1 Introduction

Many systems biology models consist of interacting sets of differential equations. Given full knowledge of the functions involved in these systems traditional methods of systems identification could be used to parameterize the models. A particularly challenging aspect of biological systems is the scarcity of the data. Some of the model components may be unobservable, and those that are observed are not normally sampled at a high rate, particularly not in the context of high throughput experiments such as gene expression microarray and RNA-sequencing. This means that important sources of information are missing. In this chapter we consider a two pronged approach to dealing with this problem. In particular we look at modeling missing components of the model through a basis function approximation to the functions that can't be observed. We look at generalized linear models as a potential approach to mapping these missing nonlinear functions. Unfortunately, as the complexity of these models increases, so does the number of parameters. We therefore consider Bayesian approaches to dealing with the increasing number of parameters. As we increase the complexity of the model further, we find that these generalized linear models converge on a particular probabilistic process known as a Gaussian process. We then explore the use of Gaussian processes in biological modeling with a particular focus on transcriptional networks and gene expression data. We will review simple cascade models of translation and transcription that can be used for genome-wide target identification. We then introduce nonlinearities into the differential equations. This means that we need to resort to sampling methods, such as Markov chain Monte Carlo, to perform inference in the systems. We briefly review an efficient approach to sampling and demonstrate its use in modeling transcriptional regulation.

1.1 A Simple Systems Biology Model

Gene expression is governed by transcription factors. To analyze expression, we need to form a mathematical idealization of this process. The exact mechanism behind transcription is not yet fully understood, but a coarse mathematical representation of the system would assume that the rate of production of a given gene's mRNA is dependent on the amount of transcription factor in the cell. When collating gene expression from many cells in a tissue we might model the rate of mRNA production through an ordinary differential equation. Barenco et al. (2006) described just such a model of gene expression,

$$\frac{\mathrm{d}m_j(t)}{\mathrm{d}t} = b_j + s_j p(t) - d_j m_j(t), \tag{1}$$

where the rate of production of the jth gene's mRNA is dependent on a basal rate of transcription, b_j, a sensitivity, s_j to the governing transcription factor concentration, $p(t)$ and a rate of decay, d_j. In this simple model each of the genes is assumed to be governed by a single input transcription factor and this is known as a single input module motif (Alon, 2006). Short of assuming that the mRNA

concentration is linearly related to the transcription factor concentration[1] this is perhaps the simplest model of transcriptional regulation that can be composed. It has a closed form solution for the mRNA concentration that depends on the transcription factor (TF) concentration. If we assume that the system starts at $t = 0$ then we have

$$m_j(t) = a_j e^{-d_j t} + \frac{b_j}{d_j} + s_j e^{-d_j t} \int_0^t e^{d_j u} p(u) \mathrm{d}u \tag{2}$$

where the initial value for the mRNA concentration is given by $m_j(0) = \frac{b_j}{d_j} + a_j$.

The simple model described above can be extended through nonlinear response, active degradation of the mRNA, stochastic effects, and so on. However, it clearly illustrates the main issue we wish to address in this chapter. Namely, how do we deal with the fact that the transcription factor concentration, $p(t)$ may be unobservable?

Our preliminary studies in this area were inspired by the work of Barenco et al. (2006) who (ignoring the transient term) reordered the equation so that for each observation time, $t_1, \ldots, t_n$, they had

$$p(t_k) = s_j^{-1} \left(\frac{\mathrm{d}m_j(t_k)}{\mathrm{d}t} - b_j + d_j m_j(t_k) \right).$$

They then created pseudo-observations of the production rates of the mRNA, $\frac{\mathrm{d}m_j(t_i)}{\mathrm{d}t}$, through fitting polynomials to their time series and computing gradients at each time point. This gave them an estimate for the TF concentration and they used Bayesian sampling to estimate the model parameters.

Khanin et al. (2006) considered a similar equation with a nonlinear repression response. They and Rogers et al. (2006), who also considered single input motifs, dealt with the missing function through fitting a piecewise constant approximation to the missing function. In this chapter our focus will be on an alternative approach: modeling the missing function with a generalized linear model.

2 Generalized Linear Model

For our purposes a generalized linear model can be seen as constructing a function through a weighted sum of nonlinearities. Formally, we assume that a function, $f(t)$ can be represented as

$$f(t) = \sum_{k=1}^{M} w_k \phi_k(t),$$

[1] This would be equivalent to assuming very high decay in the simple model we describe. Models like this are widely used. One can see clustering to find coregulated targets as making this assumption. The assumption is made more explicitly in some genomewide analysis models (e.g. Sanguinetti et al., 2006a,b).

where the basis function has a nonlinear form. One possibility is a basis derived from a Gaussian form,

$$\phi_k(t) = \frac{1}{\sqrt{\pi \ell_k^2}} \exp\left(-\frac{(t-\tau_k)^2}{\ell_k^2}\right). \tag{3}$$

Here τ_k represents a location parameter which gives the center of the basis function and ℓ represents a timescale parameter which gives the width of the basis function (or the timescale over which it is active). As the distance between t and τ_k increases the basis function approaches zero. This is therefore sometimes described as a "local" basis function. In Figure 1 we show a set of bases and some nonlinear functions that can be derived from them using different weights.

Our aim is to introduce this representation of the function into our model of the system. We assume that the transcription factor concentration can be represented through basis functions so we have,

$$p(t) = \sum_{k=1}^{M} w_k \phi_k(t).$$

Taking the simple linear differential equation model given in (1) we can substitute in our representation of the transcription factor concentration and recover

$$\frac{\mathrm{d}m_j(t)}{\mathrm{d}t} = b_j + s_j \sum_{k=1}^{M} w_k \phi_k(t) - d_j m_j(t).$$

Substituting this into the solution for $m_j(t)$ we have

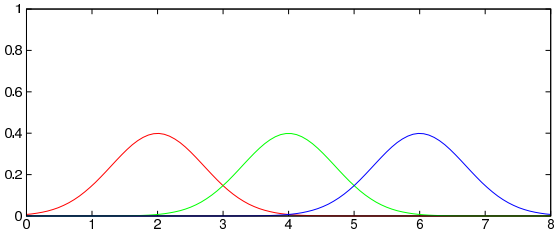
$$m_j(t) = a_j e^{-d_j t} + \frac{b_j}{d_j} + s_j \sum_{k=1}^{M} w_k e^{-d_j t} \int_0^t e^{d_j u} \phi_k(u) \mathrm{d}u \tag{4}$$

where we have been able to pull the weighted sum over the basis functions outside the integral. Our solution for the mRNA concentration is now also expressed as a generalized linear model. Now the basis set consists of a transient term, $a_j e^{-d_j t}$, a constant term, $\frac{b_j}{d_j}$, and a weighted sum of convolutions of our original basis. For some choices for $\phi_k(t)$ the integral in (4) will be tractable. In particular if we use the Gaussian form shown in (3) we can show that,

