University of Warwick institutional repository: http://go.warwick.ac.uk/wrap

A Thesis Submitted for the Degree of PhD at the University of Warwick

http://go.warwick.ac.uk/wrap/45732

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it. Our policy information is available from the repository home page.

THE UNIVERSITY OF WARWICK

Library Declaration and Deposit Agreement

1. STUDENT DETAILS

Please complete the following:

Full name: …………………………………………………………………………………………….

University ID number: ………………………………………………………………………………

2. THESIS DEPOSIT

2.1 I understand that under my registration at the University, I am required to deposit my thesis with the University in BOTH hard copy and in digital format. The digital version should normally be saved as a single pdf file.

2.2 The hard copy will be housed in the University Library. The digital version will be deposited in the University's Institutional Repository (WRAP). Unless otherwise indicated (see 2.3 below) this will be made openly accessible on the Internet and will be supplied to the British Library to be made available online via its Electronic Theses Online Service (EThOS) service.
[At present, theses submitted for a Master's degree by Research (MA, MSc, LLM, MS or MMedSci) are not being deposited in WRAP and not being made available via EthOS. This may change in future.]

2.3 In exceptional circumstances, the Chair of the Board of Graduate Studies may grant permission for an embargo to be placed on public access to the hard copy thesis for a limited period. It is also possible to apply separately for an embargo on the digital version. (Further information is available in the *Guide to Examinations for Higher Degrees by Research*.)

2.4 *If you are depositing a thesis for a Master's degree by Research, please complete section (a) below. For all other research degrees, please complete both sections (a) and (b) below:*

(a) Hard Copy

I hereby deposit a hard copy of my thesis in the University Library to be made publicly available to readers (please delete as appropriate) EITHER immediately OR after an embargo period of ……….................... months/years as agreed by the Chair of the Board of Graduate Studies.

I agree that my thesis may be photocopied. YES / NO (*Please delete as appropriate*)

(b) Digital Copy

I hereby deposit a digital copy of my thesis to be held in WRAP and made available via EThOS.

Please choose one of the following options:

EITHER My thesis can be made publicly available online. YES / NO (*Please delete as appropriate*)

OR My thesis can be made publicly available only after…..[date] (Please give date)
YES / NO (*Please delete as appropriate*)

OR My full thesis cannot be made publicly available online but I am submitting a separately identified additional, abridged version that can be made available online.
YES / NO (*Please delete as appropriate*)

OR My thesis cannot be made publicly available online. YES / NO (*Please delete as appropriate*)

3. GRANTING OF NON-EXCLUSIVE RIGHTS

Whether I deposit my Work personally or through an assistant or other agent, I agree to the following:

Rights granted to the University of Warwick and the British Library and the user of the thesis through this agreement are non-exclusive. I retain all rights in the thesis in its present version or future versions. I agree that the institutional repository administrators and the British Library or their agents may, without changing content, digitise and migrate the thesis to any medium or format for the purpose of future preservation and accessibility.

4. DECLARATIONS

(a) I DECLARE THAT:

- I am the author and owner of the copyright in the thesis and/or I have the authority of the authors and owners of the copyright in the thesis to make this agreement. Reproduction of any part of this thesis for teaching or in academic or other forms of publication is subject to the normal limitations on the use of copyrighted materials and to the proper and full acknowledgement of its source.

- The digital version of the thesis I am supplying is the same version as the final, hard-bound copy submitted in completion of my degree, once any minor corrections have been completed.

- I have exercised reasonable care to ensure that the thesis is original, and does not to the best of my knowledge break any UK law or other Intellectual Property Right, or contain any confidential material.

- I understand that, through the medium of the Internet, files will be available to automated agents, and may be searched and copied by, for example, text mining and plagiarism detection software.

(b) IF I HAVE AGREED (in Section 2 above) TO MAKE MY THESIS PUBLICLY AVAILABLE DIGITALLY, I ALSO DECLARE THAT:

- I grant the University of Warwick and the British Library a licence to make available on the Internet the thesis in digitised format through the Institutional Repository and through the British Library via the EThOS service.

- If my thesis does include any substantial subsidiary material owned by third-party copyright holders, I have sought and obtained permission to include it in any version of my thesis available in digital format and that this permission encompasses the rights that I have granted to the University of Warwick and to the British Library.

5. LEGAL INFRINGEMENTS

I understand that neither the University of Warwick nor the British Library have any obligation to take legal action on behalf of myself, or other rights holders, in the event of infringement of intellectual property rights, breach of contract or of any other right, in the thesis.

Please sign this agreement and return it to the Graduate School Office when you submit your thesis.

Student's signature: ....…… Date: ....

THE UNIVERSITY OF
WARWICK

AUTHOR: **Edward R. Morrissey** DEGREE: **Ph.D.**

TITLE: **Bayesian Inference of Causal Gene Networks**

DATE OF DEPOSIT: ....

I agree that this thesis shall be available in accordance with the regulations governing the University of Warwick theses.

I agree that the summary of this thesis may be submitted for publication.

I **agree** that the thesis may be photocopied (single copies for study purposes only).

Theses with no restriction on photocopying will also be made available to the British Library for microfilming. The British Library may supply copies to individuals or libraries. subject to a statement from them that the copy is supplied for non-publishing purposes. All copies supplied by the British Library will carry the following statement:

> "Attention is drawn to the fact that the copyright of this thesis rests with its author. This copy of the thesis has been supplied on the condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without the author's written consent."

AUTHOR'S SIGNATURE: ....

USER'S DECLARATION

1. I undertake not to quote or make use of any information from this thesis without making acknowledgement to the author.
2. I further undertake to allow no-one else to use this thesis while it is in my care.

DATE SIGNATURE ADDRESS

....

....

....

....

....

Bayesian Inference of Causal Gene Networks

by

Edward R. Morrissey

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Systems Biology DTC

2012

Contents

Acknowledgments

I would like to thank my supervisors Nigel Burroughs and David Hodgson for their supervision and patience. I am especially grateful to Miguel A. Juárez both from an academic and personal point of view. Without his presence I would never have learnt nor enjoyed myself as much: gracias amigo. I would also like to thank Mirela Domijan for proof-reading and correcting my "Gramma".

I would like to thank my family for putting up with me during my thesis writing and making my time in Madrid enjoyable. Good times.

Thanks to Prof Collins for having me as a visiting scholar at Boston University. Also thanks to Henry Lee for discussions regarding possible projects. I am grateful to Alvaro Sanchez for housing me on several occasions during my stay.

I gratefully acknowledge the studentship received from EPSRC/BBSRC through the Systems Biology Doctoral Training Centre.

Declarations

This thesis has not been submitted for a degree at any another university.

The novel work in this thesis is presented in journal paper format. Prior to each paper there is a preface indicating which elements of the paper were contributed by me.

Abstract

Genes do not act alone, rather they form part of large interacting networks with certain genes regulating the activity of others. The structure of these networks is of great importance as it can produce emergent behaviour, for instance, oscillations in the expression of network genes or robustness to fluctuations. While some networks have been studied in detail, most networks underpinning biological processes have not been fully characterised. Elucidating the structure of these networks is of paramount importance to understand these biological processes.

With the advent of whole-genome gene expression measurement technology, a number of statistical methods have been put forward to predict the structure of gene networks from the individual gene measurements. This thesis focuses on the development of Bayesian statistical models for the inference of gene regulatory networks using time-series data.

Most models used for network inference rely on the assumption that regulation is linear. This assumption is known to be incorrect and when the interactions are highly non-linear can affect the accuracy of the retrieved network. In order to address this problem we developed an inference model that allows for non-linear interactions and benchmarked the model against a linear interaction model.

Next we addressed the problem of how to infer a network when replicate measurements are available. To analyse data with replicates we proposed two models that account for measurement error. The models were compared to the standard way of analysing replicate data, that is, calculating the mean/median of the data and treating it as a noise-free time-series.

Following the development of the models we implemented GRENITS, an R/Bioconductor package that integrates the models into a single free package. The package is faster than the previous implementations and is also easier to use.

Finally GRENITS was used to fit a network to a whole-genome time-series for the bacterium *Streptomyces coelicolor*. The accuracy of a sub-network of the inferred network was assessed by comparing gene expression dynamics across datasets collected under different experimental conditions.

Chapter 1

Introduction

1.1 Thesis Overview

Gene Regulatory Networks (GRNs) lie at the heart of many biological processes, from the timing mechanism of the plant *Arabidopsis thaliana* [McClung, 2008] to the control of cell differentiation in *Caenorhabditis elegans* [Raj et al., 2010].

This thesis focuses on the problem of predicting the structure of a GRN using expression measurements of the individual network genes. We focus on the development of Bayesian inference methods for the analysis of time-series. The methods developed are used to analyse a number of synthetic datasets as well as experimental datasets. Of the experimental data analysed, the data from the STREAM consortium is noteworthy, as this data was specifically designed for a systems biology approach.

1.1.1 STREAM

STREAM (STREptomyces Analysis of Metabolism) is part of SysMo, a European funded project with members from United Kingdom, Germany, The Netherlands, Spain and Norway. The project's aim is to use a systems biology approach to study the metabolic switch of the bacterium *Streptomyces coelicolor.*

Streptomyces coelicolor is the model representative of *Streptomyces*, a group of soil-dwelling bacteria with a highly complex secondary metabolism. Streptomycetes are of great interest to the pharmaceutical industry as they are able to synthesise around two-thirds of naturally derived antibiotics currently used [Bentley et al., 2002].

When *Streptomyces coelicolor* is starved of certain key nutrients it undergoes a metabolic transition shifting its metabolism from primary to secondary metabolism. Antibiotics are synthesised and secreted. Under certain conditions it may also form spores which will disperse and form new colonies.

For the STREAM project a novel fermentation set-up was implemented leading to highly reproducible results. The set-up was used to generate large, high-quality time-series of expression data using custom built microarrays. Proteomic and metabolomic data were also measured. Data was initially collected for three different experimental conditions known to trigger metabolic switching, as well as a knock-out mutant for one of the key regulators (*phoP*).

Analysis of the STREAM consortium data is covered in chapter 5.

1.1.2 Thesis Organisation

This thesis follows the format of a thesis based on publications. That is, novel published research is presented in journal paper format. Unpublished work is also included and written in paper format. Each paper is preceded by a preface containing a summary and context of the research work. All background information is covered in the introduction.

The *Introduction* covers genes, gene regulatory networks and measurement of gene expression. Next, models used to describe networks are discussed, as well as approximations used for the inference models presented in this thesis. Following this, a summary of Bayesian inference is included. Finally, how to fit a network to a whole genome dataset is discussed.

The first paper (*non-linear interactions paper*) covers the use of non-linear regulation functions for network inference. A common approximation used for inference models is the use of linear functions to describe regulatory interactions; while this approximation is useful, it may not always hold. In this paper we develop a non-linear interactions model and compare inferred networks using the non-linear model and a linear model. Both *in-silico* and experimental data are used for the comparison.

The second paper (*measurement error models paper*) covers the use of repeated measurements for network inference. The paper presents two model variants for the inference of networks using time-series with replicate measurements.

The inference models presented in the first two papers were implemented in MATLAB and required in-depth knowledge to use. In the third paper (*GRENITS paper*) we present an integrated, improved re-implementation of the models. The re-implementation is fast and easy-to-use, as well as being freely available.

The final paper (*STREAM paper*) describes unpublished work relating to the STREAM data. Here we use the GRENITS package to fit a network to a whole-genome data-set generated by the STREAM consortium. Following this, the results are benchmarked against other inference methods comparing across several STREAM generated datasets.

The work is summarised and possible improvements are discussed in the *Conclusions* section.

1.2 Gene Networks

1.2.1 Overview

Genes contain information needed to produce molecules that serve for cell construction and interaction with the environment. In general, the information encoded in the genes is used to synthesise messenger ribonucleic acid (mRNA) in a process called transcription. The mRNA molecule will in turn be used by the cell to synthesise a specific protein (the process called translation). Depending on the gene, the synthesised protein can serve a number of purposes. For instance, it can catalyse certain reactions or act as a signalling molecule.

Genes are not necessarily always being transcribed. Aside from the drain on energy and molecular resources that constant transcription would imply, certain biological processes would be detrimental to cell survival if they are active at the wrong moment. Certain proteins, called transcription factors (TFs), are able to affect the expression of other genes, either by activating or inhibiting the activity of the target genes. TFs can regulate large numbers of genes (the regulon) amongst which can be TFs, forming a gene regulatory network (GRN). A crucial aspect of these networks is the emergent behaviour that can arise due to the structure and, the interaction types and strengths. For instance, the structure of the circadian clock regulatory network in *Arabidopsis thaliana* [McClung, 2008] causes the expression levels of all the genes in the network to oscillate with a specific frequency.

Network structure can not only underpin the dynamic of gene expression, but can also confer robustness to fluctuations. An example of this was observed in a recent study by Raj et al. [2010]. The authors studied the regulatory network controlling intestinal cell differentiation in *Caenorhabditis elegans* embryos. By perturbing a number of elements of the network they were able to show that internal fluctuations, previously buffered by the network, started to dominate cell fate.

Aside from naturally occurring networks, a number of gene networks with interesting properties can be found in the field of synthetic biology. By using electronic circuits as inspiration small networks are designed using mathematical models and later inserted into cells and tuned to produce the desired behaviour. Published networks show a wide range of behaviours, including oscillators [Elowitz and Leibler, 2000], toggle switches [Gardner et al., 2000] and discrete counters [Friedland et al., 2009].

Not only is the elucidation of regulatory networks crucial for a better understanding of organisms and how they interact with the environment, it also offers the potential to improve antibiotic and drug discovery. An example of this is the use of networks to find key network elements sensitive to combinatorial drug attacks Nelander et al. [2008].

1.2.2 Network Inference

A number of experimental techniques that allow the discovery of regulatory interactions have been developed over the years. The most popular techniques are based on chromatin immunoprecipitation (e.g. ChIP-seq [Park, 2009]).

The idea behind the ChIP technique is to temporarily fix the TF molecules that are bound to DNA, shear the DNA and then use a molecule that will specifically bind the TF of interest and allow the extraction of the fragment of DNA bound to it. The extracted DNA fragment can then be mapped to the genome, this way finding which gene the TF was regulating.

A central step to this method is the use of a molecule that specifically binds the TF. This will usually be an antibody. These highly specific proteins have to be produced individually for each TF and must bind the TF both in its DNA bound form and in free solution. This is not always possible and when it is possible, this is not a straightforward process.

Genetic perturbations (*e.g.* knock outs, RNAi, overexpressors) can also be used to elicit regulatory interactions. By perturbing individual elements of the network some information on the network connectivity can be found. However this type of approach tends to be poor at separating direct network connections from knock on effects caused by the perturbation.

Given that it is currently possible to simultaneously measure the gene expression of the whole genome of almost any organism, at a specific point in time within an experiment, another approach that has emerged in recent years is the use of mathematical/statistical models to propose plausible network structures from the expressions of its genes [De Smet and Marchal, 2010].

1.3 Measuring Gene Expression

1.3.1 Measurement Technology

Expression measurements can be separated into single cell measurements and population measurements. Single cell measurements are normally carried out by genetically modifying cells in order to produce a detectable signal when the gene of interest is transcribed, for example, a reporter gene can be inserted upstream of a chosen gene. The reporter gene will normally encode a protein that produces some sort of visual signal such as fluorescence (GFP) [Chalfie et al., 1994]. When the gene of interest is transcribed so is the reporter gene, thus giving a measure of the activity of the gene.

Reporter gene measurements have the advantage that the cells need not be destroyed and can be followed over time. The downside is that only two or three genes can be measured for each cell.

A more common approach is to collect a large number of cells, lyse them and extract the mRNA. The mRNA pool is later transformed into its corresponding cDNA form. After this step, depending on the technology used, it will either be sequenced [Wang et al., 2009], bound to a chip [Ramsay, 1998] or measured using qRT-PCR [Nolan et al., 2006].

qRT-PCR produces reliable results even for weakly expressed genes, although it is normally limited to a handful of genes. RNA-seq is a novel method that does not rely on anticipating what DNA sequences will be measured and offers sequence counts rather than harder to interpret fluorescence units. This flexibility also allows simultaneous measurements of several organisms' transcriptome in one run which can be useful, for instance, when measuring virus host interactions [Yang et al., 2010].

However, currently microarray technology dominates the transcriptome measurement field. The technology relies on the use of DNA base pairing. Homologous strands of the transcript to be measured are synthesised and placed on a chip at specific locations. Later a solution containing a pool of the fluorescently tagged DNA to be measured is washed over it. By reading the fluorescence emitted for each gene a proxy for the expression of the gene can be found.

Rather than place a single long DNA strand representing a gene, multiple short DNA strands (probes) representing different sections of the gene of interest are normally used. Shorter strands avoid secondary structures of the DNA as well as the binding of contiguous probes. Once the probes have been read, software [Turro et al., 2007; Irizarry et al., 2003; Pearson et al., 2009] is used to correct systematic effects known to occur, make different chips comparable and summarise the probes into a single gene expression level.

1.3.2 Cell Population Synchronisation

The population approaches mentioned in the previous section benefit from the fact that the total amount of DNA is far larger than in the single cell approach. The downside of this is that cells are known to display heterogeneous behaviour even under homogeneous external conditions [Wilkinson, 2009]. For example, cells may be in different stages of the cell cycle, which would mean that different sets of genes will be expressed. When all the mRNA is pooled together and measured, the different signals are averaged leading to noisy dampened signals.

Some of this variation can be diminished by careful preparation of the experimental conditions. When studying the cell cycle of *Saccharomyces cerevisiae*, Granovskaia et al. [2010] used mating hormones to synchronise the cell cycle. *Saccharomyces cerevisiae* is capable of both asexual reproduction and a form of sexual reproduction. The hormones added in the experiment are naturally secreted mating hormones and have the effect of arresting the progress of the cell cycle at a very specific point, synchronising the whole population. The hormone is then washed away and the cell cycle resumes.

For experiments with multicellular organisms, not only is it of interest to synchronise populations of cells, it is also necessary to synchronise populations of a multicellular organism, as each measurement normally requires the destruction of the whole organism. In the case of the model plant *Arabidopsis thaliana*'s circadian clock, the clock circuit can be synchronised by exposing the population of plants to controlled light/dark cycles [Morrissey et al., 2010].

Population synchronisation for the STREAM consortium experiments was achieved as follows. All the *Streptomyces coelicolor* cells were added in spore (dormant) form in a single batch, meaning the whole population underwent the same growth conditions. The experiment was conducted in fermenters where the environmental conditions can be closely controlled. Diluted oxygen was controlled and set at a constant level, which was found to produce strongly reproducible results.

One of the nutrients in the medium was introduced in a limiting concentration, so as to run out a few hours after beginning the sampling regime. For example, in the first experiment, a limiting amount of phosphate was used. Once phosphate ran out, the population responded in a synchronised manner by switching on phosphate scavenging mechanisms, shifting to secondary metabolism and later secreting antibiotics [Nieselt et al., 2010].

1.4 Gene Network Models

In order to build a network model to fit to data, a good starting point is to study GRN models used to describe known networks.

The expression dynamics of genes forming GRNs can be mathematically described in a number of ways. One of the most popular ways is by using mechanistically derived models.

1.4.1 Mechanistic Models

Mechanistic models based on Ordinary Differential Equations (ODEs) and chemical reaction kinetics have been successfully used in the past to describe known GRNs [Ronen et al., 2002; Locke et al., 2006]. ODEs are used to describe the rate of change of the concentration of the molecules involved in the network and chemical reaction kinetics are used to describe the interaction between the TF and its target gene.

Gene Regulation Model

As mentioned previously, the process of gene regulation requires the TF gene to be transcribed to mRNA. Following this, the mRNA must be translated and the resulting protein must then bind a specific region close to the target gene, either activating or inhibiting its expression.

Each of these steps can be modelled as an ODE describing the change in concentration of the molecule over time, accounting for both molecule synthesis and degradation. The first step of the process is the transcription of the TF gene:

$$\frac{dTF_m}{dt} = -\lambda_{TF_m} TF_m + h(t), \tag{1.1}$$

where TF_m is the concentration of TF mRNA, λ_{TF_m} is the rate constant of degradation and $h(t)$ is its corresponding transcription rate, which could be due to some other TF or an external factor.

The term $-\lambda_{TF_m} TF_m$ accounts for mRNA decay. The decay can be due to the mRNA's limited half-life, as well as the organism's degradation pathways that degrade the mRNA in a controlled fashion [Houseley and Tollervey, 2009].

Next the mRNA will get translated into protein

$$\frac{dTF_p}{dt} = -\lambda_{TF_p} TF_p + \alpha TF_m, \tag{1.2}$$

where TF_p is the concentration of the TF protein, λ_{TF_p} is the rate constant of degradation and α is the rate constant of translation. Finally, the concentration of the target gene's mRNA (G_m) can be described by

$$\frac{dG_m}{dt} = -\lambda_{G_m} G_m + \beta f(TF_p), \tag{1.3}$$

where λ_{G_m} is the degradation rate constant of the mRNA and the product $\beta f(TF_p)$ represents the transcription rate due to the TF. β is the rate constant of transcription and $f(\cdot)$ a function that describes either the fraction of time the TF is bound to the DNA (activation) or the faction of time the DNA binding site is unbound (inhibition), and can be derived using chemical kinetics.

Assuming that the binding and unbinding of the TF to the DNA is at steady state and that the TF binds as monomer, the fraction of time the TF is bound to the DNA can be described by $\frac{TF_p}{k+TF_p}$, whereas the fraction of time the DNA is not bound will be $\frac{1}{k+TF_p}$. The activating function is known as the Michaelis-Menten equation and k is the Michaelis-Menten constant.

It is also possible to derive more complex functions. For instance, when the TF is assumed to bind cooperatively, the resulting functions are $\frac{TF_p^n}{k^n+TF_p^n}$ for activation and $\frac{k^n}{k^n+TF_p^n}$ for inhibition. Here n represents number of TF that are required to bind the DNA and k is the Hill constant. The activating functional form is known as the Hill function.

When several TF regulate a target gene, the functional form can also be derived. For example, when an activator and inhibitor regulate a target gene, the transcription rate can be modelled as a product of the activating and inhibiting function.

Circadian Clock Model

GRNs can be modelled by coupling ODEs of the type described in the previous section. An example of this is the circadian clock network in *Arabidopsis thaliana.* Using coupled ODEs Locke et al. [2006] built a five gene network that was able to explain the period of circadian oscillations experimentally observed. The model includes the usual transcription and translation steps, although decay is modelled using Michaelis-Menten kinetics. Additionally the model accounts for protein location. The model assumes that the protein is synthesised in the cytoplasm and uses an ODE to model protein transport into and out of the nucleus. The model also includes the effect of light on certain genes in the network, a crucial aspect of the circadian clock.

Model parameters were fitted using oscillation period information from several experiments, including knock-outs (KOs). Interestingly, in order to adequately explain this data, it was necessary to include an unknown hypothetical gene. A specific gene (GI) was proposed to be the hypothetical gene and later accepted to be correct [McClung, 2008].

1.4.2 From Mechanistic Models to Approximate Models

While in general mechanistic models are able to capture fine details of the dynamics of the GRNs, from the point of view of inference, they may be too complex, especially considering the type of data used.

Microarray data is noisy, and because of its cost most experiments tend to use a handful of arrays. This kind of data is not adequate for inference of a complex model, even more so if the structure of the network is unknown and all possible interactions must be modelled and queried. A simpler approximate model will be easier to fit and it will be more computationally efficient.

TF mRNA concentration as protein concentration

Very rarely are quantitative protein measurements available, so most models use the approximation that TF mRNA measurements constitute a good proxy for TF protein measurements. This approximation is quite convenient as it removes a large number of parameters from the model and in general mRNA will be translated into protein after a time-lapse. In the case of prokaryotes, transcription and translation occurs simultaneously.

Barenco et al. [2006] and Honkela et al. [2010] argue that protein dynamics can play a prominent role and should be considered, especially in the case where the decay parameters of the mRNA and protein are very different. In the first case the authors use known targets of a TF along with mRNA decay measurements (of at least one target gene) to estimate the dynamics of the unmeasured active form of the TF protein. With this they then predict further targets of the TF. Honkela et al. [2010] do not make use of any prior regulation knowledge and use mRNA measurements of a TF and a candidate target to infer the protein concentration that best explains the potential interaction.

Linear additive approximation

A further convenient simplification is to model the regulatory interaction with a linear function. As the interaction function will be some kind of sigmoid, a linear approximation may work well, especially for values of the TF concentration away from the saturation region.

A linear approximation will not only reduce the number of parameters, but will also make the structure of the model easier to deal with when inferring the model parameters.

While a linear approximation is convenient, using this approximation can affect the retrieved network. Efforts to address this issue include using sigmoid functions [Yip et al., 2010], flexible no-linear functions (Chapter 2, [Äijö and Lähdesmäki, 2009; Kim et al., 2004]) and linear interactions truncated at a maximum and minimum value to mimic known transcription saturation and minimum concentrations effects [Bonneau et al., 2006].

These methods are equipped to deal with non-linear effects, but they tend to require more data to work adequately, as they need to infer several parameters to describe the interactions. As well as this, they are also computationally slower than their linear equivalent.

Discrete Time Approximation

Gene expression will often be measured over a small number of time-points, leading to a small discrete-time set of measurements (time-series). In order to fit this kind of data to continuous time models, most network inference models use a discrete time approximation to the derivative $\frac{x^{t+1}-x^{t}}{\Delta t}$.

Joining this with the previous approximations and re-parameterising leads to the widely used linear autoregressive model (AR1):

$$X_i^{t+1} = \mu_i + aX_i^t + \sum_{j \neq i} b_{ij} X_j^t, \tag{1.4}$$

where μ_i is the basal mRNA level of regulated gene i, X_i^t is the mRNA concentration of regulated gene i at time t, X_j^t represent the mRNA concentration of the regulator gene j at time t, aX_i^t is the fraction of mRNA remaining from the previous time-point (mass conservation) and b_ij represent regulation strength of gene j on gene i.

While the discrete time approximation is widely used, other solutions have been proposed. For instance, Honkela et al. [2010] use an integral-form solution to the ODE's to perform further inference, whereas Bansal et al. [2006] interpolate and smooth the mRNA expression profiles and use this to explicitly calculate the value of the derivative.

Another approximation used considers the system to be at steady state making $\frac{dx}{dt} = 0$ in (1.3). This has the advantage that different experiments can be pooled together allowing the use of hundreds of data-points [di Bernardo et al., 2005; Cosgrove et al., 2008]. The downside is that causality can be hard to elicit. Gardner et al. [2003]; Lorenz et al. [2009] get around the causality problem by externally perturbing the steady state with a number of controlled mRNA over-expressions which are included in the model. Bonneau et al. [2006] combine both time-series and static measurements allowing for some causality to be elicited.

1.5 Bayesian Inference

Inferring the model parameters from the data requires the use of statistical methodology. The chosen methodology for the inference models presented in this thesis is Bayesian inference.

The idea is to combine the data, a model that explains the data and prior knowledge of the model parameters, in order to derive a probability distribution of the model parameters.

A central element of Bayesian inference is the Likelihood function which will be explained next.

1.5.1 Likelihood

The Likelihood function is a conditional probability distribution that relates the data and the parameters as

$$L(\theta) \equiv P(X \mid \theta). \tag{1.5}$$

Here X represents all the data and θ all the parameters.

For the models included in this thesis, the likelihood is made up of a systematic part (e.g. AR1) and a stochastic part. The stochastic part represents random deviations from the systematic model.

We use the common assumption that the fluctuations are independent and identically distributed. Generally, the distribution chosen for the fluctuations is Gaussian.

In the case of gene regulatory networks, the likelihood function represents the joint probability of the gene expression data given the model parameters and the network structure. For each network structure we can define this joint probability using Bayesian Networks [Pearl, 1985].

Bayesian Networks

A Bayesian Network (BN) is a graphical model defined as a directed acyclic graph, where the nodes are conditionally independent of all other nodes when conditioned on their parent nodes. In other words

$$P(X_i|X_1, X_2, ..., X_{i-1}) = P(X_i|pa_i), \tag{1.6}$$

where X are nodes in the network and pa_i are all parent nodes of node X_i. Combining this property with the chain rule of probability calculus the joint probability of all nodes can be described as

$$P(X_1, X_2, ..., X_i) = \prod_i P(X_i|pa_i). \tag{1.7}$$

A simple example of a Bayesian Network can be seen in figure 1.1. The joint probability of this network can be written as:

$$P(A, B, C, D) = P(D \mid B, C)P(B \mid A)P(C \mid A)P(A). \tag{1.8}$$

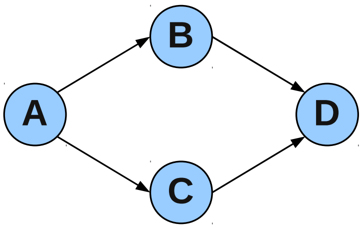


Figure 1.1: Example of Bayesian Network. Here A, B, C and D are variables and the edges are conditional relations.

In the case of GRNs, the nodes represent the expression level for each gene and the edges the regulatory relation. The conditional probabilities that define the regulatory interactions can be calculated for example using the AR1 model and a chosen distribution for the error [Grzegorczyk, 2010].

The conditionals can also be modelled using a purely probabilistic model, for example, Yu et al. [2004] discretise the data and use a multinomial distribution to describe the conditional relations.

Bayesian Networks are restricted to acyclic networks. When analysing biological networks (e.g. GRNs), this can be a severely limiting restriction, as loops are central to biological networks. A simple example is positive self regulation of a

TF, which allows a fast explosive response when triggered.

Also, when trying to infer the structure of the network, two networks with different conditional dependence structure can give the same joint probability. For instance, a given network with the same structure but with edges pointing in different directions will, in general, have the same joint probability (graph equivalence), making it impossible to elicit causality without more information (e.g. external perturbations).

If the data being analysed is a time-series, Dynamic Bayesian Networks (DBN) can be used. A DBN is a BN that accounts for time. The nodes are now the values of the variables at different time-points and the relations can only go forwards in time. For instance the BN example from figure 1.1 could be transformed to a DBN as seen in figure 1.2.

DBNs allow for the inference of causality and also allow for GRN loops.

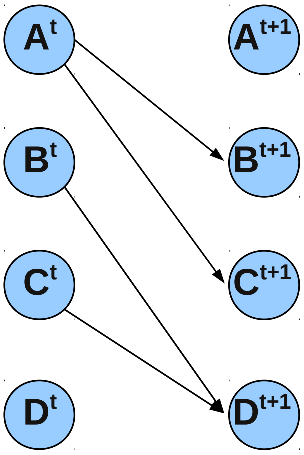


Figure 1.2: Example of Dynamic Bayesian Network. Here A^t, B^t, C^t and D^t represent network variables at time t and the edges are conditional relations.

Causal Bayesian Networks

The Bayesian Network representing the true Gene Regulatory Network is a Causal Bayesian Network. A Causal Bayesian Network or Causal Network [Pearl, 2000] is a Bayesian Network where all edges in the network are causal. This definition is particularly important when considering modifications to the network. With a Causal Bayesian Network, it is possible to predict the outcome of external interventions.

1.5.2 Posterior Distribution

As seen earlier, the likelihood is defined as the probability of the data given the parameters $P(X \mid \theta)$. In order to find the distribution of the parameters given the data $P(\theta \mid X)$ we must use Bayes' theorem [Bernardo and Smith, 1994]:

$$P(\theta \mid X) = \frac{P(X \mid \theta)P(\theta)}{P(X)}, \tag{1.9}$$

where the left hand term $P(\theta \mid X)$ is the posterior, $P(X \mid \theta)$ is the likelihood, $P(\theta)$ is the prior distribution of the parameters and $P(X)$ is the probability of the data under the model, also known as the Evidence.

The prior is a probability distribution that encodes the information about model parameters, prior to analysing the data. There is some freedom in the choice of priors (see below).

Noteworthy is the Evidence term. Calculating this term explicitly requires a potentially high dimension integral

$$P(X) = \int P(X \mid \theta)P(\theta)d\theta. \tag{1.10}$$

While it may not be possible to calculate the Evidence term analytically, several approximate techniques have been developed to find the posterior, that avoid dealing with the Evidence term.

The Evidence term is of interest, as it is the basis for a type of Bayesian Model Selection: Bayes Factors [Kass and Raftery, 1995]. Bayes Factors are used for pairwise model fit comparisons. This comparison has the appealing characteristic that it automatically includes Occam's razor. In other words, BFs account for model complexity and thus favour less complex models with similar fits.

Equation 1.9 can be interpreted as a learning step that can be iterated. The prior is seen as the distribution over the parameters using all the data available before acquiring data X and the posterior $P(\theta \mid X)$ represents the updated distribution once X is made available, thus, if further data Y becomes available, then we can use the posterior from $P(\theta \mid X)$ as prior for likelihood $P(Y|\theta)$.

Priors

Strictly speaking, the prior is a probability distribution that should convey the prior knowledge of the parameters before the data is analysed. Working on biological problems, it is sometimes possible to find information from previous studies. In some cases this information can be incorporated into the prior [Werhli and Husmeier, 2007].

A more common scenario is that of little prior knowledge of the parameter in question. In this case, it is of interest to use an uninformative prior. In other words, a prior that will have a minimum impact on the posterior. The most obvious way of achieving this is to use a distribution that is flat or as flat as possible over the range of the parameter, for example a uniform distribution. A function of this kind, if the support is not bounded, is called an improper distribution as integrating over the parameter leads to an infinite value. An improper prior must be used with caution as, depending on the amount and strength of the data, it can lead to the posterior being improper.

Finding an uninformative prior is typically not simple. For example Jeffreys [1946] noted that for a specific model a flat prior which should have been uninformative had a different effect if the physical units of the model were changed, this led him to propose a method for calculating a prior using the likelihood (Jeffreys prior), that can be shown to be invariant under change of physical units and other re-parametrisations.

Similarly Bernardo [1979] proposed the idea of Reference Priors in order to select a prior with minimal influence on the posterior. Reference Priors are calculated by maximising the Kullback - Leibler divergence between the posterior and the prior distributions.

Some authors use the prior to control parameters that are problematic or hard to estimate. In some cases this involves using the data to estimate values for the prior [Griffin and Brown, 2010].

Priors can also be used to share information between related parameters (hierarchical priors). This can be useful if, for example, data arising from experiments are regarded as somehow related. With a hierarchical prior, information can be shared without necessarily fixing all the parameters to the same value.

For practical reasons most authors choose priors that fall into the category of conjugate priors. This is, priors which yield a posterior within the same family of distributions. By choosing conjugate priors, analytical calculation of the posterior conditional distribution and integration tend to become easier.

Approximating the Posterior Distribution

As mentioned earlier, evaluation of the Evidence term is needed to obtain the normalised posterior distribution, requiring the calculation of a high-dimensional integral. In some cases the integral can be calculated analytically, although this usually requires use of very specific parametrisation.

A number of computational methods have been put forward in order to approximate the posterior distribution. By far the most used methods are Markov chain Monte Carlo (MCMC) algorithms. These sampling algorithms make use of the un-normalised posterior $P(X \mid \theta)P(\theta)$ to draw samples from the normalised posterior distribution. If run for a sufficiently large number of iterations the posterior can be well approximated. Depending on how the algorithm is constructed it may take a long time to produce an adequate approximation to the posterior. For this reason the algorithms have to be carefully constructed and tuned.

Of all MCMC algorithms, Metropolis-Hastings (MH) algorithms are the most used. These algorithms use a candidate value drawn from a proposal distribution and a rejection step. Key to this algorithm is the choice of the proposal distribution, as an inadequate choice can severely affect the efficiency of the sampler

A special case of the MH algorithm is the Gibbs sampling algorithm, where if the proposal distribution used is the conditional posterior distribution it can be shown that the proposed value will always be accepted. This is particularly convenient as it becomes computationally more efficient, although it does generate highly correlated samples [Casella and George, 1992].

While sampling methods are widely used, other methods exist for approximating the posterior distribution. Specifically in the field of network inference, Beal et al. [2005] use an AR1 model with hidden variables and find the posterior using a Variational Bayes approach.

Variational Bayes methodology is used to find an analytical approximation to the posterior distribution. The method works by starting with a factorised free form variational distribution and then using the Kullback-Leibler divergence in order to select a specific distribution and parameters that best approximate the posterior distribution [Beal and Ghahramani, 2003].

1.6 Network Topology Inference

In order to infer a network, all possible network links (i.e. regulatory interactions) must be included in the model. The manner in which the links are then queried varies. Some inference methods will fit a fully connected network and then assign zero to those connections that have a weak effect, while others search through the space of networks and use some formal criteria to decide which network configuration is better supported by the data.

GRN are typically sparse. Given that there will be potentially hundreds of candidate regulators, most of which are expected to not have an effect, it is of interest to use an approach that searches for solutions in the space of sparse networks.

Also, for most experiments the number of measured genes far outnumbers the number of experiments. A full gene network, assuming that each interaction is parametrised by a single parameter, will have at least G^2 parameters. For n experiments there will be $G \times n$ measurements, meaning that the data will be insufficient to determine the parameters and that, if using Bayesian inference, most of the information in the posterior will come from the prior. This scenario appears frequently in statistics and is referred to as $p \gg n$ scenario.

This problem can be side-stepped by constraining the solutions to sparse networks.

1.6.1 Shrinkage Methods

One way to find a sparse solution is to use a shrinkage approach. In regression problems, shrinkage attempts to draw all the coefficients towards zero. A well known case is the LASSO method [Tibshirani, 1996], a regularised version of the least squares problem. Rather than attempting to solve the Ordinary Least Squares problem

$$\hat{\beta}_{OLS} = \text{argmin}_\beta \mid\mid y - X\beta \mid\mid_2 \tag{1.11}$$

a version including a regularising term is used

$$\hat{\beta}_{LASSO} = \text{argmin}_\beta \big(\mid\mid y - X\beta \mid\mid_2 - \alpha \mid\mid \beta \mid\mid_1\big). \tag{1.12}$$

Here β are the regression coefficients, X are the regressors, y the response and α is the shrinkage parameter. The LASSO method has been used in network inference, for instance by Bonneau et al. [2006] and Cosgrove et al. [2008].

From a Bayesian perspective, LASSO regression can be viewed as a Bayesian regression using a Laplace prior $P(x) = \frac{\alpha}{2} \exp\left(-\alpha \mid\mid x \mid\mid_1\right)$ on the regression coefficients [Hans, 2009]. As the Laplace Distribution is strongly peaked at zero, it is easy to see how this would drag the coefficients towards zero (figure 1.3).

Drawing from the fact that the Laplace distribution can be described as a scale mixture of normals $P(x|0,\alpha) = \int N(x \mid 0, \tau) Exp\left(\tau \mid \alpha^2/2\right) d\tau$, other similar distributions can be derived.

For example, Griffin and Brown [2010] use a gamma mixing distribution and show that the Bayesian LASSO is a special case of this normal-gamma distribution. They also show that the gamma distribution allows for heavier tails, meaning that the non zero coefficients are less affected by the shrinkage.

These types of priors have been used extensively for QTL (quantitative trait

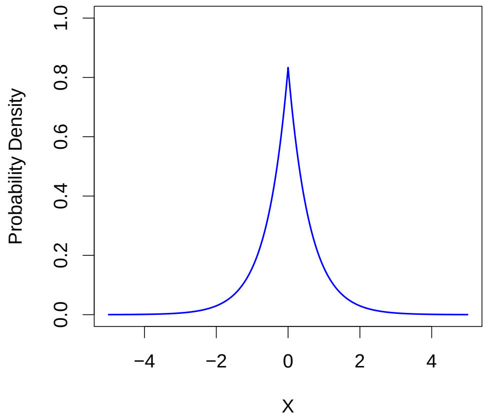


Figure 1.3: Plot of Laplace probability distribution.

loci) mapping, for example a Jeffreys scale prior was used $P(\tau) = \tau^{-1}$ [Xu, 2003], producing aggressive shrinkage. A student distribution prior was also proposed for shrinkage [Yi and Xu, 2008], which was implemented by using an inverse gamma mixing distribution and was shown to work as well as the Bayesian LASSO for the experimental data shown.

The fact that the distributions can be described as a scale mixture of normals is particularly convenient, as MCMC samplers can be easily constructed for them.

Shrinkage methods do not explicitly select variables. Some further criteria must be used to decide whether or not a coefficient can be considered to be zero. This could be done for instance by using a threshold on the scaled coefficients.

1.6.2 Bayesian Model Selection

In its general from, Bayesian model selection allows the formal comparison of model-fits to data. The idea was introduced by Harold Jeffreys (although also separately attributed to Alan Turing) [Kass and Raftery, 1995], as a method to compare two scientific theories. The comparison was formulated in terms of what are now known as Bayes Factors (BF):

$$BF_{12} = \frac{P(X \mid M_1)}{P(X \mid M_2)}, \tag{1.13}$$

where M_i is model i and $P(X \mid M_1)$ is the Evidence from M_1. Jeffreys [1998] provides a table of intervals to interpret the value of BF. The BF can also be interpreted as posterior odds for the case where the two models have equal prior probabilities.

Also referred to as Bayesian hypothesis testing, BFs are intrinsically different to classical hypothesis testing, as two models are explicitly compared rather than attempting to disprove a single model (with the exception of the likelihood ratio test). An appealing characteristic of the BF, is the fact that a balance between model fit and model complexity is struck.

A more general formulation of model selection can be found by considering the posterior probability of each model:

$$P(M_i \mid X) = \frac{P(X \mid M_i)P(M_i)}{\sum_k P(X \mid M_k)P(M_k)}, \tag{1.14}$$

where each M_i represents a different model and k is added over all possible models.

Evaluating (1.14) requires calculating high-dimensional integrals and doing so for a potentially enormous model space. MCMC alternatives exist [Brooks, 1998].

Bayesian Variable Selection

The case where Bayesian model comparison is used to determine which subset of a model's parameters best explain the data is referred to as Bayesian Variable Selection (BVS). Given that a balance will be struck between complexity and fit, when applied to the selection of a subset of regressors, under similar data-fits the sparser option will be favoured. This type of sparsity induction differs to the shrinkage methods in that the non-zero coefficients are not affected by the shrinkage mechanism and not only is the value of the coefficients inferred, but also a probability for each network or a marginal probability for each network connection is also inferred.

As mentioned earlier, while BVS requires dealing with the Evidence term, MCMC alternatives have been proposed. One such approach (used for the models described in this thesis) is Gibbs Variable Selection (GVS) [Dellaportas et al., 1997]. GVS uses a "slab and spike" prior [O'Hara, 2009] on the coefficients. The so-called "spike and slab" prior is a mixture distribution which consists of a distribution that concentrates all of its mass at zero (the "spike") and a flat distribution centred elsewhere (the "slab"). This is implemented by assigning an indicator (0 or 1) variable γ_{ij} to each coefficient in the model.

$$X_i^{t+1} = \mu_i + aX_i^t + \sum_{j \neq i} \gamma_{ij} b_{ij} X_j^t. \tag{1.15}$$

From the point of view of network inference, the γ_{ij} can be interpreted as defining the topology of the network. When $\gamma_{ij} = 0$, TF X_j has no effect on X_i, in other words it is excluded from the model.

During MCMC, all variables will be sampled from as usual including the γ_{ij}, which can be sampled from using a Gibbs update step. GVS generates an unusual situation in that, during MCMC, when $\gamma_{ij} = 0$ the value of the corresponding coefficient b_{ij} has no effect on the likelihood, but still has a value through the prior. Though the posterior remains unaffected by the prior for $\gamma_{ij} = 0$, the value of b_{ij} will affect efficiency of the MCMC. This can be taken advantage of by choosing a different prior for the case where $\gamma_{ij} = 0$. This prior, called a pseudo-prior, is chosen so as to improve the MCMC efficiency. A more efficient GVS approach can be found by using a simultaneous block update of γ_{ij} and b_{ij}, which produces a far more efficient MCMC and avoids having to deal with pseudo-priors (See Chapter 2).

It is likely that several networks can explain the data similarly. Rather than choosing the highest probability network, analysing the marginal distributions $P(\gamma_{ij} \mid X)$ allows combining the information from all networks.

BVS can also be carried out using the Evidence explicitly, though this normally requires using analytical approximations to estimate the Evidence term. For instance, the Laplace approximation assumes that the posterior is sharply peaked around a unique maximum. By making this assumption, the posterior can be approximated by a Gaussian and can be analytically integrated. This approach was used for network inference by Kim et al. [2004], where the authors used Penalised Splines and the Laplace approximation to derive a network score.

1.7 Network Inference for a Whole-Genome Data-Set

Attempting to infer a network using DBN and BVS on a whole-genome dataset can be problematic. In general, organisms will have of the order of thousands of genes, which is currently too many genes for an MCMC sampler to run in a realistic amount of time. Also, the number of genes will far outnumber the data points meaning an extreme case of the $p >> n$ scenario, which although addressed by BVS, can also have an effect on the inferred network.

1.7.1 Preprocessing and Filtering Data

While an organism can have of the order of thousands of genes, usually only a fraction will actually be expressed in a given experiment. By using a differential expression method, the dataset can be drastically reduced. For the STREAM analysis of the first time-series, BATS (Bayesian Analysis of Time Series) [Angelini et al., 2008] was run and reduced the dataset from 8000 genes to approximately 1000 genes.

Restricting the analysis to dynamically active genes will reduce the dataset considerably, though usually this will not be sufficient. The size of the networks to be fitted can be reduced by allowing only a subset of genes to act as regulators.

