

Statistical inference of the time-varying structure of gene-regulation networks.

Sophie Lèbre, Jennifer Becq, Frédéric Devaux, Michael Stumpf, Gaëlle Lelandais

▶ **To cite this version:**

Sophie Lèbre, Jennifer Becq, Frédéric Devaux, Michael Stumpf, Gaëlle Lelandais. Statistical inference of the time-varying structure of gene-regulation networks.. BMC Systems Biology, 2010, 4 (1), pp.130. 10.1186/1752-0509-4-130 . inserm-00663719

HAL Id: inserm-00663719

https://inserm.hal.science/inserm-00663719

Submitted on 27 Jan 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

BMC Systems Biology

METHODOLOGY ARTICLE Open Access

Statistical inference of the time-varying structure of gene-regulation networks

Sophie Lèbre[1,2], Jennifer Becq[3,4,5], Frédéric Devaux[6], Michael PH Stumpf[1,7*†], Gaëlle Lelandais[3,4,5*†]

Abstract

Background: Biological networks are highly dynamic in response to environmental and physiological cues. This variability is in contrast to conventional analyses of biological networks, which have overwhelmingly employed static graph models which stay constant over time to describe biological systems and their underlying molecular interactions.

Methods: To overcome these limitations, we propose here a new statistical modelling framework, the ARTIVA formalism (Auto Regressive TIme VArying models), and an associated inferential procedure that allows us to learn temporally varying gene-regulation networks from biological time-course expression data. ARTIVA simultaneously infers the topology of a regulatory network and how it changes over time. It allows us to recover the chronology of regulatory associations for individual genes involved in a specific biological process (development, stress response, etc.).

Results: We demonstrate that the ARTIVA approach generates detailed insights into the function and dynamics of complex biological systems and exploits efficiently time-course data in systems biology. In particular, two biological scenarios are analyzed: the developmental stages of *Drosophila melanogaster* and the response of *Saccharomyces cerevisiae* to benomyl poisoning.

Conclusions: ARTIVA does recover essential temporal dependencies in biological systems from transcriptional data, and provide a natural starting point to learn and investigate their dynamics in greater detail.

Background

Molecular interactions and regulatory networks underlie the development and functioning of biological systems [1-3]. These networks reliably and robustly coordinate the molecular and biochemical processes inside a cell, while remaining flexible in order to respond to physiological and environmental changes. The changing nature of regulatory and signalling interactions is beyond doubt, and a dynamical point of view is already deeply enshrined into cell and molecular biology. Illustrations of such time-varying biological systems can be provided for instance by the development of the fruitfly *Drosophila melanogaster* - which is segmented into different life stages: embryogenesis, larva, pupa and adult, or the adaptation of cellular organisms (the yeast *Saccharomyces cerevisiae* for instance) to growth defects and cellular damages induced by environmental stresses. Because they are extensively studied, considerable large-scale functional screening data exist for these examples. But while a growing number of studies report detailed and time-resolved analyses of regulatory and signalling processes [4,5], mapping these temporally changing networks systematically remains a major and increasingly pressing challenge.

From available data, *in-silico* methods can generate hypotheses about biochemical and molecular mechanisms [6,7] and guide further experimental and theoretical investigations into regulatory interactions underlying biological systems. Biological networks are usually described mathematically in a way where each gene is represented by a node and the interactions (or regulatory associations) between genes as edges. A range of approaches has been proposed, which learn or infer correlative and causal relationships among the genes from

* Correspondence: m.stumpf@imperial.ac.uk; gaelle.lelandais@univ-paris-diderot.fr
† Contributed equally
[1]Center for Bioinformatics, Imperial College London, London, UK
[3]Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB), INSERM U 665, Paris, F-75015, France
Full list of author information is available at the end of the article

BioMed Central

© 2010 Lèbre et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

high-throughput, in particular gene expression, data. However most of these approaches assume that the topology of the network, *i.e.* the sets of nodes and edges, stays constant over time. Inferring the temporal changes in biological networks is an important statistical challenge [8], but it does open up new perspectives for biological data analyses and will aid the generation of hypotheses about the dynamics of biological systems.

Serious attempts to reconstruct dynamic networks whose topology changes with time started in 2005 [9,10]. Yoshida *et al.* [9] employed a dynamic linear model with Markov switching for estimating time-dependent gene network structures from time-series gene expression data. Although promising this approach assumes that there is a fixed (user-specified) number of distinct networks or phases, and the switching between phases is modelled *via* a stochastic transition matrix that requires an estimation of many parameters. Talih and Hengarten [10] developed a Markov Chain Monte Carlo (MCMC) methodology to recover time-varying Gaussian graphical models in a financial context. Again the total number of distinct network topologies is assumed to be known *a priori* and the network evolution is restricted to changing at most a single edge at a time. More recently (2007-2008) methods in which the number of distinct regulatory phases is determined *a posteriori* have been proposed. Fujita *et al.* [11] developped a Dynamic Vector Autoregressive model to estimate time-varying gene regulatory networks. Notably, only the values of the network parameters change over time, meaning that the global topology of the network remains constant. Xuan and Murphy [12] introduced an iterative procedure based on a similar modelling *ansatz*, which switches between a convex optimization approach for determining a suitable candidate graph structure and a dynamic programming algorithm for calculating the segmentation of the time into distinct phases, *i.e.* the sets of successive time-points for which the graph structure remains unchanged. This time, the number of phases is explicitly determined, but it requires that the graph structure is decomposable. Finally, Robinson and Hartemink [13] used a MCMC sampler for the inference of non-stationary dynamical Bayesian networks, with the attractive feature that the network structure within a temporal phase depends on the structure of the contiguous phases.

The approaches cited so far produce global network topologies with global changes, meaning that all the genes of the network change their regulatory inputs simultaneously. In reality however, we would rather expect that each gene (or at most a subset of genes) has its own and characteristic regulatory pattern. To that end, Rao *et al.* [14] developed a method called *regime-SSM*, which is divided into two steps. The main idea is to first cluster the genes that share the same temporal phases before inferring, in a second step, the network topology describing the regulatory associations between genes within each cluster using an expectation-maximization (EM) algorithm. Ahmed and Xing [15] introduced in 2009 a machine learning algorithm (called TESLA) to infer time evolving networks (that are gene-specific), by solving a set of temporally smoothed l_1-regularized logistic regression problems *via* convex optimization techniques.

The challenge of inferring time-varying structures of gene regulation networks is only starting to be adressed and in this paper we present the ARTIVA algorithm (Auto Regressive TIme VArying models) that is particularly well-suited for addressing the issues raised above. Starting from time-course gene expression data, ARTIVA performs a gene-by-gene analysis and infers simultaneously (*i*) the topology of the regulatory network, and (*ii*) how it changes over time. In order to strike a balance between model refinement and the amount of information available to infer the model parameters, the ARTIVA model delimits temporal segments for each gene where the influence factors and their weights can be assumed homogeneous. For that we use a combination of efficient and robust methods: dynamical Bayesian networks (DBN) to model directed regulatory interactions between genes and Reversible Jump MCMC for inferring simultaneously the times when the network changes and the resulting network topologies. We evaluate the performance of ARTIVA on simulated data and illustrate our approach in the context of two different biological systems. We start by analyzing a commonly used dataset related to the developmental stages of *Drosophila melanogaster* and demonstrate the utility of our approach by a comparative analysis of the ARTIVA results with the TESLA results [15]. Next, we analyze the response of the yeast *Saccharomyces cerevisiae* to benomyl poisoning. This dataset represents an important challenge for the inference of time-varying networks since (*i*) the number of time-points is extremelly small (only 5 time points) and (*ii*) the expression values combine measurements obtained in wild-type and knock-out yeast strains. The biological relevance of the results obtained with ARTIVA are finally assessed using functional annotations and transcription factor binding information.

Methods

Graphical models

Bayesian Networks (BNs) have become a popular framework for representing regulatory networks [6,16] as they offer both a probabilistic interpretation of dependencies among expression of genes and a graphical representation that is more readily accessible than mathematical expressions (see Figure 1). For example if the expression