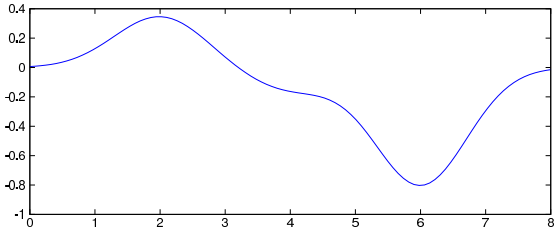
$$\begin{aligned} e^{-d_j t} \int_0^t e^{d_j u} \phi_k(u) \mathrm{d}u = & e^{-d_j (t-\tau_k)} e^{\frac{d_j^2 \ell_k^2}{4}} \frac{1}{2} \left[\operatorname{erf}\left(\frac{t - \left(\tau_k + \frac{d_j \ell_k^2}{2}\right)}{\ell_k} \right) \right. \\ & \left. - \operatorname{erf}\left(-\frac{\tau_k + \frac{d_j \ell_k^2}{2}}{\ell_k} \right) \right] \end{aligned}$$

where $\operatorname{erf}(\cdot)$ is the error function defined as

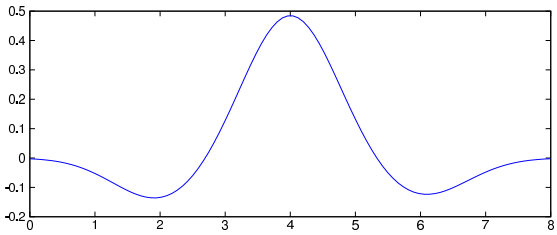
$$\operatorname{erf}(x) = \int_0^x \frac{2}{\sqrt{\pi}} \exp\left(-z^2\right) \mathrm{d}z.$$



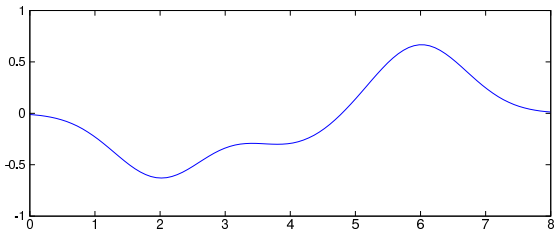
(a) Set of three basis functions.



(b) Linear combination of basis with $w_1 = 0.875$, $w_2 = -0.388$, and $w_3 = -2.01$.



(c) Linear combination of basis with $w_1 = -0.359$, $w_2 = 1.23$, and $w_3 = -0.328$.



(d) Linear combination of basis with $w_1 = -1.56$, $w_2 = -0.74$, and $w_3 = 1.69$.

Figure 1: Some functions based on a simple basis set with three members. Location parameters of the basis functions are set to $\tau_k = 2, 4, 6$ and the time scale of the bases is set to 1. Functions derived from these bases are shown for different weights. Each weight was sampled from a standard normal density.

For fixed parameters, a_j, s_j, d_j, b_j there is a deterministic relationship between the output mRNA concentration, $m_j(t)$ and a governing TF concentration, $p(t)$. In Figure 2 we show functions that correspond to high, medium and low decay rates that result from solving the differential equation for the different TF concentrations shown in Figure 1. By setting $\{a_j, b_j\}_{j=1}^3$ to zero and $\{s_j\}_{j=1}^3$ to one we ensure that the differences between all the solutions are arising only from the different responses to the underlying basis from Figure 1. The solution for each individual basis function is shown in the top row of Figure 2. The solution for each different set of weights from Figure 1 is then simply the weighted sum of the relevant weights multiplied by the convolved basis functions.

2.1 Fitting Basis Function Models

Given a basis set, $\{\phi_k(t)\}_{k=1}^M$, a set of weights, $\{w_k\}_{k=1}^M$, and the parameters of the differential equation, $\{b_j, d_j, s_j\}_{j=1}^p$, we can solve the differential equations for the mRNA concentrations, $\{m_j(t)\}_{j=1}^p$. We can determine all these parameters of the model (including the parameters of the basis functions) by fitting through maximum likelihood. If we assume Gaussian noise of variance σ^2 we can compute the log likelihood of an observed data set of mRNA concentrations, perhaps obtained from a gene expression microarray,

$$\log p(\mathbf{y}|\mathbf{w}) = -\frac{p}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{j=1}^{p}\sum_{i=1}^{n}(m_j(t_i) - y_{i,j})^2,$$

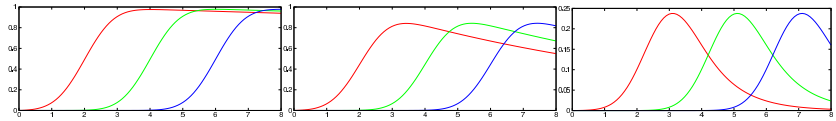
where the gene expression data has been collected at times $\{t_i\}$ and has values given by $\{y_{i,j}\}$ for the jth gene. We have used the notation $m_j(t_i)$ to denote the value of the predicted mRNA concentration for the jth gene at the ith time point. Bearing in mind that this is dependent on the parameters of the jth differential equation and the basis functions, we can maximize this log likelihood with respect to all these parameters and the noise variance. Here, in our representation of the likelihood $p(\mathbf{y}|\mathbf{w})$, we have only made explicit the conditioning on the vector of weights, $\mathbf{w} = [w_1, \ldots, w_M]^\top$, as we will mainly focus on maximization with respect to this vector. With such a probabilistic formulation we can also exploit the PUMA (Milo et al., 2003; Liu et al., 2005) framework for propagating uncertainty through microarray analysis and associate an individual variance $\sigma_{i,j}$ with each gene expression observation, $y_{i,j}$.

Working with the convolved basis set we have

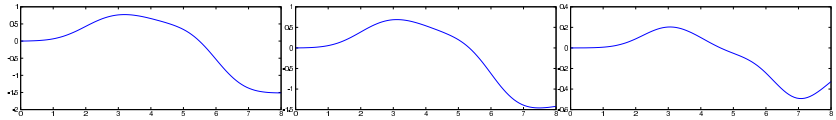
$$\phi'_{j,k}(t) = s_j e^{-d_j t}\int_0^t e^{d_j u}\phi_k(u)\mathrm{d}u.$$

We can then write a particular mRNA concentration as

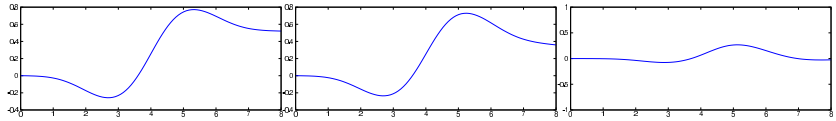
$$m_j(t) = a_i e^{-d_j t} + \frac{b_j}{d_j} + \mathbf{w}^\top \boldsymbol{\phi}'_j(t)$$



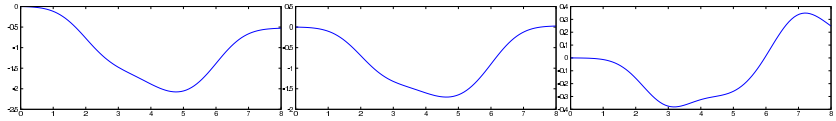
(a) Three basis functions after convolution. *Left*: $d_1 = 0.01$, *middle*: $d_2 = 0.1$, *right* $d_3 = 1$



(b) Linear combination of convolved basis with $w_1 = 0.875$, $w_2 = -0.388$, and $w_3 = -2.01$.



(c) Linear combination of convolved basis with $w_1 = -0.359$, $w_2 = 1.23$, and $w_3 = -0.328$.



(d) Linear combination of convolved basis with $w_1 = -1.56$, $w_2 = -0.74$, and $w_3 = 1.69$.