Even for non-model organisms such as *Streptomyces coelicolor*, information about the function of each gene can be found (e.g. http://strepdb.streptomyces.org.uk/). Most of the the information will be putative, as it comes from sequence similarity analysis of other organisms, as well as from the presence of certain protein domains, such as the DNA binding helix-turn-helix protein motif. By selecting as regulators those genes that have the ability to affect the expression of other genes either directly (TF) or indirectly (Kinases), the number of regulators can be reduced again. Out of the 1000 genes mentioned for the STREAM data, only ~ 70 qualified as regulators.

A final preprocessing step is to run a clustering algorithm on the regulators, as often a number of the regulators will cluster into groups of indistinguishable gene expression profiles. In these cases it is reasonable to cluster those genes into a single node and use either one of the genes or a summary statistic of the cluster for further inference. For the STREAM data, clustering the regulators was done using the Spline Cluster software [Heard et al., 2005] and led to ~ 30 dynamically distinct expression profiles. Using ontology and experts' opinion, those genes regarded as more biologically meaningful were selected to represent the rest of the genes within the corresponding cluster.

1.7.2 Co-expressed Genes and the AR1 Model

The combination of discrete time measurements and noise can render certain gene expression profiles as indistinguishable. Similarly, genes regulated by the same TF can also present indistinguishable expression profiles. Fitting an AR1 network model to data of this kind can lead to spurious connections.

When fitting the AR1 model to the STREAM data, preliminary runs showed an unusual behaviour for an operon gene (*phoP*). Operons are sets of genes that are transcribed as a single unit and therefore usually appear as co-expressed. When inferring the regulators of *phoP*, another gene in the same operon was incorrectly identified as a regulator. Interestingly, the probability of being a regulator of *phoP* was distributed between the self interaction and the operon gene. As described in section 1.4.2, a self interaction term representing mass conserved from the previous time-point is expected. If the expression profiles of two genes are identical, then the self interaction term can be replaced with a regulation term by an identical gene, leading to a (potentially) spurious connection. To overcome this problem, the self interaction term for the models presented in this thesis (chapters 2, 3, 4 and 5) is forced to be permanently present. In further runs this modification proved to remove the problem.

Fixing the self interaction has a further implication. In cases where the regulator and regulated gene share similar profiles it is likely that regulation will not be predicted. For this reason it is also advisable to use a clustering method and consider the possibility of genes co-clustered with a TF as potential regulon genes.

1.8 Other Network Inference Methods

The network inference methods referenced throughout this introduction correspond to methods similar to those presented in this thesis. The problem of network inference has been tackled by scientists from varied backgrounds, leading to a number of very different approaches.

Information theory methods have been used, with promising results. For example CLR [Faith et al., 2007] uses mutual information, an undirected measure of dependence between two random variates, to assess the dependence between a TF and a potential target. The mutual information value is then corrected for the background distributions of mutual information between each of the pair of genes with all other genes. Using a large compendium of *Escherichia coli* arrays and known interactions from regulonDB, the authors were able to show that their method outperformed the other methods considered. Mutual information will infer undirected interactions and only considers pairwise interactions, therefore indirect interactions will appear as significant. To address the problem of indirect interactions, Margolin et al. [2006] propose an algorithm (ARACNE) which first calculates a mutual information network and then uses the mutual information scores of triplets of interacting genes to prune the network in an attempt to remove indirect interactions.

Statistical Physics approaches have also been used, for example Werhli and Husmeier [2007] use Gibbs Free Energy to codify prior network information. Berg [2008] uses a thermodynamic approach to infer targets of a specific TF. From the field of econometrics, Granger Causality [Zou and Feng, 2009] has been applied to network inference. Granger Causality attempts to elicit whether or not one time-series has an effect on another, this is done via time delayed regressions and comparison of the variance of the residuals.

A number of methods address the problem of co-expression mentioned in the previous section. These methods, referred to as module based methods, either first infer modules and then predict interactions between modules [Bonneau, 2007] or simultaneously predict both [Segal et al., 2003; Hirose et al., 2008].

While not a network inference method *per se*, Liao et al. [2003] use microarray data along with known network structure to predict which parts of the network are active in a given experiment. This is an important problem to address given that although a TF may be expressed it could require activation via phosphorylation or the presence of a metabolite [Cho et al., 2008]. Sabatti and James [2006] extend this approach by using binding site predictions to propose putative connections and then use a similar approach to predict which of these binding sites might be active/real. Binding site predictions tend to produce a large number of false positives, though they have proven to predict correct binding sites.

Chapter 2

Non-Linear Interaction Model

Most network inference models make the assumption that the functional form of the regulatory interaction is linear. While the interaction is known to be better described by a sigmoid function, the linear approximation reduces the number of parameters and simplifies inference.

In this paper we propose an alternative model where the interactions are allowed to be non-linear. Contrary to other approaches that attempt to accommodate non-linearities, no specific functional form is chosen. This is achieved by using spline functions. Splines are flexible functions that are defined as a linear combination of a set of basis curves. Their flexibility is controlled by a single parameter. In the limit where the parameter is "large" the curve becomes linear. We use this property to propose a prior for this parameter that will bias the regression towards linearity unless there is strong evidence for non-linearity. This prior will avoid spurious complex functions that could lead to overfitting.

As the posterior is analytically intractable, we constructed a Markov Chain Monte Carlo algorithm to sample from the posterior. The sampler was built using Gibbs updates for all variables, except for the network topology variables. In the literature these are usually updated with Gibbs updates, but we found that these updates generated convergence problems. The problem stems from the relation between the coefficients and the topology variables. In order to improve convergence we propose a novel block update for the topology variables and the coefficients, and show that convergence is greatly improved compared to the separate Gibbs updates.

We build a similar model with linear interactions allowing us to compare the effects of inference on non-linear data using both a linear and non-linear model. Using a number of synthetic and experimental datasets we show that the networks inferred by the non-linear model are more accurate than the linear model ones.

2.1 Contribution

The initial idea was formulated by Morrissey, Juárez and Burroughs. The mathematical expressions used for the sampler were calculated separately by Morrissey and Juárez, and later compared. Both the computational and theoretical aspects of the joint update of the Gibbs Variables was carried out by Morrissey under the supervision of Burroughs. MATLAB code (serial and parallel) was in its majority (approximately 90 percent) contributed by Morrissey, with Juárez contributing the functions to calculate the design matrix for the Splines. Theoretical and sampler related issues were discussed and solved between Morrissey and Juárez, under the supervision of Burroughs. Morrissey generated the synthetic data, ran the models and analysed the outputs.

Morrissey's contribution to the writing was as follows. During the draft stage, Morrissey wrote the circadian clock section and the description of the Gibbs Variables block update, as well as contributing several sections of the introduction, discrete time examples and the conclusions. Morrissey participated in the revision of the draft suggesting changes and edits. During the journal submission process, Morrissey and Juárez worked on addressing the problems raised by the reviewers with the supervision of Burroughs.

The microarray data as well as the description of the experimental design was contributed by Denby.

Biostatistics Advance Access published May 5, 2011

Biostatistics (2011), **0**, 0, *pp.* 1–13
doi:10.1093/biostatistics/kxr009

CD

Inferring the time-invariant topology of a nonlinear sparse gene regulatory network using fully Bayesian spline autoregression

EDWARD R. MORRISSEY

Systems Biology Centre, University of Warwick, Coventry House, CV4 7AL, Coventry, UK

MIGUEL A. JUÁREZ*

School of Mathematics and Statistics, University of Sheffield, Hicks Building,
S3 7RH, Sheffield, UK
m.juarez@sheffield.ac.uk

KATHERINE J. DENBY

Warwick Life Sciences and Systems Biology Centre, University of Warwick,
Wellesbourne, CV35 9EF, Coventry, UK

NIGEL J. BURROUGHS

Systems Biology Centre, University of Warwick, Coventry House, CV4 7AL, Coventry, UK

Downloaded from biostatistics.oxfordjournals.org at University of Warwick on May 17, 2011

SUMMARY

We propose a semiparametric Bayesian model, based on penalized splines, for the recovery of the time-invariant topology of a causal interaction network from longitudinal data. Our motivation is inference of gene regulatory networks from low-resolution microarray time series, where existence of nonlinear interactions is well known. Parenthood relations are mapped by augmenting the model with kinship indicators and providing these with either an overall or gene-wise hierarchical structure. Appropriate specification of the prior is crucial to control the flexibility of the splines, especially under circumstances of scarce data; thus, we provide an informative, proper prior. Substantive improvement in network inference over a linear model is demonstrated using synthetic data drawn from ordinary differential equation models and gene expression from an experimental data set of the *Arabidopsis thaliana* circadian rhythm.

Keywords: Circadian clock; Gibbs variable selection; Markov process prior; Nonlinear gene regulatory networks; P-Splines regression; time course gene expression data.

1. INTRODUCTION

Our objective is the inference of gene regulatory networks (GRNs) from time series data; specifically, inferring the gene regulatory kinships for a particular process. To this end, we can conceptualize a GRN

*To whom correspondence should be addressed.

© The Author 2011. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

as a directed graph, with its nodes representing genes and the edges gene–gene regulation. Bayesian networks (BNs) have been used previously in gene network determination (Friedman *and others*, 2000). However, it is well known that biological processes have feedback loops and thus the validity of BNs is questionable when modeling such systems. Dynamic Bayesian networks (DBNs) have been proposed for modeling time course (longitudinal) gene expression data (Zou and Conzen, 2005). These can be thought of as "unfolding" a BN for every time point, and when folding back the network, selfregulation and cliques may be obtained. Formally, a DBN is characterized by a set of conditional relations, $p(y^{t+1}|\boldsymbol{y}^t)$. In the case of an (auto)regression-based DBN, these relations can be written as $y_i^{t+1} = f_i(\boldsymbol{y}^t) + \varepsilon_i^{t+1}$, where y_i^t is the expression measurement of gene $i = 1, \ldots, G$ at time $t = 1, \ldots, T$, $\boldsymbol{y}^t = y_1^t, y_2^t, \ldots, y_G^t$ and ε_i^t is an idiosyncratic error term. The functional forms of the interactions, $f_i(\cdot)$, are usually unknown and typically nonlinear due to the complex biochemistry behind gene regulation. Whether or not $\partial f_i(\boldsymbol{y}^t)/\partial y_j^t \equiv 0$ defines the topology of the network. The interaction topology is key in GRN, as it determines the causal relations in the gene regulatory dynamics for a given biological process. Although gene regulatory relationships can change in time, especially when dealing with varying experimental conditions (Ahmed and Xing, 2009), we assume that the data have been produced in controlled conditions and thus regulation can be suitably captured with a time-invariant network topology.

A flexible way of including unknown nonlinearities, and thus avoiding model selection issues, is to use a semiparametric specification by letting the interactions be described by spline functions. The use of splines in the estimation of GRNs has been advanced by *i.a.* Gustafsson *and others* (2005) and Kim *and others* (2004). A fundamental problem when using spline regression is knot selection that greatly influences the curve fitting. One efficient solution is to select a few well-placed knots for a given spline degree. This requires determining both the optimal number and the position of the knots, which is typically addressed by means of a transdimensional Monte Carlo Markov chain (MCMC) scheme (Ferreira *and others*, 2008; Denison *and others*, 2002) or by cross validation (Ruppert, 2002; Friedman, 1991). The efficiency gained in the modeling may be offset by mixing problems in the sampler, due mainly to the vast space that must be explored and the associated computational problems, or by the unwieldy amount of comparisons required for cross validation.

Our approach avoids such issues by relying on P-splines (Eilers and Marx, 1996; Lang and Brezger, 2004), which are characterized by specifying a rather large number of evenly spaced knots. Then, in order to avoid overfitting and also to control for the effective number of parameters to be estimated, a penalty that shrinks the spline coefficients toward the origin is specified. Such a penalty depends crucially on a so-called smoothness parameter. In this paper, we propose a fully Bayesian setup for dealing with this smoothness parameter and discuss the implications of alternative prior specifications for this key model component.

The proposed model is presented in Section 2, where we also discuss the prior specification. Implementation is briefly described in Section 3. Section 4 illustrates the application of our model to 3 examples, where we reconstruct the corresponding networks and assess their accuracy. Conclusions and possible extensions are given in Section 5. Data sets and Matlab code used in the paper are available in the supplementary material available at *Biostatistics* online and in http://majuarez.staff.shef.ac.uk/materials/index.html.

Downloaded from biostatistics.oxfordjournals.org at University of Warwick on May 17, 2011

2. THE MODEL

Let y_g^t denote the gene expression level of gene $g = 1, \ldots, G$, measured at time $t = 1, \ldots, T$. We propose to model it as $y_g^t = \eta_g^t + \varepsilon_g^t$, where η_g^t is the predictor and ε_g^t is an idiosyncratic error term, centered at zero. We assume that η_g^t is determined by some unknown subset of the genes at the previous time point, and that the error terms are Gaussian and independent for all genes and time points. Thus, we can write it as

$$y_g^t = \eta_g^t(\boldsymbol{y}^{t-1}; \boldsymbol{\theta}_g) + \varepsilon_g^t, \quad \varepsilon_g^t \sim N(\varepsilon_g^t|0, \lambda_g) \quad \text{ind.}, \tag{2.1}$$

where $\boldsymbol{y}^t = \{y_1^t, \ldots, y_G^t\}$, $\boldsymbol{\theta}_g$ is a set of parameters indexing $\eta_g^t(\cdot;\cdot)$ and $\lambda_g^{-1} = \mathrm{Var}(\varepsilon_g^t)$.

In order to accommodate nonlinearities, the regulatory relationships are modeled by

$$\eta_g^t = f_{g1}(y_1^{t-1}) + f_{g2}(y_2^{t-1}) + \ldots + f_{gG}(y_G^{t-1}) + \mu_g, \tag{2.2}$$

where μ_g is a gene-specific constant term and $f_{gi}(y_i) = \sum_{k=1}^{M} \beta_{ik}^g B_{ik}(y_i)$. Here, $\{B_{ik}(y_i)\}$ are M B-spline basis functions of degree l defined over the set of r evenly spaced knots, $\boldsymbol{\kappa}_i = \{\kappa_{i1}, \ldots, \kappa_{ir}\}$, with $\min\{y_i\} = \kappa_{i1} < \kappa_{i2} < \ldots < \kappa_{ir} = \max\{y_i\}$, and $M = r + l$. By defining the spline design row vectors $X_j^t \in \mathbb{R}^M$, such that $X_j^t(k) = B_{jk}(y_j^t)$, we can rewrite the predictor in (2.1) as $\eta_g^t = X_1^{t-1}\boldsymbol{\beta}_{1g} + \ldots + X_G^{t-1}\boldsymbol{\beta}_{Gg} + \mu_g$, with $\boldsymbol{\beta}_{jg} = \{\beta_{j1}^g, \ldots, \beta_{jM}^g\} \in \mathbb{R}^M$ a column vector of coefficients for $j = 1, \ldots, G$. If $\|\boldsymbol{\beta}_{jg}\| \approx 0$, there is negligible influence of gene j on gene g, and thus the "link" $j \rightarrow g$ is off. If the link is on, then we say that j is a "parent" of g.

Stacking the bases and the coefficients into $X^t = \{X_1^t, \ldots, X_G^t\} \in \mathbb{R}^{MG}$ and $\boldsymbol{\beta}_g = \{\boldsymbol{\beta}_{1g}, \ldots, \boldsymbol{\beta}_{Gg}\} \in \mathbb{R}^{MG}$, respectively, and after further stacking the equations over time, we have

$$\boldsymbol{y}_g = \boldsymbol{\mu}_g + \mathcal{X}\boldsymbol{\beta}_g + \boldsymbol{\varepsilon}_g, \quad g = 1, \ldots, G, \tag{2.3}$$

where $\boldsymbol{\mu}_g = \mu_g \boldsymbol{\imath}_T'$, with $\boldsymbol{\imath}_T$ a row vector of ones of size T and $\mathcal{X} = \{X^1, X^2, \ldots, X^T\}'$ a bases matrix of size $T \times MG$. This model is unidentifiable given that every potential parent spline contributes with its own constant term. To correct for this, we add the identifiability restriction $\boldsymbol{\imath}_T \times (\mathcal{X}\boldsymbol{\beta}_g) = 0$. We describe its implementation within the sampling scheme in the supplementary material available at *Biostatistics* online.

As it stands to estimate the $2 + M \times G$ parameters of each spline-regression component in (2.3) would require in excess of this number of data points per gene. If the number of time measurements is relatively small, one would need to select a rather small number of knots, thus effectively reducing the capacity of the splines to capture nonlinearities. We address this issue by performing a Gibbs variable selection as in Smith and Kohn (1996). The model is augmented with the indicators γ_{jg}, such that $\tilde{\boldsymbol{\beta}}_{jg} = \gamma_{jg} \times \boldsymbol{\beta}_{jg}$, where $\gamma_{jg} = 1$ if the link is on and $\gamma_{ij} = 0$ if the link is off and substituting these new coefficients into the model.

The practical advantage of augmenting with the indicators is that it allows us to make inference about the network topology, now parameterized by the connectivity matrix, $\Gamma = \{\gamma_{jg}\}$.

2.1 *The prior*

We use conditionally conjugate priors where suitable, which simplifies the sampling algorithm. We take particular care when specifying a shrinkage or penalty prior for the spline coefficients, as this determines the smoothness of the functional form fitted.

Precisions. We use conjugate, i.i.d. gamma priors, $\mathrm{Ga}(\lambda_g|a_\lambda, b_\lambda)$, on the gene precisions, $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_G\}$,

$$\pi(\boldsymbol{\lambda}) = \prod_{g=1}^{G} \frac{b_\lambda^{a_\lambda}}{\Gamma[a_\lambda]} \lambda_g^{a_\lambda - 1} \exp[-b_\lambda \lambda_g]. \tag{2.4}$$

Downloaded from biostatistics.oxfordjournals.org at University of Warwick on May 17, 2011

Constant term. An independent Gaussian prior, $N(\boldsymbol{\mu}|\mathbf{0}, \tau_\mu I)$, for the gene-specific constant, $\boldsymbol{\mu} = \{\mu_1, \dots, \mu_G\}$

$$\pi(\boldsymbol{\mu}) = \left(\frac{\tau_\mu}{2\pi}\right)^{G/2} \exp\left[-\frac{\tau_\mu}{2}\boldsymbol{\mu}'\boldsymbol{\mu}\right]. \tag{2.5}$$

Network structure. We provide 2 alternatives for modeling the network topology. The first is to define the "overall network connectivity," ρ, as $P[\gamma_{jg} = 1] = \rho$ and complement it with a Beta prior, $\text{Be}(\rho|a_\rho, b_\rho)$. The full specification is then,

$$\pi(\gamma_{jg}|\rho) = \rho^{\gamma_{jg}}(1-\rho)^{1-\gamma_{jg}}, \quad g, j = 1, \dots, G, \tag{2.6}$$

$$\pi(\rho) = [\text{B}(a_\rho, b_\rho)]^{-1}\rho^{a_\rho - 1}(1-\rho)^{b_\rho - 1} \quad 0 < \rho < 1. \tag{2.7}$$

It is well known that GRNs often present hub-like structures where a handful of genes control the regulation process almost completely and the rest of the genes have very few children, if any (see, e.g. Seo *and others*, 2009, and references therein). One can capture such features by allowing for "parent-wise connectivity," $P[\gamma_{jg} = 1] = \rho_j$ and complementing it with independent priors, that is,

$$\pi(\gamma_{jg}|\rho_j) = \rho_j^{\gamma_{jg}}(1-\rho_j)^{1-\gamma_{jg}}, \quad g = 1, \dots, G, \tag{2.8}$$

$$\pi(\rho_j) = [\text{B}(a_\rho, b_\rho)]^{-1}\rho_j^{a_\rho - 1}(1-\rho_j)^{b_\rho - 1} \quad j = 1, \dots, G. \tag{2.9}$$

The hyperparameters $\{a_\rho, b_\rho\}$, convey our prior knowledge about the connectivity of the network and can be set accordingly. For general purposes, we recommend setting both equal to $1/2$, as this is the reference prior for a Bernoulli experiment (Bernardo and Smith, 1994). If biological knowledge of the process demands it, it is straightforward to fix any link to be deterministically on (off) by setting $\gamma_{rl} = 1(0)$, modifying the prior accordingly.

Spline coefficients. We use a second-order Markov process prior on the coefficients $\boldsymbol{\beta}_{jg}$ to shrink them toward the origin.

$$\pi(\boldsymbol{\beta}_{jg}|\tau_{jg}) = N(\boldsymbol{\beta}_{jg}|\mathbf{0}, \tau_{jg}K), \tag{2.10}$$

where τ_{jg} are the smoothness parameters addressed below. The structure of the covariance matrix, $K = \{K_{kl}\}$, is constructed from the second-order differences between adjacent coefficients, that is, $\beta_k = 2\beta_{k-1} - \beta_{k-2}$, omitting link identifiers for simplicity (see supplementary material available at *Biostatistics* online). The prior for the 2 remaining coefficients, $\{\beta_1, \beta_2\}$, is discussed below.

Smoothness parameters. In the case of small data sets, the specification of the smoothness parameters, τ_{jg}, becomes crucial as these largely determine the fitting of the spline to the data. In the limit, when $\tau_{jg} \to 0$, an interpolating spline is fitted, while as $\tau_{jg} \to \infty$ a straight line is rendered.

A conditionally conjugate prior is the product of independent gamma distributions, $\text{Ga}\cdot|a_\tau, b_\tau)$. This specification concentrates mass around a_τ/b_τ and has a relatively large right tail for small values of b_τ. It is common to find in the literature $a_\tau = b_\tau$ and set to quite small values, for example, 0.001. This indeed is quite flat over a large range of τ, but has a mode at zero effectively giving relative importance to rougher curves and thus favoring overfitting when the data are only weakly informative. On the other hand, if mass is carried toward larger values of τ—thus favoring smoother curves—the gamma distribution tails off quite quickly to the left and experiences difficulties capturing nonlinearities, (see, e.g. Jullion and Lambert, 2007).

In order to obtain a more flexible prior specification, while retaining the conditional conjugacy, we also tried a gamma scale mixture of gammas. The resulting gamma–gamma distribution (Bernardo and Smith,

Downloaded from biostatistics.oxfordjournals.org at University of Warwick on May 17, 2011

1994, p. 120; Zellner, 1971, p. 376), can achieve a larger spread than the gamma and also has a heavier right tail. It may also not have any finite moments for certain parameter values. Despite these desirable characteristics, we found that the heavy right tail of this prior, combined with the flatness of the likelihood in regions where τ is very large can lead to identifiability issues. This can be understood since there exists a threshold value, $\tau^\star$, for which the fit of the spline is practically linear and thus indistinguishable for any $\tau > \tau^\star$.

This lead us to propose an inverted Pareto prior, $\mathrm{Ip}(\cdot|a_\tau, b_\tau)$:

$$\pi(\tau_{jg}|a_\tau, b_\tau) = \frac{a_\tau}{b_\tau}\left(\frac{\tau_{jg}}{b_\tau}\right)^{a_\tau - 1}, \quad \tau_{jg} \leqslant b_\tau, a_\tau > 0. \tag{2.11}$$

We restrict $a_\tau \geqslant 1$, to prevent concentration of mass near the origin. Setting $a_\tau = 1$ is tantamount to putting a uniform prior on $(0, b_\tau)$. The prior is concave for $1 < a_\tau \leqslant 2$ and convex for $a_\tau \geqslant 2$, gathering mass closer to b_τ as a_τ grows, thus favoring smoother curves. Values of $a_\tau > 3$ allocate too much mass close to b_τ and thus are not advisable, unless there is prior evidence for high levels of linearity. The cutoff value b_τ can be interpreted as that level of τ after which the likelihood is numerically invariant, that is, the fitted curve is practically linear.

2.2 *Posterior propriety*

In most of our intended applications, we will have a limited number of time measurements compared to the number of genes. Given that an improper prior will yield an improper posterior if the number of parents for any given gene exceeds T/M (see the supplementary material available at *Biostatistics* online), we construct a proper prior by supplying (2.10) with an independent specification for the first 2 coefficients,

$$\pi(\beta_1, \beta_2) = N(\beta_1|0, k_1)N(\beta_2|0, k_2). \tag{2.12}$$

To approximate the behavior of the improper prior, we could let $k_1, k_2 \to 0$. In situations where the data are scarce, we do not recommend this, as it will affect the stability of the posterior (Sun and Speckman, 2008). In our applications, we set $k_1 = k_2 = \tau_0$.

3. Implementation

3.1 *P-splines model algorithm*

As there is no closed-form expression for the posterior numerical methods are needed. We propose a Metropolis-within-Gibbs scheme that leads to a dramatic decrease in autocorrelation of the chain, compared to a Gibbs move. Details are given in the supplementary material available at *Biostatistics* online.

3.2 *A linear model*

In order to compare the network retrieval power of the splines model, we constructed a fully parametric, linear AR(1) model

$$y_g^{t+1} = \mu_g + \sum_{j=1}^{G} \beta_{jg} y_j^t + \varepsilon_g^t, \tag{3.1}$$

with the same prior specification as above, deleting the irrelevant terms.

Downloaded from biostatistics.oxfordjournals.org at University of Warwick on May 17, 2011

4. ILLUSTRATIONS AND APPLICATIONS

First, we analyze 2 synthetic, discrete time data sets where the data generation mechanism and the topology of the network are known. Second, we examine a synthetic data set comprising discrete time measurements drawn from a continuous time ordinary differential equation (ODE) model of a circadian clock. For our last example, we use microarray gene expression data from the *Arabidopsis thaliana* circadian clock. Details on the prior parameters specification are given in the supplementary material available at *Biostatistics* online.

4.1 *Discrete time synthetic networks*

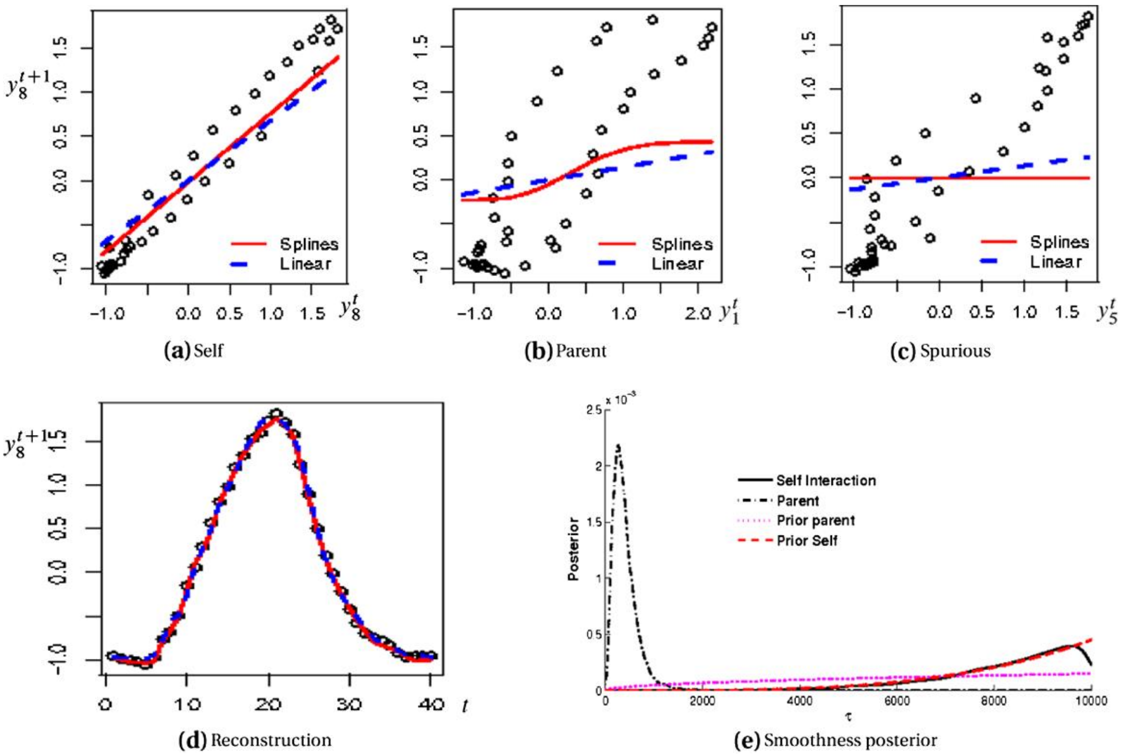
In order to assess the network topology recovery power of our model, we produced 2 synthetic, first-order autoregressive processes. One has only linear and the second a number of nonlinear relations. In the nonlinear case, all the functional relations were produced using Hill functions, except for the self-interactions that are linear. In both cases, we set $G = 16$, $T = 40$, and $\rho \approx 0.1$.

When the topology of the network is known, we can use the receiver operating characteristic (ROC) curve to assess graphically the retrieval performance of a model. A more formal comparison can be carried out by calculating the area under the ROC curve (AUC) and the mean cross entropy (MxE). For the linear data set, the AUC were 0.999 for the fully parametric model and 0.998 with the splines; and when fitting the nonlinear data set, we obtained 0.728 and 0.912, respectively. In the linear network, the MxE was 0.042 when fitting the parametric model and 0.064 when fitting the splines; with Hill interactions the values were 0.41 and 0.22, respectively. Thus, using these scores network topology retrieval from the splines model is almost as good as that from the linear when the interactions are linear and outperforms it when nonlinearities are present (ROC curves are shown in the supplementary material available at *Biostatistics* online).

To further understand the differences between the inferred networks under either model, we plot in Figure 1, the partial and full reconstructions for gene 8's trace in the nonlinear data set, along with the posterior of the corresponding smoothing parameter. Both models provide similar predictions, as illustrated by the full reconstructions that are practically undistinguishable (Figure 1(d)). However, the way this fit is achieved varies significantly. As expected, both models have a very similar fit for the self-regulation (Figure 1(a)). As the self-interaction is linear, the splines model fits it by allocating most of the posterior mass of the corresponding smoothness parameter toward high values, depicted by the solid line in Figure 1(e). Gene 8 has one parent with a nonlinear interaction and the splines model is capable of reproducing the Hill functional relationship quite precisely (Figure 1(b)), by allocating almost all posterior mass toward small values of the corresponding smoothness parameter, shown in Figure 1(e) (dot-dashed line). Obviously, the linear model cannot accommodate such behavior and may include spurious parents in order to compensate for the lack of fit, as in this case, illustrated in Figure 1(c). In contrast, the splines model does not predict Gene 5 as a parent (solid line in Figure 1(c)). Notice the mass allocation of the self-regulation link (solid) in Figure 1(e): it is basically drawn from the prior (dashed), illustrating that our specification is adequate for linear relations to be reproduced accurately.

When the network topology—that is, the biological model—is fixed, we can compare the fit of alternative statistical models using formal tools. We calculated the deviance information criterion (DIC) obtained from the different data/model combinations used in this paper by fixing the network topology to those links with posterior probability larger than 0.8 (Table 1). In the linear data case, both models produced similar estimates of the connectivity matrix and therefore their AUC and MxE scores are quite close to each other. However, the DIC indicates that the linear model is preferred to the splines, mainly due to the costs associated with the additional complexity of the splines model, unnecessary for this data set. In contrast, the DIC from the nonlinear data set favors the splines model, granting the increase in model complexity.

Downloaded from biostatistics.oxfordjournals.org at University of Warwick on May 17, 2011



(a) Self **(b)** Parent **(c)** Spurious

(d) Reconstruction **(e)** Smoothness posterior

Fig. 1. Partial and full reconstructions of Gene 8's trace using splines (solid) and linear (dashed) models. The circles represent a scatter plot of the expression values of Gene 8 against 3 potential parents (Genes 8, 1, and 5) at the previous time point. (a) Both models capture the linear self-regulation. (b) The true parent is predicted in both models, while splines is able to reproduce the Hill functional relationship. (c) The linear model predicts one spurious parent. (d) Trace reconstruction from both models is almost identical. (e) Marginal posterior distributions of the smoothing parameter for the self-regulation (solid, prior dashed) and nonlinear parent (dot-dashed, prior dotted)

Table 1. *Comparing model fit with a fixed network topology. DIC values obtained from the data sets used in the paper when fitting the linear and the splines models. The network topology is fixed by selecting those links with posterior probability above* 0.8. *The model preferred by DIC is highlighted in bold font*

Data set	Linear model	Splines model
Synthetic linear	$\mathbf{2.02 \times 10^4}$	2.56×10^4
Synthetic Hill	3.01×10^4	$\mathbf{3.64 \times 10^3}$
ODE data	2.44×10^5	$\mathbf{2.34 \times 10^5}$
Microarray data	9.95×10^3	$\mathbf{1.93 \times 10^3}$

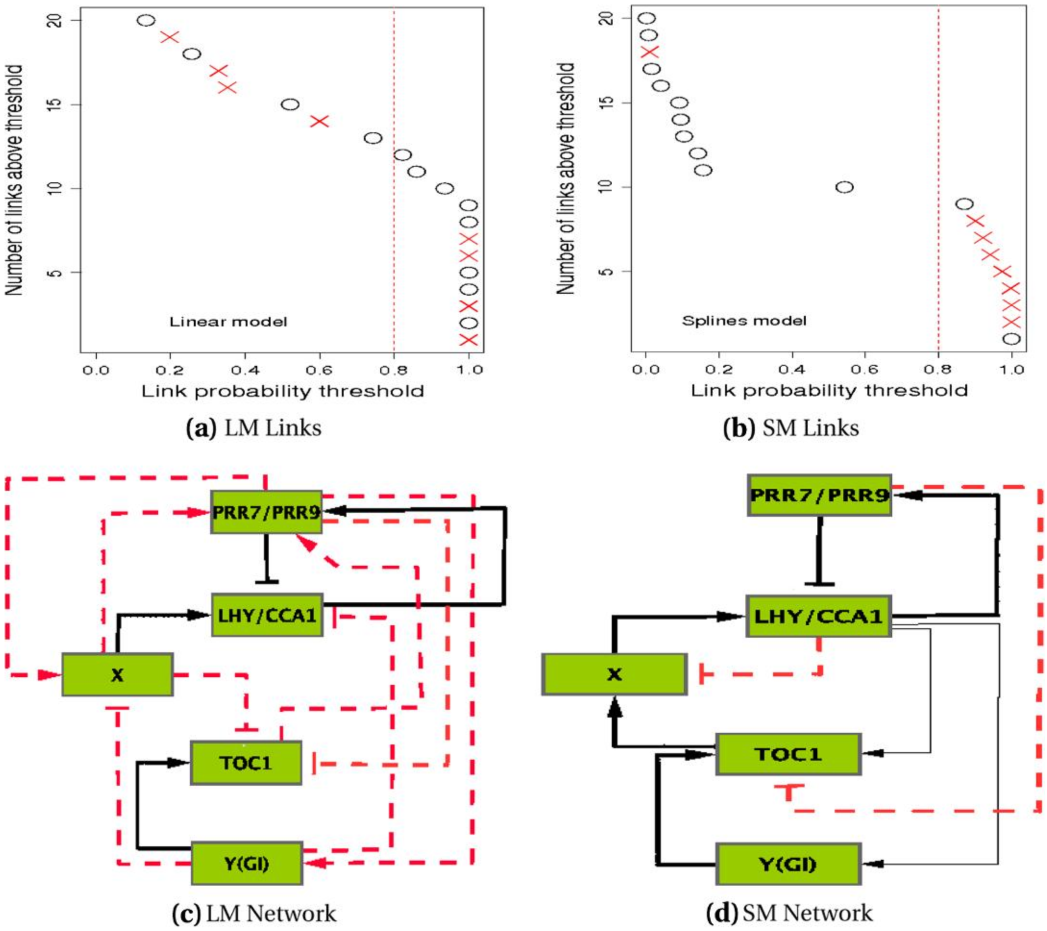
4.2 *Biological GRN: the plant Circadian clock*

In the following sections, we focus on a partially known GRN, specifically the plant *A. thaliana* circadian clock. Locke *and others* (2006) developed an ODE model of the clock, which we use below for generating synthetic observations. The current working biological model is due to McClung (2008). Both models include nodes *X* and *Y*, representing genes that are thought to be involved in the circadian clock, but

Downloaded from biostatistics.oxfordjournals.org at University of Warwick on May 17, 2011

whose identity remains unknown. These network models are shown schematically and further explained in the supplementary material available at *Biostatistics* online.

Differential equation data. We generated data from the ODE model fixing the light source to be permanently on. The data were then subsampled, logged, and standardized. The resulting data set has 50 time points with a time spacing of 1 h. We present the results obtained using the parent-wise connectivity structure (2.8)–(2.9). In order to interpret the output, rather than examining the ROC curves, we analyze the inferred network at a given threshold. This is more convenient given that there are only a few genes and therefore a more detailed comparison with the true network is possible. We plot the number of links included in the predicted network against the posterior link probability when fitting the linear model, Figure 2(a), and when using the splines model, Figure 2(b). We use a cross (circle) for a correctly



(a) LM Links (b) SM Links

(c) LM Network (d) SM Network

Fig. 2. Network topology inference on the ODE circadian clock data. (a) The number of links predicted to be present in the network versus posterior link probabilities estimated when using the linear model (LM) and (b) with the splines model (SM). Crosses (circles) represent correctly (incorrectly) predicted links. (c) The network obtained with a threshold of 0.8 using the LM and (d) when using the SM. Solid lines represent correct predictions, dashed lines incorrect predictions, and thin lines correct predictions, but with the wrong sign.

Downloaded from biostatistics.oxfordjournals.org at University of Warwick on May 17, 2011

(incorrectly) predicted link; for instance, the predicted network with the splines model using a threshold of 0.85 would have 9 links (circles and crosses with link probability above 0.85 in Figure 2(a)), 7 out of these correct. It is apparent that the splines model produces a better separation in the link probabilities, classifying all but one link into 2 populations: a low probability (below 0.2) and a high probability (above 0.8) group. This contrasts with the linear model Figure 2(a) where almost 40% of the links are in the ambiguous region between 0.2 and 0.8.

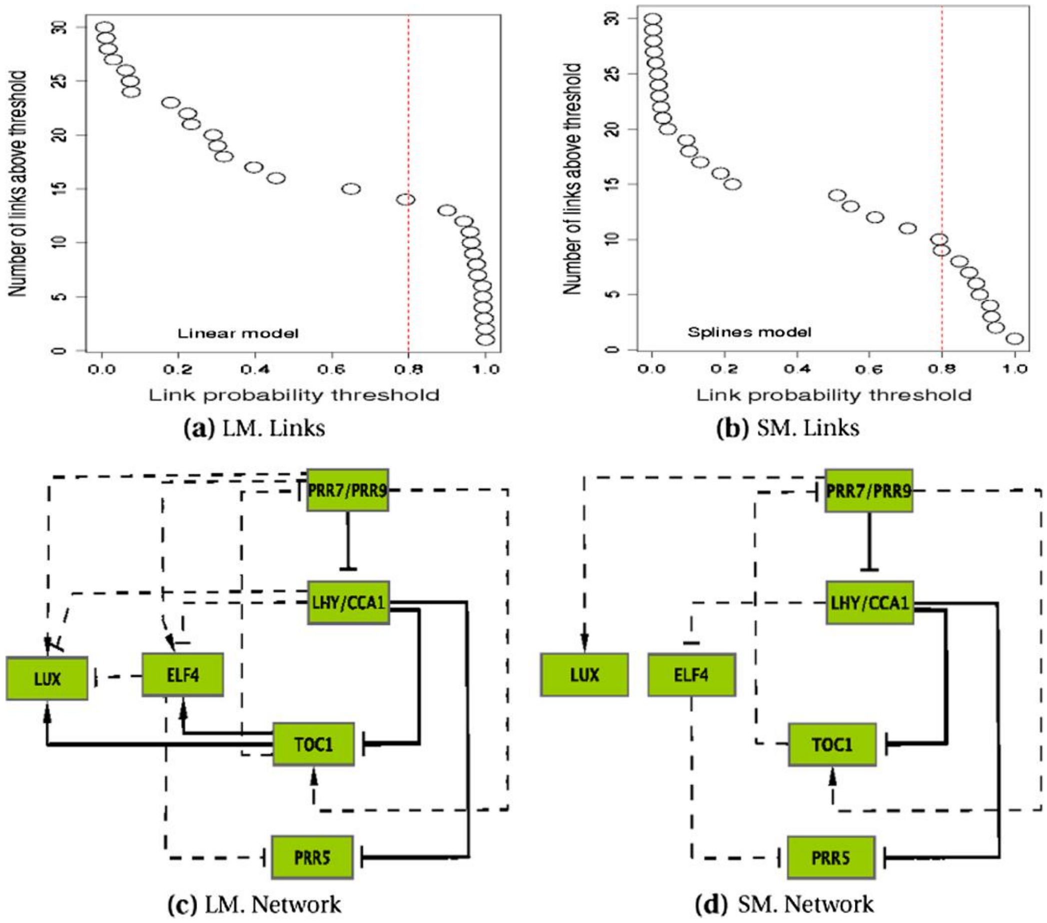
Using 0.8 as the threshold value, we plot the reconstructed networks for both models in Figure 2(c) and (d). The inferred network for splines (Figure 2(d)) contains all correct links from the network, except for the TOC1–Y link. There are 2 spurious links (LHY/CCA1–X and PRR7/PRR9–TOC1) and 2 links with incorrect signs (LHY/CCA1–TOC1 and LHY/CCA1–Y). In addition to only having half of the correct links, the inferred network for the linear model adds a large number of spurious links (false positives). Again, we found cases where the splines model correctly predicts a single parent using a nonlinear interaction, whereas the linear model predicts that link but adds extra spurious links to improve the fit (not shown). Moreover, the DIC of the estimated network topology from the splines model is smaller than that from the linear model at threshold 0.8 (see third row of Table 1).

Experimental data. We use our methods on gene expression time series for *Arabidopsis* leaves generated using microarrays and analyze the output using the parent-wise connectivity structure. The separation of posterior link probabilities into groups is no longer as pronounced as in the synthetic data—see Figure 3(a) and (b). This may be due to the combination of a high level of noise and fewer time points. The networks inferred by each model, using a threshold of 0.8, are shown in Figure 3(c) and (d). All links predicted by the splines model appear in the linear model reconstruction. However, the linear model predicts an additional 2 parents for ELF4 and an additional 3 parents for LUX. Among those additional links are TOC1–ELF4 and TOC1–LUX, which while we have marked as correct on the plot (for consistency with the current working model), are probably incorrect. Those links were included in the accepted model as an indication that TOC1 regulates some gene (X) that in turn regulates LHY, but neither of the genes are predicted to regulate LHY. Furthermore, from the previous examples, it is clear that the linear model tends to add spurious parents.

Although only mild nonlinearities are found, the DIC score indicates a better fit from the splines model than the linear alternative for the given threshold (last row of Table 1), suggesting that even mild departures from linearity can have an important effect in model fit.

For network reconstruction, we have used a subjective choice for the posterior probability threshold. Moreover, as our reconstruction is based solely on the individual link marginal probabilities, possible correlations between these are disregarded. In order to provide a graphical representation of the uncertainty in our network retrieval, in Figure 4, we plot a heatmap with the distribution of the number of parents for each gene in the clock (left), together with a heatmap with the marginal link probabilities of its top 4 potential parents (right). These complementary sources of information render a picture of the uncertainty in the retrieval of the network topology. For instance, the splines model predicts one parent for LHY with very high probability and there is only one potential parent with high marginal probability, suggesting a very confident prediction; in contrast, the linear model predicts 1 or 2 parents with a mild probability (and 3 with a very slight probability), while one of the potential parents has a high marginal probability, there are 2 more with intermediate marginal probabilities, suggesting an ambiguity in the identity of the second potential parent. Overall, there is a shift to the left of the distribution for the number of parents from the splines model compared to the linear model, strengthening the evidence for overfitting in the latter model. Likewise, marginal link probabilities for the splines model seem to be higher over a smaller number of potential parents, thus suggesting a decrease in the uncertainty in topology retrieval compared to the linear model.

Downloaded from biostatistics.oxfordjournals.org at University of Warwick on May 17, 2011



(a) LM. Links **(b)** SM. Links

(c) LM. Network **(d)** SM. Network

Fig. 3. Network topology inference on the circadian clock microarray data. (a) The number of links predicted to be present in the network versus posterior link probabilities estimated when using the LM and (b) with the SM. (c) The network obtained with a threshold of 0.8 using the LM and (d) when using the SM. Solid lines represent inferred links that are included in the currently accepted model (most of which have been experimentally confirmed) and dashed lines inferred links absent in the accepted model (though not necessarily incorrect).

5. DISCUSSION

We have presented a fully Bayesian implementation of P-spline based inference of a DBN within a sparse connectivity context. Our motivation is the inference of GRNs from longitudinal data, for instance, from microarray time series data. Despite being capable of measuring up to tens of thousands of genes simultaneously, currently available microarray time series are typically shorter than 20 time points. This introduces significant problems for analysis and modeling, particularly as it limits the complexity of the models that can be used. We addressed this issue through use of spike-and-slab type priors that limit the connectivity of the GRN. Within this context, we are able to increase regression model complexity, designing a method for exploring whether nonlinear regulatory mechanisms are present in time series data.

