CSI: A Nonparametric Bayesian Approach to Network Inference from Multiple Perturbed Time Series Gene Expression Data

Christopher A. Penfold[1], Ahmed Shifaz[2], Paul Brown[1], Ann Nicholson[2] and David L. Wild[*,1]

[1]Warwick Systems Biology Centre, University of Warwick, Coventry, UK, CV4 7AL
[2]Faculty of Information Technology, Monash University, VIC, 3800, Australia
[*]d.l.wild@warwick.ac.uk

How an organism responds to the environmental challenges it faces is heavily influenced by its gene regulatory networks (GRNs). Whilst most methods for inferring GRNs from time series mRNA expression data are only able to cope with single time series (or single perturbations with biological replicates), it is becoming increasingly common for several time series to be generated under different experimental conditions. The CSI algorithm (Klemm, 2008) represents one approach to inferring GRNs from multiple time series data, which has previously been shown to perform well on a variety of datasets (Penfold and Wild, 2011). Another challenge in network inference is the identification of condition specific GRNs i.e., identifying how a GRN is rewired under different conditions or different individuals. The Hierarchical Causal Structure Identification (HCSI) algorithm (Penfold et al., 2012) is one approach that allows inference of condition specific networks (Hickman et al., 2013), that has been shown to be more accurate at reconstructing known networks than inference on the individual datasets alone. Here we describe a MATLAB implementation of CSI/HCSI that includes fast approximate solutions to CSI as well as Markov Chain Monte Carlo implementations of both CSI and HCSI, together with a user-friendly GUI, with the intention of making the analysis of networks from multiple perturbed time series datasets more accessible to the wider community.[1] The GUI itself guides the user through each stage of the analysis, from loading in the data, to parameter selection and visualisation of networks, and can be launched by typing `>> csi` into the MATLAB command line. For each step of the analysis, links to documentation and tutorials are available within the GUI, which includes documentation on visualisation and interacting with output files.

Inferring networks with CSI

The CSI algorithm assumes the dynamics of gene expression for gene A evolves as: $\mathbf{X}_A(t+1) = f(\mathbf{X}_{\mathscr{P}a}(t)) + \varepsilon$, where $\mathbf{X}_A(t)$ represent the expression level of gene A at time t, $\mathbf{X}_{\mathscr{P}a}(t)$ represents the expression of the regulators at time t, ε some Gaussian noise, and $f(\cdot)$ represents an unknown (nonlinear) function. Here we assign a prior distribution over this function in the form of a Gaussian process (GP) prior (see Fig. 1(a)), which may be integrated out to yield the likelihood of gene expression for A given a particular parental set of genes, $\mathscr{P}a$. The posterior distribution over parental sets (and hyperparameters) can be constructed via Bayes' rule, and consists of combinatorially searching through all sets of putative regulators (up to a maximum in-degree d), and assigning a likelihood to each set. We may obtain a point estimate of the distribution via an Expectation Maximisation algorithm (Penfold and Wild, 2011), or sample from it via MCMC (Penfold et al., 2012). The Gaussian process model underpinning the dynamics of gene regulation requires a matrix inversion (which scales as $\mathscr{O}(m^3)$ where m is the number of experimental observations) for each possible parental set and each step in the gradient optimisation. Within this implementation we have also included a sparse GP using the Fully Independent Training Conditional approximation (FITC; Snelson and Ghahramani, 2006, Quinonero-Candela et al., 2005). FITC approximates the full GP prior, effectively summarising the m observations through n inducing points (see Fig. 1(a,b)), and scales as $O(mn^2)$ with $n < m$ the number

[1]This implementation is available from `http://go.warwick.ac.uk/systemsbiology/software`

of inducing points. When combined with a more efficient gradient search algorithm, this results in a significant speedup compared to previous CSI implementations (Fig. 1(c)) with no degradation to the accuracy of inferred networks.