Figure 2: The bases in Figure 1 have been pushed through the differential equation to give the convolved basis shown in (a). We took $a_j = 0$, $b_j = 0$ and $s_j = 1$ for $j = 1, 2, 3$. Decays were then set to be low, $d_1 = 0.01$, medium, $d_2 = 0.1$, and high, $d_3 = 1$. in (b)-(d) we see mRNA concentrations derived from these bases with different weights, w_k. Weights match those shown in Figure 1. In each row the leftmost plot is the result of convolving with a low decay rate ($d_1 = 0.01$), the middle plot a medium decay rate ($d_2 = 0.1$) and the rightmost a higher decay rate ($d_3 = 1$).

where $\mathbf{w}$ is a vector of the values $\{w_k\}_{k=1}^M$ and $\boldsymbol{\phi}'(t)$ is a vector valued function representing the values of the basis functions $\left\{\phi'_{j,k}(t)\right\}_{k=1}^M$. Ignoring the basal rate and a_i for the moment, we can write the log likelihood as

$$p(\mathbf{y}|\mathbf{w}) = -\frac{p}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\left(-\mathbf{w}^\top \sum_{j=1}^{p}\sum_{i=1}^{n}\boldsymbol{\phi}'_j(t_i)\boldsymbol{\phi}'_j(t_i)^\top\mathbf{w} + 2\sum_{j=1}^{p}\sum_{i=1}^{n}\mathbf{w}^\top\boldsymbol{\phi}'_j(t_i)y_{i,j} + y_{i,j}^2\right)$$

which can be maximized with respect to $\mathbf{w}$ by finding a stationary point of the log likelihood,

$$\mathbf{w} = \left[\sum_{j=1}^{p}\sum_{i=1}^{n}\boldsymbol{\phi}'_j(t_i)\boldsymbol{\phi}'_j(t_i)^\top\right]^{-1}\sum_{j=1}^{p}\sum_{i=1}^{n}\boldsymbol{\phi}'_j(t_i)y_{i,j}.$$

If we require a lot of flexibility for our model of the TF concentration, $p(t)$, we can use a large number of basis functions, M. However, as we increase the number of basis functions we may find that we are no longer able to compute the maximum likelihood solution for $\mathbf{w}$. If $M > np$ then the matrix given by $\sum_{j=1}^{p}\sum_{i=1}^{n}\boldsymbol{\phi}'_j(t_i)\boldsymbol{\phi}'_j(t_i)^\top$ will not be invertible. A solution to this problem is to handle $\mathbf{w}$ through Bayesian inference (see Lawrence and Rattray (2010) for a brief introduction). In Bayesian inference, parameters are treated with a prior distribution, in this case $p(\mathbf{w})$, and rather than being maximized they are integrated out:

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{w})p(\mathbf{w})\mathrm{d}\mathbf{w}.$$

The quantity $p(\mathbf{y})$ is called the marginal likelihood, integrated likelihood or evidence of the model, which is important for model comparison (MacKay, 2003). If we place a zero mean Gaussian prior density over $\mathbf{w}$,

$$p(\mathbf{w}) = \frac{1}{\sqrt{2\pi}\,|\mathbf{C}|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}\mathbf{w}^\top\mathbf{C}^{-1}\mathbf{w}\right)$$

we can compute the marginal likelihood of the data analytically and we find that it is given by

$$p(\mathbf{y}) = \frac{1}{\sqrt{2\pi}\,|\mathbf{K}+\sigma^2\mathbf{I}|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}\mathbf{y}^\top\left(\mathbf{K}+\sigma^2\mathbf{I}\right)^{-1}\mathbf{y}\right),$$

where $\mathbf{y}$ is a "stacked" version of the data matrix, i.e. it is composed of the columns from $\mathbf{y}$ stacked on top of one another. The matrix $\mathbf{K}$ gives the covariance between the different mRNA concentrations at different times. It is structured as a block matrix with $p \times p$ blocks, each block of size $n \times n$. The diagonal blocks give covariances of a single gene output, while off-diagonal blocks give *cross covariances* between different outputs. In general we can define each element of $\mathbf{K}$ to be $k_{i,j}(t,t') = \boldsymbol{\phi}'_i(t)^\top\mathbf{C}\boldsymbol{\phi}'_j(t')$. This denotes the covariance between the i and jth mRNA concentrations computed between times t and t'.

2.1.1 Relationship with Basis of $p(t)$

The model as we described is dependent on our choice of basis function for $p(t)$. An interesting feature is that we can represent all the elements of the marginal likelihood's covariance matrix through the inner product of the basis, $k_{0,0}(t,t') = \boldsymbol{\phi}(t)^\top \mathbf{C}\boldsymbol{\phi}(t')$:

$$k_{i,j}(t,t') = s_j s_i e^{-d_i t - d_j t'} \int_0^t e^{d_i u} \int_0^{t'} e^{d_j u'} \underbrace{\boldsymbol{\phi}(u)^\top \mathbf{C}\boldsymbol{\phi}(u')}_{k_{0,0}(u,u')} \mathrm{d}u'\mathrm{d}u. \tag{5}$$

In fact this equation holds in general: as long as $k_{0,0}(t,t')$ is a positive definite function, it represents an inner product of a basis set. If we consider a spherical prior for $\mathbf{w}$, so $\mathbf{C} = \gamma\mathbf{I}$, then we can write $k_{0,0}(t,t') = \gamma\boldsymbol{\phi}(t)^\top\boldsymbol{\phi}(t')$. Now, instead of maximizing over all different weight parameters, $\mathbf{w}$, we only need to find γ. Through marginalization we have reduced the number of parameters in the system by $M-1$. A problem remains though, each basis function has a center, τ_k, and these locations represent an additional M parameters that also need to be determined. We now show how these parameters can also be eliminated. We will consider a limit that allows us to effectively take the number of basis functions, $M \to \infty$.

2.2 An Infinite Basis

Now we have eliminated the parameters $\mathbf{w}$ we are left to decide the number of basis functions, M, and the location of these basis functions, $\{\tau_k\}_{k=1}^M$. First we will explore what happens when we place those basis functions at uniform intervals over time so we have

$$p(t) = \sum_{k=1}^{M} w_k \frac{1}{\sqrt{\pi\ell^2}} \exp\left(-\frac{(t-a-\Delta\tau\cdot k)^2}{\ell^2}\right),$$

where we have set the location parameter of each $\phi_k(t)$ to

$$\tau_k = a + \Delta\tau \cdot k.$$

We showed that the marginal likelihood of the model is entirely dependent on the inner product between basis vectors at different times. For the basis at uniform intervals this can be written as

$$k_{0,0}\left(t,t'\right) = \frac{\gamma}{\pi\ell^2} \sum_{k=1}^{M} \exp\left(-\frac{t^2 + {t'}^2 - 2\left(a+\Delta\tau\cdot k\right)\left(t+t'\right) + 2\left(a+\Delta\tau\cdot k\right)^2}{\ell^2}\right).$$

More basis functions allow greater flexibility in our model of the TF concentration. We can increase the number of basis functions we use in the interval between a and b by decreasing the interval $\Delta\tau$. However, to do this without increasing the expected variance of the resulting TF concentration we need to

scale down the variance of the prior distribution for $\mathbf{w}$. This can be done by setting the variance parameter to be proportional to the interval, so we take $\gamma = \alpha\Delta\tau$. Now we have basis functions where the location of the leftmost basis function, $k = 1$, is $\tau_1 = a + \Delta\tau$ and the rightmost basis is $\tau_M = b$ so that $b = a + \Delta\tau \cdot M$. The fixed interval distance between a and b is therefore given by $b - a = (M-1)\Delta\tau$.