Downloaded from biostatistics.oxfordjournals.org at University of Warwick on May 17, 2011

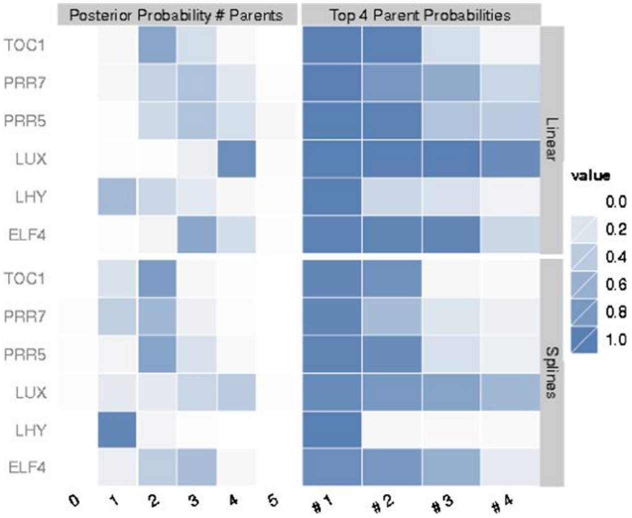


Fig. 4. Uncertainty in *Arabidopsis* circadian clock gene network reconstruction. On the left, a heatmap of the distribution of the number of parents for each gene in the clock, estimated using microarray data with the linear (top) and splines (bottom) models. On the right, a heatmap with the marginal probabilities of the top 4 potential parents.

Downloaded from biostatistics.oxfordjournals.org at University of Warwick on May 17, 2011

Our model successfully identified nonlinear interactions on simulated data (both discrete time and ODE models), while the corresponding DIC scores favored the estimated network topologies with the splines model, for a given threshold, over the linear alternative. On simulated data with nonlinear interactions, the inferred GRN under a linear model typically acquired additional parents, these incorrectly predicted parents improved the fit to a similar quality to that achieved by the P-splines model. The P-splines model also enhances network sparsity since an additional parent under a splines regression model incurs a greater penalty than a parent with a linear functional dependence model given the higher (model) complexity; thus even when the links are actually linear there is stronger control on the number of parents. This compares to artificially imposed parent number penalization, for instance, through an arbitrary weighting $\exp(-n)$ for n parents, as in Kim *and others* (2004).

Assessing the uncertainty in network topology retrieval is an active field of research. We provide a graphical representation that combines the information on the distribution on the number of parents for each gene, with the marginal posterior probabilities of the most probable regulators. In our example using microarray data from *A. thaliana* leaves, joint inspection of these estimates suggests that the splines model provides a more accurate network reconstruction compared to the linear model.

Use of splines in inference requires handling of their functional flexibility. We recommend that the number of knots is much smaller than the number of time points; here, we presented results using 10 knots for a time series with 40–50 time points. We found that doubling the number of knots (20) gave severe problems in the mixing of the chain, while using a smaller number (7) gave similar results. We also use a prior on the coefficients that effectively controls the spline curvature. This entails choice of the value of the smoothing parameter τ; previous authors have optimized and fixed it before estimating the regression. We propose a fully Bayesian approach, inferring it concomitantly with the regression and performed a sensitivity analysis to confirm our prior is sufficiently weak, further confirming that linear relations can be retrieved. Network connectivity and spline smoothness were regression/gene specific; this allowed for both heterogeneity in the nonlinearity and the number of parents across the network. We presented the results for parent-wise connectivity and the proposed Beta prior parameters that we expect to be appropriate for data sets similar as those used in this paper. Moreover, performing a sensitivity analysis by modifying these values in the region $(1/2, 2)$, and also restricting to an overall connectivity did not affect the results significantly in our examples. However, we have found that when the number of genes increases significantly with respect to the number of time points there might not be enough information for using the parent-wise prior and we suggest using an overall/global connectivity model.

Our P-splines model can be extended and modified for specific purposes. First, we model only direct, first-order filiation. One can extend the present model for allowing higher degree interactions, for example, by using tensor product splines. The main hindrance would then be the combinatorial growth of the topology space, and efficient methods for exploring it must be devised. Second, spline coefficient shrinkage can be performed in a number of ways. Additional constraints can be used, including a further term on the prior for the spline coefficients, $N(\boldsymbol{\beta}|\mathbf{0}, \omega H)$, with H derived from the first-order differences of adjacent coefficients. This effectively penalizes large first-order differences and favors less jagged curves, depending on the value of $\omega > 0$. Additionally, the shape of the functional form the spline may take can also be further restricted. For instance, many gene regulatory effects are monotonic. Extending the model to include monotonicity restrictions is feasible by providing such information through a prior (Ansley *and others*, 1993). Finally, the splines model can be utilized to infer the functional form of the regulation, and coupled with current biological knowledge, serve as a basis of a tailor-made parametric model.

Downloaded from biostatistics.oxfordjournals.org at University of Warwick on May 17, 2011

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the assistance of Kathryn Richardson. *Conflict of Interest:* None declared.

FUNDING

Engineering and Physical Sciences Research Council through the WSB-DTC (EP/d508150/1 to E.R.M.); M.A.J. carried out the research while working in WSBC and was funded by Biotechnology and Biological Sciences Research Council through the SysMO initiative (BB/ff003498/1 to M.A.J); The experimental data were provided by K.J.D. through the PRESTA project (BB/f005806/1 to K.J.D.).

REFERENCES

AHMED, A. AND XING, E. P. (2009). Recovering time-varying network dependencies in social and biological studies. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 11878–11883.

ANSLEY, C. F., KOHN, R. AND WONG, C. M. (1993). Nonparametric spline regression with prior information. *Biometrika* **80**, 75–88.

BERNARDO, J. M. AND SMITH, A. F. M. (1994). *Bayesian Theory*. Chichester: John Wiley & Sons.

DENISON, D. G. T., HOLMES, C. C., MALLICK, B. K. AND SMITH, A. F. M. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. Chichester: Wiley.

EILERS, P. H. C. AND MARX, B. D. (1996). Flexible smoothing using B-splines and penalised likelihood (with discussion). *Statistical Science* **11**, 89–121.

FERREIRA, J. T. A. S., JUÁREZ, M. A. AND STEEL, M. F. J. (2008). Directional log-spline distributions. *Bayesian Analysis* **3**, 267–315.

FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines (with discussion). *The Annals of Statistics* **19**, 1–141.

FRIEDMAN, N., LINIAL, M., NACHMAN, I. AND PE'ER, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology* **7**, 601–620.

GUSTAFSSON, M., HÖRQUIST, M. AND LOMBARDI, A. (2005). Constructing and analysing a large-scale gene-to-gene regulatory network—Lasso-constrained inference and biological validation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2**, 254–261.

JULLION, A. AND LAMBERT, P. (2007). Robust specification of the roughness penalty prior distribution in spatially adaptive Bayesian P-splines models. *Computational Statistics & Data Analysis* **51**, 2542–2558.

KIM, S. Y., IMOTO, S. AND MIYANO, S. (2004). Dynamic Bayesian network and nonparametric regression for nonlinear modelling of gene networks from time series gene expression data. *Biosystems* **75**, 57–65.

LANG, S. AND BREZGER, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics* **13**, 183–212.

LOCKE, J. C. W., KOZMA-BOGNR, L., GOULD, P. D., FEHÉR, B., KEVEI, E., NAGY, F., TURNER, M. S., HALL, A. AND MILLAR, A. J. (2006). Experimental validation of a predicted feedback loop in the multi-oscillator clock of *Arabidopsis thaliana*. *Molecular Systems Biology* **2**, 59.

MCCLUNG, C. R. (2008). Comes a time. *Current Opinion in Plant Biology* **11**, 514–520.

RUPPERT, D. (2002). Selecting the number of knots for penalised splines. *Journal of Computational and Graphical Statistics* **11**, 735–757.

SEO, C. H., KIM, J. R., KIM, M. S. AND CHO, K. H. (2009). Hub genes with positive feedbacks function as master switches in developmental gene regulatory networks. *Bioinformatics* **25**, 1898–1904.

SMITH, M. AND KOHN, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics* **75**, 317–343.

SUN, D. AND SPECKMAN, P. (2008). Bayesian hierarchical linear mixed models for additive smoothing splines. *Annals of the Institute of Statistical Mathematics* **60**, 499–517.

ZELLNER, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York: Wiley.

ZOU, M. AND CONZEN, S. D. (2005). A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics* **21**, 71–79.

[*Received August 26, 2010; revised February 16, 2011; accepted for publication February 27, 2011*]

Downloaded from biostatistics.oxfordjournals.org at University of Warwick on May 17, 2011

Biostatistics (2011), **1**, 1, *pp.* 1–13
doi:10.1093/biostatistics/bio000

Supplementary material to Inferring the time-invariant topology of a non-linear sparse gene regulatory network using fully Bayesian spline autoregression

Edward R. Morrissey
Systems Biology Centre
University of Warwick, Coventry House, CV4 7AL, Coventry, UK.

MIGUEL A. JUÁREZ*
School of Mathematics and Statistics
University of Sheffield, Hicks Building, S3 7RH, Sheffield, UK.
m.juarez@sheffield.ac.uk

KATHERINE J. DENBY
Warwick Life Sciences and Systems Biology Centre
University of Warwick, Wellesbourne, CV35 9EF, Coventry, UK.

NIGEL J. BURROUGHS
Systems Biology Centre
University of Warwick, Coventry House, CV4 7AL, Coventry, UK.

1. The model

For completeness, we repeat the model and the prior specification. Stacking the bases and the coefficients into $X^t = \left\{X_1^t, \ldots, X_G^t\right\} \in \mathbb{R}^{MG}$ and $\boldsymbol{\beta}_g = \{\boldsymbol{\beta}_{1g}, \ldots, \boldsymbol{\beta}_{Gg}\} \in \mathbb{R}^{MG}$, respectively, we can express the model as $y_g^{t+1} = \mu_g + X^t \boldsymbol{\beta}_g + \varepsilon_g^t$ and after further stacking the equations over time we have,

$$\boldsymbol{y}_g = \boldsymbol{\mu}_g + \mathcal{X}\,\boldsymbol{\beta}_g + \boldsymbol{\varepsilon}_g\ , \qquad g = 1, \ldots, G\ , \tag{1.1}$$

where $\boldsymbol{\mu}_g = \mu_g \boldsymbol{\iota}_T'$, with $\boldsymbol{\iota}_T$ a row vector of ones of size T and $\mathcal{X} = \{X^1, X^2, \ldots, X^T\}'$ a bases matrix of size $[T \times MG]$.

Precisions. We use conjugate, iid gamma priors, $\mathrm{Ga}(\lambda_g \mid a_\lambda, b_\lambda)$, on the gene preci-

[*] To whom correspondence should be addressed.

© The Author 2011. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oxfordjournals.org.

sions, $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_G\}$,

$$\pi(\boldsymbol{\lambda}) = \prod_{g=1}^{G} \frac{b_\lambda^{a_\lambda}}{\Gamma\,[a_\lambda]}\, \lambda_g^{a_\lambda - 1} \exp\left[-b_\lambda\, \lambda_g\right] \ . \tag{1.2}$$

Constant term. An independent Gaussian prior, $\mathrm{N}(\boldsymbol{\mu} \mid \mathbf{0}, \tau_\mu I)$, for the gene-specific constant, $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_G\}$

$$\pi(\boldsymbol{\mu}) = \left(\frac{\tau_\mu}{2\,\pi}\right)^{G/2} \exp\left[-\frac{\tau_\mu}{2}\ \boldsymbol{\mu}'\,\boldsymbol{\mu}\right] \ . \tag{1.3}$$

Network structure. We provide two alternatives for modelling the network topology. The first is to define the *overall network connectivity*, ρ, as $\mathrm{P}\left[\gamma_{jg} = 1\right] = \rho$ and complement it with a Beta prior, $\mathrm{Be}(\rho \mid a_\rho, b_\rho)$. The full specification is then,

$$\pi(\gamma_{jg} \mid \rho) = \rho^{\gamma_{jg}}\ (1-\rho)^{1-\gamma_{jg}}\ , \quad g, j\ = 1, \ldots, G\ , \tag{1.4}$$

$$\pi(\rho) = \left[\mathrm{B}(a_\rho, b_\rho)\right]^{-1}\ \rho^{a_\rho - 1}(1-\rho)^{b_\rho - 1} \quad 0 < \rho < 1\ . \tag{1.5}$$

Alternatively, we can accommodate *parent-wise connectivity*, $\mathrm{P}\left[\gamma_{jg} = 1\right] = \rho_j$, by letting

$$\pi(\gamma_{jg} \mid \rho_j) = \rho_j^{\gamma_{jg}}\ (1-\rho_j)^{1-\gamma_{jg}}\ , \quad g = 1, \ldots, G\ , \tag{1.6}$$

$$\pi(\rho_j) = \left[\mathrm{B}(a_\rho, b_\rho)\right]^{-1}\ \rho_j^{a_\rho - 1}(1-\rho_j)^{b_\rho - 1} \quad j\ = 1, \ldots, G\ . \tag{1.7}$$

Spline Coefficients. We use the prior on the coefficients $\boldsymbol{\beta}_{jg}$ to shrink them towards the origin specifying a second order Markov process prior

$$\pi(\boldsymbol{\beta}_{jg} \mid \tau_{jg}) = \mathrm{N}(\boldsymbol{\beta}_{jg} \mid \mathbf{0}, \tau_{jg}\ K)\ . \tag{1.8}$$

Where,

$$K_{M,M-2} = 1\,, \qquad K_{M-2,M-1} = -4\,, \qquad K_{M,M} = 1\,,$$
$$K_{M,M-1} = -2\,, \qquad K_{M-1,M-1} = 5\ ;$$

and for all $i, j \in \{3, \ldots M-2\}$,

$$K_{i,j} = \begin{cases} 0 & |i-j| > 2 \\ -4 & |i-j| = 1 \\ 1 & |i-j| = 2 \\ 6 & |i-j| = 0 \end{cases} .$$

Commonly, the two remaining coefficients are given an improper prior, $\pi(\beta_1, \beta_2) \propto 1$.

We discuss conditions for posterior propriety in Section 1.1.
Smoothness parameters. An inverted Pareto prior, $\mathrm{Ip}(\cdot \mid a_\tau, b_\tau)$:

$$\pi(\tau_{jg} \mid a_\tau, b_\tau) = \frac{a_\tau}{b_\tau}\left(\frac{\tau_{jg}}{b_\tau}\right)^{a_\tau - 1}, \qquad \tau_{jg} \leqslant b_\tau\,,\ a_\tau > 0\,. \tag{1.9}$$

1.1 *Posterior propriety*

Fahrmeir and Kneib (2009) discuss conditions for posterior propriety using the covariance structure, K, arising from the second order Markov process prior and different alternatives for the smoothing parameters within the context of structured additive models. We provide a result justifying the extension of this prior for GRN inference used in the paper. The proof is standard and therefore omitted.

Theorem 1 Consider the longitudinal data set $Y = \left\{y_g^t\right\}$, consisting of $g = 1, \dots, G$ genes measured at times $t = 1, \dots, T$, modelled as (1.1) and with prior given by (1.2)–(1.9). Let $\mathcal{K}_g = \mathrm{blkdiag}\,[\tau_{1g}K, \tau_{2g}K, \dots, \tau_{G^\star g}K]$ and $\Psi_g = \mathcal{X}_g'\mathcal{X}_g + \mathcal{K}_g$. Where $\mathcal{X}_g$ is the design sub-matrix conformable to $G_g^\star$, the number of parents of gene g. Then, the posterior distribution of $\{\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_G, \boldsymbol{\lambda}\}$ is proper if Ψ_g is positive definite for every g and $M \times G_g^\star < T$.

Given that in most of our applications we will only have a limited number of time measurements compared to the number of genes, this leads to an improper posterior if the prior was not proper, since the number of parents for any given gene only needs to exceed T/M. To construct a proper prior we supply (1.8) with an independent specification for the first two coefficients,

$$\pi(\beta_1, \beta_2) = \mathrm{N}(\beta_1 \mid 0, k_1)\,\mathrm{N}(\beta_2 \mid 0, k_2)\,. \tag{1.10}$$

Including these into the covariance structure we have

$$K_{1,1} = (1 + k_1/\tau), K_{1,2} = -2, K_{1,3} = 1, K_{2,2} = (5 + k_2/\tau) \text{ and } K_{2,3} = 1,$$

for the appropriate smoothness parameter, τ.

2. MCMC SCHEME

Combining the likelihood with the prior and letting Θ denote all the model parameters we obtain,

$$\pi(\Theta \mid \mathcal{X}, Y) \propto \left[\prod_g \mathrm{N}_T\big(\boldsymbol{y}_g \mid \boldsymbol{\mu}_g + \mathcal{X}\,\widetilde{\boldsymbol{\beta}}_g, \lambda_g I_T\big)\right] \times$$

$$\left[\prod_g \pi(\boldsymbol{\beta}_g \mid \boldsymbol{\tau}_g)\pi(\boldsymbol{\tau}_g)\pi(\lambda_g)\pi(\mu_g)\pi(\boldsymbol{\gamma}_g \mid \rho_g)\pi(\rho_g)\right],$$

where I_T is the identity matrix of size T, $\boldsymbol{\gamma}_g = \{\gamma_{1g}, \ldots, \gamma_{Gg}\}$ and $\boldsymbol{\tau}_g = \{\tau_{1g}, \ldots, \tau_{Gg}\}$. As there is no closed form expression for the posterior numerical methods are needed. We propose a Metropolis-within-Gibbs scheme which is drafted below.

Precisions The full conditional of λ_g, $g = 1, \ldots, G$ is given by

$$\pi(\lambda_g \mid \text{—}) \propto \lambda_g^{T/2+a_\lambda-1} \exp\left[-\lambda_g\left(b_\lambda + \frac{1}{2}\,\boldsymbol{e}'_g\,\boldsymbol{e}_g\right)\right]$$

which is the kernel of a gamma distribution, with $\boldsymbol{e}_g = \boldsymbol{y}_g - \boldsymbol{\mu}_g - \mathcal{X}\widetilde{\boldsymbol{\beta}}_g$.

Constant term μ_g is conditionally Gaussian, with mean and precision

$$m_g = \frac{\bar{\boldsymbol{y}}_g - \bar{\mathcal{X}}\widetilde{\boldsymbol{\beta}}_g}{\lambda_g + \tau_\mu/T} \qquad \text{and} \qquad \tau'_\mu = \tau_\mu + T\,\lambda_g\ ,$$

respectively, where $\bar{\boldsymbol{y}}_g = T^{-1}\sum_t y_g^t$ and $\bar{\mathcal{X}} = T^{-1}\sum_t \mathcal{X}$.

Connectivity The full conditionals for the gene-wise connectivity, ρ_g, are obtained as

$$\pi(\rho_g \mid \text{—}) \propto \rho_g^{S_g+a_\rho-1}(1-\rho_g)^{G+b_\rho-S_g-1} \text{ with } S_g = \sum_i^G \gamma_{gi}\ ,$$

and are sampled from a $\text{Be}(\rho_g \mid S_g + a_\rho, G + b_\rho - S_g)$, for $g = 1, \ldots, G$. The overall connectivity, ρ, is sampled from a $\text{Be}\big(\rho \mid S + a_\rho, G^2 + b_\rho - S\big)$, with $S = \sum_{g=1}^G S_g$.

Smoothness parameters When the corresponding link is on, the full conditional is given by

$$\pi(\tau_{jg} \mid \text{—}) \propto \tau_{jg}^{(M-2)/2+a_\tau-1} \exp\left[-\tau_{jg}\,\frac{1}{2}\widetilde{\boldsymbol{\beta}}'_{jg} K\,\widetilde{\boldsymbol{\beta}}_{jg}\right]\ , \qquad 0 < \tau_{jg} < b_\tau$$

and can be sampled from a truncated gamma distribution (Damien and Walker, 2001; Gentle, 2003) with parameters $\left\{(M-2)/2 + a_\tau, \widetilde{\boldsymbol{\beta}}'_{jg} K\,\widetilde{\boldsymbol{\beta}}_{jg}/2\right\}$. An observation is drawn from the prior when the link is off.

Spline Coefficients and link probabilities The update of the spline coefficients and indicator variables is performed as a block. Specifically, the update of a given indicator variable γ_{jg} and all the coefficients of the regression for gene g, $\boldsymbol{\beta}_g$, are performed simultaneously. In practice, as the regression is sparse, only a few links are actually present drastically reducing this computation. At every iteration, the individual link indicator γ_{jg} is turned on (off) if it is off (on) and the associated coefficient, $\boldsymbol{\beta}_g \in \mathbb{R}^{MG}$, for present links (on) is proposed from the joint conditional. Schematically we have,

$$\gamma : 0 \to 1 \qquad \text{and} \qquad \boldsymbol{\beta} : \boldsymbol{\beta}^a \to \boldsymbol{\beta}^b$$

with acceptance probability

$$\alpha = \min\left\{ \frac{\pi\left(\tilde{\boldsymbol{\beta}}^b\right)}{\pi\left(\tilde{\boldsymbol{\beta}}^a\right)} \frac{q(\boldsymbol{\beta}^a \mid \gamma^a) q(\gamma^a)}{q(\boldsymbol{\beta}^b \mid \gamma^b) q(\gamma^b)},\ 1 \right\} \quad ,$$

where the subscripts have been omitted for clarity. γ is proposed symmetrically, thus $q(\gamma^a)/q(\gamma^b) \ = 1$. For $q(\boldsymbol{\beta} \mid \gamma)$ we use the proposal

$$q(\boldsymbol{\beta} \mid -) = N(\mu_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}})$$

with

$$\Sigma_{\boldsymbol{\beta}} = \left[\lambda_g \mathcal{X}_g' \mathcal{X}_g + \Upsilon_g\right] \qquad \text{and} \qquad \mu_{\boldsymbol{\beta}} = \lambda_g(\boldsymbol{y}_g - \boldsymbol{\mu}_g)\mathcal{X}_g \Sigma_{\boldsymbol{\beta}}^{-1} \ ,$$

where Υ_g is the block diagonal penalty (precision) matrix, calculated by multiplying each block in $\mathcal{K}_g$ times the corresponding τ_{jg}. Note that, as only the coefficients with non zero indicator variable are updated, $\mathcal{X}_g$, $\boldsymbol{y}_g$ and Υ_g are adjusted to only include the appropriate elements. Substituting this in the Hastings ratio gives

$$\frac{\rho}{1-\rho}\ \tau_0 \tau_{jg}^{(M-2)/2} \frac{\exp\left[\frac{1}{2}\mu_{\boldsymbol{\beta}}^b \Sigma_{\boldsymbol{\beta}}^{-1b} \mu_{\boldsymbol{\beta}}^b\right]}{\exp\left[\frac{1}{2}\mu_{\boldsymbol{\beta}}^a \Sigma_{\boldsymbol{\beta}}^{-1a} \mu_{\boldsymbol{\beta}}^a\right]} \frac{\left|\Sigma_{\boldsymbol{\beta}}^b\right|^{1/2}}{\left|\Sigma_{\boldsymbol{\beta}}^a\right|^{1/2}} \ .$$

The opposite move (switching an indicator variable off) can be performed using the reciprocal of the ratio above.

In order to enforce the identifiability restriction, at each step we calculate $\bar{m}_g = \boldsymbol{\iota}_T \times \left[\mathcal{X}\ \tilde{\boldsymbol{\beta}}_g\right]$, for every gene, subtract it from the splines and add it to the constant term, μ_g.

Our sampler exploits the conditional independence structure of the model. We constructed a parallel scheme where the calculation for each parent is assigned to a CPU-node, these communicating only when the overall connectivity is updated and for sample recording. Gains in computation times can potentially be up to n-fold, with n the number of CPU-nodes used.

2.1 *Improving convergence*

The last move in Section 2 leads to a dramatic decrease in autocorrelation of the Markov chain, compared to a Gibbs move. Indeed, a common approach in these cases is to use a full Gibbs specification, with a full conditional Bernoulli distribution on the γ_{jg} and a full conditional Gaussian for the coefficients $\boldsymbol{\beta}_{jg}$. The latter requires the introduction of a so-called pseudo-prior which needs to be tuned to improve the mixing of the chain (Dellaportas *and others*, 2000; Ntzoufras, 2002; O'Hara and Sillanpää, 2009). In order to assess the gains in mixing, we implemented a full Gibbs sampler for the linear model. When chain mixing is compared, the advantage of our MH update becomes apparent as illustrated in Figure 1, obtained by running both samplers on the non-linear synthetic data described in Section 3.1. The top panels plot the number of times that link was switched during the MCMC run against the posterior probability of the link being present. One would expect that links with probabilities around $1/2$ would change more often, as in Figure 1b. However, the Gibbs strategy tends to mix more slowly, as shown in Figure 1a. Although the MH step is more computationally demanding, the benefit brought about by the improved mixing of the chain, quantified by the reduction in autocorrelation (ACF), offsets this cost easily (compare Figure 1c with Figure 1d). Given that the parameter space of the splines model is much larger than the linear one, the benefits of using this move, compared to the full Gibbs alternative are expected to be even greater.

2.2 *A linear model*

Our fully parametric, linear AR(1) model

$$y_g^{t+1} = \mu_g + \sum_{j=1}^{G} \beta_{jg} y_j^t + \varepsilon_g^t \ , \tag{2.1}$$

has similar basic interaction dynamics to those used in *e.g.* Opgen-Rhein and Strimmer (2006) and Lëbre (2009). However, the way network topology estimation is carried out varies significantly. Clearly, the latter models outperform our implementation in terms of speed; however, our Bayesian formulation is capable of providing measures of variability on all model parameters, including the network topology. Further, by using an identical prior structure (up to the relevant terms), we can focus on non-linear departures alone.

3. ILLUSTRATIONS AND APPLICATIONS

In all our applications, we include a slight modification of the structure of the network topology to that described in Section 1. We know from the context that each gene has a decay term, corresponding to mRNA decay. We include this information in the prior by fixing $\gamma_{gg} = 1$. As we also know that this decay is close to linear, we set the shape of the inverted Pareto prior for these smoothing parameters to thrice the value used for the rest, *i.e.* $a_{\tau_{gg}} = 3 \times a_{\tau_{ij}}, i \neq j$.

The splines and linear models were fitted using the overall and gene-wise connectivity specifications. Throughout, 13 bases were used, i.e. splines of degree 3 with 10 evenly spaced knots. Prior parameters were set to $\{a_\rho, b_\rho\} = \{1/2, 1/2\}$, $\{a_\lambda, b_\lambda\} = \{2, 0.01\}$, $\tau_\mu = 1/4$, $\tau_0 = 0.25$ and $\{a_\tau, b_\tau\} = \{1.5, 10^4\}$. We ran two parallel chains of length 10^5, dropping the first 10^4 steps and then recording every tenth draw. We performed some sensitivity analyses, varying a_τ from 1 (uniform prior) up to 3, setting $a_\rho = b_\rho = 1, 2$ and using flatter versions of the prior for λ by setting $a_\lambda = 1, 0.1$, without finding noteworthy differences. Convergence was assessed by comparing both chains graphically and by formal tests using the CODA package (Plummer *and others*, 2006).

3.1 *Discrete time synthetic networks*

In order to assess the network topology recovery power of our model, we produced two synthetic, first order autoregressive processes. One has only linear and the second a number of non-linear (S-shaped) relations. In the non-linear case, all the functional relations were produced using Hill functions, except for the self-interactions which are linear. In both cases we set $G = 16$, $T = 40$, and $\rho \approx 0.1$.

The models with gene-wise and overall connectivity produced almost indistinguishable estimations for the network topology and thus we report the results for the simpler model only. In Figure 2a we plot the marginal posterior and prior distributions of the model precisions, λ_g (for a selection of the genes only, to avoid clutter). We also performed a sensitivity analysis on ρ, fixing the prior parameters $a_\rho = b_\rho = 2$. As shown in Figure 2b, the posterior was practically unaffected by this change.

When the topology of the network is known, we can use the ROC curve plotted in Figure 3 below, to assess graphically the retrieval performance of a model (Pepe, 2000; Sing *and others*, 2005). A more formal comparison can be carried out by calculating the area under the ROC curve (AUC): the closer the AUC to one, the better the retrieval. For the linear data set these were 0.999 for the fully parametric model and 0.998 with the splines; and when fitting the non-linear data set we obtained 0.728 and 0.912, respectively. An alternative measure of fit is the so-called mean cross entropy (MxE), calculated as the Kullback-Leibler divergence from the known network topology to that estimated by the inferred connectivity matrix $\widehat{\Gamma}$, averaged over all possible links. MxE is bounded from below at zero, when the predicted topology is identical to the real one. Its value for a network topology predicted totally at random, *i.e.* $\widehat{\gamma}_{ij} = 1/2$, is $-\log 1/2 \approx 0.7$. In the linear network the MxE was 0.042 when fitting the parametric model and 0.064 when fitting the splines; with Hill interactions the values were 0.41 and 0.22, respectively. It is apparent that network topology prediction from both models is almost identical when fitting the linear interactions network (left panel). When non-linearities are present, the splines model achieves a better network retrieval, according to the ROC curve (right panel).

3.2 *Biological* GRN*: the plant Circadian Clock*

Most organisms have the ability to track time even in the absence of external input (*e.g.* light). This ability allows the organism to anticipate and prepare for future events, thus enabling it to optimise the interaction with the environment. In some cases, such as in *Arabidopsis,* diurnal period tracking is achieved via a regulatory network that oscillates with a period close to 24 hrs. This period then propagates through one or more of the core genes of the clock to target genes responsible for other biological processes (reviews can be found in Harmer, 2009; Más, 2008; McClung, 2006). The circadian clock is of central importance and has been extensively studied both experimentally and through mathematical modelling. It has recently been reported to regulate up to 90% of the *Arabidopsis* genome under some environmental conditions (Michael *and others,* 2008). While the circadian clock is able to maintain oscillations without the need of light, it is known that the period is modified by light exposure, allowing it to adapt to shorter and longer daylight hours. The ODE and current working network models are schematically outlined in Figure 4. As usual, genes are represented by nodes and regulation by directed edges ending either in arrows (activation) or bars (inhibition).

3.2.1 *Differential Equation Data.* We generated data from Locke *and others* (2006) using COPASI (Hoops *and others,* 2006) fixing the light source to be permanently on. The data was then subsampled, logged and standardised. The resulting data set has 50 time points with a time spacing of 1 hr, see Figure 5 a. Given that simultaneous measurement of multiple proteins is currently very hard, usually only mRNA is available for network inference. For this reason, although the ODE model outputs protein concentration and location, we used only the mRNA data. We expect the data generated from this model to be a reasonable reflection of experimental data, not only because it is a continuous time model with nonlinear interactions, but it also reflects realistic sampling regimes and interaction intensities.

3.2.2 *Experimental Data.* Whole leaves were harvested every 2 hrs for 48 hrs, with four biological replicates at each time point. To reduce variability, the same leaf (the 7th leaf to emerge) was harvested for each sample. This means that the same plant was not monitored over the entire time series but leaves of 96 distinct plants grown in identical conditions were sampled (four at each of the 24 time points). Full genome expression profiles of these leaves were generated using CATMA arrays (Sclep *and others,* 2007). Data processing and normalisation of the time series was carried out using a pipeline based on the R package MAANOVA (Wu *and others,* 2003). Given that the replicates showed some outliers we use the median of the four bioreplicates as the observed series. We use the same genes as those that appear in the ODE model, leaving some freedom to choose which genes to use for the ambiguous nodes. For the two genes that represent pairs, we selected those that showed least variability across replicates (LHY, PRR7). To represent X we chose the genes amongst the candidates in the current working model that showed the strongest signal-to-noise ratio: LUX and ELF4. For Y we chose PRR5 as it had a stronger signal than GI. Traces of the selected genes are depicted in Figure 5 b.

One way to assess the accuracy of the splines inferred network topology is to restrict our analysis to the most extensively studied genes: PRR7, LHY, TOC1 and (less so) PRR5. We can reasonably assume that the connections between these links are known. As can be seen in Figure 4b, there are 6 connections amongst these genes. The splines model predicts 5 connections, of which 3 are correctly predicted. Both of the incorrect predictions appear for genes that are missing a link in the inferred networks, indicating that the model has found the wrong parent rather than overfitting with more parents than necessary. Neither candidate gene for X (LUX and ELF4) regulates LHY, which would be evidence supporting the hypothesis that one of them is the unknown gene. On the other hand, these genes were proposed to be X as they are known to be involved in the clock and have some effect on the system, so the predictions can serve as a working hypotheses for determining the role they play within the network.

4. Computation times

Our algorithm took 2.7 hrs to run 10^5 iterations with the nonlinear synthetic data ($G = 16$, $T = 40$, $\rho \approx 0.1$) and scaling is likely to be quadratic in the number of genes and number of potential parents, and linear in the number of time points; for instance, fitting a microarray gene expression data set —not shown— with $G = 30$ and $T = 37$ ($\hat{\rho} \approx 0.15$) took 20 hrs for the same run length. Thus, for data sets with a large number of genes a parallel algorithm is available which reduces computation time approximately linearly in the number of cpu-nodes; for instance, using 31 cpus the former data set took 28.6 mins and the latter 3 hrs.

References

Damien, P. and Walker, S. G. (2001). Sampling truncated Normal, Beta and Gamma densities. *Journal of Computational and Graphical Statistics* **10**, 206–215.

Dellaportas, P., Foster, J. J. and Ntzoufras, I. (2000). Bayesian variable selection using the Gibbs sampling. In: Dey, D. K., Ghosh, S. K. and Mallick, B. K. (editors), *Generalized linear models: a Bayesian perspective*. New York: Marcel Dekker, pp. 273–286.

Fahrmeir, L. and Kneib, T. (2009). Property of posteriors in structured additive regression models: Theory and empirical evidence. *Journal of Statistical Planning and Inference* **139**, 843–859.

Gentle, J. E. (2003). *Random number generation and Monte Carlo methods*, 2 edition. New York: Springer-Verlag.

Harmer, S. L. (2009). The circadian system in higher plants. *Annual Review of Plant Biology* **60**, 357–377.

HOOPS, S., SAHLE, S., GAUGES, R., LEE, C., PAHLE, J., SIMUS, N., SINGHAL, M., XU, L., MENDES, P. AND KUMMER, U. (2006). COPASI —a COmplex PAthway SImulator. *Bioinformatics* **22**, 3067–3074.

LËBRE, S. (2009). Inferring dynamic genetic networks with low order independencies. *Statistical Applications in Genetics and Molecular Biology* **8**, Article 9.

LOCKE, J. C. W., KOZMA-BOGN·R, L., GOULD, P. D., FEHÉR, B., KEVEI, E., NAGY, F., TURNER, M. S., HALL, A. AND MILLAR, A. J. (2006). Experimental validation of a predicted feedback loop in the multi-oscillator clock of *Arabidopsis thaliana*. *Molecular Systems Biology* **2**, 59.

MÁS, P. (2008). Circadian clock function in *Arabidopsis thaliana*: time beyond transcription. *Trends in Cell Biology* **18**, 273–281.

MCCLUNG, C. R. (2006). Plant circadian rhythms. *The plant Cell* **18**, 792–803.

MCCLUNG, C. R. (2008). Comes a time. *Current Opinion in Plant Biology* **11**, 514–520.

MICHAEL, T. P., MOCKLER, T. C., BRETON, G., MCENTEE, C., BYER, A., TROUT, J. D., HAZEN, S. P., SHEN, R., PRIEST, H. D., SULLIVAN, C. M., GIVAN, S. A., YANOVSKY, M., HONG, F., KAY, S. A. *and others*. (2008). Network discovery pipeline elucidates conserved time-of-day-specific cis-regulatory modules. *PLoS Genetics* **4**, e14.

NTZOUFRAS, I. (2002). Gibbs variable selection using BUGS. *Journal of Statistical Software* **7**(7), 1–19.

O'HARA, R. B. AND SILLANPÄÄ, M. J. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Analysis* **4**, 85–118.

OPGEN-RHEIN, R. AND STRIMMER, K. (2006). Inferring gene dependency networks from genomic longitudinal data: a functional data approach. *REVSTAT* **4**, 53–65.

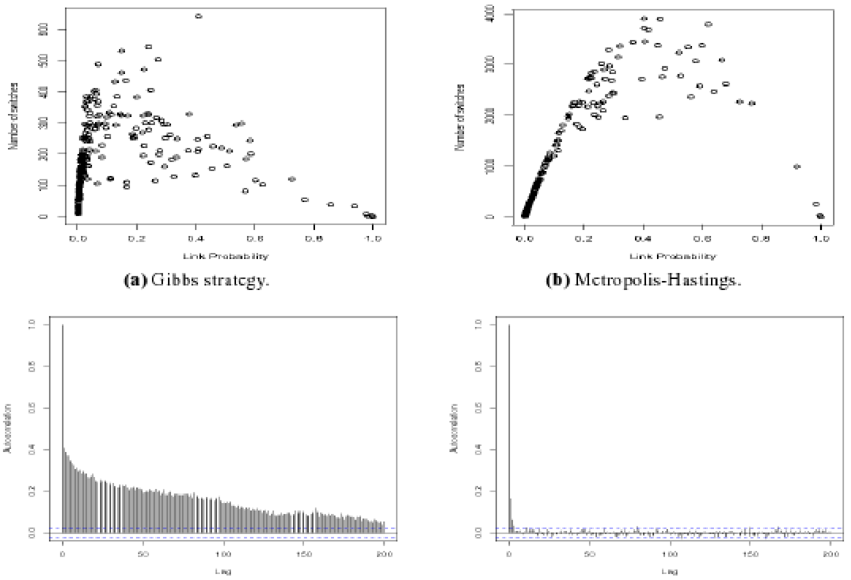
PEPE, M. S. (2000). Receiver operating characteristic methodology. *Journal of the American Statistical Association* **95**, 308–311.

PLUMMER, M., BEST, N., COWLES, K. AND VINES, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News* **6**, 7–11.

SCLEP, G., ALLEMEERSCH, J., LIECHTI, R., DEMEYER, B., BEYNON, J., BHALERAO, R., MOREAU, Y., NIETFELD, W., RENOU, J. P., REYMOND, P., KUIPER, M. T. R. *and others*. (2007). CATMA, a comprehensive genome-scale resource for silencing and transcript profiling of *Arabidopsis* genes. *BMC Bioinformatics* **8**, 400.

SING, T., SANDER, O., BEERENWINKEL, N. AND LENGAUER, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics* **21**, 3940–3941.

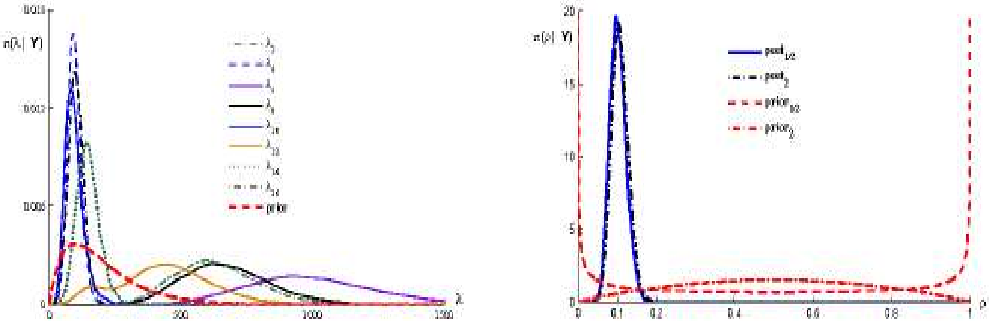
WU, HAO, KERR, M. K., CUI, XIANGQIN AND CHURCHILL, G. A. (2003). MAANOVA: A software package for the analysis of spotted cDNA microarray experiments. In: Parmigiani, G., Garrett, E. S., Irizarry, R. A. and Zeger, S. L. (editors), *The Analysis of Gene Expression Data: Methods and Software*. New York: Springer, pp. 313–341.



(a) Gibbs strategy. **(b)** Metropolis-Hastings.

(c) ACF λ_{15}, Gibbs strategy. **(d)** ACF λ_{15}, Metropolis-Hastings.

Fig. 1. Chain mixing comparison of the Gibbs and MH strategies. Top panels plot the number of state changes of a link during the MCMC run against its posterior probability. The bottom panels show the autocorrelation function (ACF) for a single link's precision (gene 15).



(a) Precision posteriors. **(b)** Connectivity posterior.

Fig. 2. Marginal posterior distributions calculated when fitting the splines model to the non-linear synthetic data set. (a) The posterior for a selection of the gene precisions, λ_g and the corresponding prior. (b) The posterior of the overall-connectivity, ρ, from two different priors, $a_\rho = b_\rho = 1/2$ and 2. In both panels priors are depicted by the thick (red) dashed lines.

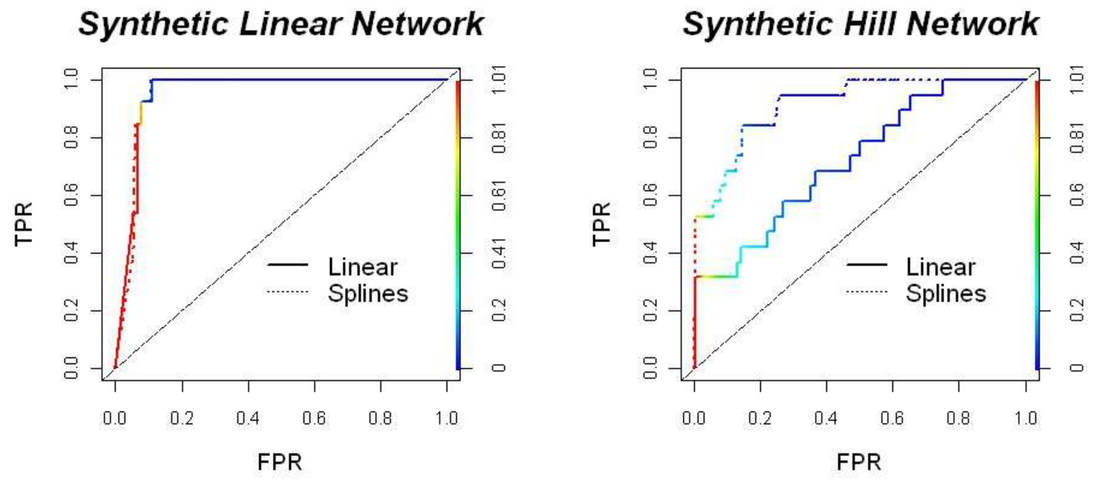
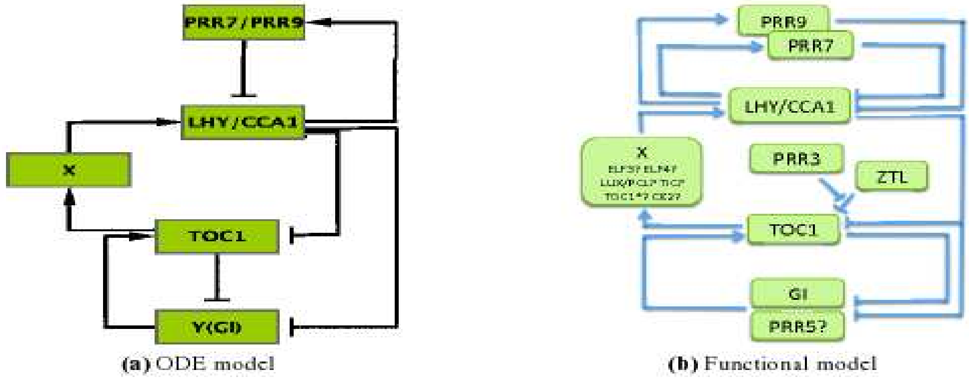
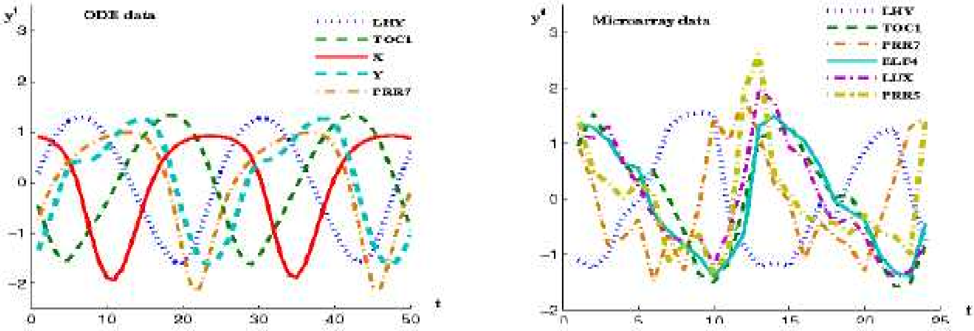


Fig. 3. ROC curves from synthetic networks. On the left panel, the ROC curves from the linear synthetic network obtained when fitting the linear AR(1) and the splines models. The right panel depicts the ROC from the non-linear, synthetic network.



(a) ODE model **(b)** Functional model

Fig. 4. Models of the Circadian Clock in *Arabidopsis thaliana*. (a) The ODE model of Locke *and others* (2006). (b) The current working model of the clock (redrawn from McClung, 2008). Nodes represent genes known (or suspected) to be part of the clock. Positive regulation is represented by an arrow and suppression by an edge with a bar end.



(a) ODE data **(b)** Microarray data

Fig. 5. Time traces of the ODE model and experimental data for the Circadian Clock in *Arabidopsis thaliana*. (a) Data simulated with the ODE clock of Locke *and others* (2006). (b) Gene expression profiles of *Arabidopsis* leaves. Data sets are standardised.

Chapter 3

Error Model

Noise is a prominent feature of biological systems with variability occurring from the molecular level all the way up to the multicellular level. When designing experiments it is common to include replicate measurements so as to have a measure of the different sources of variability.

In this paper we look at different ways of using replicate expression time-series for the retrieval of regulatory networks. Microarray gene expression measurements require destructive sampling, meaning that replicate measurements cannot be arranged into a single set of replicate time-series. This is because each replicate time-point cannot be assigned to a specific time-series, leading to a large number of different ways of arranging the data into replicate time-series, with each of these combinations potentially yielding different network predictions.