Inferring Context Dependent Networks with HCSI

HCSI represents a method for identifying condition specific regulation using multiple datasets. For example, if data is collected under different experimental conditions or in different individuals, the underlying networks may be similar, but there may also be some network rewiring resulting in condition or individual specific links. HCSI infers the upstream regulators for a gene of interest in each dataset based upon the Gaussian process model for gene expression used by CSI. Crucially, the regulators for each individual dataset are constrained to favour similar sets of regulators, allowing some differences in those regulators but in a way that favours similarity. A parameter, β, is responsible for determining how similar the sets of regulators should be in the different datasets. For $\beta = 0$, the parents in each datasets can be considered independent of the parents in another dataset; for $\beta >> 1$ parental sets will be increasingly similar across the datasets. This parameter can be automatically tuned within the algorithm or fixed by the user. Currently an MCMC approach is implemented for HCSI, that updates the parental set in each dataset in turn via a Gibbs update, followed by a Metropolis update of the GP hyperparameters and, if not fixed, a Metropolis update of the β hyperparameter.

Because CSI/HCSI infer parental sets on a gene-by-gene basis, they represents ideal companions to yeast one-hybrid (Y1H) or other transcription factor binding measurements such chromatin landscaping (Kent et al., 2011). Specifically, given a list of genes shown to bind the promotor region of a gene of interest, CSI/HCSI can be used to identify the most likely subsets driving the expression of that gene in a given dataset. Example Y1H and expression data taken from Hickman et al. (2013) are included with this package (see accompanying help files for full details). Furthermore, the gene-by-gene nature of these algorithms can be exploited when inferring networks, by identifying the parental sets for different genes in parallel, using distributed computing capabilities. The ability to connect the GUI to computational clusters to utilise high-performance computing has been included within this package via the MATLAB Parallel Computing Toolbox and Distributed Computing Server.

Discussion and Future Work

The CSI package allows easy inference of the regulators of a gene from multiple perturbed gene expression data. This approach may be of use in a number of cases: (i) CSI may be used for combining many time series data from different conditions, resulting in more accurate GRNs, but does so assuming identical network structure in each of the datasets; (ii) HCSI can be used for for inferring similarities and differences in GRNs from different biological conditions (where network rewiring might occur) or in different individuals where genetic differences give rise to related but non-identical networks. Future work that extends the concepts of HCSI will allow the leveraging of data and networks between different species, including cases in which multiple orthologues might exist (Penfold et al., 2015). This should allow the vast amounts of information obtained from model organisms to be leveraged into novel or economically/medically relevant ones, such as crops and humans. Finally, further development of CSI will allow the simulation of new time series data using the inferred dynamical models.

Funding: This work was supported by EPSRC grants EP/I036575/1 & EP/J020281/1, BBSRC grant BB/F005806/1 and Monash Undergraduate Research Projects Abroad.

References

Greenfield, A., A. Madar, H. Ostrer, and R. Bonneau (2010): "DREAM4: Combining genetic and dynamic information to identify biological networks and dynamical models," PLoS One, **5**, e13397.

Hickman, R., C. Hill, C. Penfold, E. Breeze, L. Bowden, J. Moore, P. Zhang, A. Jackson, E. Cooke, F. Bewicke-Copley, A. Mead, J. Beynon, D. Wild, K. Denby, S. Ott, and V. Buchanan-Wollaston