We are going to increase the number of basis functions by taking the limit as $\Delta\tau \to 0$. This will take us from a discrete system to a continuous system. In this limit the number of basis functions becomes infinite because we have $M = \lim_{\Delta\tau\to 0} \frac{b-a}{\Delta\tau} + 1$. In other words we are moving from a fixed number of basis functions to infinite basis functions. The inner product between the basis functions becomes

$$k_{0,0}(t,t') = \frac{\alpha}{\pi\ell^2}\int_a^b \exp\left(-\frac{t^2 + t'^2 - 2\tau(t+t') + 2\tau^2}{\ell^2}\right)\mathrm{d}\tau.$$

Completing the square gives

$$k_{0,0}(t,t') = \frac{\alpha}{\pi\ell^2}\int_a^b \exp\left(-\frac{t^2 + t'^2 + 2\left(\tau - \frac{1}{2}(t+t')\right)^2 - \frac{1}{2}(t+t')^2}{\ell^2}\right)\mathrm{d}\tau$$

$$k_{0,0}(t,t') = \frac{\alpha}{\sqrt{2\pi\ell^2}}\exp\left(-\frac{(t-t')^2}{2\ell^2}\right)\sqrt{\frac{2}{\pi\ell^2}}\int_a^b \exp\left(-\frac{2}{\ell^2}\left(\tau - \frac{t+t'}{2}\right)^2\right)\mathrm{d}\tau,$$

$$k_{0,0}(t,t') = \frac{\alpha}{\sqrt{2\pi\ell^2}}\exp\left(-\frac{(t-t')^2}{2\ell^2}\right)\frac{1}{2}\left[1 + \operatorname{erf}\left(\sqrt{\frac{2}{\ell^2}}\left(\tau - \frac{t+t'}{2}\right)\right)\right]_a^b,$$

Performing the integration leads to

$$\begin{aligned} k_{0,0}(t,t') =& \frac{\alpha}{\sqrt{2\pi\ell^2}}\exp\left(-\frac{(t-t')^2}{2\ell^2}\right)\frac{1}{2}\left[\operatorname{erf}\left(\sqrt{\frac{2}{\ell^2}}\left(b - \frac{t+t'}{2}\right)\right)\right. \\ & \left. - \operatorname{erf}\left(\sqrt{\frac{2}{\ell^2}}\left(a - \frac{t+t'}{2}\right)\right)\right], \end{aligned}$$

Finally, if we take the limit as $a \to -\infty$ and $b \to \infty$ (i.e. we have infinite basis functions distributed across the entire real line) then the square bracketed term on the right becomes 2 and we have

$$k_{0,0}(t,t') = \frac{\alpha}{\sqrt{2\pi\ell^2}}\exp\left(-\frac{(t-t')^2}{2\ell^2}\right), \tag{6}$$

which is known as the squared exponential covariance function (despite the fact that it is not a squared exponential, but an exponentiated quadratic).

The analysis above shows that if we take a one dimensional fixed basis function model, we can increase the number of basis functions to infinity and distribute them evenly across the real line. In this way we loose the requirement to specify M location parameters and further reduce the number of parameters in the system. Without the Bayesian approach we had $2(M+1)+4p$ parameters. The combination of the Bayesian approach and the use of infinite basis functions leads to $4p+3$ parameters without any loss of model flexibility. In fact the resulting model is more flexible as it is allowing for infinite basis functions. Instead of specifying the basis function directly, we now specify the covariance of the marginal through a positive definite function $k_{0,0}(t,t')$.

The procedure for moving from inner products, $\boldsymbol{\phi}(t)^\top\boldsymbol{\phi}(t')$, to covariance functions, $k_{0,0}(t,t')$, is sometimes known as kernelization (Schölkopf and Smola, 2001) due to the fact that the covariance function has the properties of a Mercer kernel. A function of two variables can be seen as a Mercer kernel when a symmetric matrix of values from $k_{0,0}(t,t')$ computed for a vector of times $\mathbf{t}$ is always positive semi-definite. In other words if $k_{i,j}=k_{0,0}(t_i,t_j)$ is the element from the ith row and jth column of $\mathbf{K}_{0,0}$ and t_i is the ith element from $\mathbf{t}$, we should have that $\mathbf{K}_{0,0}$ is a positive definite matrix for any vector of inputs $\mathbf{t}$. This same property is what allows it to be used as a covariance function: covariances must be positive definite. The matrix $\mathbf{K}_{0,0}$ specifies the covariance between instantiations of the function $p(t)$ at the times given by $\mathbf{t}$. Mercer's theorem says that underlying all such positive definite functions there is always a (possibly infinite) feature space, $\phi(t)$, which can be used to construct the covariance function. For our example the relationship between the feature space and this covariance function emerges naturally through considering a Bayesian approach to a fixed basis function model. The resulting model is known as a Gaussian process (O'Hagan, 1978). This perspective of converting from a parametric model to a covariance function is one way of introducing Gaussian processes (see also Williams (1997) and Rasmussen and Williams (2006)). For an alternative introduction which ignores the parametric interpretation see Lawrence et al. (2010b).

2.3 Gaussian Processes

Using the marginal likelihood of the model we described instead of the model parameterized by $\mathbf{w}$ can be understood as taking a Gaussian process perspective on the model for $p(t)$. Gaussian processes are powerful models for nonlinear functions. They allow us to specify the characteristics of a function through specifying its covariance function. Given a covariance function for the Gaussian process over p, the covariances between the target genes can also be computed from (5). Substituting $k_{0,0}(t,t')=\boldsymbol{\phi}(t)^\top\mathbf{C}\boldsymbol{\phi}(t')$ we have

$$k_{i,j}(t,t')=s_j s_i e^{-d_i t-d_j t'}\int_0^t e^{d_i u}\int_0^{t'} e^{d_j u'}k_{0,0}(u,u')\mathrm{d}u'\mathrm{d}u.$$

We can also compute the cross covariance between $p(t)$ and the ith output gene,

$$k_{0,i}(t,t') = s_i e^{-d_i t} \int_0^t e^{d_i u} k_{0,0}(u,t') \mathrm{d}u.$$

This arises in the joint distribution of $p(t)$ and $\{m_i(t)\}_{i=1}^p$ which is also Gaussian with covariance

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{0,0} & \mathbf{K}_{0,1} & \dots & \mathbf{K}_{0,p} \\ \mathbf{K}_{1,0} & \mathbf{K}_{1,1} & \dots & \mathbf{K}_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{K}_{p,0} & \mathbf{K}_{p,1} & \dots & \mathbf{K}_{p,p} \end{bmatrix}$$

where the matrix $\mathbf{K}_{i,j}$ is computed using $k_{i,j}(t,t')$ for the relevant observation times for the ith and jth function.