A straightforward way of analysing replicate data is to take the mean/median of the replicates and proceed to infer a network in the usual way. In this paper we give evidence that by proceeding in this way valuable information is discarded. We propose two variants of a statistical model that make use of the information in the variability. In our approaches we assume there is a "true value" of the expression data for the population and that the replicates are noisy surrogate measurements of the true value. This true value is inferred by the models simultaneously with the regulatory network. The two proposed models differ in the distribution of the replicates around the true value, in one case Gaussian and the other Student.

We benchmarked the two models using several synthetic datasets with replicates and compared the results to inference using just the mean value of the replicates. The retrieved networks showed clear improvement for the replicates models over the mean data model. Finally, we used the models to analyse an experimental dataset from the model organism *Arabidopsis thaliana*, consisting of a 24 time-point time-series with 4 replicates per time-point. The preferred model was Student-t, suggesting that outliers can have a significant impact on network inference.

3.1 Contribution

The initial idea was formulated by Morrissey, Juárez and Burroughs. The mathematical expressions used for the sampler were calculated separately by Morrissey and Juárez, and later compared. MATLAB code was contributed by Morrissey. Theoretical and sampler related issues were discussed and solved between Morrissey and Juárez, with the supervision of Burroughs. Morrissey generated the synthetic data, ran the models and analysed the outputs.

Morrissey's contribution to the writing was as follows. During the draft stage, Morrissey wrote the results section and the conclusions. Morrissey participated in the revision of the draft suggesting changes and edits. During the journal submission process, Morrissey and Juárez worked on addressing the problems raised by the reviewers with the supervision of Burroughs.

The microarray data as well as the description of the experimental design was contributed by Denby.

BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 18 2010, pages 2305–2312
doi:10.1093/bioinformatics/btq421

Systems biology

Advance Access publication July 16, 2010

On reverse engineering of gene interaction networks using time course data with repeated measurements

E. R. Morrissey[1,†], M. A. Juárez[1,*,†], K. J. Denby[1,2] and N. J. Burroughs[1]

[1]Warwick Systems Biology Centre and [2] Warwick HRI, University of Warwick, Coventry CV4 7AL, UK

Associate Editor: Trey Ideker

ABSTRACT

Motivation: Gene expression measurements are the most common data source for reverse engineering gene interaction networks. When dealing with destructive sampling in time course experiments, it is common to average any available measurements for each time point and to treat this as the actual time series data for fitting the network, neglecting the variability contained in the repeated measurements. Proceeding in such a way can affect the retrieved network topology.

Results: We propose a fully Bayesian method for reverse engineering a gene interaction network, based on time course data with repeated measurements. The observations are treated as surrogate measurements of the underlying gene expression. As these measurements often contain outliers, we use a non-Gaussian specification for dealing with measurement error. The network interactions are assumed linear and an autoregressive model is specified, augmented with indicator variables that allow inference on the topology of the network. We analyse two *in silico* and one *in vivo* experiments, the latter dealing with the circadian clock in *Arabidopsis thaliana*. A systematic attenuation of the estimated regulation strengths and a concomitant overestimation of their precision is demonstrated when measurement error is disregarded. Thus, a clear improvement in the inferred topology for the synthetic datasets is demonstrated when this is included. Also, the influence of outliers in the retrieved network is demonstrated when using the *in vivo* data.

Availability: Matlab code and data used in the article are available from http://go.warwick.ac.uk/majuarez/home/materials.

Contact: m.a.juarez@warwick.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 19, 2010; revised on June 23, 2010; accepted on July 14, 2010

1 INTRODUCTION

Identifying and understanding gene regulatory networks is of key importance in Systems Biology. Reverse engineering such networks is thus paramount and a plethora of literature dealing with the problem has developed in recent years (see Bansal *et al.*, 2007; Hache *et al.*, 2009 and references therein). Bayesian networks (BNs) have been used previously in gene network determination (Friedman *et al.*, 2000; Friedman, 2004; Hongqiang *et al.*, 2005). However, it is well known that when followed through time, biological processes have feedback loops and thus the validity of BNs is questionable when modelling such systems. Dynamic BNs (DBNs) have been proposed for modelling time course (longitudinal) gene expression data (Cao and Zhao, 2008; Murphy and Mian, 1999; Perrin *et al.*, 2003; Yu *et al.*, 2004; Zou and Conzen, 2005). These can be thought of as 'unfolding' a BN for every time point and when folding it back self-regulation and cliques may be obtained.

Formally, a DBN is characterized by a set of conditional relations, $p(y^{t+1} \mid \boldsymbol{y}^t)$. In the case of a regression-based DBN, these relations can be written as

$$y_g^{t+1} = f_g(\boldsymbol{y}^t) + \varepsilon_g^{t+1}, \qquad (1)$$

where y_g^t is the expression level of gene $g = 1,\dots,G$, measured at time $t = 1,\dots,T$, $\boldsymbol{y}^t = \{y_1^t, y_2^t, \dots, y_G^t\}$ and ε_g^t is an idiosyncratic error term.

The approaches above assume one observed time series for each gene. However, gene expression measurement normally requires destruction of the sample, e.g. microarrays, and, therefore, the idea of a longitudinal time series becomes problematic. This is because a single individual is not followed throughout the experiment, but rather a population of cells or individuals are sampled and their gene expression measured. The phenomenon is particularly acute in experiments with multicellular organisms, where not even the same population of cells can be followed through time. Thus, rather than 'real' gene expression measurements, we are faced with a set of surrogate measures. In addition to the uncertainty involved in the sampling process, it is well known that gene expression measurement technologies, such as microarrays, render noisy data and frequently exhibit outliers (Brody *et al.*, 2002; Lewin *et al.*, 2007).

When repeated measurements are available, time course data used for reverse engineering gene interaction networks are commonly obtained as a (weighted) average of these replicates and, therefore, these sources of uncertainty are ignored. Neglecting the variability within the replicates can have severe effects when fitting a linear model, with perhaps the most important being attenuation of the coefficient estimates (see Carroll *et al.*, 2006; Fuller, 1987). Working within a univariate first-order linear autoregressive setting, Schmid *et al.* (1994) demonstrated that neglecting measurement error yields severe attenuation, of the autoregressive coefficient and the variability of this estimate. Interpreting this result within the framework of (1) with a linear specification of $f_i(\boldsymbol{y}^t)$, this suggests that an averaged time series will yield attenuated estimates of the interactions within the network, with a spurious sense of security given the concomitant underestimation of the variability of these estimates.

[*]To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Downloaded from bioinformatics.oxfordjournals.org at University of Warwick on May 17, 2011

© The Author 2010. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

In the context of regulatory networks, Fujita *et al.* (2009) examine the effect of including measurement error in a linear model, both in static and dynamic contexts. They verify that the OLS estimator of the regression coefficients is attenuated and that its corresponding standard error is sub-estimated; then they provide a correction based on the asymptotic behaviour of the OLS. However, they do not assume sparsity in the network and, therefore, require more observations than the number of parameters in the model, making practical applications of their method to biological network inference difficult. Further, when dealing with time series data, they are forced to assume the measurement error precision known, omitting this source of uncertainty, which in turn will affect the estimation of the autoregression coefficients' precision. The main aim of this article is to show that by neglecting the uncertainty within the measurement process some biases can be passed on to the estimated network and, more importantly, that the confidence in the retrieved network will be spuriously large.

In the sequel, we present a model that takes into account repeated measurements of time course gene expression data for estimating the topology of a gene interaction network. We model the network topology explicitly, enabling us to control for the effective number of parameters to be estimated and, therefore, able to apply the methodology to commonly available time course gene expression datasets, where the number of parameters exceeds the number of observations. To ease the presentation, we assume an underlying first-order linear autoregressive process, AR(1), for the interaction network. As gene expression measurements frequently exhibit heavier-than-normal tail behaviour, we model the measurement process with Student-*t* errors. To this end, we present our approach in Section 2. To account for all sources of uncertainty in the model and measurement process, modelling and estimation is carried out from a Bayesian perspective, with our prior specification and estimation procedure explained in Section 3. Illustrations of model fitting and comparisons with the standard approach are conducted in Section 4. A final discussion is provided in Section 5.

2 APPROACH

Denote by y_g^t the expression level of gene $g=1,\dots,G$, measured at time $t=1,\dots,T$. We model the interaction network as a linear AR(1) process,

$$y_g^{t+1}=\mu_g+\sum_{j=1}^{G} y_j^t\tilde{\beta}_{jg}+\varepsilon_g^t, \tag{2}$$

where μ_g is the basal expression level of gene g; $\tilde{\beta}_{jg}=\gamma_{jg}\beta_{jg}$ measures the influence of gene j on gene g, with $\beta_{jg}\in\mathbb{R}$ and $\gamma_{jg}=1$, if j regulates g and $\gamma_{jg}=0$ otherwise. Finally, ε_g^t is an idiosyncratic error term, centred at zero and with precision parameter λ_g, typically assumed to be Gaussian. We augment the model with the parenthood (link) indicator variables $\Gamma=\{\gamma_{jg}\}$, which will be the basis for estimating the network topology.

Assume now that instead of measuring y_g^t directly, we are presented with R surrogate measurements $X=\{x_{gr}^t\}$, $r=1,\dots,R$. As mentioned before, the common approach is to calculate

$$\bar{x}_g^t=\left[\sum_{r=1}^{R}\omega_{gr}\right]^{-1}\sum_{r=1}^{R}\omega_{gr}x_{gr}^t\,, \tag{3}$$

with $\omega_{gr}\geq 0$; and then replace y_g^t in (2) with $\bar{x}_g^t$. Influence of possible outliers will depend on the choice of $\boldsymbol{\omega}_g=\{\omega_{g1},\dots,\omega_{gR}\}$. The pervasive choice in the literature is to set $\omega_{gr}=1$; but when outliers are suspected, sometimes a more robust alternative, such as the median, can be used. In the process of fitting (2), the probability of any given link being present depends on the relative magnitude of its associated coefficient, and hence failure of acknowledging measurement error can yield an attenuated network connectivity.

In Section 3, we provide a fully Bayesian approach to accommodating measurement error when estimating the topology of a gene interaction network within a linear AR(1) specification.

3 METHODS

To account for the additional uncertainty when repeated measurements are available, we assume that the regulation process can be captured by (2), but instead of actually observing y_g^t, we have noisy measurements, x_{gr}^t, such that

$$x_{gr}^t=y_g^t+\eta_{gr}^t,\qquad r=1,\dots,R, \tag{4}$$

with η_g^t a zero mean measurement error term, with precision parameter τ_g, independent for all g,t,r. This error term is frequently assumed Gaussian; however, given that the measurement process can potentially produce outliers, we will use a Student-*t* specification, $\mathrm{St}(\eta_g^t\mid 0,\tau_g,\nu)$, such that $\mathrm{Var}\,[\eta_g^t]=\nu\tau_g^{-1}/(\nu-2)$ provided the degrees of freedom, $\nu>2$.

When combining (2) and (4), and using the Student-*t* representation as a Gamma scale mixture of normals, one can write the likelihood as

$$\begin{aligned}\ell(\Theta;X)=\prod_{g=1}^{G}\prod_{t=1}^{T}\prod_{r=1}^{R}&N(y_g^{t+1}\mid\mu_g+\boldsymbol{y}^t\tilde{\boldsymbol{\beta}}_g,\lambda_g)\\ &\times N(x_{gr}^t\mid y_g^t,\omega_{gr}^t\tau_g)\mathrm{Ga}(\omega_{gr}^t\mid\nu/2,\nu/2)\,.\end{aligned} \tag{5}$$

where $Y=\{y_g^t\}$ are the unobserved expression levels, $X=\{x_{gr}^t\}$ denote their surrogate measurements and $\Theta=\{\boldsymbol{\mu},B,\Gamma,\boldsymbol{\lambda},\boldsymbol{\tau},\nu\}$ collects the model parameters, with $\boldsymbol{\mu}=\{\mu_1,\dots,\mu_G\}$; $B=\{\boldsymbol{\beta}'_1,\dots,\boldsymbol{\beta}'_G\}\in\mathbb{R}^{G\times G}$ and $\boldsymbol{\beta}_g=\{\beta_{1g},\dots,\beta_{Gg}\}$; $\Gamma=\{\gamma_{ij}\}$; $\boldsymbol{\lambda}=\{\lambda_1,\dots,\lambda_G\}$; and $\boldsymbol{\tau}=\{\tau_1,\dots,\tau_G\}$.

The Bayesian model is completed by specifying a prior for all the unknowns. We use a product structure

$$\pi(\Theta)=\pi(\rho)\pi(\nu)\prod_{g=1}^{G}\left[\pi(\mu_g)\pi(\boldsymbol{\beta}_g)\pi(\lambda_g)\pi(\tau_g)\pi(\boldsymbol{\gamma}_g)\right], \tag{6}$$

and specify componentwise conditionally conjugate priors where suitable. Thus,

$$\pi(\mu_g)=N(\mu_g\mid 0,k), \tag{7}$$

$$\pi(\boldsymbol{\beta}_g)=N_G(\boldsymbol{\beta}_g\mid\mathbf{0},k_\beta I),\ g=1,\dots,G, \tag{8}$$

$$\pi(\lambda_g)=\mathrm{Ga}(\lambda_g\mid a_\lambda,b_\lambda), \tag{9}$$

$$\pi(\tau_g)=\mathrm{Ga}(\tau_g\mid a_\tau,b_\tau), \tag{10}$$

$$\pi(\boldsymbol{\gamma}_g\mid\rho)=\prod_{j=1}^{G}\mathrm{Ber}(\gamma_{jg}\mid\rho),\ g=1,\dots,G, \tag{11}$$

$$\pi(\rho)=\mathrm{Be}(\rho\mid a_\rho,b_\rho), \tag{12}$$

$$\pi(\nu)=\mathrm{Ga}(\nu\mid a_\nu,b_\nu). \tag{13}$$

Of paramount importance in our modelling is the inclusion of the link indicator variables, $\Gamma=\{\gamma_{ij}\}$, and their probabilistic structure, controlled by (11) and (12). By augmenting the model in this way, we are able to switch the regulation of gene i on gene j on or off, controlling for the effective number of parameters to be estimated. Moreover, the posterior mean of the link indicators, $\hat{\gamma}_{ij}$, are interpreted as the posterior link probabilities, the building blocks of the network topology retrieval. From a formal viewpoint,

Downloaded from bioinformatics.oxfordjournals.org at University of Warwick on May 17, 2011

Γ supply the means for performing an automated link selection (see Smith and Kohn, 1996). The overall connectivity of the network is controlled by ρ and any relevant information about this key aspect can be fed into the model through its prior.

Identifiability is always a potential issue when dealing with measurement error models, 'particularly when neither gold-standard measurements or pure replicate measurements can be obtained' (Gustafson, 2004, Section 6.3). This problem is further compounded in gene network determination where the number of potential parameters to be estimated ($p = \dim\Theta \geq G(G+1)$) is typically much greater than the available data points ($n = T \times G$), i.e. the so-called $p \gg n$ case. From a Bayesian perspective, models are always identifiable as long as a proper prior is specified. However, this formal identifiability may imply not learning from the data (Poirier, 1998) and thus calls for a careful elicitation of the prior (Gustafson, 2005). Indeed this is the case with the model and measurement precision parameters, when repeated measurements are absent ($R = 1$). We take advantage of the information contained in the measurements X about $\boldsymbol{\tau}$ and use a rather flat prior on this parameter, while carefully eliciting the parameters for $\pi(\boldsymbol{\lambda})$. To the extent of our knowledge, there is no conventional prior for the degrees of freedom, ν. We decided to use a gamma distribution such that P $[\nu \leq 30] \approx 0.6$ and with mode at 15; thus giving roughly prior odds of 3 to 2 for the measurement error distribution being fat-tailed. A detailed specification of the prior parameters is given in Supplementary Section 2S.

There is no closed analytic expression for the posterior distribution, $\pi(\Theta \mid X) \propto \ell(\Theta; X)\pi(\Theta)$, and numerical methods to explore it are needed. To this end, we constructed a Markov chain Monte Carlo (MCMC) algorithm. We use Gibbs sampling for all of the parameters except B and ν. For the former, (Morrissey *et al.*, submitted for publication) showed that a Metropolis-within-Gibbs strategy improves mixing and, therefore, faster convergence of the chain, and we follow their suggestion. We use a Metropolis step for ν with a Gamma proposal centred at the previous draw and tune its coefficient of variation to control for the acceptance rate.

In passing, noteworthy is the Gibbs step used for drawing a new non-observable expression level y_g^t. These are drawn from a Gaussian distribution, $N(y_g^t \mid m_g^t, p_g^t)$, with location

$$m_g^t = \frac{\lambda_g m_{\mathrm{AR}} + \tau_g m_{\mathrm{meas}}}{p_g^t}$$

and precision

$$p_g^t = \lambda_g(1+\widetilde{\beta}_{gg}^2) + \tau_g \sum_{r=1}^{R} \omega_{gr}^t,$$

where

$$m_{\mathrm{AR}} = \sum_{i \neq g} \widetilde{\beta}_{ig}(y_i^{t-1} - \widetilde{\beta}_{gg} y_i^t) + \widetilde{\beta}_{gg}(y_g^{t-1} + y_g^{t+1})$$

and

$$m_{\mathrm{meas}} = \sum_{r=1}^{R} \omega_{gr}^t x_{gr}^t.$$

It is apparent from the expression above that draws of y_g^t depend on the weighted average of the observed measures, with the weights determined by the degrees of freedom through ω_{gr}^t. These averages are then combined with the AR(1) component of the model, thus effectively entertaining all sources of uncertainty. A detailed description of sampler is presented in Supplementary Section 3S.

Our estimation method is computationally intensive. Runtime for any of the measurement error models is not significantly longer than for the plain AR(1), though. Sampler's 2×10^5 iterations with the 16 gene linear data used in Section 4.1.1 took 2.10, 2.14 and 2.24 h for the AR(1), the Gaussian and the Student error models, respectively. In terms of scalability, the algorithms show the usual problems associated with network inference. Large datasets (i.e. thousands of genes) can take unrealistically long times to run. Runtime can be reduced by allowing only transcription factors to be regulators (and possibly other genes that could affect regulation, such as kinases), encoding such information in the prior through (11). Further, the algorithm is straightforwardly parallelized, as the parameters for each gene can be computed independently, the CPU-nodes needing to communicate only for updating the overall connectivity, ρ and for collecting the draws, thus reducing the runtime roughly proportionally to the number of available CPU-nodes.

4 RESULTS

We analyse two *in silico* and one *in vivo* datasets. The simulated experiments allow us to isolate the effect of explicitly modelling measurement error in controlled situations and highlight the attenuation effect. We then turn to a real experiment dealing with the circadian clock in *Arabidopsis thaliana*. In all cases, we fit the model with and without the measurement error component, using the same prior structure (6)–(13), deleting the relevant terms when not accounting for measurement error or when assuming it is Gaussian distributed.

4.1 *In silico* networks

The first synthetic network is linear, and thus serves as a baseline for comparisons. The second is nonlinear and will allow comparisons within a more realistic, yet still controlled setting. For each synthetic dataset, we generate a rather large number of time points, 41 and 50, respectively, so we can highlight the effect of measurement error in network reconstruction.

4.1.1 Linear interactions Here, we use (2) and (4) to generate a synthetic data set with $G = 16$ and $T = 41$. We set the network connectivity $\rho \approx 0.13$ and produced a layered network: a hub gene is perturbed by an (unmeasured) external input, the signal is then propagated to a second layer of genes with another hub which, in turn, propagates the signal to a third layer. A small amount of links feeding forward and backward between layers are also included. Expression profiles can be found in Supplementary Figure S1a.

This dataset is regarded as the 'noiseless' case. Using it as basis, we generated noisy replicates according to (4) with η_g^t either Gaussian (GD) or Student (SD) distributed. For each distribution, we generated two datasets: one with few ($R = 4$) and a second with many ($R = 20$) replicates. In both cases, we consider rather noisy scenarios by setting $\tau_g^{-1/2}$ at 50% of the maximum absolute expression value of each gene (note that in the GD case, τ_g corresponds to the measurement precision). We fixed $\nu = 5$, for the Student-t case.

We fitted three models to each dataset: one where the measurement error is assumed to be Student distributed (SM), the second assumes normal errors (GM) and the third disregards measurement error by taking the mean of the replicates as the true time series (MM). To summarize the results of the inference on the network topology, we use two threshold-independent scores: the area under the ROC curve (AUC) and mean cross entropy (M×E). The AUC provides an overall accuracy measure of network retrieval, using the link predictions sorted according to their magnitude. It thus fails to account for the strength in the predictions; in our case, the estimated link probabilities. These are key as, when performing inference on an unknown network, we will normally set a threshold above which links will be predicted as being present. For this reason, we also calculate the M×E defined as the average Kullback–Leibler

Downloaded from bioinformatics.oxfordjournals.org at University of Warwick on May 17, 2011

Table 1. Performance comparison using a synthetic linear network

	AUC				M×E			
	$R=4$		$R=20$		$R=4$		$R=20$	
	GD	SD	GD	SD	GD	SD	GD	SD
MM	0.78	0.68	0.91	0.82	0.40	0.59	0.24	0.33
GM	**0.86**	0.77	0.92	0.89	**0.32**	0.37	0.21	0.28
SM	0.85	**0.81**	**0.93**	**0.92**	**0.32**	**0.36**	**0.20**	**0.24**

AUC and M×E scores obtained by fitting the model without measurement error (MM) and those with Gaussian (GM) and Student (SM) errors to *in silico* data with Gaussian (GD) and Student (SD) distributed errors. Bold values are the best scores for each case. The smaller the M×E the better. The larger the AUC the better.

divergence from the link structure of the true network to the posterior link probabilities, over all possible links (detailed in Supplementary Section 4.1S). As a baseline for comparison, the M×E of a perfectly inferred network is 0 and that of one predicted totally at random (i.e. probability of 1/2 for each link) is $-\log(1/2)\approx 0.7$. In the case of the AUC, this corresponds to values of 1 and 1/2, respectively. For instance, we fitted the AR(1) model with the ideal, noiseless data, resulting in an AUC of 0.99 and a M×E of 0.05.

When using GD, GM and SM perform equally well under both criteria—see the corresponding columns of Table 1. This is to be expected, since a Student distribution with large degrees of freedom approaches a Gaussian. Using a small number of replicates ($R=4$ in Table 1) GM and SM outperform MM in either criteria. When a large number of replicates is considered ($R=20$ in Table 1), the AUC for all three scenarios are quite close, indicating a similar ordering of the estimated link probabilities for all models fitted. However, the M×E scores are better for SM and GM, highlighting that the inferred probabilities with measurement error are comparatively higher for existent links and lower for non-existent.

Regarding inference on data with Student error, MM performs worse under both scores—first row in Table 1. In fact, the M×E score with few replicates is very close to that of random predictions. SM shows a small improvement over GM for both small and large R. This difference is slightly larger for $R=20$, illustrating that the degrees of freedom are hard to estimate; four replicates is barely enough to infer them, while $R=20$ allows for a more precise estimation.

The effect of measurement error can also be highlighted when concentrating on the data with several replicates (columns with $R=20$ in Table 1), while relative differences in the AUC between MM and SM are reduced to <2% for GD and 11% for SD, relative differences in M×E are 17% and 28%, respectively. This can be understood since attenuation reduces with increased number of replicates and, therefore, the point estimates of the coefficients from either model will be closer to each other, resulting in a similar AUC. However, the variance of these estimates will still be underestimated when using MM and as a result the M×E, which takes into account the actual value of the estimated probabilities, will capture these differences.

Attenuation and underestimation of the variability on the coefficients estimates are illustrated in Figure 1 for a specific link, $\text{gene}_{12}\rightarrow\text{gene}_9$. As expected in the ideal, noiseless case, the posterior distribution of the corresponding coefficient, $\beta_{12,9}$, has its mode close to the true value of the coefficient (1.0) and has a rather large precision. With noisy measurements MM renders an attenuated coefficient, with its posterior distribution shifted towards the origin (dashed curve). Moreover, the corresponding posterior has negligible mass near the true value. In contrast, the posterior from SM overlaps nicely (solid). Also, the MM estimate has a posterior precision of about 72, larger than the SM, 62, illustrating the underestimation of the uncertainty in the coefficient estimate. This effect was observed more or less markedly in all predicted links (not shown).

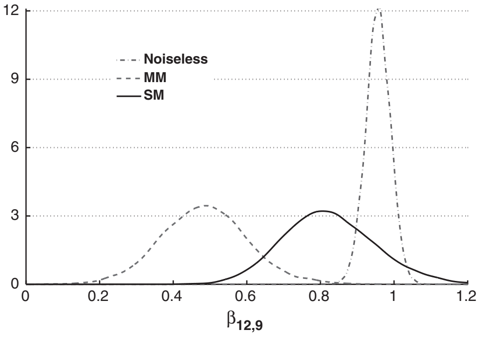


Fig. 1. Marginal posterior distributions of the coefficient, $\beta_{12,9}$ corresponding to link $\text{gene}_{12}\rightarrow\text{gene}_9$ of the *in silico* linear data. The noiseless data (dot-dashed) is the data without noise, the MM (dashed) is the mean of four replicates, neglecting measurement error and SM (solid) uses the four replicates and assumes Student measurement error.

4.1.2 Nonlinear interactions A dataset with nonlinear interactions was generated using a gene network model built with ordinary differential equations (ODEs). The network is a mathematical model of the *A.thaliana* circadian clock and consists of $G=5$ genes with eight links, and also includes protein production and transport, as well as daylight (Locke *et al.*, 2006). We generated data from this model using COPASI (Hoops *et al.*, 2006) with the light source fixed permanently on. To mimic realistic sampling regimes, we sub-sampled the data so as to have a time spacing of an hour and then took logs. The resulting time series has a total of $T=50$ time points (plots of the profiles are shown in Supplementary Fig. 1Sb). This is the 'noiseless' dataset. We generate noisy replicates using Student distributed errors. As before, we fix $\tau_g^{-1/2}$ at 50% of the maximum value of the noiseless gene expression, use $\nu=5$ and produced two datasets: one with 4 replicates and the other with 20 replicates.

Due to the small size and high connectivity of the network ($\rho\approx 0.40$), the scores used in the previous example become quite sensitive. This is because the ROC is a piecewise constant function and with a small number of genes (and therefore links), the size of the steps becomes larger and thus comparisons get more sensitive. For improved interpretability, we plot in Figure 2 the links included in the predicted network against the posterior link probability when fitting the data with $R=4$ replicates. We use a circle (cross) for an incorrect (correct) link; for instance, in the noiseless case the predicted network using a threshold of 0.95 would have nine links (circles and crosses with link probability threshold above 0.95 in Fig. 2a), four out of which (crosses) are correct.

Inspection of the inferred link probabilities using the noiseless data (Fig. 2a), shows that there is some overfitting. For instance,

Downloaded from bioinformatics.oxfordjournals.org at University of Warwick on May 17, 2011

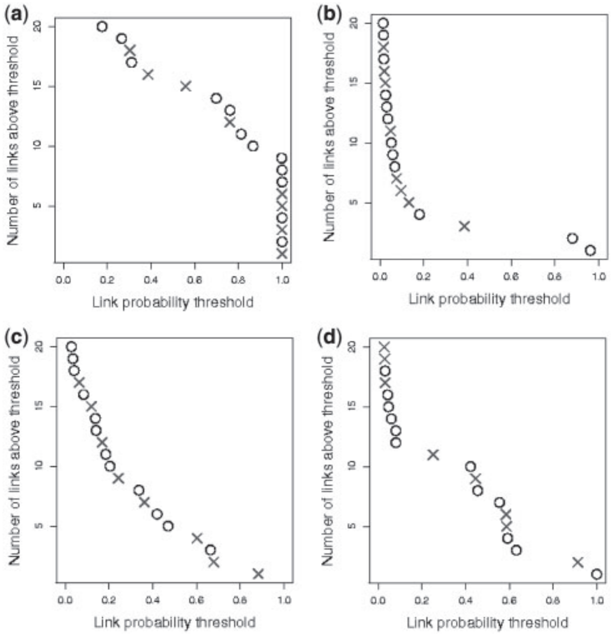


Fig. 2. Network topology retrieval for the ODE circadian clock *in silico* dataset with $R=4$ replicates. Links predicted to be present in the network versus posterior link probabilities estimated for the noiseless case and when using MM, GM and SM. A link (not) present in the ODE model is shown with a (circle) cross.

using a link probability threshold of 0.7, 14 out of 20 possible links are predicted to be present. The main reason for this is that the linear model is unable to explain the nonlinear interactions adequately, resulting in either a misprediction or a compensation by the inclusion of spurious parents (see Morrissey *et al.*, submitted for publication, for a detailed discussion).

For the case where we have few replicates, GM and SM outperform MM (Fig. 2b–d). MM predicts only two links with high probability ($\hat{\gamma}_{ij} \geq 0.8$), both incorrect. For a link probability threshold of 0.7, GM performs marginally better than SM (even though the true noise is Student). Using a threshold of 0.5, the inferred network would be the same for both models, though. If we compare model fit for low thresholds, MM predicts 17 links not present in the network with a threshold of 0.2 (i.e. those crosses and circles to the left of 0.2). Seven out of these are real (crosses); i.e. it has a high false negative rate, compared to the models with measurement error: 4 out of 10 for GM and 3 out of 9 with SM. Moreover, the true positive rate for this (low) threshold is 1 out of 3 for MM (one cross and two circles to the right of 0.2), while GM has 4 out of 10 and SM 5 out of 11. When a large number of replicates are available, the three models yield similar network reconstructions; however, the spread of posterior link probabilities is still more concentrated when using MM (see Supplementary Fig. S3).

The effect of attenuation in network retrieval can be seen by comparing Figure 2b to Figure 2c and d. Most posterior link probabilities of MM are tightly clustered towards zero; this is due in part to the combined effect of attenuated coefficients (and thus a lower overall connectivity) and the underestimation of the variability on these estimates (the tight grouping). In contrast, the posterior link probabilities in either GM or SM have a wider spread reflecting both the larger estimates of the coefficients and the increased variability in the estimates when considering measurement error.

4.2 *In vivo* data

We used a microarray time series of gene expression profiles from *A.thaliana* (Denby,K.J. unpublished data). Sampling is destructive, with a different plant used for each time-replicate. To reduce variability, the same leaf was used for each sample. A total of 96 plants were grown under a 16 h : 8 h light : dark cycle and the seventh leaf to emerge from each of $R=4$ plants was sampled every 2 h over a 48-h period, i.e. $T=24$.

To select those genes to be included in the analysis, we referred to the current working model of the circadian clock in *Arabidopsis* (Robertson and Webb, 2009), sketched in Figure 3a. Recently, a new gene (CHE) was identified as a member of the core circadian clock (Pruneda-Paz *et al.*, 2009), so we include this gene to entertain the most up-to-date version of the clock. Two nodes in Figure 3a (LHY/CCA1 and PRR7/PRR9) represent pairs of genes that perform the same role and have very similar expression profiles. To avoid collinearity, a single gene to represent each pair was selected. PRR7 is chosen over PRR9 as it shows a higher signal to noise ratio and CCA1 over LHY given that CHE is predicted to regulate CCA1 and not LHY (traces of the expression profiles are shown in Supplementary Fig. 2S).

Figure 4 depicts the distribution of the inferred link probabilities for each of the three models. As there are few links predicted with high probability, we set a link probability threshold of 0.5. Figure 3 shows the inferred networks for this threshold. GM and MM infer the same network topology, whereas SM infers a network with six links, only two of them in common with the GM/MM network.

This large difference is explained by the posterior probability distribution of the degrees of freedom (Fig. 5), where four out of the five genes are inferred to have a fat-tailed measurement error distribution, with a mode of $\nu=3$ for PRR7. The effect of attenuation in the estimation of the link probabilities is illustrated by comparing the three pictures on Figure 4. Again, the majority of probabilities estimated by MM are smaller than those estimated by either GM or SM and are more tightly grouped. On the other hand, GM and SM reflect the additional uncertainty in the measurement process by dragging these probabilities towards the centre of the plot.

The GM/MM network correctly predicts the TOC1→GI→TOC1 loop, but also predicts two incorrect links (GI→CCA1 and TOC1→PRR7). The SM model also correctly predicts TOC1→GI but incorrectly TOC1→PRR7. The other four links that are absent in the GM/MM predicted network involve either PRR7 (three links) or CCA1 (one) as a regulator. Out of these, there are two correct predictions: PRR7→CCA1, CCA1→TOC1 and two incorrect: PRR7→GI and PRR7→TOC1.

Noteworthy is the discrepancy between the SM and MM/GM inferred networks (Fig. 3). The fact that for the given threshold the MM and GM inferred networks are the same illustrates the effect of outliers in the estimation: when using GM, despite taking into account the uncertainty in the replicates through τ_g, these are treated as interchangeable in the update of y_g^t (in this case, $\omega_{gr}^t \equiv 1$). In contrast, the weights in SM depend on the degrees of freedom and will be more variable for smaller values of ν, allowing for some of the measurements to dominate the average. To verify that this is

Downloaded from bioinformatics.oxfordjournals.org at University of Warwick on May 17, 2011

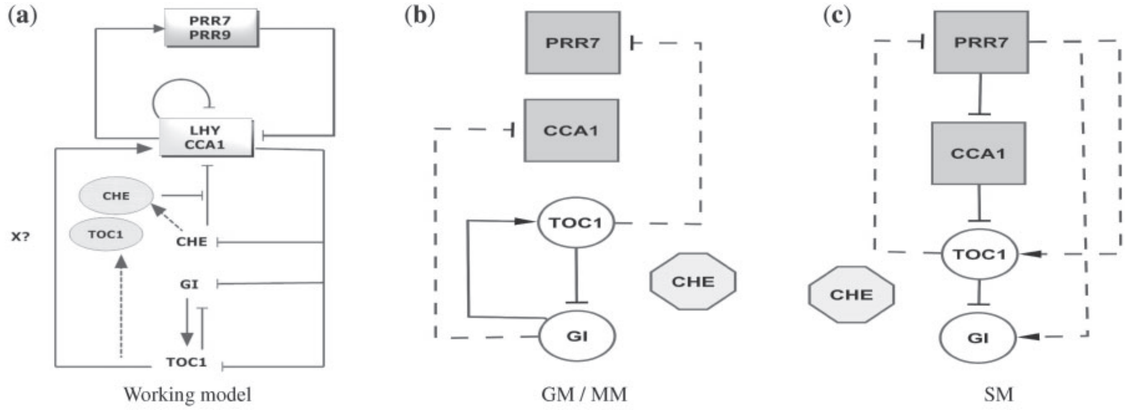


Fig. 3. *Arabidopsis* circadian clock. (**a**) Current working model (redrawn from Robertson and Webb, 2009): dotted lines represent protein production and oval shapes (binding) proteins. (**b**) and (**c**) Depict the network topologies inferred with the measurement error models with a threshold of 1/2, using the microarray data. Solid edges represent predicted links that are present in the working model and dashed represent links predicted by either model and not present in the working model. (**b**) The retrieved topology with Mean and Gaussian models and (**c**) with the Student model.

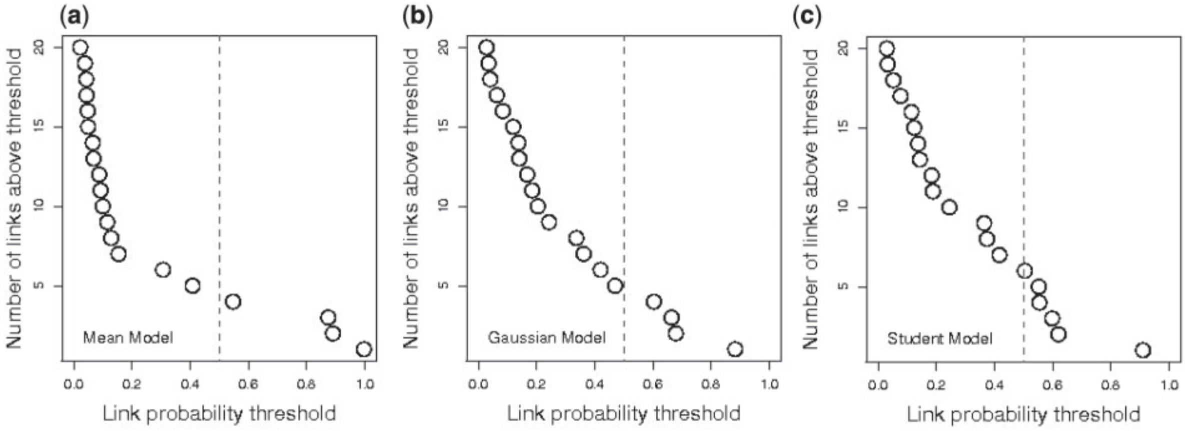


Fig. 4. Gene network link prediction for the *Arabidopsis* circadian clock microarray data. Posterior link probabilities from each model are depicted as circles. The vertical dotted line represents a threshold of 1/2.

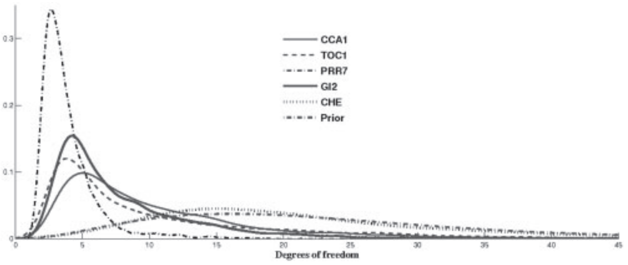


Fig. 5. Marginal posterior distributions of the degrees of freedom, ν, for each gene when fitting the Student measurement error model to the *Arabidopsis* circadian clock microarray data.

indeed the case, we fitted the AR(1) model using the median value of the replicates as the sole source of information. The predicted network topology (not shown) also has six links with only one of them differing from those inferred by SM (incorrect in both cases); nevertheless, the posterior distribution of the link probabilities was tighter, illustrating the sub-estimation of the coefficients' precision.

These comparisons are dependant on the threshold selected and as such are only a point estimate of the network. As described in the Section 1, the fundamental effect of not accounting for repeated measurements is the overconfidence in the estimation of the coefficients and its effect on the estimates of link probabilities. Even though the topology retrieved from MM and GM, for a fixed threshold, is similar, the distribution of the link probabilities is extremely different. MM predicts three links with very high probability (>0.8) and 14 with very low probability (<0.2), with only three in the central region of Figure 4a. In contrast, GM predicts only 1 link with high probability and 10 with low, leaving 9 possible links in the central zone of Figure 4b. The conclusions obtained from either model can thus be quite distinct, while MM suggests a high confidence in its predictions, GM correctly warns about the high levels of uncertainty in the recovery of the network.

Interestingly, no parents or children are predicted for the new clock gene CHE. Pruneda-Paz *et al.* (2009) showed that CHE can

Downloaded from bioinformatics.oxfordjournals.org at University of Warwick on May 17, 2011

bind TOC1 and also binds the promoter region of CCA1. This suggests that we can expect non-additive, nonlinear effects that may be difficult to capture with an additive linear model. For the regulation of CHE, we do not expect these problems. From Pruneda-Paz *et al.* (2009), we know that CHE is regulated by CCA1. Analysing the SM posterior link probabilities, we found that the other four genes appear as regulators of CHE with posterior probabilities of around 0.13, suggesting that either the regulation is highly nonlinear, or there are other regulatory mechanisms such as post-translational modifications. The latter would not be unlikely as it is well known that targeted protein degradation and sequestration plays a very important role in the circadian clock.

We further explored the effect of having less replicates available. To this end, we sub-sampled the original four replicate dataset and generated 60 datasets, 30 with three replicates and 30 with one. The datasets were sampled in a controlled manner to ensure they were not too similar to the original four replicate dataset (see Supplementary Section 4.3S).

We summarize the information by setting the same link probability threshold as in the four replicate case ($\hat{\gamma}_{ij} \geq 0.5$), and then counted the times a link was predicted to be present. These counts are shown in Table 2 for the 1-replicate case and Table 3 for the 3-replicate. As expected, the 'no replicates' scenario (Table 2) shows high variability in the inferred networks. No single link was predicted as present in more than half the datasets. A third of the predicted networks had no more than one predicted link and a further third of the networks had from four to six predicted links (not shown). Three links appeared more frequently than the rest: the two link loop TOC1→GI→TOC1 and the PRR7→TOC1 link. The two links in the loop were predicted together only in five of the datasets, reflecting the loss in estimation precision when no replicates are available.

Table 2. Circadian clock experimental data. Link prediction counts using the 30 sub-sampled datasets with one replicate and a threshold of 1/2

	CCA1	TOC1	PRR7	GI	CHE
CCA1	0	3	3	4	1
TOC1	6	0	3	14	0
PRR7	2	14	0	2	1
GI	0	13	6	0	0
CHE	1	3	0	0	0

As there is only one replicate per gene, the model without measurement error is used. Genes in columns are regulators and rows are regulatees.

Table 3 illustrates the benefits of including repeated measurements: there is a clear separation in the link prediction frequency, with a few links being predicted quite frequently and the rest barely appearing. Those links predicted in more than half of the datasets (highlighted in Table 3) are consistent with those predicted using the full dataset. Links predicted by MM are exactly the same as those with the full data (see Fig. 3). Compared to SM, GM incorrectly swaps TOC1→GI for PRR7→GI while SM misses two links (TOC1→GI and PRR7→TOC1), but is still able to correctly predict two further links (CCA1→TOC1 and PRR7→CCA1). This shows that even in the case when less data is available, SM is still able to infer the degrees of freedom and outperforms MM/GM. In a scenario where few replicates and time points are available, it may be advisable to modify SM making the degrees of freedom common to all genes.

5 DISCUSSION

We demonstrate here that the uncertainty conveyed in repeated measurements of time course gene expression data can have a strong effect when estimating a gene interaction network. In the case of a linear autoregressive network specification, not accounting for this uncertainty leads to attenuation of the autoregressive coefficients and overestimation of the precision of these estimates. This in turn can affect the network topology retrieval. To address this issue, we propose a model that explicitly includes this variability.

Our modelling is fully Bayesian, with the true gene expression unobserved and thus inferred. Inference of these expression values draws information from both the surrogate measurements and the linear AR(1) process assumed for the gene network interaction, with the influence of each source weighted by the relative value of the AR(1) precision, $\boldsymbol{\lambda}$, and the measurement precision, $\boldsymbol{\tau}$. When $\boldsymbol{\tau}$ is relatively small, inference on the expression values will be predominately determined by the AR(1) part of the model. In the case where there is little information in the data about the regulatory process, the prior on $\boldsymbol{\lambda}$ must thence be carefully elicited. We provide a benchmark prior for the kinds of datasets arising from microarrays.

Our model accommodates simultaneous inference of the network topology along with the interaction coefficients. We showed that attenuation of the network coefficients as well as the underestimation of the variability of these estimates is systematic. Such behaviour is then passed on to the estimated link probabilities, yielding a more concentrated distribution of link probabilities towards either one or zero. The retrieved networks are obtained by setting an (arbitrary) threshold on the posterior link probabilities, and thus

Table 3. Link prediction counts using the *Arabidopsis* circadian clock microarray sub-sampled datasets with three replicates

	MM					GM					SM				
	CCA1	TOC1	PRR7	GI	CHE	CCA1	TOC1	PRR7	GI	CHE	CCA1	TOC1	PRR7	GI	CHE
CCA1	0	0	1	**17**	0	0	0	2	**24**	0	0	0	**21**	7	0
TOC1	1	0	1	**29**	0	4	0	1	**22**	0	**15**	0	7	5	0
PRR7	1	**29**	0	3	0	0	**30**	0	5	0	0	**30**	0	3	0
GI	1	**29**	1	0	0	0	11	**20**	0	0	0	9	**24**	0	0
CHE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Bold values are those links that were predicted present in more than one half of the 30 subsets, using a threshold of 1/2. Genes in columns are regulators and rows are regulatees.

Downloaded from bioinformatics.oxfordjournals.org at University of Warwick on May 17, 2011

are only point estimates of the network topology. Thus, the final effect on the inferred network topology is case dependant. This is because, while searching for regulatory dependencies, the estimation process implicitly compares alternative parenthood configurations and the inclusion/exclusion of a link depends on the specific dynamics of each gene and its relative variability. However, the distribution of these probabilities can be quite different, providing the experimentalists with a more accurate description of the uncertainty contained in the model fitting, and thus will be better informed when designing further experiments.

High-throughput technologies yield noisy measurements, with the noise distribution typically exhibiting heavier than Gaussian tails. Not accounting for this behaviour can also have a negative impact when performing inference on the interaction coefficients of the network. In our examples, we showed that the inferred topology with the synthetic datasets improved when using SM over GM and MM. Even though there is not a definite network as yet for the *in vivo* dataset, the inferred topology showed no difference between MM and GM for the selected threshold, while the posterior distributions of the degrees of freedom indicate heavy tails for all but one of the genes, indicating significant outliers in the data and thus suggesting the MM/GM predictions are questionable.

Funding: Warwick Systems Biology Doctoral Training Centre (to E.R.M.); BBSRC grant BB/F003498/1 (to M.A.J.); Experimental data was provided by KJ Denby through the PRESTA Project, grant number BB/F005806/1.

Conflict of Interest: none declared.

REFERENCES

Bansal,M. *et al.* (2007) How to infer gene networks from expression profiles. *Mol. Syst. Biol.*, **3**, 78.

Brody,J.P. *et al*. (2002) Significance and statistical errors in the analysis of DNA microarray data. *Proc. Natl Acad. Sci. USA*, **99**, 12975–12978.

Cao,J. and Zhao,H. (2008) Estimating dynamic models for gene regulation networks. *Bioinformatics*, **24**, 1619–1624.

Carroll,R.J. *et al.* (2006) *Measurement error in nonlinear models: A modern perspective*, 2nd edn. Chapman & Hall/CRC, Boca Raton.

