Learning the structure of an *in vivo* gene regulatory network using Gaussian processes

{tarmo.aijo, harri.lahdesmaki}@tut.fi

TAMPERE UNIVERSITY OF TECHNOLOGY

Abstract

Revealing the structure and dynamics of gene regulatory networks (GRNs) is of great interest and represents a considerably challenging computational problem. The GRN estimation problem is complicated by the fact that the number of gene expression measurements is typically extremely small when compared to the dimension of the biological system. Many recent GRN inference methods are founded on ordinary differential equations (ODE). However, because gene regulation process is intrinsically complex, commonly used parametric models can provide too simple description of the underlying phenomena and, thus, can be unreliable.

The presented novel framework is based on the use of Bayesian analysis with ODEs and non-parametric Gaussian process modeling for the transcriptional level regulation. It is applicable to structure learning and prediction of the dynamic behavior of a GRN. The uncertainty in measurements is taken into account by using a noise model and Bayesian analysis, and uncertainty in the phenomena is tackled by utilizing non-parametric modeling.

The performance of the proposed structure and dynamics inference method is evaluated using the recently published in vivo reverse-engineering and modeling assessment (IRMA) data set. By comparing the obtained structure inference results with those of existing ODE-based inference methods we demonstrate that the proposed method provides more accurate network structure learning. By splitting the IRMA data set into training and test sets, we also demonstrate that the model is able to capture dynamics of the system.

Model

Let $x_i(t)$ denote the expression of gene i at time t and vector $\hat{\mathbf{x}}_i(t)$ denote the expressions of genes that regulate gene i. The model of gene regulation is a first-order differential equation

$$\dot{x_i}(t) \simeq \Delta x_i(t) = \alpha_i + f_i\big(\hat{\mathbf{x}}_i(t)\big) - \lambda_i x_i(t),$$

where α is the basal rate, f_i is an unknown regulatory function of gene expressions and λ_i is the degradation rate. If one wants to infer regulatory interactions from steady-state measurements, then the rate of expression is set to zero

$$\dot{x_i}(t) \simeq 0 = \Delta x_i(t) = \alpha_i + f_i\left(\hat{\mathbf{x}}_i(t)\right) - \lambda_i x_i(t).$$

In practice, the exact values of α_i and λ_i are unknown, thus we assign distribution to them

$$\Delta x_i(t) = f_i(\hat{\mathbf{x}}_i(t)) + \mathbf{h}(x_i(t))\,[\alpha_i,\, \lambda_i]^{\mathrm{T}},$$

where $\mathbf{h}(x_i(t)) = [1, -x_i(t)]$ and $[\alpha_i,\, \lambda_i]^{\mathrm{T}}$ has a normal prior $\mathcal{N}(\mathbf{b}, B)$. Later on, we will integrate them out to get the complete marginal likelihood.

Gaussian Processes

One of the key ideas behind our method is to use Gaussian processes to learn the unknown regulation function $f_i(\cdot)$ from the data. We may write

$$g_i(x_i, \hat{\mathbf{x}}_i) \sim \mathcal{GP}\left(\mathbf{h}(x_i)^{\mathrm{T}}\mathbf{b},\; k(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_i') + \mathbf{h}(x_i)^{\mathrm{T}} B \mathbf{h}(x_i')\right),$$

where $g_i(x_i, \hat{\mathbf{x}}_i)$ represents Δx_i. Under these assumptions, it is possible to derive the marginal likelihood and prediction equations in analytical form. The hyperparameters of the Matérn covariance function are optimized (ML-II) by maximizing the (log) marginal likelihood with a conjugate gradient method.

Model Selection

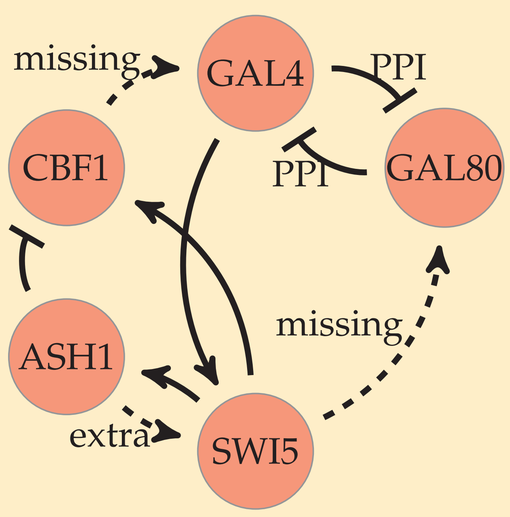
Let $\mathcal{M}_j$ denote a network structure. The posterior probability of a given model $\mathcal{M}_j$ can be obtained by applying Bayes' theorem

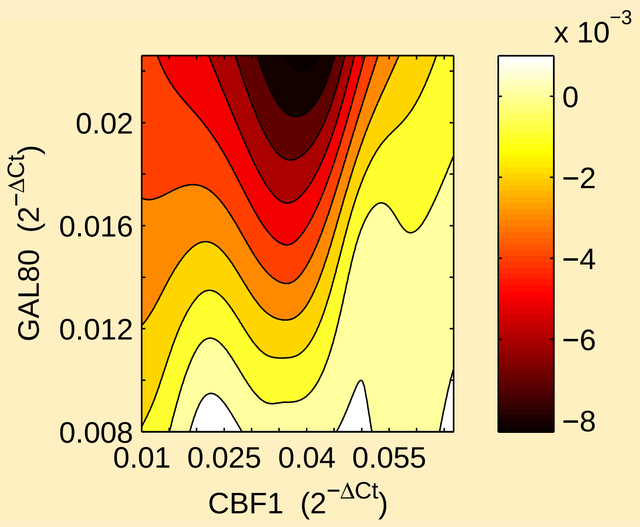
$$p(\mathcal{M}_j|\Delta\mathbf{x}_i, X) = \frac{p(\mathcal{M}_j)p(\Delta\mathbf{x}_i|X, \mathcal{M}_j)}{\sum_j p(\Delta\mathbf{x}_i|X, \mathcal{M}_j)p(\mathcal{M}_j)},$$