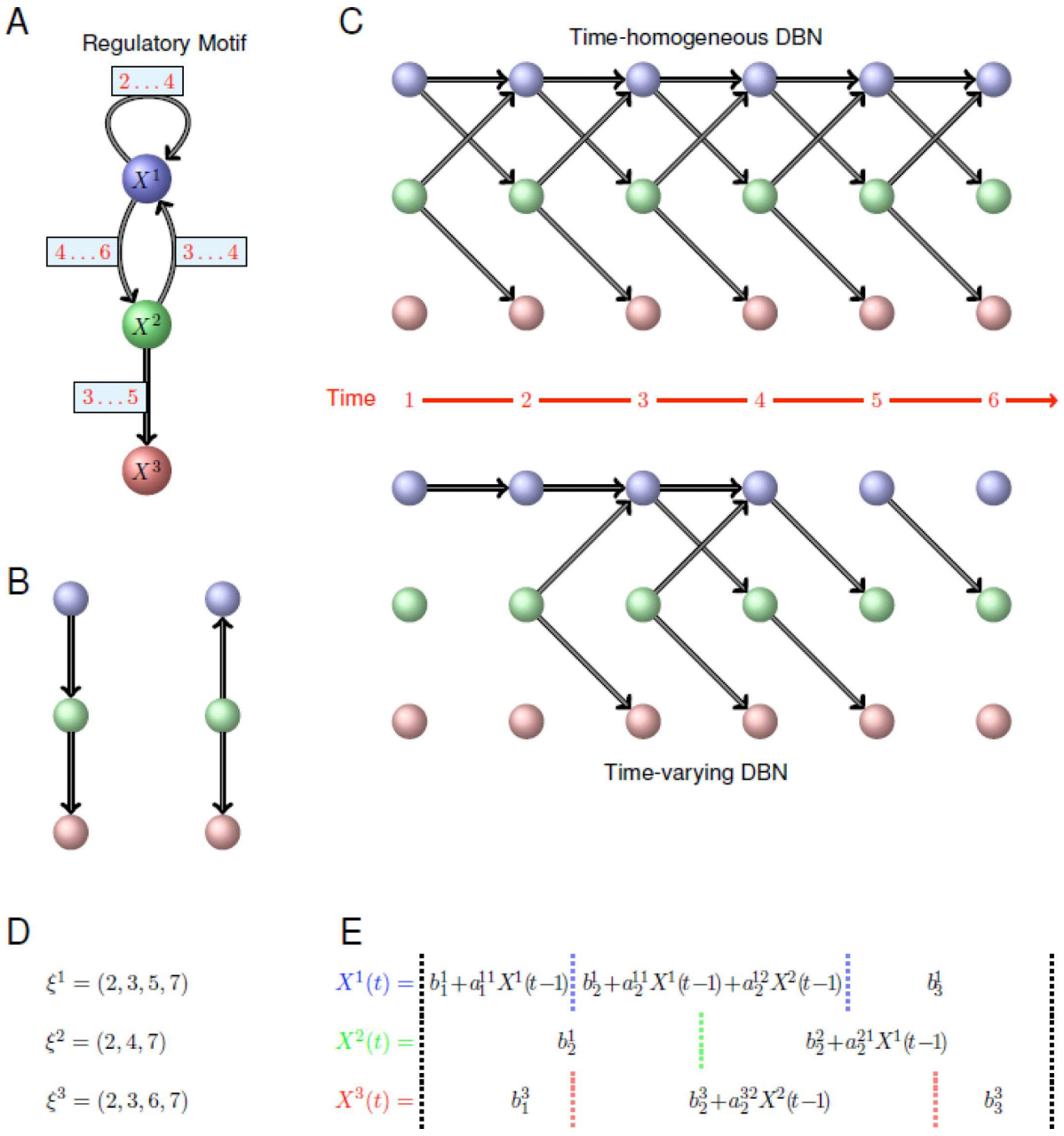


Figure 1 Illustration of the time-varying DBN formalism. (A) Regulatory motif among three genes that we wish to model. Crucially, regulatory interactions do not persist over the whole time course considered here, but are turned "on" and "off" at different times. The labels on the edges indicate at what times an edge points to or influences the expression of the target gene.(B) Because Bayesian networks (BNs) are constrained to have a directed acyclic graph (DAG) structure, they cannot contain loops or cycles. Therefore the motif in (A) can only be imperfectly represented using a conventional BN formalism which does not take temporal ordering into account; if X^3 is statistically independent of X^1 provided X^2 is known, we can construct two alternative representations, $P(X^1, X^2, X^3) = P(X^3|X^2).P(X^2|X^1).P(X^1)$ and $P(X^1, X^2, X^3) = P(X^3|X^2).P(X^1|X^2).P(X^2)$.(C) If time-course expression measurements are available we can unravel the feedback cycles and loops over time. Such dynamical Bayesian networks (DBN) represent the interactions by assuming that at each given time, all the parental nodes come from the previous time point. At the top of this panel we show the DBN constructed assuming a time-homogenous DBN; at the bottom of (C) we show the time-varying DBN constructed by the new algorithm. (D) Changepoint vectors for each of the three genes obtained for the time-varying DBN representation of the motif in (A). (E) The sets of regression models corresponding to the three nodes X^1, X^2 and X^3 in the inferred phases. Vertical dotted lines correspond to changepoints separating distinct phases for each node. Compulsory changepoints at the start and the end of the process (*i.e.* at $t = 2$ and $t = n + 1$) are indicated by the black dotted lines; inferred changepoints for each gene are shown in blue, green and red, corresponding to the colours of the genes (as used in parts (A), (B) and (C) of this figure).

level of gene i, here denoted by X^i, determines the expression level of genes j and k, a diagram such as

$$j \leftarrow i \rightarrow k$$

can be drawn and the joint distribution of gene expression levels written,

$$P\left(X^i, X^j, X^k\right) = P\left(X^k \mid X^i\right) P\left(X^j \mid X^i\right) P\left(X^i\right),$$

where $P(a|b)$ denotes the probability of a conditional on b. Because BNs aim to represent the joint probability distribution (in our case for the expression levels of p genes) the corresponding graphical representation is limited to graphs which contain no cycles (Figure 1B). This means that closed loops or complex feed-back structures (as in Figure 1A) cannot faithfully be represented, whereas they are known to pervade regulatory networks [17].

With time-course measurements, this limitation can be overcome by employing a Dynamical Bayesian Network (DBN) formalism [18], where the expression levels of all the genes in a system are modelled as a generally discrete-time stochastic process (Figure 1C). For p genes and n measurements the expression levels are written as $X^i(t)$, with $1 \le i \le p$ and $1 \le t \le n$. The joint probability distribution over the expression levels of all genes and at all times is then partitionned, $P(X^1(1), ..., X^p(1), ..., X^1(n), ..., X^p(n))$, into a product of conditional probabilities of the Markov form:

$$P(X^i(t) \mid X^r(t-1), \ldots, X^s(t-1)) \quad (1)$$

This means that the expression level of gene i at time t depends on the expression levels of genes r, ..., s at time t - 1. Genes r, ..., s are called the 'parents' of gene i and denoted by Pa^i (reciprocally gene i is called the 'target' gene of genes r, ..., s). By making the time dependence of expression levels explicit, loops and feedback interactions can be represented simply by requiring only that the expression of gene i at time t is independent of all other genes at the same time t. In conventional DBN inference approaches it is assumed that the conditional dependencies in Eqn. (1), and hence the set Pa^i, are independent of time t. Of course it is possible to allow $X^i(t)$ to depend on expression levels $X^r(t - \tau)$ with $\tau > 1$, *i.e.* allow for higher order dependencies. For computational reasons, however, our analysis is restricted to first order Markov processes.

ARTIVA network model

Let p be the number of observed genes and n the number of time-points at which expression levels are measured for each gene. In this study, the discrete-time stochastic process $X = \{X^i(t);\ 1 \le i \le p,\ 1 \le t \le n\}$ is considered, taking real values and describing the expression level of the p genes at n time-points. We start by modelling the gene expression levels at time t probabilistically by a vector-autoregressive process:

$$\forall t \ge 2, \quad X(t) = A(t)X(t-1) + B(t) + \epsilon(t) \quad \text{with} \quad \epsilon(t) \sim \mathcal{N}(0, \Sigma(t)), \quad (2)$$

where $\boldsymbol{X}(t) = (X^i(t))_{1 \le i \le p}$ and $\mathcal{N}(0, \boldsymbol{\Sigma}(t))$ is the multivariate normal distribution centered at 0 with diagonal covariance matrix $\boldsymbol{\Sigma}(t)$. Note that diagonality of $\boldsymbol{\Sigma}$ ensures that the process describing the temporal evolution of gene expression – here a first order autoregressive process – can be represented by a Directed Acyclic Graph (DAG) as in Figure 1C, *i.e.* no edges between nodes at the same time, and where the edges from time t - 1 to time t are defined by the set of non zero coefficients in matrix $\boldsymbol{A}(t)$ [19]. Furthermore the error in expression measurements of gene i does not affect the expression measurements of the other genes and off-diagonal elements in $\boldsymbol{\Sigma}$ can be set to 0.

Crucially, the coefficient matrix $\boldsymbol{A}(t) = (a^{ij}(t))_{1 \le i,\ j \le p}$ – which is the adjacency matrix of the gene regulatory network [19,20] – and the column vector $\boldsymbol{B}(t) = (b^i(t))_{1 \le i \le p}$ – which is the baseline gene expression that does not depend on the parent gene regulatory controls – are allowed to vary explicitly with time. This could for example reflect switching on or off of regulatory interactions, *e.g.* in response to developmental, physiological or environmental signals (Figures 1C, D and 1E).

For each gene, i, a set of time-points for which the regulatory inputs of the gene change is determined. These time-points are referred to as 'changepoints' and delimit homogeneous phases, *i.e.* sets of time-points for which the local network topology (edges between gene i and its parents Pa^i) remains unchanged. Assuming k changepoints, the changepoints are denoted by $\xi^i = (\xi^i_0, \ldots, \xi^i_{k+1})$, where $\xi^i_0 = 2$ (for a 1^{st} order Markov model) and $\xi^i_{k+1} = n + 1$ (to delimitate the bounds). The distinct phases are labelled by the index of their respective right changepoints. For all times t in the phase h of gene i (*i.e.* $\xi^i_{h-1} \le t < \xi^i_h$), the i-th row of matrix $\boldsymbol{A}(t)$ and coefficient $b^i(t)$ are assumed constant:

$$\forall \xi^i_{h-1} \le t < \xi^i_h, \quad b^i(t) = b^i_h, \quad \forall 1 \le j \le p,\ a^{ij}(t) = a^{ij}_h \quad (3)$$

For phase h of gene i, the parents Pa^i_h of gene i include every gene j such that the coefficient a^{ij}_h differs from 0: $\mathrm{Pa}^i_h = \{j; \forall 1 \le j \le p, a^{ij}_h \ne 0\}$. Hence for $\xi^i_{h-1} \le t < \xi^i_h$ the expression of gene i is modelled in a regression framework as:

$$X^i(t) = \sum_{j \in \mathrm{Pa}_h^i} a_h^{ij} X^j(t-1) + b_h^i + e^i(t), \quad \text{with } e^i(t) \sim \mathcal{N}(0, (\sigma_h^i)^2), \tag{4}$$

where $X^j(t - 1)$ is the expression level of gene j at time $t - 1$. This defines a multiple changepoint regulatory network, with changepoint positions $\xi = (\xi_0^i, \ldots, \xi_{k+1}^i)_{1 \le i \le p}$, and the phase-specific regression model parameters, $\{a_h^{ij}, b_h^i, \sigma_h^i\}$ for all h, i, j. All non-zero coefficients, a_h^{ij}, indicate relationships between expression levels of genes i and j, and hence are good indicators of putative biological interactions between those genes.