In a parametric model, given observations of the mRNA concentrations, we would compute the posterior distribution for $\mathbf{w}$ and use it to compute expectations of $p(t)$ and other gene concentrations. In the Gaussian process setting it may be that there are infinitely many parameters in $\mathbf{w}$ which makes it difficult to express distributions over this space. Instead, we simply condition on the observations in the probabilistic model. Our data is taken to be composed of noise corrupted observations of the mRNA concentrations (for example gene expression measurements).

$$y_j(t_i) = m_j(t_i) + \epsilon_{i,j}$$

where ϵ is a corruptive noise term which would be drawn independently, perhaps from a Gaussian $\epsilon_{i,j} \sim \mathcal{N}\left(0, \sigma_i^2 \mathbf{I}\right)$ for each observation. We augment this matrix with any potential observations of the TF concentration,

$$y_0(t_i) = p(t_i) + \epsilon_{i,0}.$$

An observation of the TF may be in the form of a constraint that the TF concentration is known to be zero at time zero: $p(0) = 0, \quad \sigma_0^2 = 0$. Or in some cases it could be a direct measurement. This gives us the full data set which we place in a stacked vector storing all the values,

$$\mathbf{y} = \left[\mathbf{y}_0^\top, \mathbf{y}_1^\top, \dots, \mathbf{y}_p^\top\right]^\top.$$

Note that the vectors of observations, $\mathbf{y}_0$, $\mathbf{y}_1$ ... need not be the same length or contain observations taken at the same times.

The joint distribution for the corrupted observations of the mRNA concentrations and protein is given by a Gaussian process with covariance $\mathbf{K}$,

$$\mathbf{y} \sim \mathcal{N}\left(\boldsymbol{\mu}, \mathbf{K}\right).$$

The mean vector, $\boldsymbol{\mu}$, here is obtained by computing the mean function, $\mu_j(t) = a_j e^{-d_j t} + \frac{b_j}{d_j}$ for the relevant times $\mathbf{t}$. The mean vector can be removed by

placing a zero mean Gaussian prior over a_j and b_j. In this case we obtain a Gaussian process with a zero mean function and a modified covariance function,

$$\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}')$$

where $k'_{0,0}(t,t') = k_{0,0}(t,t')$ but for $i > 0$, $j > 0$ we have

$$k'_{i,j}(t,t') = k_{i,j}(t,t') + \delta_{i,j}\left(\alpha_{a_j} e^{-\delta_j(t+t')} + \frac{\alpha_{b_j}}{d_j}\right),$$

where $\delta_{i,j}$ is the Kronecker delta function[2] and α_{a_j} and α_{b_j} are the variances of the priors over the a_j and b_j parameters respectively. These may be shared across all genes: $\alpha_{a_j} = \alpha_a$ and $\alpha_{b_j} = \alpha_b$. Similarly we can use a shared noise variance $\sigma_j^2 = \sigma^2$. This reduces the number of parameters sought to $2p + 5$.

The parameters of the model can be found by minimizing the negative log likelihood,

$$E(\boldsymbol{\theta}) = -\frac{1}{2}\log p(\mathbf{y}|\boldsymbol{\theta}) \tag{7}$$

$$= \frac{1}{2}\log|\mathbf{K}| + \frac{1}{2}\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y} + \text{const.} \tag{8}$$

with respect to the parameters $\boldsymbol{\theta} = \left\{\{s_j, d_j\}, \alpha, \alpha_a, \alpha_b, \ell, \sigma^2\right\}$. This requires us to compute the gradient of the covariance matrix with respect to each parameter, $\frac{\mathrm{d}\mathbf{K}}{\mathrm{d}\theta_i}$ and combine it with the gradient of the negative log likelihood,

$$\frac{\mathrm{d}E(\boldsymbol{\theta})}{\mathrm{d}\mathbf{K}} = -\frac{1}{2}\mathbf{K}^{-1} + \frac{1}{2}\mathbf{K}^{-1}\mathbf{y}\mathbf{y}^\top\mathbf{K}^{-1}\ .$$

The resulting gradients of the negative log likelihood can be used in a gradient based minimizer to find a local minimum. We often make use of scaled conjugate gradients (Møller, 1993), but other optimizers such as quasi-Newton approaches (see e.g. Zhu et al., 1997) or conjugate gradients can also be used. If there are few points in the time series the parameters may be badly determined. As a diagnostic the curvature of the log likelihood can be computed or alternatively we can place appropriate prior distributions over the parameters and the gradients of the corresponding joint distribution can be used in a hybrid Monte Carlo, also known as Hamiltonian Monte Carlo, algorithm (see e.g. MacKay, 2003, for details on hybrid Monte Carlo) to obtain samples from the posterior distribution for $\boldsymbol{\theta}$.

The framework we've described allows us to determine, from a set of known targets, the parameters of a set of linear ordinary differential equations that best govern those targets, assuming a single regulator. By retaining linear differential equations and specifying the TF concentration through a generalized linear model we can marginalize many of the model parameters. The use of Gaussian processes and prior distributions for the basal rate and a_j further reduce the number of parameters in the model to $2p+5$. In Section 3 we describe

[2]The Kronecker delta is defined as one if $i = j$ and zero otherwise.

an approach to ranking candidate targets of a given transcription factor based on this approach. However, we may be interested in more general modeling formulations. For example, our current model: a generalized linear model with Gaussian prior on the weights, gives us a prior distribution for the TF concentration that can become negative. However, even the simple fix of modeling the TF as a Gaussian process in log space,

$$\log p(t) \sim \mathcal{N}\left(\boldsymbol{\mu}_0, \mathbf{K}_{0,0}\right),$$

renders the elegant solution given above intractable. In the generalized linear model formulation the weights, $\mathbf{w}$, now appear inside the exponent,

$$p(t) = e^{\sum_{k=1}^{M} w_k \phi_k(t)},$$

denying us the ability to bring the weights outside the integral when convolving.

Even such a simple modification forces us to seek approximations to deal with this intractability. In this review we focus on the sampling approximations we have considered (Titsias et al., 2009, 2011), although we have also explored Laplace's approximation[3] (see Lawrence et al., 2007; Gao et al., 2008; Lawrence et al., 2010b) and variational approximations could also be applicable.

2.4 Sampling Approximations

Regardless of how we specify the prior distribution for the TF concentration, for a given sample from this function, $p^{(i)}(t)$ we should be able to solve the system of differential equations. For linear differential equations this requires solving the convolution integral, and for non-linear differential equations we need to apply numerical methods such as Runge-Kutta. Irrespective of the manner in which we find the solution, the sample $p^{(i)}(t)$ is associated with a set of solutions $\left\{m_j^{(i)}(t)\right\}_{j=1}^{p}$ for a given parameterization of the system $\boldsymbol{\theta}^{(i)}$.

Whilst in practice we cannot sample the full function $p^{(i)}(t)$ we will assume that we can obtain points, $\mathbf{p}$, from the function that are so closely spaced relative to the time scale of $p^{(i)}(t)$ that we are not losing any information. However, because the prior over $p(t)$ is smooth then elements of the vector $\mathbf{p}$ that are close neighbors in time will be very strongly correlated. This can present a serious obstacle to efficient sampling in these systems. Under Markov chain Monte Carlo, for such a sample to be accepted it must respect the smoothness constraints imposed by the prior. However, random draws from the prior are highly unlikely to be accepted as they will result in values for $m_j(t)$ that do not match the data. The key challenge is to develop a sampling approach that respects the constraints imposed by the prior and can rapidly explore the space of values of $p(t)$ that are plausible under the data $\mathbf{y}$. With this task in mind a control point strategy for sampling from Gaussian processes was developed (Titsias et al., 2009, 2011).