Friedman,N. *et al.* (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, **7**, 601–620.

Friedman,N. (2004) Inferring cellular networks using probabilistic graphical models. *Science*, **303**, 799–805.

Fujita,A. *et al.* (2009) The impact of measurement errors in the identification of regulatory networks. *BMC Bioinformatics*, **10**, 412.

Fuller,W.A. (1987) *Measurement error models*. Wiley, New York.

Gustafson,P. (2004) *Measurement error and Misclassification in Statistics and Epidemiology. Impacts and Bayesian adjustments*. Chapman & Hall/CRC, Boca Raton.

Gustafson,P. (2005) On model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables. *Stat. Sci.*, **20**, 111–140.

Hache,H. *et al.* (2009) Reverse engineering of gene regulatory networks: a comparative study. *EURASIP J. Bioinform. Syst. Biol.*, **2009**, 1–12.

Hongqiang,L. *et al.* (2005) Inferring gene transcriptional modulatory relations: a genetical genomics approach. *Human Mol. Genet.*,, **14**, 1119–1125.

Hoops,S. *et al.* (2006) COPASI: a COmplex PAthway SImulator. *Bioinformatics*, **22**, 3067–3074.

Lewin,A. *et al.* (2007) Fully Bayesian mixture model for differential gene expression: simulations and model checks. *Stat. Appl. Genet. Mol. Biol.*, **6**, 36.

Locke,J.C.W. *et al.* (2006) Experimental validation of a predicted feedback loop in the multi-oscillator clock of *Arabidopsis thaliana*. *Mol. Syst. Biol.*, **2**, 59.

Murphy,K. and Mian,S. (1999) Modelling gene expression data using dynamic Bayesian networks. *Technical report*, Computer Science Division, University of California, Berkeley.

Perrin,B. *et al.* (2003) Gene network inference using dynamic Bayesian networks. *Bioinformatics*, **19**, ii138–ii148.

Poirier,D.J. (1998) Revising beliefs in nonidentified models. *Econometric Theory*, **14**, 483–509.

Pruneda-Paz,J.L. *et al.* (2009) A functional genomics approach reveals CHE as a component of the *Arabidopsis* circadian clock. *Science*, **323**, 1481–1485.

Robertson,F.C. and Webb,A.A.R. (2009) Revolutionary functional genomics liberates CHE. *Nat. Chem. Biol.*, **5**, 276 – 277.

Schmid,C.H. *et al.* (1994) Incorporating measurement error in the estimation of autoregressive models for longitudinal data. *J. Stat. Plan. Inference*, **42**, 1–18.

Smith,M. and Kohn,R. (1996) Nonparametric regression using Bayesian variable selection. *J. Econom.*, **75**, 317–343.

Yu,J. *et al.* (2004) Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, **20**, 3594–3603.

Zou,M. and Conzen,S.D. (2005) A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, **21**, 71–79.

Downloaded from bioinformatics.oxfordjournals.org at University of Warwick on May 17, 2011

On reverse engineering of gene interaction networks using time course data with repeated measurements

Supplementary material

Morrissey, E.R., Juárez, M.A.
Systems Biology Centre
University of Warwick, Coventry CV4 7AL, UK

K. J. Denby
Warwick HRI, Systems Biology Centre
University of Warwick, Coventry CV4 7AL, UK

Burroughs, N.J.
Systems Biology Centre,
University of Warwick, Coventry CV4 7AL, UK

These supplementary materials are presented in order of appearance within the main paper. For completeness, we briefly recall the model specification.

1S The model

Denote by y_g^t the expression level of gene $g = 1, \dots, G$, measured at time $t = 1, \dots, T$. We model the interaction network as a linear AR(1) process,

$$y_g^{t+1} = \mu_g + \sum_{j=1}^{G} \tilde{\beta}_{jg} y_j^t + \varepsilon_g^t \,, \tag{1}$$

where μ_g is the basal expression level of gene g; $\tilde{\beta}_{jg} = \gamma_{jg}\beta_{jg}$ measures the influence of gene j on gene g, with $\beta_{jg} \in \mathbb{R}$ and $\gamma_{jg} = 1$ if j regulates g and $\gamma_{jg} = 0$ otherwise; finally, ε_g^t is an idiosyncratic error term, centred at zero and with precision parameter λ_g, typically assumed to be Gaussian. We augmented the model with the parenthood (link) indicator variables $\Gamma = \{\gamma_{jg}\}$ which will be the basis for estimating the network topology.

In order to account for the additional uncertainty when having repeated measurements we assume that the regulation process can be captured by (1), but instead of actually observing y_g^t, we have noisy measurements, x_{gr}^t, such that

$$x_{gr}^t = y_g^t + \eta_{gr}^t \,, \qquad r = 1 \dots, R \,, \tag{2}$$

with η_g^t a zero mean error measurement term, with precision parameter τ_g, independent for all t, g, r. This error term is frequently assumed Gaussian; however, given that the measurement process can potentially produce outliers, we will use a Student-t specification, $\mathrm{St}\big(\eta_g^t \mid 0, \tau_g, \nu\big)$, such that $\mathrm{Var}\left[\eta_g^t\right] = \nu\, \tau_g^{-1}/(\nu - 2)$ provided the degrees of freedom, $\nu > 2$.

2S THE PRIOR

The likelihood for the Student-t measurement AR(1) model is,

$$\ell(\Theta; X) = \prod_{g=1}^{G} \prod_{t=1}^{T} \prod_{r=1}^{R} \mathrm{N}\left(y_g^{t+1} \mid \mu_g + \tilde{\boldsymbol{\beta}}_g \boldsymbol{y}^t, \lambda_g\right) \mathrm{N}\left(x_{gr}^t \mid y_g^t, \omega_{gr}^t \tau_g\right) \mathrm{Ga}\left(\omega_{gr}^t \mid \nu/2, \nu/2\right) . \tag{3}$$

Where $Y = \left\{y_g^t\right\}$ are the unobserved expression levels, $X = \left\{x_{gr}^t\right\}$ denote their surrogate measurements and $\Theta = \{\boldsymbol{\mu}, B, \Gamma, \boldsymbol{\lambda}, \boldsymbol{\tau}, \nu\}$ collects all the parameters involved, with $\boldsymbol{\mu} = \{\mu_1, \dots, \mu_G\}$; $B = \left\{\boldsymbol{\beta}_1', \dots, \boldsymbol{\beta}_G'\right\} \in \mathbb{R}^{G \times G}$ and $\boldsymbol{\beta}_g = \{\beta_{1g}, \dots, \beta_{Gg}\}$; $\Gamma = \{\gamma_{ij}\}$; $\boldsymbol{\lambda} = \{\lambda_1, \dots, \lambda_G\}$; and $\boldsymbol{\tau} = \{\tau_1, \dots, \tau_G\}$.

We specify a product form (independent) prior,

$$\pi(\Theta) = \pi(\rho)\, \pi(\nu) \left[\prod_{g=1}^{G} \pi(\mu_g)\, \pi(\boldsymbol{\beta}_g)\, \pi(\lambda_g)\, \pi(\tau_g)\, \pi(\boldsymbol{\gamma}_g) \right] , \tag{4}$$

where,

$$\pi(\mu_g) = \mathrm{N}(\mu_g \mid 0, k) , \tag{5}$$

$$\pi(\boldsymbol{\beta}_g) = \mathrm{N}_G\left(\boldsymbol{\beta}_g \mid \mathbf{0}, k_\beta I\right), \quad g = 1, \dots, G , \tag{6}$$

$$\pi(\lambda_g) = \mathrm{Ga}(\lambda_g \mid a_\lambda, b_\lambda) , \tag{7}$$

$$\pi(\tau_g) = \mathrm{Ga}(\tau_g \mid a_\tau, b_\tau) , \tag{8}$$

$$\pi(\boldsymbol{\gamma}_g \mid \rho) = \prod_{j=1}^{g} \mathrm{Ber}(\gamma_{jg} \mid \rho), \quad g = 1, \dots, G , \tag{9}$$

$$\pi(\rho) = \mathrm{Be}(\rho \mid a_\rho, b_\rho) , \tag{10}$$

$$\pi(\nu) = \mathrm{Ga}(\nu \mid a_\nu, b_\nu) . \tag{11}$$

Given that the data is standardised before performing the estimation (zero mean and unitary standard deviation for each time series), we set $k_\mu = k_\beta = 1/4$; i.e. the prior variance of any component of $\boldsymbol{\mu}$ and B is four. In our experience, this is typically not over-informative for microarray data.

As mentioned in the paper, when repeated measurements are available it is easier to estimate $\boldsymbol{\tau}$ than $\boldsymbol{\lambda}$. Thus, we set $\{a_\tau, b_\tau\} = \{2, 1/100\}$ which is renders a rather flat prior with mode at 100 and variance of 20000.

For the autoregressive precision $\boldsymbol{\lambda}$, we used $\{a_\lambda, b_\lambda\} = \{1/10, 1/10\}$. Thus setting the prior mean at one and the variance at 10. The mode now does not exist.

Derived from the conditions given in the paper: $\mathrm{P}\left[\nu \leq 30\right] \approx 0.6$ and $\mathrm{Mode}[\nu] = 15$, it is straightforward to verify that $\{a_\nu, b_\nu\} = \{3.5, 0.15\}$.

In the absence of any prior information, we treat ρ as the probability of any given link to be present and thus use the corresponding reference prior, $\mathrm{Be}(\rho \mid 1/2, 1/2)$ (Bernardo and Smith, 1994, p. 315) .

3S THE SAMPLER

AR(1) Precisions The full conditional of $\lambda_g, g = 1, \ldots, G$ is given by

$$\pi(\lambda_g \mid -) \propto \lambda_g^{T/2+a_\lambda-1} \exp\left[-\lambda_g\left(b_\lambda + \frac{1}{2}\left(y_g^{t+1} - \mu_g - \mathbf{y}^t\widetilde{\boldsymbol{\beta}}_g\right)'\left(y_g^{t+1} - \mu_g - \mathbf{y}^t\widetilde{\boldsymbol{\beta}}_g\right)\right)\right]$$

and thus can be sampled from a gamma distribution.

Constant term μ_g is conditionally Gaussian, with mean and precision

$$m_g = \frac{\overline{y}_g^{t+1} - \overline{\mathbf{y}}^t\widetilde{\boldsymbol{\beta}}_g}{\lambda_g + k_\mu/T} \qquad \text{and} \qquad \tau'_\mu = k_\mu + T\,\lambda_g \ ,$$

respectively, where $\overline{y}_g^{t+1} = T^{-1}\sum_t y_g^{t+1}$ and $\overline{\mathbf{y}} = T^{-1}\sum_t \mathbf{y}^t$.

Connectivity The overall connectivity, ρ, is sampled from a $\text{Be}\big(\rho \mid S + a_\rho, G^2 + b_\rho - S\big)$, with $S = \sum_{i,j=1}^{G}\gamma_{ji}$.

Measurement precision For each gene $g = 1, \ldots, G$, the measurement precision, τ_g are updated from a gamma distribution $\text{Ga}\big(\tau_g \mid a'_\tau, b'_\tau\big)$ with

$$a'_\tau = R\,T/2 + a_\tau \qquad \text{and} \qquad b'_\tau = b_\tau + \frac{1}{2}\sum_{t=1}^{T}\sum_{r=1}^{R}\omega_{gr}^t\left(x_{gr}^t - y_{gr}^t\right)^2 .$$

Degrees of freedom We use a Metropolis-within-Gibbs strategy to draw a new value, $\nu^{(m)}$, with a gamma proposal with its mean fixed at the previous draw, $\nu^{(m-1)}$. We control for the acceptance rate to lie around 1/3 by tuning the proposal's coefficient of variation, cv. Thus, we propose a new $\nu^{(m)}$ from

$$\text{Ga}\big(\nu^{(m)} \mid cv^{-2}, cv^{-2}/\nu^{(m-1)}\big).$$

Coefficients and link probabilities The update of each indicator variable γ_{jg} is performed jointly with all the corresponding coefficients

$$\boldsymbol{\beta} : \boldsymbol{\beta}^a \to \boldsymbol{\beta}^b \qquad \text{and} \qquad \gamma : 0 \to 1$$

with acceptance probability

$$\alpha = \min\left\{\frac{\pi(\widetilde{\boldsymbol{\beta}}^b)}{\pi(\widetilde{\boldsymbol{\beta}}^a)}\,\frac{q(\boldsymbol{\beta}^a \mid \gamma^a)\,q(\gamma^a)}{q(\boldsymbol{\beta}^b \mid \gamma^b)\,q(\gamma^b)}, 1\right\} \ ,$$

where the subscripts have been removed for clarity. Given that we propose γ symmetrically, $q(\gamma^a)/q(\gamma^b) = 1$. The Hastings ratio is then

$$\frac{q(\boldsymbol{\beta}^a \mid \gamma^a)}{q(\boldsymbol{\beta}^b \mid \gamma^b)} = \frac{\rho}{1-\rho}\,k_{\boldsymbol{\beta}}^{1/2}\,\frac{\exp\left[\frac{1}{2}\mu_{\boldsymbol{\beta}}^b\Sigma_{\boldsymbol{\beta}}^{-1b}\mu_{\boldsymbol{\beta}}^b\right]\left|\Sigma_{\boldsymbol{\beta}}^b\right|^{1/2}}{\exp\left[\frac{1}{2}\mu_{\boldsymbol{\beta}}^a\Sigma_{\boldsymbol{\beta}}^{-1a}\mu_{\boldsymbol{\beta}}^a\right]\left|\Sigma_{\boldsymbol{\beta}}^a\right|^{1/2}} .$$

with $\Sigma_{\cdot}$ the covariance matrix obtained by considering only the relevant gene expression vectors. For the opposite move *i.e.* switching a link off, we use the reciprocal of the ratio above.

Non-observables These are drawn from a Gaussian distribution, $\mathrm{N}\big(y_g^t \mid m_g^t, p_g^t\big)$, with location

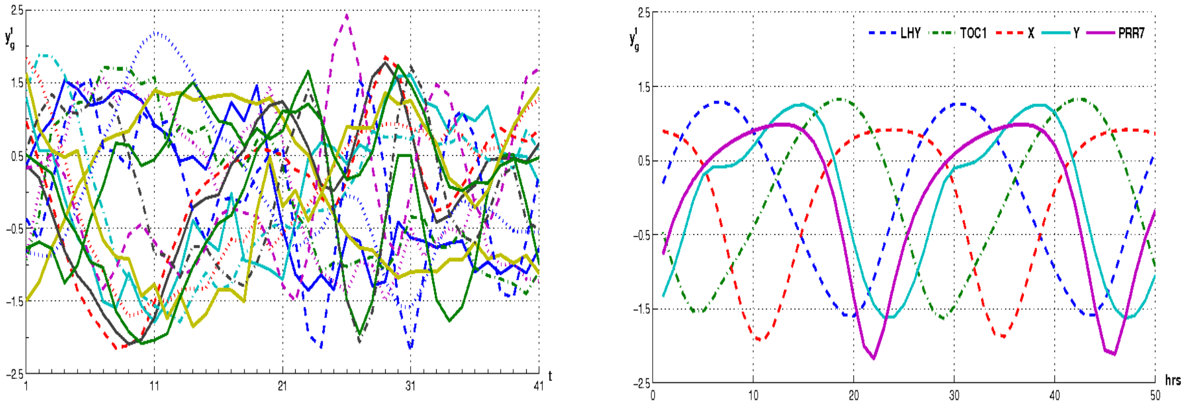
$$m_g^t = \frac{\lambda_g m_{\mathrm{AR}} + \tau_g m_{\mathrm{meas}}}{p_g^t} \qquad \text{and precision} \qquad p_g^t = \lambda_g\Big(1+\tilde{\beta}_{gg}^2\Big) + \tau_g \sum_{r=1}^{R} \omega_{gr}^t \,,$$

where

$$m_{\mathrm{AR}} = \sum_{i\neq g} \tilde{\beta}_{ig}\big(y_i^{t-1} - \tilde{\beta}_{gg}\, y_i^t\big) + \tilde{\beta}_{gg}\big(y_g^{t-1} + y_g^{t+1}\big) \qquad \text{and} \qquad m_{\mathrm{meas}} = \sum_{r=1}^{R} \omega_{gr}^t x_{gr}^t \,.$$

4S DATA SETS

Time traces of the data sets used in Section 4 of the paper. The linear *in silico* data, Figure-S 1a, comprises 16 genes measured at 41 time points. The ODE data has five genes and 50 measurements in time, Figure-S 1b.



(a) *Linear interactions*

(b) *ODE data*

Figure-S 1. In silico *data sets. Traces of the noiseless synthetic linear and ODE, non-linear data sets.*

The *Arabidopsis* data set has 5 genes with 24 time points and four repetitions.

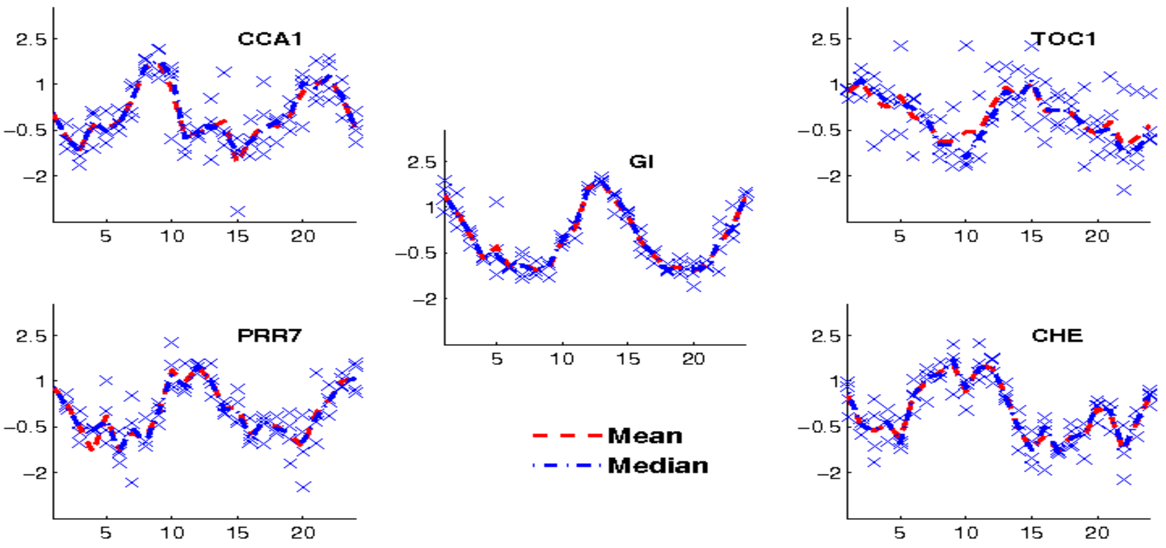


Figure-S 2. *Circadian clock related genes in* Arabidopsis thaliana. *Gene expression repeated measurements. The mean (dashed) and median (dot-dashed) of each time point are plotted as time series.*

4.1S Mean cross entropy

The score used in the paper, MxE, is simply the Kullback-Liebler divergency from any specific link to the true network configuration, $KL(\widehat{\gamma}_{ij} \mid p_{ij})$,

$$KL(\widehat{\gamma}_{ij} \mid p_{ij}) = p_{ij} \log \frac{p_{ij}}{\widehat{\gamma}_{ij}} + (1 - p_{ij}) \log \frac{1 - p_{ij}}{1 - \widehat{\gamma}_{ij}}$$

averaged over all possible links, $i, j, = 1, \ldots, G$, where $p_{ij} = 1$ if the link is present and zero otherwise; with the convention of $0 \log 0 = 0$.

4.2S *In silico* data

As expected, when a large number of replicates are available the three models yield similar networks, for a given threshold. However, the effect of overestimation in the regression coefficients posterior precisions is apparent when looking at the probabilities predicted by MM: again, they are less disperse than those predicted by either GM or SM.

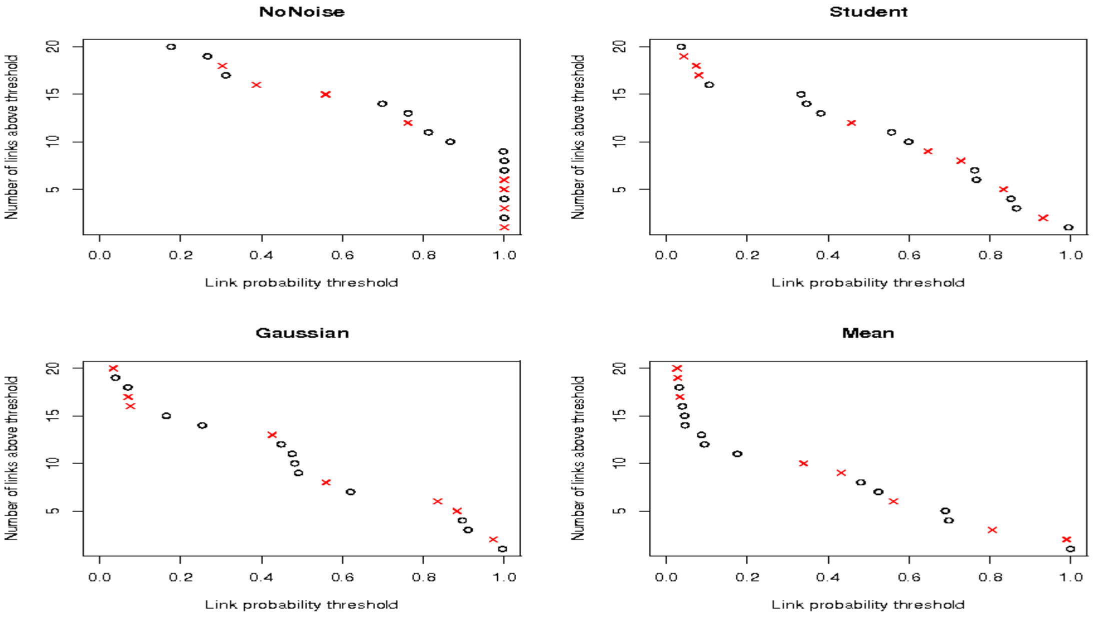


Figure-S 3. *ODE circadian clock* in silico *data set with $R = 20$ replicates. Number of links predicted present in the network versus posterior link probabilities for each model considered. A link present in the ODE model is highlighted with a cross.*

4.3S *In vivo* data

To have a more or less representative sample, we calculated all the possible subsamples with three replicates, $\mathcal{P}_3$, and then measured the Euclidean distance between the standard deviations of the original data and $\mathcal{P}_3$. These were classified into large, medium and small, based on their empirical distribution and ten series were selected from each region. For the 1-replicate case we used the Euclidean distance between the mean of the four replicate data set and the single data set. Interestingly we found no apparent effect of the Euclidean distance in the retrieved topologies and therefore we joined them when calculating the counts tables.

REFERENCES

Bernardo, J.M. and Smith, A.F.M. (1994) *Bayesian Theory.* Chichester: John Wiley & Sons.

Chapter 4

GRENITS: An R/Bioconductor Package

The network inference models described in the previous sections were published along with the papers. The models were written in MATLAB, a high level programming language. While MATLAB can be efficient, it suffers greatly with loops, an unavoidable element of Markov Chain Monte Carlo (MCMC) samplers. As well as this, it is not open source, free nor is it the programming language of choice for most gene expression analysis software.

In this paper we present GRENITS, a R/Bioconductor package that collates the inference models described in the two previous chapters. That is, the two models with measurement errors (Student-t and Gaussian), the non-linear model and the linear model. The models have been rewritten in `C++` and where possible, the algebra has been rewritten in a more efficient manner. These changes have led to speed increases of up to 60 fold compared to the original MATLAB code.

The package has been designed to cater for both experienced users and casual users. To this end, the package offers some basic analysis functions as well as "reasonable" default parameters for the prior distributions and MCMC sampler. That is to say it is possible to run a full MCMC inference run with a single function and two arguments. The output can then be analysed with a single function with one argument. The analysis includes basic convergence checks, as well as plots and files that we have found to be useful in our analyses. For the experienced user we also make available functions to modify and plot priors, as well as functions to read MCMC chains for further analysis.

A further improvement is that there is an option for fixing connections in the network to be on/off. This feature can be used to input a list of regulators, allowing for the analysis of networks of a significantly larger size. Fixing network links can also be useful, for example, to fix a full network and use the non-linear model to infer the functional form of the interactions.

4.1 Contribution

This paper is a submitted paper. The initial idea was formulated by Morrissey. Morrissey contributed both the `C++` code and the R code. Morrissey submitted the package to Bioconductor and addressed the issues raised by the reviewer.

Morrissey wrote the draft of the paper. Juárez helped with testing. Both Juárez and Burroughs contributed with suggestions both for the package and the paper.

BIOINFORMATICS

Vol. 00 no. 00 2011
Pages 1–2

GRENITS: An R/Bioconductor package for the inference of gene regulatory networks using Bayesian Networks

E. R. Morrissey [1,2]*, M. A. Juárez [3], STREAM Consortium and N. J. Burroughs [2]

[1] Department of Oncology, University of Cambridge, CRUK Cambridge Research Institute, Cambridge CB2 0RE, UK
[2]Warwick Systems Biology Centre, University of Warwick, Coventry CV4 7AL, UK
[3]School of Mathematics and Statistics, University of Sheffield, Sheffield S3 7RH, UK

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: GRENITS is an R/Bioconductor package for the inference of gene regulatory networks using time-series gene-expression data. The package implements four published network inference statistical models using Dynamic Bayesian Networks and Bayesian Variable Selection. The four methods include a linear interactions model, two models that account for experimental noise under linear interactions (useful when replicates are available) and a non-linear interactions model. The implementations are in C++ with an R interface and optimised for speed, taking respectively seconds to hours to run on typical networks of size 10 to 1000 nodes. The package is intended to be used by both users with a background in Bayesian Inference, as well as casual users.

Availability: GRENITS is available as an R/Bioconductor package from http://bioconductor.org/packages/2.9/bioc/html/GRENITS.html.

Contact: edward.morrissey@gmail.com

1 INTRODUCTION

With the ability to measure the expression of the whole genome, there is significant interest in studying biological processes using a top-down approach (Nieselt *et al.*, 2010; Amit *et al.*, 2009). Time-series are particularly attractive given their ability to capture causal events, the relative expression profiles between genes indicating the presence of a potential regulatory interaction (mediated through expressed proteins). A number of computational/statistical approaches have been proposed to produce plausible hypothesis about which genes are key regulators and how they contribute to the biological process, as well as how they may affect (influence, interact with) each other. One such approach is network inference (De Smet *et al.*, 2010), where a mathematical/statistical model of how the genes interact is fitted to the data, allowing a network structure to be predicted.

Here we present GRENITS, an R/Bioconductor package implementing four previously published statistical models (Morrissey *et al.*, 2010, 2011) that allow network inference from time-series expression data. The models are dynamic Bayesian network models based on regression between time points. There are three linear models available, a linear regression model with Gaussian noise appropriate for longitudinal time-series, e.g. sampling from a fermentor, and two with measurement error models included where replicates are required. These measurement error models are appropriate for time-series studies where there is no continuity between sequential samples, e.g. sampling involves different organisms under a synchronised experimental protocol. The fourth model implements non-linear dependencies between genes. The fastest and simplest of the four models is the linear model. Here the regulation function is assumed linear. This model is particularly useful for short time-series (10 to 20 time-points) and capable of handling a large number of genes. For the case where repeated measurements are available, the package has two models that account for measurement errors (Morrissey *et al.*, 2010). As in the previous case, the interactions are taken to be linear. Both measurement error models accommodate biological noise in the same way: we assume that the replicates represent noisy realisations of a true unmeasured gene expression value whose dynamic is then driven by the network. The two error models differ in the underlying measurement error distribution assumptions: either Gaussian or Student-t (heavy tailed in order to capture outliers). The true expression values are modelled as hidden variables. The number of replicates per time point do not have to be identical. Also available is a model that accounts for non-linear interactions (Morrissey *et al.*, 2011). Non-linear interactions are constructed using B-Splines, a flexible non-parametric type of function. Using these functions, the model infers simultaneously the functional form of the interactions along with the structure of the network.

All of the models rely on Markov chain Monte Carlo (MCMC) algorithms to sample from the posterior. Thus, all parameters are inferred together; this includes the network structure variables. The network structure variables tend to have the poorest mixing; thus to improve the speed of the samplers we use a novel block sampling step described in (Morrissey *et al.*, 2010).

*to whom correspondence should be addressed

© Oxford University Press 2011.

2 IMPLEMENTATION

Although MATLAB implementations of the inference algorithms were distributed with the papers, the samplers rely heavily on loops– a weak point of high level programming languages. The samplers implemented in GRENITS are written in C++ and called from R, making the models considerably faster than their MATLAB equivalents whilst retaining the advantages of the Bioconductor platform. For matrix algebra, Armadillo (Sanderson, 2010) the fast C++ library was used. A further speed increase was achieved by optimising the code structure and matrix algebra where possible. All in all, this has lead to speed improvements of up to 60 fold with respect to the MATLAB versions.

2.1 User Interface

The package is intended to be used by both experienced and casual users. For casual users, we provide default parameter settings both for the priors and the MCMC algorithms, that in our experience, have worked well for a range of data sets. This means that the sampler should be usable 'off the shelf' on most data sets. Caution however is recommended, as with all MCMC samplers as results are only trustworthy if the chains have converged. We therefore strongly recommend running parallel analyses (default is 2 chains) and provide a function for the analysis (comparison) of those chains. The function provides basic convergence plots and issues a warning to the user if there is evidence for a lack of convergence; a longer burn-in is then advised. It also produces some plots with information on the network structure, and outputs files with link probabilities that can be read either into R or used with network analysis software such as cytoscape (Shannon *et al.*, 2003).

For the more experienced user, functions to adjust and plot the prior distributions are available, along with functions to read the MCMC chains for further analyses.

2.2 Speed and Network Size

Although the R/Bioconductor implementation of the algorithms is much faster than the MATLAB versions, the run time is still considerable (12+ hrs depending on number of time points and model) on networks over 1000 genes. However, often when inferring a network the identity of potential regulators is known. Regulating genes can be either transcription factors or proteins able to indirectly affect the expression of other genes. GRENITS allows this information to be used by letting the user provide a list of potential regulators and only allowing these genes to be parents in the network. This considerably reduces the algorithm runtime. For instance, the linear model was used to infer a network comprising 988 genes with 35 regulators, 41 time points, in about 3 hours per chain (run on a dual-core 2.53GHz laptop).

2.3 Limitations

Sampling times need to be spaced at equal intervals, ideally with a time spacing that is finer than the time scale of the regulatory dynamics. The measurement error models can deal with missing data since the number of replicates need not be identical per time point. Individual missing data on genes is not dealt with. Provided these are small in number linear interpolation can be used.

3 DISCUSSION

Here we have presented GRENITS, an R/Bioconductor package for the inference of regulatory networks from gene expression data, such as that available from microarrays, quantitative PCR, nanostring and RNAseq. The package reimplements previously published methods, integrating them into the same common framework. This implementation is much faster than the original MATLAB versions thanks to a faster programming language, better code structure and optimised matrix algebra. As well as this, all functions are fully documented and a number of convenient functions to read, analyse and plot the output from the samplers are provided. Further, prior information on the potential regulators can be utilised, allowing realistic data sets to be run in an acceptable amount of time. The package is freely available from Bioconductor.

ACKNOWLEDGEMENT

Funding: Both ER Morrissey and MA Juárez were funded by BBSRC, grant number BB/F003498/1 under the SysMO Initiative.

REFERENCES

Amit, I., Garber, M., Chevrier, N., Leite, A., Donner, Y., Eisenhaure, T., Guttman, M., Grenier, J., Li, W., Zuk, O., Schubert, L., Birditt, B., Shay, T., Goren, A., Zhang, X., Smith, Z., Deering, R., McDonald, R., Cabili, M., Bernstein, B., Rinn, J., Meissner, A., Root, D., Hacohen, N., Regev, A. (2009) Unbiased Reconstruction of a Mammalian Transcriptional Network Mediating Pathogen Responses, *Science*, **326**, 257-263.

De Smet, R. and Marchal, K. (2010) Advantages and limitations of current network inference methods, *Nat Rev Microbiol*, **8**, 717-729.

Morrissey, E.R., Juarez, M.A., Denby, K.J. and Burroughs, N.J. (2010) On reverse engineering of gene interaction networks using time course data with repeated measurements, *Bioinformatics*, **26**, 2305-2312.

Morrissey, E.R., Juarez, M.A., Denby, K.J. and Burroughs, N.J. (2011) Inferring the time-invariant topology of a non-linear sparse gene regulatory network using fully Bayesian spline autoregression, *Biostatistics*, (to appear).

Nieselt, K., Battke, F., Herbig, A., Bruheim,P., Wentzel, A., Jakobsen, O., Sletta, H., Alam, M., Merlo, M., Moore, J, Omara, W. , Morrissey, E.R. , Juarez, M.A., Rodriguez-Garcia, A., Nentwich, M., Thomas, L., Iqbal, M., Legaie, R., Gaze, W., Challis, G., Jansen, R., Dijkhuizen, L., Rand, D., Wild, D., Bonin, M., Reuther, J., Wohlleben, W., Smith, M., Burroughs, N., Martin, J., Hodgson, D., Takano, E., Breitling, R., Ellingsen, T., Wellington, E. (2010) The dynamic architecture of the metabolic switch in Streptomyces coelicolor, *BMC Genomics*, **11**.

Sanderson C. (2010) Armadillo: An Open Source C++ Linear Algebra Library for Fast Prototyping and Computationally Intensive Experiments, *NICTA Technical Report*

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks, *Genome Research*, **13**, 2498-2504.

Chapter 5

Network Inference for STREAM Data

The following paper addresses the analysis of the STREAM data. The focus of the paper is on the regulon of *phoP*, a key regulator in the response to phosphate starvation. The data used to study *phoP*'s regulon were three high resolution gene expression time-series. Two of the time-series were measured in exactly the same experimental conditions of phosphate starvation, with one of the experiments being performed with wild type *Streptomyces coelicolor* and the other with a knock out (KO) mutant missing *phoP*. The third time-series was generated by subjecting a wild-type strain to glutamate starvation conditions. All three time-series trigger metabolic switching and the secretion of antibiotics, though *phoP* is only active in the wild type phosphate limitation experiment.

By comparing the two phosphate starvation time-series, it was possible to find a list of genes showing evidence of *phoP* regulation. The glutamate limitation time-series was then used to categorise the type of regulation.

Several computational methods were used to predict *phoP*'s regulon: two types of clustering, a network inference algorithm and two variants of a sequence analysis algorithm. For the clustering and network inference methods only the wild type phosphate limitation time-series was used. The computational methods were then benchmarked against the across-time-series predictions.

The networks fitted in the previous chapters were small networks of between five and twenty genes. For the STREAM time-series we fitted a network to a whole-genome dataset with close to 8000 genes. Given the speed of the methods described in the previous chapters, fitting a full network to a dataset of this size would be impractical. In order to fit a network in a reasonable amount of time we first filtered the dataset to those genes that appeared as dynamic during the experiment. Next we identified the genes known to be regulators and clustered these regulators so as to have a set of dynamically distinct regulators/clusters. A gene was chosen from each cluster to represent the whole cluster. Finally we used GRENITS to fit a linear network restricting regulation to the list of regulators.

Of the computational predictions the clustering methods performed best, followed by the sequence analysis methods and finally the network inference predictions. When including co-clustered genes, the network inference method showed a considerable improvement, indicating that a number of the inferred relations are different from the co-clustered genes.

5.1 Network Inference Pre-analysis

As mentioned in the introduction (section 1.7.2) the network inference methods presented in this thesis all have the self interaction term fixed on. The decision to do this, as well as consider co-clustered genes as potential regulators, arose from the pre-analysis of the STREAM data.

As an initial exploratory analysis, we extracted from the first time-series a set of genes with known interactions. The genes chosen were two TFs (*redD* and *redZ*) and a non-TF gene (*redM*) belonging to one of the antibiotic biosynthesis pathways [Williamson et al., 2006]. We fitted networks using the linear and splines model with no interactions fixed and found that of the inferred networks, the splines network was slightly more accurate. We then fitted a full network to the data and found connections being predicted between genes in the same operon (*phoP,phoR*). When looking in more detail at the BVS parameters, we found that regulation was being switched between just a self-interaction term and just a regulation term by an operon gene. In other words, the data can be equally well explained by using the self-interaction term or by predicting regulation by a gene that is co-expressed. Given that a self interaction term is expected (mass conservation) and that co-clustered genes, a very common occurrence, can be spuriously predicted as regulators, we opted to fix the self-interaction on.

When we fitted networks to the three gene dataset (*redD*, *redZ* and *redM*), we found that fixing the self-interaction term degraded the quality of the inferred network, as the expression of the three genes is very similar. As expected, for the full dataset we found less connections and loss of regulation of co-clustered genes, including genes we knew from the literature were regulated by co-expressed genes. This lead us to include information of all clustered genes within the predictions.

5.2 Contribution

This paper is an early-stage draft. Morrissey contributed the inferred network. Morrissey did not contribute any of the writing.

Morrissey was involved in the pre-analysis stage of this paper. This involved analysis of clustering and assessing the performance of the inferred network by comparing the WT time-series to the $\Delta phoP$ for the predicted genes. Morrissey was also involved in discussing the biological interpretation of the clusters and inferred networks with Hodgson.

Morrissey contributed suggestions on paper content and structure.

Coregulation profiling of causal, correlation and bioinformatic predictions of PhoP dependent genes in *S. coelicolor* reveals methodology biases

Miguel A. Juárez[1], Alexander Wentzel[2,3], Roxane Legaie[1], Mudassar Iqbal[1], Edward R. Morrissey[1], Jonathan D. Moore[1], Antonio Rodríguez-García[5], Juan F. Martín[5], Øyvind M Jakobsen[2], Kay Nieselt[6], Trond E. Ellingsen[2], STREAM[7] consortium, *Nigel Burroughs[1].

[1]Warwick Systems Biology Centre, University of Warwick, Coventry, CV4 7AL, UK
[2]Department of Biotechnology, SINTEF Materials and Chemistry, Sem Sælandsvei 2a, N-7465 Trondheim, Norway
[3]Department of Biotechnology, Norwegian University of Science and Technology (NTNU), Sem Sælandsvei 6-8, N-7491 Trondheim
[5]Instituto de Biotecnología de León, INBIOTEC, Parque Científico de León, Av. Real 1, 24006 León, Spain, and Área de Microbiología, Universidad de León, Spain
[6]Center for Bioinformatics Tübingen, Department of Information and Cognitive Sciences, University of Tübingen, Sand 14, D-72076 Tübingen, Germany
[7]STREAM consortium www.sysmo.net/index.php?index=62, BMC Genomics. 2010, 11:10.

Corresponding author. N.J.Burroughs@warwick.ac.uk

Unravelling cross-regulation and elucidating regulatory networks is a major theme of systems biology. Many methods have been advocated; however there is no consensus on which methods are most accurate, or if performance is species or condition specific. We compare 3 computational methods to ascertain if there is a bias towards particular forms of coregulation or functional categories. By using comparative dynamic signatures on multiple high time resolution experimental time-series data we validated computational predictions for PhoP dependent genes in *Streptomyces coelicolor*. Our analysis shows that the prediction methods have little overlap with no single method predicting more than 52% of the PhoP dependent genes, whilst all methods contribute unique predictions demonstrating that pooling predictions across information sources is necessary. In our system correlation methods had greater accuracy than motif based methods, whilst causal signal based inference (using 1 hr time difference) had the lowest sensitivity, although it identified a rich set of coregulated genes. The coregulation profiles of these prediction methods were distinct indicating methodology biases. By integrating predictions from these methods we constructed a refined PhoP regulon and demonstrated that it has a mixed coregulation profile with genes either solely regulated by PhoP (within the frame of reference of our experiments), are jointly regulated by PhoP and a glutamate specific response or coregulated by a phosphate PhoP independent response pathway.

Key words: regulatory network, flow chart, PhoURP, phosphate depletion, glutamate depletion.

Introduction.

Systems biology has the potential to dramatically accelerate knowledge acquisition, high-throughput techniques being able to reverse engineer global regulatory networks within a comparatively short time frame either from time-series data, multiple conditions/strains comparison experiments or more recently, measuring regulator binding directly by chromatin immuno-precipitation (ChIP-Seq/chip). A number of methods have been developed to extract regulatory networks from these data, predominantly sequence based prediction algorithms (building binding site motif models using target searches) and network inference methods utilising correlates in gene expression data. There are a large number of available methods under the latter, differing in how they interpret correlations in gene expression profiles as evidence of regulatory links. Bayesian network models typically model regression correlates, *e.g.* either expression correlations (1) or causal signals in time series (2), while they are highly flexible allowing considerable model expansion, such as incorporation of hidden (unmeasured) variables (3) and nonlinearity (4). Information based methods interpret the conditional probability between genes as evidence of links, *e.g.* BANJO (http://www.cs.duke.edu/~amink/software/banjo) and ARACNE (5). The primary distinction between these methods, and their applicability to a given data set, depends on whether there is a structure to the patterns within the data, *e.g.* causal signals in time-series data; this structure can be utilised to extract informative signals through use of appropriate models. This is qualified by the need to match model complexity with the information levels in the data. There is thus a method hierarchy through model complexity, information based methods having the lowest complexity and consequently, fewest assumptions. However, there remain major unanswered questions on all these methods, specifically the reliability of these methods, the appropriate levels of interpretation that can be applied to patterns in the data, and the relative weighting of data types (6). A thorough comparison of these computational prediction methods has not been undertaken in any biological system, while the validity of extensive comparisons on simulated data (7) is unclear given the difficulty of simulating experiments with sufficient (unknown) realism. In particular, it is unknown if there are classes of targets that are preferentially detected by particular methods, and, as is typical for discovery studies, the accuracy of the resulting predictions is difficult to assess. This is in part because of 3 difficulties, specifically, separating ambiguous signals in correlated dynamic patterns, distinguishing direct and indirect (grandchild) regulation and decoupling the dynamic influences from multiple regulators of a gene. A major hindrance to solving these issues is the lack of a reference data set where the truth is known, synthetic networks (8) still being too small for extensive validation.

We propose a new validation methodology where we examine the performance of predictions against a set of comparative dynamic signatures that either support or contradict the claim of a particular regulatory dependence. The idea is to utilize the high levels of information inherent in the changes in temporal gene expression under different conditions or strains, essentially providing additional information/data to score predictions. We analyse microarray data from three time-series experiments performed on *Streptomyces coelicolor,* a model actinomycete producing a range of secondary metabolites including antibiotics under nutrient starvation. Upon phosphate depletion the PhoP two-component system (PhoURP, SCO4228-30) is the primary response pathway (9), affecting the expression of a swathe of genes. Importantly for our analysis, the PhoP regulon, and more broadly, the set of PhoP dependent genes is large, allowing a statistical assessment of prediction methods.

This paper is organised as follows. We examined seven predictions for PhoP dependent genes - a regulon constructed from previous biochemical and genetic studies, denoted G^{lit}, and six groups constructed by computational prediction methods utilising at most a single time-series, Supplementary TableS1. All the computational groups constitute legitimate predictions for PhoP dependent/influenced genes, and, ideally, should identify the same targets. Comparison of these predictions shows that they have little overlap. By scoring predictions against the comparative signatures we obtain an estimate of their accuracy whilst also decomposing the predictions into sole and joint (co-) regulation by PhoP. Integration across these predictions led to a list of genes with significant support for being involved in the nutrient limited response regulatory network that ultimately drives secondary metabolite production.

Results.

Transcriptome analysis: dynamics and differentially expressed (DE) genes. We analyse microarray data from three time-series (TS) experiments performed on *S. coelicolor* using the submerged batch cultivation system as described previously (10). Specifically, the TSs are $TS^{wt}{}_{PO4}$: wildtype (strain M145), phosphate depletion conditions (10); $TS^{ko}{}_{PO4}$: *phoP* knock-out (INB201), phosphate depletion conditions (this study); and, $TS^{wt}{}_{glu}$: wildtype, glutamate depletion conditions (Mast et al., submitted). Gene expression was measured approximately hourly over a period of 40 hrs as previously described (10). Transcription of the *phoP* operon dramatically increased within 1hr of phosphate depletion in wt (experiment $TS^{wt}{}_{PO4}$), but is not activated under glutamate depletion conditions, $TS^{wt}{}_{glu}$, Fig. 1. Thus, the PhoP pathway is only active in $TS^{wt}{}_{PO4}$, as *phoP* is deleted in $TS^{ko}{}_{PO4}$ and phosphate is not the limiting resource in $TS^{wt}{}_{glu}$.

We identified differentially expressed (DE) genes (11), typically giving ~2000 genes (25% of genome) per TS. We also added an additional 148 genes from the prediction groups deemed DE by eye but below threshold in the BATS analysis. We denote by $DE^{wt}{}_{PO4}$ genes that are DE in TS $^{wt}{}_{PO4}$, etc and by DE genes that are DE in any of the 3 TSs. There were 2867 genes considered DE in any time-series out of the genomic complement of 7814[1] genes on the array (37%). There are similar numbers of DE genes in each TS whilst nearly 40% of the DE genes were DE in all TSs, Fig. 2A. DE genes unique to a TS ranged from 11-24% of the DE genes, $TS^{wt}{}_{PO4}$ having the fewest (200), while overlaps between any 2 were of the same order (188-277). This already suggests that unravelling PhoP regulation is a complex problem since any pairwise comparison has similar levels of complexity; the comparison with Δ*pho* alone suggesting that there may be as many as 388 genes for which PhoP expression is essential, these genes losing DE under KO of *phoP*.