Model inference via reversible jump MCMC sampling

General principle

We want to infer the autoregressive time-varying network model, which belongs to the overall parameter space that is the union of the parameter spaces of all phases delimited by k changepoints ($k = \{0, \ldots, \overline{k}\}$). Furthermore, for each phase the number of incoming edges on each node (or the network topology) is unknown. Adding or removing a changepoint results in a change in the dimension of the system's state-space: for each additional changepoint a new network topology has to be estimated, and for each deleted changepoint the results previously obtained for the two distinct phases have to be reconciled. Thus, the dimension of the model is unknown and can vary substantially. In order to infer the posterior distribution $\Pr(k, \xi, s, \mathrm{Pa}, \theta, \sigma|x)$ given the observed data x over all of the system's parameters, we used a Reversible Jump Markov Chain Monte Carlo (RJ-MCMC) procedure. The principle of RJ-MCMC lies in constructing a reversible Markov chain sampler that can jump between parameter subspaces of different dimensions; thus allowing the generation of an ergodic Markov chain whose equilibrium distribution is the desired posterior distribution [21,22].

Presented in Figure 2, our inference procedure allows us to simultaneously consider all possible combinations of changepoints and network topologies within the different phases. In the RJ-MCMC procedure, the likelihood of the expression measurements $x(1)$ observed at time-point $t = 1$ is denoted by $\Pr(x(1))$. From the hierarchical structure of the overall parameter space, the joint probability distribution over all parameters can thus be written as the product:

$$\Pr(k, \xi, s, \mathrm{Pa}, \theta, \sigma, x) = \Pr(x(1)) \prod_{i=1}^{p} \{\Pr(k^i, \xi^i, s^i, \mathrm{Pa}^i, \theta^i, \sigma^i, x^i)\}. \tag{5}$$

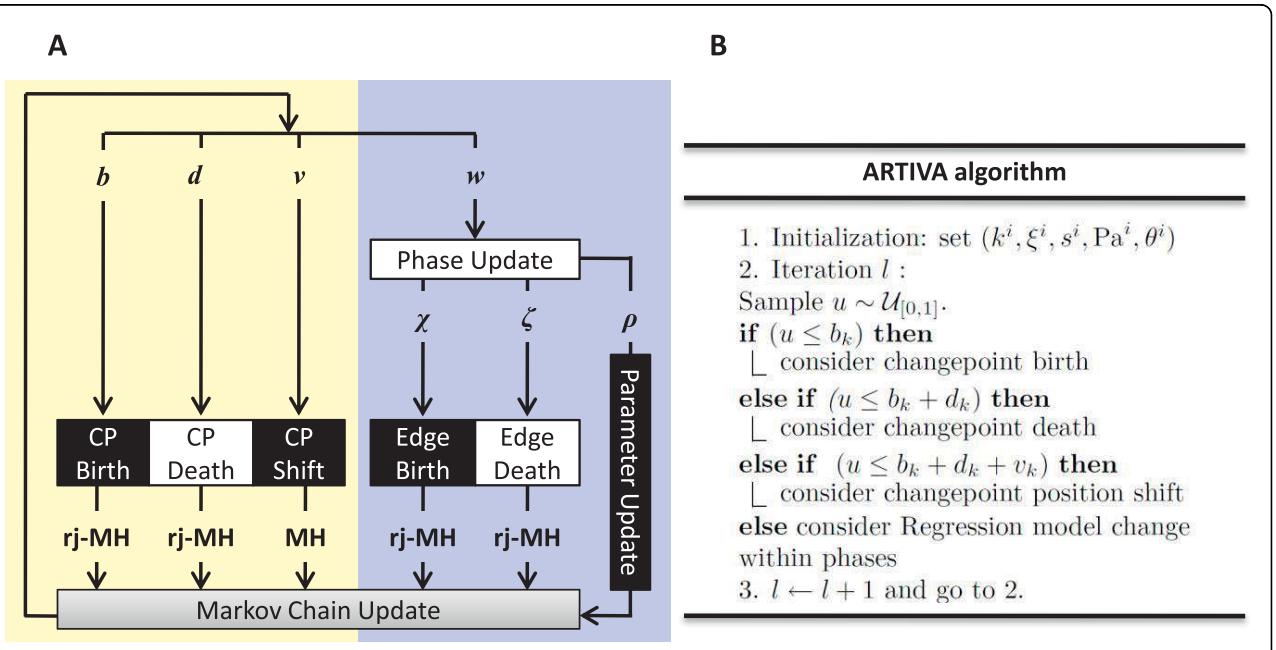


Figure 2 Illustration of the ARTIVA procedure. (A) Schematic illustration of the two-step RJ-MCMC scheme used for determining the stationary distribution of time varying dynamic Bayesian networks. With probabilities b, d and v, we propose the birth, death or shift of a changepoint (CP) respectively; with probability w we propose an update of the regression model describing regulatory interactions for a gene within a temporal phase. Varying the number of CPs or the number of edges (network topology) corresponds to a change in the dimension of the state-space and is dealt with by using Green's RJ-MCMC formalism [21]. Proposed shifts in changepoint positions are accepted according to a standard Metropolis-Hastings step. Because of conservation of probability we necessarily have $b + d + v + w = 1$ and $\chi + \zeta + \rho = 1$. (B) Outline of the ARTIVA inference procedure.

Posterior distribution

For each gene, i, we construct a RJ-MCMC sampler that directly samples from the joint distribution:

$$\Pr(k^i, \xi^i, s^i, \mathrm{Pa}^i, \theta^i, \sigma^i, x^i) = \Pr_{\bar{k}}(k^i)\Pr(\xi^i \mid k^i) \prod_{h=1}^{k^i+1} \Pr(s_h^i, \mathrm{Pa}_h^i, \theta_h^i, \sigma_h^i)\ \Pr(x_h^i \mid \xi_{h-1}^i, \xi_h^i, s_h^i, \mathrm{Pa}_h^i, \theta_h^i, \sigma_h^i), \quad (6)$$

Where $\Pr_{\bar{k}}(k^i)$ $\Pr(\xi^i|k^i)$ and are respectively the prior probabilities of the number of changepoints k^i and of the changepoint position vector ξ^i for gene i, and where $\Pr(s_h^i, \mathrm{Pa}_h^i, \theta_h^i, \sigma_h^i)$ is the prior probability of the parameters defining the regression model of the phase h of gene i. Finally:

$$\Pr(x_h^i \mid \xi_{h-1}^i, \xi_h^i, s_h^i, \mathrm{Pa}_h^i, \theta_h^i, \sigma_h^i) = \prod_{\xi_h^i \le t < \xi_{h+1}^i} \Pr(x^i(t) \mid \xi_{h-1}^i, \xi_h^i, s_h^i, \mathrm{Pa}_h^i, \theta_h^i, \sigma_h^i) \quad (7)$$

is the likelihood of the expression levels $x_h^i = (x^i(t))_{\xi_h^i \le t < \xi_{h+1}^i}$ of gene i observed during phase h, and is a realization of the Gaussian distribution defined in Equation(4).

Priors

In order to reinforce sparsity of the network and following multiple changepoint approaches involving RJ-MCMC [23,24], we assume the number of changepoints k^i to be distributed *a priori* as a truncated Poisson random variable with mean λ and maximum $\bar{k}$,

$$\forall k^i \le \bar{k}, \quad \Pr_{\bar{k}}(k^i) \propto e^{-\lambda} \frac{\lambda^{k^i}}{k^i!} 1_{\{k^i \le \bar{k}\}}. \quad (8)$$

Similarly, the prior probability for the number of parents is a truncated Poisson distribution $\Pr_{\bar{s}}(s_h^i)$ with mean Λ and maximum $\bar{s}$. Here λ and Λ can be interpreted as the expected number of changepoints and parent variables, respectively. Following [25], λ and Λ are drawn according to a Gamma distribution: $\lambda, \Lambda \sim \mathcal{G}a(\alpha, \beta)$ where the shape parameter α and the scale parameter β are chosen so that the prior probability decreases when the numbers of changepoints or parents increase (we set $\alpha = 1$, $\beta = 0.5$, see Additional file 1 for an illustration of the corresponding distribution). Conditional on there being k^i changepoints, we assume that the changepoint positions vector ξ^i takes only non-overlapping and uniformly distributed integer values. The prior for the regression model parameters (s^i, Pa^i, θ^i, σ^i) are chosen following Andrieu and Doucet' RJ-MCMC procedure for regression model selection [25], based on a work proposed in [26]. The sets of parents $\mathrm{Pa}^h(i)$ are assumed to be uniformly distributed conditional on $|\mathrm{Pa}^h(i)| = s_h^i$. The variance, σ_h^i, is assumed to be distributed according to a conjugate inverse-Gamma prior distribution with shape parameter $v_0/2$ and scale parameter $\gamma_0/2$, $(\sigma_h^i)^2 \sim \mathcal{IG}(v_0/2, \gamma_0/2)$. By choosing $v_0 = 1$ and $\gamma_0 = 0.1$, we set up to Jeffrey's vague prior, $\Pr((\sigma_h^i)^2) \propto 1/(\sigma_h^i)^2$ [25]. Finally, conditional on σ_h^i, the prior distribution for the regression model parameters can be written as,

$$\Pr(s_h^i, \mathrm{Pa}_h^i, \theta_h^i \mid \sigma_h^i) = \Pr_{\bar{s}}(s_h^i)\Pr(\mathrm{Pa}_h^i \mid s_h^i)\Pr(\theta_h^i \mid \sigma_h^i, s_h^i, \mathrm{Pa}_h^i). \quad (9)$$

Given the parent gene set Pa_h^i of size s_h^i, the $s_h^i + 1$ regression coefficients, $\theta_{\mathrm{Pa}_h^i} = (b_h^i, (a_h^{ij})_{j \in \mathrm{Pa}_h^i})$, are assumed to be drawn from zero-mean Gaussian distributions with covariance $(\sigma_h^i)^2 \Sigma_{\mathrm{Pa}_h^i}$,

$$\Pr(\theta_h^i \mid \sigma_h^i, s_h^i, \mathrm{Pa}_h^i) = |2\pi(\sigma_h^i)^2 \Sigma_{\mathrm{Pa}_h^i}|^{-1/2} \exp\left[-\frac{\theta_{\mathrm{Pa}_h^i} \Sigma_{\mathrm{Pa}_h^i}^{-1} \theta_{\mathrm{Pa}_h^i}}{2(\sigma_h^i)^2} \right], \quad (10)$$

where the symbol † denotes matrix transposition, $\sum_{\mathrm{Pa}_h^i} = \delta^{-2} D_{\mathrm{Pa}_h^i}^{\dagger}(x) D_{\mathrm{Pa}_h^i}(x)$ and $D_{\mathrm{Pa}_h^i}(x)$ is a matrix of size $m^i(\xi_h^i - \xi_{h-1}^i) \times (s_h^i + 1)$, whose first column is a vector of 1 when the regression model includes a constant, and each $j + 1^{th}$ column contains the observed (eventually repeated) value $(x_{tl}^j)_{(\xi_{h-1}^i - 1) \le t < (\xi_h^i - 1); 1 \le l \le m}$ for all parent gene j in Pa_h^i. We did not use shrinkage priors here because the truncated Poisson prior for the number s_h^i of parents already favours dimension reduction. The term δ^2 represents the expected signal-to-noise ratio and is sampled according to an Inverse Gamma distribution $\delta^2 \sim \mathcal{IG}(\alpha_{\delta^2}, \beta_{\delta^2})$ with $\alpha_{\delta^2} = 2$ and $\beta_{\delta^2} = 0.2$.