[3]Laplace's approximation involves second order Taylor expansion around the mode of the posterior distribution.

3 Model Based Target Ranking

We used Gaussian process inference over the linear activation model in (1) to identify targets of two TFs, Mef2 and Twist, regulating mesoderm and muscle development in Drosophila (Honkela et al., 2010). These TFs are thought to be primarily regulated by differential expression at the mRNA level and therefore their measured mRNA expression levels are highly informative about their protein concentration in the nucleus. We therefore include a model of translation from the TF mRNA concentration $f(t)$ to the TF protein concentration $p(t)$:

$$\frac{\mathrm{d}p(t)}{\mathrm{d}t} = f(t) - \delta p(t) \tag{9}$$

where δ is the decay rate of the TF protein. The differential equation can be solved to give,

$$p(t) = \exp(-\delta t) \int_0^t f(v) \exp(\delta v) \mathrm{d}v$$

and we see that the TF protein concentration $p(t)$ is a linear function of the TF expression $f(t)$. There is also a linear relationship between the TF protein and the regulated target gene expression levels $m_j(t)$ under the linear activation model (recall (2)) and therefore $m_j(t)$ is also a linear function of both $p(t)$ and $f(t)$. We place a Gaussian process prior on $f(t)$ and the linear models of translation and activation define a joint Gaussian process over $m_j(t)$ and $p(t)$ as well as $f(t)$. If we use the squared exponential covariance in (6) for $f(t)$ then all the terms in the covariance function for the multivariate Gaussian process $\{f, p, m_j\}$ can be calculated analytically (Honkela et al., 2010).

The parameters of the TF mRNA covariance and the differential equation models parameterize the covariance function: $\boldsymbol{\theta} = [\delta, \alpha, \ell, \{b_j, s_j, d_j\}]$. These are estimated by maximising the likelihood by gradient-based optimisation (recall (8)). In this example we assume that $a_j = 0$ for all genes. After estimating the model parameters we use the likelihood as a score to rank genes according to their fit to the model. Before the final ranking weakly expressed genes need to be filtered because they often attain high likelihoods from any TF with an uninformative model. This can be accomplished using average z-scores computed using the variance information from PUMA preprocessing.

In Figure 3 we show examples of the model fit to the data for putative targets of the TF Twist. We fitted two different classes of models. Figure 3(a) shows examples where we fit independent models to three different genes. Since the models are fitted independently for each gene there is no reason for the models to infer a consistent TF protein concentration profile. We refer to this as the single-target model approach. While somewhat unrealistic, this approach is very attractive due to its computational advantage: fitting of independent models is thus trivially parallelisable. In Figure 3(b) we fit one model to the same three genes by sharing the TF protein profile across target genes. This conforms more with our belief that the TF protein profile should be consistent across targets. The model is therefore less flexible as can be observed in the example

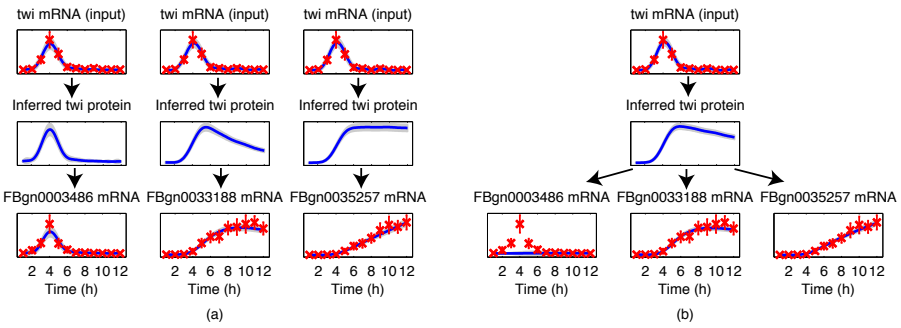


Figure 3: Examples of the model fit for two different classes of Gaussian process model fitted to potential targets of the TF Twist (from Honkela et al., 2010). (a) Three independent single-target models for likely targets. Red marks denote observed expression levels from Tomancak et al. (2002) with 2 s.d. error bars. The inferred posterior means of the functions are shown in blue and the shaded regions denote 2 s.d. posterior confidence intervals. (b) A joint multiple-target model for the same set of target genes as in (a). Note that the multiple-target models used in evaluation have more targets than the one shown here.

where gene FBgn0003486 cannot be fitted by the multiple-target model. The multiple-target model relies on the assumption that the whole set of targets considered are genuine. We used the five top-ranked targets identified by the single-target models as a "training set" in the multiple-target approach, adding in each putative target one-by-one to the training set and using the likelihood of the resulting six-target model for ranking. In practice one might also use additional biological prior knowledge to choose a small set of confident targets.

Figure 4 shows evaluation of the model-based ranking results using data from a genome-wide Chromatin-immunoprecipitation (ChIP-chip) experiment (Zinzen et al., 2009) to determine whether there is evidence of TF binding within 2000 base-pairs of a putative target. As well as showing results for the single-target and multiple-target Gaussian process methods, we also show the performance obtained using evidence of differential expression in mutant embryos (knock-outs), correlation with TF expression (correlation) and an alternative maximum likelihood model-based approach (quadrature, see Honkela et al., 2010, for details). In the global validation we score all genes showing significant variability in the expression data. In the focused evaluation we only consider genes with annotated expression in mesoderm or muscle tissue according to the *in situ* hybridization data (Tomancak et al., 2002).

We find that the model-based approach provides a significant advantage over the simpler correlation-based approach for Twist, whose targets display a diverse range of temporal profiles (Figure 3). For Mef2 the correlation-based approach works well. We observe that most of the predicted Mef2 targets have profiles very similar to Mef2 itself, in which case the model-based approach would not be expected to provide an advantage. The single-target Gaussian process approach

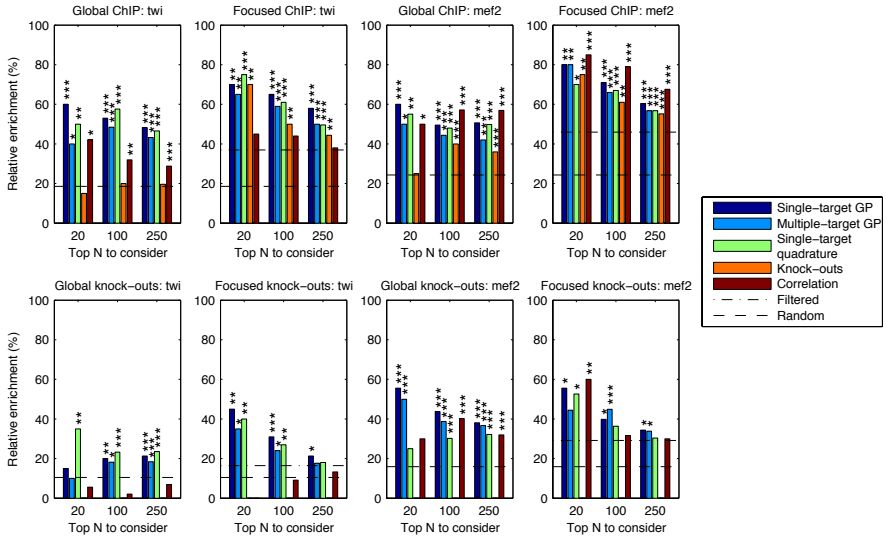


Figure 4: Evaluation results of different rankings (from Honkela et al., 2010) as discussed in the text showing the relative frequency of positive predictions among N top-ranking targets ("global" evaluations) or among N top genes with annotated expression in mesoderm or muscle tissue ("focused" evaluations). The dashed line denotes the frequency in the full population and the dash-dot line within the population considered in focused evaluation. The first row shows the frequency of targets with ChIP-chip binding within 2000 base pairs of the gene while the second row shows the frequency of predicted targets with significant differential expression in TF knock-outs. p-values of results significantly different from random are denoted by '***': $p < 0.001$, '**': $p < 0.01$, '*': $p < 0.05$. Comparison to knock-out ranking is obviously omitted for knock-out validation.

is found to out-perform other methods in most cases (Honkela et al., 2010).