Performance of a literature compiled PhoP predicted regulon, G^{lit}. We examine a group of predictions, denoted G^{lit} consisting of 60 genes identified from the literature as regulated by PhoP and independent of our data sets (Table S1A). Of these 60 G^{lit}

[1] Based on annotation available in 2007. The new Uniprot-GOA annotation Barrell et al 2009 has 11 more genes.

genes, only 42 are DE in TS^{wt}_{PO4}, with 19 in TS^{ko}_{PO4} and 24 in TS^{wt}_{glu}; 54 are DE in at least one of the TSs, Fig. 2B. This bias of DE towards TS^{wt}_{PO4} provides confirmation of the enrichment of PhoP dependent genes in G^{lit} since PhoP is only active in this experiment. However, since only 70% of the predictions are active in TS^{wt}_{PO4} there is clearly a lack of support for a PhoP dependence for the 18 genes that are not active in this TS. Analysis of the DE overlap, Fig. 2B, identifies 27 genes that no longer display temporal variation in TS^{ko}_{PO4}. These genes, all of which are activated upon phosphate depletion, are clearly verified as PhoP dependent genes; *i.e.* PhoP activity is essential for their expression. The 15 genes that are DE in TS^{wt}_{PO4} and in TS^{ko}_{PO4} need further analysis to ascertain if they are likely to be PhoP dependent, *i.e.* these could have a second coregulator. Clustering of expression of the G^{lit} genes in TS^{wt}_{PO4} demonstrates they display a variety of behaviours with only 12 genes having a *phoP*-like profile, Supplementary Fig. S3. None show an inhibition profile.

Computational prediction methods. We analysed three computational prediction methods for PhoP dependent genes, Supplementary TableS1B, consisting of firstly, sequence based predictions using the PHO box motif, $G^{dr/word}$, secondly, methods identifying genes with similar dynamic patterns to *phoP*, $G^{simphoP/phoPclust}$, and thirdly, an inferred causal network, G^{causal}. In brief, G^{dr} and G^{word} differ in that either a directed repeat PHO box (Supplementary Fig. S1A) or a single word PHO box (approximately GTTCA) is used in the genome search, the latter allowing for the fact that the directed repeat (dr) motif may have a variable spacer or a weak second word, as is in fact evident in our case, Supplementary Fig. S1A. $G^{simphoP}$, $G^{phoPclust}$ are based on similar principles but implemented by a distance metric (searching across all gene expression profiles in TS^{wt}_{PO4} for those similar to *phoP*) and by clustering in TS^{wt}_{PO4} (restricted to the 1000 most significant DE genes in TS^{wt}_{PO4}) respectively.

We also define a second causal group that includes possible targets that are activated too rapidly by PhoP to leave a causal signature in the data. To infer dependent genes based on causal signals (G^{causal}) there must be a sufficiently strong causal signal (evident as a time displacement) between the expression profiles. Thus, genes with an expression similar to *phoP* cannot be ascribed a causal relationship. Thus, genes in the same cluster as *phoP* are probably regulated by PhoP given the lack of any other suitable regulator. We define the extended causal prediction group $G^{causal+}$ as the union of G^{causal} and the *phoP* cluster, Supplementary TableS1B.

To define comparative signatures on gene expression profiles, the gene must be DE in at least one of the TS. Thus, we restrict these computational predictions to the set of DE genes giving a combined total of 288 predicted PhoP dependent genes from the computational methods. There are a further 71 and 38 genes that are not DE in the prediction groups G^{dr} and G^{word}, respectively, Table 1. These genes are probably not PhoP dependent given the lack of dynamics in TS^{wt}_{PO4} and this should be borne in mind in the following analysis, *i.e.* these groups have another large group of false positives. Genes in these groups are given in Supplementary Tables TableS2_Causal.xls, etc.

Comparing the PhoP influenced gene predictions. There are 307 genes predicted to be PhoP regulated by any of the methods (including the 19 unique predictions from G^{lit}), with $G^{simphoP}$, $G^{dr/word}$ of similar size at 78, 113 and 112 resp. $G^{phoPclust}$ and G^{causal} are smaller which is probably due to the restricted target search within a smaller set of genes (1000; to enable clustering in TS^{wt}_{PO4}). The overlaps of these prediction groups are shown in Table 2. The 78 genes in $G^{simphoP}$ include most of the genes that cluster with *phoP*, *i.e.* >91% of $G^{phoPclust}$ are in $G^{simphoP}$. The causally predicted regulon G^{causal} has little overlap with $G^{simphoP}$ (6%); this is because genes with similar dynamics to *phoP* have no causal signal since regulation by PhoP is indistinguishable from a causal dependence on self, see Supplementary Methods. This suggests that correlation methods may fail to detect a proportion of PhoP dependent genes (~30%, G^{causal} comprising an additional 34 genes over and above the 78 contained in $G^{simphoP}$). The two motif predictions have overlaps of 21%; this is because there are genes with a weak second repeat which have insufficient conservation to be identified in the dr search. If we interpret the presence of a dr PHO box as identifying directly regulated genes (G^{dr}), the proportion of targets in each group that are directly regulated by PhoP are low at 38% for $G^{phoPclust}$, 28% $G^{simphoP}$ and only 8% for G^{causal}. Finally, comparing the computational predictions to G^{lit} shows that there are major differences to these other groups. Specifically, it has 19 unique predictions (35%), with only 35 predictions that are in any of our computationally predicted groups. The largest overlap is with G^{dr}, (26), closely followed by the correlation groups $G^{simphoP}$, $G^{phoPclust}$, while there is near negligible overlap with G^{causal}, Table 2.

This high disparity between predictions clearly calls into question the accuracy of all methods, suggesting that the computational predictive methods identify different targets. In the following we analyse these groups against 2 additional TS experiments and construct a framework to validate predictions and thus score prediction method accuracy.

Dynamic signature categorisation of (co)regulation. High-resolution time sampling provides a rich resource for ascertaining regulatory dependencies. PhoP dependent genes are expected to show significant dynamic differences between the cases where the PhoP pathway is active (TS^{wt}_{PO4}) and the two time-series where it is inactive (TS^{ko}_{PO4} and TS^{wt}_{glu}), Fig. 1. However, pleiotrophic effects complicate the interpretation of changes except in the simplest of cases. For instance, genes that are solely activated by PhoP can be identified because they are dynamic in TS^{wt}_{PO4} but inactive (non DE/flat) in the other time-series. The key issue therefore is whether genes with less dramatic dynamic changes between experiments can be interpreted in terms of regulatory relationships. To tackle this complexity we define 10 comparative signatures, see Methods & Supplementary TableS5. Combinations of signatures were then used to identify whether there is evidence of a PhoP dependence, and secondly if there is evidence of co-regulation, see flow chart in Fig. 3. Specifically we define categories **$\mathbf{Solo^{PhoP}}$**: sole regulation by PhoP (genes are only active (DE) in TS^{wt}_{PO4}); **$\mathbf{CR^{PhoP,glu}}$**: co-regulated by PhoP and Glu (respond under both phosphate and glutamate depletion but show no activity (non DE) under KO of *phoP*); **$\mathbf{CR^{PhoP,PO}}$**, phosphate specific but involving a PhoP independent response (inactive under glutamate depletion but have different dynamics under phosphate in the wt/KO); **Integ:** Multiple pathway integration (active, and profiles distinct, in all TSs).

We ran this categorisation on the DE genes, and on the best 1600 predictions the category was checked by inspection and reassigned if necessary, an assignment we denote as *Expert* in the following, see Methods. This reassignment was necessary since the above flow-chart does not distinguish dynamic changes around nutrient depletion from that elsewhere, either early or late in the time series. Thus, we reassigned the category based on changes within -1 to 5 hrs of nutrient depletion. Further, many ambiguous assignments could be resolved by eye. We identified 245 genes with strong support from the comparative signatures for being PhoP dependent, *i.e.* had an assignment to one of the regulatory categories **Solo$^{\mathbf{PhoP}}$**, **CR$^{\mathbf{PhoP,glu}}$**, **CR$^{\mathbf{PhoP,PO}}$** or **Integ**. These genes separated into 5 distinct types of expression patterns in TS^{wt}_{PO4}, (13 clusters), Supplementary Fig. S7, with most genes showing activation at, or near nutrient depletion (149) compared to inhibition (96). The profile patterns also had distinctive regulatory category profiles, with **Solo$^{\mathbf{PhoP}}$** dominating 5 clusters, while the inhibitory expression profiles displayed high levels of coregulation.

The flow chart categorisation was 46% accurate (relative to *Expert*) for designating a gene as PhoP dependent based on the regulatory categories. This could undoubtably be improved by threshold optimisation but this was not explored. The frequency of genes with a PhoP dependence was low outside the pooled predictions of Table 1, specifically 139 (48%) of the 288 computationally PhoP dependent predictions of Table 1 had evidence of being PhoP dependent from the Expert categorisation, which decreased to <8% for the remaining manually inspected genes.

Scoring predictions against dynamic signatures. To ascertain which computational methods are most accurate, we utilised the regulation category assignment above to score genes as having support for being PhoP dependent, or otherwise from the comparative analysis. Specifically we define a regulatory category as providing *evidence of comparative support* (ECS) if the category supports a PhoP dependence, *e.g.* **Solo$^{\mathbf{PhoP}}$**, *counter comparative evidence* (CCE) if there is evidence against PhoP regulation, or *unassigned* if PhoP dependence is unclear, see Supplementary TableS6.

Using this comparative analysis we find in all predictive groups evidence to both support and contradict claims of a PhoP dependence, Table 1. The performance of the group G^{lit} is good, with a 78% classification, although 29% of these genes are contradicted as being PhoP regulated since they are classified as CCE. The prediction method with the highest consistency under the comparative dynamics analysis is that based on correlation, $G^{simphoP/phoPclust}$ both having high assignment rates (>95%) and a high accuracy (>96%) of assigned genes. The causal network predictions had similar performance to G^{lit}, while the sequence predictions $G^{dr/word}$ were the poorest with only a 67-70% assignment rate and 55-56% accuracy, Table 1. Only the $G^{phoPclust}$ group is redundant, each of the others uniquely identifying a number of PhoP dependent genes (unique and with ECS, 14 G^{dr}, 29 G^{word}, 26 $G^{simphoP}$, 12 G^{causal}). We pool all the accepted predictions over the 7 prediction groups and denote this set G^{Dep}, 142 genes in total. Only $G^{simphoP}$ predicts more than ½ of these accepted PhoP dependent predictions (52%), Fig. 4A.

There are distinct profiles of co-regulation for the prediction methods, Fig. 4A. **Solo$^{\mathbf{PhoP}}$** dominates the PhoP dependent genes in $G^{simPhoP}$ and G^{lit}, whilst the former also contributes most of the targets that exhibit coregulation by PhoP and a phosphate PhoP independent response, **CR$^{\mathbf{PhoP,PO}}$**. The causal prediction is dominated by targets that have evidence of multiple regulation, particularly by a glutamate dependent response. This analysis confirms the intuition that genes with an expression profile similar to *phoP* in TS^{wt}_{PO4} ($G^{simphoP}$) are unlikely to have additional regulators; additional regulators, if active, would contribute to the expression profile and make it diverge from that of *phoP*. A regulatory category breakdown of the predicted groups is given in Supplementary Tables S3. We also examined whether there were functional enrichment within the predictive groups (GO analysis). We failed to find broad categorisations; all significant enrichments were due to a small number of genes (or an operon).

Improving the predicted PhoP dependents by integrating across data. There are 288 genes in the union of the three computational predictions (groups G^{dr}, G^{word}, $G^{simphoP}$, $G^{phoPclust}$, G^{causal}). 139 (48%) are confirmed as PhoP dependent. Of the remaining 149 predictions, 74 (26%) are inconsistent with being PhoP dependent (CCE) while the remaining 75 (26%) have no signature either way. A further 3 genes were identified from G^{lit} as PhoP dependent, and an additional 103 were identified from the comparison method alone by searching within DE (above) but not predicted by any of the 7 groups of Table 1. This high number of additional targets indicates that the computational methods are incomplete predictive methods; the failure to detect these targets is likely explained in part because these targets tend to have a later response to phosphate depletion, potentially indicating that they may be downstream of the PhoP pathway, or they show a relatively small change at nutrient depletion and thus their regulatory signals are weak.

To distinguish direct (versus indirect) regulation by PhoP we assume that genes in the regulon have a PHO box dr motif; this defines the refined regulon G^{Reg} (Supplementary Table S1B) as the intersection of G^{Dep} and G^{dr} giving 42 genes with ECS, Table 1. 21 of these genes have evidence of being solely PhoP regulated while 21 have evidence of also being regulated by a glutamate pathway (3 genes have joint classification). The rapidity of the response to phosphate depletion should enrich for direct regulation, *i.e.* those genes with an immediate (within 1hr) response are more likely to be in the PhoP regulon. This identifies clusters A-E and G (Supplementary Fig. S7) as the regulon, 102 genes (85 predicted) of which 30 possess a dr motif (33 either a dr or single PHO box), while 17 were not identified by any of the computational methods or in G^{lit}; all inhibitory profiles are excluded since these all have a delayed response.

For the 142 genes in G^{Dep}, there are 46-61 $\mathbf{Solo^{PhoP}}$, 13-16 $\mathbf{CR^{PhoP,PO}}$ and the remaining have evidence of coregulation by a glutamate response pathway. This contrasts to the predictions containing a dr PHO box, G^{Reg}, which have a higher proportion of $\mathbf{Solo^{PhoP}}$ (20-23) relative to G^{Dep}, Fig. 4 Thus, there is a correlation of being $\mathbf{Solo^{PhoP}}$ and having a dr PHO box; specifically 43% of $\mathbf{Solo^{PhoP}}$ genes have a PHO box compared to 27% of those PhoP dependent genes not categorised as $\mathbf{Solo^{PhoP}}$.

Robustness to thresholds: ROC curves. To estimate the performance (sensitivity and specificity) of our predicted groups a reference group of genes definitely regulated by PhoP is needed. Such a group is not known, but we can approximate it as the group G^{Dep}, *i.e.* predicted genes for which the comparative analysis confirms their PhoP dependence. This refined PhoP dependent gene set allows us to estimate the sensitivity of our predictive groups, Table 1. The correlation methods, $G^{simphoP/phoPclust}$ are again identified as the best performing methodology with a true positive rate TPR >44% and false positive rate FPR<0.12% (FP=5, 1 resp., the number of negatives being large at 2725 within DE). This conclusion is robust to the choice of (stringency) threshold used in the group definition. Specifically, we can carry out the same analysis for the 4 prediction groups G^{dr}, G^{word}, $G^{simphoP}$ and G^{causal} under a change of stringency thresholds spanning from the most confident predictions (high stringent threshold) to low quality predictions (low threshold), see Methods. This allows us to map out the performance relative to our best estimate of the PhoP influenced genes, G^{Dep} by tracing out the Receiver Operating Characteristic (ROC) curve, Fig. 5. The causal predictions G^{causal} had poorest performance, whilst $G^{simphoP}$ was the best, in fact reaching a 70% prediction rate with a reasonable level of false positives. The sequence prediction methods $G^{dr/word}$ had similar (intermediate) performance. When the causal network is corrected for similar genes ($G^{causal+}$) it had good predictivity, a performance that could evidently be further improved by taking a larger group of *phoP*-similar genes. All methods show an initially high performance at high stringency that drops off rapidly after a method specific threshold is reached, the ROC gradient thereafter being close to, or less than 1 (random selection of a FP or TP). The threshold separating this change in behaviour is similar to that chosen independently for defining the predictive groups as given in Table 1 suggesting that these thresholds are well chosen for each method.

For the unparametrised groups, G^{lit}, $G^{PhoPclust}$ and the DE genes in each TS we find the former two perform well on FPR but relatively low on TPR (low sensitivity), suggesting they are conservative. This is expected for G^{lit} since it is a prediction of PhoP directly regulated genes (the regulon). Correspondingly, the set of DE genes in each experiment are effectively random, although DE^{wt}_{PO4} has a higher TP rate as expected since the PhoP regulon is activated.

Discussion

We identified 139 genes with evidence of support of being PhoP dependent from our comparative analysis (having ECS) amongst the original 288 PhoP dependent computational predictions, Table 1. These predictions can be ranked by the number of methods predicting each gene. There are 30 PhoP dependent genes (including operons SCO4139-42, SCO4877-82, SCO5010-11) identified by at least 2 computational methods; most of these are a match between G^{dr} and $G^{simphoP}$, and thus are part of the PhoP regulon. Amongst the 3 computational methods, (sequence motifs, correlated profiles, causal), each have unique predictions, 167, 47 and 27 respectively, with 45, 43, 12 resp. unique predictions accepted as PhoP dependent (ECS), *i.e.* most of the validated predictions (72%) are unique to one method. This suggests that detection of downstream targets cannot be solely based on one method, since there is no universal characteristic that can be used to detect these genes, no single method detecting more than 52% of the dependent genes, Fig. 4. A further concern was that these methods may not detect all PhoP dependent genes. Our comparative analysis revealed an additional 103 targets (17 immediate responders) as PhoP dependent over and above those from the predictions in Table 1. A ROC analysis of the prediction methods, Supplementary Fig. S5, demonstrated that the specificity of the computational methods for these additional targets was very poor. This indicates that the methodologies are incomplete, and there remain another set of PhoP dependent genes that are only evident from a comparative (KO) analysis. The regulon faired no better, of the 42 genes at most 53% were predicted by a single ($G^{simphoP}$) computational method (other than G^{dr}). The immediate responsive genes in TS^{wt}_{PO4}, identified 102 genes as potentially comprising the PhoP regulon; however only 85 (83%) were in any of our prediction groups. This set of genes are enriched for the directed repeat PHO box (30%, $p<10^{-30}$) or single PHO box (either, 33%).

Of the previously predicted targets (G^{lit}), we find evidence to support 30 predictions out of the 54 that are DE, whilst 12 show CCE and thus there is significant counter evidence of a PhoP dependence. The remaining 6 show no DE and thus should also be regarded as lacking support. These contradictions to previous studies may be a reflection of our higher resolution experimental data and thus more informative, easier to interpret patterns, but could also be a result of an environment/culture condition dependence of the PhoP targets. Specifically, in other studies only 50-58% of ChIP-Seq/chip binding sites have been reported to be functional (12, 13), this possibly being explained through the lack of expression of an essential cofactor under the given environmental conditions, *i.e.* the targets are coregulated in a cooperative (logical AND) fashion. This implies that further testing of these predicted PhoP dependent genes is necessary to identify whether context is playing a role. This difficulty is not unique to PhoP, but relevant to all determinations of regulons.

The poorest performing predictive method as measured by the sensitivity was the causal signal based network inference method, group G^{causal}. Of the 36 predictions, 21 genes were ECS, while 7 demonstrated CCE. This is a high accuracy compared to sequence based methods. Thus, the poor sensitivity is only a reflection of the small number of initial predictions. A ROC analysis however demonstrated that lowering the stringency failed to detect additional targets with any level of accuracy, and in fact these additional targets were worse than random, Fig. 5. This is partly because of the inability of the method to identify targets with expression profiles that are very similar to that of *phoP*; the large contribution of $G^{simphoP}$ to G^{Dep} demonstrates that *phoP* similar genes comprise a large proportion of the PhoP dependent genes, G^{Dep}, Fig. 4. Despite the low sensitivity, there were 12 ECS confirmed predictions unique to this method, and 1 in G^{Reg}. This method showed a distinct bias towards identifying genes with a rich coregulation, *i.e.* many of the target genes have a glutamate dependent response which contrasts to the high dominance of $\mathbf{Solo^{PhoP}}$ targets in the correlation methods ($G^{simphoP}$, $G^{PhoPclust}$). Most of the genes in G^{causal} are in fact inhibited at phosphate depletion, Supplementary Fig. S4, consistent with there being additional regulators. These genes also typically did not have a dr (or GTTCA) motif, and therefore are unlikely to be directly regulated by PhoP; this is supported by their response timing, inhibition typically starting 2hrs after nutrient depletion. Thus, the poor sensitivity of the causal predictive group is probably a mix of 2 effects: in bacteria transcriptional regulation is rapid and so a 1hr time resolution is on the verge of being too long an interval to detect direct causality, (compare the significant improvement of $G^{causal+}$ with respect to G^{causal}, Fig. 5), downstream genes in the pathway in fact being detected, and secondly, the complexity of coincident responses under phosphate depletion means that PhoP dependent and PhoP independent phosphate responses are inseparable by examination of gene expression dynamics in TS^{wt}_{PO4} alone, these genes having similar expression profiles under these conditions.

Conclusions

In this paper we have utilized three high resolution time-series of microarray derived gene expression data to evaluate, and distinguish genes regulated by PhoP and other stress pathways. We developed a methodology based on comparison of high resolution temporal profiles to classify genes based on 3 regulatory influences- PhoP, PO_4 (possibly multiple PhoP independent phosphate depletion responses) and Glu (possible multiple glutamate depletion specific responses). We were able to distinguish sole regulation by PhoP (**Solo$^{\mathbf{PhoP}}$**) and coregulation by PO_4 (**CR$^{\mathbf{PhoP,PO}}$**) and/or Glu (**Integ/CR$^{\mathbf{PhoP,glu}}$**) in conjunction with PhoP. This identification of a gene's regulatory influences allowed us to firstly, score computational prediction methods for their accuracy, secondly, construct a refined prediction for PhoP dependent/regulated genes based on confirmed computational predictions against comparative signatures, and thirdly, analyse prediction methods against the type of regulation. We demonstrated that causal, correlation and comparative information have fundamental (coregulation) biases in their predictions, and thus an integration across methods is necessary to achieve a coverage of both PhoP dependent genes and the regulon, Fig. 4. Specifically, causal signals (G^{causal}) detect a high prevalence of phosphate and glutamate influenced genes (**Integ**, **CR$^{\mathbf{PhoP,glu}}$**), while correlation methods ($G^{simphoP}$) are dominated by solely PhoP dependent genes (**Solo$^{\mathbf{PhoP}}$**). The dr motif search G^{dr} identifies a high number of **Integ** and **Solo$^{\mathbf{PhoP}}$** targets. Of the three computational methods, the correlation based methods have the best performance with a sensitivity of 45-51%, followed by sequence search methods (30-31% TPR), Table 1. The number of FP predictions was low for correlation methods, but the sequence prediction methods had FP/TP ratios close to 1. This high error rate was also evident in the CCE statistic, Table 1, indicating that the motif methods had low accuracy. The dr and single word motifs performed similarly indicating that the dr PHO box signature had no more predictive power than a single GTTCA motif, identifying predominantly different sets (only 21% overlap, 14 (12%) with ECS). This might reflect the difficulty in performing a flexible dr search (only the first word is well conserved, see Supplementary Fig. S1), or it may reflect biological functionality of highly imbalanced dr word components, genes with a poor/absent second word being more likely to be coregulated, Fig. 4. Our analysis supports the hypothesis that PhoP is predominantly an activator of transcription; most of the PhoP dependent genes have an activation profile at nutrient depletion. The mechanism of inhibition has been suggested to be indirect through prevention of RNA polymerase binding by obstructive binding of PhoP in the -10 region (14).

Our analysis is very dependent on having a sufficiently large diverse, and importantly, coherent data set on which we can utilise a number of prediction methods. This contrasts to other studies where only a single method is typically implemented, usually because of the inappropriateness of other methods. For instance, analysis of large compilations of microarray data such as M^{3D}, (15) is restricted because of compatibility issues and the low information per dataset (few samples). Fundamental to our analysis is a standardised experimental methodology across experiments (minimising variation due to experimental protocols) and a high time resolution coupled with an ability to synchronise the time-series; in our case the time of nutrient depletion is a natural point of reference. Our integration across methodologies is thus highly dependent on the ability to apply multiple methods to our data and utilise detailed modelling/categorisation of patterns to interpret the fine detail in the expression profiles. Through this comparative dynamic analysis we demonstrate that there is a methodology bias towards certain types of gene regulation, although no method is redundant, with unique targets being identified from each. Further, we identified methodology limitations in causal network inference that stem from inadequate time resolution in the experimental time-series.

Materials and Methods.

Transcriptome data. Cultivations were carried out as (10) with the exception of TS^{wt}_{glu} where glutamate was reduced from 55 to 15 g/L and phosphate doubled from 4.6 to 9.2 mM. Samples were distributed over 20-60 hrs post inoculation with time separation as low as ½ hr. mRNA was extracted and hybridised to the microarrays as (10). Arrays were processed using RMA with loess normalisation, separately for each time-series. Differential expression was ascertained using the BATS package (www.na.iac.cnr.it/bats). **Microarray Data** for TS^{wt}_{PO4} and TS^{ko}_{PO4} are available on Gene Expression Omnibus (GEO) at http://www.ncbi.nlm.nih.gov/geo/ with series identifiers GSE18489 and GSE31068[2] respectively. **Prediction groups.** $\mathbf{G^{lit}}$ is a set of directly PhoP regulated genes identified from the literature, see Supplementary TableS1A for provenance information. $\mathbf{G^{dr}}$ is a genome-wide search (MEME software suite) for the directed repeat PHO box (16 bases) generated from targets in G^{lit}, motif model as Supplementary Figure S1A. $\mathbf{G^{word}}$ is the corresponding search for the single binding site approximating GTTCA (first word of above motif). All genes in the operon are considered as having these motifs, (operon definitions as (16)). Thresholds 9 and 12.5 for the dr and word motif respectively were used in the groups used in Tables 1,2 (see Supplementary data for a discussion). $\mathbf{G^{causal}}$ comprises all genes predicted to be regulated by *phoP* with a posterior link probability >0.4 in the sparse linear causal model of (4), generated from the GRENITS bioconductor package (submitted). $\mathbf{G^{simphoP}}$ comprises genes with similar profiles to SCO4230 (*phoP*) based on a distance metric lying within the q-th lower quantile of the distribution (Tables 1,2 use q=0.01). **Comparative Signature Analysis and scoring.** The regulation categories are defined as composite logical relations on the signatures defined in Supplementary TableS5. The regulatory category assignments of the best 20% of predictions from each method were inspected by eye (Expert categorisation). Predicted genes were scored for being in a PhoP dependent regulatory category. ROC curves were plotted using as the reference set either G^{Dep} or G^{Reg} running over the group defining threshold: G^{causal} the posterior link probability, $G^{dr/word}$ the motif threshold and $G^{simphoP}$ the tail percentile q.

Acknowledgments

This project was supported by grants of the ERA-NET SysMO Project [GEN2006-27745-E/SYS]: (P-UK-01-11-3i) and the Research Council of Norway [project no. 181840/I30]. MJ, MI, JM, RL were funded by BBSRC, grant (SysMO programme, BB/F003439/1), ER by EPSRC through the Warwick Systems Biology DTC. We acknowledge with thanks EMH Wellington as coordinator and PI of STREAM. We further acknowledge important contributions to the development of the cultivation system by Håvard Sletta (SINTEF) and Per Bruheim (NTNU) and the excellent technical help of Anders Øverby, Elin Hansen and Sunniva Hoel during fermentation experiments.

Author contributions.

NJB wrote the manuscript with contributions from the other authors, completed the analysis and conceived the ideas. RL, NJB built the processing pipeline. JM built a Biomart database and undertook general data handling. MJ performed the pilot analysis with help from EM, RL, NJB. EM, MJ wrote the causal analysis software (deriving group G^{causal}). MI performed the motif analysis (deriving groups $G^{dr/word}$). TE, ØMJ and AW were responsible for all aspects related to the time-series fermentation experiments. Arrays were hybridized in Tuebingen with pipeline construction and quality control overseen by KN, as (10). AR-G, JM provided the *phoP* deletion mutant strain INB201, and compiled the literature based PhoP regulon prediction (G^{lit}), Table S1A.

[2] Not public. See letter for access weblink.

Tables.

TABLE 1. Predictive group performance against the comparative dynamic signatures. Rows are group size, the number of genes that are DE in any of the time-series, Evidence for comparative signature (ECS), Comparative counter evidence (CCE), assignment rate p (either ECS, CCE), error and precision Q, all measured relative to DE (not Size). Performance is calibrated against the accepted PhoP dependent genes G^{Dep} giving FP, TP, FPR and sensitivity TPR (relative to the universe of genes used by the prediction method, *i.e*. the genome, DE^{wt}_{PO4} or 1000 top ranking DE^{wt}_{PO4}). Max TPR refers to the size of the universes used in the prediction group relative to DE.

Group	Size	DE	ECS	CCE	AssignmentRate	Error	PrecisionQ	FP	TP	FPR	rescaledTPR	MaxTPR
Motif dr	184	113	42	34	67%	45%	55%	71	42	0.93%	30%	100%
Motif Word	150	112	44	34	70%	44%	56%	68	44	0.89%	31%	100%
phoPClusters	47	47	46	1	100%	2%	98%	1	46	0.11%	45%	73%
SimilarPhoP	78	78	73	2	96%	3%	97%	5	73	0.07%	51%	100%
Causal	36	36	21	7	78%	25%	75%	15	21	1.70%	20%	74%
Causalplus	77	77	61	8	90%	12%	88%	16	61	1.81%	59%	73%
Literature	60	54	30	12	78%	29%	71%	24	30	0.31%	21%	100%

TABLE 2. Overlaps of predicted PhoP dependent genes of Table 1 (pairwise), restricted to genes that are DE. Overlap percentage of each row group is shown in parentheses. There are 307 genes in the union across all predictions. The last column shows how many predicted genes are unique to each group ($G^{causal+}$ is removed from the uniqueness analysis).

Observations	Motif dr	Motif Word	phoPClusters	SimilarPhoP	Causal	Causalplus	Literature	Unique
Motif dr	113 (100.0%)	24 (21.4%)	18 (38.3%)	22 (28.2%)	3 (8.3%)	20 (26.0%)	26 (48.1%)	72 (63.7%)
Motif Word	24 (21.2%)	112 (100.0%)	8 (17.0%)	9 (11.5%)	2 (5.6%)	10 (13.0%)	14 (25.9%)	87 (77.7%)
phoPClusters	18 (15.9%)	8 (7.1%)	47 (100.0%)	43 (55.1%)	6 (16.7%)	47 (61.0%)	23 (42.6%)	0 (0.0%)
SimilarPhoP	22 (19.5%)	9 (8.0%)	43 (91.5%)	78 (100.0%)	2 (5.6%)	43 (55.8%)	24 (44.4%)	29 (37.2%)
Causal	3 (2.7%)	2 (1.8%)	6 (12.8%)	2 (2.6%)	36 (100.0%)	36 (46.8%)	3 (5.6%)	27 (75.0%)
Causalplus	20 (17.7%)	10 (8.9%)	47 (100.0%)	43 (55.1%)	36 (100.0%)	77 (100.0%)	24 (44.4%)	-
Literature	26 (23.0%)	14 (12.5%)	23 (48.9%)	24 (30.8%)	3 (8.3%)	24 (31.2%)	54 (100.0%)	19 (35.2%)

Figures.

Figure 1. Gene expression and nutrient dynamics. **A.** mRNA time course (log base 2) *phoP* (SCO4230) expression for TS^{wt}_{PO4} (blue), TS^{ko}_{PO4} (red), TS^{wt}_{glu} (black). **B.** Phosphate and glutamate levels (dashed) for the 3 experiments (same colours as A). Time is measured from nutrient depletion (35 hrs from inoculation for TS^{wt}_{PO4}).

Figure 2. Overlaps of differentially expressed genes and predicted groups. **A**. Overlap of DE genes in each TS relative to whole genome, **B**. decomposition of G^{lit} predictions by DE, **C**. overlaps of predictive groups G^{causal}, G^{dr} and $G^{simphoP}$ within the set of DE genes.

Figure 3. Flow chart classification of PhoP dependence by the comparative dynamic signatures analysis. See text and Supplementary Data for regulatory category description. Sig. dyn. change stands for significant dynamic change, $S^{dyn}{}_{ko}$ and $S^{dyn}{}_{env}$ of Supplementary Table S5.

Figure 4. Regulation profiles of the PhoP dependent genes by prediction method. **A**. Decomposition of predictions in G^{Dep} and each group's accepted PhoP predictions (based on the comparative analysis (ECS)). CSA are the genes identified from the comparative signatures as PhoP dependent but not predicted from any of the groups. **B**. Decomposition of predictions in G^{Reg}, and the respective members from each prediction group. Only the principle categories are shown (joint categories are partitioned proportionally to components); see Supplementary Data for definitions of categories.

Figure 5. ROC curves for the groups of predictions. **A.** ROC of prediction of G^{Dep} for the 4 parametrised prediction groups G^{dr} (red), G^{word} (green), $G^{simphoP}$ (blue), G^{causal} (black). **B.** Detail of A. Red dashed line is equality of TP and FP. **C.** ROC of prediction of G^{Reg} for the 3 parametrised prediction groups G^{word} (green), $G^{simphoP}$ (blue), G^{causal} (black). The performance of groups at the thresholds used in Table 2 are shown as dots while the performance of the groups $G^{causal+}$ (black diamond), $G^{phoPclust}$ (blue diamond), G^{lit} (magenta star) and $DE^{wt}{}_{PO4}$ (black square), $DE^{ko}{}_{PO4}$ (red square), $DE^{wt}{}_{glu}$ (blue square), are shown. All ROC curves are computed relative to the prediction group universe DE, thus G^{causal} is based on a smaller set of possible TPs, Table 2. ROCs are incomplete for the motif searches as the score jumps to zero at around 2000 genes.

Supplementary materials.

Text and Figures.
SupplementaryTextECS.pdf

Tables.
Table S1A. Provenance of PhoP regulated genes identified in the literature.
Table S1B. Predictive group definitions.

Tables S2. Excel files TableS2_GroupName.xls.
Tables of genes in each computational predicted group G^{dr}, G^{word}, G^{causal}, $G^{simphoP}$. Column two contains the respective score for each gene where relevant. Common names and annotation are included. The motif files are corrected for operons, see Methods.

Tables S3. Excel file TableS3_GroupName.xls.
Performance Tables of genes in each predicted group G^{dr}, G^{word}, G^{causal}, $G^{simphoP}$, G^{lit} of Table 1. Thresholds correspond to 0.4, G^{causal}, 9, G^{dr}, 12.5, G^{word}, 0.01, $G^{simphoP}$ relative to TableS2, whilst all groups are restricted to DE. Columns are: the regulation category according Flow Chart (FC) and Expert, PhoP dependence (which can be fractional if multiple regulation categories given), the number of signatures supporting activation (nECS+) or reduction (nECS-) of dynamic variation on activation of PhoP, differential expression in each of the time series (column DE with a 3 letter code, D for DE, N for not DE, order TS^{wt}_{PO4}, TS^{ko}_{PO4}, TS^{wt}_{glu}), annotation as (17) and COG class when available. For G^{causal}, G^{lit} membership of clusters (column Cluster) shown in Figs. S3/4 is indicated (Singleton indicates the gene is in a cluster by itself).

Table S4. Excel file TableS4_CSA_PhoPDeptargets.xlsx.
Predicted PhoP dependent genes G^{Dep}, and G^{Reg} Columns as above in S3. The groups identifying that target are also indicated under columns named by each group (0 absent, 1 present); NoGrps gives the number of groups predicting gene, and No. Methods identifies the number of computational method (motif, similarity, causal) identifying the gene. Targets identified only from the comparative analysis (Expert assessment) are also included and identified as having NoGrps=0.

TABLE S5. Definition of comparative signatures defined on the 3 TSs TS^{wt}_{PO4}, TS^{ko}_{PO4}, TS^{wt}_{glu} and associated methodology. ko/env refer to the strain/environment comparison respectively. All comparisons are carried out with matched time at nutrient depletion using common time points.

Bibliography

1. **Castelo R & Roverato A (2009) Reverse Engineering Molecular Regulatory Networks from Microarray Data with qp-Graphs. *Journal of Computational Biology* 16(2):213-227.**

2. Kim S, Imoto, S., and Miyano, S (2003) Inferring gene networks from time series microarray data using dynamic Bayesian networks. *Briefings in Bioinformatics* 4:228-235.
3. Beal MJ, Falciani F, Ghahramani Z, Rangel C, & Wild DL (2004) A Bayesian approach to reconstructing genetic regulatory networks with hidden factors *Bioinformatics* 21(3):349-356.
4. Morrissey E, Juarez M, Denby K, & Burroughs N (2011) Inferring the time-invariant topology of a nonlinear sparse gene regulatory network using fully Bayesian spline autoregression. *Biostatistics* 12(4):682-694.
5. Margolin AA*, et al.* (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7 Suppl 1:S7.
6. Geier F, Timmer J, & Fleck C (2007) Reconstructing gene-regulatory networks from time series, knock-out data, and prior knowledge. *BMC systems biology* 1:11.
7. Schaffter T, Marbach D, & Floreano D (2011) GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics* 27(16):2263-2270.
8. Cantone I*, et al.* (2009) A Yeast Synthetic Network for In Vivo Assessment of Reverse-Engineering and Modeling Approaches. *Cell* 137(1):172-181.
9. Rodriguez-Garcia A, Barreiro C, Santos-Beneit F, Sola-Landa A, & Martin JF (2007) Genome-wide transcriptomic and proteomic analysis of the primary response to phosphate limitation in Streptomyces coelicolor M145 and in a Delta *phoP* mutant. *Proteomics* 7(14):2410-2429.
10. Nieselt K*, et al.* (2010) The dynamic architecture of the metabolic switch in Streptomyces coelicolor. *BMC Genomics* 11:10.
11. Angelini C, De Canditiis D, Mutarelli M, & Pensky M (2007) A Bayesian approach to estimation and testing in time-course microarray experiments. *Statistical Applications in Genetics and Molecular Biology* 6.
12. Gao F, Foat BC, & Bussemaker HJ (2004) Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics* 5.
13. Ucar D, Beyer A, Parthasarathy S, & Workman CT (2009) Predicting functionality of protein-DNA interactions by integrating diverse evidence. *Bioinformatics* 25(12):i137-144.
14. Sola-Landa A, Rodriguez-Garcia A, Apel AK, & Martin JF (2008) Target genes and structure of the direct repeats in the DNA-binding sequences of the response regulator PhoP in *Streptomyces coelicolor*. *Nucleic Acids Research* 36(4):1358-1368.
15. Faith JJ*, et al.* (2008) Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Research* 36(Database issue):D866-870.
16. Charaniya S*, et al.* (2007) Transcriptome dynamics-based operon prediction and verification in *Streptomyces coelicolor*. *Nucleic Acids Res* 35(21):7222-7236.

17. **Barrell D*, et al.* (2009) The GOA database in 2009-an integrated Gene Ontology Annotation resource. *Nucleic Acids Research* 37:D396-D403.**

Figures.

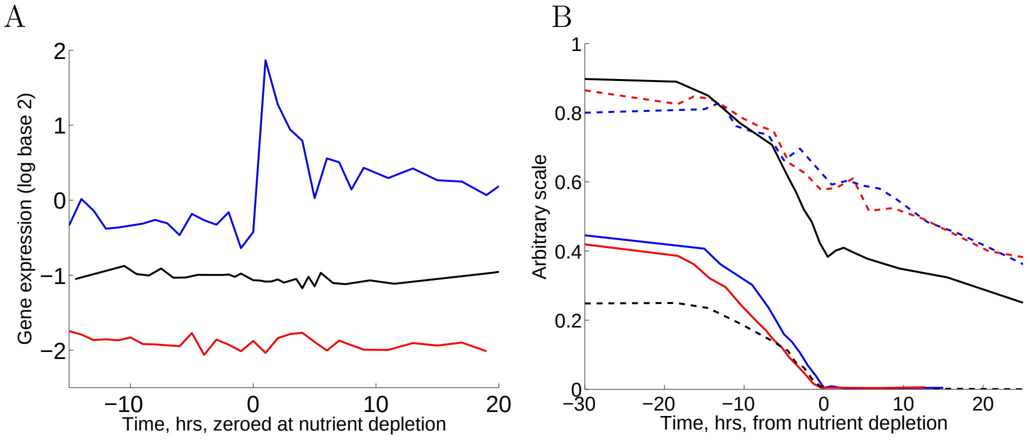


Figure 1.

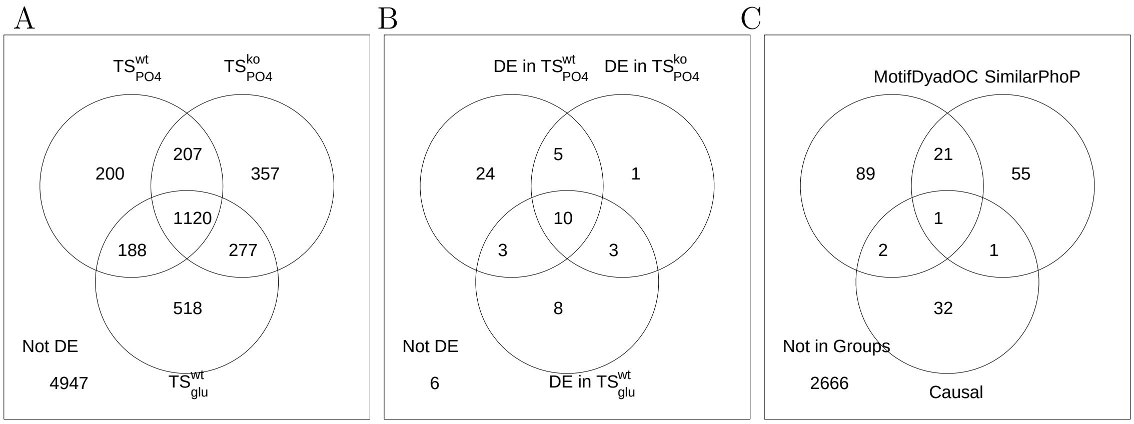

A
TS^{wt}_{PO4}
TS^{ko}_{PO4}
200
207
357
1120
188
277
518
Not DE
4947
TS^{wt}_{glu}
B
DE in TS^{wt}_{PO4}
DE in TS^{ko}_{PO4}
24
5
1
10
3
3
8
Not DE
6
DE in TS^{wt}_{glu}
C
MotifDyadOC
SimilarPhoP
89
21
55
1
2
1
32
Not in Groups
2666
Causal

Figure 2.

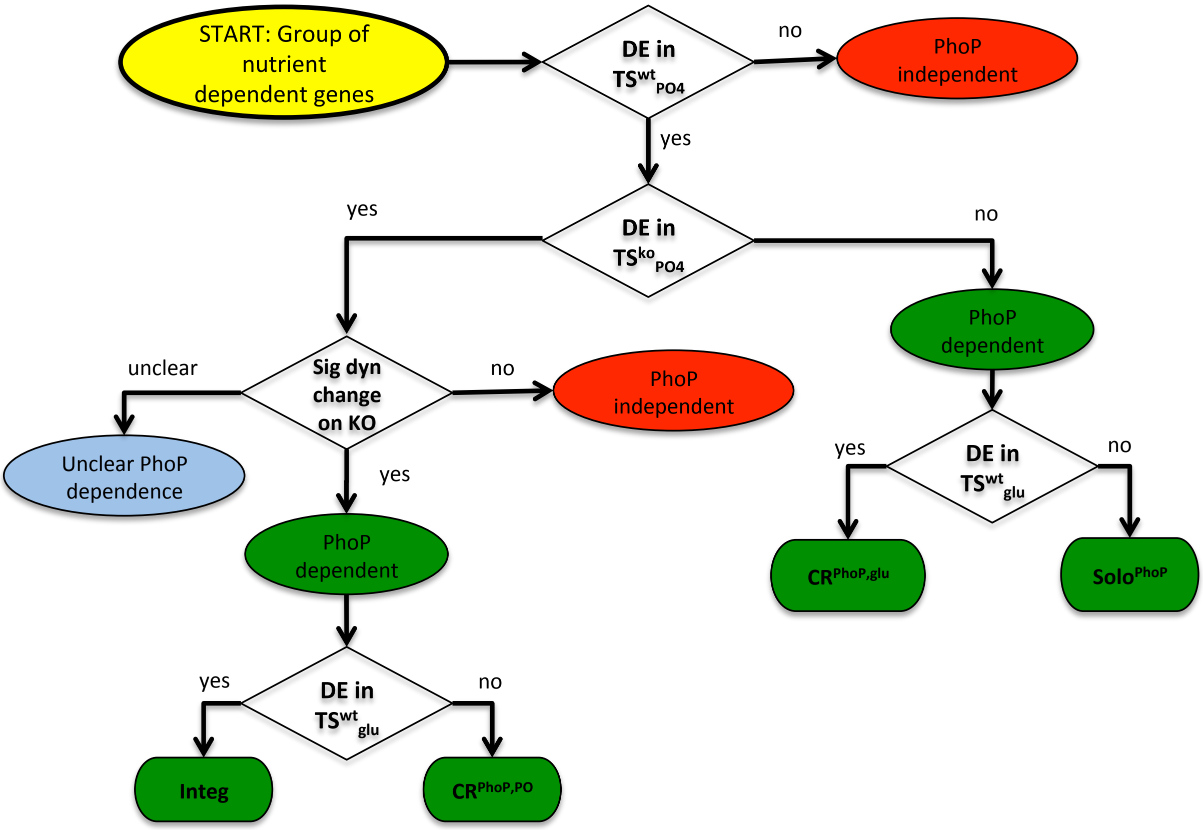

START: Group of nutrient dependent genes
DE in TS^{wt}_{PO4}
no
PhoP independent
yes
DE in TS^{ko}_{PO4}
yes
no
PhoP dependent
Sig dyn change on KO
unclear
no
PhoP independent
Unclear PhoP dependence
yes
PhoP dependent
DE in TS^{wt}_{glu}
yes
no
$CR^{PhoP,glu}$
$Solo^{PhoP}$
DE in TS^{wt}_{glu}
yes
no
Integ
$CR^{PhoP,PO}$

Figure 3.