A noticeable advantage of the model is that the marginalization over the regression parameters (θ^i, σ^i) in the posterior distribution is analytically tractable,

$$\Pr(k^i, \xi^i, s^i, \mathrm{Pa}^i, x^i) = \int\int \Pr(k^i, \xi^i, s^i, \mathrm{Pa}^i, \theta^i, \sigma^i \mid x^i) d\theta^i d\sigma^i \quad (11)$$

(see Additional file 1, Section 2 for more details). Then the proposals are sampled from the analytical expression of the network topology posterior distribution (11) – which is proportional to $\Pr(k^i, \xi^i, s^i, \mathrm{Pa}^i|x)$ – and the acceptance probability depends on the network topology (ξ^i, Pa^i) only.

Moves

In order to traverse the parameter space of unknown dimension we propose here four different update moves (see Figure 2 and Additional file 1): birth of a new changepoint (*B*); death (removal) of an existing changepoint (*D*); shift of a changepoint to a different time-point (*S*); and update of the regression model defining the network topology within the phases (*R*). These moves occur with probabilities b_k for *B*, d_k for *D*, ν_k for *S* and w_k for *R*, depending only on the current number of changepoints k^i and satisfying $b_k + d_k + \nu_k + w_k = 1$. The changepoint birth and death moves represent changes from, respectively, k^i to $k^i + 1$ phases and k^i to k^i - 1 phases. We impose $d_0 = \nu_0 = 0$ and $b_{\bar{k}} = 0$ to preserve the restriction on the number of changepoints. Otherwise, these probabilities are chosen as follows:

$$b_k = c\min\left\{1, \frac{\Pr_{\bar{k}}(k^i+1)}{\Pr_{\bar{k}}(k^i)}\right\}, \quad d_{k+1} = c\min\left\{1, \frac{\Pr_{\bar{k}}(k^i)}{\Pr_{\bar{k}}(k^i+1)}\right\} \qquad (12)$$

where $\Pr_{\bar{k}}$ is the prior distribution for the number of changepoints and the constant *c* is chosen to be smaller than 14 so that the regression model updates and changepoint position shifts are proposed more frequently than births and deaths of changepoints. This improves our ability to infer changepoint positions and the network structures (using the vector-autoregressive framework) within the different phases. Proposed shifts in changepoint positions are accepted using a standard Metropolis-Hastings step, while regression model updates within phases invoke a second RJ-MCMC criterion, which was adapted from the model selection approach of Andrieu and Doucet [25]. As proposals are sampled from the analytical network topology posterior distribution (11), the generation of the regression model parameters (θ_h^i, σ_h^i) is optional. Together the four moves *B*, *D*, *S* and *R* allow the generation of samples from probability distributions defined on unions of spaces of different dimensions for both the number, k^i, of changepoints and the number s_h^i of parents within each phase for gene *i*.

Model selection

Given *a priori* probabilities, the ARTIVA algorithm produces posterior probability estimations over the algorithm iterations for changepoint vectors and network topologies. These posterior probabilities give a detailed picture of all the results and allow in depth analyses of the entire regulatory network architecture. In this study we use in complement to posterior probabilities, the Bayes factor, *i.e.* the ratio of the posterior odds of an hypothesis over its prior odds [27]. The Bayes factor has the advantage to consider the posterior distribution with respect to the priors and to obtain quantitative measurements of the statistical significance of the ARTIVA results which are comparable between different datasets. As an indication, according to Kass and Raftery [27], a proposition is (*i*) *not* supported when it has a Bayes factor below 3, (*ii*) *positively* supported for a Bayes factor between 3 and 20 and (*iii*) *strongly* supported for a Bayes factor over 20. The performance of ARTIVA is evaluated on synthetic and real data (see the following section) by selecting the network structure according to the following procedure. For each gene *i* we first choose the number k^i of changepoints having the greatest Bayes factor. Then the k^i changepoint positions having the highest Bayes factors are selected, and for each resulting phase we finally compute the Bayes factor for the possible parent genes and choose the ones with a Bayes factor greater than 3 (see Additional file 1 for a description of the Bayes factor computation).

Simulation study

In order to evaluate the accuracy of the ARTIVA algorithm to recover changepoints and network topologies correctly, expression data for randomly defined dynamic networks were generated. With respect to the experimental datasets analyzed later (see the following section), two types of expression data were produced. The first type – referred to Wild-Type (WT) simulations – match the 'Drosophila life cycle' data. This dataset contains time-series expression data of several genes and the algorithm must find the correlations between unknown parent genes and each target gene. The second type – referred to as Knock-Out (KO) simulations – is equivalent to the 'benomyl' dataset. This dataset only contains time-series expression data of target genes in different genetic contexts: wild-type and knock-out mutants for several transcription factors (TFs).

The simulation procedure for a given target gene, also presented in detail in Additional file 2, involves three main steps. First, the structure of the dynamic regulatory network is defined. This consists of randomly setting the number and the localization of changepoints, thereby defining regulatory phases. Then the parent genes and the corresponding coefficients are chosen for each phase. Once the regulatory network is defined, expression data can be generated from this network model. The expression values of the parent genes are first generated randomly (uniformly drawn from [-2,

-0.1] ∪ [0.1, 2]) and subsequently used to calculate the target gene expression according to the autoregressive model presented in Eqn. (4). The whole procedure is repeated (to represent experimental replicates) and noise is added (to represent experimental variability). Because the simulations use the ARTIVA hypothesis concerning the expression associations between parent and target gene expression profiles (autoregressive model), we expect all the results to be correct under ideal conditions (like absence of noise). Therefore, this simulation protocol evaluates the ARTIVA performance and studies the influence of the following parameters:

- the quantity of noise in the data. For all phases h of gene i, the noise $e^i(t)$ is drawn from a Gaussian distribution $\mathcal{N}(0, (\sigma_h^i)^2)$ with standard deviation $\sigma_h^i (\sigma_h^i = 0.2, 0.4, \ldots, 1.8)$,
- the size of the temporal phases (*phasesize* = 1, 2, ..., 5, 12), and
- for WT simulations only, the number of potential parent genes (*Pa#* = 5, 10, 20, 40). This is not necessary for KO simulations because the potential parent genes are obviously restricted to the transcription factors for which KO data is being generated. Regardless of the number of potential parent genes, a maximum of 5 edges from parent genes to a target gene is allowed.

For each parameter value, 200 gene time-series of length = 12 timepoints were randomly generated with 8 and 4 replicates for WT and KO data respectively. Note that KO simulations present less replicates because each replicate already comprises the measurements of the gene expression levels for each knock-out mutant. All other parameters were set to their default values and are

Table 1 Performance of ARTIVA on simulated data

		WT simulations				KO Simulations			
Parameter		**Changepoint sensitivity**	**Changepoint PPV**	**Edges sensitivity**	**Edges PPV**	**Changepoint sensitivity**	**Changepoint PPV**	**Edges sensitivity**	**Edges PPV**
	0.2	94.2	95.1	73.9	99.2	100	100	100	98.5
	0.4	90.8	94.1	79.3	97.9	100	99	100	98.2
	0.6	87.8	92.5	73.9	96.9	96	97.1	99.2	95.4
	0.8	75.4	96.3	78.8	96.4	81.4	97.8	97.6	91.4
Noise	**1**	80.5	96.6	74.7	97.5	69.1	95	95.1	88.8
	1.2	71.7	96.2	78.4	97.6	28.1	82.4	97.2	87.6
	1.4	58.7	94.6	79	95.9	23.6	89.7	92.7	87.5
	1.6	52.9	91.8	74.4	97.2	10.8	75.9	79.1	86.8
	1.8	60.8	94.5	76	95.6	4.8	81.8	76.8	85.6
	1	78.5	97.5	1.3	100	79	98.8	18.1	82.4
	2	92	92	24.7	98.5	97	96.5	98.7	92.5
Phase size	**3**	90	94.2	50.8	98.6	99.5	99	100	93.6
	4	94.5	94	74.5	98.6	99.5	99.5	100	96.4
	5	96	99	76.9	96.8	99.5	97.5	100	94.7
	12	100	99	92.6	98	99	97.1	100	97.8
	5	93.7	95.5	82.4	99.1	_	_	_	_
# of parent	**10**	81.1	88.3	69.7	96.4	_	_	_	_
genes	**20**	62.4	83.4	51.2	97.3	_	_	_	_
	40	54	77.2	33.1	96.4	_	_	_	_

To evaluate ARTIVA performances, two types of data were used: WT simulations corresponding to time-series expression data with no knowledge of potential transcription factors and KO simulations corresponding to several time-series expression data of a simulated *wild-type* strain and *knock-out* strains for each known transcription factor (see Methods and Additional file 2 for more details). The default values of the parameters used for the simulation study are: # of timepoints n = 12; # of changepoints $k \sim \mathcal{U}(\{1, .., n/4\})$; maximal # of edges = 5; parent to target coefficient $\sim \mathcal{U}([-2, -0.1] \cup [0.1, 2])$; phase sizes $\sim \mathcal{U}(\{3, .., n\})$; # # of replicates r = 8 for WT simulations, r = 4 for KO simulations; noise standard deviation $\sim \mathcal{N}(0, 0.5)$; total # of simulations = 200 for each condition. In this table, the value of noise intensity, phase size and number of parent genes change according to the parameter under study (all other parameters were set to default). In each condition, the ability of ARTIVA to detect *all* true phase changepoints and model edges (Sensitivity) and to detect *only* true positives (Positive Predictive Value, PPV) was calculated. Overall, this simulation study allows us to gain confidence in ARTIVA results (≈ 80% of PPV and sensitivity) for a given set of parameters (noise standard error is on the order of the mean value of the regression coefficients, number of measurements in a regulatory phase > 8 and less than 20 parent genes).

specified in Table 1. The ARTIVA algorithm was run on each expression data set and we compared the proposed network model (selected as described in the previous subsection 'Model selection') with the original one. The ability of the algorithm to recover changepoints was evaluated via the Positive Predictive Value (PPV) and the Sensitivity,

$$PPV = \frac{TP}{(TP+FP)} \quad Sensitivity = \frac{TP}{(TP+FN)} \tag{13}$$

with TP = True Positives, FP = False Positives and FN = False Negatives. The edges PPV and Sensitivity was computed for the phases whose changepoints were correctly inferred.