4 Multiple Tanscription Factors

The models discussed in previous sections are based on the idealized assumption that the gene regulation process is driven by a single trascription factor. However, gene regulation in complex biological systems can involve several TFs. More precisely, to initiate trascription of a gene, multiple TFs may be required to simultaneously bind the related DNA sequence. Thus, in order to take into account the effect of multiple TFs, we need to generalize the basic single-TF model described previously. Furthermore, to deal with the non-negativity of the mRNA and protein TF functions, along with saturation effects that are naturally encountered in their dynamics, it is essential to assume biologically

plausible non-linearities.

The basic single-TF ODE model from Equation (1) can be generalized to include multiple TFs according to

$$\frac{\mathrm{d}m_j(t)}{\mathrm{d}t} = b_j + s_j G\left(p_1(t), \ldots, p_I(t); \mathbf{w}_j, w_{j0}\right) - d_j m_j(t), \tag{10}$$

where $p_i(t)$, $i = 1, \ldots, I$, are the TF protein functions. The function $G(\cdot)$ represents a non-linear response that allows the TFs to competitively or co-operatively activate or repress the transcription. A reasonable assumption for this non-linear response is to follow a sigmoidal form, similarly to the Michaelis-Menten and hill-climbing functions used in single input motifs (Alon, 2006). For instance, we can choose

$$G(p_1(t), \ldots, p_I(t); \mathbf{w}_j, w_{j0}) = \frac{1}{1 + e^{-w_{j0} - \sum_{i=1}^{I} w_{ji} \log p_i(t)}}, \tag{11}$$

which is the standard sigmoid function that takes values in $[0, 1]$ and recieves as inputs the logarithm of the TF activities. Here, w_{j0} is a real-valued bias parameter and the I-dimensional real-valued vector $\mathbf{w}_j = [w_{j1} \ldots w_{jI}]^T$ represents the interaction weights between the j^{th} target gene and the I TFs. These parameters quantify the strength of the network links between TFs and genes in the underlying regulatory network. Specifically, when $w_{ji} = 0$ the link between the j^{th} gene and the i^{th} TF is absent while when w_{ji} is negative or positive the TF acts as a repressor or activator respectively. Notice that by estimating the values of the interaction weights, we can infer the network links between TFs and genes. This can lead to a very flexible framework for target identification which generalizes the method described in section 3.

The multiple TF model can contain a much larger number of unknown parameters and unobserved protein functions compared to the single input motif model. Given the scarcity of the data, it is therefore essential that inference is carried out by a fully Bayesian approach using Markov chain Monte Carlo (MCMC). This poses significant computational challenges since we need to infer via sampling all the unknown quantities which are the protein functions and the remaining model parameters. We have developed an MCMC approach that infers the TF profiles from a small set of training data, associated with a moderate-sized network, and then it performs target identification at a genome-wide scale. This Bayesian approach is based on placing GP priors on the logged TF concentrations and suitable priors on the remaining model parameters (i.e. interaction weights and kinetic parameters) and then simulate from the posterior distribution by applying suitable Metropolis-Hastings updates.

When observations of the TF mRNAs are available, the above framework can be combined with the protein translation ODE model presented in section 3. This can further facilitate the estimation of the TF profiles and resolve identifiability problems caused by the scarcity of the data and experimental conditions.

We applied the multiple TF dynamical model to an artifical dataset consisting of time-series mRNA mesearuments associated with 1000 genes and two

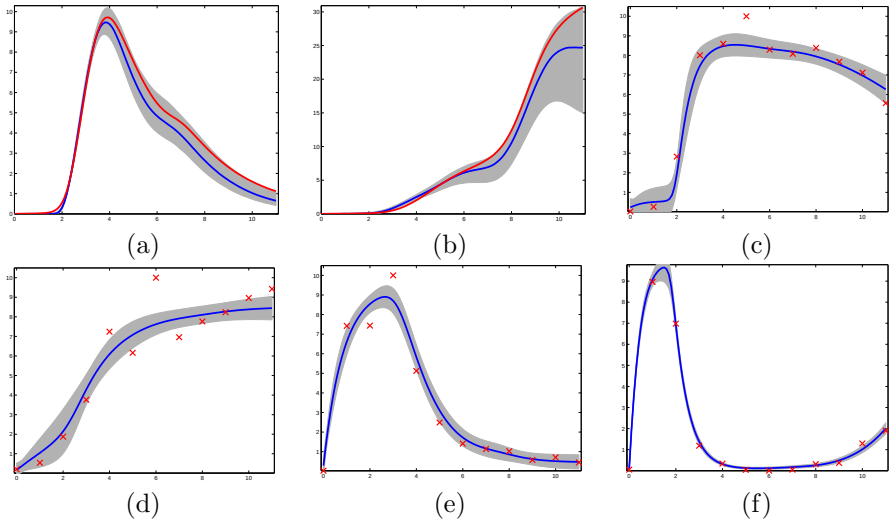


Figure 5: Panels (a) and (b) show the inferred TFs. The estimated means are plotted as the blue solid lines and the shaded areas represent 95% uncertainty around the estimated means. The red solid lines show the ground-truth TFs that generated the data. Panels (c)-(f) show several examples of how the multiple TF model predicts gene mRNA functions. The red crosses denote observed measurements, while blue lines and shaded areas represented noisy-free (without adding the observation noise from the likelihood) Bayesian model predictions of the mRNA functions.

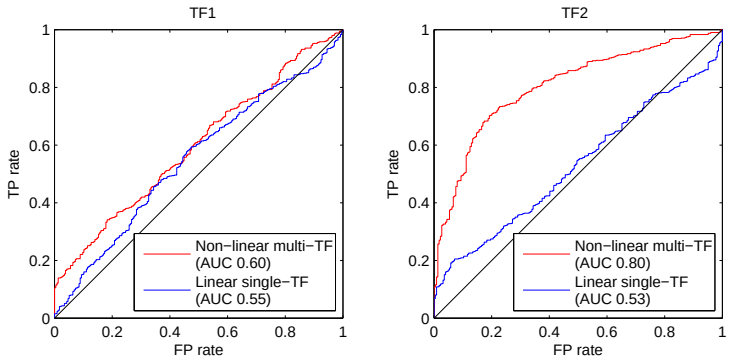


Figure 6: ROC curves for predicting the network connections in the synthetic data. The left panel shows the performance when predicting the first TF, while the right panel shows performance for predicting the second TF. Red curves show the results by using the multiple TF model and blue lines the results by using the linear single-TF model. In both plots, random prediction is represented by the diagonal black line.