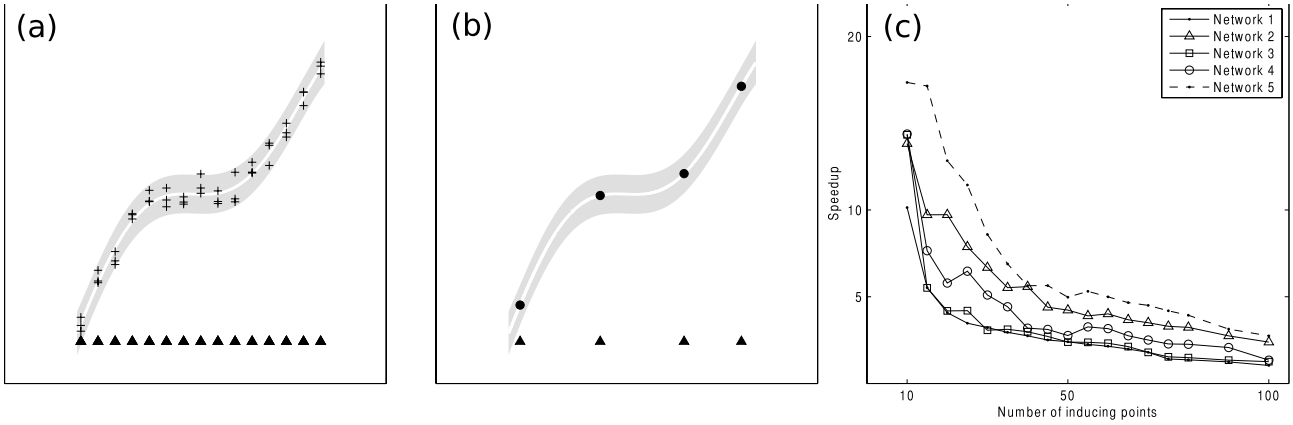

Figure 1: (a) A Gaussian process (GP) represents a prior over functions that captures the underlying behaviour of the system via the training data (+) at input locations (▲) and hyperparameters. (b) Sparse GPs such as the FITC approximation attempt to encode the behaviour underlying the system via a smaller number of inducing points (•) at input locations (▲). (c) The speedup of the CSI algorithm versus the version implementation of Penfold and Wild (2011) is shown for an increasing number of the inducing points, n, using the DREAM4 10-gene networks (Prill et al., 2010, Greenfield et al., 2010). When the number of inducing points is small (10) the speedup is between 10-20 fold with no degradation to the accuracy of the network reconstruction.

(2013): "A local regulatory network around three NAC transcription factors in stress responses and senescence in Arabidopsis leaves," Plant J., 75, 26–39.

Kent, N., S. Adams, A. Moorhouse, and K. Paszkiewicz (2011): "Chromatin particle spectrum analysis: a method for comparative chromatin structure analysis using paired-end mode next-generation dna sequencing," *Nucleic acids research*, 39, e26.

Klemm, S. L. (2008): *'Causal Structure Identification in Nonlinear Dynamical Systems'*, MPhil thesis, Department of Engineering, University of Cambridge, UK.

Penfold, C.A, J.B.A. Millar, and D.L. Wild (2015): "Inferring orthologous gene regulatory networks using interspecies data fusion," *Bioinformatics*, 10.1093/bioinformatics/btv267.

Penfold, C. A., V. Buchanan-Wollaston, K. Denby, and D. L. Wild (2012): "Nonparametric Bayesian inference for perturbed and orthologous gene regulatory networks," Bioinformatics, 28, i233–241.

Penfold, C. A. and D. L. Wild (2011): "How to infer gene networks from expression profiles, revisited," J R Soc Interface Focus, 1, 857–870.

Prill, R. J., D. Marbach, J. Saez-Rodriguez, P. K. Sorger, L. G. Alexopoulos, X. Xue, N. D. Clarke, G. Altan-Bonnet, and G. Stolovitzky (2010): "Towards a rigorous assessment of systems biology models: The DREAM3 challenges," PLoS One, **5**, e9202.

Quinonero-Candela, J., C. E. Ramussen, and C. K. I. Williams (2005): "Approximation methods for Gaussian process regression," J Mach Learn Res., 6, 1939–1959.

Snelson, E. and Z. Ghahramani (2006): "Sparse gaussian processes using pseudo-inputs," in Y. Weiss, B. Schölkopf, and J. Platt, eds., *Advances in Neural Information Processing Systems 18*, MIT Press, 1257–1264.