Microarray data

The first microarray dataset – referred to as '*Drosophila* life cycle' data – was produced by Arbeitman *et al*. [28]. It includes the mRNA expression levels of 4028 genes at 67 successive time-points spanning the four stages of the *D. melanogaster* life cycle: the embryonic (31 time-points), larval (10 time-points) and pupal stage (18 time-points) and the first 30 days of adulthood (8 time-points). Expression data were collected from the Gene Expression Omnibus database: http://www.ncbi.nlm.nih.gov/geo/. 4005 genes with consistent annotation are used for the analysis. Potential parent genes were restricted to genes with known transcriptional activity based on Gene Ontology information [29]. Hence, 136 genes were selected as potential parents. They belong to one of the four following Gene Ontology molecular functions: 'Transcription activator activity' (GO:0016563), 'Transcription repressor activity' (GO:0016564), 'Transcription factor activity' (GO:0003700) and 'Transcription cofactor activity' (GO:0003712). For each target gene, we gave priority to the 10 potential parent genes with the most highly correlated gene expression profiles over any successive 10 time-points.

The second microarray dataset – referred to as 'benomyl' data – was published by Lucau-Danila *et al*. [30]. In this study, the authors measured the changes in mRNA concentrations for each gene at successive times after addition of benomyl (an antimitotic drug) in the growth media of *Saccharomyces cerevisiae* cells. Parallel experiments were conducted in different genetic contexts: the wild type strain and knock-out (KO) strains in which the genes coding for different transcription factors connected to drug response, *YAP1*, *PDR1*, *PDR3*, and *YRR1*, were deleted. For each yeast strain, the measured expression values for 5 time-points (at 30 s, 2 min, 4 min, 10 min, 20 min) were obtained from the website: http://www.biologie.ens.fr/lgmgml/publication/benomyl. We only considered genes that (*i*) showed significant changes in mRNA levels during the time-course analysis in the WT strain (119 genes presented by Lucau-Danila *et al*. [30]), and (*ii*) had less than 20% of missing expression measurement data in the four KO strains. The resulting expression table comprised data for 78 genes (see Additional file 3 for complete list of genes). Hierarchical clustering was performed applying the 'hclust' function available in the R programming langage http://cran.r-project.org/, using Euclidean distance between gene expression profiles and the 'ward' method for gene agglomeration (see also Additional file 4).

Technical information

The ARTIVA algorithm is implemented in R programming language. The source code is freely distributed to academic users upon request to the authors. A 50,000 iterations procedure lasts around 5 min times the number of genes for the analysis of 100 time-course measurements (for example 5 replicates over 20 time-points) with a 2.66 GHz Intel(R) Xeon(R) CPU and 4 G RAM.

Results

Evaluation of the algorithm performances

To evaluate the performance of our ARTIVA approach, simulations are run in order to assess the impact of three major factors on the algorithm performances: noise in the data, minimal length of phases, and number of proposed parent genes (the latter for WT simulations only, see Methods). Sensitivity and Positive Predictive Value (PPV) calculated for the detection of changepoints and of models, *i.e.* the topology of the network within the phases, are presented in Table 1. In WT simulations, the changepoint sensitivity is greater than 80% when the noise standard error reaches σ_i = 1. As noise increases further, the ability of the algorithm to recover changepoints decreases in terms of sensitivity, but still, the changepoint sensitivity remains greater than 70% when the noise standard deviation reaches σ_i = 1.2 (a value that is larger than the mean value of the regression coefficients, uniformly sampled from [-2; -0.1] ∪ [0.1; 2]). The WT data was generated with r = 8 repeated measurements for each time point, whereas the KO data were simulated with only 4 repeated measurements for each time point (because each measurement includes data from different genetic contexts, see Methods). That is the reason why the changepoint sensitivity with KO simulations starts to decrease with smaller noise standard deviation compared to WT simulations. Nevertheless, the changepoint sensitivity is still greater than 80% even when noise reaches σ_i = 0.8. The number of measurements for each phase also plays an important role for the changepoint detection sensitivity. Indeed, during a phase reduced to a single timepoint, there are only

r repeated measurements to estimate the autoregressive models. Interestingly, the ARTIVA algorithm here succeeds in finding the correct dynamic networks with a sensitivity value of 79% for a phase size of 1, in both WT and KO simulations (default noise standard error $\sigma_i = 0.5$). With phases of size 2, the changepoint sensitivity is greater than 90%. For all noise levels considered here the changepoint PPV is greater than 95%; furthermore changepoint PPV appears to be stable and not to be affected by the phase size either. Knock-out data are usually collected for a restricted number of knock-out genes and the number of possible parents is limited. However, wild-type experiments give expression time series data for a large number of genes at once. The number of proposed parents increases the dimension of the model and the estimation procedure accuracy is expected to be affected as the dimension increases. Here, the changepoint sensitivity obtained with ARTIVA is still 54% when the parent genes are chosen from among a set of 40 proposed parents. The changepoint sensitivity goes up to 81% when the number of potential parents is reduced to 10. The changepoint PPV is only slightly affected by the number of proposed parents. The PPV is still greater than 75% when the number of potential parents is 40.

The edge detection in Table 1 was evaluated when the correct changepoint segmentation was recovered. Once the correct changepoints are recovered, neither noise nor short phases appear to strongly affect the detection of edges. The edge sensitivity deteriorates for extreme situations only. Indeed, the edge sensitivity is equal to 18% when phase size is 1 for KO simulations. For WT simulations, the edge sensitivity is about 50% when phase size is 3 or when the number of proposed parents is 20. In all other cases, the edge sensitivity is greater than 75% and the edge PPV is greater than 95%.

Simulation studies such as the one performed here do, of course, only provide a partial insights into an algorithms performance and robustness. They are nevertheless essential to gain confidence in the performance of novel algorithms and to develop understanding of their likely limitations. Together these results serve to illustrate of the robustness of the ARTIVA algorithm. In particular, ARTIVA can deal with some of generic problems encountered in real experimental data. It still performs well when noise standard error is on the order of the mean value of the regression coefficients, when the number of measurement per phase is reduced to 8 or when the number of possible parents reaches 20. At some point, the ARTIVA algorithm misses some changepoints, but the PPV is still very large, meaning that we can have great confidence in the changepoints having a high posterior probability.

Temporal variation of the Drosophila development transcriptional program

In light of the simulation analysis, we then apply our method to the well-studied expression datasets produced by Arbeitman *et al*. [28]. In this study, the authors report gene expression patterns for nearly one-third of all *D. melanogaster* genes during a complete time-course of development. The ARTIVA algorithm is run for each gene for 50,000 iterations, looking for parental relationships with the 10 transcription factors for which gene expression profiles were most highly correlated over any successive 10 time-points (see Methods). Out of the 4005 analyzed genes, 1704 (42%) were found to be involved in the time-varying networks spanning the whole *Drosophila* life-cycle (134 were identified as parent genes, 1623 as target genes and 53 were both parent and target genes). Interestingly, 2583 changepoints were also identified. The distribution over the time-points and with respect to the developmental stages is shown Figure 3. We observe that time intervals {18 to 19}, {31 to 33}, {41 to 43} and {59 to 61} contain more than 40% of the changepoints. Notably the intervals {31 to 33}, {41 to 43} and {59 to 61} include the developmental stage transitions from embryo to larva, from larva to pupa and from pupa to adult, respectively. The high number of changepoints at mid-embryogenesis (interval {18 to 19}) corresponds to a major morphological change related to a modification of transcriptional regulations, as described in [28].