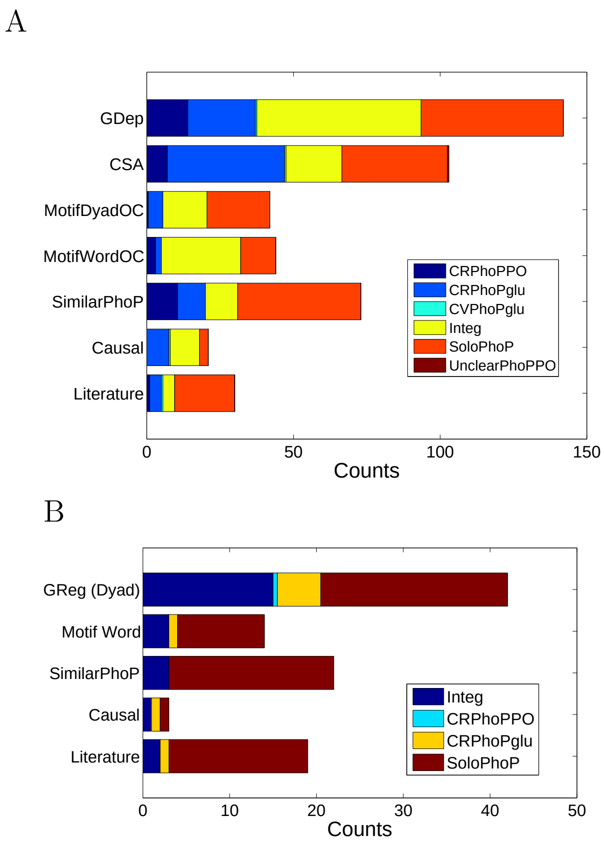

A
GDep
CSA
MotifDyadOC
MotifWordOC
SimilarPhoP
Causal
Literature
CRPhoPPO
CRPhoPglu
CVPhoPglu
Integ
SoloPhoP
UnclearPhoPPO
0
50
100
150
Counts
B
GReg (Dyad)
Motif Word
SimilarPhoP
Causal
Literature
Integ
CRPhoPPO
CRPhoPglu
SoloPhoP
0
10
20
30
40
50
Counts

Figure 4.

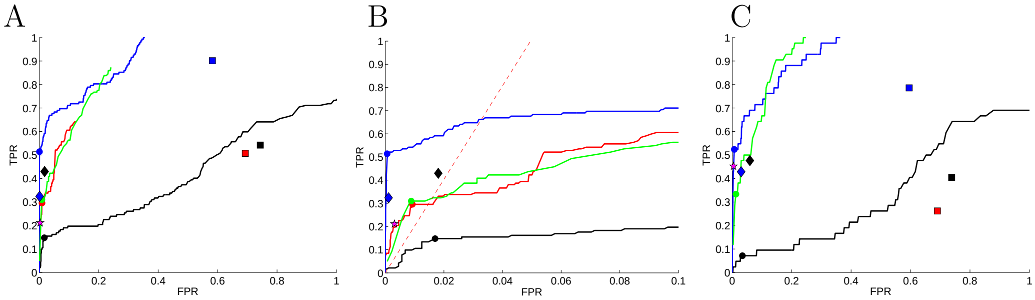

A
B
C
TPR
FPR
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
1
0.02
0.04
0.06
0.08
0.1

Figure 5. ROC.

Supplementary Text.

Coregulation profiling of causal, correlation and bioinformatic predictions of PhoP dependent genes in *S. coelicolor* reveals methodology biases

Contact: Nigel Burroughs. Email: N.J.Burroughs@warwick.ac.uk.

1 Contents.

2 PHO box motifs constructed from G^{lit}

The gene group G^{lit} (60 genes) was used to construct position weight matrix models of PhoP binding sites using the MEME suite, Fig. S1. The resulting motif was a directed repeat (dr), with a weaker second repeat. We subsequently used the dr motif and its first word as PhoP binding site models. Although an alignment (Supplementary Table S1A) showed the presence of a common motif in all operons, the constructed motif models revealed that specificity dropped rapidly with increasing stringency, Fig. S2, with only 20% (dr), 4% (word) of genes in the genome with the respective motifs actually being in G^{lit} at a score of 10, a score where the sensitivity (the proportion of G^{lit}) plateaued at 45% (dr) and 38% (word). The sensitivity remained at that level for the dr and increased for the single word motif to about 80% but at a significant loss of specificity. At high stringency (score), the dr demonstrates that it picks out only members of the training set within the genome for thresholds >17, identifying 20% of the target set. In contrast the single word motif is unable to identify only targets of the training set. This analysis demonstrated that there were rapid changes in sensitivity for our 2 motifs above the thresholds 13 and 10 respectively. The thresholds used in the analysis (Table 2) were thus reasonable and would not unfairly disadvantage G^{lit} performance.

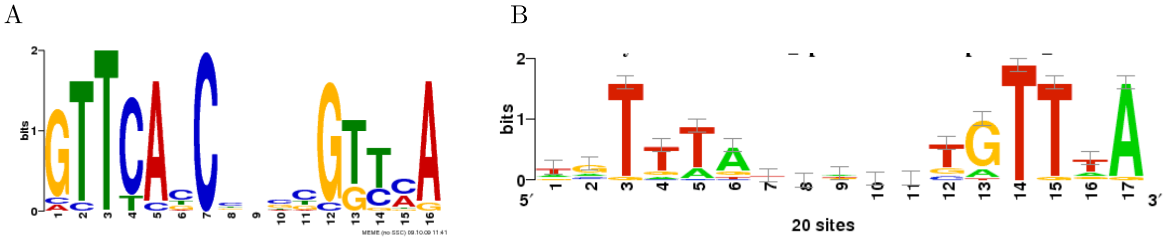


Figure S1. PhoP binding site models. **A.** PHO box directed repeat (dyad) motif, constructed using MEME on genes in Table S1A (G^{lit}). **B**. PHO box for *E. coli* K12 (RegulonDB).

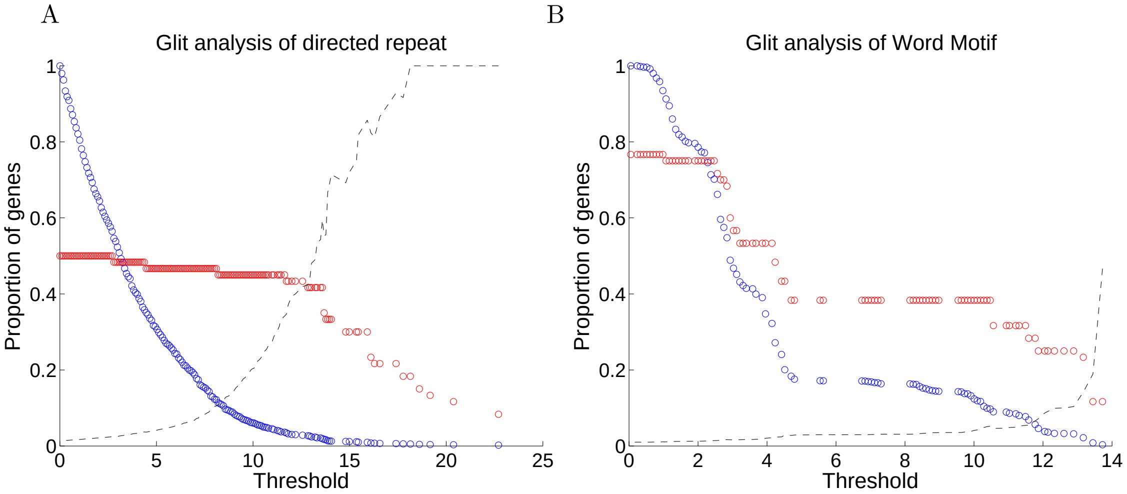


Figure S2. Performance of PhoP motifs against the training set. The proportion of genes in each of G^{lit} and the genome with a motif score above the given threshold is shown for, **A**. Directed repeat motif. Red G^{lit}, Blue 2188 highest scoring genes in the genome, **B** Single word motif. Red G^{lit}, Blue 4551 highest scoring genes in the genome. On each plot the specificity is shown (black dashed), *i.e.* the ratio of genes in G^{lit} to the genes in the genome with a motif score above the given threshold. Genes are operon corrected.

3 G^{lit} clusters

The 54 DE genes belonging to the prediction group G^{lit} were clustered using the SplineCluster package (Heard et al., 2005), Fig. S3. There were 6 clusters and 7 singletons. Of the clusters, only 4 clusters showed an expression change at nutrient depletion, all showing activation, whilst the other non-singleton clusters (5,6 genes, not shown) had an expression profile with changes during the growth phase but were flat after 31 hrs post inoculation, and exhibited gradual decay throughout respectively. Cluster A (12 genes) is similar to the expression of *phoP* with a substantial transient activation followed by a sustained activation level, Cluster B (13 genes) shows sustained activation from nutrient depletion, Cluster C (3 genes) increased expression over 5 hours prior to nutrient depletion then exhibits a further increase in activation at nutrient depletion, and Cluster D (8) shows weak transient activation at nutrient depletion.

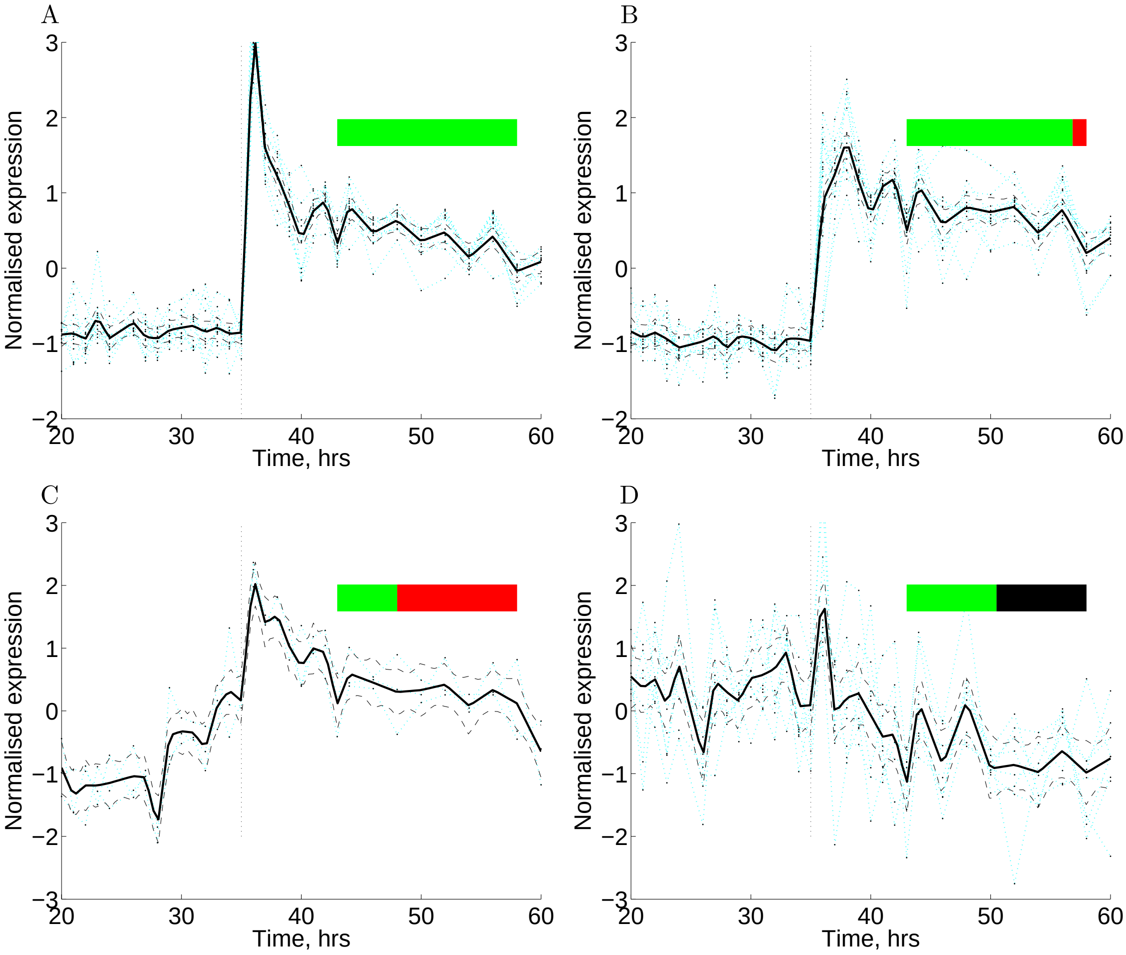


Figure S3 Clusters of prediction group G^{lit} of the PhoP regulon. Six distinct clusters were found (and 7 singletons), four (A-D) with dynamics around nutrient depletion. **A.** Peak activated genes (12), **B.** sustained activation genes (13). **C.** early activated genes (3), **D.** weak transient activation (8). Numbers in parentheses are the number in each cluster. Mean cluster profile (black), confidence intervals (dashed) and gene expression of individual genes (cyan) are shown. Bar indicates proportion of ECS (green), UA (black), CE (blue). Gene membership of these clusters is given in Table S3. Clustering performed using SplineCluster software.

4 G^{causal} clusters

The genes belonging to the causal regulon were clustered using the SplineCluster package (Heard et al., 2005), Fig. S4. Four clearly distinct clusters were identified, Cluster A, comprising activated genes (6 genes), cluster B, has a transient activation then inhibition profile (4 genes), clusters C and D, two inhibited expression clusters (26 genes). The first cluster (A) comprises genes presenting a similar but delayed expression profile to *phoP* with an expression peak at approx. 38h post inoculation. Only 3 of these genes show activity in TS^{wt}_{glu}, SCO1048, SCO0919, SCO1196 (categorised as $CR^{PhoP,glu}$) with transient activation on glutamate depletion over 3-8 hrs; all the rest (classed as $Solo^{PhoP}$) are inactive. All are inactive in TS^{ko}_{PO4}. The second cluster (B) contains 4 genes showing a high peak in expression at 36h, dynamics again delayed relative to *phoP*. This high expression lasts only for around two hours before returning rapidly to a lower inhibited level (relative to growth phase expression). They are all inhibited in TS^{wt}_{glu}at glutamate inhibition, while only SCO2052 (Integ) is active in TS^{ko}_{PO4}, with transient activation at PO_4 depletion, whilst SCO1845 ($CR^{PhoP,glu}$) is the only one that shows distinct loss of expression. Cluster C (predominantly Integ) comprises genes activated during growth phase but inhibited on phosphate depletion. All show similar profiles in the other times series; transient inhibition on TS^{wt}_{glu}, and inhibition in TS^{ko}_{PO4}, although all are typically over expressed relative to TS^{wt}_{PO4}up to nutrient depletion. This suggests that the early *phoP* expression in TS^{wt}_{PO4}, Fig. 1, is in fact inhibiting these genes. The last, largest cluster D contains 17 genes (most showing regulation by glutamate either as $\mathrm{CR}^{PO,glu}$ or $\mathrm{CR}^{PhoP,glu}$), showing rapid inhibition 1hr post depletion in TS^{wt}_{PO4}. About half show similar (invariant) profiles in TS^{wt}_{PO4} as TS^{ko}_{PO4}, except the decline at depletion is more gentle; the others are approximately flat although expression levels of all genes remain higher than in TS^{wt}_{PO4}, similar to pre-depletion values. In TS^{wt}_{glu}, all except SCO6108 show a transient dip in expression after glutamate (depletion). Typically there is a predominant lowering of expression after phosphate depletion in TS^{wt}_{PO4} relative to TS^{ko}_{PO4} and TS^{wt}_{glu}, suggesting these genes may be switched off. This indicates that most of these genes are not PhoP dependent but responding through generic stress pathways, *i.e.* the CCE classification is correct.

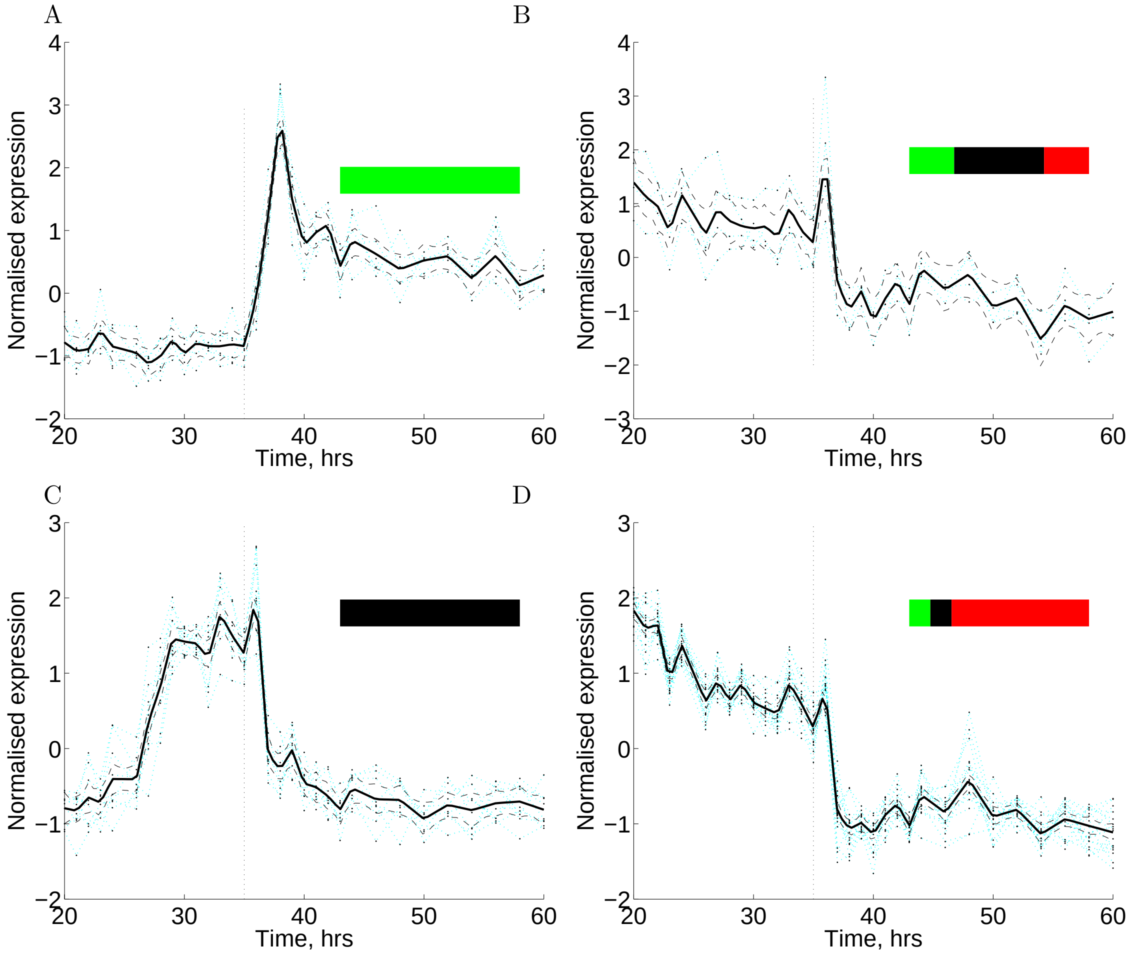


Figure S4. Clusters for the causal network predicted *phoP* target genes (posterior probability of phoP being a parent $p > 0.4$). Clustering of the target genes show four distinct clusters, **A.** Activated genes (6), **B.** Transiently activated inhibited genes (4). **C.** late exponential inhibited genes (9) **D.** Inhibited genes (17). Numbers in parentheses are the number in each cluster. Mean

cluster profile (black), confidence intervals (dashed) and gene expression of individual genes (cyan) are shown. Bar indicates proportion of ECS (green), UA (black), CCE (red). Gene membership of these clusters is given in Table S3_Causal. Clustering performed using SplineCluster software.

5 Robustness analysis

Robustness of the results to choices and thresholds in the analysis are further examined here. A basic ROC analysis was given in Fig. 5 of main text demonstrating that results were invariant to the choice of predictive group thresholds. However, additional analyses are needed, specifically showing an invariance to -

- Universe size. The universe for the prediction groups differ, specifically the causal network is based on the top 1000 most DE genes in TS^{wt}_{PO4}, while motifs and similar profiles can be defined across the genome. The ROC can thus be either computed relative to a common universe (such as DE genes, as shown in Fig. 5, main text), or relative to each universe, Supplementary Fig. S5A.
- Indeterminancy in regulatory category assignment/indeterminancy in G^{Dep}. The Expert assignment allowed for a joint assignment, *i.e.* cases where the assignment was unclear. This weighting can be allowed for in the ROC analysis, Fig. S5B, giving essentially similar results.
- An analysis of the sensitivity and specificity of the computational methods to the additional targets found by the comparative signature analysis (CSA).

These analyses demonstrated that there was an effect, but the results are essentially similar. The common versus relative universe has a distinct effect in that for methods utilising a smaller universe their sensitivity drops in absolute terms since there are targets that are inaccessible (lying outside the universe). This is clearly an important issue because there are 184 genes in the full G^{dr} prediction, only 113 are in DE, and only 62 are DE in TS^{wt}_{PO4}. Since our signatures are not defined outside DE, we cannot reliably score the 71 predictions outside of DE. It is likely however, that the number of TP in this unscored set of predictions is low suggesting G^{dr}, has a substantially larger number of FP than discussed here. The probabilistic weighting performance is very similar to the analysis using only uniquely categorised targets, Fig. 5. For detection of the CSA only targets we find that performance is poor, far below a 1:1 detection TP:FP, while the motif and causal predictions are close to random (equal TPR, FPR).

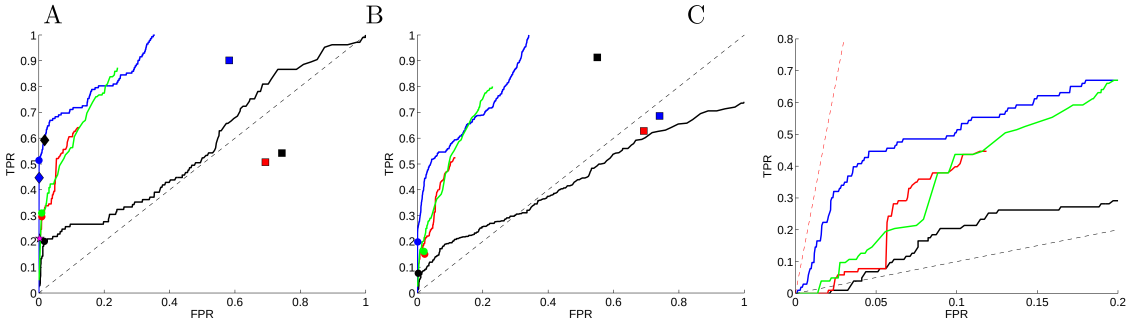


Figure S5. Performance (ROC) of prediction groups under variation of **A** relative universe size, **B** weighted ROC with probabilistic categorisations (by Expert). **C** performance on the additional CSA targets, CSA^{PhoP}. Genes in the group G^{Dep} were removed. The dashed lines are equal TPR and FPR (black, random performance), and equal TP and FP counts (red).

6 Extended Materials and Methods

Cultivations. Three cultivation experiments with high resolution time-series sampling for transcriptome analysis were performed using *S. coelicolor* strains M145 (WT, (Kieser et al., 1999) and

INB201 (*phoP* deletion mutant). Time-series experiments TS^{wt}_{PO4}(WT, phosphate depletion) and TS^{ko}_{PO4}(INB201, phosphate depletion) were performed using the cultivation conditions as described in (Nieselt et al., 2010), featuring phosphate depletion during cultivation after 35hrs and 41hrs, respectively. TS^{wt}_{glu}(WT, glutamate depletion) was performed using identical cultivation conditions except for altered concentrations of glutamate (reduced from 55 to 15 g/L) and phosphate (doubled from 4.6 to 9.2 mM) in the medium. As a consequence, the sole nitrogen and the most important carbon source (glutamate) in this medium is depleted after 35hrs of cultivation, while phosphate remained in excess until the end of the fermentation. These data are further discussed in (Nieselt et al., 2010), and Mast et al., submitted.

Microarray processing. Arrays were processed using RMA with loess normalisation, separately for each time-series. A subsequent normalisation was used to collate the time-series. Quality control measures RLE, NUSE (Brettschneider et al., 2008) were within acceptable ranges. Some arrays were removed because of batch effects that could not be removed by normalisation. Differential expression was ascertained using the BATS package (www.na.iac.cnr.it/bats). SplineCluster software (Heard et al., 2005) was used to cluster groups of genes, priors $10^{-4} - 10^{-2}$ depending on experiment. Gene Ontology analysis (for the differentially expressed genes between experiments and clusters) was performed using BINGO (Maere et al., 2005).

Literature based predicted regulon. G^{lit} is a set of PhoP directly regulated genes identified from 1) experimentally proven PhoP binding sites (gel-retardation and footprinting assays for all operators), 2) previous transcriptomic results (Rodríguez-García et al., 2009), 3) promoter-probe assays, 4) putative transcriptional units. See Supplementary Table S1A for provenance information.

Bioinformatic prediction groups $\mathbf{G}^{dr/word}$. A binding signature (the PHO box) has been identified for PhoP comprising a direct repeat (two linked instances) of the motif GTTCACC in *S. coelicolor* (Sola-Landa et al., 2008), a motif that is conserved throughout the actinomycetes (Mendes et al., 2007), although distinct from those of *E.coli* (CTGTCAT; (Wanner, 1996) and other bacteria (Bacillus subtillus, (TT[A/T]ACA; (Liu and Hulett, 1998)), see Supplementary Fig. S1. This allows bioinformatic (sequence) methods to be used to search for PhoP targets. We constructed the motif using MEME software suite from sequences of experimentally determined PhoP binding sites, Table S1A; this gave a motif (the PHO box) comprising a directed repeat of GTTCA with a spacer of 6 bp, Supplementary Fig. S1. The first word is more conserved than the second. This structure is typical of bacterial transcription factor binding sites which are often of the dyad-type, *i.e.* having two conserved words separated by a variable non-conserved spacer between them. We then searched the *S. coelicolor* genome for the binding sites matching that motif using MEME (http://meme.nbcr.net, (Bailey and C., 1994)). We created the group G^{dr} as genome-wide matches of this model of the PHO box (16 bases) as described above. Genes in the same operon are also included as having this motif (all assigned the highest score of genes in the operon), using the operon definitions of (Charaniya et al., 2007). We also defined a single word motif group G^{word}, *i.e.* genes with just a single binding site comprising the first word in Fig. S1A, specifically an approx. GTTCA word plus the flanking sequences (two letters with background genome frequency on each side of this word). We justify this single-word search since the constrained directed repeat search may miss potential targets because of a noisy middle spacer between the conserved words and the weakness of the second GTTCA motif. Genes in the same operon are again included. The score of each motif in both groups was used to quantify motif quality as used in the ROC analysis. Thresholds 9 and 12.5 for the directed repeat and word motif respectively were used in the groups used in Tables 1,2 (see Supplementary data for a discussion).

$\mathbf{G}^{causal}$. Sparse AR(1) model. We used the sparse linear causal model of (Morrissey et al., 2011), implimented in the GRENITS bioconductor package (submitted), under Gaussian noise extracting the subnetwork of depth 1 outgoing from *phoP* in the full directed gene-gene interaction network. This model implicitly determines the (posterior) probability of a regulatory link being present which was used to quantify quality in the ROC analysis. G^{causal} comprises all genes predicted to be regulated by *phoP* with a posterior link probability > 0.4. Members are given in Supplementary Table S3_Causal.

Similar expression profiles $\mathbf{G}^{simphoP}$. We use a metric methodology to define genes with similar profiles to SCO4230 (*phoP*). We subtract the trace of SCO4230 from each gene profile.

To remove effects such as probe efficiency, we rescale each gene to mean zero, variance 1 prior to subtraction. Those genes within q-th lower quantile of the distribution are selected as G^{simphoP}. For the standard group of Tables 1,2 we use q=0.01. Members are given in Supplementary Table S3_SimilarphoP.

Signature definitions. All signature definitions are dependent on expression profile information, and thus the analysis is restricted to DE genes only. We use the DE groups in each TS to define two of the signatures, *i.e.* a shift to non DE from being DE in TS$^{wt}_{PO4}$. The remaining are defined below. To ascertain appropriate thresholds we used a variety of methods. On all signatures we compared DE genes and non DE, the latter acting as a control set. Typically we used a threshold corresponding to a tail with 1-5% in the non DE gene statistic. Thresholds were ultimately chosen based on an assessment of their discrimination ability (by eye).

Invariant signatures, S^{inv}. We used significant levels of correlation between the TSs, matching the time of nutrient depletion between the TSs. Typically there was a more compact distribution of correlation in both the DE and non DE genes relative to a random walk process, with the former having higher levels of correlation as expected. To define invariant profiles S^{inv} we used a correlation coefficient threshold corresponding to a highly significant correlation, a Bonferroni multiple testing correction at significance 1% giving a distribution tail of the order of 17-19% of the population. A stringent threshold was indicated since the non DE genes had high levels of correlation.

Differential activity signatures S$^{LH/HL}$. Signatures S$^{LH/HL}$ were defined for each gene by taking the ratio of the standard deviation of the TS in each experiment and using a tail threshold on the distribution to determine genes with extremely different activities, from high to low (HL) variation and *vice versa*. To establish an appropriate level for the threshold we used an F-test for the variance ratio, assuming all samples are independent. Tail fractions between 5-23% were used with p-values (F-test) of $10^{-3} - 10^{-4}$. These thresholds gave reasonable discrimination compared to the statistic computed on the non DE genes.

Significant difference in expression profiles S^{diff}. We determined whether genes have a significant difference in their dynamics between two given time-series by subtracting one time-series from the other. Thus, to identify those genes with invariant dynamics from TS$^{wt}_{PO4}$to TS$^{ko}_{PO4}$/TS$^{wt}_{glu}$, we matched the nutrient depletion times in both time-series and then calculated the standard deviation of the difference time-series (sdD); we restrict to common time points, giving approximately 27, 22 time points per comparison, TS$^{wt/ko}_{PO4}$, TS$^{wt}_{PO4/glu}$ respectively. The significantly different genes were those within the q-th higher quantile of the sdD distribution. To determine the threshold we compared to the non DE genes.

Flow chart regulation categories. The regulation categories are defined in terms of the dynamic signatures as follows,

- Solo^{PhoP}: S$^{ko}_{DEnDE}$ & S$^{env}_{DEnDE}$
- CR^{PhoP,glu}: S$^{ko}_{DEnDE}$ & !S$^{env}_{DEnDE}$ This category can be further decomposed into *convergent* coregulation CV^{PhoP,glu}= S$^{ko}_{DEnDE}$ & !S$^{env}_{DEnDE}$ & S$^{env}_{inv}$ where dynamics is identical in TS$^{wt}_{PO4}$and TS$^{wt}_{glu}$
- CR^{PhoP,PO}: S$^{env}_{DEnDE}$ & DE$^{ko}_{PO4}$ & S$^{ko}_{dyn}$
- Integ: DE$^{wt}_{PO4}$ & DE$^{ko}_{PO4}$ & DE$^{wt}_{glu}$ & S$^{ko}_{dyn}$

Here & denotes logical AND, || logical OR, ! NOT. The signature S^{dyn} is a composite signature, S^{dyn}=S^{sigdiff} || S^{HL} || S^{LH} identifying genes with large changes in the dynamic profile. Subscripts on Signatures indicate the comparison pair, ko= TS$^{wt/ko}_{PO4}$, env= TS$^{wt}_{PO4/glu}$.

Expert evaluation of category assignment. The best 20% of predictions from each computational method, and G^{lit} were examined by eye and reassigned based on an assessment of the expression profile for the type of regulation. The regulation category was restricted to apply to dynamics between -1 to 5 hrs of nutrient depletion.

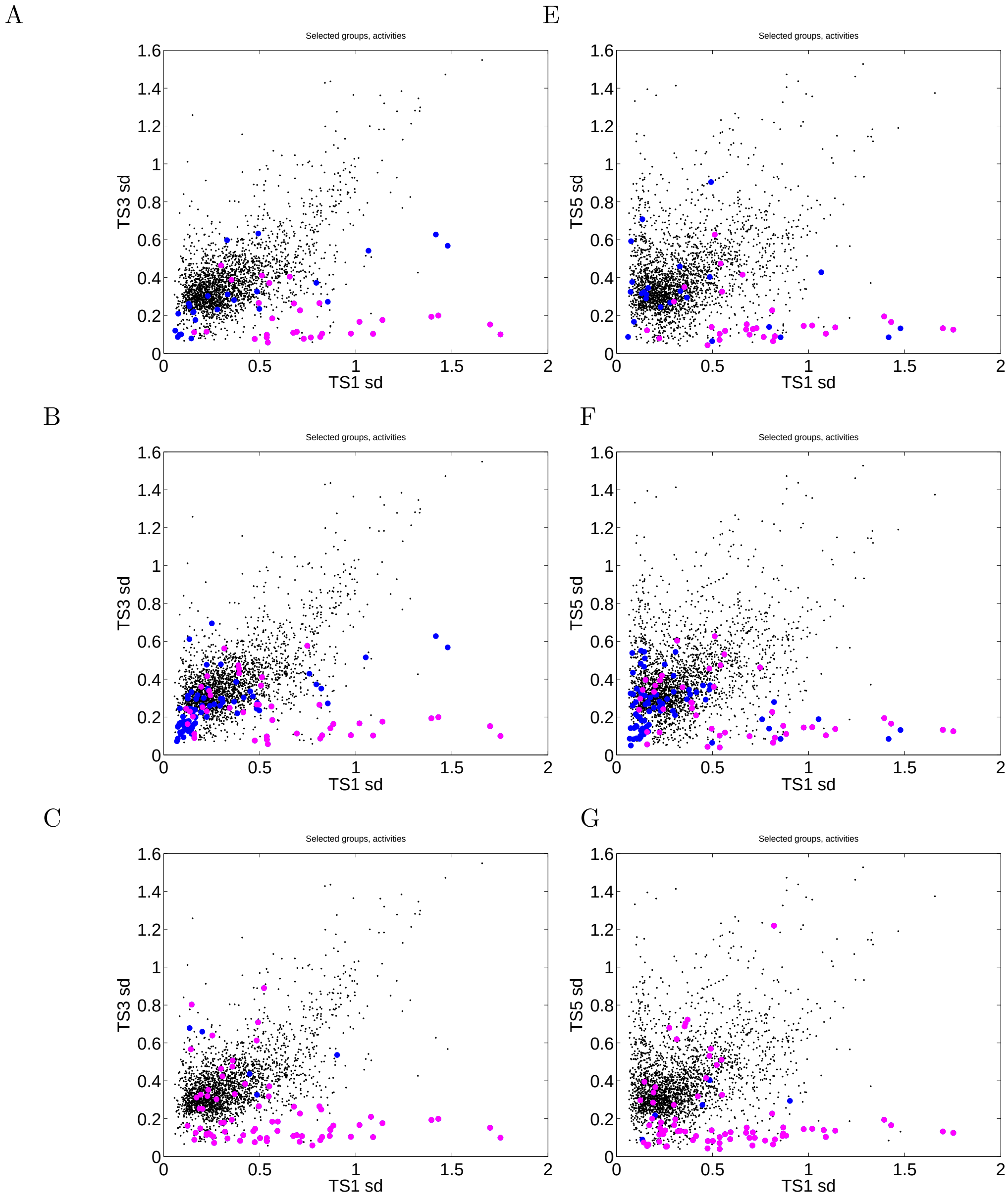
Scoring predictions and Receiver Operating Characteristics (ROC). Under these signatures we score the genes in each group and count the number of genes in the group for which there is ECS, CCE or a lack of an assignment UA relative to PhoP dependence (see text). We can thus quantify the assignment fraction p= (ECS+CCE)/n, where there are n genes in the group, and the precision (positive predictive value) of the group Q=1-E, with error E=CCE/(ECS+CCE). Clearly high assignment and low levels of incompatibility are preferable. To estimate the sensitivity S (true positive rate, TPR) and specificity F (1-FPR, false positive rate FPR) of each group we use our improved predictions of PhoP regulated genes, G^{Dep}, G^{Reg}. Sensitivity is defined by S=ECS_D/m, ECS_D the hits in the reference group G^{Dep} (of size m), and FPR=CCE/m_N, where m_N is the number of true negatives, *i.e.* compliment of G^{Dep} (G^{Reg}) in the universe. The universe differs depending on the set- for $G^{dr/word}$, $G^{simphoP}$ it is the set of all genes (in the genome), and the set of 1000 top differentially expressed genes in $TS^{wt}_{PO4}$for G^{causal}, $G^{phoPclust}$.

ROC curves were plotted using as the reference set either G^{Dep} or as indicated. We used parameters: G^{causal} the posterior link probability, $G^{dr/word}$ the motif threshold and $G^{simphoP}$ the tail percentile q.

7 Supplementary Tables

See main text.

8 Figures



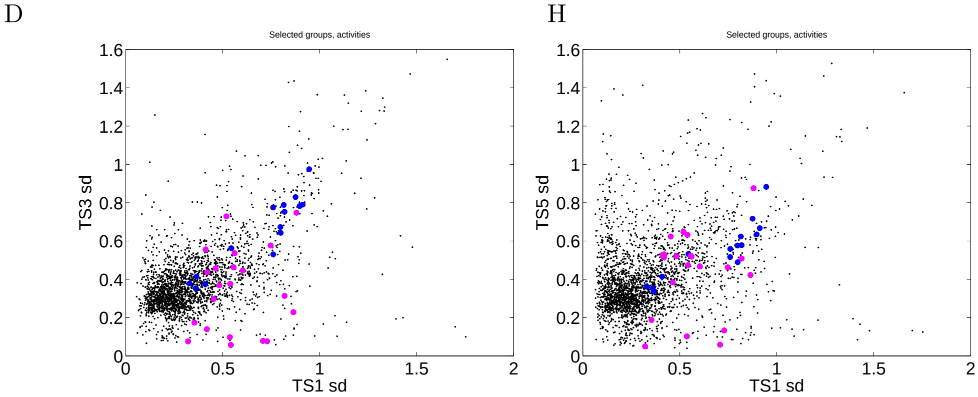


Figure S6. Gene profile variation for TS^{wt}_{PO4}versus $\mathrm{TS}^{ko}_{PO4}$and TS^{wt}_{glu}showing selected prediction groups. **A/E.** 47 genes of the previously compiled *phoP* regulon gene list, G^{lit}, Table S1A. , **B/F.** 88 genes with a dr motif PHO box, G^{dr}, **C/G.** 73 genes with similar dynamic profile to *phoP*, $\mathrm{G}^{simphoP}$, **D/H.** the 36 genes of the causally identified regulon G^{causal}. Only genes DE in either TS^{wt}_{PO4} or TS^{ko}_{PO4} (resp. TS^{wt}_{glu}) are shown (black), genes in each group that are DE in any of the TSs are shown (blue), while those in G^{Dep} are plotted in magenta. For all the prediction groups there is a significant reduction of profile variance for a number of genes in $\mathrm{TS}^{ko}_{PO4}$and TS^{wt}_{glu} relative to TS^{wt}_{PO4}. The average variance across all genes in each experiment is set to 1 giving the scale.

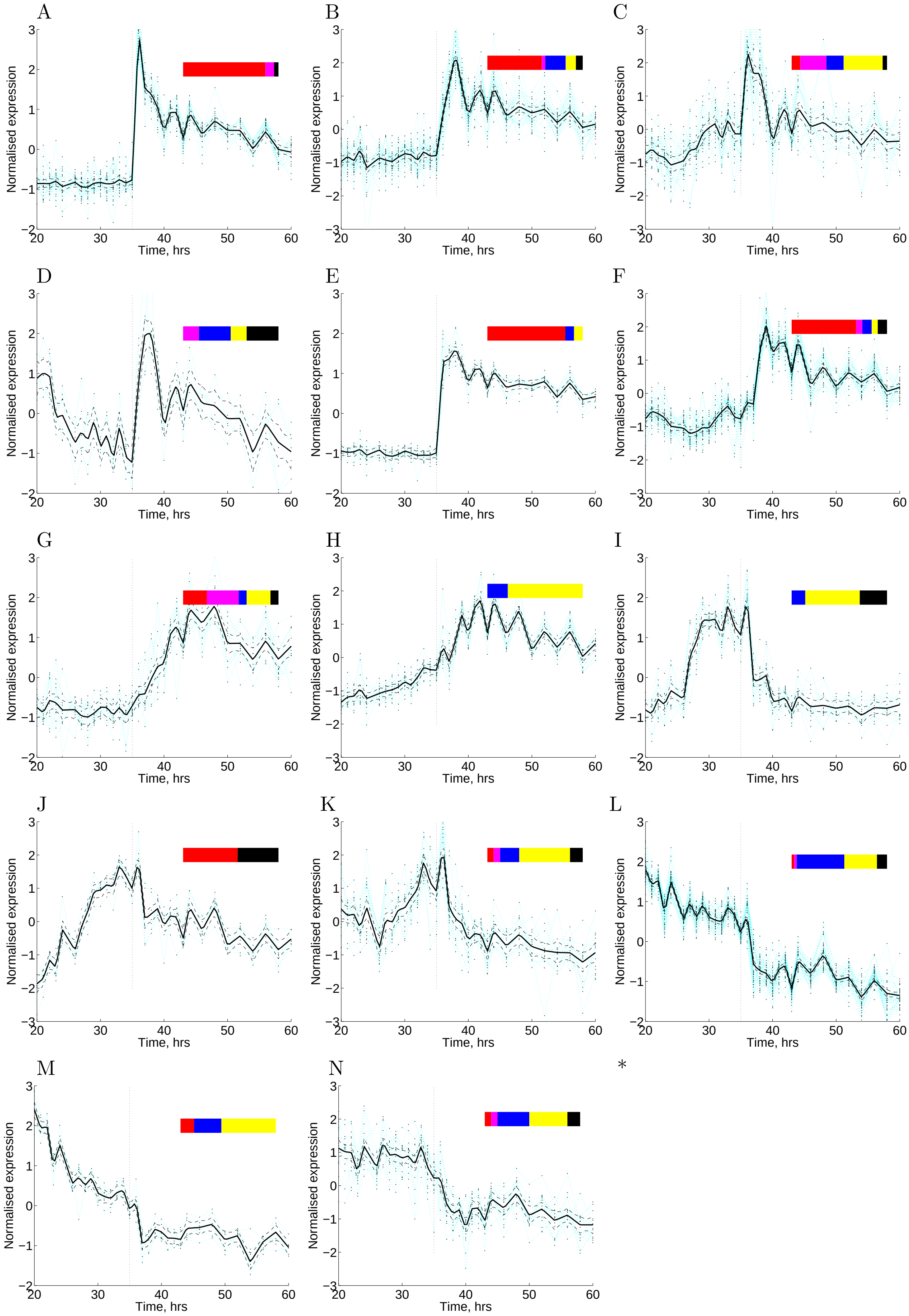


Figure S7. Clusters of gene expression profiles in $\mathrm{TS}^{wt}_{PO4}$of all genes identified as PhoP dependent

(245) from the comparative analysis. Cluster patterns can be grouped as *peak profile*: A (22, all predicted), B (28, 26 predicted), C (22, 19 predicted), D (6, 3 predicted), *activation profile*: E (11, all predicted), F (31, 2 predicted), *progressive activation*: G (12, 3 predicted), H (14, 13 predicted), *inhibition*: I(14, 9 predicted), J(7, none predicted), K(15, 6 predicted), *decreasing expression*: L (38, 16 predicted), M(7, 5 predicted), N(15, 6 predicted). Numbers in parentheses relate to cluster size and the number that are predicted as PhoP dependent from the 7 prediction groups (Table 2). Cluster pattern is shown (solid) with confidence (dashed). Depletion time indicated by the grey vertical line. Fractional decomposition by regulation type is shown as a bar on right with categories Solo^{PhoP} (red), CR^{PhoP,PO} (magenta), CR/CV^{PhoP,glu} (blue), Integ (yellow). Genes without a clear category are shown black. Gene membership is given in Supplementary Table S4_GDep_breakdown.

References

Bailey, T. L. and C., E. (1994) , *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* pp. 28–36

Brettschneider, J., Collin, F., Bolstad, B. M., and Speed, T. P. (2008) , *Technometrics* **50(3)**, 241, 403ID Times Cited:8 Cited References Count:99

Charaniya, S., Mehra, S., Lian, W., Jayapal, K. P., Karypis, G., and Hu, W. S. (2007) , *Nucleic Acids Res* **35(21)**, 7222, Charaniya, Salim Mehra, Sarika Lian, Wei Jayapal, Karthik P Karypis, George Hu, Wei-Shou GM55850/GM/NIGMS NIH HHS/ RLM008713A/PHS HHS/ England Nucleic Acids Res. 2007;35(21):7222-36. Epub 2007 Oct 24.

Heard, N. A., Holmes, C. C., Stephens, D. A., Hand, D. J., and Dimopoulos, G. (2005) , *Proceedings of the National Academy of Sciences of the United States of America* **102(47)**, 16939, 986PW Times Cited:22 Cited References Count:18

Kieser, T., Bibb, M., Buttner, M., Chater, K., and Hopwood, D. (1999) , *John Innes Foundation*

Liu, W. and Hulett, F. M. (1998) , *Microbiology-UK* **144**, 1443, Part 5 Zn779 Times Cited:39 Cited References Count:35

Maere, S., Heymans, K., and Kuiper, M. (2005) , *Bioinformatics* **21(16)**, 3448, Maere, Steven Heymans, Karel Kuiper, Martin England Oxford, England Bioinformatics. 2005 Aug 15;21(16):3448-9. Epub 2005 Jun 21.

