Nagarajan, M. Scutari, and S. Lèbre. *Bayesian Networks in R with Applications in Systems Biology*. Springer, 2013.

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

[6] A. Statnikov, N. I. Lytkin, J. Lemeire, and C. F. Aliferis. Algorithms for Discovery of Multiple Markov Boundaries. *Journal of Machine Learning Research*, 14:499–566, 2013.

[7] W. Valdar, L. C. Solberg, D. Gauguier, S. Burnett, P. Klenerman, W. O. Cookson, M. S. Taylor, J. N. Rawlins, R. Mott, and J. Flint. Genome-Wide Genetic Association of Complex Traits in Heterogeneous Stock Mice. *Nature Genetics*, 8:879–887, 2006.

Mathematical and Statistical Aspects of Molecular Biology (MASAMB), London, Imperial College, London, 11-12 April 2013

Email: m.scutari@ucl.ac.uk