To further evaluate ARTIVA, we compared our results with those obtained using the TESLA algorithm [15]. TESLA has been recently published (2009) and to our knowledge it is with ARTIVA, the only other procedure which recovers time varying regulatory networks where the changepoints are gene specific. As described in [15], we first discretized the expression measurements into two levels: 1 for up-regulation and 0 for down-regulation. The TESLA procedure requires specification of two parameters, λ_1, which is a sparsity coefficient, and λ_2, which is a smoothness penalty coefficient. Several combinations of (λ_1, λ_2) parameters were tested (data not shown), and we finally retained the average values presented by the authors in their simulation study [15], *i.e.* $\lambda_1 = 0.01$, $\lambda_2 = 1$. The TESLA analysis was run using the same subset of *Drosophila* genes used with ARTIVA, and the 2583 most significant temporal changes identified with TESLA are compared to the 2583 ARTIVA changepoints (Figure 3, dashed line). In agreement with the ARTIVA results, an important number of regulatory changes (28%) occurred during the developmental stage transitions (mid-embryogenesis, embryo to larva, larva to pupa and pupa to adult), but notably this number is significantly lower than the one obtained with ARTIVA (40%, see previous paragraph). This is especially

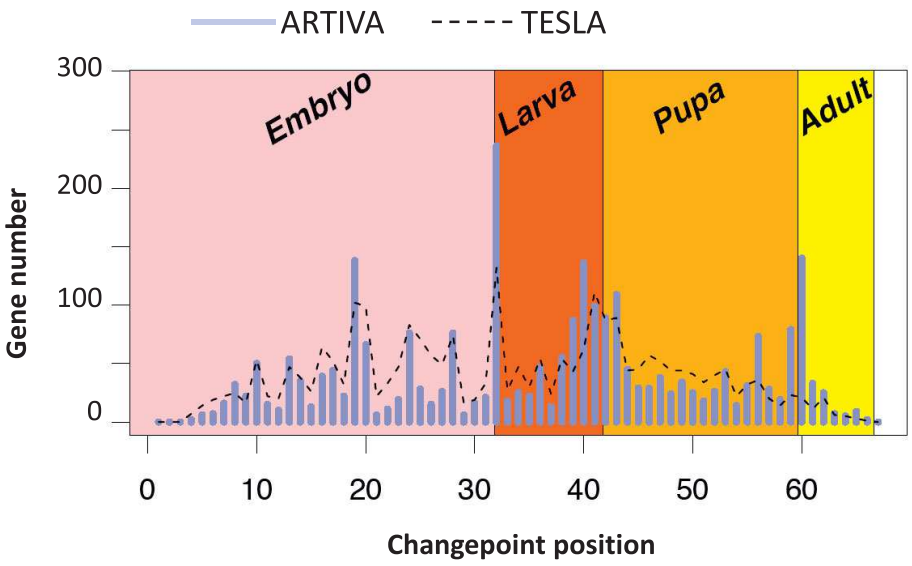


Figure 3 Changepoints of gene regulation networks across the Drosophila melanogaster development. Microarray results for the time courses of *Drosophila* life cycle [28] were analysed using ARTIVA. The number of identified changepoints using ARTIVA are shown in blue for each of the 67 time-points. They are compared with the most significant changes identified with the TESLA algorithm [15], shown in black dashed line. Time-intervals for each developmental stage are represented with the following color-code: pink = Embryo (31 time-points), red = Larva (10 time-points), orange = Pupa (17 time-points), yellow = Adult (8 time-points).

remarkable considering the last phase transition from pupa to adult. The observation of a significant number of changepoints at developmental stage transitions lends credibility and supports our ARTIVA results. Our method appears powerful in inferring the timepoints at which transcriptional control of individual genes switches.

Time-varying regulatory network involved in the response of Saccharomyces cerevisiae to benomyl poisoning

In our second example, we apply ARTIVA to a selected set of 78 gene expression profiles from *Saccharomyces cerevisiae* cells grown under benomyl-induced stress conditions [30] (see Methods). A hierarchical cluster analysis identifies 18 clusters of genes with concordant transcription profiles (see Additional file 5). For each cluster, time varying networks are inferred using the included gene expressions measured in the wild type and four deletion strains (Yap1, Pdr1, Pdr3 and Yrr1), running the RJ-MCMC scheme for 50,000 iterations. Regulatory associations between parent and target genes are proposed if the deletion of a parent gene significantly alters the expression measurements of its target genes (compared to the WT situation) (see Methods). The results are presented in Figure 4 and Dataset S1. As an illustration, the cluster #1 comprises 10 genes (Figure 4A) for which two changepoints are detected at the 4 and 10 minute time-points (Figure 4B), when these genes fall under the regulatory control of Yap1 (Figure 4C). Even if Yap1 is the only transcription factor identified here, its regulatory interactions with the target genes in the third phase are highly significant (Bayes factor = 9.10^3) compared to those in the second phase (Bayes factor = 14.22). This explains the detection of two changepoints. The results obtained for all other clusters are combined to obtain a global view of the time-varying regulatory network involved in benomyl stress response (Figure 4D). In agreement with the pioneering study of Lucau-Danila *et al*. [30], the transcription factor Yap1 appears to have the predominant role in the benomyl stress response as ARTIVA identified edges with 79% of the analyzed genes (62 associations with clusters # {1, 2, 5, 6, 7, 8, 9, 13, 18, 17}). Also PDR1, being the parent gene of 24% of the genes, exerts significant control (19 associations with clusters # {5, 6}). Pdr3 and Yrr1 present only a small number of target genes (10 associations with cluster #6 and 2 associations with cluster #13, respectively).

Furthermore, our ARTIVA model provides a dynamic classification of the benomyl responsive-genes, based on their time of induction. Such a dynamical point of view can elucidate the chronology of events, especially regarding the Yap1 activity. ARTIVA identified three classes of Yap1 targets, depending on their time of induction: 4 minutes (clusters # {1, 7, 18}, orange arrows Figure 4D); 10 minutes (clusters # {2, 8, 9, 13, 17}, yellow arrows Figure 4D); and 20 minutes (cluster # {5, 6},

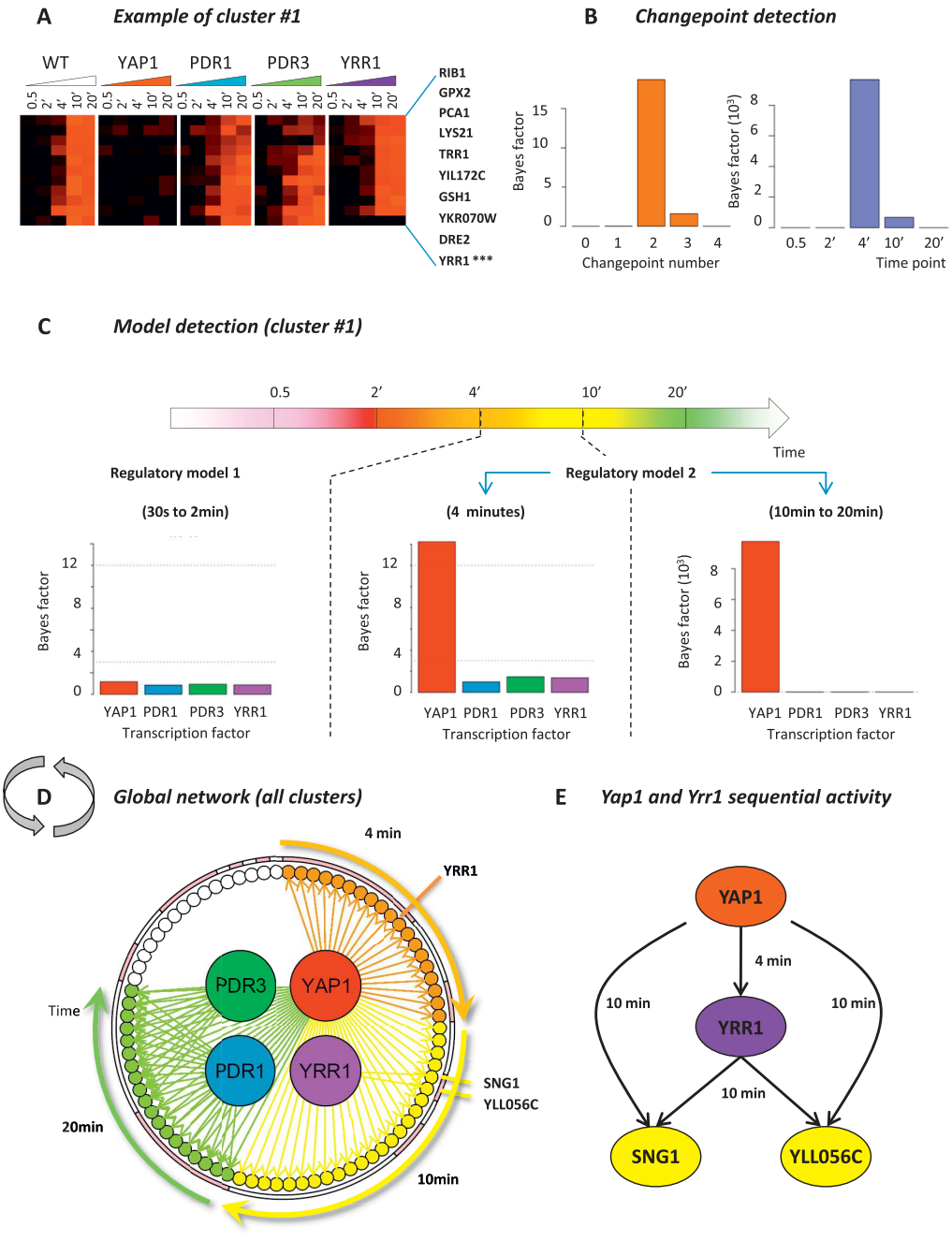


Figure 4 Time-dependent regulatory network involved in yeast chemical stress response. Microarray results for the kinetics of benomyl action [30] were analyzed using ARTIVA. A hierarchical clustering analysis was carried out for the time-course responses of the wild-type strain and deletion strains for four transcription factors (TFs): Yap1, Pdr1, Pdr3 and Yrr1. The resulting 18 clusters have low intra-cluster variability and comprise genes whose expression is identically modified in TFs deletion strains compared to the wild-type strain (see Methods). Results for cluster #1 are presented here. (A) Gene expression measurements represented using the common color code (black for expression values around 0 and red for positive values). Bayes factors for changepoint (CP) and edge detection are respectively shown in (B) and (C). Two CP were identified at the 4 min and 10 min time-points, and regulatory associations with the TF Yap1 were identified in the second temporal phase (from 4 min to 20 min). (D) All the identified regulatory associations are shown here, after analyzing the 18 clusters of co-expressed genes independently. They are all positive, meaning that each transcription factor activates the expression of their respectives target genes. Regulatory interactions are color-coded according to their starting time-point: orange = 4 minutes, yellow = 10 minutes and green = 20 minutes. We found 62 regulatory interactions for Yap1; 19 for Pdr1; 10 for Pdr3; and 2 for Yrr1. Pink and white segments on the surrounding circle indicate genes belonging to the same gene expression cluster, the clusters are ordered as follows (starting at 4 min): 1, 18, 7, 9, 13, 8, 17, 2, 5, 6, 3, 11, 4, 10, 12, 14, 15, 16.