Mendes, M. V., Tunca, S., Anton, N., Recio, E., Sola-Landa, A., Aparicio, J. F., and Martin, J. F. (2007) , *Metabolic Engineering* **9(2)**, 217, 151SD Times Cited:24 Cited References Count:45

Morrissey, E., Juarez, M., Denby, K., and Burroughs, N. (2011) , *Biostatistics* **12(4)**, 682

Nieselt, K., Battke, F., Herbig, A., Bruheim, P., Wentzel, A., Jakobsen, O. M., Sletta, H., Alam, M. T., Merlo, M. E., Moore, J., Omara, W. A. M., Morrissey, E. R., Juarez-Hermosillo, M. A., Rodriguez-Garcia, A., Nentwich, M., Thomas, L., Iqbal, M., Legaie, R., Gaze, W. H., Challis, G. L., Jansen, R. C., Dijkhuizen, L., Rand, D. A., Wild, D. L., Bonin, M., Reuther, J., Wohlleben, W., Smith, M. C. M., Burroughs, N. J., Martin, J. F., Hodgson, D. A., Takano, E., Breitling, R., Ellingsen, T. E., and Wellington, E. M. H. (2010) , *BMC Genomics* **11**, 10, 565GY Times Cited:10 Cited References Count:30

Rodríguez-García, A., Sola-Landa, A., Apel, K., Santos-Beneit, F., and Martin, J. F. (2009) , *Nucleic Acids Research* **37(10)**, 3230, Rodriguez-Garcia, Antonio Sola-Landa, Alberto Apel, Kristian Santos-Beneit, Fernando Martin, Juan F England Nucleic Acids Res. 2009 Jun;37(10):3230-42. Epub 2009 Mar 24.

Sola-Landa, A., Rodriguez-Garcia, A., Apel, A. K., and Martin, J. F. (2008) , *Nucleic Acids Research* **36(4)**, 1358, Sola-Landa, Alberto Rodriguez-Garcia, Antonio Apel, Alexander Kristian Martin, Juan F England Nucleic Acids Res. 2008 Mar;36(4):1358-68. Epub 2008 Jan 10.

Wanner, B. L. (1996) , *Kidney Int* **49(4)**, 964, Wanner, B L GM35392/GM/NIGMS NIH HHS/ Kidney Int. 1996 Apr;49(4):964-7.

Chapter 6

Conclusions

6.1 Thesis Summary

Gene regulatory network inference has emerged as a key challenge in the field of systems biology. Network inference refers to the process of determining the structure of a network using expression measurements from the individual network genes. This thesis has focused on network inference, specifically inference using time-series gene expression data. The statistical framework chosen is Dynamic Bayesian Networks coupled with Bayesian Variable Selection.

Network inference models usually model regulatory interactions with linear functions. As the functional form of the interaction is known to be non-linear, we developed a non-linear interaction model and benchmarked it against a linear model using several datasets. The non-linear model outperformed the linear model for most cases studied, with the exception of data with linear interactions, where the performance was similar.

Next we addressed the issue of how to infer a network when replicate measurements are available. An initial approach would be to approximate the true expression value by the mean/median of the replicates. In other words, calculate the mean/median and then run a regular network inference method. By doing this, information on the variability within replicates is discarded. To use the information in the variability we developed two alternative models that account for measurement error; the two models differ in the assumed distribution of the replicates around the true value.

As we used Bayesian statistics for our inference models, it was necessary to choose priors for all parameters. A crucial step of the model building process was to carefully chose these priors and calibrate them adequately. It was necessary to test different types of priors and measure the effect on the posterior. This was in general a non-trivial process given that some of the priors had a very strong effect on the posterior, requiring a large amount of tuning in order for the inference model to work adequately.

The network inference models developed were implemented using MATLAB. While MATLAB is an easy-to-use high-level programming language, it is highly inefficient when evaluating loops. As loops are an unavoidable feature of the inference algorithms, there was room for improvement. We reimplemented the models using `C++` and wrapped them in R. The software is available as the R/Bioconductor package GRENITS. Due to the reimplementation, the algorithms are considerably faster than their MATLAB equivalent. In some cases the increase in speed can be up to 60 fold.

Finally, we used GRENITS to fit a network to a whole-genome gene expression time-series generated by the STREAM consortium. The dataset consisted of 1000 DE genes. In order to run the algorithm on the full dataset, we restricted regulation to those genes known to be able to regulate the expression of other genes. The resulting inferred network was then compared to predictions from other inference methods for the regulon of a specific TF, for which biological knowledge and further experiments were available.

6.1.1 Non-Linear Interactions

In general, the functional form relating the mRNA transcription rate of a given gene to the concentration of the TF regulating it is a sigmoidal curve. If the regulation is positive, for large values of the TF concentration the production rate will saturate and essentially remain unchanged. Depending on the way regulation occurs (e.g multimers) there might also be an initial TF concentration range where little transcription takes place.

An explicit functional form reflecting this regulatory behaviour (Michaelis-Menten function or Hill function) can be derived using enzyme kinetics [Alon, 2006]. Such functional forms have been used to fit simple experimental setups and have proven to describe the behaviour adequately (e.g. [Ronen et al., 2002]). However using Michaelis-Menten or Hill functions for a network inference model has a number of inconvenient features. The data used for inference will be sparse and noisy, which could lead to over-fitting. Also, given the structure of the function, it is not possible to use conjugate priors and Gibbs updates for all parameters.

In order to avoid these problems we chose to use Splines, a family of piecewise polynomial functions frequently used for curve fitting and smoothing. An advantage of using these functions is that they are defined as a linear combination of basis functions, thus the model parameters are the coefficients of a linear combination, meaning that the parameters can be sampled from using Gibbs updates.

Spline functions are highly flexible and if not constrained may over-fit the data. To avoid overfitting it is common to use a prior that will curb the flexibility by biasing the function towards a linear form. This prior in turn introduces a further hyper-parameter that controls the smoothness of the spline function. In our work the prior on the smoothness parameter proved to be crucial. The parameter has two limiting behaviours. When the parameter is very large, the function becomes linear, whereas for small values the function becomes highly non-linear. When using a typical conjugate prior the resulting posterior became very sensitive to the parameters being used, either accumulating too much mass close to the non-linear region or the linear region. After trying different priors we settled on a inverted Pareto distribution. This selection enabled us to select the scale parameter such that linear relationships can be described. Moreover, by carefully selecting its shape parameter we managed to bias the spline fit towards linearity, unless there is strong evidence against it in the data.

A further improvement we developed is related to the uncertainty in network estimation. The usual way of predicting a network from the posterior is to calculate the marginal probability of each network link and then use a threshold to select which links are present. By proceeding in this way, cases where the probability of regulation is split between several genes, thus diluting the individual probabilities, are lost. In order to highlight these cases we developed a network plot that combines marginal link probabilities and the probability of total number of regulators. This plot provides a visual representation of the uncertainty in the retrieval of the network structure/topology.

The resulting non-linear inference model was benchmarked against the AR1 model using different synthetic data sets. We found that both models performed equally well in the case of data with linear interactions. Whereas for non-linear interaction data the non-linear model outperformed the linear model.

6.1.2 Replicate Measurements

Both the linear model and the non-linear model described in the previous sections account for biological noise. The noise represents deviations from the deterministic linear/non-linear model. This form of noise is produced by fluctuations in the concentration of mRNA. In other words, the fluctuation will propagate through the network via the genes regulated by it. When replicates are available for time-series data, if they can be considered as biological replicates, they can be analysed jointly with the linear/non-linear model. Experimental noise on the other hand is not accounted for by the previous models. Fluctuations associated to experimental noise do not propagate through the network. If replicate measurements are available for this kind of data, a network can be inferred by taking the mean/median and producing a single averaged time-series and proceeding as usual. However this approach will discard valuable information about the variability of the expression measurements. In order to use all the information contained in the replicates, we extended the AR1 model to account for experimental noise. We achieved this by assuming that the replicates are centred around a true unmeasured expression value and follow some distribution. We considered two cases. The first case is that the replicates are distributed following a Gaussian distribution and second a Student-t distribution, this way accounting for possible outliers.

The priors used for the newly introduced parameters were all conjugate, except for the degrees of freedom of the Student-t distribution. By testing the model with several datasets and performing a sensitivity analysis, we found that the model is highly sensitive to the prior on the AR1 precision; especially so when the replicates are few and noisy. With this scenario, if a prior concentrates mass at high values of the regression precision, we are essentially implying that there is a network topology that explains the data well. As the information in the replicates is very weak, there will be several combinations of networks and true expression values that fit the data and the prior. All this leads to severe convergence problems. To solve this we selected a prior that is very flat but still concentrates mass close to zero. The resulting model will make weak predictions when there is little information in the replicates and will not have a strong effect when the replicates are informative.

Using several synthetic datasets with replicates we showed that the measurement error models produce improved inference results with respect to running the regular AR1 model on replicate-averaged data. Also, we showed that the Student error model outperformed the Gaussian error model when the data analysed presented outliers. This result was confirmed for synthetic data as well as microarray data.

6.1.3 GRENITS

Four inference models have been presented in this thesis: a linear model, a non-linear model, a Gaussian measurement error model and a Student-t measurement error model. MATLAB code for these models was published with the corresponding paper.

MATLAB can be in general quite fast, especially if the code can be written in a vectorised manner (i.e. avoiding explicit loops). In the case where loops cannot be avoided MATLAB becomes quite inefficient. This is the case for MCMC algorithms, where loops are entirely necessary. Algorithm speed is an issue for the inference models used, as it limits the size of the networks that can be analysed. I reimplemented the four models using `C++`. For matrix algebra I used armadillo, a fast linear algebra library that serves as a wrapper to LAPACK and BLAS functions. A bottleneck of the algorithm and a common function to all models is the Gibbs/MH update. This function was optimised, including using a combination of Cholesky decomposition and triangular optimised functions. The triangular optimised functions were not included in armadillo, so it was necessary to call the LAPACK functions directly.

The full code for all four models was profiled and optimised. The resulting algorithms are considerably faster than the MATLAB code. For instance the STREAM dataset of 1000 genes and 35 regulators, run on a desktop, took just over 6 hours to fit a network. This represents a 60 fold speed up with respect to the MATLAB implementation.

Using the `C++` code I built an R package (GRENITS). I structured the package so that it would be useful to both casual and advanced users. Any of the four methods can be run with a single function, with a time-series data matrix and output folder name as arguments. The hyperparameters and MCMC parameters are by default fixed to values that in our experience work well for a large range of datasets. Once the MCMC has completed, an analysis function can be run that will produce plots and output files. For advanced users I implemented functions to allow modification of the prior parameters, plots of the priors, as well as a function to read MCMC chains. The package has been accepted in Bioconductor and is available from the Bioconductor repository.

6.1.4 Bayesian Network Inference For Large Networks

The datasets used for inference in the previous chapters corresponded to small networks of up to twenty genes. Inference of small networks can be useful if, for example, a set of genes were known to form a network but the wiring of the network was only partially known. A more general scenario is that of a dataset measuring the response of an organism to a specific perturbation or environmental change. This is the case for the dataset generated by the STREAM consortium. The data was generated by subjecting the bacterium *Streptomyces coelicolor* to a environment with a limiting amount of phosphate and measuring gene expression at regular interval during phosphate starvation.

Attempting to fit a network to a whole-genome dataset would be problematic given the speed of the sampling methods described in this thesis. In order to fit a network to the STREAM dataset we first filtered the data to those genes that are dynamically active. This was done by running the BATS differential expression algorithm, which reduced the dataset from around 8000 genes to around 1000 genes. Next we used biological knowledge to find those genes able to regulate the expression of other genes. The total number of regulators found was around 70. The regulators were then clustered so as to group those genes with indistinguishable expression patterns. After clustering, the regulators list was comprised of total of 35 clusters/genes. Using biological criteria we selected one gene from within each cluster to represent the whole cluster. Finally we used GRENITS to fit an AR1 model to the 1000 gene dataset and limited regulation to the 35 regulator-genes. The algorithm ran in approximately 5 hours on a desktop computer. To the best of our knowledge this is the first time such a large network has been run using BVS.

Along with the time-series used for network inference another time-series was generated under the same experimental conditions using a *phoP* KO strain. A third time-series generated using a WT strain subjected to glutamate starvation conditions. By comparing gene expression dynamics across the three time-series, it was possible to compile a list of genes with evidence for regulation by *phoP*. The list was then used to compare the performance of the inferred network against correlation methods and sequence analysis methods. The best method appeared to be the sequence based method, although if the network inference predictions were complemented with genes strongly correlated with *phoP* they outperformed the rest. Merging the network inference predictions with genes strongly correlated with *phoP* is justified by the fact that co-expressed genes are known to be potential regulated genes and are not detectable by an AR1 network inference method.

6.2 Comparison to Other Published Models

The main focus of this thesis has been the implementation of a computationally efficient AR1 model with BVS and the development of improvements to this model in order to address key issues (e.g. non-linear interactions). The improved models were benchmarked against the AR1 model, thus allowing for the explicit assessment of the effect of the model improvements. However, an issue that has not been addressed is how well the basic AR1 model and the improved models compare to other published methods.

method	algorithm	network 1	network 2	network 3	network 4	network 5
DBN	G1DBN	0.73(0.37)	0.64(0.34)	0.68(0.45)	0.85(0.69)	**0.92**(0.77)
	VBSSM	0.73(0.38)	0.66(0.41)	**0.77**(0.49)	0.80(0.46)	0.84(0.64)
	LinearNet	***0.85(0.69)***	*0.73(**0.57**)*	*0.75(0.49)*	*0.90(0.82)*	*0.71(0.54)*
	NonLinearNet	***0.85***_(0.64)_	*0.74(0.53)*	*0.76(**0.55**)*	***0.91(0.83)***	*0.76(0.57)*
ODE	TSNI	0.62(0.27)	0.63(0.32)	0.58(0.21)	0.63(0.23)	0.68(0.25)
NDS	GP4GRN	0.66(0.42)	0.69(0.44)	0.70(0.47)	0.62(0.35)	0.86(0.65)
	CSI^d	0.72(0.64)	**0.75**(0.54)	0.67(0.45)	0.83(0.67)	0.90(**0.78**)
	CSI^c	0.78(0.42)	0.73(0.40)	0.66(0.29)	0.64(0.26)	0.75(0.27)
GC	GCCA	0.67(0.30)	0.70(0.47)	0.62(0.26)	0.80(0.56)	0.80(0.58)
random	random	0.55(0.18)	0.55(0.19)	0.55(0.17)	0.57(0.17)	0.56(0.16)

Table 6.1: Performance of different network inference methods. The data used was generated *in-silico* using five different ten gene networks. Scores used are AUROC/AUPR, with AUROC appearing first and AUPR next in brackets. Bold font indicates best score. CSI^c evolves over continuous time and CSI^d over discrete time. Italicised algorithms/scores indicate thesis algorithms/scores.

We can, to some extent, address this question by drawing on the model comparison work published in Penfold and Wild [2011]. In brief, for this work the authors used time-series data generated from ten *in-silico* networks (five ten gene networks and five one hundred-gene networks) and two *in-vivo* networks. With this data they then inferred networks using seven different network inference methods and used the known true networks to assess the quality of the predictions. The scores used to assess the quality of the predictions were the area under the receiver operating curve (AUROC) and the area under the precision recall curve (AUPR).

The seven methods chosen were separated into the categories of: ordinary differential equations (TSNI), Dynamic Bayesian Networks (G1DBN, VBSSM), nonlinear dynamical systems (GP4GRN, CSI) and Granger causality (GCCA). For the sake of brevity the reader is referred to the original paper [Penfold and Wild, 2011] for details on the models and the data sets.

Using GRENITS we fitted the models presented in this thesis to data from Penfold and Wild [2011] (data provided by C. Penfold). As GRENITS does not allow for the use of multiple time-series representing different experimental condition, the time-series were concatenated. The single concatenated time-series has inconsistent time-steps where the individual time-series concatenate–though this represents a small percentage of the time-steps ($\approx 4\%$).

As the network from the *Arabidopsis thaliana* data-set is only partially known, we did not include it in the comparison. Also, of the data used, only the yeast data has replicates, thus we could only run the replicates model with the yeast

method	algorithm	network 1	network 2	network 3	network 4	network 5
DBN	G1DBN	0.68(0.11)	0.64(0.10)	0.68(0.13)	0.66(0.10)	0.72(0.11)
	VBSSM^1	0.59(0.08)	0.56(0.05)	0.59(0.11)	0.67(0.10)	0.71(0.09)
	VBSSM	0.56(0.09)	0.57(0.06)	0.62(0.12)	0.64(0.12)	0.70(0.09)
	LinearNet	*0.78(0.26)*	*0.69(0.11)*	*0.75(0.17)*	*0.71(0.23)*	*0.69(0.16)*
	NonLinearNet	***0.82(0.27)***	*__0.72__(0.13)*	*__0.76__(0.20)*	*__0.74__(0.22)*	*__0.74__(0.15)*
ODE	TSNI	0.55(0.02)	0.55(0.03)	0.60(0.03)	0.54(0.02)	0.59(0.03)
NDS	GP4GRN	0.72(0.22)	0.62(0.10)	0.70(0.16)	0.70(0.21)	0.69(0.12)
	CSI^d	0.71(0.25)	0.67(**0.17**)	0.71(**0.25**)	**0.74**(**0.24**)	0.73(**0.26**)
	CSI^c	0.65(0.13)	0.56(0.03)	0.63(0.07)	0.61(0.07)	0.60(0.05)
GC	GCCA	0.60(0.04)	0.57(0.04)	0.60(0.07)	0.58(0.07)	0.57(0.03)
random	random	0.50(0.002)	0.50(0.002)	0.50(0.002)	0.50(0.002)	0.50(0.002)

Table 6.2: Performance of different network inference methods. The data used was generated *in silico* using five different hundred gene networks. Scores used are AUROC/AUPR, with AUROC appearing first and AUPR next in brackets. Bold font indicates best score. CSI^c evolves over continuous time and CSI^d over discrete time. VBSSM^1 uses a single hidden state. Italicised algorithms/scores indicate thesis algorithms/scores.

data.

The results of the analysis have been added to the original tables presented in Penfold and Wild [2011] and reproduced in tables 6.1, 6.2 and 6.3.

Table 6.1 shows the results of inference on the ten gene network data. In general the performance of the two methods is good, coming first in two out of the five networks with respect to the AUROC score and four out of five with respect to the AUPR score. Of the two methods, the NonLinearNet model performs better than the LinearNet model for most networks under both scores.

For the case of the hundred gene networks (table 6.2) the NonLinearNet ranked first or joint first for all networks under the AUC score. Under the AUPR score, the NonLinearNet model ranked first in one network and second or third in all other networks. The results of the Linear model are also good, scoring second best under the AUROC score in three out of five cases and also three out of five for the AUPR score.

As mentioned earlier the experimental data was measured in replicate (four replicates for the switch off experiment and five for the switch on) allowing the use of all the methods available through GRENITS. The results can be seen in table 6.3. Here the results are less favourable for the thesis methods, especially for the non-replicate methods (LinearNet and NonLinearNet), both of which score close to random under both AUROC and AUPR scores for the switch-off data. On the

method	algorithm	switch on	switch off
DBN	G1DBN	0.78(0.64)	0.61(0.34)
	VBSSM	0.79(**0.70**)	0.76(0.60)
	LinearNet	*0.76(0.48)*	*0.65(0.38)*
	NonLinearNet	*0.74(0.46)*	*0.65(0.38)*
	GaussReps	***0.80**(0.65)*	*0.72(0.45)*
	StudentReps	*0.78(0.59)*	*0.73(0.52)*
ODE	TSNI	0.68(0.51)	0.68(0.42)
NDS	GP4GRN	0.73(0.61)	0.76(0.57)
	CSI^d	0.63(0.46)	**0.86**(**0.72**)
	CSI^c	0.64(0.39)	0.73(0.59)
GC	GCCA	0.71(0.55)	0.74(0.65)
random	random	0.65(0.45)	0.65(0.45)

Table 6.3: Performance of different network inference methods. The data corresponds to two experiments measuring genes in a synthetic biology engineered five gene network. Scores used are AUROC/AUPR, with AUROC appearing first and AUPR next in brackets. Bold font indicates best score. CSI^c evolves over continuous time and CSI^d over discrete time. Italicised algorithms/scores indicate thesis algorithms/scores.

other hand, the replicates methods do reasonably well, scoring better than both the non replicates methods for both data sets. For the switch-on data, the GaussianReps model shows the best score out of all methods compared under the AUROC score and scores second under the AUPR score.

In terms of timing, one of two chains run on a single 2.53GHz processor took approximately 4 seconds, 12 seconds and 35 minutes for the LinearNet to run a five gene data-set, ten gene data-set and one hundred gene data-set respectively. The NonLinearNet took 1.2 minutes to run a five gene data-set, 6.7 minutes to run a ten gene data-set and 3 days for a one hundred gene data-set. Both the replicates models took under 15 seconds to run a five gene data-set.

A small fraction ($\approx 0.3\%$) of the network links did not fully converge with the default parameters for the NonLinearNet 100 gene networks. To reduce this problem, we pooled the samples from both chains.

Worthy of mention is the fact that the models that we are benchmarking against share similarities with the thesis models. For instance, G1DBN uses an AR1 model identical to the Linear model, though the way the model is fitted to the data is completely different. Also the replicates models share some similarities to VBSSM [Beal et al., 2005] as the replicates models classify as State Space Models (SSM). Though for the replicates models only the hidden factors follow an AR1 process. In this sense, the models are more similar to the SSM of Hirose et al. [2008] and Wu et al. [2004], though in the replicates model the number of hidden factors is fixed and each map to a unique set of replicates.

6.3 Final Reflections

From the analysis of the STREAM data it is clear that network inference can be a useful tool, as can be seen by the fact that within the *phoP* predicted targets a small number were not accounted for in the literature and were later proven to be correct by the KO data. The analysis of the STREAM data also showed that the predictions have to be handled with caution as they contain a large number of false positives, implying that expert biological knowledge is also needed.

The methods presented in this thesis aimed at improving upon certain approximations commonly used, specifically the use of linear interactions and discarding the information within the replicates. The improved models showed good performance for synthetic data that met the model assumptions; even in the case where the data-generating-model was different to the inference model (Locke data, Locke et al. [2006]). When the network models were fit to experimental data the improvements were appreciable though small. As the model improvements reduce the algorithm speed, from a practical point of view it may be better to try to find other approximations that have a stronger negative effect and address them; for example including unmeasured protein concentrations.

6.4 Future Work

6.4.1 Extensions of Presented Work

The work presented in this thesis has relied on a number of assumptions, for instance, the replicates model assumes that the individual replicates are not correlated across time. This is not always the case. It would be possible to extend the model to account for this correlation by assuming the correlated replicates follow an AR1 process with two forms of noise, biological and experimental.

Also, for the network analysis of the STREAM data clustering was performed and a gene from each cluster was selected to represent the cluster. This was done by choosing a gene thought to be most biologically relevant. It would be interesting to assess the robustness of the inferred network by selecting different genes to represent the clusters.

6.4.2 Regulatory Protein Concentration

One of the approximations used for the models included in this thesis, is to assume that the concentration of regulating protein can be approximated by a scaled version of the concentration of mRNA of the corresponding regulator. This approximation will not always hold and may lead to scenarios that affect the accuracy of the retrieved network.

The effect of approximating the concentration of TF protein from its mRNA concentration can be visually explored for Locke et al.'s circadian clock model. By plotting the mRNA production rate for genes with a single regulator against protein concentration of the TF regulating it the regulating function can be reconstructed. If instead we plot the production rate against a surrogate measurement, such as the TF mRNA concentration, we can visualise the distorting effect.

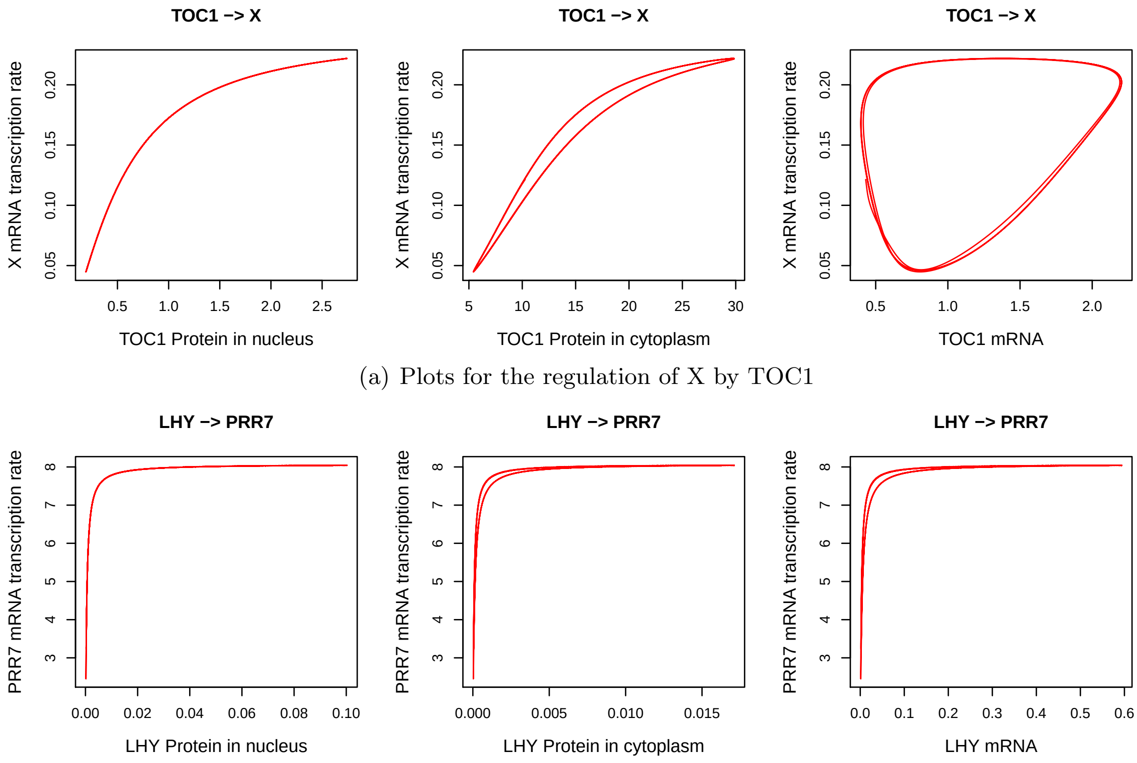
For Locke et al's model, regulation is directly dependant on TF protein in the nucleus, although values for protein in the cytoplasm are also available. Figure 6.1(a) and figure 6.1(b) show the transcription rate plotted against protein in the nucleus (left), protein in the cytoplasm (middle) and mRNA (right) for the regulation of TOC1 by gene X and the regulation of PRR7 by LHY respectively. In both cases the left hand plot (protein in the nucleus) shows the correct Hill function. When using protein in the cytoplasm as a surrogate for protein in the nucleus, neither regulation shows a large disruption, but when using mRNA (right plot), the approximation is clearly very poor for the TOC1-X interaction. On the other hand for the PRR7-LHY interaction, the approximation appears to be reasonable.

The AR1 model could be expanded to include the concentration of the regulatory proteins. The equations would be:

$$\begin{aligned} X^{t+1} &= Ay^t + \mu + \nu \\ y^{t+1} &= ay^t + X^t + \epsilon, \end{aligned}$$

where A is the network coefficients matrix, μ is the baseline transcription level for the regulated gene, X is the measured expression data, y is the unmeasured protein concentration, a is a protein decay term and both ν and ϵ are Gaussian distributed noise. While the translation rate does not appear explicitly in the equations, it has been accounted for by rescaling A, y and ϵ.

Given the lack of information on the protein measurements, it is possible that several networks could give similar data-fits, which could lead to strong convergence issues. This implies that the priors would have to be carefully chosen and tuned.



(a) Plots for the regulation of X by TOC1

(b) Plots for the regulation of PRR7 by LHY

Figure 6.1: Plots of the transcription rate (see equation 1.3) for a regulated gene with a single regulator versus: the concentration of the regulator protein in the nucleus, in the cytoplasm and the regulator's mRNA. All the data was generated using the ODE model from Locke et al. [2006].

6.4.3 Data Integration

Even in the case that a time-series has many more time-points than parameters, it may not contain sufficient information to determine the full network structure. For example if two regulators share the same dynamic trace, it will be impossible to determine which of the two might be regulating a third gene.

Potentially more information may be found by combining datasets from different experimental conditions or perturbations. This raises the question of how to integrate disparate datasets. For the case of multiple time-series the plain AR1 model could be used to jointly analyse the data, thus assuming that the AR1 parameters are identical for each time-series (see for example Bonneau et al. [2006]). This assumption can be weakened by using a hierarchical structure to allow different parameters in each experiment while still sharing information across time-series. Another approach could be to just allow the network topology to be the same across datasets and let each dataset have its own AR1 parameters.

A problem that could hamper the integration of different experiments is topology rewiring. Certain molecules, such as metabolites, can modify the activity of a TF by either activating, inhibiting or changing the sign of the regulation (see for example [Cho et al., 2008]). As only the expression is used to infer a network, a regulator may be expressed but not regulating in one experiment whilst it could be regulating in another experiment, depending on whether a specific molecule is present or not. In other words the topology changes due to the presence or absence of unmeasured molecules. If the molecule is present in certain experiments and not in others, the data being merged contains contradictory information.

Network inference using expression data could be improved by integrating different data types. Complementary data that could be integrated in the model include metabolomic data, in order to predict TF activity modification. Proteomic data could also help improve predictions, especially if the data has information on protein state (e.g phosphorylation state).

TF binding information could also be integrated as priors on the BVS variables. Binding information can be found using for example ChIP data or putative binding site data when a consensus motif is available.

6.4.4 Biologically Derived Priors

So far, the priors used for the coefficients of the AR1 model are chosen to be as uninformative as possible. The data is scaled, which will in turn scale the coefficients. Following this, a Gaussian centred at zero and with a large variance is used as prior for all coefficients. This could lead to inference of a model with biologically infeasible parameters, for example a regulatory interaction with an unrealistically fast transcription rate. It would be interesting to construct informative priors for the coefficients using information such as number of nucleotides per gene or maximum transcription speed.

Biologically motivated priors could also be used to reduce problems with the self interaction term. The trace of mRNA concentration is a continuous (noisy) function, meaning that the correlation between one time-point and the next for the same gene will be high. When fitting an AR1 model this high correlation can lead to the self interaction term playing an excessively prominent role. As the self interaction term comes from decay and mass conservation, including information on mRNA decay could help reduce this problem. Using a measured decay term in the context of network inference has been done before by Barenco et al. [2006]. Using a transcription blocking molecule, mRNA decay traces were measured in the same experimental conditions as the experiments used for network inference. Exponential functions were fitted to the data, thus finding the decay parameters for each gene. The inferred parameters were then fixed for the inference model and used to fit to data from other experimental conditions. Whole genome studies on mRNA decay can be found for different organisms (see for example [Selinger et al., 2003]). It would be interesting to study whether the use of measured decay can improve the quality of inferred networks.

6.4.5 Quantitative Networks

While the models covered in this thesis have concentrated on inference of the network topology, it would be interesting to see if the quantitative aspect can be used to solve further problems. An example of this would be the work done by di Bernardo et al. [2005]. The authors fit a linear network model to a compendium of about 500 microarrays. The resulting network was used as a tool to analyse microarrays where certain genes had been perturbed. The network allowed the separation between genes that had been perturbed directly and genes affected by a network knock on effect.

The models developed in this thesis could be modified to analyse data in a similar way. Using the AR1 model it would be possible to elicit priors of known network interactions, this way inferring a more accurate network. Also the non-linear model could be used, potentially improving the predictions.

Bibliography

Tarmo Äijö and Harri Lähdesmäki. Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics. *Bioinformatics*, (22):2937–2944, 2009.

Uri Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman and Hall, Boca Raton, 2006.

Claudia Angelini, Luisa Cutillo, Daniela De Canditiis, Margherita Mutarelli, and Marianna Pensky. BATS: a Bayesian user-friendly software for Analyzing Time Series microarray experiments. *BMC Bioinformatics*, (1):415+, 2008.

Mukesh Bansal, Giusy D. Gatta, and Diego di Bernardo. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics*, (7):815–822, 2006.

Martino Barenco, Daniela Tomescu, Daniel Brewer, Robin Callard, Jaroslav Stark, and Michael Hubank. Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome Biology*, (3):R25+, 2006.

Matthew J. Beal and Zoubin Ghahramani. The Variational Bayesian EM Algorithm for Incomplete Data: with Application to Scoring Graphical Model Structures. *Bayesian Statistics*, 2003.

Matthew J. Beal, Francesco Falciani, Zoubin Ghahramani, Claudia Rangel, and David L. Wild. A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics*, (3):349–356, 2005.

S. D. Bentley, K. F. Chater, A. M. Cerdeno-Tarraga, G. L. Challis, N. R. Thomson, K. D. James, D. E. Harris, M. A. Quail, H. Kieser, D. Harper, A. Bateman, S. Brown, G. Chandra, C. W. Chen, M. Collins, A. Cronin, A. Fraser, A. Goble, J. Hidalgo, T. Hornsby, S. Howarth, C. H. Huang, T. Kieser, L. Larke, L. Murphy, K. Oliver, S. O'Neil, E. Rabbinowitsch, M. A. Rajandream, K. Rutherford,

S. Rutter, K. Seeger, D. Saunders, S. Sharp, R. Squares, S. Squares, K. Taylor, T. Warren, A. Wietzorrek, J. Woodward, B. G. Barrell, J. Parkhill, and D. A. Hopwood. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature*, (6885):141–147, 2002.

Johannes Berg. Out-of-equilibrium dynamics of gene expression and the Jarzynski equality. *Physical review letters*, (18), 2008.

Jose M. Bernardo. Reference Posterior Distributions for Bayesian Inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, (2):113–147, 1979.

Jose M. Bernardo and Adrian F. M. Smith. *Bayesian Theory.* Wiley, Chichester, 1994.

R. Bonneau. A Predictive Model for Transcriptional Control of Physiology in a Free Living Cell. *Cell*, (7):1354–1365, 2007.

Richard Bonneau, David Reiss, Paul Shannon, Marc Facciotti, Leroy Hood, Nitin Baliga, and Vesteinn Thorsson. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biology*, (5):R36+, 2006.

Stephen P. Brooks. Markov Chain Monte Carlo Method and Its Application. *Journal of the Royal Statistical Society. Series D*, (1):69–100, 1998.

George Casella and Edward I. George. Explaining the Gibbs Sampler. *The American Statistician*, (3):167–174, 1992.

M. Chalfie, Y. Tu, G. Euskirchen, W. W. Ward, and D. C. Prasher. Green fluorescent protein as a marker for gene expression. *Science*, (5148):802–805, 1994.

Byung-Kwan Cho, Christian L. Barrett, Eric M. Knight, Young S. Park, and Bernhard Palsson. Genome-scale reconstruction of the Lrp regulatory network in *Escherichia coli*. *Proceedings of the National Academy of Sciences*, (49):19462–19467, 2008.

Elissa J. Cosgrove, Yingchun Zhou, Timothy S. Gardner, and Eric D. Kolaczyk. Predicting gene targets of perturbations via network-based filtering of mRNA expression compendia. *Bioinformatics*, (21):2482–2490, 2008.

Riet De Smet and Kathleen Marchal. Advantages and limitations of current network inference methods. *Nature Reviews Microbiology*, (10):717–729, 2010.

P. Dellaportas, J. Forster, and I. Ntzoufras. On Bayesian model and variable selection using MCMC. *Technical report*, 1997.

Diego di Bernardo, Michael J. Thompson, Timothy S. Gardner, Sarah E. Chobot, Erin L. Eastwood, Andrew P. Wojtovich, Sean J. Elliott, Scott E. Schaus, and James J. Collins. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nature Biotechnology*, (3):377–383, 2005.

Michael B. Elowitz and Stanislas Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, (6767):335–338, 2000.

Jeremiah J. Faith, Boris Hayete, Joshua T. Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J. Collins, and Timothy S. Gardner. Large-Scale Mapping and Validation of *Escherichia coli* Transcriptional Regulation from a Compendium of Expression Profiles. *PLoS Biol*, (1):e8+, 2007.

Ari E. Friedland, Timothy K. Lu, Xiao Wang, David Shi, George Church, and James J. Collins. Synthetic Gene Networks That Count. *Science*, (5931):1199–1202, 2009.

Timothy S. Gardner, Charles R. Cantor, and James J. Collins. Construction of a genetic toggle switch in *Escherichia coli*. *Nature*, (6767):339–342, 2000.

Timothy S. Gardner, Diego di Bernardo, David Lorenz, and James J. Collins. Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling. *Science*, (5629):102–105, 2003.

Marina Granovskaia, Lars Jensen, Matthew Ritchie, Joern Toedling, Ye Ning, Peer Bork, Wolfgang Huber, and Lars Steinmetz. High-resolution transcription atlas of the mitotic cell cycle in budding yeast. *Genome Biology*, (3):R24+, 2010.

Jim E. Griffin and Philip J. Brown. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, (1):171–188, 2010.

Marco Grzegorczyk. An introduction to Gaussian Bayesian networks. *Methods in molecular biology*, pages 121–147, 2010.

Chris Hans. Bayesian lasso regression. *Biometrika*, (4):835–845, 2009.

Nicholas A. Heard, Christopher C. Holmes, David A. Stephens, David J. Hand, and George Dimopoulos. Bayesian coclustering of Anopheles gene expression time series: Study of immune defense response to multiple experimental challenges.

Proceedings of the National Academy of Sciences of the United States of America, (47):16939–16944, 2005.

Osamu Hirose, Ryo Yoshida, Seiya Imoto, Rui Yamaguchi, Tomoyuki Higuchi, D. Stephen Charnock-Jones, Cristin Print, and Satoru Miyano. Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models. *Bioinformatics*, 2008.

Antti Honkela, Charles Girardot, E. Hilary Gustafson, Ya-Hsin Liu, Eileen E. M. Furlong, Neil D. Lawrence, and Magnus Rattray. Model-based method for transcription factor target identification with limited data. *Proceedings of the National Academy of Sciences*, (17):7793–7798, 2010.

Jonathan Houseley and David Tollervey. The Many Pathways of RNA Degradation. *Cell*, (4):763–776, 2009.

Rafael A. Irizarry, Bridget Hobbs, Francois Collin, Yasmin D. Beazer-Barclay, Kristen J. Antonellis, Uwe Scherf, and Terence P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, (2):249–264, 2003.

H. Jeffreys. *Theory of probability.* Clarendon Press, Oxford, 1998.

Harold Jeffreys. An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, (1007):453–461, 1946.

Robert E. Kass and Adrian E. Raftery. Bayes Factors. *Journal of the American Statistical Association*, (430):773–795, 1995.

Sunyong Kim, Seiya Imoto, and Satoru Miyano. Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. *Bio Systems*, (1-3):57–65, 2004.

James C. Liao, Riccardo Boscolo, Young-Lyeol Yang, Linh M. Tran, Chiara Sabatti, and Vwani P. Roychowdhury. Network component analysis: Reconstruction of regulatory signals in biological systems. *Proceedings of the National Academy of Sciences*, (26):15522–15527, 2003.

James C. Locke, László Kozma-Bognár, Peter D. Gould, Balázs Fehér, Eva Kevei, Ferenc Nagy, Matthew S. Turner, Anthony Hall, and Andrew J. Millar. Experimental validation of a predicted feedback loop in the multi-oscillator clock of Arabidopsis thaliana. *Molecular systems biology*, 2006.

David R. Lorenz, Charles R. Cantor, and James J. Collins. A network biology approach to aging in yeast. *Proceedings of the National Academy of Sciences*, (4): 1145–1150, 2009.

Adam Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Favera, and Andrea Califano. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*, (Suppl 1):S7+, 2006.

C. Robertson McClung. Comes a time. *Current Opinion in Plant Biology*, (5): 514–520, 2008.

E. R. Morrissey, M. A. Juárez, K. J. Denby, and N. J. Burroughs. On reverse engineering of gene interaction networks using time course data with repeated measurements. *Bioinformatics*, (18):2305–2312, 2010.

Sven Nelander, Weiqing Wang, Bjorn Nilsson, Qing-Bai She, Christine Pratilas, Neal Rosen, Peter Gennemark, and Chris Sander. Models from experiments: combinatorial drug perturbations of cancer cells. *Molecular Systems Biology*, 2008.

Kay Nieselt, Florian Battke, Alexander Herbig, Per Bruheim, Alexander Wentzel, Oyvind Jakobsen, Havard Sletta, Mohammad Alam, Maria Merlo, Jonathan Moore, Walid Omara, Edward Morrissey, Miguel J. Hermosillo, Antonio R. Garcia, Merle Nentwich, Louise Thomas, Mudassar Iqbal, Roxane Legaie, William Gaze, Gregory Challis, Ritsert Jansen, Lubbert Dijkhuizen, David Rand, David Wild, Michael Bonin, Jens Reuther, Wolfgang Wohlleben, Margaret Smith, Nigel Burroughs, Juan Martin, David Hodgson, Eriko Takano, Rainer Breitling, Trond Ellingsen, and Elizabeth Wellington. The dynamic architecture of the metabolic switch in *Streptomyces coelicolor*. *BMC Genomics*, (1):10+, 2010.

Tania Nolan, Rebecca E. Hands, and Stephen A. Bustin. Quantification of mRNA using real-time RT-PCR. *Nature Protocols*, (3):1559–1582, 2006.

R O'Hara. A Review of Bayesian Variable Selection Methods: What, How and Which . *Bayesian Analysis*, (1):85 – 117, 2009.

Peter J. Park. ChIPseq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, (10):669–680, 2009.

Judea Pearl. Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning. *Proceedings of the 7th Conference of the Cognitive Science Society*, 1985.

Judea Pearl. *Causality: models, reasoning, and inference.* Cambridge University Press, Cambridge, 2000.

Richard Pearson, Xuejun Liu, Guido Sanguinetti, Marta Milo, Neil Lawrence, and Magnus Rattray. puma: a Bioconductor package for propagating uncertainty in microarray analysis. *BMC Bioinformatics*, (1):211+, 2009.

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

Arjun Raj, Scott A. Rifkin, Erik Andersen, and Alexander van Oudenaarden. Variability in gene expression underlies incomplete penetrance. *Nature*, (7283):913–918, 2010.

Graham Ramsay. DNA chips: State-of-the art. *Nature Biotechnology*, (1):40–44, 1998.

Michal Ronen, Revital Rosenberg, Boris I. Shraiman, and Uri Alon. Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proceedings of the National Academy of Sciences of the United States of America*, (16):10555–10560, 2002.

Chiara Sabatti and Gareth M. James. Bayesian sparse hidden components analysis for transcription regulation networks. *Bioinformatics*, (6):739–746, 2006.

Eran Segal, Michael Shapira, Aviv Regev, Dana Pe'er, David Botstein, Daphne Koller, and Nir Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, (2):166–176, 2003.

Douglas W. Selinger, Rini M. Saxena, Kevin J. Cheung, George M. Church, and Carsten Rosenow. Global RNA Half-Life Analysis in *Escherichia coli* Reveals Positional Patterns of Transcript Degradation. *Genome Research*, (2):216–223, 2003.

Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, (1):267–288, 1996.

Ernest Turro, Natalia Bochkina, Anne-Mette M. Hein, and Sylvia Richardson. BGX: a Bioconductor package for the Bayesian integrated analysis of Affymetrix GeneChips. *BMC Bioinformatics*, (1):439+, 2007.

Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, (1):57–63, 2009.

Adriano V. Werhli and Dirk Husmeier. Reconstructing gene regulatory networks with Bayesian Networks by combining expression data with multiple sources of prior knowledge. *Statistical Applications in Genetics and Molecular Biology*, 2007.

Darren J. Wilkinson. Stochastic modelling for quantitative description of heterogeneous biological systems. *Nature Reviews Genetics*, (2):122–133, 2009.

Neil R. Williamson, Peter C. Fineran, Finian J. Leeper, and George P. C. Salmond. The biosynthesis and regulation of bacterial prodiginines. *Nature Reviews Microbiology*, pages 887–899, 2006.

F.X. Wu, W.J. Zhang, and A.J. Kusalik. Modeling Gene Expression from Microarray Expression Data with State-Space Equations. *Pacific Symposium on Biocomputing*, 2004.

Shizhong Xu. Estimating Polygenic Effects Using Markers of the Entire Genome. *Genetics*, (2):789–801, 2003.

Zhilong Yang, Daniel P. Bruno, Craig A. Martens, Stephen F. Porcella, and Bernard Moss. Simultaneous high-resolution analysis of vaccinia virus and host cell transcriptomes by deep RNA sequencing. *Proceedings of the National Academy of Sciences*, (25):11513–11518, 2010.

Nengjun Yi and Shizhong Xu. Bayesian LASSO for quantitative trait loci mapping. *Genetics*, (2):1045–1055, 2008.

Kevin Y. Yip, Roger P. Alexander, Koon-Kiu Yan, and Mark Gerstein. Improved Reconstruction of In Silico Gene Regulatory Networks by Integrating Knockout and Perturbation Data. *PLoS ONE*, (1):e8121+, 2010.

Jing Yu, V. Anne Smith, Paul P. Wang, Alexander J. Hartemink, and Erich D. Jarvis. Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, (18):3594–3603, 2004.

Cunlu Zou and Jianfeng Feng. Granger causality vs. dynamic Bayesian network inference: a comparative study. *BMC Bioinformatics*, (1):122+, 2009.