where terms $p(\Delta\mathbf{x}_i|X, \mathcal{M}_j)$ are obtained by evaluating the optimized marginal likelihood for each gene i (corresponding to the explanatory variables specified by $\mathcal{M}_j$) and $p(\mathcal{M}_j)$ is the prior probability of the network structure $\mathcal{M}_j$. Because the model selection relies on the marginal likelihood, the variable selection automatically favors models that are explanatory but at the same time not too complex.

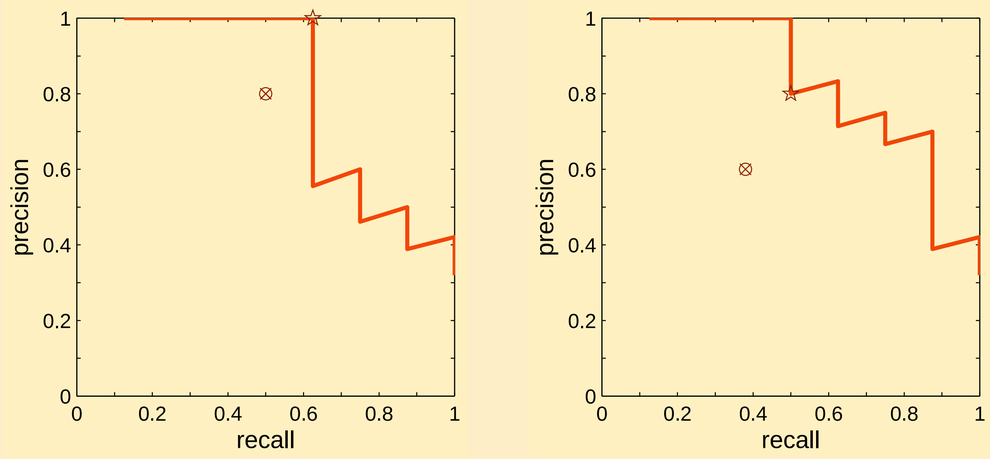
Results

The method was evaluated on the IRMA dataset which contains two time-series and two steady-state data sets.

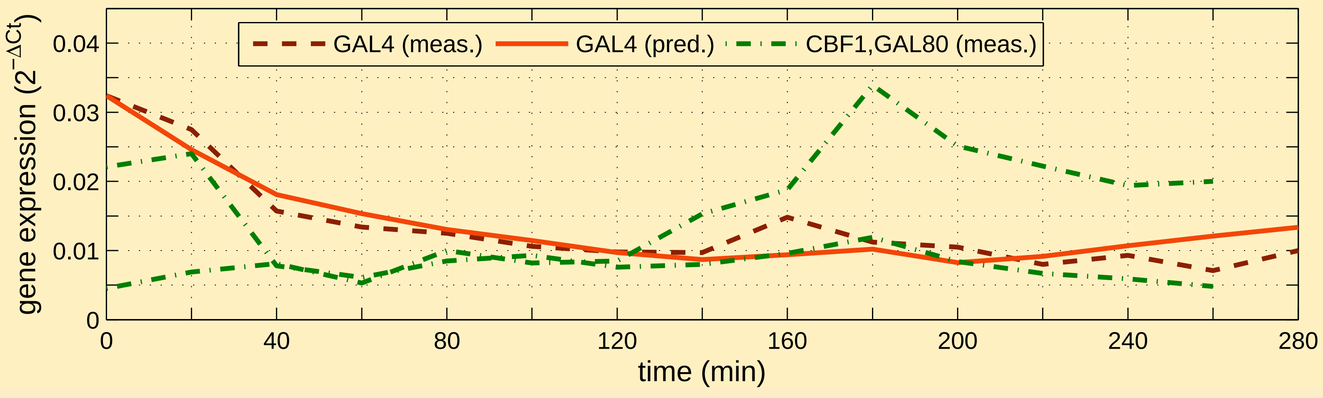




The topology of the inferred IRMA network ($p = 0.5$) from switch-on and switch-off data set and the regulatory function of GAL4 learned from switch-off data set. Our method found out correctly the regulatory genes of GAL4.. It can be seen that the method found out correctly that gene CBF1 is an activator and gene GAL80 is a repressor. This demonstrates that the proposed non-parametric model is able to learn the regulatory role of different explanatory variables even in the case of combinatorial regulation.



Plotted are the P-ROC curves on switch-on and switch-off data sets and the TSNI results.



Predicted expression profile of gene GAL4. Switch-off data set was used as a training set and switch-on as a test set. As before, the method found out correctly the regulatory genes, i.e., this combination of regulatory genes had highest posterior probability.

References

[1] T. Äijö and H. Lähdesmäki, Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics, Submitted

[2] I. Cantone *et al.* (2009), A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches, In *Cell*

Acknowledgments

This work was supported by the Academy of Finland, project no. 213462 (Finnish Programme for Centres of Excellence in Research 2006-2011) and the Finnish Foundation for Technology Promotion.