green arrows Figure 4D). Almost all the genes included in the earliest group are known to be transcriptionally controlled by Yap1 (95% based on YEASTRACT information [31]). They encode proteins involved in redox control (*GPX2*, *TRR1*, *GSH1*, *GTT2*) or vacuolar transporters (*YCF1*). The middle group contains also an important rate of Yap1 targets (87%), which act at the level of the plasma membrane (*FLR1* and *FRM2*) or encode proteins involved in response to toxins (for instance *AAD6*, *AAD16*, *ECM4*). Yap1 activity in the last group is partially overlapping with the actions of Pdr1 and Pdr3. Most of the genes in this group have unknown functions, but some of them are still labelled in the YEASTRACT database as being targets for Yap1 (74%), Pdr1 (32%) and Pdr3 (20%). Finally, *YRR1* deserves a special mention. Unlike the genes that encode the transcription factors Yap1, Pdr1 and Pdr3, the *YRR1* gene is transcriptionnally activated during the benomyl response. As a consequence, ARTIVA identified *YRR1* (*i*) as a Yap1 target whose expression was induced 4 minutes after benomyl addition in the cell growth culture (see *** Figure 4A); and (*ii*) as a parent for genes *SNG1* and *YLL056C* at 10 minutes. Interestingly these observations highlight a sequential activity of Yap1 and Yrr1 transcription factors together with an overlap of their targets (Figure 4E). This regulatory model, in which Yrr1 seconds Yap1, is fully supported by recent experimental data [32].

Discussion

ARTIVA: a new statistical modelling framework to learn temporally varying gene-regulation networks

The ARTIVA approach allows us to reverse engineer the temporally varying structure of transcriptional networks by inferring simultaneously the times at which regulatory inputs of genes change and the nature of these incoming inputs. Our approach is computationally efficient and can exploit powerful search heuristics to scan the space of potential incoming edges. Compared to others methodologies recently proposed in the literature, ARTIVA has the major advantage of combining efficient and well-tried techniques (Bayesian networks and RJ-MCMC sampler) in order to solve several related problems. First, with ARTIVA there is no need for prior information regarding either the number of regulatory phases or the number of regulatory interactions between parent and target genes. Starting from uninformative priors (such as truncated Poisson or uniform distributions, see Methods), the posterior distribution for the number of changepoints, their positions and the regulatory models within each recovered phase is directly obtained from the ARTIVA runs. Also, ARTIVA allows the detection of regulatory phases for individual genes. Finally, whereas many approaches – like Bayesian Dirichlet Equivalent (BDE) score in a dynamic context [13] or the TESLA framework [15]– require the expression measurements to be discretized, the ARTIVA procedure has the advantage to work directly with continuous datasets. Thus there is no need to set arbitrary thresholds to define up- and down-regulated groups of genes.

We demonstrate the performance of the ARTIVA algorithm by (*i*) applying it to simulated data (Table 1) and (*ii*) performing a comparative analysis of the ARTIVA and TESLA [15] results (Figure 3). Because the simulations were such that they mirror the biological data analyzed afterwards as much as possible, we gain considerable confidence in the output of the ARTIVA approach when used on the two datasets considered here. Overall, the algorithm shows very good performance in retrieving the simulated dynamic networks, except in extremely unfavourable conditions, such as too much noise in the data or time series that are not sufficiently long and dense. These exploratory studies allow us to interpret the ARTIVA outputs more reliably.

New biological insights into the Drosophila development and the yeast stress response

The two biological networks presented in this study (Figures 3 and 4) are very different, both from a biological and a technical point of view. Their respective analyses represent different challenges for the application of the ARTIVA algorithm. The '*Drosophila* life cycle' data is representative of data used for classical regulatory network inference; successive gene expression measurements spanning a given biological process - here the *Drosophila* development - in order to detect potential regulatory interactions from gene expression profiles. This data is particularly suited for the inference of a temporally varying regulation network, since (*i*) the number of time-points is large (more than 80% of all published time series expression datasets are short with 8 time-points or fewer [33]) and (*ii*) the transitions between the distinct stages of *Drosophila* development {Embryo (E), Larva (L), Pupa (P), Adult (A)} are well-described in the literature [28]. We can thus reasonably expect to identify changepoints precisely at transitions between life stages (Figure 3). On the other hand, discrimination between parent and target genes represents an important additional step towards a complete description of the genetic networks that control development. These inferred temporal changes can form hypotheses as to how we can interfere rationally with developmental processes; *e.g.* arresting development in a given state by selectively knocking down transcription factors or targets at a given developmental stage.

The 'benomyl' dataset represents a particular challenge for ARTIVA to retrieve a dynamic regulatory network

for two reasons. First, the number of time-points is extremely small (only 5 time-points), and no replicate data points are available. To manage the lack of data, we cluster genes with concordant transcription profiles and analyze them jointly with ARTIVA. This cluster analysis was possible because the maximal intracluster variability did not exceed 0.2 (see Additional file 4), a value that ARTIVA is able to manage based on our simulation results (Table 1). Second, in this *S. cerevisiae* dataset, it is known that the genes coding for key regulators of the stress response system, *i.e.* transcription factors Yap1, Pdr1 and Pdr3, exhibit flat expression patterns during stress condition (Additional file 4); this prevents the use of correlation measures with their expression profiles to identify causal relationships with their potential target genes. In this context, we needed to adapt the ARTIVA inference procedure in order to integrate gene expression profiles measured in the wild type and KO strains. Regulatory associations between parent and target genes are thus proposed if the deletion of a parent gene significantly alters the expression measurements of its target genes (compared to the WT situation). Compared to the previous study of Lucau-Danila *et al.* [30], the main benefit of ARTIVA analyses is that it provides a dynamic classification of the benomyl response genes (Figure 4). It also points out contributions of the Yrr1 and Pdr3 transcription factors, which were ignored in previous analyses. Interestingly, the versatile and non exclusive joint action of Pdr1 and Pdr3 in chemical stress response, together with the overlap with Yap1 activity, is supported by recent experimental data available on these two factors [32,34,35].

Conclusions

The comprehensive analysis suggests that the ARTIVA approach allows us to describe and reverse-engineer the dynamic aspects of molecular networks. Such time-varying networks provide a middle ground between networks homogeneous in time and explicit dynamical models. The latter require substantial further information in order to model the dynamics of biological systems [36]. Inferring such systems is a considerable statistical challenge and it has recently been shown that some parameters cannot be inferred with any degree of certainty from time-course data. This so-called sloppy behaviour [37,38] has been identified even in very simple dynamical systems. In contrast to classical network reverse engineering approaches such as dynamic Bayesian networks [18] and graphical Gaussian models [39], ARTIVA also allows us to construct more complex hypotheses where interactions may depend on time.

As no particular constraint is imposed to the change-point positions or to the succession in network topologies within phases, the ARTIVA model appears to be highly flexible. The results are not *a priori* directed toward any particular regulatory associations between genes. This flexibility can be extremely valuable, especially when no information regarding the studied biological process is available. But the rapid accumulation of data obtained with different experimental approaches gives the opportunity to acquire a more comprehensive picture of all the interactions between cellular components. To understand the biology of the studied systems better, the trend is clearly towards the aggregation of multiple sources of information. A natural future direction in the development of ARTIVA will be to incorporate data originating from different sources in the model. In particular, protein/DNA interaction data (ChIP-chip or ChIP-seq experiments) could be effective by replacing the uniform prior for the edges with a prior favouring edges that correspond to the experimentally identified interactions (see [40,41] for an illustration). Also, ARTIVA assumes independent network topologies within successive phases and can identify very different regulatory associations between two phases, even if the time delay between the phases is very short. This assumption was appropriate in case of biological models like the *Drosophila* development and the yeast stress response, mainly because those are processes in which transcriptional regulations are highly dynamic. However, when considering systems that evolve more smoothly or in case of datasets with a small number of time points, it would be interesting to incorporate a regularization scheme into ARTIVA in order to favour slight changes from one phase to the next one. Such an approach has already been initiated in [13] for discretized data and in [42] where the regularization scheme is based on a common network structure. There are still huge gaps in our knowledge of biological networks and of the dynamics they mediate. What triggers whether or not an interaction is present depends subtly on the cellular context, the complement of molecules inside a cell (if we focus attention of intra-cellular processes and networks) and their respective molecular interactions. Understanding all of these factors and their interplay will ultimately be crucial in order to design biological interventions rationally. But statistically inferring them poses a set of formidable challenges. The use of relatively simple mathematical models (such as vector-autoregressive processes) allows us to distil the essential dynamics of complex temporal processes in biological systems. Thus ARTIVA provides a platform for the analysis of transcriptomic data, which could be straightforwardly expanded to include other data, *e.g.* transcription factor activities or other proteomic measurements.

Additional material

Additional file 1: Supplementary Text S1 - Priors illustration and complete mathematical description of the RJMCMC procedure and of the Bayes factor computation.

Additional file 2: Supplementary Figure S1 - Principle of the simulation study.

Additional file 3: Supplementary Dataset S1 - Full edge list of the inferred time varying networks of the 'benomyl' data.

Additional file 4: Supplementary Text S2 - Supplementary results related to the 'benomyl' analyses.

Additional file 5: Supplementary Figure S2 - Expression measurements for the 18 clusters used in the 'benomyl' analyses.

Acknowledgements
We thank Liam Kelly and Tina Toni for helpful comments on this manuscript, Catherine Etchebest for helpful discussions and the anonymous reviewers for their constructive advices. SL and MPHS gratefully acknowledge support from the BBSRC. JB and GL gratefully acknowledge support from the Institut National de la Transfusion Sanguine. MPHS is a Royal Society Wolfson Research Merit Award holder.