TFs. The network links between TFs and genes are unknown apart from a small set of 20 genes which are assumed to have known connections. This small set of genes is used to infer the TF concentration functions. Figure 5(a) and (b) show the inferred TF profiles where blue lines represent the inferred means, shaded areas represent 95% uncertainty and the red lines are the actual profiles that generated the data. Figure 5(c)-(f) give several examples on how the model fits the mRNA expression data. Specifically, the panel in (c) shows a gene that is activated only by the first TF while the second TF is inactive. Panel (d) and (e) show two genes that are regulated (activated in (d) and repressed in (e)) only by the second TF. The last panel (f) shows a gene that is jointly regulated by the two TFs so that it is repressed by the first TF and activated by the second one. Figure 6 shows ROC curves of predictive performance for identifying the network links between the two TFs and the 980 test genes. The performance of a linear single-TF model is also displayed and is clearly inferior to the performance of the multiple TF model.

5 Conclusion

The difficulty of measuring every biochemical species involved in cellular interaction networks will mean that we will often be faced with the problem of missing functions. In this chapter we have reviewed an elegant approach to dealing with such missing functions, we started with parameteric generalized linear models for dealing with the missing functions, and extended them through Bayesian treatments and considering the limit of infinite basis functions to arrive at sGaussian process models. We reviewed a simple model of transcription and translation and showed how it can be used to rank putative targets of transcription factors. The Gaussian process framework provides an elegant framework for analytically dealing with missing functions if the differential equation responds in a linear way to the driving function. However, if we consider nonlinear responses, which we must necessarily do if we wish to constrain the concentration of the missing functions to be positive, the analytic approach is no longer tractable. We reviewed an efficient sampling based approach that allows us to consider non linear responses and can be applied to the situation where there are multiple transcription factors regulating the targets of interest. Both approaches can be applied for genomewide ranking of putative targets.xs Together these approaches present a powerful set of diagnostic tools for unravelling the transcriptional interactions that underpin a biological time series.xs

References

Uri Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman and Hall/CRC, London, 2006. ISBN 1-58488-642-0.

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*, 7(3):R25, 2006.

Pei Gao, Antti Honkela, Magnus Rattray, and Neil D. Lawrence. Gaussian process modelling of latent chemical species: Applications to inferring transcription factor activities. *Bioinformatics*, 24:i70–i75, 2008. doi: 10.1093/bioinformatics/btn278.

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. *Proc. Natl. Acad. Sci. USA*, 107(17):7793–7798, Apr 2010. doi: 10.1073/pnas.0914285107.

Raya Khanin, Veronica Viciotti, and Ernst Wit. Reconstructing repressor protein levels from expression of gene targets in *E. Coli*. *Proc. Natl. Acad. Sci. USA*, 103(49):18592–18596, 2006. doi: 10.1073/pnas.0603390103.

Neil D. Lawrence and Magnus Rattray. A brief introduction to Bayesian inference. In Lawrence et al. (2010a), chapter 5. ISBN 0-262-01386-X.

Neil D. Lawrence, Guido Sanguinetti, and Magnus Rattray. Modelling transcriptional regulation using Gaussian processes. In Bernhard Schölkopf, John C. Platt, and Thomas Hofmann, editors, *Advances in Neural Information Processing Systems*, volume 19, pages 785–792, Cambridge, MA, 2007. MIT Press.

Neil D. Lawrence, Mark Girolami, Magnus Rattray, and Guido Sanguinetti, editors. *Learning and Inference in Computational Systems Biology*. MIT Press, Cambridge, MA, 2010a. ISBN 0-262-01386-X.

Neil D. Lawrence, Magnus Rattray, Pei Gao, and Michalis K. Titsias. Gaussian processes for missing species in biochemical systems. In Lawrence et al. (2010a), chapter 9. ISBN 0-262-01386-X.

Xuejun Liu, Marta Milo, Neil D. Lawrence, and Magnus Rattray. A tractable probabilistic model for Affymetrix probe-level analysis across multiple chips. *Bioinformatics*, 21(18):3637–3644, 2005. doi: 10.1093/bioinformatics/bti583.

David J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, Cambridge, U.K., 2003. ISBN 0-52164-298-1.

Marta Milo, Alireza Fazeli, Mahesan Niranjan, and Neil D. Lawrence. A probabilistic model for the extraction of expression levels from oligonucleotide arrays. *Biochemical Transations*, 31(6):1510–1512, 2003.

Martin F. Møller. A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, 6(4):525–533, 1993.

Anthony O'Hagan. Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society, B*, 40:1–42, 1978.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006. ISBN 0-262-18253-X.

Simon Rogers, Raya Khanin, and Mark Girolami. Model based identification of transcription factor activity from microarray data. In *Probabilistic Modeling and Machine Learning in Structural and Systems Biology*, Tuusula, Finland, 17-18th June 2006.

Guido Sanguinetti, Neil D. Lawrence, and Magnus Rattray. Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities. *Bioinformatics*, 22(22):2275–2281, 2006a. doi: 10.1093/bioinformatics/btl473.

Guido Sanguinetti, Magnus Rattray, and Neil D. Lawrence. A probabilistic dynamical model for quantitative inference of the regulatory mechanism of transcription. *Bioinformatics*, 22(14):1753–1759, 2006b. doi: 10.1093/bioinformatics/btl154.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2001.

Michalis K. Titsias, Neil D. Lawrence, and Magnus Rattray. Efficient sampling for Gaussian process inference using control variables. In Daphne Koller, Dale Schuurmans, Yoshua Bengio, and Leon Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21, pages 1681–1688, Cambridge, MA, 2009. MIT Press.

Michalis K. Titsias, Magnus Rattray, and Neil D. Lawrence. Markov chain monte carlo algorithms for gaussian processes. In David Barber, A. Taylan Cemgil, and Silvia Chiappa, editors, *Bayesian Time Series Models*, chapter 14. Cambridge University Press, 2011. ISBN 9780521196765.

P. Tomancak, A. Beaton, R. Weiszmann, E. Kwan, S. Shu, S. E. Lewis, S. Richards, M. Ashburner, V. Hartenstein, S. E. Celniker, and G. M. Rubin. Systematic determination of patterns of gene expression during Drosophila embryogenesis. *Genome Biology*, 3(12):RESEARCH0088, 2002.

Christopher K. I. Williams. Regression with Gaussian processes. In S. W. Ellacott, J. C. Mason, and I. J. Anderson, editors, *Mathematics of Neural Networks: Models, Algorithms and Applications*. Kluwer, Dordrecht, The Netherlands, 1997. Paper presented at the *Mathematics of Neural Networks and Applications conference*, Oxford, UK, July 1995.

Ciyou Zhu, Richard H. Byrd, and Jorge Nocedal. L-BFGS-B: Algorithm 778: L-BFGS-B, FORTRAN routines for large scale bound constrained optimization. *ACM Transactions on Mathematical Software*, 23(4):550–560, 1997.

Robert P Zinzen, Charles Girardot, Julien Gagneur, Martina Braun, and Eileen E M Furlong. Combinatorial binding predicts spatio-temporal cis-regulatory

activity. *Nature*, 462(7269):65–70, Nov 2009. doi: 10.1038/nature08531. URL http://dx.doi.org/10.1038/nature08531.