Author details
[1]Center for Bioinformatics, Imperial College London, London, UK. [2]Laboratoire des Sciences de l'Image de l'Informatique et de la télédétection (LSIIT), UMR UdS-CNRS 7005, Université de Strasbourg, Strasbourg, France. [3]Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB), INSERM U 665, Paris, F-75015, France. [4]Université Paris Diderot - Paris 7, UMR-S665, Paris, F-75015, France. [5]INTS, Paris, F-75015, France. [6]Laboratoire de Génomique des Microorganismes, CNRS FRE 3214, Université Pierre et Marie Curie, Institut des Cordeliers, Paris, France. [7]Institute of Mathematical Sciences, Imperial College London, London, UK.

Authors' contributions
SL conceived, implemented the first version of ARTIVA algorithm and drafted the manuscript. JB optimized the ARTIVA source code and performed simulations. FD provided help with data analysis. MPHS and GL contributed equally to this work. MPHS provided help with the model selection formalism and drafted the manuscript. GL designed the experiments, performed data analysis and drafted the manuscript. All the authors read and approved the final manuscript.

Received: 22 March 2010 Accepted: 22 September 2010
Published: 22 September 2010

References

1. Luscombe N, Babu M, Yu H, Snyder M, Teichmann S, Gerstein M: **Genomic analysis of regulatory network dynamics reveals large topological change.** *Nature* 2004, **431**:308-312.
2. Seshasayee A, Bertone P, Fraser G, Luscombe N: **Transcriptional regulatory networks in bacteria: from input signals to output responses.** *Curr Opin Microbiol* 2006, **9(5)**:511-519.
3. Zhu J, B Z, Smith E, Drees B, Brem R, Kruglyak L, Bumgarner R, Schadt E: **Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks.** *Nature Genetics* 2008, **40(7)**:854-861.
4. Barenco M, Tomescu D, Brewer D, Callard R, Stark J, Hubank M: **Ranked prediction of p53 targets using hidden variable dynamic modeling.** *Genome Biology* 2006, **7**:R27.
5. Bonneau R, Facciotti M, Reiss D, Schmid A, Pan M, Kaur A, Thorsson V, Shannon P, Johnson M, Bare J, Longabaugh W, Vuthoori M, Whitehead K, Madar A, Suzuki L, Mori T, Chang D, Diruggiero J, Johnson C, Hood L, Baliga N: **A predictive model for transcriptional control of physiology in a free living cell.** *Cell* 2007, **131(7)**:1354-1365.
6. Friedman N: **Inferring Cellular Networks Using Probabilistic Graphical Models.** *Science* 2004, **303**:799-805.
7. Hartemink A: **Reverse engineering gene regulatory networks.** *Nature Biotechnology* 2005, **23(5)**:554-555.
8. Werhli A, Grzegorczyk M, Husmeier D: **Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks.** *Bioinformatics* 2006, **22(20)**:2523-2531.
9. Yoshida R, Imoto S, Higuchi T: **Estimating Time-Dependent Gene Networks from Time Series Microarray Data by Dynamic Linear Models with Markov Switching.** *CSB '05: Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference* Washington, DC, USA: IEEE Computer Society 2005, 289-298.
10. Talih M, Hengartner N: **Structural learning with time-varying components: tracking the cross-section of financial time series.** *Journal of the Royal Statistical Society Series B* 2005, **67(3)**.
11. Fujita A, Sato J, Garay-Malpartida H, Morettin P, Sogayar M, Ferreira C: **Time-varying modeling of gene expression regulatory networks using the wavelet dynamic vector autoregressive method.** *Bioinformatics* 2007, **23(13)**:1623-1630.
12. Xuan X, Murphy KP: **Modeling changing dependency structure in multivariate time series.** *ICML* ACM International Conference Proceeding Series 2007, **227**:1055-1062.
13. Robinson J, Hartemink A: **Non-Stationary Dynamic Bayesian Networks.** *NIPS '08: Neural Information Processing Systems 2008* 2008, 1369-1376.
14. Rao A, Hero AO, States DJ, Engel JD: **Inferring Time-Varying Network Topologies from Gene Expression Data.** *EURASIP Journal on Bioinformatics and Systems Biology* 2007.
15. Ahmed A, Xing EP: **Recovering time-varying networks of dependencies in social and biological studies.** *Proceedings of the National Academy of Sciences* 2009, **106(29)**:11878-11883.
16. Needham C, Bradford J, Bulpitt A, Westhead D: **Inference in Bayesian networks.** *Nature Biotechnology* 2006, **24**:51-53.
17. Alon U: **Network motifs: theory and experimental approaches.** *Nature Reviews Genetics* 2007, **8(6)**:450-461.
18. Sachs K, Perez O, Pe'er D, Lauffenburger D, Nolan G: **Causal protein-signaling networks derived from multiparameter single-cell data.** *Science* 2005, **308(5721)**:523-529.
19. Lebre S: **Inferring dynamic genetic networks with low order independencies.** *Statistical Applications in Genetics and Molecular Biology* 2009, **8**.
20. de Silva E, Stumpf M: **Complex networks and simple models in biology.** *J Roy Soc Interface* 2005, **2**:419-340.
21. Green P: **Reversible jump Markoc chain Monte Carlo computation and Bayesian model determination.** *Biometrika* 1995, **82**:711-732.
22. Robert C, Ryden T, Titterington D: **Bayesian inference in hidden Markov models through reversible jump Markov chain Monte Carlo.** *Journal of the Royal Statistical Society B* 2000, **62**:57-75.
23. Green PJ: **Reversible jump Markov chain Monte Carlo computation and Bayesian model determination.** *Biometrika* 1995, **82**:711-732.
24. Suchard M, Weiss R, Dorman K, Sinsheimer J: **Inferring spatial phylogenetic variation along nucleotide sequences: a multiple change-point model.** *Journal of the American Statistical Assocation* 2003, **98**:427-437.
25. Andrieu C, Doucet A: **Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC.** *IEEE Trans. on Signal Processing* 1999, **47(10)**:2667-2676.
26. Zellner A: **On assessing prior distributions and Bayesian regression analysis with g-prior distribution.** In *Bayesian Inference and Decision Techniques.* Edited by: Goel PK, Zellner A. New York: Elsevier; 1986:233-243.
27. Kass RE, Raftery AE: **Bayes Factors.** *Journal of the American Statistical Association* 1995, **90**:773-795.
28. Arbeitman MN, Furlong F, Imam E Eand, Johnson E, Null BH, Baker BS, Krasnow MA, Scott MP, Davis RW, P WK: **Gene expression during the life cycle of drosophila melanogaster.** *Science* 2002, **297(5590)**:2270-2275.
29. The Gene Ontology Consortium: **Gene ontology: tool for the unification of biology.** *Nature Genetics* 2000, **25**:25-29 [http://www.geneontology.org].
30. Lucau-Danila A, Lelandais G, Kozovska Z, Tanty V, Delaveau T, Devaux F, Jacq C: **Early Expression of Yeast Genes Affected by Chemical Stress.** *Mol Cell Biol* 2005, **25(5)**:1860-1868.
31. Monteiro PT, Mendes ND, Teixeira MC, d'Orey S, Tenreiro S, Mira NP, Pais H, Francisco AP, Carvalho AM, Lourenco AB, Sa-Correia I, Oliveira AL, Freitas AT: **YEASTRACT-DISCOVERER: new tools to improve the analysis of transcriptional regulatory associations in Saccharomyces cerevisiae.** *Nucl*

Acids Res 2008, **36(suppl_1)**:D132-136 [http://nar.oxfordjournals.org/cgi/content/abstract/36/suppl_1/D132].
32. Salin H, Fardeau V, Piccini E, Lelandais G, Tanty V, Lemoine S, Jacq C, Devaux F: **Structure and properties of transcriptional networks driving selenite stress response in yeasts.** *BMC Genomics* 2008, **9**:333 [http://www.biomedcentral.com/1471-2164/9/333].
33. Ernst J, Nau GJ, Bar-Joseph Z: **Clustering short time series gene expression data.** *Bioinformatics (Oxford, England)* 2005, **21(Suppl 1)**: i159-168, [PMID: 15961453].
34. Banerjee D, Lelandais G, Shukla S, Mukhopadhyay G, Jacq C, Devaux F, Prasad R: **Responses of Pathogenic and Nonpathogenic Yeast Species to Steroids Reveal the Functioning and Evolution of Multidrug Resistance Transcriptional Networks.** *Eukaryotic Cell* 2008, **7**:68-77 [http://ec.asm.org/cgi/content/abstract/7/1/68].
35. Fardeau V, Lelandais G, Oldfield A, Salin H, Lemoine S, Garcia M, Tanty V, Crom SL, Jacq C, Devaux F: **The Central Role of PDR1 in the Foundation of Yeast Drug Resistance.** *J Biol Chem* 2007, **282(7)**:5063-5074 [http://www.jbc.org/cgi/content/abstract/282/7/5063].
36. Toni T, Welch D, Strelkowa N, Ipsen D, Stumpf M: **Approximate Bayesian Computation scheme for parameter inference and model selection in dynamical systems.** *J Roy Soc Interface* 2009, **6**:187-202.
37. Gutenkunst RN, Waterfall JJ, Casey FP, Brown KS, Myers CR, Sethna JP: **Universally sloppy parameter sensitivities in systems biology models.** *PLoS Comput Biol* 2007, **3**:1871-1878.
38. Gutenkunst RN, Casey FP, Waterfall JJ, Myers CR, Sethna JP: **Extracting falsifiable predictions from sloppy models.** *Ann N Y Acad Sci* 2007, **1115**:203-11.
39. Lauritzen SL: *Graphical models* Oxford University Press 1996.
40. Werhli A, Husmeier D: **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, **6.**
41. Mukherjee S, Speed TP: **Network inference using informative priors.** *Proceedings of the National Academy of Sciences (PNAS)* 2008, **105(38)**:14313-14318.
42. Grzegorczyk M, Husmeier D: **Non-stationary continuous dynamic Bayesian networks.** In *Advances in Neural Information Processing Systems (NIPS)* Edited by: Bengio Y, Schuurmans D, Lafferty J, Williams CKI, Culotta A 2009, **22**:682-690.

doi:10.1186/1752-0509-4-130
Cite this article as: Lèbre *et al.*: Statistical inference of the time-varying structure of gene-regulation networks. *BMC Systems Biology* 2010 **4**:130.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central